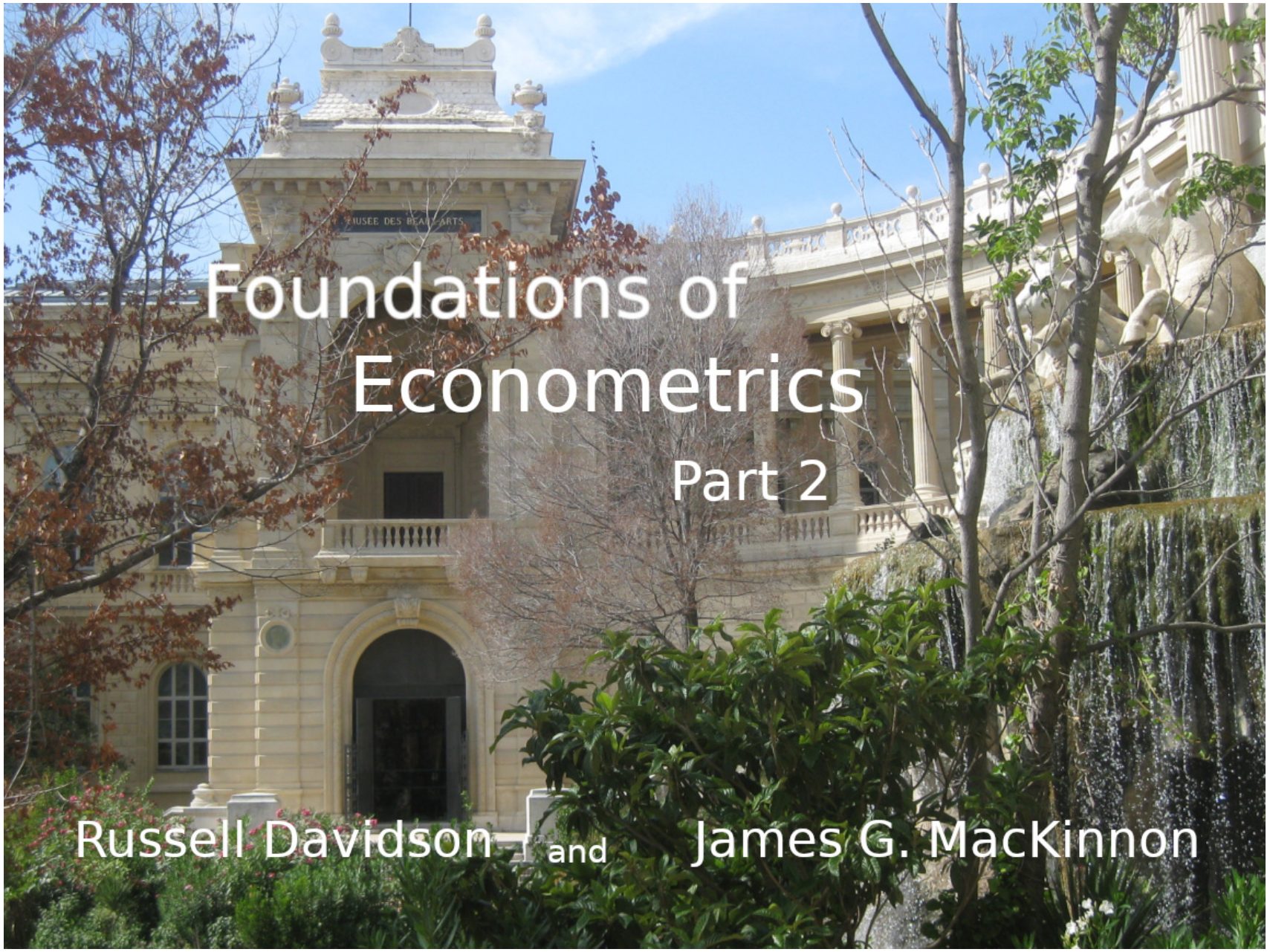


Foundations of Econometrics

Part 2

Russell Davidson and James G. MacKinnon

Contents

Preface

This book is part two of an updated and abbeviated version of our 2004 textbook *Econometric Theory and Methods* (ETM). A plan to create a full second edition of that book never came to fruition, but the first several chapters of the book have served both of us well, not only as a text for a first, one-term, graduate course, but also for the Honours course in econometrics at McGill University. But even in those early chapters, there have been more and more things, over the years since the book was published, that we wished to update and change. In the first book, of which this book is the second part, we included only those chapters actually used in our one-term courses. For this book, four more chapters are included, duly updated, for the purposes of a graduate course that Davidson teaches at the Aix-Marseille School of Economics (AMSE). Some remarks follow, taken from the preface to the first book, intended to provide information on how best to use this second book.

Some of the exercises provided at the end of each chapter are really quite challenging, as we discovered many years ago while preparing solutions to them. These exercises are starred, as are a number of other exercises for which we think that the solutions are particularly illuminating, even if they are not especially difficult. In some cases, these starred exercises allow us to present important results without proving them in the text. In other cases, they are designed to allow instructors to cover advanced material that is not in the text itself. Because the solutions to the starred exercises should be of considerable value to students, they are available from the website for ETM. All the data needed for the exercises are also available from the website, although these are necessarily not at all recent. Instructors might prefer to ask students to go online themselves to find more recent data that they can use instead of the older data on the website.

There are several types of exercises, intended for different purposes. Some of the exercises are empirical, designed to give students the opportunity to become familiar with a variety of practical econometric methods. Others involve simulation, including some that ask students to conduct small Monte Carlo experiments. Many are fairly straightforward theoretical exercises that good students should find illuminating and, we hope, not too difficult. Although most of the exercises have been taken over unchanged from ETM, some have been modified, and a fair number of new exercises introduced.

An instructor's manual was provided for ETM, with solutions to all the exercises of that book. It can be found online by anyone willing to spend a little time with Google. In a sense this is a shame, as it means that instructors can no longer safely use exercises given here for exams or assignments, since some students may be tempted to copy the solutions from the manual. We fear that this is likely to be a problem that university instructors will have

to face more and more frequently, even if most instruction is no longer being given online. However, for our purposes here, the most important point is that solutions for the starred exercises are readily available without access to Google or any other search engine.

Organization

In the first book of this update of parts of ETM, no nonlinear models are considered. Some of these are given a reasonably full treatment here. The first chapter deals with nonlinear least squares (NLS), the simplest of the nonlinear models discussed here. As in part 1, much use is made of estimating functions and estimating equations. The asymptotic properties of NLS are developed in this chapter, and serve as a model for those of the other nonlinear models dealt with in later chapters. Similarly, attention is given to the methods of estimation used for nonlinear models, which, being iterative, are quite different from the methods presented in part 1 for linear models. Artificial regressions play a large part in nonlinear estimation, and it is here that the Gauss-Newton regression is presented. It too serves as a model for the other artificial regressions found in later chapters. Inference can be based either on asymptotic methods or the bootstrap – both are treated here.

The topic of Chapter 2 is the Generalized Method of Moments (GMM). This very general technique of estimation, and, to a lesser extent, also of inference, was developed much more recently than the classical method of Maximum Likelihood (ML), and so it might seem appropriate to discuss ML before GMM. We chose to do things in the opposite order, because some of the classical theory of ML is specific to that technique, while the theory and methods of GMM are much more general. Nonetheless, we begin by developing the linear regression model, as seen from the viewpoint of GMM, before looking at other, nonlinear, models. In this chapter, the very intimate connection between GMM and linear models estimated by use of instrumental variables (IV) is presented. The concept of over-identifying restrictions is seen to be as important in this more general context as with linear IV models, and several methods for testing these restrictions are presented. It is in this chapter, too, that inference robust to heteroskedasticity, or to that and serial correlation, is explained for nonlinear models.

Maximum Likelihood is the topic of Chapter 3. As with GMM, the starting point here is the linear regression model, for which the distribution of the disturbances must be specified more completely than for least squares. Here, we make the usual assumption of normal disturbances. After a brief exposition of the asymptotic properties of ML, we present the three classical tests associated with ML, likelihood ratio, Wald, and Lagrange Multiplier. Some special topics broached in this chapter are how the first observations in an autoregressive time-series model can be handled by ML, and how ML can handle transformations of the dependent variable in regression models.

Chapter 4 discusses some examples of models with discrete, or otherwise limited, dependent variables. The first of these models are binary-choice, or binary-response, models, where the most frequently used are the Probit and Logit. Estimation and inference for these are treated in this chapter, and the artificial regression suitable for use with binary-response models is presented. The rest of the chapter discusses other models with limited dependent variables: models with more than two discrete responses, models with count data, with censored or truncated data, with duration data. Another important topic found in this chapter is that of sample selectivity.

Even with this second part, the abbreviated version of the original ETM cannot treat more than a limited number of topics of current interest to econometricians. We may hope, however, that what is presented in these two books is enough for students to embark on study of the recent research literature, and to find research problems of their own.

Acknowledgements

It took us nearly six years to write ETM, the principal source for this book, and during that time we received assistance and encouragement from a large number of people. Bruce McCullough, of Drexel University, read every chapter, generally at an early stage, and made a great many valuable comments and suggestions. Thanasis Stengos, of the University of Guelph, also read the entire book and made several suggestions that have materially improved the final version. Richard Parks, of the University of Washington, taught out of the unfinished manuscript and found an alarming number of errors that were subsequently corrected. Others who read at least part of the book with care and provided us with useful feedback include John Galbraith (McGill University), the late Ramazan Gencay (University of Windsor), Sílvia Gonçalves (McGill University), Manfred Jäger (Martin Luther University), Richard Startz (University of Washington), Arthur Sweetman (McMaster University), and Tony Wirjanto (University of Waterloo). There are numerous other colleagues who also deserve our thanks, but the list is much too long to include here.

We made use of draft chapters of ETM in courses at both the Master's and Ph.D. levels at Queen's University, McGill University, and the University of Toronto in Canada, as well as at Aix-Marseille Université and AMSE in France. A great many students found errors or pointed out aspects of the exposition that were unclear. The book has been improved enormously by addressing their comments, and we are very grateful to all of them. We are also grateful to the many students at the above institutions, and also at the University of Washington, who expressed enthusiasm for the book when it was still a long way from being finished and thereby encouraged us to finish a task that, at times, seemed almost incapable of completion.

This book and its predecessor have seen much less use than has ETM as of this writing (January 2022), but were in heavy use for online teaching at McGill and at AMSE. This experience has led to numerous improvements in the exposition, for many of which we thank the students who were taking the course online.

We also owe a debt of gratitude to the thousands of talented programmers who have contributed to the development of free, open-source, software, without which it would have been much more difficult to write this book. The book was typeset entirely by us using the TEX LIVE distribution of TEX running on the Debian distribution of the Linux operating system. We also made extensive use of the gcc and g77 compilers, and of many other excellent free programs that run on Linux.

Notation

We have tried our best to use a consistent set of notation throughout the book. It has not always been possible to do so, but below we list most of the notational conventions used in the book.

n	sample size; number of observations
k	number of regressors
l	number of instrumental variables
$\boldsymbol{A}$ (upper-case bold letter)	a matrix or row vector
$\boldsymbol{a}$ (lower-case bold letter)	a column vector
$\boldsymbol{y}$	dependent variable; regressand
$\boldsymbol{X}$	matrix of explanatory variables; regressors
$\boldsymbol{W}$	matrix of instrumental variables
$\boldsymbol{\beta}$	vector of regression parameters
$\hat{\boldsymbol{\beta}}$	vector of estimated parameters
$\tilde{\boldsymbol{\beta}}$	vector of parameters estimated under restrictions
$\boldsymbol{u}$	vector of disturbances
σ^2	variance (usually of disturbances)
F	cumulative distribution function (CDF)
f	probability density function
Φ	CDF of standard normal distribution N(0,1)
ϕ	density of standard normal distribution
$\mathbf{I}$	identity matrix
$\mathbf{0}$	column vector of zeros
$\mathbf{O}$	matrix of zeros
$\boldsymbol{\iota}$	column vector of ones
$\mathcal{S}(\boldsymbol{X})$	linear span of the columns of $\boldsymbol{X}$
$\mathcal{S}^{\perp}(\boldsymbol{X})$	orthogonal complement of $\mathcal{S}(\boldsymbol{X})$
$\boldsymbol{P}_{\boldsymbol{X}}$	orthogonal projection matrix on to $\mathcal{S}(\boldsymbol{X})$
$\boldsymbol{M}_{\boldsymbol{X}}$	complementary orthogonal projection; $\boldsymbol{M}_{\boldsymbol{X}} = \mathbf{I} - \boldsymbol{P}_{\boldsymbol{X}}$
μ	a data-generating process (DGP)
$\mathbb{M}$	model, a set of DGPs
$\mathbb{R}$	the real line
$\mathbb{R}^n$	set of n–vectors
E^n	n–dimensional Euclidean space
E	expectation operator
Var	a variance or a covariance matrix
$\boldsymbol{\Omega}$	a covariance matrix
$\boldsymbol{\Gamma}(j)$	autocovariance matrix at lag j
L	lag operator
$\overset{a}{=}$	asymptotic equality
$\overset{a}{\sim}$	asymptotically distributed as
$\xrightarrow{d}$	convergence in distribution

Data, Solutions, and Corrections

The website for ETM, and for this book, is located at

http://qed.econ.queensu.ca/ETM/

This website provides all the data needed for the exercises, solutions to the starred exercises, and corrections made since the book was printed. The solutions and corrections are provided as PDF files, which almost all modern computers should have the software to view and print.

The website provides solutions only to the starred exercises. In addition, there is an instructor's manual, available to instructors only, which comes on a CD-ROM and contains solutions to all the exercises. For information about how to obtain it, please contact your local Oxford University Press sales representative or visit the Oxford higher education website at

http://www.oup-usa.org/highered

The authors are happy to hear from readers who have found errors that should be corrected. Their affiliations are given below:

Russell Davidson
<russell.davidson@mcgill.ca>

Aix-Marseille Université
CNRS, EHESS, AMSE
13205 Marseille cedex 01
France

Department of Economics
McGill University
855 Sherbrooke St. West
Montreal, Quebec, H3A 2T7
Canada

James G. MacKinnon
<jgm@econ.queensu.ca>

Department of Economics
Queen's University
Kingston, Ontario, K7L 3N6
Canada

Chapter 1

Nonlinear Regression

1.1 Introduction

For each observation t of any regression model, there is an information set Ω_t and a suitably chosen vector $\boldsymbol{X}_t$ of explanatory variables that belong to Ω_t. A linear regression model consists of all DGPs for which the expectation of the dependent variable y_t conditional on Ω_t can be expressed as a *linear* combination $\boldsymbol{X}_t\boldsymbol{\beta}$ of the components of $\boldsymbol{X}_t$, and for which the disurbances satisfy suitable requirements, such as being IID. Since the elements of $\boldsymbol{X}_t$ may be nonlinear functions of the variables originally used to define Ω_t, many types of nonlinearity can be handled within the framework of the linear regression model. However, many other types of nonlinearity cannot be handled within this framework. In order to deal with them, we often need to estimate **nonlinear regression models**. These are models for which $\mathrm{E}(y_t \mid \Omega_t)$ is a nonlinear function of the *parameters*.

A typical nonlinear regression model can be written as

$$y_t = x_t(\boldsymbol{\beta}) + u_t, \quad u_t \sim \mathrm{IID}(0, \sigma^2), \quad t = 1, \dots, n, \tag{1.01}$$

where, just as for the linear regression model, y_t is the t^{th} observation on the dependent variable, and $\boldsymbol{\beta}$ is a k–vector of parameters to be estimated. The scalar function $x_t(\boldsymbol{\beta})$ is a **nonlinear regression function**. It determines the expectation of y_t conditional on Ω_t, which is made up of some set of explanatory variables. These explanatory variables, which may include lagged values of y_t as well as exogenous variables, are not shown explicitly in (1.01). However, the t subscript of $x_t(\boldsymbol{\beta})$ indicates that the regression function varies from observation to observation. This variation usually occurs because $x_t(\boldsymbol{\beta})$ depends on explanatory variables, but it can also occur because the functional form of the regression function actually changes over time. The number of explanatory variables, all of which must belong to Ω_t, need not be equal to k.

The disturbances in (1.01) are specified to be IID. By this, we mean something very similar to, but not precisely the same as, the two conditions in (F5.46). In order for the disturbances to be identically distributed, the distribution of each disturbance u_t, conditional on the corresponding information set Ω_t, must be the same for all t. In order for them to be independent, the distribution of u_t, conditional not only on Ω_t but also on all the other disturbances, should be

the same as its distribution conditional on Ω_t alone, without any dependence on the other disturbances.

Another way to write the nonlinear regression model (1.01) is

$$\boldsymbol{y} = \boldsymbol{x}(\boldsymbol{\beta}) + \boldsymbol{u}, \quad \boldsymbol{u} \sim \text{IID}(\boldsymbol{0}, \sigma^2 \mathbf{I}), \tag{1.02}$$

where $\boldsymbol{y}$ and $\boldsymbol{u}$ are n–vectors with typical elements y_t and u_t, respectively, and $\boldsymbol{x}(\boldsymbol{\beta})$ is an n–vector of which the t^{th} element is $x_t(\boldsymbol{\beta})$. Thus $\boldsymbol{x}(\boldsymbol{\beta})$ is the nonlinear analog of the vector $\boldsymbol{X\beta}$ in the linear case.

As a very simple example of a nonlinear regression model, consider the model

$$y_t = \beta_1 + \beta_2 z_{t1} + \frac{1}{\beta_2} z_{t2} + u_t, \quad u_t \sim \text{IID}(0, \sigma^2), \tag{1.03}$$

where z_{t1} and z_{t2} are explanatory variables. For this model,

$$x_t(\boldsymbol{\beta}) = \beta_1 + \beta_2 z_{t1} + \frac{1}{\beta_2} z_{t2}.$$

Although the regression function $x_t(\boldsymbol{\beta})$ is linear in the explanatory variables, it is nonlinear in the parameters, because the coefficient of z_{t2} is constrained to equal the inverse of the coefficient of z_{t1}. In practice, many nonlinear regression models, like (1.03), can be expressed as linear regression models in which the parameters must satisfy one or more nonlinear restrictions.

The Linear Regression Model with AR(1) Disturbances

We now consider a particularly important example of a nonlinear regression model that is also a linear regression model subject to nonlinear restrictions on the parameters. This arises in connection with the phenomenon of **serial correlation**, in which nearby disturbances in a regression model are (or appear to be) correlated. Serial correlation is very commonly encountered in applied work using time-series data, and many techniques for dealing with it have been proposed. One of the simplest and most popular ways of dealing with serial correlation is to assume that the disturbances follow the **first-order autoregressive**, or **AR(1)**, process

$$u_t = \rho u_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim \text{IID}(0, \sigma_\varepsilon^2), \quad |\rho| < 1. \tag{1.04}$$

According to this model, the disturbance at time t is equal to ρ times the disturbance at time $t-1$, plus a new disturbance ε_t. The vector $\boldsymbol{\varepsilon}$ with typical component ε_t satisfies the IID condition we discussed above. This condition is enough for ε_t to be an innovation. Thus the ε_t are homoskedastic and independent of all past and future innovations. We see from (1.04) that, in each period, part of the disturbance u_t is the previous period's disturbance, shrunk somewhat toward zero and possibly changed in sign, and part is the

innovation ε_t. We discussed serial correlation, including the AR(1) process and other autoregressive processes, in Chapter 9 of Part 1. At present, we are concerned solely with the nonlinear regression model that results when the disturbances of a linear regression model are assumed to follow an AR(1) process.

If we combine (1.04) with the linear regression model

$$y_t = \boldsymbol{X}_t\boldsymbol{\beta} + u_t \tag{1.05}$$

by substituting $\rho u_{t-1} + \varepsilon_t$ for u_t and then replacing u_{t-1} by $y_{t-1} - \boldsymbol{X}_{t-1}\boldsymbol{\beta}$, we obtain the nonlinear regression model

$$y_t = \rho y_{t-1} + \boldsymbol{X}_t\boldsymbol{\beta} - \rho \boldsymbol{X}_{t-1}\boldsymbol{\beta} + \varepsilon_t, \quad \varepsilon_t \sim \text{IID}(0, \sigma_\varepsilon^2). \tag{1.06}$$

Since the lagged dependent variable y_{t-1} appears among the regressors, this is a **dynamic** model. As with the other dynamic models that are treated in the exercises, we have to drop the first observation, because y_0 and $\boldsymbol{X}_0$ are assumed not to be available. The model is linear in the regressors but nonlinear in the parameters $\boldsymbol{\beta}$ and ρ, and it therefore needs to be estimated by nonlinear least squares or some other nonlinear estimation method.

In the next section, we study estimators for nonlinear regression models generated by the method of moments, and we establish conditions for asymptotic identification, asymptotic normality, and asymptotic efficiency. Then, in Section 1.3, we show that, under the assumption that the disturbances are IID, the most efficient estimator is **nonlinear least squares**, or **NLS**. In Section 1.4, we discuss various methods by which NLS estimates may be computed. The method of choice in most circumstances is some variant of Newton's Method. One commonly-used variant is based on an artificial linear regression called the Gauss-Newton regression. We introduce this artificial regression in Section 1.5 and show how to use it to compute NLS estimates and estimates of their covariance matrix. In Section 1.6 we introduce the important concept of one-step estimation. Then, in Section 1.7, we show how to use the Gauss-Newton regression to compute hypothesis tests.

1.2 Estimating Equations for Nonlinear Models

The OLS estimator for linear models may be derived by using the fact that, for each observation, the expectation of the disturbance in the regression model is zero conditional on the vector of explanatory variables. This implied that

$$\text{E}(\boldsymbol{X}_t u_t) = \text{E}\big(\boldsymbol{X}_t(y_t - \boldsymbol{X}_t\boldsymbol{\beta})\big) = \boldsymbol{0}. \tag{1.07}$$

The sample analog of the middle expression here is $n^{-1}\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X\beta})$. Setting this to zero and ignoring the factor of n^{-1}, we obtained the vector of **estimating equations**

$$\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X\beta}) = \boldsymbol{0}, \tag{1.08}$$

and these conditions were easily solved to yield the OLS estimator $\hat{\boldsymbol{\beta}}$. We now want to employ the same type of argument for nonlinear models.

An information set Ω_t is typically characterized by a set of variables that belong to it. But, since the realization of any deterministic function of these variables is known as soon as the variables themselves are realized, Ω_t must contain not only the variables that characterize it but also all deterministic functions of them. As a result, an information set Ω_t contains precisely those variables which are equal to their expectations conditional on Ω_t. In Exercise 1.1, readers are asked to show that the conditional expectation of a random variable is also its expectation conditional on the set of all deterministic functions of the conditioning variables.

For the nonlinear regression model (1.01), the disturbance u_t has expectation 0 conditional on all variables in Ω_t. Thus, if $\boldsymbol{W}_t$ denotes any $1 \times k$ vector of which all the components belong to Ω_t,

$$\mathrm{E}(\boldsymbol{W}_t u_t) = \mathrm{E}\Big(\boldsymbol{W}_t\big(y_t - x_t(\boldsymbol{\beta})\big)\Big) = \mathbf{0}. \tag{1.09}$$

Just as the estimating equations that correspond to (1.07) are (1.08), the estimating equations that correspond to (1.09) are

$$\boldsymbol{W}^\top\big(\boldsymbol{y} - \boldsymbol{x}(\boldsymbol{\beta})\big) = \mathbf{0}, \tag{1.10}$$

where $\boldsymbol{W}$ is an $n \times k$ matrix with typical row $\boldsymbol{W}_t$. There are k nonlinear equations in (1.10). These equations can, in principle, be solved to yield an estimator of the k–vector $\boldsymbol{\beta}$. Geometrically, the estimating equations (1.10) require that the vector of residuals should be orthogonal to all the columns of the matrix $\boldsymbol{W}$.

How should we choose $\boldsymbol{W}$? There are infinitely many possibilities. Using almost any matrix $\boldsymbol{W}$, of which the t^{th} row depends only on variables that belong to Ω_t, and which has full column rank k asymptotically, yields a consistent estimator of $\boldsymbol{\beta}$. However, these estimators in general have different asymptotic covariance matrices, and it is therefore of interest to see if any particular choice of $\boldsymbol{W}$ leads to an estimator with smaller asymptotic variance than the others. Such a choice would then lead to an efficient estimator, judged by the criterion of the asymptotic variance.

Identification and Asymptotic Identification

Let us denote by $\hat{\boldsymbol{\beta}}$ the estimator defined implicitly by (1.10). In order to show that $\hat{\boldsymbol{\beta}}$ is consistent, we must assume that the parameter vector $\boldsymbol{\beta}$ in the model (1.01) is **asymptotically identified**. In general, a vector of parameters is said to be **identified** by a given data set and a given estimation method if, for that data set, the estimation method provides a unique way to determine the parameter estimates. In the present case, $\boldsymbol{\beta}$ is identified by a given data set if equations (1.10) have a unique solution.

For the parameters of a model to be asymptotically identified by a given estimation method, we require that the estimation method should provide a unique way to determine the parameter estimates in the limit as the sample size n tends to infinity. In the present case, asymptotic identification can be formulated in terms of the probability limit of the vector $n^{-1}\boldsymbol{W}^\top\big(\boldsymbol{y} - \boldsymbol{x}(\boldsymbol{\beta})\big)$ as $n \to \infty$. Suppose that the true DGP is a special case of the model (1.02) with parameter vector $\boldsymbol{\beta}_0$. Then we have

$$\frac{1}{n}\boldsymbol{W}^\top\big(\boldsymbol{y} - \boldsymbol{x}(\boldsymbol{\beta}_0)\big) = \frac{1}{n}\sum_{t=1}^{n}\boldsymbol{W}_t^\top u_t. \tag{1.11}$$

By (1.09), every term in the sum above has expectation 0, and the IID assumption in (1.02) is enough to allow us to apply a law of large numbers to that sum. It follows that the right-hand side, and therefore also the left-hand side, of (1.11) tends to zero in probability as $n \to \infty$.

Let us now define the k–vector of deterministic functions $\boldsymbol{\alpha}(\boldsymbol{\beta})$ as

$$\boldsymbol{\alpha}(\boldsymbol{\beta}) = \operatorname*{plim}_{n\to\infty}\frac{1}{n}\boldsymbol{W}^\top\big(\boldsymbol{y} - \boldsymbol{x}(\boldsymbol{\beta})\big), \tag{1.12}$$

where we continue to assume that $\boldsymbol{y}$ is generated by the model (1.02) with $\boldsymbol{\beta} = \boldsymbol{\beta}_0$. Since a law of large numbers can be applied to the right-hand side of equation (1.12) whatever the value of $\boldsymbol{\beta}$, the components of $\boldsymbol{\alpha}$ are deterministic. In the preceding paragraph, we explained why $\boldsymbol{\alpha}(\boldsymbol{\beta}_0) = \mathbf{0}$. The parameter vector $\boldsymbol{\beta}$ is asymptotically identified if $\boldsymbol{\beta}_0$ is the *unique* solution to the equations $\boldsymbol{\alpha}(\boldsymbol{\beta}) = \mathbf{0}$, that is, if $\boldsymbol{\alpha}(\boldsymbol{\beta}) \neq \mathbf{0}$ for all $\boldsymbol{\beta} \neq \boldsymbol{\beta}_0$.

Although most parameter vectors that are identified by data sets of reasonable size are also asymptotically identified, neither of these concepts implies the other. It is possible for an estimator to be asymptotically identified without being identified by many data sets, and it is possible for an estimator to be identified by every data set of finite size without being asymptotically identified. To see this, consider the following two examples.

As an example of the first possibility, suppose that $y_t = \beta_1 + \beta_2 z_t$, where z_t is a random variable which follows the **Bernoulli distribution**. Such a random variable is often called a **binary variable**, because there are only two possible values it can take on, 0 and 1. The probability that $z_t = 1$ is p, and so the probability that $z_t = 0$ is $1 - p$. If p is small, there could easily be samples of size n for which every z_t was equal to 0. For such samples, the parameter β_2 cannot be identified, because changing β_2 can have no effect on $y_t - \beta_1 - \beta_2 z_t$. However, provided that $p > 0$, both parameters are identified asymptotically. As $n \to \infty$, a law of large numbers guarantees that the proportion of the z_t that are equal to 1 tends to p.

As an example of the second possibility, consider the following model

$$y_t = \beta_1 + \beta_2\frac{1}{t} + u_t, \tag{1.13}$$

where t is a time trend. The OLS estimators of β_1 and β_2 can, of course, be computed for any finite sample of size at least 2, and so the parameters are identified by any data set with at least 2 observations. But β_2 is not identified asymptotically. Suppose that the true parameter values are β_1^0 and β_2^0. Let us use the two regressors for the variables in the information set Ω_t, so that $\boldsymbol{W}_t = [1 \;\; 1/t]$ and the Z-estimator is the same as the OLS estimator. Then, using the definition (1.12), we obtain

$$\boldsymbol{\alpha}(\beta_1, \beta_2) = \operatorname*{plim}_{n\to\infty} \begin{bmatrix} n^{-1}\sum_{t=1}^n \big((\beta_1^0 - \beta_1) + 1/t(\beta_2^0 - \beta_2) + u_t\big) \\ n^{-1}\sum_{t=1}^n \big(1/t(\beta_1^0 - \beta_1) + 1/t^2(\beta_2^0 - \beta_2) + 1/t u_t\big) \end{bmatrix}. \tag{1.14}$$

It is known that the deterministic sums $n^{-1}\sum_{t=1}^n (1/t)$ and $n^{-1}\sum_{t=1}^n (1/t^2)$ both tend to 0 as $n \to \infty$. Further, the law of large numbers tells us that the limits in probability of $n^{-1}\sum_{t=1}^n u_t$ and $n^{-1}\sum_{t=1}^n (u_t/t)$ are both 0. Thus the right-hand side of (1.14) simplifies to

$$\boldsymbol{\alpha}(\beta_1, \beta_2) = \begin{bmatrix} \beta_1^0 - \beta_1 \\ 0 \end{bmatrix}.$$

Since $\boldsymbol{\alpha}(\beta_1, \beta_2)$ vanishes for $\beta_1 = \beta_1^0$ and for any value of β_2 whatsoever, we see that β_2 is not asymptotically identified. It can be shown that, although the OLS estimator of β_2 is unbiased, it is not consistent. The simultaneous failure of consistency and asymptotic identification in this example is not a coincidence: It will turn out that asymptotic identification is a necessary and sufficient condition for consistency.

Consistency

Suppose that the DGP is a special case of the model (1.02) with true parameter vector $\boldsymbol{\beta}_0$. Under the assumption of asymptotic identification, the equations $\boldsymbol{\alpha}(\boldsymbol{\beta}) = \mathbf{0}$ have a unique solution, namely, $\boldsymbol{\beta} = \boldsymbol{\beta}_0$. This can be shown to imply that, as $n \to \infty$, the probability limit of the estimator $\hat{\boldsymbol{\beta}}$ defined by (1.10) is precisely $\boldsymbol{\beta}_0$. We will not attempt a formal proof of this result, since it would have to deal with a number of technical issues that are beyond the scope of this book. See Amemiya (1985, Section 4.3) or Davidson and MacKinnon (1993, Section 5.3) for more detailed treatments.

However, an intuitive, heuristic, proof is not at all hard to provide. If we make the assumption that $\hat{\boldsymbol{\beta}}$ has a deterministic probability limit, say $\boldsymbol{\beta}_\infty$, the result follows easily. What makes a formal proof more difficult is showing that $\boldsymbol{\beta}_\infty$ exists. Let us suppose that $\boldsymbol{\beta}_\infty \neq \boldsymbol{\beta}_0$. We will derive a contradiction from this assumption, and we will thus be able to conclude that $\boldsymbol{\beta}_\infty = \boldsymbol{\beta}_0$, in other words, that $\hat{\boldsymbol{\beta}}$ is consistent.

For all finite samples large enough for $\boldsymbol{\beta}$ to be identified by the data, we have, by the definition (1.10) of $\hat{\boldsymbol{\beta}}$, that

$$\frac{1}{n}\boldsymbol{W}^\top\big(\boldsymbol{y} - \boldsymbol{x}(\hat{\boldsymbol{\beta}})\big) = \mathbf{0}. \tag{1.15}$$

If we take the limit of this as $n \to \infty$, we have $\mathbf{0}$ on the right-hand side. On the left-hand side, because we assume that $\operatorname{plim} \hat{\boldsymbol{\beta}} = \boldsymbol{\beta}_\infty$, the limit is the same as the limit of

$$\frac{1}{n}\boldsymbol{W}^\top\big(\boldsymbol{y} - \boldsymbol{x}(\boldsymbol{\beta}_\infty)\big).$$

By (1.12), the limit of this expression is $\boldsymbol{\alpha}(\boldsymbol{\beta}_\infty)$. We assumed that $\boldsymbol{\beta}_\infty \neq \boldsymbol{\beta}_0$, and so, by the asymptotic identification condition, $\boldsymbol{\alpha}(\boldsymbol{\beta}_\infty) \neq \mathbf{0}$. But this contradicts the fact that the limits of both sides of (1.15) are equal, since the limit of the right-hand side is $\mathbf{0}$.

We have shown that, if we assume that a deterministic $\boldsymbol{\beta}_\infty$ exists, then asymptotic identification is sufficient for consistency. Although we will not attempt to prove it, asymptotic identification is also necessary for consistency. The key to a proof is showing that, if the parameters of a model are not asymptotically identified by a given estimation method, then no deterministic limit like $\boldsymbol{\beta}_\infty$ exists in general. An example of this is provided by the model (1.13); see also Exercise 1.2.

The identifiability of a parameter vector, whether asymptotic or by a data set, depends on the estimation method used. In the present context, this means that certain choices of the variables in $\boldsymbol{W}_t$ may identify the parameters of a model like (1.01), while others do not. We can gain some intuition about this matter by looking a little more closely at the limiting functions $\boldsymbol{\alpha}(\boldsymbol{\beta})$ defined by (1.12). We have

$$\begin{aligned} \boldsymbol{\alpha}(\boldsymbol{\beta}) &= \operatorname*{plim}_{n\to\infty} \frac{1}{n}\boldsymbol{W}^\top\big(\boldsymbol{y} - \boldsymbol{x}(\boldsymbol{\beta})\big) \\ &= \operatorname*{plim}_{n\to\infty} \frac{1}{n}\boldsymbol{W}^\top\big(\boldsymbol{x}(\boldsymbol{\beta}_0) - \boldsymbol{x}(\boldsymbol{\beta}) + \boldsymbol{u}\big) \\ &= \boldsymbol{\alpha}(\boldsymbol{\beta}_0) + \operatorname*{plim}_{n\to\infty} \frac{1}{n}\boldsymbol{W}^\top\big(\boldsymbol{x}(\boldsymbol{\beta}_0) - \boldsymbol{x}(\boldsymbol{\beta})\big) \\ &= \operatorname*{plim}_{n\to\infty} \frac{1}{n}\boldsymbol{W}^\top\big(\boldsymbol{x}(\boldsymbol{\beta}_0) - \boldsymbol{x}(\boldsymbol{\beta})\big). \end{aligned} \tag{1.16}$$

Therefore, for asymptotic identification, and so also for consistency, the last expression in (1.16) must be nonzero for all $\boldsymbol{\beta} \neq \boldsymbol{\beta}_0$.

Evidently, a necessary condition for asymptotic identification is that there is no $\boldsymbol{\beta}_1 \neq \boldsymbol{\beta}_0$ such that $\boldsymbol{x}(\boldsymbol{\beta}_1) = \boldsymbol{x}(\boldsymbol{\beta}_0)$. This condition is the nonlinear analog of the requirement of linearly independent regressors for linear regression models. We can now see that this requirement is in fact a condition necessary for the identification of the model parameters, both by a data set and asymptotically. Suppose that, for a linear regression model, the columns of the regressor matrix $\boldsymbol{X}$ are linearly dependent. This implies that there is a nonzero vector $\boldsymbol{b}$ such that $\boldsymbol{X}\boldsymbol{b} = \mathbf{0}$. Then it follows that $\boldsymbol{X}\boldsymbol{\beta}_0 = \boldsymbol{X}(\boldsymbol{\beta}_0 + \boldsymbol{b})$. For a linear regression model, $\boldsymbol{x}(\boldsymbol{\beta}) = \boldsymbol{X}\boldsymbol{\beta}$. Therefore, if we set $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_0 + \boldsymbol{b}$, the linear dependence means that $\boldsymbol{x}(\boldsymbol{\beta}_1) = \boldsymbol{x}(\boldsymbol{\beta}_0)$, in violation of the necessary condition stated at the beginning of this paragraph.

For a linear regression model, linear independence of the regressors is both necessary and sufficient for identification by any data set. We saw above that it is necessary, and sufficiency follows from the fact that $\boldsymbol{X}^\top\boldsymbol{X}$ is nonsingular if the columns of $\boldsymbol{X}$ are linearly independent. If $\boldsymbol{X}^\top\boldsymbol{X}$ is nonsingular, the OLS estimator $(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}$ exists and is unique for any $\boldsymbol{y}$, and this is precisely what is meant by identification by any data set.

For nonlinear models, however, things are more complicated. In general, more is needed for identification than the condition that no $\boldsymbol{\beta}_1 \neq \boldsymbol{\beta}_0$ exists such that $\boldsymbol{x}(\boldsymbol{\beta}_1) = \boldsymbol{x}(\boldsymbol{\beta}_0)$. The relevant issues will be easier to understand after we have derived the asymptotic covariance matrix of the estimator defined by (1.10), and so we postpone study of them until later.

The estimator $\hat{\boldsymbol{\beta}}$ defined by (1.10) is actually consistent under considerably weaker assumptions about the disturbances than those we have made. The key to the consistency proof is the requirement that the disturbances satisfy the condition

$$\operatorname*{plim}_{n\to\infty} \frac{1}{n}\boldsymbol{W}^\top\boldsymbol{u} = \boldsymbol{0}. \tag{1.17}$$

Under reasonable assumptions, it is not difficult to show that this condition holds even when the u_t are heteroskedastic, and it may also hold even when they are serially correlated. However, difficulties can arise when the u_t are serially correlated and $x_t(\boldsymbol{\beta})$ depends on lagged dependent variables. In this case, it will be seen later that the expectation of u_t conditional on the lagged dependent variable is nonzero in general. Therefore, in this circumstance, condition (1.17) does not hold whenever $\boldsymbol{W}$ includes lagged dependent variables, and such estimators are generally not consistent.

Asymptotic Normality

The estimator $\hat{\boldsymbol{\beta}}$ defined by (1.10) for different possible choices of $\boldsymbol{W}$ is asymptotically normal under appropriate conditions. This means that the vector $n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ follows the multivariate normal distribution with expectation vector $\boldsymbol{0}$ and a covariance matrix that will be determined shortly.

Before we start our analysis, we need some notation, which will be used extensively in the remainder of this chapter. In formulating the generic nonlinear regression model (1.01), we deliberately used $x_t(\cdot)$ to denote the regression function, rather than $f_t(\cdot)$ or some other notation, because this notation makes it easy to see the close connection between the nonlinear and linear regression models. It is natural to let the derivative of $x_t(\boldsymbol{\beta})$ with respect to β_i be denoted $X_{ti}(\boldsymbol{\beta})$. Then we can let $\boldsymbol{X}_t(\boldsymbol{\beta})$ denote a $1 \times k$ vector, and $\boldsymbol{X}(\boldsymbol{\beta})$ denote an $n \times k$ matrix, each having typical element $X_{ti}(\boldsymbol{\beta})$. These are the analogs of the vector $\boldsymbol{X}_t$ and the matrix $\boldsymbol{X}$ for the linear regression model. In the linear case, when the regression function is $\boldsymbol{X}\boldsymbol{\beta}$, it is easy to see that $\boldsymbol{X}_t(\boldsymbol{\beta}) = \boldsymbol{X}_t$ and $\boldsymbol{X}(\boldsymbol{\beta}) = \boldsymbol{X}$. The big difference between the linear and nonlinear cases is that, in the latter case, $\boldsymbol{X}_t(\boldsymbol{\beta})$ and $\boldsymbol{X}(\boldsymbol{\beta})$ depend on $\boldsymbol{\beta}$.

If we multiply equation (1.10) by $n^{-1/2}$, replace $\boldsymbol{y}$ by what it is equal to under the DGP (1.01) with parameter vector $\boldsymbol{\beta}_0$, and replace $\boldsymbol{\beta}$ by $\hat{\boldsymbol{\beta}}$, we obtain

$$n^{-1/2}\boldsymbol{W}^\top\big(\boldsymbol{u} + \boldsymbol{x}(\boldsymbol{\beta}_0) - \boldsymbol{x}(\hat{\boldsymbol{\beta}})\big) = \boldsymbol{0}. \tag{1.18}$$

The next step is to apply Taylor's Theorem at the point $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ to the components of the vector $\boldsymbol{x}(\hat{\boldsymbol{\beta}})$. We apply the formula (F6.58) with $m = k$, replacing $f(\boldsymbol{x})$ by $x_t(\boldsymbol{\beta}_0)$ and $\boldsymbol{h}$ by the vector $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0$. We thus obtain, for $t = 1, \dots, n$,

$$x_t(\hat{\boldsymbol{\beta}}) = x_t(\boldsymbol{\beta}_0) + \sum_{i=1}^{k} X_{ti}(\bar{\boldsymbol{\beta}}_t)(\hat{\beta}_i - \beta_{0i}), \tag{1.19}$$

where β_{0i} is the i^{th} element of $\boldsymbol{\beta}_0$, and the $\bar{\boldsymbol{\beta}}_t$, which play the role of $\boldsymbol{x} + \lambda\boldsymbol{h}$ in equation (F6.58), satisfy the condition

$$\big\|\bar{\boldsymbol{\beta}}_t - \boldsymbol{\beta}_0\big\| \leq \big\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\big\|. \tag{1.20}$$

Substituting the Taylor expansion (1.19) into (1.18) yields

$$n^{-1/2}\boldsymbol{W}^\top\boldsymbol{u} - n^{-1/2}\boldsymbol{W}^\top\boldsymbol{X}(\bar{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) = \boldsymbol{0}. \tag{1.21}$$

The notation $\boldsymbol{X}(\bar{\boldsymbol{\beta}})$ is convenient, but slightly inaccurate. According to (1.19), we need different parameter vectors $\bar{\boldsymbol{\beta}}_t$ for each row of that matrix. But, since all of these vectors satisfy (1.20), it is not necessary to make this fact explicit in the notation. Thus here, and in subsequent chapters, we will refer to a vector $\bar{\boldsymbol{\beta}}$ that satisfies (1.20), without implying that it must be the *same* vector for every row of the matrix $\boldsymbol{X}(\bar{\boldsymbol{\beta}})$. This is a legitimate notational convenience, because, since $\hat{\boldsymbol{\beta}}$ is consistent, as we have seen that it is under the requirement of asymptotic identification, then so too are all of the $\bar{\boldsymbol{\beta}}_t$. Consequently, (1.21) remains true asymptotically if we replace $\bar{\boldsymbol{\beta}}$ by $\boldsymbol{\beta}_0$. Doing this, and rearranging factors of powers of n so as to work only with quantities which have suitable probability limits, yields the result that

$$n^{-1/2}\boldsymbol{W}^\top\boldsymbol{u} - n^{-1}\boldsymbol{W}^\top\boldsymbol{X}(\boldsymbol{\beta}_0)\, n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \stackrel{a}{=} \boldsymbol{0}. \tag{1.22}$$

This result is the starting point for all our subsequent analysis.

We need to apply a law of large numbers to the first factor of the second term of (1.22), namely, $n^{-1}\boldsymbol{W}^\top\boldsymbol{X}_0$, where for notational ease we write $\boldsymbol{X}_0 \equiv \boldsymbol{X}(\boldsymbol{\beta}_0)$. Under reasonable regularity conditions, not unlike those needed for (F4.23) to hold, we have

$$\operatorname*{plim}_{n\to\infty} \frac{1}{n}\boldsymbol{W}^\top\boldsymbol{X}_0 = \lim_{n\to\infty} \frac{1}{n}\boldsymbol{W}^\top\mathrm{E}\big(\boldsymbol{X}(\boldsymbol{\beta}_0)\big) \equiv \boldsymbol{S}_{\boldsymbol{W}^\top\boldsymbol{X}},$$

where $\boldsymbol{S}_{\boldsymbol{W}^\top\boldsymbol{X}}$ is a deterministic $k \times k$ matrix. It turns out that a sufficient condition for the parameter vector $\boldsymbol{\beta}$ to be asymptotically identified by the

estimator $\hat{\boldsymbol{\beta}}$ defined by the estimating equations (1.10) is that $\boldsymbol{S}_{\boldsymbol{W}^\top\boldsymbol{X}}$ should have full rank. To see this, observe that (1.22) implies that

$$\boldsymbol{S}_{\boldsymbol{W}^\top\boldsymbol{X}}\, n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \stackrel{a}{=} n^{-1/2}\boldsymbol{W}^\top\boldsymbol{u}. \tag{1.23}$$

Because $\boldsymbol{S}_{\boldsymbol{W}^\top\boldsymbol{X}}$ is assumed to have full rank, its inverse exists. Thus we can multiply both sides of (1.23) by this inverse to obtain a well-defined expression for the limit of $n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$:

$$n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \stackrel{a}{=} (\boldsymbol{S}_{\boldsymbol{W}^\top\boldsymbol{X}})^{-1} n^{-1/2}\boldsymbol{W}^\top\boldsymbol{u}. \tag{1.24}$$

From this, we conclude that $\boldsymbol{\beta}$ is asymptotically identified by $\hat{\boldsymbol{\beta}}$. The condition that $\boldsymbol{S}_{\boldsymbol{W}^\top\boldsymbol{X}}$ be nonsingular is called **strong asymptotic identification**. It is a sufficient but not necessary condition for ordinary asymptotic identification.

The second factor on the right-hand side of (1.24) is a vector to which we should, under appropriate regularity conditions, be able to apply a central limit theorem. Since, by (1.09), $\mathrm{E}(\boldsymbol{W}_t u_t) = \boldsymbol{0}$, we can show that $n^{-1/2}\boldsymbol{W}^\top\boldsymbol{u}$ is asymptotically multivariate normal, with expectation vector $\boldsymbol{0}$ and a finite covariance matrix. To do this, we can show that the vector $\boldsymbol{v}$ of (F5.48) is asymptotically multivariate normal. Because the components of $n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ are, asymptotically, linear combinations of the components of a vector that follows the multivariate normal distribution, we conclude that $n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ itself must be asymptotically normally distributed with expectation vector zero and a finite covariance matrix. This implies that $\hat{\boldsymbol{\beta}}$ is root-n consistent.

Asymptotic Efficiency

The asymptotic covariance matrix of $n^{-1/2}\boldsymbol{W}^\top\boldsymbol{u}$, the second factor on the right-hand side of (1.24), is,

$$\sigma_0^2 \operatorname*{plim}_{n\to\infty} \frac{1}{n}\boldsymbol{W}^\top\boldsymbol{W} = \sigma_0^2 \boldsymbol{S}_{\boldsymbol{W}^\top\boldsymbol{W}}, \tag{1.25}$$

where σ_0^2 is the disturbance variance for the true DGP, and where we make the definition $\boldsymbol{S}_{\boldsymbol{W}^\top\boldsymbol{W}} \equiv \operatorname{plim} n^{-1}\boldsymbol{W}^\top\boldsymbol{W}$. From (1.24) and (1.25), it follows immediately that the asymptotic covariance matrix of the vector $n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ is

$$\sigma_0^2 (\boldsymbol{S}_{\boldsymbol{W}^\top\boldsymbol{X}})^{-1} \boldsymbol{S}_{\boldsymbol{W}^\top\boldsymbol{W}} (\boldsymbol{S}_{\boldsymbol{W}^\top\boldsymbol{X}}^\top)^{-1}, \tag{1.26}$$

which has the form of a sandwich. By the definitions of $\boldsymbol{S}_{\boldsymbol{W}^\top\boldsymbol{W}}$ and $\boldsymbol{S}_{\boldsymbol{W}^\top\boldsymbol{X}}$, expression (1.26) can be rewritten as

$$\begin{aligned}
&\sigma_0^2 \operatorname*{plim}_{n\to\infty} \big((n^{-1}\boldsymbol{W}^\top\boldsymbol{X}_0)^{-1} n^{-1}\boldsymbol{W}^\top\boldsymbol{W} (n^{-1}\boldsymbol{X}_0^\top\boldsymbol{W})^{-1}\big) \\
&\quad = \sigma_0^2 \operatorname*{plim}_{n\to\infty} \big(n^{-1}\boldsymbol{X}_0^\top\boldsymbol{W}(\boldsymbol{W}^\top\boldsymbol{W})^{-1}\boldsymbol{W}^\top\boldsymbol{X}_0\big)^{-1} \\
&\quad = \sigma_0^2 \operatorname*{plim}_{n\to\infty} (n^{-1}\boldsymbol{X}_0^\top\boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{X}_0)^{-1},
\end{aligned} \tag{1.27}$$

where $\boldsymbol{P}_{\boldsymbol{W}}$ is the orthogonal projection on to $\mathcal{S}(\boldsymbol{W})$, the subspace spanned by the columns of $\boldsymbol{W}$. Expression (1.27) is the asymptotic covariance matrix of the vector $n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$. However, it is common to refer to it as the asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}$, and we will allow ourselves this slight abuse of terminology when no confusion can result.

It is clear from the result (1.27) that the asymptotic covariance matrix of the estimator $\hat{\boldsymbol{\beta}}$ depends on the variables $\boldsymbol{W}$ used to obtain it. Most choices of $\boldsymbol{W}$ lead to an inefficient estimator by the criterion of the asymptotic covariance matrix, as we would be led to suspect by the fact that (1.26) has the form of a sandwich. It is not hard to show that an estimator which is **asymptotically efficient** is given by the choice $\boldsymbol{W} = \boldsymbol{X}_0$. To demonstrate this, we need to show that this choice of $\boldsymbol{W}$ minimizes the asymptotic covariance matrix, in the sense used in the Gauss-Markov theorem. Recall that one covariance matrix is said to be "greater" than another if the difference between it and the other is a positive semidefinite matrix.

If we set $\boldsymbol{W} = \boldsymbol{X}_0$ to define the estimator, the asymptotic covariance matrix (1.27) becomes $\sigma_0^2 \operatorname{plim}(n^{-1}\boldsymbol{X}_0^\top\boldsymbol{X}_0)^{-1}$. It is often easier to establish efficiency by reasoning in terms of the precision matrix, that is, the inverse of the covariance matrix, rather than in terms of the (asymptotic) covariance matrix itself. Since

$$\boldsymbol{X}_0^\top\boldsymbol{X}_0 - \boldsymbol{X}_0^\top\boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{X}_0 = \boldsymbol{X}_0^\top\boldsymbol{M}_{\boldsymbol{W}}\boldsymbol{X}_0,$$

which is a positive semidefinite matrix, it follows at once that the precision of the estimator obtained by setting $\boldsymbol{W} = \boldsymbol{X}_0$ is greater than or equal to that of the estimator obtained by using any other choice of $\boldsymbol{W}$. The same precision can be obtained only if $\boldsymbol{M}_{\boldsymbol{W}}\boldsymbol{X}_0 = \boldsymbol{O}$, that is, if every column of the matrix $\boldsymbol{X}_0$ is in the subspace $\mathcal{S}(\boldsymbol{W})$. In other words, the estimator is asymptotically efficient whenever $\boldsymbol{X}_0$ belongs to $\mathcal{S}(\boldsymbol{W})$.

Of course, we cannot actually use $\boldsymbol{X}_0$ for $\boldsymbol{W}$ in practice, because $\boldsymbol{X}_0 \equiv \boldsymbol{X}(\boldsymbol{\beta}_0)$ depends on the unknown true parameter vector $\boldsymbol{\beta}_0$. The estimator that uses $\boldsymbol{X}_0$ for $\boldsymbol{W}$ is therefore said to be **infeasible**. In the next section, we will see how to overcome this difficulty. The nonlinear least-squares estimator that we will obtain turns out to have exactly the same asymptotic properties as the infeasible estimator.

1.3 Nonlinear Least Squares

There are at least two ways in which we can approximate the asymptotically efficient, but infeasible, estimator that uses $\boldsymbol{X}_0$ for $\boldsymbol{W}$. The first, and perhaps the simpler of the two, is to begin by choosing any $\boldsymbol{W}$ for which $\boldsymbol{W}_t$ belongs to the information set Ω_t and using this $\boldsymbol{W}$ to obtain a preliminary consistent estimate, say $\acute{\boldsymbol{\beta}}$, of the model parameters. We can then estimate $\boldsymbol{\beta}$ once more, setting $\boldsymbol{W} = \acute{\boldsymbol{X}} \equiv \boldsymbol{X}(\acute{\boldsymbol{\beta}})$. The consistency of $\acute{\boldsymbol{\beta}}$ ensures that $\acute{\boldsymbol{X}}$ tends to the efficient choice $\boldsymbol{X}_0$ as $n \to \infty$.

A more subtle approach is to recognize that the above procedure estimates the same parameter vector twice, and to compress the two estimation procedures into one. Consider the estimating equations

$$\boldsymbol{X}^\top(\boldsymbol{\beta})\big(\boldsymbol{y} - \boldsymbol{x}(\boldsymbol{\beta})\big) = \boldsymbol{0}. \tag{1.28}$$

If the estimator $\hat{\boldsymbol{\beta}}$ obtained by solving the k equations (1.28) is consistent, then $\hat{\boldsymbol{X}} \equiv \boldsymbol{X}(\hat{\boldsymbol{\beta}})$ tends to $\boldsymbol{X}_0$ as $n \to \infty$. Therefore, it must be the case that, for sufficiently large samples, $\hat{\boldsymbol{\beta}}$ is very close to the infeasible, efficient estimator.

The estimator $\hat{\boldsymbol{\beta}}$ based on (1.28) is known as the **nonlinear least-squares**, or **NLS**, estimator. The name comes from the fact that the estimating equations (1.28) are just the first-order conditions for the minimization with respect to $\boldsymbol{\beta}$ of the sum-of-squared-residuals (or SSR) function. The SSR function is defined just as in (F2.48), but for a nonlinear regression function:

$$\mathrm{SSR}(\boldsymbol{\beta}) = \sum_{t=1}^{n}\big(y_t - x_t(\boldsymbol{\beta})\big)^2 = \big(\boldsymbol{y} - \boldsymbol{x}(\boldsymbol{\beta})\big)^\top\big(\boldsymbol{y} - \boldsymbol{x}(\boldsymbol{\beta})\big). \tag{1.29}$$

It is easy to check (see Exercise 1.4) that the estimating equations (1.28) are equivalent to the first-order conditions for minimizing (1.29).

Equations (1.28), which define the NLS estimator, closely resemble equations (1.08), which define the OLS estimator. Like the latter, the former can be interpreted as orthogonality conditions: They require that the columns of the matrix of derivatives of $\boldsymbol{x}(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$ should be orthogonal to the vector of residuals. There are, however, two major differences between (1.28) and (1.08). The first difference is that, in the nonlinear case, $\boldsymbol{X}(\boldsymbol{\beta})$ is a matrix of functions that depend on the explanatory variables and on $\boldsymbol{\beta}$, instead of simply a matrix of explanatory variables. The second difference is that equations (1.28) are nonlinear in $\boldsymbol{\beta}$, because both $\boldsymbol{x}(\boldsymbol{\beta})$ and $\boldsymbol{X}(\boldsymbol{\beta})$ are, in general, nonlinear functions of $\boldsymbol{\beta}$. Thus there is no closed-form expression for $\hat{\boldsymbol{\beta}}$ comparable to the famous formula (F2.45). As we will see in Section 1.4, this means that it is substantially more difficult to compute NLS estimates than it is to compute OLS ones.

Consistency of the NLS Estimator

Since it has been assumed that every variable on which $x_t(\boldsymbol{\beta})$ depends belongs to Ω_t, it must be the case that $x_t(\boldsymbol{\beta})$ itself belongs to Ω_t for any choice of $\boldsymbol{\beta}$. Therefore, the partial derivatives of $x_t(\boldsymbol{\beta})$, that is, the elements of the row vector $\boldsymbol{X}_t(\boldsymbol{\beta})$, must belong to Ω_t as well, and so

$$\mathrm{E}\big(\boldsymbol{X}_t(\boldsymbol{\beta})\, u_t\big) = \boldsymbol{0}. \tag{1.30}$$

If we define the limiting functions $\boldsymbol{\alpha}(\boldsymbol{\beta})$ for the estimator based on (1.28) analogously to (1.12), we have

$$\boldsymbol{\alpha}(\boldsymbol{\beta}) = \operatorname*{plim}_{n\to\infty} \frac{1}{n}\boldsymbol{X}^\top(\boldsymbol{\beta})\big(\boldsymbol{y} - \boldsymbol{x}(\boldsymbol{\beta})\big).$$

It follows from (1.30) and the law of large numbers that $\boldsymbol{\alpha}(\boldsymbol{\beta}_0) = \boldsymbol{0}$ if the true parameter vector is $\boldsymbol{\beta}_0$. Thus the NLS estimator is consistent provided that it is asymptotically identified. We will have more to say in the next section about identification and the NLS estimator.

Asymptotic Normality of the NLS Estimator

The discussion of asymptotic normality in the previous section needs to be modified slightly for the NLS estimator. Equation (1.21), which resulted from applying Taylor's Theorem to $\boldsymbol{x}(\hat{\boldsymbol{\beta}})$, is no longer true, because the matrix $\boldsymbol{W}$ is replaced by $\boldsymbol{X}(\boldsymbol{\beta})$, which, unlike $\boldsymbol{W}$, depends on the parameter vector $\boldsymbol{\beta}$. When we take account of this fact, we obtain a rather messy additional term in (1.21) that depends on the second derivatives of $\boldsymbol{x}(\boldsymbol{\beta})$. However, it can be shown that this extra term vanishes asymptotically. Therefore, equation (1.22) remains true, but with $\boldsymbol{X}_0 \equiv \boldsymbol{X}(\boldsymbol{\beta}_0)$ replacing $\boldsymbol{W}$. This implies that, for NLS, the analog of equation (1.24) is

$$n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \stackrel{a}{=} \Big(\operatorname*{plim}_{n\to\infty} \frac{1}{n}\boldsymbol{X}_0^\top\boldsymbol{X}_0\Big)^{-1} n^{-1/2}\boldsymbol{X}_0^\top\boldsymbol{u}, \tag{1.31}$$

from which the asymptotic normality of the NLS estimator follows by essentially the same arguments as before.

Slightly modified versions of the arguments for estimators of the previous section also yield expressions for the asymptotic covariance matrix of the NLS estimator $\hat{\boldsymbol{\beta}}$. The consistency of $\hat{\boldsymbol{\beta}}$ means that

$$\operatorname*{plim}_{n\to\infty} \frac{1}{n}\hat{\boldsymbol{X}}^\top\hat{\boldsymbol{X}} = \operatorname*{plim}_{n\to\infty} \frac{1}{n}\boldsymbol{X}_0^\top\boldsymbol{X}_0 \quad \text{and} \quad \operatorname*{plim}_{n\to\infty} \frac{1}{n}\hat{\boldsymbol{X}}^\top\boldsymbol{X}_0 = \operatorname*{plim}_{n\to\infty} \frac{1}{n}\boldsymbol{X}_0^\top\boldsymbol{X}_0.$$

Thus, on setting $\boldsymbol{W} = \hat{\boldsymbol{X}}$, (1.27) gives for the asymptotic covariance matrix of $n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ the matrix

$$\sigma_0^2 \operatorname*{plim}_{n\to\infty}\Big(\frac{1}{n}\boldsymbol{X}_0^\top\boldsymbol{P}_{\hat{\boldsymbol{X}}}\boldsymbol{X}_0\Big)^{-1} = \sigma_0^2 \operatorname*{plim}_{n\to\infty}\Big(\frac{1}{n}\boldsymbol{X}_0^\top\boldsymbol{X}_0\Big)^{-1}, \tag{1.32}$$

from which we see that the NLS estimator $\hat{\boldsymbol{\beta}}$ is asymptotically efficient. Moreover, it follows that a consistent estimator of the covariance matrix of $\hat{\boldsymbol{\beta}}$ is

$$\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}) = s^2(\hat{\boldsymbol{X}}^\top\hat{\boldsymbol{X}})^{-1}, \tag{1.33}$$

where, by analogy with (F4.63),

$$s^2 \equiv \frac{1}{n-k}\sum_{t=1}^{n}\hat{u}_t^2 = \frac{1}{n-k}\sum_{t=1}^{n}\big(y_t - x_t(\hat{\boldsymbol{\beta}})\big)^2. \tag{1.34}$$

Of course, s^2 is not the only consistent estimator of σ^2 that we might reasonably use. Another possibility is to use

$$\hat{\sigma}^2 \equiv \frac{1}{n}\sum_{t=1}^{n}\hat{u}_t^2. \tag{1.35}$$

However, we will see shortly that (1.34) has particularly attractive properties.

NLS Residuals and the Variance of the Disturbances

Not very much can be said about the finite-sample properties of nonlinear least squares. The techniques that we used in Part 1, Chapter 4 to obtain the finite-sample properties of the OLS estimator simply cannot be used for the NLS one. However, it is easy to show that, if the DGP is

$$\boldsymbol{y} = \boldsymbol{x}(\boldsymbol{\beta}_0) + \boldsymbol{u}, \quad \boldsymbol{u} \sim \text{IID}(\boldsymbol{0}, \sigma_0^2 \mathbf{I}), \tag{1.36}$$

which means that it is a special case of the model (1.02) that is being estimated, then

$$\text{E}\big(\text{SSR}(\hat{\boldsymbol{\beta}})\big) \leq n\sigma_0^2. \tag{1.37}$$

The argument is just this. From (1.36), $\boldsymbol{y} - \boldsymbol{x}(\boldsymbol{\beta}_0) = \boldsymbol{u}$. Therefore,

$$\text{E}\big(\text{SSR}(\boldsymbol{\beta}_0)\big) = \text{E}(\boldsymbol{u}^\top \boldsymbol{u}) = n\sigma_0^2.$$

Since $\hat{\boldsymbol{\beta}}$ minimizes the sum of squared residuals and $\boldsymbol{\beta}_0$ in general does not, it must be the case that $\text{SSR}(\hat{\boldsymbol{\beta}}) \leq \text{SSR}(\boldsymbol{\beta}_0)$. The inequality (1.37) follows immediately. Thus, just like OLS residuals, NLS residuals have variance less than the variance of the disturbances.

The consistency of $\hat{\boldsymbol{\beta}}$ implies that the NLS residuals $\hat{u}_t$ converge to the disturbances u_t as $n \to \infty$. This means that it is valid asymptotically to use either s^2 from (1.34) or $\hat{\sigma}^2$ from (1.35) to estimate σ^2. However, we see from (1.37) that the NLS residuals are too small on average. Therefore, by analogy with exact results for the OLS case, it seems plausible to divide by $n-k$ instead of by n when we estimate σ^2. In fact, as we now show, there is an even stronger justification for doing this.

Now let us apply Taylor's Theorem to a typical residual, $\hat{u}_t = y_t - x_t(\hat{\boldsymbol{\beta}})$. If we expand this quantity around the true value $\boldsymbol{\beta}_0$ and substitute $u_t + x_t(\boldsymbol{\beta}_0)$ for y_t, we obtain

$$\begin{aligned}
\hat{u}_t &= y_t - x_t(\boldsymbol{\beta}_0) - \bar{\boldsymbol{X}}_t(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \\
&= u_t + x_t(\boldsymbol{\beta}_0) - x_t(\boldsymbol{\beta}_0) - \bar{\boldsymbol{X}}_t(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \\
&= u_t - \bar{\boldsymbol{X}}_t(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0),
\end{aligned}$$

where $\bar{\boldsymbol{X}}_t$ denotes the t^{th} row of the matrix $\boldsymbol{X}(\bar{\boldsymbol{\beta}})$, for some $\bar{\boldsymbol{\beta}}$ that satisfies condition (1.20); recall the discussion of that condition. This implies that, for the entire vector of residuals, we have

$$\hat{\boldsymbol{u}} = \boldsymbol{u} - \bar{\boldsymbol{X}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0). \tag{1.38}$$

For the NLS estimator $\hat{\boldsymbol{\beta}}$, the asymptotic result (1.24) becomes

$$n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \stackrel{a}{=} (\boldsymbol{S}_{\boldsymbol{X}^\top \boldsymbol{X}})^{-1} n^{-1/2} \boldsymbol{X}_0^\top \boldsymbol{u}, \tag{1.39}$$

where

$$\boldsymbol{S}_{\boldsymbol{X}^\top \boldsymbol{X}} \equiv \operatorname*{plim}_{n\to\infty} \frac{1}{n} \boldsymbol{X}_0^\top \boldsymbol{X}_0. \tag{1.40}$$

We have redefined $\boldsymbol{S}_{\boldsymbol{X}^\top \boldsymbol{X}}$ here. The old definition, (F4.23), applies only to linear regression models. The new definition, (1.40), applies to both linear and nonlinear regression models, since it reduces to the old one when the regression function is linear.

When we substitute $n^{-1/2}$ times the right-hand side of equation (1.39) into equation (1.38) and replace $\bar{\boldsymbol{X}}$ with $\boldsymbol{X}_0$ because $\bar{\boldsymbol{\beta}}$ tends asymptotically to $\boldsymbol{\beta}_0$, we find that

$$\begin{aligned}
\hat{\boldsymbol{u}} &\stackrel{a}{=} \boldsymbol{u} - n^{-1/2} \boldsymbol{X}_0 (\boldsymbol{S}_{\boldsymbol{X}^\top \boldsymbol{X}})^{-1} n^{-1/2} \boldsymbol{X}_0^\top \boldsymbol{u} \\
&\stackrel{a}{=} \boldsymbol{u} - n^{-1} \boldsymbol{X}_0 (n^{-1} \boldsymbol{X}_0^\top \boldsymbol{X}_0)^{-1} \boldsymbol{X}_0^\top \boldsymbol{u} \\
&= \boldsymbol{u} - \boldsymbol{X}_0 (\boldsymbol{X}_0^\top \boldsymbol{X}_0)^{-1} \boldsymbol{X}_0^\top \boldsymbol{u} \\
&= \boldsymbol{u} - \boldsymbol{P}_{\boldsymbol{X}_0} \boldsymbol{u} = \boldsymbol{M}_{\boldsymbol{X}_0} \boldsymbol{u},
\end{aligned} \tag{1.41}$$

where $\boldsymbol{P}_{\boldsymbol{X}_0}$ and $\boldsymbol{M}_{\boldsymbol{X}_0}$ project orthogonally on to $\mathcal{S}(\boldsymbol{X}_0)$ and $\mathcal{S}^\perp(\boldsymbol{X}_0)$, respectively. This asymptotic result for NLS looks very much like the exact result that $\hat{\boldsymbol{u}} = \boldsymbol{M}_{\boldsymbol{X}} \boldsymbol{u}$ for OLS. A somewhat more intricate argument can be used to show that the difference between $\hat{\boldsymbol{u}}^\top \hat{\boldsymbol{u}}$ and $\boldsymbol{u}^\top \boldsymbol{M}_{\boldsymbol{X}_0} \boldsymbol{u}$ tends to zero as $n \to \infty$; see Exercise 1.8. Since $\boldsymbol{X}_0$ is an $n \times k$ matrix, precisely the same argument that was used for the linear case in (F4.62) shows that $\text{E}(\hat{\boldsymbol{u}}^\top \hat{\boldsymbol{u}}) \stackrel{a}{=} \sigma_0^2(n-k)$. Thus we see that, in the case of nonlinear least squares, s^2 provides an approximately unbiased estimator of σ^2.

1.4 Computing NLS Estimates

We have not yet said anything about how to compute nonlinear least-squares estimates. This is by no means a trivial undertaking. Computing NLS estimates is always much more expensive than computing OLS ones for a model with the same number of observations and parameters. Moreover, there is a risk that the program may fail to converge or may converge to values that do not minimize the SSR. However, with modern computers and well-written software, NLS estimation is usually not excessively difficult.

In order to find NLS estimates, we need to minimize the sum-of-squared-residuals function $\text{SSR}(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$. Since $\text{SSR}(\boldsymbol{\beta})$ is not a quadratic function of $\boldsymbol{\beta}$, there is no analytic solution like the classic formula (F2.45) for the linear regression case. What we need is a general algorithm for minimizing a sum of squares with respect to a vector of parameters. In this section, we discuss methods for unconstrained minimization of a smooth function $Q(\boldsymbol{\beta})$.

It is easiest to think of $Q(\boldsymbol{\beta})$ as being equal to $\text{SSR}(\boldsymbol{\beta})$, but much of the discussion is applicable to minimizing any sort of **criterion function**. Since minimizing $Q(\boldsymbol{\beta})$ is equivalent to maximizing $-Q(\boldsymbol{\beta})$, it is also applicable to maximizing any sort of criterion function, such as the loglikelihood functions that we will encounter in Chapter 3.

We will give an overview of how numerical minimization algorithms work, but we will not discuss many of the important implementation issues that can substantially affect the performance of these algorithms when they are incorporated into computer programs. References on the art and science of numerical optimization, especially as it applies to nonlinear regression, include Bard (1974), Gill, Murray, and Wright (1981), Quandt (1983), Bates and Watts (1988), Seber and Wild (1989, Chapter 14), Press, Flannery, Teukolsky, and Vetterling (1992a; 1992b, Chapter 10), McCullough (2003), and McCullough and Vinod (2003).

There are many algorithms for minimizing a smooth function $Q(\boldsymbol{\beta})$. Most of these operate in essentially the same way. The algorithm goes through a series of iterations, or steps, at each of which it starts with a particular value of $\boldsymbol{\beta}$ and tries to find a better one. It first chooses a direction in which to search and then decides how far to move in that direction. After completing the move, it checks to see whether the current value of $\boldsymbol{\beta}$ is sufficiently close to a local minimum of $Q(\boldsymbol{\beta})$. If it is, the algorithm stops. Otherwise, it chooses another direction in which to search, and so on. There are three principal differences among minimization algorithms: the way in which the direction to search is chosen, the way in which the size of the step in that direction is determined, and the stopping rule that is employed. Numerous choices for each of these are available.

Newton's Method

All of the techniques that we will discuss are based on **Newton's Method**. Suppose that we wish to minimize a function $Q(\boldsymbol{\beta})$, where $\boldsymbol{\beta}$ is a k-vector and $Q(\boldsymbol{\beta})$ is assumed to be twice continuously differentiable. Given any initial value of $\boldsymbol{\beta}$, say $\boldsymbol{\beta}_{(0)}$, we can perform a second-order Taylor expansion of $Q(\boldsymbol{\beta})$ around $\boldsymbol{\beta}_{(0)}$ in order to obtain an approximation $Q^*(\boldsymbol{\beta})$ to $Q(\boldsymbol{\beta})$:

$$Q^*(\boldsymbol{\beta}) = Q(\boldsymbol{\beta}_{(0)}) + \boldsymbol{g}_{(0)}^\top(\boldsymbol{\beta} - \boldsymbol{\beta}_{(0)}) + \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}_{(0)})^\top \boldsymbol{H}_{(0)}(\boldsymbol{\beta} - \boldsymbol{\beta}_{(0)}), \tag{1.42}$$

where $\boldsymbol{g}(\boldsymbol{\beta})$, the **gradient** of $Q(\boldsymbol{\beta})$, is a column vector of dimension k with typical element $\partial Q(\boldsymbol{\beta})/\partial\beta_i$, and $\boldsymbol{H}(\boldsymbol{\beta})$, the **Hessian** of $Q(\boldsymbol{\beta})$, is a $k \times k$ matrix with typical element $\partial^2 Q(\boldsymbol{\beta})/\partial\beta_i\partial\beta_l$. For notational simplicity, $\boldsymbol{g}_{(0)}$ and $\boldsymbol{H}_{(0)}$ denote $\boldsymbol{g}(\boldsymbol{\beta}_{(0)})$ and $\boldsymbol{H}(\boldsymbol{\beta}_{(0)})$, respectively.

It is easy to see that the first-order conditions for a minimum of $Q^*(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$ can be written as

$$\boldsymbol{g}_{(0)} + \boldsymbol{H}_{(0)}(\boldsymbol{\beta} - \boldsymbol{\beta}_{(0)}) = \mathbf{0}.$$

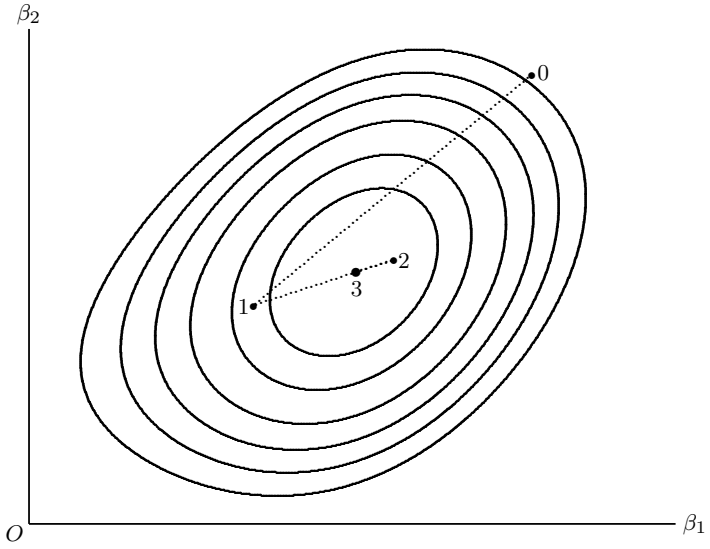


Figure 1.1 Newton's Method in two dimensions

Solving these yields a new value of $\boldsymbol{\beta}$, which we will call $\boldsymbol{\beta}_{(1)}$:

$$\boldsymbol{\beta}_{(1)} = \boldsymbol{\beta}_{(0)} - \boldsymbol{H}_{(0)}^{-1}\boldsymbol{g}_{(0)}. \tag{1.43}$$

Equation (1.43) is the heart of Newton's Method. If the quadratic approximation $Q^*(\boldsymbol{\beta})$ is a strictly convex function, which it is if and only if the Hessian $\boldsymbol{H}_{(0)}$ is positive definite, then $\boldsymbol{\beta}_{(1)}$ is the global minimum of $Q^*(\boldsymbol{\beta})$. If, in addition, $Q^*(\boldsymbol{\beta})$ is a good approximation to $Q(\boldsymbol{\beta})$, $\boldsymbol{\beta}_{(1)}$ should be close to $\hat{\boldsymbol{\beta}}$, the minimum of $Q(\boldsymbol{\beta})$. Newton's Method involves using equation (1.43) repeatedly to find a succession of values $\boldsymbol{\beta}_{(1)}$, $\boldsymbol{\beta}_{(2)}\ldots$. When the original function $Q(\boldsymbol{\beta})$ is quadratic and has a global minimum at $\hat{\boldsymbol{\beta}}$, Newton's Method evidently finds $\hat{\boldsymbol{\beta}}$ in a single step, since the quadratic approximation is then exact. When $Q(\boldsymbol{\beta})$ is approximately quadratic, as all sum-of-squares functions are when sufficiently close to their minima, Newton's Method generally converges very quickly.

Figure 1.1 illustrates how Newton's Method works. It shows the contours of the function $Q(\boldsymbol{\beta}) = \text{SSR}(\beta_1, \beta_2)$ for a regression model with two parameters. Notice that these contours are not precisely elliptical, as they would be if the function were quadratic. The algorithm starts at the point marked "0" and then jumps to the point marked "1." On the next step, it goes in almost exactly the right direction, but it goes too far, moving to "2." It then retraces its own steps to "3," which is essentially the minimum of $\text{SSR}(\beta_1, \beta_2)$. After one more step, which is too small to be shown in the figure, it has essentially converged.

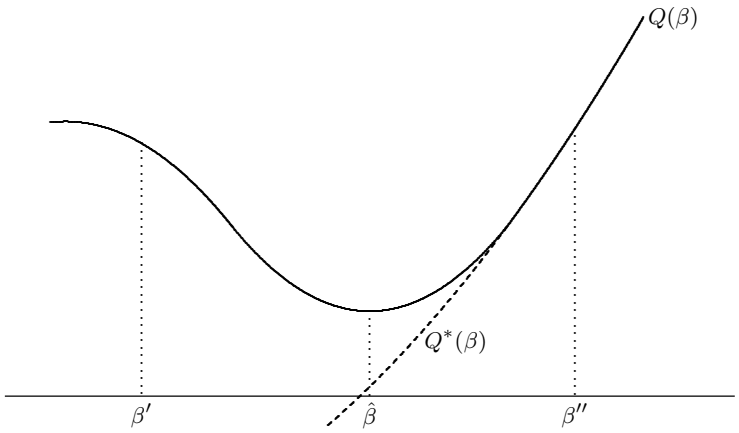


Figure 1.2 Cases for which Newton's Method will not work

Although Newton's Method works very well in this example, there are many cases in which it fails to work at all, especially if $Q(\boldsymbol{\beta})$ is not convex in the neighborhood of $\boldsymbol{\beta}_{(j)}$ for some j in the sequence. Some of the possibilities are illustrated in Figure 1.2. The one-dimensional function shown there has a global minimum at $\hat{\beta}$, but when Newton's Method is started at points such as β' or β'', it may never find $\hat{\beta}$. In the former case, $Q(\beta)$ is concave at β' instead of convex, and this causes Newton's Method to head off in the wrong direction. In the latter case, the quadratic approximation at β'', $Q^*(\beta)$, which is shown by the dashed curve, is extremely poor for values away from β'', because $Q(\beta)$ is very flat near β''. It is evident that $Q^*(\beta)$ must have a minimum far to the left of $\hat{\beta}$. Thus, after the first step, the algorithm is very much further away from $\hat{\beta}$ than it was at its starting point.

One important feature of Newton's Method and algorithms based on it is that they must start with an initial value of $\boldsymbol{\beta}$. It is impossible to perform a Taylor expansion around $\boldsymbol{\beta}_{(0)}$ without specifying $\boldsymbol{\beta}_{(0)}$. As Figure 1.2 illustrates, where the algorithm starts may determine how well it performs, or whether it converges at all. In most cases, it is up to the econometrician to specify the starting values.

Quasi-Newton Methods

Most effective nonlinear optimization techniques for minimizing smooth criterion functions are variants of Newton's Method. These **quasi-Newton methods** attempt to retain the good qualities of Newton's Method while surmounting problems like those illustrated in Figure 1.2. They replace (1.43) by the slightly more complicated formula

$$\boldsymbol{\beta}_{(j+1)} = \boldsymbol{\beta}_{(j)} - \alpha_{(j)} \boldsymbol{D}_{(j)}^{-1} \boldsymbol{g}_{(j)}, \tag{1.44}$$

which determines $\boldsymbol{\beta}_{(j+1)}$, the value of $\boldsymbol{\beta}$ at step $j+1$, as a function of $\boldsymbol{\beta}_{(j)}$. Here $\alpha_{(j)}$ is a scalar which is determined at each step, and $\boldsymbol{D}_{(j)} \equiv \boldsymbol{D}(\boldsymbol{\beta}_{(j)})$ is a matrix which approximates $\boldsymbol{H}_{(j)}$ near the minimum but is constructed so that it is always positive definite. In contrast to quasi-Newton methods, **modified Newton methods** set $\boldsymbol{D}_{(j)} = \boldsymbol{H}_{(j)}$, and Newton's Method itself sets $\boldsymbol{D}_{(j)} = \boldsymbol{H}_{(j)}$ and $\alpha_{(j)} = 1$.

Quasi-Newton algorithms involve three operations at each step. Let us denote the current value of $\boldsymbol{\beta}$ by $\boldsymbol{\beta}_{(j)}$. If $j = 0$, this is the starting value, $\boldsymbol{\beta}_{(0)}$; otherwise, it is the value reached at iteration j. The three operations are

1. Compute $\boldsymbol{g}_{(j)}$ and $\boldsymbol{D}_{(j)}$ and use them to determine the direction $\boldsymbol{D}_{(j)}^{-1} \boldsymbol{g}_{(j)}$.
2. Find $\alpha_{(j)}$. Often, this is done by solving a one-dimensional minimization problem. Then use (1.44) to determine $\boldsymbol{\beta}_{(j+1)}$.
3. Decide whether $\boldsymbol{\beta}_{(j+1)}$ provides a sufficiently accurate approximation to $\hat{\boldsymbol{\beta}}$. If so, stop. Otherwise, return to 1.

Because they construct $\boldsymbol{D}(\boldsymbol{\beta})$ in such a way that it is always positive definite, quasi-Newton algorithms can handle problems where the function to be minimized is not globally convex. The various algorithms choose $\boldsymbol{D}(\boldsymbol{\beta})$ in a number of ways, some of which are quite ingenious and may be tricky to implement on a digital computer. As we will shortly see, however, for sum-of-squares functions there is a very easy and natural way to choose $\boldsymbol{D}(\boldsymbol{\beta})$.

The scalar $\alpha_{(j)}$ is often chosen so as to minimize the function

$$Q^\dagger(\alpha) \equiv Q\big(\boldsymbol{\beta}_{(j)} - \alpha \boldsymbol{D}_{(j)}^{-1} \boldsymbol{g}_{(j)}\big),$$

regarded as a one-dimensional function of α. It is fairly clear that, for the example in Figure 7.1, choosing α in this way would produce even faster convergence than setting $\alpha = 1$. Some algorithms do not actually minimize $Q^\dagger(\alpha)$ with respect to α, but merely choose $\alpha_{(j)}$ so as to ensure that $Q(\boldsymbol{\beta}_{(j+1)})$ is less than $Q(\boldsymbol{\beta}_{(j)})$. It is essential for this to be the case if we are to be sure that the algorithm always makes progress at each step. The best algorithms, which are designed to economize on computing time, may choose α quite crudely when they are far from $\hat{\boldsymbol{\beta}}$, but they almost always perform an accurate one-dimensional minimization when they are close to $\hat{\boldsymbol{\beta}}$.

Stopping Rules

No minimization algorithm running on a digital computer ever finds $\hat{\boldsymbol{\beta}}$ exactly. Without a rule telling it when to stop, the algorithm would just keep on going forever. There are many possible **stopping rules**. We could, for example, stop when $Q(\boldsymbol{\beta}_{(j-1)}) - Q(\boldsymbol{\beta}_{(j)})$ is very small, when every element of $\boldsymbol{g}_{(j)}$ is very small, or when every element of the vector $\boldsymbol{\beta}_{(j)} - \boldsymbol{\beta}_{(j-1)}$ is very small. However, none of these rules is entirely satisfactory, in part because they depend on the magnitude of the parameters. This means that they yield different results

if the units of measurement of any variable are changed or if the model is reparametrized in some other way. A more logical rule is to stop when

$$\boldsymbol{g}_{(j)}^{\top} \boldsymbol{D}_{(j)}^{-1} \boldsymbol{g}_{(j)} < \varepsilon, \tag{1.45}$$

where ε, the **convergence tolerance**, is a small positive number that is chosen by the user. Sensible values of ε might range from 10^{-12} to 10^{-4}. The advantage of (1.45) is that it weights the various components of the gradient in a manner inversely proportional to the precision with which the corresponding parameters are estimated. We will see why this is so in the next section.

Of course, any stopping rule may work badly if ε is chosen incorrectly. If ε is too large, the algorithm may stop too soon, when $\boldsymbol{\beta}_{(j)}$ is still far away from $\hat{\boldsymbol{\beta}}$. On the other hand, if ε is too small, the algorithm may keep going long after $\boldsymbol{\beta}_{(j)}$ is so close to $\hat{\boldsymbol{\beta}}$ that any differences are due solely to round-off error. It may therefore be a good idea to experiment with the value of ε to see how sensitive to it the results are. If the reported $\hat{\boldsymbol{\beta}}$ changes noticeably when ε is reduced, then either the first value of ε was too large, or the algorithm is having trouble finding an accurate minimum.

Local and Global Minima

Numerical optimization methods based on Newton's Method generally work well when $Q(\boldsymbol{\beta})$ is globally convex. For such a function, there can be at most one local minimum, which is also the global minimum. When $Q(\boldsymbol{\beta})$ is not globally convex but has only a single local minimum, these methods also work reasonably well in many cases. However, if there is more than one local minimum, optimization methods of this type often run into trouble. They generally converge to a local minimum, but there is no guarantee that it is the global one. In such cases, the choice of the **starting values**, that is, the vector $\boldsymbol{\beta}_{(0)}$, can be extremely important.

This problem is illustrated in Figure 7.3. The one-dimensional criterion function $Q(\beta)$ shown in the figure has two local minima. One of these, at $\hat{\beta}$, is also the global minimum. However, if a Newton or quasi-Newton algorithm is started to the right of the local maximum at β'', it is likely to converge to the local minimum at β' instead of to the global one at $\hat{\beta}$.

In practice, the usual way to guard against finding the wrong local minimum when the criterion function is known, or suspected, not to be globally convex is to minimize $Q(\boldsymbol{\beta})$ several times, starting at a number of different starting values. Ideally, these should be quite dispersed over the interesting regions of the parameter space. This is easy to achieve in a one-dimensional case like the one shown in Figure 1.3. However, it is not feasible when $\boldsymbol{\beta}$ has more than a few elements: If we want to try just 10 starting values for each of k parameters, the total number of starting values is 10^k. Thus, in practice, the starting values generally cover only a very small fraction of the parameter space. Nevertheless, if several different starting values all lead to the same

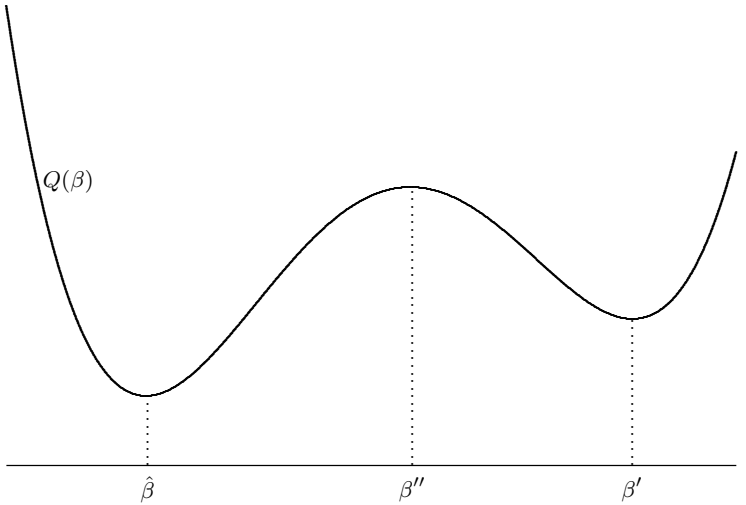


Figure 1.3 A criterion function with multiple minima

local minimum $\hat{\boldsymbol{\beta}}$, with $Q(\hat{\boldsymbol{\beta}})$ less than the value of $Q(\boldsymbol{\beta})$ observed at any other local minimum, then it is plausible, but by no means certain, that $\hat{\boldsymbol{\beta}}$ is actually the global minimum.

Numerous more formal methods of dealing with multiple minima have been proposed. See, among others, Veall (1990), Goffe, Ferrier, and Rogers (1994), Dorsey and Mayer (1995), and Andrews (1997). In difficult cases, one or more of these methods should work better than simply using a number of starting values. However, they tend to be computationally expensive, and none of them works well in every case.

Many of the difficulties of computing NLS estimates are related to the identification of the model parameters by different data sets. The identification condition for NLS is rather different from the identification condition for the estimators discussed in Section 1.2. For NLS, it is simply the requirement that the function $\mathrm{SSR}(\boldsymbol{\beta})$ should have a unique minimum with respect to $\boldsymbol{\beta}$. This is not at all the same requirement as the condition that the estimating equations (1.28) should have a unique solution. In the example of Figure 1.3, the estimating equations, which for NLS are first-order conditions, are satisfied not only at the local minima $\hat{\beta}$ and β', but also at the local maximum β''. However, $\hat{\beta}$ is the unique global minimum of $\mathrm{SSR}(\beta)$, and so β is identified by the NLS estimator.

The analog for NLS of the strong asymptotic identification condition that $\boldsymbol{S}_{\boldsymbol{W}^\top \boldsymbol{X}}$ should be nonsingular is the condition that $\boldsymbol{S}_{\boldsymbol{X}^\top \boldsymbol{X}}$ should be nonsingular, since the variables $\boldsymbol{W}$ of the estimator are replaced by $\boldsymbol{X}_0$ for NLS. The strong condition for identification by a given data set is simply that the

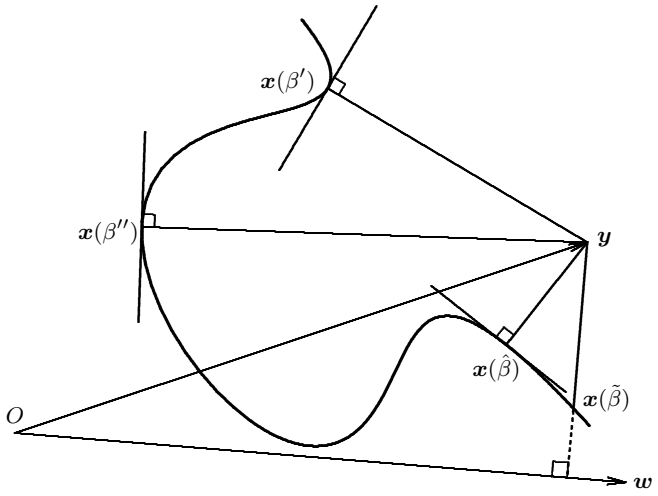


Figure 1.4 NLS and MM estimation of a nonlinear model

matrix $\hat{\boldsymbol{X}}^\top\hat{\boldsymbol{X}}$ should be nonsingular, and therefore positive definite. It is easy to see that this condition is just the sufficient second-order condition for a minimum of the sum-of-squares function at $\hat{\boldsymbol{\beta}}$.

The Geometry of Nonlinear Regression

For nonlinear regression models, it is not possible, in general, to draw faithful geometrical representations of the estimation procedure in just two or three dimensions, as we can for linear models. Nevertheless, it is often useful to illustrate the concepts involved in nonlinear estimation geometrically, as we do in Figure 1.4. Although the vector $\boldsymbol{x}(\boldsymbol{\beta})$ lies in E^n, we have supposed for the purposes of the figure that, as the scalar parameter β varies, $\boldsymbol{x}(\beta)$ traces out a curve that we can visualize in the plane of the page. If the model were linear, $\boldsymbol{x}(\beta)$ would trace out a straight line rather than a curve. In the same way, the dependent variable $\boldsymbol{y}$ is represented by a point in the plane of the page, or, more accurately, by the vector in that plane joining the origin to that point.

For NLS, we seek the point on the curve generated by $\boldsymbol{x}(\beta)$ that is closest in Euclidean distance to $\boldsymbol{y}$. We see from the figure that, although the first-order conditions are satisfied at three points, only one of them yields the NLS estimator. Geometrically, the sum-of-squares function is just the square of the Euclidean distance from $\boldsymbol{y}$ to $\boldsymbol{x}(\beta)$. Its global minimum is achieved at $\boldsymbol{x}(\hat{\beta})$, not at either $\boldsymbol{x}(\beta')$ or $\boldsymbol{x}(\beta'')$.

We can also use Figure 1.4 to see how estimation with a fixed matrix $\boldsymbol{W}$ works. Since there is just one parameter, we need a single variable $\boldsymbol{w}$ that does not depend on the model parameters, and such a variable is shown in the figure. The estimating equation defining the estimator is that the residuals should be orthogonal to $\boldsymbol{w}$. It can be seen that this condition is satisfied only by the residual vector $\boldsymbol{y} - \boldsymbol{x}(\tilde{\beta})$. In the figure, a dotted line is drawn continuing this residual vector so as to show that it is indeed orthogonal to $\boldsymbol{w}$. There are cases, like the one in the figure, in which the NLS first-order conditions can be satisfied for more than one value of $\boldsymbol{\beta}$ while the conditions for estimation are satisfied for just one value, and there are cases in which the reverse is true. Readers are invited to use their geometrical imaginations.

Instrumental Variables

The results of Section 1.2 are almost all conserved, subject to slight modifications, if some or all of the regression functions contain endogenous explanatory variables. Nonlinear least squares no longer gives a consistent estimator, for the same reasons that OLS fails to be consistent with endogenous regressors. But if a matrix $\boldsymbol{W}$ can be found that satisfies the condition (1.09), the estimating equations (1.10) yield a consistent estimator. When $\boldsymbol{W}$ has more columns than the parameter vector $\boldsymbol{\beta}$ has elements, over-identified estimation works just as in the linear case, and leads to the following estimating equations:

$$\boldsymbol{X}^\top(\boldsymbol{\beta})\boldsymbol{P}_{\boldsymbol{W}}(\boldsymbol{y} - \boldsymbol{x}(\boldsymbol{\beta})) = \boldsymbol{0}. \tag{1.46}$$

A consistent estimate of the covariance matrix of the estimator $\hat{\boldsymbol{\beta}}_{\mathrm{IV}}$ that solves (1.46) is given by

$$\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}_{\mathrm{IV}}) = s^2(\hat{\boldsymbol{X}}^\top\boldsymbol{P}_{\boldsymbol{W}}\hat{\boldsymbol{X}})^{-1}, \tag{1.47}$$

where $\hat{\boldsymbol{X}} \equiv \boldsymbol{X}(\hat{\boldsymbol{\beta}}_{\mathrm{IV}})$.

1.5 The Gauss-Newton Regression

When the function we are trying to minimize is a sum-of-squares function, we can obtain explicit expressions for the gradient and the Hessian used in Newton's Method. It is convenient to write the criterion function itself as $\mathrm{SSR}(\boldsymbol{\beta})$ divided by the sample size n:

$$Q(\boldsymbol{\beta}) = n^{-1}\mathrm{SSR}(\boldsymbol{\beta}) = \frac{1}{n}\sum_{t=1}^{n}\big(y_t - x_t(\boldsymbol{\beta})\big)^2.$$

Therefore, using the fact that the partial derivative of $x_t(\boldsymbol{\beta})$ with respect to β_i is $X_{ti}(\boldsymbol{\beta})$, we find that the i^{th} element of the gradient is

$$g_i(\boldsymbol{\beta}) = -\frac{2}{n}\sum_{t=1}^{n} X_{ti}(\boldsymbol{\beta})\big(y_t - x_t(\boldsymbol{\beta})\big).$$

The gradient can be written more compactly in vector-matrix notation as

$$\boldsymbol{g}(\boldsymbol{\beta}) = -2n^{-1}\boldsymbol{X}^\top(\boldsymbol{\beta})\big(\boldsymbol{y} - \boldsymbol{x}(\boldsymbol{\beta})\big). \tag{1.48}$$

Similarly, it can be shown that the Hessian $\boldsymbol{H}(\boldsymbol{\beta})$ has typical element

$$H_{ij}(\boldsymbol{\beta}) = -\frac{2}{n}\sum_{t=1}^{n}\Big(\big(y_t - x_t(\boldsymbol{\beta})\big)\frac{\partial X_{ti}(\boldsymbol{\beta})}{\partial \beta_j} - X_{ti}(\boldsymbol{\beta})X_{tj}(\boldsymbol{\beta})\Big). \tag{1.49}$$

When this expression is evaluated at $\boldsymbol{\beta}_0$, it is asymptotically equivalent to

$$\frac{2}{n}\sum_{t=1}^{n} X_{ti}(\boldsymbol{\beta}_0)X_{tj}(\boldsymbol{\beta}_0). \tag{1.50}$$

The reason for this asymptotic equivalence is that, since $y_t = x_t(\boldsymbol{\beta}_0) + u_t$, the first term inside the large parentheses in (1.49) becomes

$$-\frac{2}{n}\sum_{t=1}^{n}\frac{\partial X_{ti}(\boldsymbol{\beta})}{\partial \beta_j}\,u_t. \tag{1.51}$$

Because $x_t(\boldsymbol{\beta})$ and all its first- and second-order derivatives belong to Ω_t, the expectation of each term in (1.51) is 0. Therefore, by a law of large numbers, expression (1.51) tends to 0 as $n \to \infty$.

Gauss-Newton Methods

The above results make it clear that a natural choice for $\boldsymbol{D}(\boldsymbol{\beta})$ in a quasi-Newton minimization algorithm based on (1.44) is

$$\boldsymbol{D}(\boldsymbol{\beta}) = 2n^{-1}\boldsymbol{X}^\top(\boldsymbol{\beta})\boldsymbol{X}(\boldsymbol{\beta}). \tag{1.52}$$

By construction, this $\boldsymbol{D}(\boldsymbol{\beta})$ is positive definite whenever $\boldsymbol{X}(\boldsymbol{\beta})$ has full rank. Substituting equations (1.52) and (1.48) into equation (1.44) yields

$$\begin{aligned}\boldsymbol{\beta}_{(j+1)} &= \boldsymbol{\beta}_{(j)} + \alpha_{(j)}\big(2n^{-1}\boldsymbol{X}_{(j)}^\top\boldsymbol{X}_{(j)}\big)^{-1}\big(2n^{-1}\boldsymbol{X}_{(j)}^\top(\boldsymbol{y} - \boldsymbol{x}_{(j)})\big)\\ &= \boldsymbol{\beta}_{(j)} + \alpha_{(j)}\big(\boldsymbol{X}_{(j)}^\top\boldsymbol{X}_{(j)}\big)^{-1}\boldsymbol{X}_{(j)}^\top(\boldsymbol{y} - \boldsymbol{x}_{(j)}).\end{aligned} \tag{1.53}$$

The classic **Gauss-Newton method** would set $\alpha_{(j)} = 1$, so that

$$\boldsymbol{\beta}_{(j+1)} = \boldsymbol{\beta}_{(j)} + \big(\boldsymbol{X}_{(j)}^\top\boldsymbol{X}_{(j)}\big)^{-1}\boldsymbol{X}_{(j)}^\top(\boldsymbol{y} - \boldsymbol{x}_{(j)}), \tag{1.54}$$

but it is generally better to use a good one-dimensional search routine to choose α optimally at each iteration. This modified type of Gauss-Newton procedure often works quite well in practice.

The second term on the right-hand side of (1.54) can most easily be computed by means of an **artificial regression** called the **Gauss-Newton regression**, or **GNR**. This artificial regression can be expressed as follows:

$$\boldsymbol{y} - \boldsymbol{x}(\boldsymbol{\beta}) = \boldsymbol{X}(\boldsymbol{\beta})\boldsymbol{b} + \text{residuals}. \tag{1.55}$$

This is the simplest version of the Gauss-Newton regression. It is called "artificial" because the variables that appear in it are not the dependent and explanatory variables of the nonlinear regression (1.02). Instead, they are functions of these variables and of the model parameters. Before (1.55) can be run as a regression, it is necessary to choose the parameter vector $\boldsymbol{\beta}$ at which the regressand and regressors are to be evaluated.

The regressand in (1.55) is the difference between the actual values of the dependent variable and the values predicted by the regression function $\boldsymbol{x}(\boldsymbol{\beta})$ evaluated at the chosen $\boldsymbol{\beta}$. There are k regressors, each of which is a vector of derivatives of $\boldsymbol{x}(\boldsymbol{\beta})$ with respect to one of the elements of $\boldsymbol{\beta}$. It therefore makes sense to think of the i^{th} regressor as being associated with β_i. The vector $\boldsymbol{b}$ is a vector of artificial parameters, and we write "+ residuals" rather than the usual "+ $\boldsymbol{u}$" to emphasize the fact that (1.55) is not a statistical model in the usual sense.

The connection between the Gauss-Newton method of numerical optimization and the Gauss-Newton regression should now be clear. If the variables in (1.55) are evaluated at $\boldsymbol{\beta}_{(j)}$, the OLS parameter estimates of the artificial parameters are

$$\boldsymbol{b}_{(j)} = \big(\boldsymbol{X}_{(j)}^\top\boldsymbol{X}_{(j)}\big)^{-1}\boldsymbol{X}_{(j)}^\top(\boldsymbol{y} - \boldsymbol{x}_{(j)}),$$

from which it follows using (1.53) that the Gauss-Newton method gives

$$\boldsymbol{\beta}_{(j+1)} = \boldsymbol{\beta}_{(j)} + \alpha_{(j)}\boldsymbol{b}_{(j)}.$$

Thus the GNR conveniently and cheaply performs two of the operations necessary for a step of the Gauss-Newton method. It yields a matrix which approximates the Hessian of SSR($\boldsymbol{\beta}$) and is always positive semidefinite. In addition, it computes a vector of artificial parameter estimates which is equal to $-\boldsymbol{D}_{(j)}^{-1}\boldsymbol{g}_{(j)}$, the direction in which the algorithm looks at iteration j.

One potential difficulty with the Gauss-Newton method is that the matrix $\boldsymbol{X}^\top(\boldsymbol{\beta})\boldsymbol{X}(\boldsymbol{\beta})$ may sometimes be very close to singular, even though the model is reasonably well identified by the data. If the strong identification condition is satisfied by a given data set, then $\hat{\boldsymbol{X}}^\top\hat{\boldsymbol{X}}$ is positive definite. However, when $\boldsymbol{X}^\top(\boldsymbol{\beta})\boldsymbol{X}(\boldsymbol{\beta})$ is evaluated far away from $\hat{\boldsymbol{\beta}}$, it may well be close to singular. When that happens, the algorithm gets into trouble, because $\boldsymbol{b}$ no longer lies in the same k–dimensional space as $\boldsymbol{\beta}$, but rather in a subspace of dimension equal to the effective rank of $\boldsymbol{X}^\top(\boldsymbol{\beta})\boldsymbol{X}(\boldsymbol{\beta})$. In this event, a Gauss-Newton algorithm can cycle indefinitely without making any progress. The best algorithms for nonlinear least squares check whether this is happening and replace $\boldsymbol{X}^\top(\boldsymbol{\beta})\boldsymbol{X}(\boldsymbol{\beta})$ with another estimate of $\boldsymbol{H}(\boldsymbol{\beta})$ whenever it does. See the references cited at the beginning of Section 1.4.

Properties of the GNR

As we have seen, when $\boldsymbol{x}(\boldsymbol{\beta})$ is a linear regression model with $\boldsymbol{X}$ being the matrix of independent variables, $\boldsymbol{X}(\boldsymbol{\beta})$ is simply equal to $\boldsymbol{X}$. Thus, in the case of a linear regression model, the GNR is simply a regression of the vector $\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}$ on $\boldsymbol{X}$. A special feature of the GNR for linear models is that the classic Gauss-Newton method converges in one step from an arbitrary starting point. To see this, let $\boldsymbol{\beta}_{(0)}$ be the starting point. The GNR is

$$\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_{(0)} = \boldsymbol{X}\boldsymbol{b} + \text{residuals},$$

and the artificial parameter estimates are

$$\hat{\boldsymbol{b}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_{(0)}) = \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_{(0)},$$

where $\hat{\boldsymbol{\beta}}$ is the OLS estimator. It follows at once that

$$\boldsymbol{\beta}_{(1)} = \boldsymbol{\beta}_{(0)} + \hat{\boldsymbol{b}} = \hat{\boldsymbol{\beta}}. \tag{1.56}$$

This property has a very useful analog for nonlinear models that we will explore in the next section.

The properties of the GNR (1.55) depend on the choice of $\boldsymbol{\beta}$. One interesting choice is $\hat{\boldsymbol{\beta}}$, the vector of NLS parameter estimates. With this choice, regression (1.55) becomes

$$\boldsymbol{y} - \hat{\boldsymbol{x}} = \hat{\boldsymbol{X}}\boldsymbol{b} + \text{residuals}, \tag{1.57}$$

where $\hat{\boldsymbol{x}} \equiv \boldsymbol{x}(\hat{\boldsymbol{\beta}})$ and $\hat{\boldsymbol{X}} \equiv \boldsymbol{X}(\hat{\boldsymbol{\beta}})$. The OLS estimate of $\boldsymbol{b}$ from (1.57) is

$$\hat{\boldsymbol{b}} = (\hat{\boldsymbol{X}}^\top\hat{\boldsymbol{X}})^{-1}\hat{\boldsymbol{X}}^\top(\boldsymbol{y} - \hat{\boldsymbol{x}}). \tag{1.58}$$

Because $\hat{\boldsymbol{\beta}}$ must satisfy the first-order conditions (1.28), the factor $\hat{\boldsymbol{X}}^\top(\boldsymbol{y} - \hat{\boldsymbol{x}})$ must be a zero vector. Therefore, $\hat{\boldsymbol{b}} = \boldsymbol{0}$, and the GNR (1.57) can have no explanatory power whatsoever.

This may seem an uninteresting result. After all, why would anyone want to run an artificial regression all the coefficients of which are known in advance to be zero? There are in fact two very good reasons for doing so.

The first reason is to check that the vector $\hat{\boldsymbol{\beta}}$ reported by a program for NLS estimation really does satisfy the first-order conditions (1.28). Computer programs use many different techniques for calculating NLS estimates, and many programs do not yield reliable answers in every case; see McCullough (1999). By running the GNR (1.57), we can see whether the first-order conditions are satisfied reasonably accurately. If all the t statistics are less than about 10^{-4}, and the R^2 is less than about 10^{-8}, then the value of $\hat{\boldsymbol{\beta}}$ reported by the program should be reasonably accurate. If not, there may be a problem. Possibly the estimation should be performed again using a tighter convergence criterion, possibly we should switch to a more accurate program, or possibly the model in question simply cannot be estimated reliably with the data set we are using. Of course, some programs run the GNR (1.57) and perform the requisite checks automatically. Once we have verified that they do so, we need not bother doing it ourselves.

Computing Covariance Matrices

The second reason to run the GNR (1.57) is to calculate an estimate of $\text{Var}(\hat{\boldsymbol{\beta}})$. The usual OLS covariance matrix from this regression is, by (F4.64),

$$\widehat{\text{Var}}(\hat{\boldsymbol{b}}) = s^2(\hat{\boldsymbol{X}}^\top\hat{\boldsymbol{X}})^{-1}, \tag{1.59}$$

where, since the regressors have no explanatory power, s^2 is the same as the one defined in equation (1.34). It is equal to the SSR from the original nonlinear regression, divided by $n - k$. Evidently, the right-hand side of equation (1.59) is identical to the right-hand side of equation (1.33), which is the standard estimator of $\text{Var}(\hat{\boldsymbol{\beta}})$. Thus running the GNR (1.57) provides an easy way to calculate $\widehat{\text{Var}}(\hat{\boldsymbol{\beta}})$.

Good programs for NLS estimation normally use the right-hand side of equation (1.59) to estimate the covariance matrix of $\hat{\boldsymbol{\beta}}$. Not all programs can be relied upon to do this, however, and running the GNR (1.57) is a simple way to check whether they do so and get better estimates if they do not. Sometimes, $\hat{\boldsymbol{\beta}}$ may be obtained by a method other than fully nonlinear estimation. For example, the regression function may be linear conditional on one parameter, and NLS estimates may be obtained by searching over that parameter and performing OLS estimation conditional on it. In such a case, it will be necessary to calculate the matrix (1.59) explicitly, and running the GNR (1.57) is an easy way to do so.

The GNR (1.57) can also be used to compute a heteroskedasticity-consistent covariance matrix estimate. Any HCCME for the parameters $\hat{\boldsymbol{b}}$ of the GNR is also perfectly valid for $\hat{\boldsymbol{\beta}}$. To see this, we start from the result (1.39), rewritten as

$$\operatorname*{plim}_{n\to\infty}\left[n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) - (\boldsymbol{S}_{\boldsymbol{X}^\top\boldsymbol{X}})^{-1}n^{-1/2}\boldsymbol{X}_0^\top\boldsymbol{u}\right] = \boldsymbol{0}.$$

If $\text{E}(\boldsymbol{u}\boldsymbol{u}^\top) = \boldsymbol{\Omega}$, then

$$n^{-1/2}\boldsymbol{X}_0^\top\boldsymbol{u} \xrightarrow{d} \text{N}(\boldsymbol{0}, n^{-1}\boldsymbol{X}_0^\top\boldsymbol{\Omega}\boldsymbol{X}_0),$$

and this shows that

$$n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{d} \text{N}(\boldsymbol{0}, (\boldsymbol{S}_{\boldsymbol{X}^\top\boldsymbol{X}})^{-1}n^{-1}\boldsymbol{X}_0^\top\boldsymbol{\Omega}\boldsymbol{X}_0(\boldsymbol{S}_{\boldsymbol{X}^\top\boldsymbol{X}})^{-1}).$$

Therefore, a reasonable way to estimate $\text{Var}(\hat{\boldsymbol{\beta}})$ is to use the sandwich covariance matrix

$$\widehat{\text{Var}}_{\text{h}}(\hat{\boldsymbol{\beta}}) \equiv (\hat{\boldsymbol{X}}^\top\hat{\boldsymbol{X}})^{-1}\hat{\boldsymbol{X}}^\top\hat{\boldsymbol{\Omega}}\hat{\boldsymbol{X}}(\hat{\boldsymbol{X}}^\top\hat{\boldsymbol{X}})^{-1}, \tag{1.60}$$

where $\hat{\boldsymbol{\Omega}}$ is an $n \times n$ diagonal matrix with the squared residual $\hat{u}_t^2$ as the t^{th} diagonal element. This is precisely the HCCME (F6.29) for the GNR (1.57). Of course, $\hat{\boldsymbol{\Omega}}$ can, and probably should, be replaced by a modified version with better finite-sample properties.

1.6 One-Step Estimation

The result (1.56) for linear regression models has a counterpart for nonlinear models: If we start with estimates that are root-n consistent but inefficient, a single Newton, or quasi-Newton, step is all that is needed to obtain estimates that are asymptotically equivalent to NLS estimates. This important result may initially seem astonishing, but the intuition behind it is not difficult.

Let $\acute{\boldsymbol{\beta}}$ denote the initial root-n consistent estimates. The GNR (1.55) evaluated at these estimates is

$$\boldsymbol{y} - \acute{\boldsymbol{x}} = \acute{\boldsymbol{X}}\boldsymbol{b} + \text{residuals},$$

where $\acute{\boldsymbol{x}} \equiv \boldsymbol{x}(\acute{\boldsymbol{\beta}})$ and $\acute{\boldsymbol{X}} \equiv \boldsymbol{X}(\acute{\boldsymbol{\beta}})$. The estimate of $\boldsymbol{b}$ from this regression is

$$\acute{\boldsymbol{b}} = (\acute{\boldsymbol{X}}^\top \acute{\boldsymbol{X}})^{-1} \acute{\boldsymbol{X}}^\top (\boldsymbol{y} - \acute{\boldsymbol{x}}). \tag{1.61}$$

Then a **one-step estimator** is defined by the equation

$$\grave{\boldsymbol{\beta}} = \acute{\boldsymbol{\beta}} + \acute{\boldsymbol{b}}. \tag{1.62}$$

This one-step estimator turns out to be asymptotically equivalent to the NLS estimator $\hat{\boldsymbol{\beta}}$, by which we mean that the difference between $n^{1/2}(\grave{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ and $n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ tends to zero as $n \to \infty$. In other words, after both are centered and multiplied by $n^{1/2}$, the one-step estimator $\grave{\boldsymbol{\beta}}$ and the NLS estimator $\hat{\boldsymbol{\beta}}$ tend to the same random variable asymptotically. In particular, this means that the asymptotic covariance matrix of $\grave{\boldsymbol{\beta}}$ is the same as that of $\hat{\boldsymbol{\beta}}$. Thus $\grave{\boldsymbol{\beta}}$ shares with $\hat{\boldsymbol{\beta}}$ the property of asymptotic efficiency. For this reason, $\grave{\boldsymbol{\beta}}$ is sometimes called a **one-step efficient estimator**.

In order to demonstrate the asymptotic equivalence of $\grave{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}$, we begin by Taylor expanding the expression $n^{-1/2}\acute{\boldsymbol{X}}^\top(\boldsymbol{y} - \acute{\boldsymbol{x}})$ around $\boldsymbol{\beta} = \boldsymbol{\beta}_0$. This yields

$$n^{-1/2}\acute{\boldsymbol{X}}^\top(\boldsymbol{y} - \acute{\boldsymbol{x}}) = n^{-1/2}\boldsymbol{X}_0^\top(\boldsymbol{y} - \boldsymbol{x}_0) + \boldsymbol{\Delta}(\bar{\boldsymbol{\beta}})\, n^{1/2}(\acute{\boldsymbol{\beta}} - \boldsymbol{\beta}_0), \tag{1.63}$$

where $\boldsymbol{x}_0 \equiv \boldsymbol{x}(\boldsymbol{\beta}_0)$, $\bar{\boldsymbol{\beta}}$ is a parameter vector that satisfies (1.20), in the sense explained just after that equation, with $\acute{\boldsymbol{\beta}}$ in place of $\hat{\boldsymbol{\beta}}$, and $\boldsymbol{\Delta}(\boldsymbol{\beta})$ is the $k \times k$ matrix with typical element

$$\begin{aligned}\Delta_{ij}(\boldsymbol{\beta}) &\equiv \frac{\partial}{\partial \beta_j}\left(\frac{1}{n}\sum_{t=1}^{n} X_{ti}(\boldsymbol{\beta})\big(y_t - x_t(\boldsymbol{\beta})\big)\right) \\ &= -\frac{1}{n}\sum_{t=1}^{n} X_{ti}(\boldsymbol{\beta})X_{tj}(\boldsymbol{\beta}) + \frac{1}{n}\sum_{t=1}^{n}\frac{\partial X_{ti}(\boldsymbol{\beta})}{\partial \beta_j}\big(y_t - x_t(\boldsymbol{\beta})\big). \end{aligned} \tag{1.64}$$

It can be shown that, when (1.64) is evaluated at $\bar{\boldsymbol{\beta}}$, or at any root-n consistent estimator of $\boldsymbol{\beta}_0$, the second term tends to zero but the first term does not. We

have seen why this is so if we evaluate (1.64) at $\boldsymbol{\beta}_0$. In that case, the second term, like expression (1.51), becomes an average of quantities each of which has expectation zero, while the first term is an average of quantities each of which has a nonzero expectation. Essentially the same result holds when we evaluate (1.64) at any root-n consistent estimator. Thus we conclude that

$$\boldsymbol{\Delta}(\bar{\boldsymbol{\beta}}) \stackrel{a}{=} -n^{-1}\bar{\boldsymbol{X}}^\top\bar{\boldsymbol{X}} \stackrel{a}{=} -n^{-1}\boldsymbol{X}_0^\top\boldsymbol{X}_0, \tag{1.65}$$

where the second equality is also a consequence of the consistency of $\bar{\boldsymbol{\beta}}$.

Using the result (1.65) in (1.63) shows that

$$n^{-1/2}\acute{\boldsymbol{X}}^\top(\boldsymbol{y} - \acute{\boldsymbol{x}}) \stackrel{a}{=} n^{-1/2}\boldsymbol{X}_0^\top\boldsymbol{u} - n^{-1}\boldsymbol{X}_0^\top\boldsymbol{X}_0\, n^{1/2}(\acute{\boldsymbol{\beta}} - \boldsymbol{\beta}_0),$$

which can be solved to yield

$$\begin{aligned} n^{1/2}(\acute{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) &\stackrel{a}{=} (n^{-1}\boldsymbol{X}_0^\top\boldsymbol{X}_0)^{-1}\big(n^{-1/2}\boldsymbol{X}_0^\top\boldsymbol{u} - n^{-1/2}\acute{\boldsymbol{X}}^\top(\boldsymbol{y} - \acute{\boldsymbol{x}})\big) \\ &\stackrel{a}{=} (\boldsymbol{S}_{\boldsymbol{X}^\top\boldsymbol{X}})^{-1}n^{-1/2}\boldsymbol{X}_0^\top\boldsymbol{u} - (\boldsymbol{S}_{\boldsymbol{X}^\top\boldsymbol{X}})^{-1}n^{-1/2}\acute{\boldsymbol{X}}^\top(\boldsymbol{y} - \acute{\boldsymbol{x}}). \end{aligned} \tag{1.66}$$

By (1.39), the first term in the second line here is asymptotically equal to $n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$. By (1.61), the second term is asymptotically equivalent to $-n^{1/2}\acute{\boldsymbol{b}}$. Thus (1.66) implies that

$$n^{1/2}(\acute{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \stackrel{a}{=} n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) - n^{1/2}\acute{\boldsymbol{b}}.$$

Rearranging this and using the definition (1.62), we see that

$$n^{1/2}(\grave{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) = n^{1/2}(\acute{\boldsymbol{\beta}} + \acute{\boldsymbol{b}} - \boldsymbol{\beta}_0) \stackrel{a}{=} n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0), \tag{1.67}$$

which is the result that we wished to show.

Despite the rather complicated asymptotic theory needed to prove (1.67), the fundamental reason that makes a one-step efficient estimator based on the GNR asymptotically equivalent to the NLS estimator is really quite simple. The GNR minimizes a quadratic approximation to $\text{SSR}(\boldsymbol{\beta})$ around $\acute{\boldsymbol{\beta}}$. Asymptotically, the function $\text{SSR}(\boldsymbol{\beta})$ is quadratic in the neighborhood of $\boldsymbol{\beta}_0$. If the sample size is large enough, the consistency of $\acute{\boldsymbol{\beta}}$ implies that we must be taking the quadratic approximation at a point very near $\boldsymbol{\beta}_0$. Therefore, the approximation coincides with $\text{SSR}(\boldsymbol{\beta})$ itself asymptotically.

Although this result is of great theoretical interest, it is typically of limited practical utility with modern computing equipment. Once the GNR, or some other method for taking Newton or quasi-Newton steps, has been programmed for a particular model, we might as well let it iterate to convergence, because the savings in computer time from stopping after a single step are rarely substantial. Moreover, a one-step estimator is consistent if and only if we start from an initial estimator that is consistent, while NLS is consistent no matter where we start from, provided we converge to a global minimum of $\text{SSR}(\boldsymbol{\beta})$. Therefore, it may well require more effort on the part of the investigator to obtain one-step estimates than to obtain NLS ones.

One-step estimators may be useful when the sample size is very large and each step in the minimization process is, perhaps in consequence, very expensive. The large sample size often ensures that the initial, consistent estimates are reasonably close to the NLS ones. If they are, then the one-step estimates should be very close to the latter. One-step estimators can also be useful when the estimation needs to be repeated many times, as is often required by the bootstrap and other simulation-based methods. Bootstrap methods that use one-step estimators are discussed by Davidson and MacKinnon (1999a).

The Linear Regression Model with AR(1) Disturbances

An excellent example of one-step efficient estimation is provided by the model (1.06), which is a linear regression model with AR(1) disturbances. The GNR that corresponds to (1.06) is

$$\begin{aligned} y_t - \rho y_{t-1} - \boldsymbol{X}_t\boldsymbol{\beta} + \rho\boldsymbol{X}_{t-1}\boldsymbol{\beta} \\ = (\boldsymbol{X}_t - \rho\boldsymbol{X}_{t-1})\boldsymbol{b} + b_\rho(y_{t-1} - \boldsymbol{X}_{t-1}\boldsymbol{\beta}) + \text{residual}, \end{aligned} \tag{1.68}$$

where $\boldsymbol{b}$ corresponds to $\boldsymbol{\beta}$ and b_ρ corresponds to ρ. As with every GNR, the regressand is y_t minus the regression function for (1.06). The last regressor, which is the derivative of the regression function with respect to ρ, looks very much like a lagged residual from the original linear regression model (1.05). The remaining k regressors are the derivatives of the regression function with respect to the elements of $\boldsymbol{\beta}$.

It is easy to obtain root-n consistent estimates of the parameters ρ and $\boldsymbol{\beta}$ of the model (1.06), because it can be written as a linear regression subject to nonlinear restrictions on its parameters. The linear regression is

$$y_t = \rho y_{t-1} + \boldsymbol{X}_t\boldsymbol{\beta} + \boldsymbol{X}_{t-1}\boldsymbol{\gamma} + \varepsilon_t. \tag{1.69}$$

If we impose the nonlinear restrictions that $\boldsymbol{\gamma} + \rho\boldsymbol{\beta} = \boldsymbol{0}$, this regression is just (1.06). Thus the model (1.06) is a special case of the model (1.69). Therefore, if (1.06) is a correctly specified model, that is, if the true DGP is a special case of (1.06), then (1.69) must be a correctly specified model as well, because every DGP in (1.06) automatically belongs to (1.69). Since (1.69) is correctly specified, the standard theory of the linear regression with predetermined regressors applies to it, with the consequence that the OLS estimates $\acute{\rho}$ and $\acute{\boldsymbol{\beta}}$ obtained from (1.69) are root-n consistent.

If we evaluate the variables of the GNR (1.68) at $\acute{\rho}$ and $\acute{\boldsymbol{\beta}}$, we obtain

$$\begin{aligned} y_t - \acute{\rho} y_{t-1} - \boldsymbol{X}_t\acute{\boldsymbol{\beta}} + \acute{\rho}\boldsymbol{X}_{t-1}\acute{\boldsymbol{\beta}} \\ = (\boldsymbol{X}_t - \acute{\rho}\boldsymbol{X}_{t-1})\boldsymbol{b} + b_\rho(y_{t-1} - \boldsymbol{X}_{t-1}\acute{\boldsymbol{\beta}}) + \text{residual}. \end{aligned} \tag{1.70}$$

We can run this regression to obtain the artificial parameter estimates $\acute{\boldsymbol{b}}$ and $\acute{b}_\rho$, and the one-step efficient estimates are just $\acute{\boldsymbol{\beta}} + \acute{\boldsymbol{b}}$ and $\acute{\rho} + \acute{b}_\rho$.

1.7 Hypothesis Testing

Hypotheses about the parameters of nonlinear regression models can be formulated in much the same way as hypotheses about the parameters of linear regression models. Let us partition the parameter vector $\boldsymbol{\beta}$ as $\boldsymbol{\beta} = [\boldsymbol{\beta}_1 \vdots \boldsymbol{\beta}_2]$, where $\boldsymbol{\beta}_1$ is $k_1 \times 1$, $\boldsymbol{\beta}_2$ is $k_2 \times 1$, and $\boldsymbol{\beta}$ is $k \times 1$, with $k = k_1 + k_2$. Then the generic nonlinear regression model (1.02) can be written as

$$\boldsymbol{y} = \boldsymbol{x}(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) + \boldsymbol{u}, \quad \boldsymbol{u} \sim \text{IID}(\boldsymbol{0}, \sigma^2\boldsymbol{\mathrm{I}}).$$

If we wish to test the hypothesis that $\boldsymbol{\beta}_2 = \boldsymbol{0}$, we can set up the models that correspond to the null and alternative hypotheses as follows:

$$H_0\colon \quad \boldsymbol{y} = \boldsymbol{x}(\boldsymbol{\beta}_1, \boldsymbol{0}) + \boldsymbol{u}; \tag{1.71}$$

$$H_1\colon \quad \boldsymbol{y} = \boldsymbol{x}(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) + \boldsymbol{u}. \tag{1.72}$$

Here, H_0 denotes the null hypothesis, and H_1 denotes the alternative.

If the regression models (1.71) and (1.72) were linear, we could test the null hypothesis by means of the F statistic (F5.27). In fact, we can do this even though they are nonlinear. The test statistic

$$F_{\boldsymbol{\beta}_2} \equiv \frac{(\text{RSSR} - \text{USSR})/r}{\text{USSR}/(n-k)} \tag{1.73}$$

is computed in exactly the same way as (F5.27), but with RSSR and USSR the sums of squared residuals from NLS estimation of (1.71) and (1.72), respectively. Here $r = k_2$, since the hypothesis that $\boldsymbol{\beta}_2 = \boldsymbol{0}$ imposes k_2 restrictions. It is not difficult to show that (1.73) is asymptotically valid: Under the null hypothesis, it follows the $F(r, \infty)$ distribution asymptotically.

First, we establish some notation. Let $\boldsymbol{X}(\boldsymbol{\beta})$ denote the $n \times k$ matrix of partial derivatives of the vector of regression functions $\boldsymbol{x}(\boldsymbol{\beta}) = \boldsymbol{x}(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$ of (1.72). Similarly, let $\boldsymbol{X}_1(\boldsymbol{\beta})$ and $\boldsymbol{X}_2(\boldsymbol{\beta})$ denote the $n \times k_1$ and $n \times k_2$ submatrices of partial derivatives with respect to the components of $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$, respectively. Finally, let $\boldsymbol{M}_1$ denote the orthogonal projection on to $\mathcal{S}^\perp(\boldsymbol{X}(\boldsymbol{\beta}_0))$, which we previously called $\boldsymbol{M}_{\boldsymbol{X}_0}$, and let $\boldsymbol{M}_0$ denote the orthogonal projection on to $\mathcal{S}^\perp(\boldsymbol{X}_1(\boldsymbol{\beta}_0))$. The projection $\boldsymbol{M}_0$ corresponds to the null hypothesis H_0, and the projection $\boldsymbol{M}_1$ corresponds to the alternative hypothesis H_1.

By the result (1.41), under both the null and alternative hypotheses, the vector of residuals $\hat{\boldsymbol{u}}$ from NLS estimation of H_1 is asymptotically equal to $\boldsymbol{M}_1\boldsymbol{u}$. By essentially the same argument, under the null hypothesis, the vector of residuals $\tilde{\boldsymbol{u}}$ from NLS estimation of H_0 is asymptotically equal to $\boldsymbol{M}_0\boldsymbol{u}$. This implies (see Exercise 1.8) that $\hat{\boldsymbol{u}}^\top\hat{\boldsymbol{u}} \stackrel{a}{=} \boldsymbol{u}^\top\boldsymbol{M}_1\boldsymbol{u}$ and $\tilde{\boldsymbol{u}}^\top\tilde{\boldsymbol{u}} \stackrel{a}{=} \boldsymbol{u}^\top\boldsymbol{M}_0\boldsymbol{u}$. Therefore, under H_0, r times the numerator of (1.73) is asymptotically equal to

$$\boldsymbol{u}^\top\boldsymbol{M}_0\boldsymbol{u} - \boldsymbol{u}^\top\boldsymbol{M}_1\boldsymbol{u} = \boldsymbol{u}^\top(\boldsymbol{M}_0 - \boldsymbol{M}_1)\boldsymbol{u} = \boldsymbol{u}^\top(\boldsymbol{P}_1 - \boldsymbol{P}_0)\boldsymbol{u},$$

where $\boldsymbol{P}_0$ and $\boldsymbol{P}_1$ are the projections complementary to $\boldsymbol{M}_0$ and $\boldsymbol{M}_1$. By the result of Part 1, Exercise 3.F18, $\boldsymbol{P}_1 - \boldsymbol{P}_0$ is an orthogonal projection matrix, which projects on to a space of dimension $k - k_1 = k_2$. Thus the numerator of (1.73) is asymptotically σ_0^2 times a χ^2 variable with k_2 degrees of freedom, divided by $r = k_2$; recall Part 1, Exercise 5.F15. The denominator of (1.73) is just a consistent estimate of σ_0^2, and so, under H_0, the statistic (1.73) itself is asymptotically distributed as $F(k_2, \infty) = \chi^2(k_2)/k_2$.

For linear models, it can be seen that the F statistic could be written as (F6.18). Not surprisingly, it is also possible to calculate test statistics to test the hypothesis that $\boldsymbol{\beta}_2 = \boldsymbol{0}$ in the nonlinear model (1.72). This type of test statistic is often called a **Wald statistic**, because the approach was suggested by Wald (1943). It can be written as

$$W_{\boldsymbol{\beta}_2} \equiv \hat{\boldsymbol{\beta}}_2^\top \big(\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}_2)\big)^{-1} \hat{\boldsymbol{\beta}}_2, \tag{1.74}$$

where $\hat{\boldsymbol{\beta}}_2$ is a vector of NLS estimates from the unrestricted model (1.72), and $\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}_2)$ is the NLS estimate of its covariance matrix. This is just a quadratic form in the vector $\hat{\boldsymbol{\beta}}_2$ and the inverse of an estimate of its covariance matrix. When $k_2 = 1$, the signed square root of (1.74) is equivalent to a t statistic. We will see below that the Wald statistic (1.74) is asymptotically equivalent to the F statistic (1.73), except for the factor of $1/k_2$.

Tests Based on the Gauss-Newton Regression

Since the GNR provides a one-step estimator asymptotically equivalent to the NLS estimator, and it also provides the NLS estimate of the covariance matrix of $\hat{\boldsymbol{\beta}}_2$, a statistic asymptotically equivalent to (1.74) can be computed by means of a GNR. This statistic also turns out to be asymptotically equivalent to the F statistic (1.73), except for the factor of $1/k_2$.

The Gauss-Newton regression corresponding to the model (1.72) is

$$\boldsymbol{y} - \boldsymbol{x}(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) = \boldsymbol{X}_1(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)\boldsymbol{b}_1 + \boldsymbol{X}_2(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)\boldsymbol{b}_2 + \text{residuals}, \tag{1.75}$$

where the vector of artificial parameters $\boldsymbol{b}$ has been partitioned as $[\boldsymbol{b}_1 \vdots \boldsymbol{b}_2]$, conformably with the partition of $\boldsymbol{X}(\boldsymbol{\beta})$. If the GNR is to be used to test the null hypothesis that $\boldsymbol{\beta}_2 = \boldsymbol{0}$, the regressand and regressors must be evaluated at parameter estimates which satisfy the null. We will suppose that they are evaluated at the point $\acute{\boldsymbol{\beta}} \equiv [\acute{\boldsymbol{\beta}}_1, \boldsymbol{0}]$, where $\acute{\boldsymbol{\beta}}_1$ may be any root-n consistent estimator of $\boldsymbol{\beta}_1$. Then the one-step estimator of $\boldsymbol{\beta}$ can be written as

$$\acute{\boldsymbol{\beta}} + \acute{\boldsymbol{b}} = \begin{bmatrix} \acute{\boldsymbol{\beta}}_1 + \acute{\boldsymbol{b}}_1 \\ \acute{\boldsymbol{b}}_2 \end{bmatrix}. \tag{1.76}$$

By the results of Section 1.6, $n^{1/2}\acute{\boldsymbol{b}}_2$ is asymptotically equivalent to $n^{1/2}\hat{\boldsymbol{\beta}}_2$ under the null, where $\hat{\boldsymbol{\beta}}_2$ is the NLS estimator of $\boldsymbol{\beta}_2$ from (1.72).

In practice, the two estimators that are most likely to be used for $\acute{\boldsymbol{\beta}}_1$ are $\tilde{\boldsymbol{\beta}}_1$, the restricted NLS estimator, and $\hat{\boldsymbol{\beta}}_1$, a subvector of the unrestricted NLS estimator. Here we are once more adopting the convention, previously used in Part 1, Chapter 5, whereby a tilde denotes restricted estimates and a hat denotes unrestricted ones. Both these estimators are root-n consistent under the null hypothesis, but $\tilde{\boldsymbol{\beta}}_1$ is generally more efficient than $\hat{\boldsymbol{\beta}}_1$. Whether we want to use $\tilde{\boldsymbol{\beta}}_1$, $\hat{\boldsymbol{\beta}}_1$, or some other root-$n$ consistent estimator when performing GNR-based tests depends on how difficult the various estimators are to compute and on the finite-sample properties of the test statistics that result from the various choices.

Now consider the vector of residuals $\acute{\boldsymbol{u}}$ from OLS estimation of the GNR (1.75) evaluated at $\acute{\boldsymbol{\beta}}$, when the true DGP is characterized by the parameter vector $\boldsymbol{\beta}_0 \equiv [\boldsymbol{\beta}_1^0 \vdots \boldsymbol{0}]$. Under the null, we have

$$\begin{aligned} \acute{\boldsymbol{u}} &= \boldsymbol{y} - \boldsymbol{x}(\acute{\boldsymbol{\beta}}_1, \boldsymbol{0}) - \acute{\boldsymbol{X}}_1\acute{\boldsymbol{b}}_1 - \acute{\boldsymbol{X}}_2\acute{\boldsymbol{b}}_2 \\ &= \boldsymbol{y} - \boldsymbol{x}(\boldsymbol{\beta}_1^0, \boldsymbol{0}) - \boldsymbol{X}_1(\bar{\boldsymbol{\beta}})(\acute{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^0) - \acute{\boldsymbol{X}}_1\acute{\boldsymbol{b}}_1 - \acute{\boldsymbol{X}}_2\acute{\boldsymbol{b}}_2 \\ &\overset{a}{=} \boldsymbol{u} - \acute{\boldsymbol{X}}_1(\acute{\boldsymbol{\beta}}_1 + \acute{\boldsymbol{b}}_1 - \boldsymbol{\beta}_1^0) - \acute{\boldsymbol{X}}_2\acute{\boldsymbol{b}}_2. \end{aligned} \tag{1.77}$$

Here, $\bar{\boldsymbol{\beta}}$ is a parameter vector between $\boldsymbol{\beta}_0$ and $\acute{\boldsymbol{\beta}}$. To obtain the asymptotic equality in the last line, we have used the fact that $\boldsymbol{X}_1(\bar{\boldsymbol{\beta}}) \overset{a}{=} \acute{\boldsymbol{X}}_1$. The one-step estimator (1.76) is consistent, and so the last two terms in (1.77) tend to zero as $n \to \infty$. Thus the residuals $\acute{u}_t$ are asymptotically equal to the disturbances u_t, and so $n^{-1}\acute{\boldsymbol{u}}^\top\acute{\boldsymbol{u}}$ is asymptotically equal to σ_0^2, the true variance of the disturbances. In fact, because of the asymptotic equivalence of the one-step estimator $\acute{\boldsymbol{\beta}}$ and the NLS estimator $\hat{\boldsymbol{\beta}}$, (1.77) tells us that $\acute{\boldsymbol{u}} \overset{a}{=} \boldsymbol{u} - \acute{\boldsymbol{X}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$. An argument like that of (1.41) then shows that $\acute{\boldsymbol{u}}$ is asymptotically equivalent to $\boldsymbol{M}_{\boldsymbol{X}_0}\boldsymbol{u}$. For the moment, however, we do not need this more refined result.

The GNR (1.75) evaluated at $\acute{\boldsymbol{\beta}}$ is

$$\boldsymbol{y} - \acute{\boldsymbol{x}} = \acute{\boldsymbol{X}}_1\boldsymbol{b}_1 + \acute{\boldsymbol{X}}_2\boldsymbol{b}_2 + \text{residuals}. \tag{1.78}$$

Since this is a linear regression, we can apply the FWL Theorem to it. Writing $\boldsymbol{M}_{\acute{\boldsymbol{X}}_1}$ for the projection on to $\mathcal{S}^\perp(\acute{\boldsymbol{X}}_1)$, we see that the FWL regression can be written as

$$\boldsymbol{M}_{\acute{\boldsymbol{X}}_1}(\boldsymbol{y} - \acute{\boldsymbol{x}}) = \boldsymbol{M}_{\acute{\boldsymbol{X}}_1}\acute{\boldsymbol{X}}_2\boldsymbol{b}_2 + \text{residuals}.$$

This FWL regression yields the same estimates $\acute{\boldsymbol{b}}_2$ as does (1.78). Thus, inserting the factors of powers of n that are needed for asymptotic analysis, we find that

$$n^{1/2}\acute{\boldsymbol{b}}_2 = (n^{-1}\acute{\boldsymbol{X}}_2^\top\boldsymbol{M}_{\acute{\boldsymbol{X}}_1}\acute{\boldsymbol{X}}_2)^{-1}n^{-1/2}\acute{\boldsymbol{X}}_2^\top\boldsymbol{M}_{\acute{\boldsymbol{X}}_1}(\boldsymbol{y} - \acute{\boldsymbol{x}}). \tag{1.79}$$

In addition to yielding the same parameter estimates $\acute{\boldsymbol{b}}_2$, the FWL regression has the same residuals as regression (1.78) and the same estimated covariance matrix for $\acute{\boldsymbol{b}}_2$. The latter is $\acute{\sigma}^2(\acute{\boldsymbol{X}}_2^\top \boldsymbol{M}_{\acute{\boldsymbol{X}}_1}\acute{\boldsymbol{X}}_2)^{-1}$, where $\acute{\sigma}^2$ is the disturbance variance estimator from (1.78), which, as we just saw, is asymptotically equal to σ_0^2. If $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ denote $\boldsymbol{X}_1(\boldsymbol{\beta}_0)$ and $\boldsymbol{X}_2(\boldsymbol{\beta}_0)$, respectively, we see that

$$\begin{aligned} n^{-1}\acute{\boldsymbol{X}}_2^\top \boldsymbol{M}_{\acute{\boldsymbol{X}}_1}\acute{\boldsymbol{X}}_2 &= n^{-1}\acute{\boldsymbol{X}}_2^\top \acute{\boldsymbol{X}}_2 - n^{-1}\acute{\boldsymbol{X}}_2^\top \acute{\boldsymbol{X}}_1 (n^{-1}\acute{\boldsymbol{X}}_1^\top \acute{\boldsymbol{X}}_1)^{-1} n^{-1}\acute{\boldsymbol{X}}_1^\top \acute{\boldsymbol{X}}_2 \\ &\stackrel{a}{=} n^{-1}\boldsymbol{X}_2^\top \boldsymbol{X}_2 - n^{-1}\boldsymbol{X}_2^\top \boldsymbol{X}_1 (n^{-1}\boldsymbol{X}_1^\top \boldsymbol{X}_1)^{-1} n^{-1}\boldsymbol{X}_1^\top \boldsymbol{X}_2 \\ &= n^{-1}\boldsymbol{X}_2^\top \boldsymbol{M}_{\boldsymbol{X}_1}\boldsymbol{X}_2, \end{aligned}$$

where the asymptotic equality follows, as usual, from the consistency of $\acute{\boldsymbol{\beta}}$. Thus n times the covariance matrix estimator for $\acute{\boldsymbol{b}}_2$ given by the GNR (1.78) provides a consistent estimate of the asymptotic covariance matrix of the vector $n^{1/2}(\hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2^0)$, as would be given by the lower right block of (1.32) if that matrix were partitioned appropriately.

The Wald test statistic (1.74) can be rewritten as

$$n^{1/2}\hat{\boldsymbol{\beta}}_2^\top \big(n\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}_2)\big)^{-1} n^{1/2}\hat{\boldsymbol{\beta}}_2. \tag{1.80}$$

Under the null, this is asymptotically equivalent to the statistic

$$\frac{1}{\acute{\sigma}^2} n^{1/2}\acute{\boldsymbol{b}}_2^\top (n^{-1}\acute{\boldsymbol{X}}_2^\top \boldsymbol{M}_{\acute{\boldsymbol{X}}_1}\acute{\boldsymbol{X}}_2) n^{1/2}\acute{\boldsymbol{b}}_2, \tag{1.81}$$

which is based entirely on quantities from the GNR (1.78). That (1.80) and (1.81) are asymptotically equal relies on (1.79) and the fact, which we have just shown, that the covariance matrix estimator for $\acute{\boldsymbol{b}}_2$ is also valid for $\hat{\boldsymbol{\beta}}_2$.

By equation (1.79), the GNR-based statistic (1.81) can also be expressed as

$$\frac{1}{\acute{\sigma}^2} n^{-1/2}(\boldsymbol{y} - \acute{\boldsymbol{x}})^\top \boldsymbol{M}_{\acute{\boldsymbol{X}}_1}\acute{\boldsymbol{X}}_2 (n^{-1}\acute{\boldsymbol{X}}_2^\top \boldsymbol{M}_{\acute{\boldsymbol{X}}_1}\acute{\boldsymbol{X}}_2)^{-1} n^{-1/2}\acute{\boldsymbol{X}}_2^\top \boldsymbol{M}_{\acute{\boldsymbol{X}}_1}(\boldsymbol{y} - \acute{\boldsymbol{x}}). \tag{1.82}$$

When this statistic is divided by $r = k_2$, we can see by comparison with (F5.30) that it is precisely the F statistic for a test of the artificial hypothesis that $\boldsymbol{b}_2 = \boldsymbol{0}$ in the GNR (1.78). In particular, $\acute{\sigma}^2$ is just the sum of squared residuals from equation (1.78), divided by $n-k$. Thus a valid test statistic can be computed as an ordinary F statistic using the sums of squared residuals from the "restricted" and "unrestricted" GNRs,

$$\mathrm{GNR}_0\colon \quad \boldsymbol{y} - \acute{\boldsymbol{x}} = \acute{\boldsymbol{X}}_1\boldsymbol{b}_1 + \text{residuals}, \text{ and} \tag{1.83}$$

$$\mathrm{GNR}_1\colon \quad \boldsymbol{y} - \acute{\boldsymbol{x}} = \acute{\boldsymbol{X}}_1\boldsymbol{b}_1 + \acute{\boldsymbol{X}}_2\boldsymbol{b}_2 + \text{residuals}. \tag{1.84}$$

In Exercise 1.9, readers are invited to show that such an F statistic is asymptotically equivalent to the F statistic computed from the sums of squared residuals from the two nonlinear regressions (1.71) and (1.72).

In the quite common event that $\acute{\boldsymbol{\beta}}_1 = \tilde{\boldsymbol{\beta}}_1$, the first-order conditions for $\tilde{\boldsymbol{\beta}}_1$ imply that regression (1.83) can have no explanatory power. There is no need to run regression (1.83) in this case, because its SSR is always identical to the SSR from NLS estimation of the restricted model. We will see an example of this in the next subsection.

The principal advantage of tests based on the GNR is that they can be calculated without computing two nonlinear regressions, one for each of the null and alternative hypotheses. The principal disadvantage is that a number of derivatives must be calculated, one for each parameter of the unrestricted model. In many cases, it is necessary to run one nonlinear regression, so as to obtain root-n consistent estimates of the parameters under the null. However, it may sometimes happen that either the null or the alternative hypothesis corresponds to a linear model. In such cases, no nonlinear estimation at all is necessary to carry out a GNR-based test.

The IV Variant of the GNR

In many circumstances, the easiest way to obtain asymptotically valid test statistics for models estimated using instrumental variables is to use a variant of the Gauss-Newton regression. For the model (1.02), this variant, called the **IVGNR**, takes the form

$$\boldsymbol{y} - \boldsymbol{x}(\boldsymbol{\beta}) = \boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{X}(\boldsymbol{\beta})\boldsymbol{b} + \text{residuals}. \tag{1.85}$$

As with the usual GNR, the variables of the IVGNR must be evaluated at some prespecified value of $\boldsymbol{\beta}$ before the regression can be run, in the usual way, using ordinary least squares.

The IVGNR has the same properties relative to model (1.02) as the ordinary GNR has relative to linear and nonlinear regression models estimated by least squares. The first property is that, if (1.85) is evaluated at $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}_{\mathrm{IV}}$, then the regressors $\boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{X}(\hat{\boldsymbol{\beta}}_{\mathrm{IV}})$ are orthogonal to the regressand, because the orthogonality conditions, namely,

$$\boldsymbol{X}^\top(\hat{\boldsymbol{\beta}}_{\mathrm{IV}})\boldsymbol{P}_{\boldsymbol{W}}(\boldsymbol{y} - \boldsymbol{x}(\hat{\boldsymbol{\beta}}_{\mathrm{IV}}) = \boldsymbol{0},$$

are just the estimating equations (1.46) that define $\hat{\boldsymbol{\beta}}_{\mathrm{IV}}$.

The second property is that, if (1.85) is again evaluated at $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}_{\mathrm{IV}}$, the estimated OLS covariance matrix is asymptotically valid. This matrix is

$$s^2(\hat{\boldsymbol{X}}^\top \boldsymbol{P}_{\boldsymbol{W}}\hat{\boldsymbol{X}})^{-1}. \tag{1.86}$$

Here s^2 is the sum of squared residuals from (1.85), divided by $n-k$, and $\hat{\boldsymbol{X}} \equiv \boldsymbol{X}(\hat{\boldsymbol{\beta}}_{\mathrm{IV}})$. Since $\hat{\boldsymbol{b}} = \boldsymbol{0}$ because of the orthogonality of the regressand and the regressors, those residuals are the components of the vector $\boldsymbol{y} - \boldsymbol{x}(\hat{\boldsymbol{\beta}}_{\mathrm{IV}})$, that is, the IV residuals from (1.02). It follows that (1.86), which is equal

to (1.47), is a consistent estimator of the covariance matrix of $\hat{\boldsymbol{\beta}}_{\text{IV}}$. As with the ordinary GNR, the estimator $\acute{s}^2$ obtained by running (1.85) with $\boldsymbol{\beta} = \acute{\boldsymbol{\beta}}$ is consistent for the variance σ^2 of the disturbances if $\acute{\boldsymbol{\beta}}$ is root-n consistent.

The third property is that, like the ordinary GNR, the IVGNR permits one-step efficient estimation. For linear models, this is true if *any* value of $\boldsymbol{\beta}$ is used in (1.85). If we set $\boldsymbol{\beta} = \acute{\boldsymbol{\beta}}$, then running (1.85) gives the artificial parameter estimates

$$\acute{\boldsymbol{b}} = (\boldsymbol{X}^\top \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{P}_{\boldsymbol{W}} (\boldsymbol{y} - \boldsymbol{X}\acute{\boldsymbol{\beta}}) = \hat{\boldsymbol{\beta}}_{\text{IV}} - \acute{\boldsymbol{\beta}},$$

from which it follows that $\acute{\boldsymbol{\beta}} + \acute{\boldsymbol{b}} = \hat{\boldsymbol{\beta}}_{\text{IV}}$ for all $\acute{\boldsymbol{\beta}}$. In the context of nonlinear IV estimation, this result, like the one above for $\acute{s}^2$, becomes an approximation that is asymptotically valid only if $\acute{\boldsymbol{\beta}}$ is a root-n consistent estimator of the true $\boldsymbol{\beta}_0$.

GNR-Based Tests for Autoregressive Disturbances

An example of a model which is linear under the null hypothesis is furnished by the linear regression model with autoregressive disturbances. With time-series data, serial correlation of the disturbances is a frequent occurrence, and so one of the most frequently performed tests in all of econometrics is a test in which the null hypothesis is a linear regression model with serially uncorrelated disturbances and the alternative is the same model with AR(1) disturbances. In this case, we may think of H_1 as being the model (1.06) and H_0 as being the model

$$y_t = \boldsymbol{X}_t \boldsymbol{\beta} + u_t, \quad u_t \sim \text{IID}(0, \sigma^2). \tag{1.87}$$

When GNRs like (1.83) and (1.84) are used for testing, all the variables in them must be evaluated at a parameter vector $\acute{\boldsymbol{\beta}}$ which satisfies the null hypothesis. In this case, the null hypothesis corresponds to the restriction that $\rho = 0$. Therefore, we must set $\acute{\rho} = 0$ in the GNRs corresponding to the restricted model (1.87) and the unrestricted model (1.06). The natural choice for $\acute{\boldsymbol{\beta}}$ is then $\tilde{\boldsymbol{\beta}}$, the vector of OLS parameter estimates for (1.87).

The GNR for (1.06) was given in (1.68). If this artificial regression is evaluated at $\boldsymbol{\beta} = \tilde{\boldsymbol{\beta}}$ and $\rho = 0$, it becomes

$$y_t - \boldsymbol{X}_t \tilde{\boldsymbol{\beta}} = \boldsymbol{X}_t \boldsymbol{b} + b_\rho (y_{t-1} - \boldsymbol{X}_{t-1} \tilde{\boldsymbol{\beta}}) + \text{residual}, \tag{1.88}$$

where $\boldsymbol{b}$ corresponds to $\boldsymbol{\beta}$ and b_ρ corresponds to ρ. If we denote the OLS residuals from (1.87) by $\tilde{u}_t$, the GNR (1.88) takes on the very simple form

$$\tilde{u}_t = \boldsymbol{X}_t \boldsymbol{b} + b_\rho \tilde{u}_{t-1} + \text{residual}. \tag{1.89}$$

This is just a linear regression of the residuals from (1.87) on the regressors of (1.87) and one more regressor, namely, the residuals lagged once. Since

only one restriction is to be tested, a suitable test statistic is the t statistic for the artificial parameter b_ρ in (1.89) to equal 0. This is the square root of the F statistic, which we have seen to be asymptotically valid.

Almost as simple as the above test is a test of the null hypothesis (1.87) against an alternative in which the disturbances follow the **AR(2) process**

$$u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + \varepsilon_t, \quad \varepsilon_t \sim \text{IID}(0, \sigma^2).$$

It is not hard to show that an appropriate artificial regression for testing (1.87) against the AR(2) alternative that is analogous to (1.06) is

$$\tilde{u}_t = \boldsymbol{X}_t \boldsymbol{b} + b_{\rho_1} \tilde{u}_{t-1} + b_{\rho_2} \tilde{u}_{t-2} + \text{residual}; \tag{1.90}$$

see Exercise 1.10. Since, in this case, we have a test with two degrees of freedom, we cannot use a t test. However, it is still not necessary to run two regressions in order to compute an F statistic. Consider the form taken by GNR_0 in this case:

$$\tilde{u}_t = \boldsymbol{X}_t \boldsymbol{b} + \text{residual}. \tag{1.91}$$

This is just the GNR corresponding to the linear regression (1.87). Since the regressand is the vector of residuals from estimating (1.87), it is orthogonal to the explanatory variables. Therefore, by (1.58), the artificial parameter estimates $\tilde{\boldsymbol{b}}$ are zero, and (1.91) has no explanatory power. As a result, the SSR from (1.91) is equal to the total sum of squares (TSS). But this is also the TSS from the GNR (1.90) corresponding to the alternative. Thus the difference between the SSRs from (1.91) and (1.90) is the difference between the TSS and the SSR from (1.90), or, more conveniently, the explained sum of squares (ESS) from (1.90). The GNR-based F statistic can therefore be computed by running (1.90) alone. In fact, since the denominator is just the estimate $\tilde{s}^2$ of the disturbance variance from (1.90), the F statistic is simply[1]

$$F = \frac{\text{ESS}}{r \tilde{s}^2} = \frac{n - k - r}{r} \times \frac{\text{ESS}}{\text{SSR}}, \tag{1.92}$$

where k is the number of regressors in (1.87) and $r = 2$ in this particular case.

Asymptotically, we can obtain a valid test statistic by using any consistent estimate of the true disturbance variance σ_0^2 as the denominator. If we were to use the estimate under the null rather than the estimate under the alternative, the denominator of the test statistic would be $(n - k_1)^{-1} \sum_{t=1}^n \tilde{u}_t^2$. Asymptotically, it makes no difference whether we divide by $n - k_1$ or n when

[1] We are assuming here that regression (1.90) is run over all n observations. This requires either that data for observations 0 and -1 are available, or that the unobserved residuals $\tilde{u}_0$ and $\tilde{u}_{-1}$ are replaced by zeros.

we estimate σ^2. Therefore, if R^2 is the uncentered R squared from (1.90), another perfectly valid test statistic is

$$nR^2 = \frac{n\text{ESS}}{\text{TSS}} = \frac{\text{ESS}}{n^{-1}\sum_{t=1}^{n}\tilde{u}_t^2}, \tag{1.93}$$

which follows the $\chi^2(2)$ distribution asymptotically. If the regressors include a constant, the residuals $\tilde{u}_t$ must have expectation zero, and the uncentered R^2 (1.93) is identical to the centered R^2 that is printed by most regression packages.

Whether we use the F statistic (1.92) or the nR^2 statistic (1.93), the GNR provides a very easy way to test the null hypothesis that the disturbances are serially uncorrelated against all sorts of autoregressive alternatives. Of course, neither statistic follows its asymptotic distribution exactly in finite samples. However, there is some evidence — for example, Kiviet (1986) — that the former tends to have better finite-sample properties than the latter. This evidence accords with theory, because, as (1.41) shows, the relationship between NLS residuals and disturbances is approximately the same as the relationship between OLS residuals and disturbances. Therefore, it makes sense to use the F form of the statistic, which treats the estimate $\tilde{s}^2$ based on the GNR as if it were based on an ordinary OLS regression.

The above example generalizes to all cases in which $\hat{\boldsymbol{\beta}}$ is taken to be $\tilde{\boldsymbol{\beta}}$ from estimating the null hypothesis, whether or not the restricted model is linear. In such cases, because GNR_0 has no explanatory power, its SSR is equal to its TSS, which in turn is equal to the TSS of GNR_1. In consequence, we only need to run GNR_1, which in this case is

$$\boldsymbol{y} - \tilde{\boldsymbol{x}} = \tilde{\boldsymbol{X}}_1\boldsymbol{b}_1 + \tilde{\boldsymbol{X}}_2\boldsymbol{b}_2 + \text{residuals}.$$

Under the null hypothesis, nR^2 from this test regression is asymptotically distributed as $\chi^2(r)$. This is not the case for GNR_1 when $\hat{\boldsymbol{\beta}} \neq \tilde{\boldsymbol{\beta}}$. However, the F test of (1.83) against (1.84) is asymptotically valid even when $\hat{\boldsymbol{\beta}} \neq \tilde{\boldsymbol{\beta}}$. It is merely required that $\hat{\boldsymbol{\beta}}$ should satisfy the null hypothesis and be root-n consistent.

Most GNR-based tests are like the ones for serial correlation that we have just discussed, in which the GNR is evaluated at least-squares estimates under the null hypothesis. However, it is also possible to evaluate the GNR at estimates obtained under the alternative hypothesis. We will encounter tests of this type when we discuss common factor restrictions in Chapter 6.

Bootstrap Tests

Because none of the tests discussed in this section is exact in finite samples, it is often desirable to compute bootstrap P values, which, in most cases, are more accurate than ones based on asymptotic theory. The procedures for

computing bootstrap P values for nonlinear regression models are essentially the same as the ones for linear models. We use estimates under the null to generate B bootstrap samples, usually either generating the disturbances from the $\text{N}(0, \tilde{s}^2)$ distribution or resampling the rescaled residuals, and we then compute a bootstrap test statistic τ_j^* using each of the bootstrap samples. For a test that rejects when the test statistic $\hat{\tau}$ is large, the bootstrap P value is then $1 - \hat{F}^*(\hat{\tau})$, where $\hat{F}^*(\hat{\tau})$ denotes the EDF of the τ_j^* evaluated at $\hat{\tau}$. Of course, this procedure can sometimes be computationally expensive; see Davidson and MacKinnon (1999a) for a way of making it somewhat less so.

1.8 Final Remarks

In this chapter, we have dealt only with the estimation of nonlinear regression models by nonlinear least squares. However, many of the results will reappear, in slightly different forms, when we consider estimation methods for other sorts of models. The NLS estimator is an **extremum estimator**, or **M-estimator**, that is, an estimator obtained by minimizing or maximizing a criterion function. In the next few chapters, we will encounter several other extremum estimators: the generalized method of moments (Chapter 2) and maximum likelihood (Chapter 3). Most of these estimators, like the NLS estimator, can be derived from the principles of estimating functions. All extremum estimators share a number of common features. Similar asymptotic results, and similar methods of proof, apply to all of them.

1.9 Exercises

1.1 Let the expectation of a random variable Y conditional on a set of other random variables $X_1, \ldots, X_k$ be the deterministic function $h(X_1, \ldots, X_k)$ of the conditioning variables. Let Ω be the information set consisting of all deterministic functions of the X_i, $i = 1, \ldots, k$. Show that $\text{E}(Y \mid \Omega) = h(X_1, \ldots, X_k)$. **Hint:** Use the Law of Iterated Expectations for Ω and the information set defined by the X_i.

⋆1.2 Consider a model similar to (F4.18), but with disturbances that are normally distributed:

$$y_t = \beta_1 + \beta_2 1/t + u_t, \quad u_t \sim \text{NID}(0, \sigma^2),$$

where $t = 1, 2, \ldots, n$. If the true value of β_2 is β_2^0 and $\hat{\boldsymbol{\beta}}_2$ is the OLS estimator, show that the limit in probability of $\hat{\beta}_2 - \beta_2^0$ is a normal random variable with expectation 0 and variance $6\sigma^2/\pi^2$. In order to obtain this result, you will need to use the results that

$$\sum_{t=1}^{\infty}(1/t)^2 = \pi^2/6,$$

and that, if $s(n) = \sum_{t=1}^{n}(1/t)$, then $\lim n^{-1}s(n) = 0$ and $\lim n^{-1}s^2(n) = 0$.

1.3 Show that the estimator defined by (1.10) depends on $\boldsymbol{W}$ only through the span $\mathcal{S}(\boldsymbol{W})$ of its columns. This is equivalent to showing that the estimator depends on $\boldsymbol{W}$ only through the orthogonal projection matrix $\boldsymbol{P}_{\boldsymbol{W}}$.

1.4 Show algebraically that the first-order conditions for minimizing the SSR function (1.29) have the same solutions as the moment conditions (1.28).

1.5 Apply Taylor's Theorem to n^{-1} times the left-hand side of the estimating equations (1.28), expanding around the true parameter vector $\boldsymbol{\beta}_0$. Show that the extra term which appears here, but was absent in (1.21), where the instruments are fixed and we multiply by $n^{-1/2}$, tends to zero as $n \to \infty$. Make clear where and how you use a law of large numbers in your demonstration.

1.6 For the nonlinear regression model

$$y_t = \beta_1 z_t^{\beta_2} + u_t, \quad u_t \sim \text{IID}(0, \sigma^2),$$

write down the sum of squared residuals as a function of β_1, β_2, y_t, and z_t. Then differentiate it to obtain two first-order conditions. Show that these equations are equivalent to special cases of the estimating equations (1.28).

1.7 In each of the following regressions, y_t is the dependent variable, x_t and z_t are explanatory variables, and α, β, and γ are unknown parameters.

(a) $y_t = \alpha + \beta x_t + \gamma/x_t + u_t$

(b) $y_t = \alpha + \beta x_t + x_t/\gamma + u_t$

(c) $y_t = \alpha + \beta x_t + z_t/\gamma + u_t$

(d) $y_t = \alpha + \beta x_t + z_t/\beta + u_t$

(e) $y_t = \alpha + \beta x_t z_t + u_t$

(f) $y_t = \alpha + \beta\gamma x_t z_t + \gamma z_t + u_t$

(g) $y_t = \alpha + \beta\gamma x_t + \gamma z_t + u_t$

(h) $y_t = \alpha + \beta x_t + \beta x_t^2 + u_t$

(i) $y_t = \alpha + \beta x_t + \gamma x_t^2 + u_t$

(j) $y_t = \alpha + \beta\gamma x_t^3 + u_t$

(k) $y_t = \alpha + \beta x_t + (1-\beta) z_t + u_t$

(l) $y_t = \alpha + \beta x_t + (\gamma - \beta) z_t + u_t$

For each of these regressions, is it possible to obtain a least-squares estimator of the parameters? In other words, is each of these models identified? If not, explain why not. If so, can the estimator be obtained by ordinary (that is, linear) least squares? If it can, write down the regressand and regressors for the linear regression to be used.

***1.8** Show that a Taylor expansion to second order of an NLS residual gives

$$\hat{u}_t = u_t - \boldsymbol{X}_t(\boldsymbol{\beta}_0)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) - \tfrac{1}{2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^\top \bar{\boldsymbol{H}}_t(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0), \tag{1.94}$$

where $\boldsymbol{\beta}_0$ is the parameter vector of the DGP, and the $k \times k$ matrix $\bar{\boldsymbol{H}}_t \equiv \boldsymbol{H}_t(\bar{\boldsymbol{\beta}})$ is the matrix of second derivatives with respect to $\boldsymbol{\beta}$ of the regression function $x_t(\boldsymbol{\beta})$, evaluated at some $\bar{\boldsymbol{\beta}}$ that satisfies (1.20).

Define $\boldsymbol{b} \equiv n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$. As $n \to \infty$, $\boldsymbol{b}$ tends to the normal random variable $\text{plim}(n^{-1}\boldsymbol{X}_0^\top\boldsymbol{X}_0)^{-1}n^{-1/2}\boldsymbol{X}_0^\top\boldsymbol{u}$. By expressing equation (1.94) in terms of $\boldsymbol{b}$, show that the difference between $\hat{\boldsymbol{u}}^\top\hat{\boldsymbol{u}}$ and $\boldsymbol{u}^\top\boldsymbol{M}_{\boldsymbol{X}_0}\boldsymbol{u}$ tends to 0 as $n \to \infty$. Here $\boldsymbol{M}_{\boldsymbol{X}_0} \equiv \boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}_0}$ is the orthogonal projection on to $\mathcal{S}^\perp(\boldsymbol{X}_0)$.

1.9 Using the result (1.41) on NLS residuals, show that the F statistic computed using the sums of squared residuals from the two GNRs (1.83) and (1.84) is asymptotically equivalent to the F statistic computed using the sums of squared residuals from the nonlinear regressions (1.71) and (1.72).

1.10 Consider a linear regression with AR(2) disturbances. This can be written as

$$y_t = \boldsymbol{X}_t\boldsymbol{\beta} + u_t, \quad u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + \varepsilon_t, \quad \varepsilon_t \sim \text{IID}(0, \sigma^2).$$

Explain how to test the null hypothesis that $\rho_1 = \rho_2 = 0$ by means of a GNR.

1.11 Consider again the ADL model (F4.90) of Part 1, Exercise 4.F32, which is reproduced here with a minor notational change:

$$c_t = \alpha + \beta c_{t-1} + \gamma_0 y_t + \gamma_1 y_{t-1} + \varepsilon_t. \tag{1.95}$$

Recall that c_t and y_t are the logarithms of consumption and income, respectively. Show that this model contains as a special case the following linear model with AR(1) disturbances:

$$c_t = \delta_0 + \delta_1 y_t + u_t, \quad \text{with} \quad u_t = \rho u_{t-1} + \varepsilon_t, \tag{1.96}$$

where ε_t is IID. Write down the relation between the parameters δ_0, δ_1, and ρ of this model and the parameters α, β, γ_0, and γ_1 of (1.95). How many and what restrictions are imposed on the latter set of parameters by the model (1.96)?

1.12 Using the data in the file **consumption.data**, estimate the nonlinear model defined implicitly by (1.96) for the period 1953:1 to 1996:4 by nonlinear least squares. Since pre-sample data are available, you should use all 176 observations for the estimation. Do not use a specialized procedure for AR(1) estimation. For starting values, use the estimates of δ_0, δ_1, and ρ implied by the OLS estimates of equation (1.95). Finding them requires the solution to the previous exercise.

Repeat this exercise, using 0 as the starting value for all three parameters. Does the algorithm converge as rapidly as it did before? Do you obtain the same estimates? If not, which ones are actually the NLS estimates?

Test the restrictions that the nonlinear model imposes on the model (1.95) by means of an asymptotic F test.

1.13 Using the estimates of the model (1.96) from the previous question, generate a single set of simulated data c_t^* for the period 1953:1 to 1996:4. The simulation should be conditional on the pre-sample value (that is, the value for 1952:4) of log consumption. Do this in two different ways. First, generate disturbances u_t^* that follow an AR(1) process, and then generate the c_t^* in terms of these u_t^*. Next, perform the simulation directly in terms of the innovations ε_t^*, using the nonlinear model obtained by imposing the appropriate restrictions on (1.95). Show that, if you use the same realizations for the ε_t^*, the simulated values c_t^* are identical. Estimate the model (1.96) using your simulated data.

1.14 The nonlinear model obtained from (1.96) has just three parameters: δ_0, δ_1, and ρ. It can therefore be estimated by the method of moments using three exogenous or predetermined variables. Estimate the model using the constant and the three possible choices of two variables from the set of nonconstant explanatory variables in (1.95).

1.15 Formulate a GNR, based on estimates under the null hypothesis, that allows you to use a t test to test the restriction imposed on the model (1.95) by the model (1.96). Compare the P value for this (asymptotic) t test with the one for the F test of Part 1, Exercise 7.12.

1.16 Starting from the unconstrained estimates provided by (1.95), obtain one-step efficient estimates of the parameters of (1.96) using the GNR associated with that model. Use the GNR iteratively so as to approach the true NLS estimates more closely, until such time as the sum of squared residuals from the GNR is within 10^{-8} of the one obtained by NLS estimation. Compare the number of iterations of this GNR-based procedure with the number used by the NLS algorithm of your software package.

1.17 Formulate a GNR, based on estimates under the alternative hypothesis, to test the restriction imposed on the model (1.95) by the model (1.96). Your test procedure should just require two OLS regressions.

1.18 Using 199 bootstrap samples, compute a parametric bootstrap P value for the test statistic obtained in Part 1, Exercise 7.17. Assume that the disturbances are normally distributed.

1.19 Test the hypothesis that $\gamma_0 + \gamma_1 = 0$ in (1.95). Do this in three different ways, two of which are valid in the presence of heteroskedasticity of unknown form.

1.20 For the nonlinear regression model defined implicitly by (1.96) and estimated using the data in the file **consumption.data**, perform three different tests of the hypothesis that all the coefficients are the same for the two subsamples 1953:1 to 1970:4 and 1971:1 to 1996:4. Firstly, use an asymptotic F test based on nonlinear estimation of both the restricted and unrestricted models. Secondly, use an asymptotic F test based on a GNR which requires nonlinear estimation only under the null. Finally, use a test that is robust to heteroskedasticity of unknown form.

1.21 The original HRGNR proposed by Davidson and MacKinnon (1985a) is

$$\boldsymbol{\iota} = \acute{\boldsymbol{U}}\boldsymbol{M}_{\acute{\boldsymbol{X}}_1}\acute{\boldsymbol{X}}_2\boldsymbol{b}_2 + \text{residuals}, \tag{1.97}$$

where $\acute{\boldsymbol{U}}$, $\acute{\boldsymbol{X}}_1$, and $\acute{\boldsymbol{X}}_2$ are as defined in Section 1.8, $\boldsymbol{b}_2$ is a k_2–vector, and $\boldsymbol{M}_{\acute{\boldsymbol{X}}_1}$ is the matrix that projects orthogonally on to $\mathcal{S}^\perp(\acute{\boldsymbol{X}}_1)$. The test statistic for the null hypothesis that $\boldsymbol{\beta}_2 = \boldsymbol{0}$ is n minus the SSR from regression (1.97).

Use regression (1.97), where all the matrices are evaluated at restricted NLS estimates, to retest the hypothesis of the previous question. Comment on the relationship between the test statistic you obtain and the heteroskedasticity-robust test statistic of the previous question.

1.22 Suppose that $\boldsymbol{P}$ is a projection matrix with rank r. Without loss of generality, we can assume that $\boldsymbol{P}$ projects on to the span of the columns of an $n \times r$ matrix $\boldsymbol{Z}$. Suppose further that the n–vector $\boldsymbol{z}$ is distributed as IID$(\boldsymbol{0}, \boldsymbol{I})$. Show that the quadratic form $\boldsymbol{z}^\top\boldsymbol{P}\boldsymbol{z}$ follows the $\chi^2(r)$ distribution asymptotically as $n \to \infty$. (**Hint:** See the proof of Theorem 4.1.)

Chapter 2

The Generalized Method of Moments

2.1 Introduction

The models we have considered in earlier chapters have all been regression models of one sort or another. In this chapter and the next, we introduce more general types of models, along with a general method for performing estimation and inference on them. This technique is called the **generalized method of moments**, or **GMM**, and it includes as special cases all the methods we have so far developed for regression models.

As we explained in Part 1, Section 4.1, a model is represented by a set of DGPs. Each DGP in the model is characterized by a parameter vector, which we will normally denote by $\boldsymbol{\beta}$ in the case of regression functions and by $\boldsymbol{\theta}$ in the general case. The starting point for GMM estimation is to specify functions, which, for any DGP in the model, depend both on the data generated by that DGP and on the model parameters. When these functions are evaluated at the parameters that correspond to the DGP that generated the data, their expectation must be zero.

As a simple example, consider the linear regression model $y_t = \boldsymbol{X}_t\boldsymbol{\beta} + u_t$. An important part of the model specification is that the disturbances have mean zero. These disturbances are unobservable, because the parameters $\boldsymbol{\beta}$ of the regression function are unknown. But we can define the residuals $u_t(\boldsymbol{\beta}) \equiv y_t - \boldsymbol{X}_t\boldsymbol{\beta}$ as functions of the observed data and the unknown model parameters, and these functions provide what we need for GMM estimation. If the residuals are evaluated at the parameter vector $\boldsymbol{\beta}_0$ associated with the true DGP, they have mean zero under that DGP, but if they are evaluated at some $\boldsymbol{\beta} \neq \boldsymbol{\beta}_0$, they do not have mean zero. In Part 1, Chapter 2, we used this fact to develop a method-of-moments (MM) estimator for the parameter vector $\boldsymbol{\beta}$ of the regression function. As we will see in the next section, the various GMM estimators of $\boldsymbol{\beta}$ include as a special case the OLS estimator developed in Chapter 2.

In Chapter 1, when we dealt with nonlinear regression models, and again in Part 1, Chapter 8, we used instrumental variables along with residuals in order to develop estimating functions. The use of instrumental variables is

also an essential aspect of GMM, and in this chapter we will once again make use of the various kinds of optimal instruments that were useful in Chapter 1 and Part 1, Chapter 8 in order to develop a wide variety of estimators that are asymptotically efficient for a wide variety of models.

We begin by considering, in the next section, a linear regression model with endogenous explanatory variables and a disturbance covariance matrix that is not proportional to the identity matrix. Such a model requires us to combine the insights of both Part 1, Chapter 8 and Part 1, Chapter 9 in order to obtain asymptotically efficient estimates. In the process of doing so, we will see how GMM estimation works more generally, and we will be led to develop ways to estimate models with both heteroskedasticity and serial correlation of unknown form. In Section 2.3, we study in some detail the **heteroskedasticity and autocorrelation consistent**, or **HAC**, covariance matrix estimators that we introduced in Part 1, Section 6.5. Then, in Section 2.4, we introduce a set of tests, based on **GMM criterion functions**, that are widely used for inference in conjunction with GMM estimation. In Section 2.5, we move beyond regression models to give a more formal and advanced presentation of GMM, and we postpone to this section most of the proofs of consistency, asymptotic normality, and asymptotic efficiency for GMM estimators.

2.2 GMM Estimators for Linear Regression Models

Consider the linear regression model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}, \quad \mathrm{E}(\boldsymbol{u}\boldsymbol{u}^\top) = \boldsymbol{\Omega}, \tag{2.01}$$

where there are n observations, and $\boldsymbol{\Omega}$ is an $n \times n$ covariance matrix. As in the previous chapter, some of the explanatory variables that form the $n \times k$ matrix $\boldsymbol{X}$ may not be predetermined with respect to the disturbances $\boldsymbol{u}$. However, there is assumed to exist an $n \times l$ matrix of predetermined instrumental variables, $\boldsymbol{W}$, with $n > l$ and $l \geq k$, satisfying the condition $\mathrm{E}(u_t \,|\, \boldsymbol{W}_t) = 0$ for each row $\boldsymbol{W}_t$ of $\boldsymbol{W}$, $t = 1, \ldots, n$. Any column of $\boldsymbol{X}$ that is predetermined must also be a column of $\boldsymbol{W}$. In addition, we assume that, for all $t, s = 1, \ldots, n$, $\mathrm{E}(u_t u_s \,|\, \boldsymbol{W}_t, \boldsymbol{W}_s) = \omega_{ts}$, where ω_{ts} is the ts^{th} element of $\boldsymbol{\Omega}$. We will need this assumption later, because it allows us to see that

$$\begin{aligned}
\mathrm{Var}(n^{-1/2}\boldsymbol{W}^\top\boldsymbol{u}) &= \tfrac{1}{n}\mathrm{E}(\boldsymbol{W}^\top\boldsymbol{u}\boldsymbol{u}^\top\boldsymbol{W}) = \tfrac{1}{n}\sum_{t=1}^{n}\sum_{s=1}^{n}\mathrm{E}(u_t u_s \boldsymbol{W}_t^\top\boldsymbol{W}_s) \\
&= \tfrac{1}{n}\sum_{t=1}^{n}\sum_{s=1}^{n}\mathrm{E}\big(\mathrm{E}(u_t u_s \boldsymbol{W}_t^\top\boldsymbol{W}_s \,|\, \boldsymbol{W}_t, \boldsymbol{W}_s)\big) \\
&= \tfrac{1}{n}\sum_{t=1}^{n}\sum_{s=1}^{n}\mathrm{E}(\omega_{ts}\boldsymbol{W}_t^\top\boldsymbol{W}_s) = \tfrac{1}{n}\mathrm{E}(\boldsymbol{W}^\top\boldsymbol{\Omega}\boldsymbol{W}).
\end{aligned} \tag{2.02}$$

The assumption that $\mathrm{E}(u_t \,|\, \boldsymbol{W}_t) = 0$ implies that, for all $t = 1, \ldots, n$,

$$\mathrm{E}\big(\boldsymbol{W}_t^\top(y_t - \boldsymbol{X}_t\boldsymbol{\beta})\big) = \boldsymbol{0}. \tag{2.03}$$

These equations form a set of what we may call **theoretical moment conditions**. They were used in Part 1, Chapter 8 as the starting point for estimation of the regression model (2.01). Each theoretical moment condition corresponds to a sample moment, or **empirical moment**, of the form

$$\frac{1}{n}\sum_{t=1}^{n} w_{ti}^\top(y_t - \boldsymbol{X}_t\boldsymbol{\beta}) = \frac{1}{n}\boldsymbol{w}_i^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}), \tag{2.04}$$

where $\boldsymbol{w}_i$, $i = 1, \ldots, l$, is the i^{th} column of $\boldsymbol{W}$, and w_{ti} is the ti^{th} element. When $l = k$, we can set these sample moments equal to zero and solve the resulting k estimating equations to obtain the simple IV estimator (F8.13). When $l > k$, we must do as we did in Part 1, Chapter 8 and select k independent linear combinations of the sample moments (2.04) in order to obtain an estimator.

Of course, what we have been calling a sample moment is in fact an estimating function. The terminology in the above paragraph is nevertheless more familiar to most econometricians than that of estimating functions, and so we will continue to use it, although in Section 2.5 we discuss estimating functions more formally than we have done so far, and explain the correspondences between the terminology of moment conditions and that of estimating functions.

Now let $\boldsymbol{J}$ be an $l \times k$ matrix with full column rank k, and consider the estimator obtained by using the k columns of $\boldsymbol{W}\boldsymbol{J}$ as instruments. This estimator solves the k equations

$$\boldsymbol{J}^\top\boldsymbol{W}^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) = \boldsymbol{0}, \tag{2.05}$$

which are referred to as **sample moment conditions**, or just **moment conditions** when there is no ambiguity. They are also sometimes called **orthogonality conditions**, since they require that the vector of residuals should be orthogonal to the columns of $\boldsymbol{W}\boldsymbol{J}$. Let us assume that the data are generated by a DGP which belongs to the model (2.01), with coefficient vector $\boldsymbol{\beta}_0$ and covariance matrix $\boldsymbol{\Omega}_0$. Under this assumption, we have the following explicit expression, suitable for asymptotic analysis, for the estimator $\hat{\boldsymbol{\beta}}$ that solves (2.05):

$$n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) = \big(n^{-1}\boldsymbol{J}^\top\boldsymbol{W}^\top\boldsymbol{X}\big)^{-1}n^{-1/2}\boldsymbol{J}^\top\boldsymbol{W}^\top\boldsymbol{u}. \tag{2.06}$$

From this, recalling (2.02), we find that the asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}$, that is, the limiting covariance matrix of $n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$, is

$$\Big(\operatorname*{plim}_{n\to\infty}\frac{1}{n}\boldsymbol{J}^\top\boldsymbol{W}^\top\boldsymbol{X}\Big)^{-1}\Big(\operatorname*{plim}_{n\to\infty}\frac{1}{n}\boldsymbol{J}^\top\boldsymbol{W}^\top\boldsymbol{\Omega}_0\boldsymbol{W}\boldsymbol{J}\Big)\Big(\operatorname*{plim}_{n\to\infty}\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{W}\boldsymbol{J}\Big)^{-1}. \tag{2.07}$$

This matrix has the familiar sandwich form that we expect to see when an estimator is not asymptotically efficient.

The next step, as in Part 1, Section 8.3, is to choose $\boldsymbol{J}$ so as to minimize the covariance matrix (2.07). We may reasonably expect that, with such a choice of $\boldsymbol{J}$, the covariance matrix would no longer have the form of a sandwich. The simplest choice of $\boldsymbol{J}$ that eliminates the sandwich in (2.07) is

$$\boldsymbol{J} = (\boldsymbol{W}^\top \boldsymbol{\Omega}_0 \boldsymbol{W})^{-1} \boldsymbol{W}^\top \boldsymbol{X}; \tag{2.08}$$

notice that, in the special case in which $\boldsymbol{\Omega}_0$ is proportional to $\mathbf{I}$, this expression reduces to the result (F8.25) that we found in Part 1, Section 8.3 as the solution for that special case. We can see, therefore, that (2.08) is the appropriate generalization of (F8.25) when $\boldsymbol{\Omega}$ is not proportional to an identity matrix. With $\boldsymbol{J}$ defined by (2.08), the covariance matrix (2.07) becomes

$$\operatorname*{plim}_{n\to\infty} \left(\frac{1}{n} \boldsymbol{X}^\top \boldsymbol{W} (\boldsymbol{W}^\top \boldsymbol{\Omega}_0 \boldsymbol{W})^{-1} \boldsymbol{W}^\top \boldsymbol{X} \right)^{-1}, \tag{2.09}$$

and the **efficient GMM estimator** is

$$\hat{\boldsymbol{\beta}}_{\text{GMM}} = \left(\boldsymbol{X}^\top \boldsymbol{W} (\boldsymbol{W}^\top \boldsymbol{\Omega}_0 \boldsymbol{W})^{-1} \boldsymbol{W}^\top \boldsymbol{X} \right)^{-1} \boldsymbol{X}^\top \boldsymbol{W} (\boldsymbol{W}^\top \boldsymbol{\Omega}_0 \boldsymbol{W})^{-1} \boldsymbol{W}^\top \boldsymbol{y}. \tag{2.10}$$

When $\boldsymbol{\Omega}_0 = \sigma^2 \mathbf{I}$, this estimator reduces to the generalized IV estimator (F8.30). In Exercise 2.1, readers are invited to show that the difference between the covariance matrices (2.07) and (2.09) is a positive semidefinite matrix, thereby confirming (2.08) as the optimal choice for $\boldsymbol{J}$. The estimator $\hat{\boldsymbol{\beta}}_{\text{GMM}}$ is efficient in the class of estimators defined by the moment conditions (2.05), but we will see that a more efficient estimator is available if we know $\boldsymbol{\Omega}_0$ and are prepared to exploit that knowledge.

The GMM Criterion Function

With both GLS and IV estimation, we showed that the efficient estimators could also be derived by minimizing an appropriate criterion function; this function was (F9.06) for GLS and (F8.31) for IV. Similarly, the efficient GMM estimator (2.10) minimizes the **GMM criterion function**

$$Q(\boldsymbol{\beta}, \boldsymbol{y}) \equiv (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top \boldsymbol{W} (\boldsymbol{W}^\top \boldsymbol{\Omega}_0 \boldsymbol{W})^{-1} \boldsymbol{W}^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}), \tag{2.11}$$

as can be seen at once by noting that the first-order conditions for minimizing (2.11) are

$$\boldsymbol{X}^\top \boldsymbol{W} (\boldsymbol{W}^\top \boldsymbol{\Omega}_0 \boldsymbol{W})^{-1} \boldsymbol{W}^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) = \mathbf{0}.$$

If $\boldsymbol{\Omega}_0 = \sigma_0^2 \mathbf{I}$, (2.11) reduces to the IV criterion function (F8.31), divided by σ_0^2. In Part 1, Section 9.6, we saw that the minimized value of the IV criterion function, divided by an estimate of σ^2, serves as the statistic for the

Sargan test for overidentification. We will see in Section 2.4 that the GMM criterion function (2.11), with the usually unknown matrix $\boldsymbol{\Omega}_0$ replaced by a suitable estimate, can also be used as a test statistic for overidentification. The criterion function (2.11) is a quadratic form in the vector $\boldsymbol{W}^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$ of sample moments and the inverse of the matrix $\boldsymbol{W}^\top \boldsymbol{\Omega}_0 \boldsymbol{W}$. Equivalently, it is a quadratic form in $n^{-1/2} \boldsymbol{W}^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$ and the inverse of $n^{-1} \boldsymbol{W}^\top \boldsymbol{\Omega}_0 \boldsymbol{W}$, since the powers of n cancel. Under the sort of regularity conditions we have used in earlier chapters, $n^{-1/2} \boldsymbol{W}^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_0)$ satisfies a central limit theorem, and so tends, as $n \to \infty$, to a normal distribution, with mean vector $\mathbf{0}$ and covariance matrix the limit of $n^{-1} \boldsymbol{W}^\top \boldsymbol{\Omega}_0 \boldsymbol{W}$. It follows that (2.11) evaluated using the true $\boldsymbol{\beta}_0$ and the true $\boldsymbol{\Omega}_0$ is asymptotically distributed as χ^2 with l degrees of freedom; recall Part 1, Theorem 5.1, and see Exercise 2.2.

This property of the GMM criterion function is simply a consequence of its structure as a quadratic form in the sample moments used for estimation and the inverse of the asymptotic covariance matrix of these moments evaluated at the true parameters. As we will see in Section 2.4, this property is what makes the GMM criterion function useful for testing. The argument leading to (2.10) shows that this same property of the GMM criterion function leads to the asymptotic efficiency of the estimator that minimizes it.

Provided the instruments are predetermined, so that they satisfy the condition that $\mathrm{E}(u_t \mid \boldsymbol{W}_t) = 0$, we still obtain a consistent estimator, even when the matrix $\boldsymbol{J}$ used to select linear combinations of the instruments is different from (2.08). Such a consistent, but in general inefficient, estimator can also be obtained by minimizing a quadratic criterion function of the form

$$(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top \boldsymbol{W} \boldsymbol{\Lambda} \boldsymbol{W}^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}), \tag{2.12}$$

where the **weighting matrix** $\boldsymbol{\Lambda}$ is $l \times l$, positive definite, and must be at least asymptotically nonrandom. Without loss of generality, $\boldsymbol{\Lambda}$ can be taken to be symmetric; see Exercise 2.3. The inefficient GMM estimator is

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{\Lambda} \boldsymbol{W}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{\Lambda} \boldsymbol{W}^\top \boldsymbol{y}, \tag{2.13}$$

from which it can be seen that the use of the weighting matrix $\boldsymbol{\Lambda}$ corresponds to the implicit choice $\boldsymbol{J} = \boldsymbol{\Lambda} \boldsymbol{W}^\top \boldsymbol{X}$. For a given choice of $\boldsymbol{J}$, there are various possible choices of $\boldsymbol{\Lambda}$ that give rise to the same estimator; see Exercise 2.4.

When $l = k$, the model is exactly identified, and $\boldsymbol{J}$ is a nonsingular square matrix which has no effect on the estimator. This is most easily seen by looking at the moment conditions (2.05), which are equivalent, when $l = k$, to those obtained by premultiplying them by $(\boldsymbol{J}^\top)^{-1}$. Similarly, if the estimator is defined by minimizing a quadratic form, it does not depend on the choice of $\boldsymbol{\Lambda}$ whenever $l = k$. To see this, consider the first-order conditions for minimizing (2.12), which, up to a scalar factor, are

$$\boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{\Lambda} \boldsymbol{W}^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) = \mathbf{0}.$$

If $l = k$, $\boldsymbol{X}^\top\boldsymbol{W}$ is a square matrix, and the first-order conditions can be premultiplied by $\boldsymbol{\Lambda}^{-1}(\boldsymbol{X}^\top\boldsymbol{W})^{-1}$. Therefore, the estimator is the solution to the equations $\boldsymbol{W}^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) = \boldsymbol{0}$, independently of $\boldsymbol{\Lambda}$. This solution is just the simple IV estimator defined in (F8.13).

When $l > k$, the model is overidentified, and the estimator (2.13) depends on the choice of $\boldsymbol{J}$ or $\boldsymbol{\Lambda}$. The efficient GMM estimator, for a given set of instruments, is defined in terms of the true covariance matrix $\boldsymbol{\Omega}_0$, which is usually unknown. If $\boldsymbol{\Omega}_0$ is known up to a scalar multiplicative factor, so that $\boldsymbol{\Omega}_0 = \sigma^2\boldsymbol{\Delta}_0$, with σ^2 unknown and $\boldsymbol{\Delta}_0$ known, then $\boldsymbol{\Delta}_0$ can be used in place of $\boldsymbol{\Omega}_0$ in either (2.10) or (2.11). This is true because multiplying $\boldsymbol{\Omega}_0$ by a scalar leaves (2.10) invariant, and it also leaves invariant the $\boldsymbol{\beta}$ that minimizes (2.11).

GMM Estimation with Heteroskedasticity of Unknown Form

The assumption that $\boldsymbol{\Omega}_0$ is known, even up to a scalar factor, is often too strong. What makes GMM estimation practical more generally is that, in both (2.10) and (2.11), $\boldsymbol{\Omega}_0$ appears only through the $l \times l$ matrix product $\boldsymbol{W}^\top\boldsymbol{\Omega}_0\boldsymbol{W}$. As we saw first in Part 1, Section 6.4, in the context of heteroskedasticity consistent covariance matrix estimation, n^{-1} times such a matrix can be estimated consistently if $\boldsymbol{\Omega}_0$ is a diagonal matrix. What is needed is a preliminary consistent estimate of the parameter vector $\boldsymbol{\beta}$, which furnishes residuals that are consistent estimates of the disturbances.

The preliminary estimates of $\boldsymbol{\beta}$ must be consistent, but they need not be asymptotically efficient, and so we can obtain them by using any convenient choice of $\boldsymbol{J}$ or $\boldsymbol{\Lambda}$. One choice that is often convenient is $\boldsymbol{\Lambda} = (\boldsymbol{W}^\top\boldsymbol{W})^{-1}$, in which case the preliminary estimator is the generalized IV estimator (F8.30). We then use the preliminary estimates $\hat{\boldsymbol{\beta}}$ to calculate the residuals $\hat{u}_t \equiv y_t - \boldsymbol{X}\hat{\boldsymbol{\beta}}$. A typical element of the matrix $n^{-1}\boldsymbol{W}^\top\boldsymbol{\Omega}_0\boldsymbol{W}$ can then be estimated by

$$\frac{1}{n}\sum_{t=1}^{n}\hat{u}_t^2 w_{ti} w_{tj}. \tag{2.14}$$

This estimator is very similar to (F6.26), and the estimator (2.14) can be proved to be consistent by using arguments just like those employed in Part 1, Section 6.4.

The matrix with typical element (2.14) can be written as $n^{-1}\boldsymbol{W}^\top\hat{\boldsymbol{\Omega}}\boldsymbol{W}$, where $\hat{\boldsymbol{\Omega}}$ is an $n \times n$ diagonal matrix with typical diagonal element $\hat{u}_t^2$. Then the **feasible efficient GMM estimator** is

$$\hat{\boldsymbol{\beta}}_{\text{FGMM}} = \big(\boldsymbol{X}^\top\boldsymbol{W}(\boldsymbol{W}^\top\hat{\boldsymbol{\Omega}}\boldsymbol{W})^{-1}\boldsymbol{W}^\top\boldsymbol{X}\big)^{-1}\boldsymbol{X}^\top\boldsymbol{W}(\boldsymbol{W}^\top\hat{\boldsymbol{\Omega}}\boldsymbol{W})^{-1}\boldsymbol{W}^\top\boldsymbol{y}, \tag{2.15}$$

which is just (2.10) with $\boldsymbol{\Omega}_0$ replaced by $\hat{\boldsymbol{\Omega}}$. Since $n^{-1}\boldsymbol{W}^\top\hat{\boldsymbol{\Omega}}\boldsymbol{W}$ consistently estimates $n^{-1}\boldsymbol{W}^\top\boldsymbol{\Omega}_0\boldsymbol{W}$, it follows that $\hat{\boldsymbol{\beta}}_{\text{FGMM}}$ is asymptotically equivalent to (2.10). It should be noted that, in calling (2.15) efficient, we mean that

it is asymptotically efficient within the class of estimators that use the given instrument set $\boldsymbol{W}$.

Like other procedures that start from a preliminary estimate, this one can be iterated. The GMM residuals $y_t - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\text{FGMM}}$ can be used to calculate a new estimate of $\boldsymbol{\Omega}$, which can then be used to obtain second-round GMM estimates, which can then be used to calculate yet another estimate of $\boldsymbol{\Omega}$, and so on. We will refer to this iterative procedure as **continuously updated GMM**, although it is not quite the same as the procedure by that name investigated by Hansen, Heaton, and Yaron (1996). Whether we stop after one round or continue until the procedure converges, the estimates have the same asymptotic distribution if the model is correctly specified. However, there is evidence that performing more iterations improves finite-sample performance. In practice, the covariance matrix is estimated by

$$\widehat{\text{Var}}(\hat{\boldsymbol{\beta}}_{\text{FGMM}}) = \big(\boldsymbol{X}^\top\boldsymbol{W}(\boldsymbol{W}^\top\hat{\boldsymbol{\Omega}}\boldsymbol{W})^{-1}\boldsymbol{W}^\top\boldsymbol{X}\big)^{-1}. \tag{2.16}$$

It is not hard to see that n times the estimator (2.16) tends to the asymptotic covariance matrix (2.09) as $n \to \infty$.

Fully Efficient GMM Estimation

In choosing to use a particular matrix of instrumental variables $\boldsymbol{W}$, we are choosing a particular representation of the information sets Ω_t appropriate for each observation in the sample. It is required that $\boldsymbol{W}_t \in \Omega_t$ for all t, and it follows from this that any deterministic function, linear or nonlinear, of the elements of $\boldsymbol{W}_t$ also belongs to Ω_t. It is quite clearly impossible to use all such deterministic functions as actual instrumental variables, and so the econometrician must make a choice. What we have established so far is that, once the choice of $\boldsymbol{W}$ is made, (2.08) gives the optimal set of linear combinations of the columns of $\boldsymbol{W}$ to use for estimation. What remains to be seen is how best to choose $\boldsymbol{W}$ out of all the possible valid instruments, given the information sets Ω_t.

In Part 1, Section 8.3, we saw that, for the model (2.01) with $\boldsymbol{\Omega} = \sigma^2\mathbf{I}$, the best choice, by the criterion of the asymptotic covariance matrix, is the matrix $\bar{\boldsymbol{X}}$ given in (F8.19) by the defining condition that $\mathrm{E}(\boldsymbol{X}_t \,|\, \Omega_t) = \bar{\boldsymbol{X}}_t$, where $\boldsymbol{X}_t$ and $\bar{\boldsymbol{X}}_t$ are the t^{th} rows of $\boldsymbol{X}$ and $\bar{\boldsymbol{X}}$, respectively. However, it is easy to see that this result does not hold unmodified when $\boldsymbol{\Omega}$ is not proportional to an identity matrix. Consider the GMM estimator (2.10), of which (2.15) is the feasible version, in the special case of exogenous explanatory variables, for which the obvious choice of instruments is $\boldsymbol{W} = \boldsymbol{X}$. If, for notational ease,

we write $\boldsymbol{\Omega}$ for the true covariance matrix $\boldsymbol{\Omega}_0$, (2.10) becomes

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{\text{GMM}} &= \big(\boldsymbol{X}^\top\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{\Omega}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{X}\big)^{-1}\boldsymbol{X}^\top\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{\Omega}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}\\ &= (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\Omega}\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{\Omega}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}\\ &= (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\Omega}\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{\Omega}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}\\ &= (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y} = \hat{\boldsymbol{\beta}}_{\text{OLS}}.\end{aligned}$$

However, we know from the results of Part 1, Section 9.2 that the efficient estimator is actually the GLS estimator

$$\hat{\boldsymbol{\beta}}_{\text{GLS}} = (\boldsymbol{X}^\top\boldsymbol{\Omega}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\Omega}^{-1}\boldsymbol{y}, \tag{2.17}$$

which, except in special cases, is different from $\hat{\boldsymbol{\beta}}_{\text{OLS}}$.

The GLS estimator (2.17) can be interpreted as an IV estimator, in which the instruments are the columns of $\boldsymbol{\Omega}^{-1}\boldsymbol{X}$. Thus it appears that, when $\boldsymbol{\Omega}$ is not a multiple of the identity matrix, the optimal instruments are no longer the explanatory variables $\boldsymbol{X}$, but rather the columns of $\boldsymbol{\Omega}^{-1}\boldsymbol{X}$. This suggests that, when at least some of the explanatory variables in the matrix $\boldsymbol{X}$ are not predetermined, the optimal choice of instruments is given by $\boldsymbol{\Omega}^{-1}\bar{\boldsymbol{X}}$. This choice combines the result of Part 1, Chapter 9 about the optimality of the GLS estimator with that of Part 1, Chapter 8 about the best instruments to use in place of explanatory variables that are not predetermined. It leads to the theoretical moment conditions

$$\mathrm{E}\big(\bar{\boldsymbol{X}}^\top\boldsymbol{\Omega}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\big) = \boldsymbol{0}. \tag{2.18}$$

Unfortunately, this solution to the optimal instruments problem does not always work, because the moment conditions in (2.18) may not be correct. To see why not, suppose that the disturbances are serially correlated, and that $\boldsymbol{\Omega}$ is consequently not a diagonal matrix. The i^{th} element of the matrix product in (2.18) can be expanded as

$$\sum_{t=1}^{n}\sum_{s=1}^{n}\bar{\boldsymbol{X}}_{ti}\,\omega^{ts}(y_s - \boldsymbol{X}_s\boldsymbol{\beta}), \tag{2.19}$$

where ω^{ts} is the ts^{th} element of $\boldsymbol{\Omega}^{-1}$. If we evaluate at the true parameter vector $\boldsymbol{\beta}_0$, we find that $y_s - \boldsymbol{X}_s\boldsymbol{\beta}_0 = u_s$. But, unless the columns of the matrix $\bar{\boldsymbol{X}}$ are exogenous, it is not in general the case that $\mathrm{E}(u_s \,|\, \bar{\boldsymbol{X}}_t) = 0$ for $s \neq t$, and, if this condition is not satisfied, the expectation of (2.19) is not zero in general. This issue was discussed at the end of Part 1, Section 9.3, and in more detail in Part 1, Section 9.8, in connection with the use of GLS when one of the explanatory variables is a lagged dependent variable.

Choosing Valid Instruments

As in Part 1, Section 9.2, we can construct an $n \times n$ matrix $\boldsymbol{\Psi}$, usually triangular, that satisfies the equation $\boldsymbol{\Omega}^{-1} = \boldsymbol{\Psi}\boldsymbol{\Psi}^\top$. As in equation (F9.03) of Part 1, Section 9.2, we can premultiply regression (2.01) by $\boldsymbol{\Psi}^\top$ to get

$$\boldsymbol{\Psi}^\top\boldsymbol{y} = \boldsymbol{\Psi}^\top\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\Psi}^\top\boldsymbol{u}, \tag{2.20}$$

with the result that the covariance matrix of the transformed disturbance vector, $\boldsymbol{\Psi}^\top\boldsymbol{u}$, is just the identity matrix. Suppose that we propose to use a matrix $\boldsymbol{Z}$ of instruments in order to estimate the transformed model, so that we are led to consider the theoretical moment conditions

$$\mathrm{E}\big(\boldsymbol{Z}^\top\boldsymbol{\Psi}^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\big) = \boldsymbol{0}. \tag{2.21}$$

If these conditions are to be correct, then what we need is that, for each t, $\mathrm{E}\big((\boldsymbol{\Psi}^\top\boldsymbol{u})_t \,|\, \boldsymbol{Z}_t\big) = 0$, where the subscript t is used to select the t^{th} row of the corresponding vector or matrix.

If $\boldsymbol{X}$ is exogenous, the optimal instruments are given by the matrix $\boldsymbol{\Omega}^{-1}\boldsymbol{X}$, and the moment conditions for efficient estimation are $\mathrm{E}\big(\boldsymbol{X}^\top\boldsymbol{\Omega}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\big) = \boldsymbol{0}$, which can also be written as

$$\mathrm{E}\big(\boldsymbol{X}^\top\boldsymbol{\Psi}\boldsymbol{\Psi}^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\big) = \boldsymbol{0}. \tag{2.22}$$

Comparison with (2.21) shows that the optimal choice of $\boldsymbol{Z}$ is $\boldsymbol{\Psi}^\top\boldsymbol{X}$. Even if $\boldsymbol{X}$ is not exogenous, (2.22) is a correct set of moment conditions if

$$\mathrm{E}\big((\boldsymbol{\Psi}^\top\boldsymbol{u})_t \,|\, (\boldsymbol{\Psi}^\top\boldsymbol{X})_t\big) = 0. \tag{2.23}$$

But this is not true in general when $\boldsymbol{X}$ is not exogenous. Consequently, we seek a new definition for $\bar{\boldsymbol{X}}$, such that (2.23) becomes true when $\boldsymbol{X}$ is replaced by $\bar{\boldsymbol{X}}$.

In most cases, it is possible to choose $\boldsymbol{\Psi}$ so that $(\boldsymbol{\Psi}^\top\boldsymbol{u})_t$ is an innovation, that is, so that $\mathrm{E}\big((\boldsymbol{\Psi}^\top\boldsymbol{u})_t \,|\, \Omega_t\big) = 0$. As an example, see the analysis of models with AR(1) disturbances in Part 1, Section 9.8, especially the discussion surrounding (F9.44). What is then required for condition (2.23) is that $(\boldsymbol{\Psi}^\top\bar{\boldsymbol{X}})_t$ should be predetermined in period t. If $\boldsymbol{\Omega}$ is diagonal, and so also $\boldsymbol{\Psi}$, the old definition of $\bar{\boldsymbol{X}}$ works, because $(\boldsymbol{\Psi}^\top\bar{\boldsymbol{X}})_t = \Psi_{tt}\bar{\boldsymbol{X}}_t$, where Ψ_{tt} is the t^{th} diagonal element of $\boldsymbol{\Psi}$, and this belongs to Ω_t by construction. If $\boldsymbol{\Omega}$ contains off-diagonal elements, however, the old definition of $\bar{\boldsymbol{X}}$ no longer works in general. Since what we need is that $(\boldsymbol{\Psi}^\top\bar{\boldsymbol{X}})_t$ should belong to Ω_t, we instead define $\bar{\boldsymbol{X}}$ implicitly by the equation

$$\mathrm{E}\big((\boldsymbol{\Psi}^\top\boldsymbol{X})_t \,|\, \Omega_t\big) = (\boldsymbol{\Psi}^\top\bar{\boldsymbol{X}})_t. \tag{2.24}$$

This implicit definition must be implemented on a case-by-case basis. One example is given in Exercise 2.5.

By setting $\boldsymbol{Z} = \boldsymbol{\Psi}^\top \bar{\boldsymbol{X}}$, we find that the moment conditions (2.21) become

$$\mathrm{E}\big(\bar{\boldsymbol{X}}^\top \boldsymbol{\Psi}\boldsymbol{\Psi}^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\big) = \mathrm{E}\big(\bar{\boldsymbol{X}}^\top \boldsymbol{\Omega}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\big) = \boldsymbol{0}. \tag{2.25}$$

These conditions do indeed use $\boldsymbol{\Omega}^{-1}\bar{\boldsymbol{X}}$ as instruments, albeit with a possibly redefined $\bar{\boldsymbol{X}}$. The estimator based on (2.25) is

$$\hat{\boldsymbol{\beta}}_{\mathrm{EGMM}} \equiv (\bar{\boldsymbol{X}}^\top \boldsymbol{\Omega}^{-1}\boldsymbol{X})^{-1}\bar{\boldsymbol{X}}^\top \boldsymbol{\Omega}^{-1}\boldsymbol{y}, \tag{2.26}$$

where EGMM denotes "fully efficient GMM." The asymptotic covariance matrix of (2.26) can be computed using (2.09), in which, on the basis of (2.25), we see that $\boldsymbol{W}$ is to be replaced by $\boldsymbol{\Psi}^\top\bar{\boldsymbol{X}}$, $\boldsymbol{X}$ by $\boldsymbol{\Psi}^\top\boldsymbol{X}$, and $\boldsymbol{\Omega}$ by $\mathbf{I}$. We cannot apply (2.09) directly with instruments $\boldsymbol{\Omega}^{-1}\bar{\boldsymbol{X}}$, because there is no reason to suppose that the result (2.02) holds for the untransformed disturbances $\boldsymbol{u}$ and the instruments $\boldsymbol{\Omega}^{-1}\bar{\boldsymbol{X}}$. The result is

$$\operatorname*{plim}_{n\to\infty}\left(\frac{1}{n}\bar{\boldsymbol{X}}^\top \boldsymbol{\Omega}^{-1}\boldsymbol{X}\left(\frac{1}{n}\bar{\boldsymbol{X}}^\top \boldsymbol{\Omega}^{-1}\bar{\boldsymbol{X}}\right)^{-1}\frac{1}{n}\boldsymbol{X}^\top \boldsymbol{\Omega}^{-1}\bar{\boldsymbol{X}}\right)^{-1}. \tag{2.27}$$

By exactly the same argument as that used in (F8.21), we find that, for any matrix $\boldsymbol{Z}$ that satisfies $\boldsymbol{Z}_t \in \Omega_t$,

$$\operatorname*{plim}_{n\to\infty}\frac{1}{n}\boldsymbol{Z}^\top\boldsymbol{\Psi}^\top\boldsymbol{X} = \operatorname*{plim}_{n\to\infty}\frac{1}{n}\boldsymbol{Z}^\top\boldsymbol{\Psi}^\top\bar{\boldsymbol{X}}. \tag{2.28}$$

Since $(\boldsymbol{\Psi}^\top\bar{\boldsymbol{X}})_t \in \Omega_t$, this implies that

$$\begin{aligned}\operatorname*{plim}_{n\to\infty}\frac{1}{n}\bar{\boldsymbol{X}}^\top\boldsymbol{\Omega}^{-1}\boldsymbol{X} &= \operatorname*{plim}_{n\to\infty}\frac{1}{n}\bar{\boldsymbol{X}}^\top\boldsymbol{\Psi}\boldsymbol{\Psi}^\top\boldsymbol{X}\\ &= \operatorname*{plim}_{n\to\infty}\frac{1}{n}\bar{\boldsymbol{X}}^\top\boldsymbol{\Psi}\boldsymbol{\Psi}^\top\bar{\boldsymbol{X}} = \operatorname*{plim}_{n\to\infty}\frac{1}{n}\bar{\boldsymbol{X}}^\top\boldsymbol{\Omega}^{-1}\bar{\boldsymbol{X}}.\end{aligned}$$

Therefore, the asymptotic covariance matrix (2.27) simplifies to

$$\operatorname*{plim}_{n\to\infty}\left(\frac{1}{n}\bar{\boldsymbol{X}}^\top\boldsymbol{\Omega}^{-1}\bar{\boldsymbol{X}}\right)^{-1}. \tag{2.29}$$

Although the matrix (2.09) is less of a sandwich than (2.07), the matrix (2.29) is still less of one than (2.09). This is a clear indication of the fact that the instruments $\boldsymbol{\Omega}^{-1}\bar{\boldsymbol{X}}$, which yield the estimator $\hat{\boldsymbol{\beta}}_{\mathrm{EGMM}}$, are indeed optimal. Readers are asked to check this formally in Exercise 2.7.

In most cases, $\bar{\boldsymbol{X}}$ is not observed, but it can often be estimated consistently. The usual state of affairs is that we have an $n \times l$ matrix $\boldsymbol{W}$ of instruments, such that $\mathcal{S}(\bar{\boldsymbol{X}}) \subseteq \mathcal{S}(\boldsymbol{W})$ and

$$(\boldsymbol{\Psi}^\top\boldsymbol{W})_t \in \Omega_t. \tag{2.30}$$

This last condition is the form taken by the **predeterminedness condition** when $\boldsymbol{\Omega}$ is not proportional to the identity matrix. The theoretical moment conditions used for (overidentified) estimation are then

$$\mathrm{E}\big(\boldsymbol{W}^\top\boldsymbol{\Omega}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\big) = \mathrm{E}\big(\boldsymbol{W}^\top\boldsymbol{\Psi}\boldsymbol{\Psi}^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\big) = \boldsymbol{0}, \tag{2.31}$$

from which it can be seen that what we are in fact doing is estimating the transformed model (2.20) using the transformed instruments $\boldsymbol{\Psi}^\top\boldsymbol{W}$. The result of Exercise 2.8 shows that, if indeed $\mathcal{S}(\bar{\boldsymbol{X}}) \subseteq \mathcal{S}(\boldsymbol{W})$, the asymptotic covariance matrix of the resulting estimator is still (2.29). Exercise 2.9 investigates what happens if this condition is not satisfied.

The main obstacle to the use of the efficient estimator $\hat{\boldsymbol{\beta}}_{\mathrm{EGMM}}$ is thus not the difficulty of estimating $\bar{\boldsymbol{X}}$, but rather the fact that $\boldsymbol{\Omega}$ is usually not known. As with the GLS estimators we studied in Part 1, Chapter 9, $\hat{\boldsymbol{\beta}}_{\mathrm{EGMM}}$ cannot be calculated unless we either know $\boldsymbol{\Omega}$ or can estimate it consistently, usually by knowing the form of $\boldsymbol{\Omega}$ as a function of parameters that can be estimated consistently. But whenever there is heteroskedasticity or serial correlation of unknown form, this is impossible. The best we can then do, asymptotically, is to use the feasible efficient GMM estimator (2.15). Therefore, when we later refer to GMM estimators without further qualification, we will normally mean feasible efficient ones.

2.3 HAC Covariance Matrix Estimation

Up to this point, we have seen how to obtain feasible efficient GMM estimates only when the matrix $\boldsymbol{\Omega}$ is known to be diagonal, in which case we can use the estimator (2.15). In this section, we also allow for the possibility of serial correlation of unknown form, which causes $\boldsymbol{\Omega}$ to have nonzero off-diagonal elements. When the pattern of the serial correlation is unknown, we can still, under fairly weak regularity conditions, estimate the covariance matrix of the sample moments by using a **heteroskedasticity and autocorrelation consistent**, or **HAC**, estimator of the matrix $n^{-1}\boldsymbol{W}^\top\boldsymbol{\Omega}\boldsymbol{W}$. This estimator, multiplied by n, can then be used in place of $\boldsymbol{W}^\top\hat{\boldsymbol{\Omega}}\boldsymbol{W}$ in the feasible efficient GMM estimator (2.15).

The asymptotic covariance matrix of the vector $n^{-1/2}\boldsymbol{W}^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$ of sample moments, evaluated at $\boldsymbol{\beta} = \boldsymbol{\beta}_0$, is defined as follows:

$$\boldsymbol{\Sigma} \equiv \operatorname*{plim}_{n\to\infty}\frac{1}{n}\boldsymbol{W}^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_0)(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_0)^\top\boldsymbol{W} = \operatorname*{plim}_{n\to\infty}\frac{1}{n}\boldsymbol{W}^\top\boldsymbol{\Omega}\boldsymbol{W}. \tag{2.32}$$

A HAC estimator of $\boldsymbol{\Sigma}$ is a matrix $\hat{\boldsymbol{\Sigma}}$ constructed so that $\hat{\boldsymbol{\Sigma}}$ consistently estimates $\boldsymbol{\Sigma}$ when the disturbances u_t display any pattern of heteroskedasticity and/or autocorrelation that satisfies certain, generally quite weak, conditions.

In order to derive such an estimator, we begin by rewriting the definition of $\boldsymbol{\Sigma}$ in an alternative way:

$$\boldsymbol{\Sigma} = \lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} \sum_{s=1}^{n} \mathrm{E}\big(u_t u_s \boldsymbol{W}_t^\top \boldsymbol{W}_s\big), \tag{2.33}$$

in which we assume that a law of large numbers can be used to justify replacing the probability limit in (2.32) by the expectations in (2.33).

For regression models with heteroskedasticity but no autocorrelation, only the terms with $t = s$ contribute to (2.33). Therefore, for such models, we can estimate $\boldsymbol{\Sigma}$ consistently by simply ignoring the expectation operator and replacing the disturbances u_t by least-squares residuals $\hat{u}_t$, possibly with a modification designed to offset the tendency for such residuals to be too small. The obvious way to estimate (2.33) when there may be serial correlation is again simply to drop the expectations operator and replace $u_t u_s$ by $\hat{u}_t \hat{u}_s$, where $\hat{u}_t$ denotes the t^{th} residual from some consistent but inefficient estimation procedure, such as generalized IV. Unfortunately, this approach does not work. To see why not, we need to rewrite (2.33) in yet another way. Let us define the **autocovariance matrices** of the $\boldsymbol{W}_t^\top u_t$ as follows:

$$\boldsymbol{\Gamma}(j) \equiv \begin{cases} \frac{1}{n} \sum_{t=j+1}^{n} \mathrm{E}(u_t u_{t-j} \boldsymbol{W}_t^\top \boldsymbol{W}_{t-j}) & \text{for } j \geq 0, \\ \frac{1}{n} \sum_{t=-j+1}^{n} \mathrm{E}(u_{t+j} u_t \boldsymbol{W}_{t+j}^\top \boldsymbol{W}_t) & \text{for } j < 0. \end{cases} \tag{2.34}$$

Because there are l moment conditions, these are $l \times l$ matrices. It is easy to check that $\boldsymbol{\Gamma}(j) = \boldsymbol{\Gamma}^\top(-j)$. Then, in terms of the matrices $\boldsymbol{\Gamma}(j)$, expression (2.33) becomes

$$\boldsymbol{\Sigma} = \lim_{n\to\infty} \sum_{j=-n+1}^{n-1} \boldsymbol{\Gamma}(j) = \lim_{n\to\infty} \Big(\boldsymbol{\Gamma}(0) + \sum_{j=1}^{n-1} \big(\boldsymbol{\Gamma}(j) + \boldsymbol{\Gamma}^\top(j)\big)\Big). \tag{2.35}$$

Therefore, in order to estimate $\boldsymbol{\Sigma}$, we apparently need to estimate all of the autocovariance matrices for $j = 0, \ldots, n-1$.

If $\hat{u}_t$ denotes a typical residual from some preliminary estimator, the **sample autocovariance matrix** of order j, $\hat{\boldsymbol{\Gamma}}(j)$, is just the appropriate expression in (2.34), without the expectation operator, and with the random variables u_t and u_{t-j} replaced by $\hat{u}_t$ and $\hat{u}_{t-j}$, respectively. For any $j \geq 0$, this is

$$\hat{\boldsymbol{\Gamma}}(j) = \frac{1}{n} \sum_{t=j+1}^{n} \hat{u}_t \hat{u}_{t-j} \boldsymbol{W}_t^\top \boldsymbol{W}_{t-j}. \tag{2.36}$$

Unfortunately, the sample autocovariance matrix $\hat{\boldsymbol{\Gamma}}(j)$ of order j is not a consistent estimator of the true autocovariance matrix for arbitrary j. Suppose, for instance, that $j = n-2$. Then, from (2.36), we see that $\hat{\boldsymbol{\Gamma}}(j)$ has only two terms, and no conceivable law of large numbers can apply to only two terms. In fact, $\hat{\boldsymbol{\Gamma}}(n-2)$ must tend to zero as $n \to \infty$ because of the factor of n^{-1} in its definition.

The solution to this problem is to restrict our attention to models for which the actual autocovariances mimic the behavior of the sample autocovariances, and for which therefore the actual autocovariance of order j tends to zero as $j \to \infty$. A great many stochastic processes generate disturbances for which the $\boldsymbol{\Gamma}(j)$ do have this property. In such cases, we can drop most of the sample autocovariance matrices that appear in the sample analog of (2.35) by eliminating ones for which $|j|$ is greater than some chosen threshold, say p. This yields the following estimator for $\boldsymbol{\Sigma}$:

$$\hat{\boldsymbol{\Sigma}}_{\text{HW}} = \hat{\boldsymbol{\Gamma}}(0) + \sum_{j=1}^{p} \big(\hat{\boldsymbol{\Gamma}}(j) + \hat{\boldsymbol{\Gamma}}^\top(j)\big), \tag{2.37}$$

We refer to this estimator as the **Hansen-White estimator**, because it was originally proposed by Hansen (1982) and White and Domowitz (1984); see also White (2000).

For the purposes of asymptotic theory, it is necessary to let the parameter p, which is called the **lag truncation parameter**, go to infinity in (2.37) at some suitable rate as the sample size goes to infinity. A typical rate would be $n^{1/4}$. This ensures that, for large enough n, all the nonzero $\boldsymbol{\Gamma}(j)$ are estimated consistently. Unfortunately, this type of result does not say how large p should be in practice. In most cases, we have a given, finite, sample size, and we need to choose a specific value of p.

The Hansen-White estimator (2.37) suffers from one very serious deficiency: In finite samples, it need not be positive definite or even positive semidefinite. If one happens to encounter a data set that yields a nondefinite $\hat{\boldsymbol{\Sigma}}_{\text{HW}}$, then, since the weighting matrix for GMM must be positive definite, (2.37) is unusable. Luckily, there are numerous ways out of this difficulty. The one that is most widely used was suggested by Newey and West (1987). The **Newey-West estimator** they propose is

$$\hat{\boldsymbol{\Sigma}}_{\text{NW}} = \hat{\boldsymbol{\Gamma}}(0) + \sum_{j=1}^{p} \Big(1 - \frac{j}{p+1}\Big) \big(\hat{\boldsymbol{\Gamma}}(j) + \hat{\boldsymbol{\Gamma}}^\top(j)\big), \tag{2.38}$$

in which each sample autocovariance matrix $\hat{\boldsymbol{\Gamma}}(j)$ is multiplied by a weight $1 - j/(p+1)$ that decreases linearly as j increases. The weight is $p/(p+1)$ for $j = 1$, and it then decreases by steps of $1/(p+1)$ down to a value of $1/(p+1)$ for $j = p$. This estimator evidently tends to underestimate the autocovariance

matrices, especially for larger values of j. Therefore, p should almost certainly be larger for (2.38) than for (2.37). As with the Hansen-White estimator, p must increase as n does, and the appropriate rate is $n^{1/3}$. A procedure for selecting p automatically was proposed by Newey and West (1994), but it is too complicated to discuss here.

Both the Hansen-White and the Newey-West HAC estimators of $\boldsymbol{\Sigma}$ can be written in the form

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\boldsymbol{W}^\top\hat{\boldsymbol{\Omega}}\boldsymbol{W} \tag{2.39}$$

for an appropriate choice of $\hat{\boldsymbol{\Omega}}$. This fact, which we will exploit in the next section, follows from the observation that there exist $n \times n$ matrices $\boldsymbol{U}(j)$ such that the $\hat{\boldsymbol{\Gamma}}(j)$ can be expressed in the form $n^{-1}\boldsymbol{W}^\top\boldsymbol{U}(j)\boldsymbol{W}$, as readers are asked to check in Exercise 2.10.

The Newey-West estimator is by no means the only HAC estimator that is guaranteed to be positive definite. Andrews (1991) provides a detailed treatment of HAC estimation, suggests some alternatives to the Newey-West estimator, and shows that, in some circumstances, they may perform better than it does in finite samples. A different approach to HAC estimation is suggested by Andrews and Monahan (1992). Since this material is relatively advanced and specialized, we will not pursue it further here. Interested readers may wish to consult Hamilton (1994, Chapter 10) as well as the references already given.

Feasible Efficient GMM Estimation

In practice, efficient GMM estimation in the presence of heteroskedasticity and serial correlation of unknown form works as follows. As in the case with only heteroskedasticity that was discussed in Section 2.2, we first obtain consistent but inefficient estimates, probably by using generalized IV. These estimates yield residuals $\hat{u}_t$, from which we next calculate a matrix $\hat{\boldsymbol{\Sigma}}$ that estimates $\boldsymbol{\Sigma}$ consistently, using (2.37), (2.38), or some other HAC estimator. The feasible efficient GMM estimator, which generalizes (2.15), is then

$$\hat{\boldsymbol{\beta}}_{\text{FGMM}} = (\boldsymbol{X}^\top\boldsymbol{W}\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{W}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{W}\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{W}^\top\boldsymbol{y}. \tag{2.40}$$

As before, this procedure may be iterated. The first-round GMM residuals may be used to obtain a new estimate of $\boldsymbol{\Sigma}$, which may be used to obtain second-round GMM estimates, and so on. For a correctly specified model, iteration should not affect the asymptotic properties of the estimates.

We can estimate the covariance matrix of (2.40) by

$$\widehat{\text{Var}}(\hat{\boldsymbol{\beta}}_{\text{FGMM}}) = n(\boldsymbol{X}^\top\boldsymbol{W}\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{W}^\top\boldsymbol{X})^{-1}, \tag{2.41}$$

which is the analog of (2.16). The factor of n here is needed to offset the factor of n^{-1} in the definition of $\hat{\boldsymbol{\Sigma}}$. We do not need to include such a factor

in (2.40), because the two factors of n^{-1} cancel out. As usual, the covariance matrix estimator (2.41) can be used to construct pseudo-t tests and other Wald tests, and asymptotic confidence intervals and confidence regions may also be based on it. The GMM criterion function that corresponds to (2.40) is

$$\frac{1}{n}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top\boldsymbol{W}\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{W}^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}). \tag{2.42}$$

Once again, we need a factor of n^{-1} here to offset the one in $\hat{\boldsymbol{\Sigma}}$.

The feasible efficient GMM estimator (2.40) can be used even when all the columns of $\boldsymbol{X}$ are valid instruments and OLS would be the estimator of choice if the disturbances were not heteroskedastic and/or serially correlated. In this case, $\boldsymbol{W}$ typically consists of $\boldsymbol{X}$ augmented by a number of functions of the columns of $\boldsymbol{X}$, such as squares and cross-products, and $\hat{\boldsymbol{\Omega}}$ has squared OLS residuals on the diagonal. This estimator, which was proposed by Cragg (1983) for models with heteroskedastic disturbances, is asymptotically more efficient than OLS whenever $\boldsymbol{\Omega}$ is not proportional to an identity matrix.

2.4 Tests Based on the GMM Criterion Function

For models estimated by instrumental variables, we saw in Part 1, Section 8.5 that any set of r equality restrictions can be tested by taking the difference between the minimized values of the IV criterion function for the restricted and unrestricted models, and then dividing it by a consistent estimate of the disturbance variance. The resulting test statistic is asymptotically distributed as $\chi^2(r)$. For models estimated by (feasible) efficient GMM, a very similar testing procedure is available. In this case, as we will see, the difference between the constrained and unconstrained minima of the GMM criterion function is asymptotically distributed as $\chi^2(r)$. There is no need to divide by an estimate of σ^2, because the GMM criterion function already takes account of the covariance matrix of the disturbances.

Tests of Overidentifying Restrictions

Whenever $l > k$, a model estimated by GMM involves $l - k$ overidentifying restrictions. As in the IV case, tests of these restrictions are even easier to perform than tests of other restrictions, because the minimized value of the optimal GMM criterion function (2.11), with $n^{-1}\boldsymbol{W}^\top\boldsymbol{\Omega}_0\boldsymbol{W}$ replaced by a HAC estimate, provides an asymptotically valid test statistic. When the HAC estimate $\hat{\boldsymbol{\Sigma}}$ is expressed as in (2.39), the GMM criterion function (2.42) can be written as

$$Q(\boldsymbol{\beta}, \boldsymbol{y}) \equiv (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top\boldsymbol{W}(\boldsymbol{W}^\top\hat{\boldsymbol{\Omega}}\boldsymbol{W})^{-1}\boldsymbol{W}^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}). \tag{2.43}$$

Since HAC estimators are consistent, the asymptotic distribution of (2.43), for given $\boldsymbol{\beta}$, is the same whether we use the unknown true $\boldsymbol{\Omega}_0$ or a matrix $\hat{\boldsymbol{\Omega}}$

that provides a HAC estimate. For simplicity, we therefore use the true $\boldsymbol{\Omega}_0$, omitting the subscript 0 for ease of notation. The asymptotic equivalence of the $\hat{\boldsymbol{\beta}}_{\text{FGMM}}$ of (2.15) or (2.40) and the $\hat{\boldsymbol{\beta}}_{\text{GMM}}$ of (2.10) further implies that what we will prove for the criterion function (2.43) evaluated at $\hat{\boldsymbol{\beta}}_{\text{GMM}}$, with $\hat{\boldsymbol{\Omega}}$ replaced by $\boldsymbol{\Omega}$, is equally true for (2.43) evaluated at $\hat{\boldsymbol{\beta}}_{\text{FGMM}}$.

We remarked in Section 2.2 that $Q(\boldsymbol{\beta}_0, \boldsymbol{y})$, where $\boldsymbol{\beta}_0$ is the true parameter vector, is asymptotically distributed as $\chi^2(l)$. In contrast, the minimized criterion function $Q(\hat{\boldsymbol{\beta}}_{\text{GMM}}, \boldsymbol{y})$ is distributed as $\chi^2(l-k)$, because we lose k degrees of freedom as a consequence of having estimated k parameters. In order to demonstrate this result, we first express (2.43) in terms of an orthogonal projection matrix. This allows us to reuse many of the calculations performed in Part 1, Chapter 8.

As in Section 2.2, we make use of a possibly triangular matrix $\boldsymbol{\Psi}$ that satisfies the equation $\boldsymbol{\Omega}^{-1} = \boldsymbol{\Psi}\boldsymbol{\Psi}^\top$, or, equivalently,

$$\boldsymbol{\Omega} = (\boldsymbol{\Psi}^\top)^{-1}\boldsymbol{\Psi}^{-1}. \tag{2.44}$$

If the $n \times l$ matrix $\boldsymbol{A}$ is defined as $\boldsymbol{\Psi}^{-1}\boldsymbol{W}$, and $\boldsymbol{P}_{\boldsymbol{A}} \equiv \boldsymbol{A}(\boldsymbol{A}^\top\boldsymbol{A})^{-1}\boldsymbol{A}^\top$, then

$$\begin{aligned} Q(\boldsymbol{\beta}, \boldsymbol{y}) &= (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top\boldsymbol{\Psi}\boldsymbol{\Psi}^{-1}\boldsymbol{W}\big(\boldsymbol{W}^\top(\boldsymbol{\Psi}^\top)^{-1}\boldsymbol{\Psi}^{-1}\boldsymbol{W}\big)^{-1}\boldsymbol{W}^\top(\boldsymbol{\Psi}^\top)^{-1}\boldsymbol{\Psi}^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \\ &= (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top\boldsymbol{\Psi}\boldsymbol{P}_{\boldsymbol{A}}\boldsymbol{\Psi}^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}). \end{aligned} \tag{2.45}$$

Since $\hat{\boldsymbol{\beta}}_{\text{GMM}}$ minimizes (2.45), we see that one way to write it is

$$\hat{\boldsymbol{\beta}}_{\text{GMM}} = (\boldsymbol{X}^\top\boldsymbol{\Psi}\boldsymbol{P}_{\boldsymbol{A}}\boldsymbol{\Psi}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\Psi}\boldsymbol{P}_{\boldsymbol{A}}\boldsymbol{\Psi}^\top\boldsymbol{y}; \tag{2.46}$$

compare (2.10). Expression (2.46) makes it clear that $\hat{\boldsymbol{\beta}}_{\text{GMM}}$ can be thought of as a generalized IV estimator for the regression of $\boldsymbol{\Psi}^\top\boldsymbol{y}$ on $\boldsymbol{\Psi}^\top\boldsymbol{X}$ using instruments $\boldsymbol{A} \equiv \boldsymbol{\Psi}^{-1}\boldsymbol{W}$. As in (F8.62), it can be shown that

$$\boldsymbol{P}_{\boldsymbol{A}}\boldsymbol{\Psi}^\top(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\text{GMM}}) = \boldsymbol{P}_{\boldsymbol{A}}(\mathbf{I} - \boldsymbol{P}_{\boldsymbol{P}_{\boldsymbol{A}}\boldsymbol{\Psi}^\top\boldsymbol{X}})\boldsymbol{\Psi}^\top\boldsymbol{y},$$

where $\boldsymbol{P}_{\boldsymbol{P}_{\boldsymbol{A}}\boldsymbol{\Psi}^\top\boldsymbol{X}}$ is the orthogonal projection on to the subspace $\mathcal{S}(\boldsymbol{P}_{\boldsymbol{A}}\boldsymbol{\Psi}^\top\boldsymbol{X})$. It follows that

$$Q(\hat{\boldsymbol{\beta}}_{\text{GMM}}, \boldsymbol{y}) = \boldsymbol{y}^\top\boldsymbol{\Psi}(\boldsymbol{P}_{\boldsymbol{A}} - \boldsymbol{P}_{\boldsymbol{P}_{\boldsymbol{A}}\boldsymbol{\Psi}^\top\boldsymbol{X}})\boldsymbol{\Psi}^\top\boldsymbol{y}, \tag{2.47}$$

which is the analog for GMM estimation of expression (F8.62) for generalized IV estimation.

Now notice that

$$\begin{aligned} &(\boldsymbol{P}_{\boldsymbol{A}} - \boldsymbol{P}_{\boldsymbol{P}_{\boldsymbol{A}}\boldsymbol{\Psi}^\top\boldsymbol{X}})\boldsymbol{\Psi}^\top\boldsymbol{X} \\ &\quad= \boldsymbol{P}_{\boldsymbol{A}}\boldsymbol{\Psi}^\top\boldsymbol{X} - \boldsymbol{P}_{\boldsymbol{A}}\boldsymbol{\Psi}^\top\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{\Psi}\boldsymbol{P}_{\boldsymbol{A}}\boldsymbol{\Psi}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\Psi}\boldsymbol{P}_{\boldsymbol{A}}\boldsymbol{\Psi}^\top\boldsymbol{X} \\ &\quad= \boldsymbol{P}_{\boldsymbol{A}}\boldsymbol{\Psi}^\top\boldsymbol{X} - \boldsymbol{P}_{\boldsymbol{A}}\boldsymbol{\Psi}^\top\boldsymbol{X} = \mathbf{O}. \end{aligned}$$

Since $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{u}$ if the model we are estimating is correctly specified, this implies that (2.47) is equal to

$$Q(\hat{\boldsymbol{\beta}}_{\text{GMM}}, \boldsymbol{y}) = \boldsymbol{u}^\top\boldsymbol{\Psi}(\boldsymbol{P}_{\boldsymbol{A}} - \boldsymbol{P}_{\boldsymbol{P}_{\boldsymbol{A}}\boldsymbol{\Psi}^\top\boldsymbol{X}})\boldsymbol{\Psi}^\top\boldsymbol{u}. \tag{2.48}$$

This expression can be compared with the value of the criterion function evaluated at $\boldsymbol{\beta}_0$, which can be obtained directly from (2.45):

$$Q(\boldsymbol{\beta}_0, \boldsymbol{y}) = \boldsymbol{u}^\top\boldsymbol{\Psi}\boldsymbol{P}_{\boldsymbol{A}}\boldsymbol{\Psi}^\top\boldsymbol{u}. \tag{2.49}$$

The two expressions (2.48) and (2.49) show clearly where the k degrees of freedom are lost when we estimate $\boldsymbol{\beta}$. We know that $\mathrm{E}(\boldsymbol{\Psi}^\top\boldsymbol{u}) = \mathbf{0}$ and that $\mathrm{E}(\boldsymbol{\Psi}^\top\boldsymbol{u}\boldsymbol{u}^\top\boldsymbol{\Psi}) = \boldsymbol{\Psi}^\top\boldsymbol{\Omega}\boldsymbol{\Psi} = \mathbf{I}$, by (2.44). The dimension of the space $\mathcal{S}(\boldsymbol{A})$ is equal to l. Therefore, the extension of Theorem 4.1 treated in Exercise 2.2 allows us to conclude that (2.49) is asymptotically distributed as $\chi^2(l)$. Since $\mathcal{S}(\boldsymbol{P}_{\boldsymbol{A}}\boldsymbol{\Psi}^\top\boldsymbol{X})$ is a k–dimensional subspace of $\mathcal{S}(\boldsymbol{A})$, it follows (see Part 1, Exercise 3.F18) that $\boldsymbol{P}_{\boldsymbol{A}} - \boldsymbol{P}_{\boldsymbol{P}_{\boldsymbol{A}}\boldsymbol{\Psi}^\top\boldsymbol{X}}$ is an orthogonal projection on to a space of dimension $l - k$, from which we see that (2.48) is asymptotically distributed as $\chi^2(l-k)$. Replacing $\boldsymbol{\beta}_0$ by $\hat{\boldsymbol{\beta}}_{\text{GMM}}$ in (2.48) thus leads to the loss of the k dimensions of the space $\mathcal{S}(\boldsymbol{P}_{\boldsymbol{A}}\boldsymbol{\Psi}^\top\boldsymbol{X})$, which are "used up" when we obtain $\hat{\boldsymbol{\beta}}_{\text{GMM}}$.

The statistic $Q(\hat{\boldsymbol{\beta}}_{\text{GMM}}, \boldsymbol{y})$ is the analog, for efficient GMM estimation, of the Sargan test statistic that was discussed in Part 1, Section 8.6. This statistic was suggested by Hansen (1982) in the famous paper that first proposed GMM estimation under that name. It is often called **Hansen's overidentification statistic** or **Hansen's *J* statistic**. However, we prefer to call it the **Hansen-Sargan statistic** to stress its close relationship with the Sargan test of overidentifying restrictions in the context of generalized IV estimation.

As in the case of IV estimation, a Hansen-Sargan test may reject the null hypothesis for more than one reason. Perhaps the model is misspecified, either because one or more of the instruments should have been included among the regressors, or for some other reason. Perhaps one or more of the instruments is invalid because it is correlated with the disturbances. Or perhaps the finite-sample distribution of the test statistic just happens to differ substantially from its asymptotic distribution. In the case of feasible GMM estimation, especially involving HAC covariance matrices, this last possibility should not be discounted. See, among others, Hansen, Heaton, and Yaron (1996) and West and Wilcox (1996).

Tests of Linear Restrictions

Just as in the case of generalized IV, both linear and nonlinear restrictions on regression models can be tested by using the difference between the constrained and unconstrained minima of the GMM criterion function as a test statistic. Under weak conditions, this test statistic is asymptotically distributed as χ^2 with as many degrees of freedom as there are restrictions to

be tested. For simplicity, we restrict our attention to zero restrictions on the linear regression model (2.01). This model can be rewritten as

$$\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{u}, \quad \mathrm{E}(\boldsymbol{u}\boldsymbol{u}^\top) = \boldsymbol{\Omega}, \tag{2.50}$$

where $\boldsymbol{\beta}_1$ is a k_1–vector and $\boldsymbol{\beta}_2$ is a k_2–vector, with $k = k_1 + k_2$. We wish to test the restrictions $\boldsymbol{\beta}_2 = \boldsymbol{0}$.

If we estimate (2.50) by feasible efficient GMM using $\boldsymbol{W}$ as the matrix of instruments, subject to the restriction that $\boldsymbol{\beta}_2 = \boldsymbol{0}$, we obtain the restricted estimates $\tilde{\boldsymbol{\beta}}_{\mathrm{FGMM}} = [\tilde{\boldsymbol{\beta}}_1 \vdots \boldsymbol{0}]$. By the reasoning that leads to (2.48), we see that, if indeed $\boldsymbol{\beta}_2 = \boldsymbol{0}$, the constrained minimum of the criterion function is

$$\begin{aligned} Q(\tilde{\boldsymbol{\beta}}_{\mathrm{FGMM}}, \boldsymbol{y}) &= (\boldsymbol{y} - \boldsymbol{X}_1\tilde{\boldsymbol{\beta}}_1)^\top \boldsymbol{W}(\boldsymbol{W}^\top\hat{\boldsymbol{\Omega}}\boldsymbol{W})^{-1}\boldsymbol{W}^\top(\boldsymbol{y} - \boldsymbol{X}_1\tilde{\boldsymbol{\beta}}_1) \\ &= \boldsymbol{u}^\top\boldsymbol{\Psi}(\boldsymbol{P}_{\boldsymbol{A}} - \boldsymbol{P}_{\boldsymbol{P}_{\boldsymbol{A}}\boldsymbol{\Psi}^\top\boldsymbol{X}_1})\boldsymbol{\Psi}^\top\boldsymbol{u}. \end{aligned} \tag{2.51}$$

If we subtract (2.48) from (2.51), we find that the difference between the constrained and unconstrained minima of the criterion function is

$$Q(\tilde{\boldsymbol{\beta}}_{\mathrm{FGMM}}, \boldsymbol{y}) - Q(\hat{\boldsymbol{\beta}}_{\mathrm{FGMM}}, \boldsymbol{y}) = \boldsymbol{u}^\top\boldsymbol{\Psi}(\boldsymbol{P}_{\boldsymbol{P}_{\boldsymbol{A}}\boldsymbol{\Psi}^\top\boldsymbol{X}} - \boldsymbol{P}_{\boldsymbol{P}_{\boldsymbol{A}}\boldsymbol{\Psi}^\top\boldsymbol{X}_1})\boldsymbol{\Psi}^\top\boldsymbol{u}. \tag{2.52}$$

Since $\mathcal{S}(\boldsymbol{P}_{\boldsymbol{A}}\boldsymbol{\Psi}^\top\boldsymbol{X}_1) \subseteq \mathcal{S}(\boldsymbol{P}_{\boldsymbol{A}}\boldsymbol{\Psi}^\top\boldsymbol{X})$, we see that $\boldsymbol{P}_{\boldsymbol{P}_{\boldsymbol{A}}\boldsymbol{\Psi}^\top\boldsymbol{X}} - \boldsymbol{P}_{\boldsymbol{P}_{\boldsymbol{A}}\boldsymbol{\Psi}^\top\boldsymbol{X}_1}$ is an orthogonal projection matrix of which the image is of dimension $k - k_1 = k_2$. Once again, the result of Exercise 2.2 shows that the test statistic (2.52) is asymptotically distributed as $\chi^2(k_2)$ if the null hypothesis that $\boldsymbol{\beta}_2 = \boldsymbol{0}$ is true. This result continues to hold if the restrictions are nonlinear, as we will see in Section 2.5.

The result that the statistic $Q(\tilde{\boldsymbol{\beta}}_{\mathrm{FGMM}}, \boldsymbol{y}) - Q(\hat{\boldsymbol{\beta}}_{\mathrm{FGMM}}, \boldsymbol{y})$ is asymptotically distributed as $\chi^2(k_2)$ depends on two critical features of the construction of the statistic. The first is that the same matrix of instruments $\boldsymbol{W}$ is used for estimating both the restricted and unrestricted models. This was also required in Part 1, Section 8.5, when we discussed testing restrictions on linear regression models estimated by generalized IV. The second essential feature is that the same weighting matrix $(\boldsymbol{W}^\top\hat{\boldsymbol{\Omega}}\boldsymbol{W})^{-1}$ is used when estimating both models. If, as is usually the case, this matrix has to be estimated, it is important that the *same* estimate is used in both criterion functions. If different instruments or different weighting matrices are used for the two models, (2.52) is no longer in general asymptotically distributed as $\chi^2(k_2)$.

One interesting consequence of the form of (2.52) is that we do not always need to bother estimating the unrestricted model. The test statistic (2.52) must always be less than the constrained minimum $Q(\tilde{\boldsymbol{\beta}}_{\mathrm{FGMM}}, \boldsymbol{y})$. Therefore, if $Q(\tilde{\boldsymbol{\beta}}_{\mathrm{FGMM}}, \boldsymbol{y})$ is less than the critical value for the $\chi^2(k_2)$ distribution at our chosen significance level, we can be sure that the actual test statistic is even smaller and would not lead us to reject the null.

The result that tests of restrictions may be based on the difference between the constrained and unconstrained minima of the GMM criterion function

holds only for efficient GMM estimation. It is not true for nonoptimal criterion functions like (2.12), which do not use an estimate of the inverse of the covariance matrix of the sample moments as a weighting matrix. When the GMM estimates minimize a nonoptimal criterion function, the easiest way to test restrictions is probably to use a Wald test; see Section 1.7 and Part 1, Section 8.5. However, we do not recommend performing inference on the basis of nonoptimal GMM estimation.

2.5 GMM Estimators for Nonlinear Models

The principles underlying GMM estimation of nonlinear models are the same as those we have developed for GMM estimation of linear regression models. For every result that we have discussed in the previous three sections, there is an analogous result for nonlinear models. In order to develop these results, we will take a somewhat more general and abstract approach than we have done up to this point. This approach, which is based on the theory of **estimating functions**, was originally developed by Godambe (1960) and Durbin (1960); see also Godambe and Thompson (1978).

The method of estimating functions employs the concept of an **elementary zero function**. Such a function plays the same role as a residual in the estimation of a regression model. It depends on observed variables, at least one of which must be endogenous, and on a k–vector of parameters, $\boldsymbol{\theta}$. As with a residual, the expectation of an elementary zero function must vanish if it is evaluated at the true value of $\boldsymbol{\theta}$, but not in general otherwise.

We let $f_t(\boldsymbol{\theta}, y_t)$ denote an elementary zero function for observation t. It is called "elementary" because it applies to a single observation. In the linear regression case that we have been studying up to this point, $\boldsymbol{\theta}$ would be replaced by $\boldsymbol{\beta}$ and we would have $f_t(\boldsymbol{\beta}, y_t) \equiv y_t - \boldsymbol{X}_t\boldsymbol{\beta}$. In general, we may well have more than one elementary zero function for each observation.

We consider a model $\mathbb{M}$, which, as usual, is to be thought of as a set of DGPs. To each DGP in $\mathbb{M}$, there corresponds a unique value of $\boldsymbol{\theta}$, which is what we often call the "true" value of $\boldsymbol{\theta}$ for that DGP. It is important to note that the uniqueness goes just one way here: A given parameter vector $\boldsymbol{\theta}$ may correspond to many DGPs, perhaps even to an infinite number of them, but each DGP corresponds to just one parameter vector. In order to express the key property of elementary zero functions, we must introduce a symbol for the DGPs of the model $\mathbb{M}$. It is conventional to use the Greek letter μ for this purpose, but then it is necessary to avoid confusion with the conventional use of μ to denote a population mean. It is usually not difficult to distinguish the two uses of the symbol.

The key property of elementary zero functions can now be written as

$$\mathrm{E}_\mu\big(f_t(\boldsymbol{\theta}_\mu, y_t)\big) = 0, \tag{2.53}$$

where $\mathrm{E}_\mu(\cdot)$ denotes the expectation under the DGP μ, and $\boldsymbol{\theta}_\mu$ is the (unique) parameter vector associated with μ. It is assumed that property (2.53) holds for all t and for all $\mu \in \mathbb{M}$.

If estimation based on elementary zero functions is to be possible, these functions must satisfy a number of conditions in addition to condition (2.53). Most importantly, we need to ensure that the model is asymptotically identified. We therefore assume that, for some observations, at least,

$$\mathrm{E}_\mu\big(f_t(\boldsymbol{\theta}, y_t)\big) \neq 0 \quad \text{for all } \boldsymbol{\theta} \neq \boldsymbol{\theta}_\mu. \tag{2.54}$$

This just says that, if we evaluate f_t at a $\boldsymbol{\theta}$ that is different from the $\boldsymbol{\theta}_\mu$ that corresponds to the DGP under which we take expectations, then the expectation of $f_t(\boldsymbol{\theta}, y_t)$ must be nonzero. Condition (2.54) does not have to hold for every observation, but it must hold for a fraction of the observations that does not tend to zero as $n \to \infty$.

In the case of the linear regression model, if we write $\boldsymbol{\beta}_0$ for the true parameter vector, condition (2.54) is satisfied for observation t if, for all $\boldsymbol{\beta} \neq \boldsymbol{\beta}_0$,

$$\mathrm{E}(y_t - \boldsymbol{X}_t\boldsymbol{\beta}) = \mathrm{E}\big(\boldsymbol{X}_t(\boldsymbol{\beta}_0 - \boldsymbol{\beta}) + u_t\big) = \mathrm{E}\big(\boldsymbol{X}_t(\boldsymbol{\beta}_0 - \boldsymbol{\beta})\big) \neq 0. \tag{2.55}$$

It is clear from (2.55) that condition (2.54) must be satisfied whenever the fitted values actually depend on all the components of the vector $\boldsymbol{\beta}$ for at least some fraction of the observations. This is equivalent to the more familiar condition that

$$\boldsymbol{S}_{\boldsymbol{X}^\top\boldsymbol{X}} \equiv \operatorname*{plim}_{n\to\infty} \frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X}$$

is a positive definite matrix; see Section 1.2 for a discussion of similar asymptotic identification conditions.

We also need to make some assumption about the variances and covariances of the elementary zero functions. If there is just one elementary zero function per observation, we let $\boldsymbol{f}(\boldsymbol{\theta}, \boldsymbol{y})$ denote the n–vector with typical element $f_t(\boldsymbol{\theta}, y_t)$. If there are $m > 1$ elementary zero functions per observation, then we can group all of them into a vector $\boldsymbol{f}(\boldsymbol{\theta}, \boldsymbol{y})$ with nm elements. In either event, we then assume that

$$\mathrm{E}\big(\boldsymbol{f}(\boldsymbol{\theta}, \boldsymbol{y})\boldsymbol{f}^\top(\boldsymbol{\theta}, \boldsymbol{y})\big) = \boldsymbol{\Omega}, \tag{2.56}$$

where $\boldsymbol{\Omega}$, which implicitly depends on μ, is a finite, positive definite matrix. Thus we are assuming that, under every DGP $\mu \in \mathbb{M}$, each of the f_t has a finite variance and a finite covariance with every f_s for $s \neq t$.

Estimating Functions and Estimating Equations

The method of estimating functions replaces relationships like (2.53) that hold in expectation with their empirical, or sample, counterparts. Because $\boldsymbol{\theta}$ is a k–vector, we need k **estimating functions** in order to estimate it. In general,

these are weighted averages of the elementary zero functions. Equating the estimating functions to zero yields k **estimating equations**, which must be solved in order to obtain the GMM estimator.

As for the linear regression model, the estimating equations are, in fact, just sample moment conditions which, in most cases, are based on instrumental variables. There are generally more instruments than parameters, and so we need to form linear combinations of the instruments in order to construct precisely k estimating equations. Let $\boldsymbol{W}$ be an $n \times l$ matrix of instruments, which are assumed to be predetermined. Usually, one column of $\boldsymbol{W}$ is a vector of 1s. Now define $\boldsymbol{Z} \equiv \boldsymbol{W}\boldsymbol{J}$, where $\boldsymbol{J}$ is an $l \times k$ matrix with full column rank k. Later, we will discuss how $\boldsymbol{J}$, and hence $\boldsymbol{Z}$, should optimally be chosen, but, for the moment, we take $\boldsymbol{Z}$ as given.

If $\boldsymbol{\theta}_\mu$ is the parameter vector for the DGP μ under which we take expectations, the theoretical moment conditions are

$$\mathrm{E}\big(\boldsymbol{Z}_t^\top f_t(\boldsymbol{\theta}_\mu, y_t)\big) = \boldsymbol{0}, \tag{2.57}$$

where $\boldsymbol{Z}_t$ is the t^{th} row of $\boldsymbol{Z}$. Later on, when we take explicit account of the covariance matrix $\boldsymbol{\Omega}$ in formulating the estimating equations, we will need to modify these conditions so that they take the form of conditions (2.31), but (2.57) is all that is required at this stage. In fact, even (2.57) is stronger than we really need. It is sufficient to assume that $\boldsymbol{Z}_t$ and $f_t(\boldsymbol{\theta})$ are asymptotically uncorrelated, which, together with some regularity conditions, implies that

$$\operatorname*{plim}_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} \boldsymbol{Z}_t^\top f_t(\boldsymbol{\theta}_\mu, y_t) = \boldsymbol{0}. \tag{2.58}$$

The vector of estimating functions that corresponds to (2.57) or (2.58) is the k–vector $n^{-1}\boldsymbol{Z}^\top\boldsymbol{f}(\boldsymbol{\theta}, \boldsymbol{y})$. Equating this vector to zero yields the system of estimating equations

$$\frac{1}{n}\boldsymbol{Z}^\top\boldsymbol{f}(\boldsymbol{\theta}, \boldsymbol{y}) = \boldsymbol{0}, \tag{2.59}$$

and solving this system yields $\hat{\boldsymbol{\theta}}$, the **nonlinear GMM estimator**.

Consistency

If we are to prove that the nonlinear GMM estimator is consistent, we must assume that a law of large numbers applies to the vector $n^{-1}\boldsymbol{Z}^\top\boldsymbol{f}(\boldsymbol{\theta}, \boldsymbol{y})$. This allows us to define the k–vector of **limiting estimating functions**,

$$\boldsymbol{\alpha}(\boldsymbol{\theta}; \mu) \equiv \operatorname*{plim}_{n\to\infty,\,\mu} \frac{1}{n}\boldsymbol{Z}^\top\boldsymbol{f}(\boldsymbol{\theta}, \boldsymbol{y}). \tag{2.60}$$

In words, $\boldsymbol{\alpha}(\boldsymbol{\theta}; \mu)$ is the probability limit, under the DGP μ, of the vector of estimating functions. Setting $\boldsymbol{\alpha}(\boldsymbol{\theta}; \mu)$ to $\boldsymbol{0}$ yields a set of **limiting estimating equations**.

Either (2.57) or the weaker condition (2.58) implies that $\boldsymbol{\alpha}(\boldsymbol{\theta}_\mu; \mu) = \mathbf{0}$ for all $\mu \in \mathbb{M}$. We then need an asymptotic identification condition strong enough to ensure that $\boldsymbol{\alpha}(\boldsymbol{\theta}; \mu) \neq \mathbf{0}$ for all $\boldsymbol{\theta} \neq \boldsymbol{\theta}_\mu$. In other words, we require that the vector $\boldsymbol{\theta}_\mu$ must be the unique solution to the system of limiting estimating equations. If we assume that such a condition holds, it is straightforward to prove consistency in the nonrigorous way we used in Section 1.2 and Part 1, Section 8.3. Evaluating equations (2.59) at their solution $\hat{\boldsymbol{\theta}}$, we find that

$$\frac{1}{n} \boldsymbol{Z}^\top \boldsymbol{f}(\hat{\boldsymbol{\theta}}, \boldsymbol{y}) = \mathbf{0}. \tag{2.61}$$

As $n \to \infty$, the left-hand side of this system of equations tends under μ to the vector $\boldsymbol{\alpha}(\operatorname{plim}_\mu \hat{\boldsymbol{\theta}}; \mu)$, and the right-hand side remains a zero vector. Given the asymptotic identification condition, the equality in (2.61) can hold asymptotically only if

$$\operatorname*{plim}_{n\to\infty, \mu} \hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_\mu.$$

Therefore, we conclude that the nonlinear GMM estimator $\hat{\boldsymbol{\theta}}$, which solves the system of estimating equations (2.59), consistently estimates the parameter vector $\boldsymbol{\theta}_\mu$, for all $\mu \in \mathbb{M}$, provided the asymptotic identification condition is satisfied.

Asymptotic Normality

For ease of notation, we now fix the DGP $\mu \in \mathbb{M}$ and write $\boldsymbol{\theta}_\mu = \boldsymbol{\theta}_0$. Thus $\boldsymbol{\theta}_0$ has its usual interpretation as the "true" parameter vector. In addition, we suppress the explicit mention of the data vector $\boldsymbol{y}$. As usual, the proof that $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ is asymptotically normally distributed is based on a Taylor series approximation, a law of large numbers, and a central limit theorem. For the purposes of the first of these, we need to assume that the zero functions f_t are continuously differentiable in the neighborhood of $\boldsymbol{\theta}_0$. If we perform a first-order Taylor expansion of $n^{1/2}$ times (2.59) around $\boldsymbol{\theta}_0$ and introduce some appropriate factors of powers of n, we obtain the result that

$$n^{-1/2} \boldsymbol{Z}^\top \boldsymbol{f}(\boldsymbol{\theta}_0) + n^{-1} \boldsymbol{Z}^\top \boldsymbol{F}(\bar{\boldsymbol{\theta}}) n^{1/2} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = \mathbf{0}, \tag{2.62}$$

where the $n \times k$ matrix $\boldsymbol{F}(\boldsymbol{\theta})$ has typical element

$$F_{ti}(\boldsymbol{\theta}) \equiv \frac{\partial f_t(\boldsymbol{\theta})}{\partial \theta_i}, \tag{2.63}$$

where θ_i is the i^{th} element of $\boldsymbol{\theta}$. This matrix, like $\boldsymbol{f}(\boldsymbol{\theta})$ itself, depends implicitly on the vector $\boldsymbol{y}$ and is therefore stochastic. The notation $\boldsymbol{F}(\bar{\boldsymbol{\theta}})$ in (2.62) is the convenient shorthand we introduced in Section 1.2: Row t of the matrix is the corresponding row of $\boldsymbol{F}(\boldsymbol{\theta})$ evaluated at $\boldsymbol{\theta} = \bar{\boldsymbol{\theta}}_t$, where the $\bar{\boldsymbol{\theta}}_t$ all satisfy the inequality

$$\|\bar{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_0\| \leq \|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|.$$

The consistency of $\hat{\boldsymbol{\theta}}$ then implies that the $\bar{\boldsymbol{\theta}}_t$ also tend to $\boldsymbol{\theta}_0$ as $n \to \infty$. The consistency of the $\bar{\boldsymbol{\theta}}_t$ implies that

$$\operatorname*{plim}_{n\to\infty} \frac{1}{n} \boldsymbol{Z}^\top \boldsymbol{F}(\bar{\boldsymbol{\theta}}) = \operatorname*{plim}_{n\to\infty} \frac{1}{n} \boldsymbol{Z}^\top \boldsymbol{F}(\boldsymbol{\theta}_0). \tag{2.64}$$

Under reasonable regularity conditions, we can apply a law of large numbers to the right-hand side of (2.64), and the probability limit is then deterministic. For asymptotic normality, we also require that it should be nonsingular. This is a condition of **strong asymptotic identification**, of the sort used in Section 1.2. By a first-order Taylor expansion of $\boldsymbol{\alpha}(\boldsymbol{\theta}; \mu)$ around $\boldsymbol{\theta}_0$, where it is equal to $\mathbf{0}$, we see from the definition (2.60) that

$$\boldsymbol{\alpha}(\boldsymbol{\theta}; \mu) \stackrel{a}{=} \operatorname*{plim}_{n\to\infty} \frac{1}{n} \boldsymbol{Z}^\top \boldsymbol{F}(\boldsymbol{\theta}_0)(\boldsymbol{\theta} - \boldsymbol{\theta}_0). \tag{2.65}$$

Therefore, the condition that the right-hand side of (2.64) is nonsingular is a strengthening of the condition that $\boldsymbol{\theta}$ is asymptotically identified. Because it is nonsingular, the system of equations

$$\operatorname*{plim}_{n\to\infty} \frac{1}{n} \boldsymbol{Z}^\top \boldsymbol{F}(\boldsymbol{\theta}_0)(\boldsymbol{\theta} - \boldsymbol{\theta}_0) = \mathbf{0}$$

has no solution other than $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. By (2.65), this implies that $\boldsymbol{\alpha}(\boldsymbol{\theta}; \mu) \neq \mathbf{0}$ for all $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$, which is the asymptotic identification condition.

Applying the results just discussed to equation (2.62), we find that

$$n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \stackrel{a}{=} -\Big(\operatorname*{plim}_{n\to\infty} \frac{1}{n} \boldsymbol{Z}^\top \boldsymbol{F}(\boldsymbol{\theta}_0)\Big)^{-1} n^{-1/2} \boldsymbol{Z}^\top \boldsymbol{f}(\boldsymbol{\theta}_0). \tag{2.66}$$

Next, we apply a central limit theorem to the second factor on the right-hand side of (2.66). Doing so demonstrates that $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ is asymptotically normally distributed. By (2.57), the vector $n^{-1/2} \boldsymbol{Z}^\top \boldsymbol{f}(\boldsymbol{\theta}_0)$ must have mean $\mathbf{0}$, and, by (2.56), its covariance matrix is $\operatorname{plim} n^{-1} \boldsymbol{Z}^\top \boldsymbol{\Omega} \boldsymbol{Z}$. In stating this result, we assume that (2.02) holds with the $\boldsymbol{f}(\boldsymbol{\theta}_0)$ in place of the disturbances. Then (2.66) implies that the vector $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ is asymptotically normally distributed with mean vector $\mathbf{0}$ and covariance matrix

$$\Big(\operatorname*{plim}_{n\to\infty} \frac{1}{n} \boldsymbol{Z}^\top \boldsymbol{F}(\boldsymbol{\theta}_0)\Big)^{-1} \Big(\operatorname*{plim}_{n\to\infty} \frac{1}{n} \boldsymbol{Z}^\top \boldsymbol{\Omega} \boldsymbol{Z}\Big) \Big(\operatorname*{plim}_{n\to\infty} \frac{1}{n} \boldsymbol{F}^\top(\boldsymbol{\theta}_0) \boldsymbol{Z}\Big)^{-1}. \tag{2.67}$$

Since this is a sandwich covariance matrix, it is evident that the nonlinear GMM estimator $\hat{\boldsymbol{\theta}}$ is not, in general, an asymptotically efficient estimator.

Asymptotically Efficient Estimation

In order to obtain an asymptotically efficient nonlinear GMM estimator, we need to choose the estimating functions $n^{-1} \boldsymbol{Z}^\top \boldsymbol{f}(\boldsymbol{\theta})$ optimally. This is equivalent to choosing $\boldsymbol{Z}$ optimally. How we should do this depends on what

assumptions we make about $\boldsymbol{F}(\boldsymbol{\theta})$ and $\boldsymbol{\Omega}$, the covariance matrix of $\boldsymbol{f}(\boldsymbol{\theta})$. Not surprisingly, we will obtain results very similar to the results for linear GMM estimation obtained in Section 2.2.

We begin with the simplest possible case, in which $\boldsymbol{\Omega} = \sigma^2\mathbf{I}$, and $\boldsymbol{F}(\boldsymbol{\theta}_0)$ is predetermined in the sense that

$$\mathrm{E}\big(\boldsymbol{F}_t(\boldsymbol{\theta}_0) f_t(\boldsymbol{\theta}_0)\big) = \mathbf{0}, \tag{2.68}$$

where $\boldsymbol{F}_t(\boldsymbol{\theta}_0)$ is the t^{th} row of $\boldsymbol{F}(\boldsymbol{\theta}_0)$. If we ignore the probability limits and the factors of n^{-1}, the sandwich covariance matrix (2.67) is in this case proportional to

$$(\boldsymbol{Z}^\top\boldsymbol{F}_0)^{-1}\boldsymbol{Z}^\top\boldsymbol{Z}(\boldsymbol{F}_0^\top\boldsymbol{Z})^{-1}, \tag{2.69}$$

where, for ease of notation, $\boldsymbol{F}_0 \equiv \boldsymbol{F}(\boldsymbol{\theta}_0)$. The inverse of (2.69), which is proportional to the asymptotic precision matrix of the estimator, is

$$\boldsymbol{F}_0^\top\boldsymbol{Z}(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\boldsymbol{Z}^\top\boldsymbol{F}_0 = \boldsymbol{F}_0^\top\boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{F}_0. \tag{2.70}$$

If we set $\boldsymbol{Z} = \boldsymbol{F}_0$, (2.69) is no longer a sandwich, and (2.70) simplifies to $\boldsymbol{F}_0^\top\boldsymbol{F}_0$. The difference between $\boldsymbol{F}_0^\top\boldsymbol{F}_0$ and the general expression (2.70) is

$$\boldsymbol{F}_0^\top\boldsymbol{F}_0 - \boldsymbol{F}_0^\top\boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{F}_0 = \boldsymbol{F}_0^\top\boldsymbol{M}_{\boldsymbol{Z}}\boldsymbol{F}_0,$$

which is a positive semidefinite matrix because $\boldsymbol{M}_{\boldsymbol{Z}} \equiv \mathbf{I} - \boldsymbol{P}_{\boldsymbol{Z}}$ is an orthogonal projection matrix. Thus, in this simple case, the optimal instrument matrix is just $\boldsymbol{F}_0$.

Since we do not know $\boldsymbol{\theta}_0$, it is not feasible to use $\boldsymbol{F}_0$ directly as the matrix of instruments. Instead, we use the trick that leads to the estimating equations (1.28) which define the NLS estimator. This leads us to solve the estimating equations

$$\frac{1}{n}\boldsymbol{F}^\top(\boldsymbol{\theta})\boldsymbol{f}(\boldsymbol{\theta}) = \mathbf{0}. \tag{2.71}$$

If $\boldsymbol{\Omega} = \sigma^2\mathbf{I}$, and $\boldsymbol{F}(\boldsymbol{\theta}_0)$ is predetermined, solving these equations yields an asymptotically efficient GMM estimator.

It is not valid to use the columns of $\boldsymbol{F}(\boldsymbol{\theta})$ as instruments if condition (2.68) is not satisfied. In that event, the analysis of Part 1, Section 8.3, taken up again in Section 2.2, suggests that we should replace the rows of $\boldsymbol{F}_0$ by their expectations conditional on the information sets Ω_t generated by variables that are exogenous or predetermined for observation t. Let us define an $n \times k$ matrix $\bar{\boldsymbol{F}}$, in terms of its typical row $\bar{\boldsymbol{F}}_t$, and another $n \times k$ matrix $\boldsymbol{V}$, as follows:

$$\bar{\boldsymbol{F}}_t \equiv \mathrm{E}\big(\boldsymbol{F}_t(\boldsymbol{\theta}_0) \,|\, \Omega_t\big) \quad \text{and} \quad \boldsymbol{V} \equiv \boldsymbol{F}_0 - \bar{\boldsymbol{F}}. \tag{2.72}$$

The matrices $\bar{\boldsymbol{F}}$ and $\boldsymbol{V}$ are entirely analogous to the matrices $\bar{\boldsymbol{X}}$ and $\boldsymbol{V}$ used in Part 1, Section 8.3. The definitions (2.72) imply that

$$\operatorname*{plim}_{n\to\infty}\frac{1}{n}\bar{\boldsymbol{F}}^\top\boldsymbol{F}_0 = \operatorname*{plim}_{n\to\infty}\frac{1}{n}\bar{\boldsymbol{F}}^\top(\bar{\boldsymbol{F}} + \boldsymbol{V}) = \operatorname*{plim}_{n\to\infty}\frac{1}{n}\bar{\boldsymbol{F}}^\top\bar{\boldsymbol{F}}. \tag{2.73}$$

The term $\operatorname{plim} n^{-1}\bar{\boldsymbol{F}}^\top\boldsymbol{V}$ equals $\mathbf{O}$ because (2.72) implies that $\mathrm{E}(\boldsymbol{V}_t \,|\, \Omega_t) = \mathbf{0}$, and the conditional expectation $\bar{\boldsymbol{F}}_t$ belongs to the information set Ω_t.

To find the asymptotic covariance matrix of $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ when $\bar{\boldsymbol{F}}$ is used in place of $\boldsymbol{Z}$ and the covariance matrix of $\boldsymbol{f}(\boldsymbol{\theta})$ is $\sigma^2\mathbf{I}$, we start from expression (2.67). Using (2.73), we obtain

$$\begin{aligned}
&\sigma^2\Big(\operatorname*{plim}_{n\to\infty}\frac{1}{n}\bar{\boldsymbol{F}}^\top\boldsymbol{F}_0\Big)^{-1}\Big(\operatorname*{plim}_{n\to\infty}\frac{1}{n}\bar{\boldsymbol{F}}^\top\bar{\boldsymbol{F}}\Big)\Big(\operatorname*{plim}_{n\to\infty}\frac{1}{n}\boldsymbol{F}_0^\top\bar{\boldsymbol{F}}\Big)^{-1} \\
&= \sigma^2\Big(\operatorname*{plim}_{n\to\infty}\frac{1}{n}\bar{\boldsymbol{F}}^\top\bar{\boldsymbol{F}}\Big)^{-1}.
\end{aligned} \tag{2.74}$$

For any other choice of instrument matrix $\boldsymbol{Z}$, the argument giving (2.73) shows that $\operatorname{plim} n^{-1}\boldsymbol{Z}^\top\boldsymbol{F}_0 = \operatorname{plim} n^{-1}\boldsymbol{Z}^\top\bar{\boldsymbol{F}}$, and so the covariance matrix (2.67) becomes

$$\sigma^2\Big(\operatorname*{plim}_{n\to\infty}\frac{1}{n}\boldsymbol{Z}^\top\bar{\boldsymbol{F}}\Big)^{-1}\Big(\operatorname*{plim}_{n\to\infty}\frac{1}{n}\boldsymbol{Z}^\top\boldsymbol{Z}\Big)\Big(\operatorname*{plim}_{n\to\infty}\frac{1}{n}\bar{\boldsymbol{F}}^\top\boldsymbol{Z}\Big)^{-1}. \tag{2.75}$$

The inverse of (2.75) is $1/\sigma^2$ times the probability limit of

$$\frac{1}{n}\bar{\boldsymbol{F}}^\top\boldsymbol{Z}(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\boldsymbol{Z}^\top\bar{\boldsymbol{F}} = \frac{1}{n}\bar{\boldsymbol{F}}^\top\boldsymbol{P}_{\boldsymbol{Z}}\bar{\boldsymbol{F}}. \tag{2.76}$$

This expression is analogous to expression (F8.22) for the asymptotic precision of the IV estimator for linear regression models with endogenous explanatory variables. Since the difference between $n^{-1}\bar{\boldsymbol{F}}^\top\bar{\boldsymbol{F}}$ and (2.76) is the positive semidefinite matrix $n^{-1}\bar{\boldsymbol{F}}^\top\boldsymbol{M}_{\boldsymbol{Z}}\bar{\boldsymbol{F}}$, we conclude that (2.74) is indeed the asymptotic covariance matrix that corresponds to the optimal choice of $\boldsymbol{Z}$. Therefore, when $\boldsymbol{F}_t(\boldsymbol{\theta})$ is not predetermined, we should use its expectation conditional on Ω_t in the matrix of instruments.

In practice, of course, the matrix $\bar{\boldsymbol{F}}$ is rarely observed. We therefore need to estimate it. The natural way to do so is to regress $\boldsymbol{F}(\boldsymbol{\theta})$ on an $n \times l$ matrix of instruments $\boldsymbol{W}$, where $l \geq k$, with the inequality holding strictly in most cases. This yields fitted values $\boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{F}(\boldsymbol{\theta})$. If we estimate $\bar{\boldsymbol{F}}$ in this way, the optimal estimating equations become

$$\frac{1}{n}\boldsymbol{F}^\top(\boldsymbol{\theta})\boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{f}(\boldsymbol{\theta}) = \mathbf{0}. \tag{2.77}$$

By reasoning like that which led to (F8.28) and (2.73), it can be seen that these estimating equations are asymptotically equivalent to the same equations with $\bar{\boldsymbol{F}}$ in place of $\boldsymbol{F}(\boldsymbol{\theta})$. In particular, if $\mathcal{S}(\bar{\boldsymbol{F}}) \subseteq \mathcal{S}(\boldsymbol{W})$, the estimator obtained by solving (2.77) is asymptotically equivalent to the one obtained using the optimal instruments $\bar{\boldsymbol{F}}$.

The estimating equations (2.77) generalize the first-order conditions (F8.29) for linear IV estimation. As readers are asked to show in Exercise 2.14, the solution to (2.77) in the case of the linear regression model is simply the

generalized IV estimator (F8.30). As can be seen from (2.67), the asymptotic covariance matrix of the estimator $\hat{\boldsymbol{\theta}}$ defined by (2.77) can be estimated by

$$\hat{\sigma}^2(\hat{\boldsymbol{F}}^\top \boldsymbol{P}_{\boldsymbol{W}} \hat{\boldsymbol{F}})^{-1},$$

where $\hat{\boldsymbol{F}} \equiv \boldsymbol{F}(\hat{\boldsymbol{\theta}})$, and $\hat{\sigma}^2 \equiv n^{-1} \sum_{t=1}^n f_t^2(\hat{\boldsymbol{\theta}})$, the average of the squares of the elementary zero functions evaluated at $\hat{\boldsymbol{\theta}}$, is a natural estimator of σ^2.

Efficient Estimation with an Unknown Covariance Matrix

When the covariance matrix $\boldsymbol{\Omega}$ is unknown, the GMM estimators defined by the estimating equations (2.71) or (2.77), according to whether or not $\boldsymbol{F}(\boldsymbol{\theta})$ is predetermined, are no longer asymptotically efficient in general. But, just as we did in Section 2.3 with regression models, we can obtain estimates that are efficient for a given set of instruments by using a heteroskedasticity-consistent or a HAC estimator.

Suppose there are $l > k$ instruments which form an $n \times l$ matrix $\boldsymbol{W}$. As in Section 2.2, we can construct estimating equations with instruments $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{J}$, using a full-rank $l \times k$ matrix $\boldsymbol{J}$ to select k linear combinations of the full set of instruments. The asymptotic covariance matrix of the estimator obtained by solving these equations is then, by (2.67),

$$\Big(\operatorname*{plim}_{n\to\infty} \frac{1}{n} \boldsymbol{J}^\top \boldsymbol{W}^\top \boldsymbol{F}_0\Big)^{-1} \Big(\operatorname*{plim}_{n\to\infty} \frac{1}{n} \boldsymbol{J}^\top \boldsymbol{W}^\top \boldsymbol{\Omega} \boldsymbol{W} \boldsymbol{J}\Big) \Big(\operatorname*{plim}_{n\to\infty} \frac{1}{n} \boldsymbol{F}_0^\top \boldsymbol{W} \boldsymbol{J}\Big)^{-1}. \tag{2.78}$$

This looks just like (2.07) with $\boldsymbol{F}_0$ in place of the regressor matrix $\boldsymbol{X}$. The optimal choice of $\boldsymbol{J}$ is therefore just (2.08) with $\boldsymbol{F}_0$ in place of $\boldsymbol{X}$. Since (2.08) depends on the unknown true $\boldsymbol{\Omega}$, we replace $n^{-1}\boldsymbol{W}^\top \boldsymbol{\Omega} \boldsymbol{W}$ by an estimator $\hat{\boldsymbol{\Sigma}}$, which could be either a heteroskedasticity-consistent or a HAC estimator. This yields the estimating equations

$$\boldsymbol{F}^\top(\boldsymbol{\theta}) \boldsymbol{W} \hat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{W}^\top \boldsymbol{f}(\boldsymbol{\theta}) = \boldsymbol{0}, \tag{2.79}$$

and the asymptotic covariance matrix (2.78) simplifies to

$$\Big(\operatorname*{plim}_{n\to\infty} n^{-2} \boldsymbol{F}_0^\top \boldsymbol{W} \hat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{W}^\top \boldsymbol{F}_0\Big)^{-1}, \tag{2.80}$$

in which, if $\boldsymbol{F}(\boldsymbol{\theta})$ is not predetermined, we may write $\bar{\boldsymbol{F}}$ instead of $\boldsymbol{F}_0$ without changing the limit. In practice, we can use

$$\widehat{\operatorname{Var}}(\hat{\boldsymbol{\theta}}) = n(\hat{\boldsymbol{F}}^\top \boldsymbol{W} \hat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{W}^\top \hat{\boldsymbol{F}})^{-1}, \tag{2.81}$$

where $\hat{\boldsymbol{F}} \equiv \boldsymbol{F}(\hat{\boldsymbol{\theta}})$, to estimate the covariance matrix of $\hat{\boldsymbol{\theta}}$. As with the estimator (2.41) for the linear regression case, the factor of n is needed to offset the factor of n^{-1} in $\hat{\boldsymbol{\Sigma}}$. The matrix (2.81) can be used to construct Wald tests and asymptotic confidence intervals in the usual way.

Efficient Estimation with a Known Covariance Matrix

When the covariance matrix $\boldsymbol{\Omega}$ is known, we can obtain a fully efficient GMM estimator. As before, we let $\boldsymbol{\Psi}$ denote an $n \times n$ matrix which satisfies the equation $\boldsymbol{\Omega}^{-1} = \boldsymbol{\Psi}\boldsymbol{\Psi}^\top$. The variance of the vector $\boldsymbol{\Psi}^\top \boldsymbol{f}(\boldsymbol{\theta}_0)$, where $\boldsymbol{\theta}_0$ is the true parameter vector for the DGP that generates the data, is then

$$\mathrm{E}\big(\boldsymbol{\Psi}^\top \boldsymbol{f}(\boldsymbol{\theta}_0) \boldsymbol{f}^\top(\boldsymbol{\theta}_0) \boldsymbol{\Psi}\big) = \boldsymbol{\Psi}^\top \boldsymbol{\Omega} \boldsymbol{\Psi} = \mathbf{I}.$$

Thus the components of the vector $\boldsymbol{\Psi}^\top \boldsymbol{f}(\boldsymbol{\theta})$ form a set of zero functions that are homoskedastic and serially uncorrelated. As we mentioned in Section 2.2, it is often possible to choose $\boldsymbol{\Psi}$ in such a way that these components can be thought of as innovations, and in this case $\boldsymbol{\Psi}$ is usually upper triangular.

The matrix $\boldsymbol{\Psi}$ does not depend on the parameters $\boldsymbol{\theta}$. Therefore, the matrix of derivatives of the transformed zero functions in the vector $\boldsymbol{\Psi}^\top \boldsymbol{f}(\boldsymbol{\theta})$ is just $\boldsymbol{\Psi}^\top \boldsymbol{F}(\boldsymbol{\theta})$. Consequently, if the t^{th} row of $\boldsymbol{\Psi}^\top \boldsymbol{F}(\boldsymbol{\theta})$ is predetermined with respect to the t^{th} component of $\boldsymbol{\Psi}^\top \boldsymbol{f}(\boldsymbol{\theta})$, the optimal estimating equations are constructed using the columns of $\boldsymbol{\Psi}^\top \boldsymbol{F}(\boldsymbol{\theta}_0)$ as instruments. Because $\boldsymbol{\theta}_0$ is not known, the optimal instruments are estimated along with the parameters by using the estimating equations

$$\frac{1}{n} \boldsymbol{F}^\top(\boldsymbol{\theta}) \boldsymbol{\Psi} \boldsymbol{\Psi}^\top \boldsymbol{f}(\boldsymbol{\theta}) = \frac{1}{n} \boldsymbol{F}^\top(\boldsymbol{\theta}) \boldsymbol{\Omega}^{-1} \boldsymbol{f}(\boldsymbol{\theta}) = \boldsymbol{0}, \tag{2.82}$$

as in (2.71). The asymptotic covariance matrix of the resulting estimator is

$$\operatorname*{plim}_{n\to\infty} \Big(\frac{1}{n} \boldsymbol{F}_0^\top \boldsymbol{\Omega}^{-1} \boldsymbol{F}_0\Big)^{-1}, \tag{2.83}$$

where, as usual, $\boldsymbol{F}_0 \equiv \boldsymbol{F}(\boldsymbol{\theta}_0)$. The derivation of (2.83) from (2.67) is quite straightforward; see Exercise 2.15. In practice, the covariance matrix of $\hat{\boldsymbol{\theta}}$ is normally estimated by

$$\widehat{\operatorname{Var}}(\hat{\boldsymbol{\theta}}) = (\hat{\boldsymbol{F}}^\top \boldsymbol{\Omega}^{-1} \hat{\boldsymbol{F}})^{-1}. \tag{2.84}$$

If the t^{th} row of $\boldsymbol{\Psi}^\top \boldsymbol{F}(\boldsymbol{\theta})$ is not predetermined with respect to the t^{th} component of $\boldsymbol{\Psi}^\top \boldsymbol{f}(\boldsymbol{\theta})$, and if this component is an innovation, then we can determine the optimal instruments just as we did in Section 2.2. By analogy with (2.24), we define the matrix $\bar{\boldsymbol{F}}(\boldsymbol{\theta})$ implicitly by the equation

$$\mathrm{E}\big((\boldsymbol{\Psi}^\top \boldsymbol{F}(\boldsymbol{\theta}))_t \,|\, \Omega_t\big) = (\boldsymbol{\Psi}^\top \bar{\boldsymbol{F}}(\boldsymbol{\theta}))_t. \tag{2.85}$$

As in Section 2.2, making this definition explicit depends on the details of the particular model under study. The estimating equations for fully efficient estimation are then given by (2.82) with $\boldsymbol{F}(\boldsymbol{\theta})$ replaced by $\bar{\boldsymbol{F}}(\boldsymbol{\theta})$. The asymptotic covariance matrix is (2.83) with $\boldsymbol{F}_0$ replaced by $\bar{\boldsymbol{F}}_0$, and the covariance matrix of $\hat{\boldsymbol{\theta}}$ can be estimated by (2.84) with $\hat{\boldsymbol{F}}$ replaced by $\bar{\boldsymbol{F}}(\hat{\boldsymbol{\theta}})$. All of these

claims are proved in the same way as were the corresponding ones for linear regressions in Section 2.2.

When the matrix $\bar{\boldsymbol{F}}(\boldsymbol{\theta})$ is not observable, as is frequently the case, we can often find an $n \times l$ matrix of instruments $\boldsymbol{W}$, where usually $l > k$, such that $\boldsymbol{W}$ satisfies the predeterminedness condition in its form (2.30), and such that $\mathcal{S}(\boldsymbol{F}(\boldsymbol{\theta}_0)) \subseteq \mathcal{S}(\boldsymbol{W})$. In such cases, overidentified estimation that makes use of the transformed zero functions $\boldsymbol{\Psi}^\top \boldsymbol{f}(\boldsymbol{\theta})$ and the transformed instruments $\boldsymbol{\Psi}^\top \boldsymbol{W}$ yields asymptotically efficient estimates. The results of Exercises 2.8 and 2.9 can also be readily extended to the present nonlinear case.

Minimizing Criterion Functions

The nonlinear GMM estimators we have discussed in this section can all, like the ones for linear regression models, be obtained by minimizing appropriately chosen quadratic forms. We restrict our attention to cases in which $\operatorname{plim} n^{-1}\boldsymbol{F}^\top(\boldsymbol{\theta})\boldsymbol{f}(\boldsymbol{\theta}) \neq \boldsymbol{0}$, and we employ an $n \times l$ matrix of instruments, $\boldsymbol{W}$. When the covariance matrix $\boldsymbol{\Omega}$ of the elementary zero functions is unknown, but a heteroskedasticity-consistent or HAC estimator $\hat{\boldsymbol{\Sigma}}$ is available, the appropriate GMM criterion function is

$$\frac{1}{n}\boldsymbol{f}^\top(\boldsymbol{\theta})\boldsymbol{W}\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{W}^\top\boldsymbol{f}(\boldsymbol{\theta}). \tag{2.86}$$

Minimizing this function with respect to $\boldsymbol{\theta}$ is equivalent to solving the estimating equations (2.79).

In the case in which the matrix $\boldsymbol{\Omega}$ is known, or can be estimated consistently, the fully efficient estimators of the previous subsection can be obtained by minimizing the quadratic form

$$\boldsymbol{f}^\top(\boldsymbol{\theta})\boldsymbol{\Psi}\boldsymbol{P}_{\boldsymbol{\Psi}^\top\boldsymbol{W}}\boldsymbol{\Psi}^\top\boldsymbol{f}(\boldsymbol{\theta}), \tag{2.87}$$

where $\boldsymbol{\Psi}\boldsymbol{\Psi}^\top = \boldsymbol{\Omega}^{-1}$, the components of $\boldsymbol{\Psi}^\top\boldsymbol{f}(\boldsymbol{\theta}_0)$ are innovations, and the matrix $\boldsymbol{W}$ satisfies the predeterminedness condition in the form (2.30). For full efficiency, the span $\mathcal{S}(\boldsymbol{W})$ of the instruments must (asymptotically) include as a subspace the span of the $\bar{\boldsymbol{F}}(\boldsymbol{\theta}_0)$, as defined in (2.85). In Exercise 2.16, readers are asked to check that minimizing (2.87) is asymptotically equivalent to solving the optimal estimating equations.

Fortunately, we need not treat (2.86) and (2.87) separately. As in Section 2.4, expression (2.86) is asymptotically unchanged if we replace $\hat{\boldsymbol{\Sigma}}$ by $n^{-1}\boldsymbol{W}^\top\boldsymbol{\Omega}\boldsymbol{W}$, where $\boldsymbol{\Omega}$ is the true covariance matrix of the zero functions. Making this replacement, we see that both (2.86) and (2.87) can be written as

$$Q(\boldsymbol{\theta}, \boldsymbol{y}) \equiv \boldsymbol{f}^\top(\boldsymbol{\theta})\boldsymbol{\Psi}\boldsymbol{P}_{\boldsymbol{A}}\boldsymbol{\Psi}^\top\boldsymbol{f}(\boldsymbol{\theta}), \tag{2.88}$$

where $\boldsymbol{A} = \boldsymbol{\Psi}^{-1}\boldsymbol{W}$ and $\boldsymbol{A} = \boldsymbol{\Psi}^\top\boldsymbol{W}$ for the criterion functions (2.86) and (2.87), respectively. Note how closely (2.88) resembles expression (2.45) for the linear regression case.

It is often more convenient to compute GMM estimators by minimizing a criterion function than by directly solving a set of estimating equations. One advantage is that algorithms for minimizing functions tend to be more stable numerically than algorithms for solving sets of nonlinear equations. Another advantage is that the criterion function may have more than one stationary point. In this event, the estimating equations are satisfied at each of these stationary points, although the criterion function may have a unique global minimum, which then corresponds to the solution of interest.

However, the main advantage of working with criterion functions is that the minimized value of a GMM criterion function can be used for testing, as we have already discussed for the linear regression case in Section 2.4. Notice that the factor of n^{-1} in (2.86), which does not matter for estimation, is essential when the criterion function is being used for testing. Its role is to offset the factor of n^{-1} in the definition of $\hat{\boldsymbol{\Sigma}}$.

Tests Based on the GMM Criterion Function

The Hansen-Sargan overidentification test statistic is $Q(\hat{\boldsymbol{\theta}}, \boldsymbol{y})$, the minimized value of the GMM criterion function. Up to an irrelevant scalar factor, the first-order conditions for the minimization of (2.88) are

$$\boldsymbol{F}^\top(\hat{\boldsymbol{\theta}})\boldsymbol{\Psi}\boldsymbol{P}_{\boldsymbol{A}}\boldsymbol{\Psi}^\top\boldsymbol{f}(\hat{\boldsymbol{\theta}}) = \boldsymbol{0}, \tag{2.89}$$

and it follows from this, either by a Taylor expansion or directly by using the result (2.66), that

$$n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \stackrel{a}{=} -\Big(\frac{1}{n}\boldsymbol{F}_0^\top\boldsymbol{\Psi}\boldsymbol{P}_{\boldsymbol{A}}\boldsymbol{\Psi}^\top\boldsymbol{F}_0\Big)^{-1} n^{-1/2}\boldsymbol{F}_0^\top\boldsymbol{\Psi}\boldsymbol{P}_{\boldsymbol{A}}\boldsymbol{\Psi}^\top\boldsymbol{f}_0,$$

where, as usual, $\boldsymbol{F}_0$ and $\boldsymbol{f}_0$ denote $\boldsymbol{F}(\boldsymbol{\theta}_0)$ and $\boldsymbol{f}(\boldsymbol{\theta}_0)$, respectively. We now follow quite closely the calculations of Section 2.4 in order to show that the minimized quadratic form $Q(\hat{\boldsymbol{\theta}}, \boldsymbol{y})$ is asymptotically distributed as $\chi^2(l-k)$. By a short Taylor expansion, we see that

$$\begin{aligned}
\boldsymbol{P}_{\boldsymbol{A}}\boldsymbol{\Psi}^\top\boldsymbol{f}(\hat{\boldsymbol{\theta}}) &\stackrel{a}{=} \boldsymbol{P}_{\boldsymbol{A}}\boldsymbol{\Psi}^\top\boldsymbol{f}_0 + n^{-1/2}\boldsymbol{P}_{\boldsymbol{A}}\boldsymbol{\Psi}^\top\boldsymbol{F}_0\, n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \\
&\stackrel{a}{=} \boldsymbol{P}_{\boldsymbol{A}}\boldsymbol{\Psi}^\top\boldsymbol{f}_0 - n^{-1/2}\boldsymbol{P}_{\boldsymbol{A}}\boldsymbol{\Psi}^\top\boldsymbol{F}_0\big(\tfrac{1}{n}\boldsymbol{F}_0^\top\boldsymbol{\Psi}\boldsymbol{P}_{\boldsymbol{A}}\boldsymbol{\Psi}^\top\boldsymbol{F}_0\big)^{-1} n^{-1/2}\boldsymbol{F}_0^\top\boldsymbol{\Psi}\boldsymbol{P}_{\boldsymbol{A}}\boldsymbol{\Psi}^\top\boldsymbol{f}_0 \\
&= (\mathbf{I} - \boldsymbol{P}_{\boldsymbol{P}_{\boldsymbol{A}}\boldsymbol{\Psi}^\top\boldsymbol{F}_0})\boldsymbol{P}_{\boldsymbol{A}}\boldsymbol{\Psi}^\top\boldsymbol{f}_0,
\end{aligned}$$

where $\boldsymbol{P}_{\boldsymbol{P}_{\boldsymbol{A}}\boldsymbol{\Psi}^\top\boldsymbol{F}_0}$ projects orthogonally on to $\mathcal{S}(\boldsymbol{P}_{\boldsymbol{A}}\boldsymbol{\Psi}^\top\boldsymbol{F}_0)$. Thus $Q(\hat{\boldsymbol{\theta}}, \boldsymbol{y})$, the minimized value of the criterion function (2.88), is

$$\begin{aligned}
\boldsymbol{f}^\top(\hat{\boldsymbol{\theta}})\boldsymbol{\Psi}\boldsymbol{P}_{\boldsymbol{A}}\boldsymbol{\Psi}^\top\boldsymbol{f}(\hat{\boldsymbol{\theta}}) &\stackrel{a}{=} \boldsymbol{f}_0^\top\boldsymbol{\Psi}\boldsymbol{P}_{\boldsymbol{A}}(\mathbf{I} - \boldsymbol{P}_{\boldsymbol{P}_{\boldsymbol{A}}\boldsymbol{\Psi}^\top\boldsymbol{F}_0})\boldsymbol{P}_{\boldsymbol{A}}\boldsymbol{\Psi}^\top\boldsymbol{f}_0 \\
&= \boldsymbol{f}_0^\top\boldsymbol{\Psi}\big(\boldsymbol{P}_{\boldsymbol{A}} - \boldsymbol{P}_{\boldsymbol{P}_{\boldsymbol{A}}\boldsymbol{\Psi}^\top\boldsymbol{F}_0}\big)\boldsymbol{\Psi}^\top\boldsymbol{f}_0.
\end{aligned} \tag{2.90}$$

Because $\mathcal{S}(\boldsymbol{P_A \Psi^\top F}_0) \subseteq \mathcal{S}(\boldsymbol{A})$, the difference of projection matrices in the last expression above is itself an orthogonal projection matrix, of which the image is of dimension $l - k$. As with (2.48), we see that estimating $\boldsymbol{\theta}$ uses up k degrees of freedom. By essentially the same argument as was used for (2.48), it can be shown that (2.90) is asymptotically distributed as $\chi^2(l-k)$. Thus, as expected, $Q(\hat{\boldsymbol{\theta}}, \boldsymbol{y})$ is the Hansen-Sargan test statistic for nonlinear GMM estimation.

As in the case of linear regression models, the difference between the GMM criterion function (2.88) evaluated at restricted estimates and evaluated at unrestricted estimates is asymptotically distributed as $\chi^2(r)$ when there are r equality restrictions. We will not prove this result, which was proved for the linear case in Section 2.3. However, we will present a very simple argument which provides an intuitive explanation.

Let $\tilde{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}$ denote, respectively, the vectors of restricted and unrestricted (feasible) efficient GMM estimates. From the result for the Hansen-Sargan test that was just proved, we know that $Q(\tilde{\boldsymbol{\theta}}, \boldsymbol{y})$ and $Q(\hat{\boldsymbol{\theta}}, \boldsymbol{y})$ are asymptotically distributed as $\chi^2(l-k+r)$ and $\chi^2(l-k)$, respectively. Therefore, since a random variable that follows the $\chi^2(m)$ distribution is equal to the sum of m independent $\chi^2(1)$ variables,

$$Q(\tilde{\boldsymbol{\theta}}, \boldsymbol{y}) \stackrel{a}{=} \sum_{i=1}^{l-k+r} x_i^2 \quad \text{and} \quad Q(\hat{\boldsymbol{\theta}}, \boldsymbol{y}) \stackrel{a}{=} \sum_{i=1}^{l-k} y_i^2, \tag{2.91}$$

where the x_i and y_i are independent, standard normal random variables. Now suppose that the first $l-k$ of the x_i are equal to the corresponding y_i. If so, (2.91) implies that

$$Q(\tilde{\boldsymbol{\theta}}, \boldsymbol{y}) - Q(\hat{\boldsymbol{\theta}}, \boldsymbol{y}) \stackrel{a}{=} \sum_{i=1}^{l-k+r} x_i^2 - \sum_{i=1}^{l-k} x_i^2 = \sum_{i=l-k+1}^{l-k+r} x_i^2. \tag{2.92}$$

Since the leftmost expression here is the test statistic we are interested in and the rightmost expression is evidently distributed as $\chi^2(r)$, we have apparently proved the result. The proof is not complete, of course, because we have not shown that the first $l-k$ of the x_i are, in fact, equal to the corresponding y_i. To prove this, we would need to show that, asymptotically, $Q(\tilde{\boldsymbol{\theta}}, \boldsymbol{y})$ is equal to $Q(\hat{\boldsymbol{\theta}}, \boldsymbol{y})$ plus another random variable independent of $Q(\hat{\boldsymbol{\theta}}, \boldsymbol{y})$. This other random variable would then be equal to the rightmost expression in (2.92).

Nonlinear GMM Estimators: Overview

We have discussed a large number of nonlinear GMM estimators, and it can be confusing to keep track of them all. We therefore conclude this section with a brief summary of the principal cases that are likely to be encountered in applied econometric work.

Case 1. Scalar covariance matrix: $\boldsymbol{\Omega} = \sigma^2 \mathbf{I}$.

When $\operatorname{plim} n^{-1} \boldsymbol{F}^\top(\boldsymbol{\theta}) \boldsymbol{f}(\boldsymbol{\theta}) = \mathbf{0}$, we solve the estimating equations (2.71) to obtain an efficient estimator. This is equivalent to minimizing $\boldsymbol{f}^\top(\boldsymbol{\theta}) \boldsymbol{f}(\boldsymbol{\theta})$. The estimated covariance matrix of $\hat{\boldsymbol{\theta}}$ is

$$\widehat{\operatorname{Var}}(\hat{\boldsymbol{\theta}}) = \hat{\sigma}^2 (\hat{\boldsymbol{F}}^\top \hat{\boldsymbol{F}})^{-1},$$

where $\hat{\sigma}^2$ consistently estimates σ^2. If the model is a nonlinear regression model, then $\hat{\boldsymbol{\theta}}$ is really the nonlinear least-squares estimator discussed in Section 1.3.

When $\operatorname{plim} n^{-1} \boldsymbol{F}^\top(\boldsymbol{\theta}) \boldsymbol{f}(\boldsymbol{\theta}) \neq \mathbf{0}$, we must replace $\boldsymbol{F}(\boldsymbol{\theta})$ by an estimate of its conditional expectation. This means that we solve the estimating equations (2.77), which is equivalent to minimizing $\boldsymbol{f}^\top(\boldsymbol{\theta}) \boldsymbol{P_W} \boldsymbol{f}(\boldsymbol{\theta})$. The estimated covariance matrix of $\hat{\boldsymbol{\theta}}$ is

$$\widehat{\operatorname{Var}}(\hat{\boldsymbol{\theta}}) = \hat{\sigma}^2 (\hat{\boldsymbol{F}}^\top \boldsymbol{P_W} \hat{\boldsymbol{F}})^{-1}.$$

If the model is a nonlinear regression model, then $\hat{\boldsymbol{\theta}}$ is really the nonlinear instrumental variables estimator.

Case 2. Covariance matrix known up to a scalar factor: $\boldsymbol{\Omega} = \sigma^2 \boldsymbol{\Delta}$.

When $\operatorname{plim} n^{-1} \boldsymbol{F}^\top(\boldsymbol{\theta}) \boldsymbol{f}(\boldsymbol{\theta}) = \mathbf{0}$, we solve the estimating equations (2.82), with $\boldsymbol{\Omega}$ replaced by $\boldsymbol{\Delta}$, to obtain an efficient estimator. This is equivalent to minimizing $\boldsymbol{f}^\top(\boldsymbol{\theta}) \boldsymbol{\Delta}^{-1} \boldsymbol{f}(\boldsymbol{\theta})$. The estimated covariance matrix is

$$\widehat{\operatorname{Var}}(\hat{\boldsymbol{\theta}}) = \hat{\sigma}^2 (\hat{\boldsymbol{F}}^\top \boldsymbol{\Delta}^{-1} \hat{\boldsymbol{F}})^{-1},$$

where $\hat{\sigma}^2$ consistently estimates σ^2. If the underlying model is a nonlinear regression model, then $\hat{\boldsymbol{\theta}}$ is really the nonlinear GLS estimator discussed in Section x.x

When $\operatorname{plim} n^{-1} \boldsymbol{F}^\top(\boldsymbol{\theta}) \boldsymbol{f}(\boldsymbol{\theta}) \neq \mathbf{0}$, we again must replace $\boldsymbol{F}(\boldsymbol{\theta})$ by an estimate of its conditional expectation. This means that we should solve the estimating equations (2.89) with $\boldsymbol{A} = \boldsymbol{\Psi}^\top \boldsymbol{W}$, where $\boldsymbol{\Psi}$ satisfies $\boldsymbol{\Delta}^{-1} = \boldsymbol{\Psi} \boldsymbol{\Psi}^\top$. This is equivalent to minimizing (2.88) with the same definition of $\boldsymbol{A}$. The estimated covariance matrix is

$$\hat{\sigma}^2 (\hat{\boldsymbol{F}}^\top \boldsymbol{\Psi} \boldsymbol{P}_{\boldsymbol{\Psi}^\top \boldsymbol{W}} \boldsymbol{\Psi}^\top \hat{\boldsymbol{F}})^{-1}.$$

If the model is a linear regression model, then $\hat{\boldsymbol{\theta}}$ is the fully efficient GMM estimator (2.26) whenever the span of the instruments $\boldsymbol{W}$ includes the span of the optimal instruments $\bar{\boldsymbol{X}}$.

When the matrix $\boldsymbol{\Delta}$ is unknown but depends on a fixed number of parameters that can be estimated consistently, we can replace $\boldsymbol{\Delta}$ by a consistent estimator $\hat{\boldsymbol{\Delta}}$ and proceed as if it were known, as in feasible GLS estimation.

Case 3. Unknown diagonal or general covariance matrix.

This is the most commonly encountered case in which GMM estimation is explicitly used. Fully efficient estimation is no longer possible, but we can still obtain estimates that are efficient for a given set of instruments by using a consistent estimator $\hat{\boldsymbol{\Sigma}}$ of the matrix $\boldsymbol{\Sigma}$ defined in (2.33). This estimator is heteroskedasticity-consistent if $\boldsymbol{\Omega}$ is assumed to be diagonal and some sort of HAC estimator otherwise. Whether or not $\operatorname{plim} n^{-1}\boldsymbol{F}^\top(\boldsymbol{\theta})\boldsymbol{f}(\boldsymbol{\theta}) = \boldsymbol{0}$, we solve the estimating equations (2.79), which is equivalent to minimizing (2.86). The estimated covariance matrix is (2.81). If there is to be any gain in efficiency relative to NLS or nonlinear IV, it is essential that l, the number of columns of $\boldsymbol{W}$, is greater than k, the number of parameters to be estimated.

The consistent estimator $\hat{\boldsymbol{\Sigma}}$ is usually obtained from initial estimates that are consistent but inefficient. These may be NLS estimates, nonlinear IV estimates, or GMM estimates that do not use the optimal weighting matrix. The efficient GMM estimates are usually obtained by minimizing the criterion function (2.86), and the minimized value of this criterion function then serves as a Hansen-Sargan test statistic.

The first-round estimates $\hat{\boldsymbol{\theta}}$ can be used to obtain a new estimate of $\boldsymbol{\Sigma}$, which can then be used to obtain a second-round estimate of $\boldsymbol{\theta}$, which can be used to obtain yet another estimate of $\boldsymbol{\Sigma}$, and so on, until the process converges or the investigator loses patience. For a correctly specified model, all of these estimators have the same asymptotic distribution. However, performing more than one iteration often improves the finite-sample properties of the estimator. Thus, if computing cost is not a problem, it may well be best to use the continuously updated estimator that has been iterated to convergence.

For a more thorough treatment of the asymptotic theory of GMM estimation, see Newey and McFadden (1994).

2.6 Final Remarks

As its name implies, the generalized method of moments is a very general estimation method indeed, and numerous other methods can be thought of as special cases. These include all of the ones we have discussed so far: OLS, NLS, GLS, and IV. Thus the number of techniques that can legitimately be given the label "GMM" is bewilderingly large. To avoid bewilderment, it is best not to attempt to enumerate all the possibilities, but simply to list some of the ways in which various GMM estimators differ:

- Methods for which the explanatory variables are exogenous or predetermined (including OLS, NLS, and GLS), and for which no extra instruments are required, versus methods that do require additional exogenous or predetermined instruments (including linear and nonlinear IV).
- Methods for linear models (including OLS, GLS, linear IV, and the GMM techniques discussed in Section 2.2) versus methods for nonlinear models (including NLS, GNLS, nonlinear IV, and the GMM techniques discussed in Section 2.5).
- Methods that are inefficient for a given set of moment conditions, which have sandwich covariance matrices, versus methods that are efficient for the same set of moment conditions, which do not.
- Methods that are fully efficient, because they are based on optimal instruments, versus methods that are not fully efficient.
- Methods based on a covariance matrix that is known, at least up to a finite number of parameters which can be estimated consistently, versus methods that require an HCCME or a HAC estimator. The latter can never be fully efficient.
- Univariate models versus multivariate models. We have not yet discussed any methods for estimating the latter, but we will do so in Chapter 5.

2.7 Exercises

2.1 Show that the difference between the matrix

$$(\boldsymbol{J}^\top\boldsymbol{W}^\top\boldsymbol{X})^{-1}\boldsymbol{J}^\top\boldsymbol{W}^\top\boldsymbol{\Omega}\boldsymbol{W}\boldsymbol{J}(\boldsymbol{X}^\top\boldsymbol{W}\boldsymbol{J})^{-1}$$

and the matrix

$$(\boldsymbol{X}^\top\boldsymbol{W}(\boldsymbol{W}^\top\boldsymbol{\Omega}\boldsymbol{W})^{-1}\boldsymbol{W}^\top\boldsymbol{X})^{-1}$$

is a positive semidefinite matrix. **Hints:** Recall Part 1, Exercise 4.F14. Express the second of the two matrices in terms of the projection matrix $\boldsymbol{P}_{\boldsymbol{\Omega}^{1/2}\boldsymbol{W}}$, and then find a similar projection matrix for the first of them.

2.2 Let the n–vector $\boldsymbol{u}$ be such that $\mathrm{E}(\boldsymbol{u}) = \boldsymbol{0}$ and $\mathrm{E}(\boldsymbol{u}\boldsymbol{u}^\top) = \boldsymbol{I}$, and let the $n \times l$ matrix $\boldsymbol{W}$ be such that $\mathrm{E}(\boldsymbol{W}_t u_t) = \boldsymbol{0}$ and that $\mathrm{E}(u_t u_s \mid \boldsymbol{W}_t, \boldsymbol{W}_s) = \delta_{ts}$, where δ_{ts} is the Kronecker delta. Assume that $\boldsymbol{S}_{\boldsymbol{W}^\top\boldsymbol{W}} \equiv \operatorname{plim} n^{-1}\boldsymbol{W}^\top\boldsymbol{W}$ is finite, deterministic, and positive definite. Explain why the quadratic form $\boldsymbol{u}^\top\boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{u}$ must be asymptotically distributed as $\chi^2(l)$.

2.3 Consider the quadratic form $\boldsymbol{x}^\top\boldsymbol{A}\boldsymbol{x}$, where $\boldsymbol{x}$ is a $p \times 1$ vector and $\boldsymbol{A}$ is a $p \times p$ matrix, which may or may not be symmetric. Show that there exists a symmetric $p \times p$ matrix $\boldsymbol{B}$ such that $\boldsymbol{x}^\top\boldsymbol{B}\boldsymbol{x} = \boldsymbol{x}^\top\boldsymbol{A}\boldsymbol{x}$ for all $p \times 1$ vectors $\boldsymbol{x}$, and give the explicit form of a suitable $\boldsymbol{B}$.

⋆**2.4** For the model (2.01) and a specific choice of the $l \times k$ matrix $\boldsymbol{J}$, show that minimizing the quadratic form (2.12) with weighting matrix $\boldsymbol{\Lambda} = \boldsymbol{J}\boldsymbol{J}^\top$ gives the same estimator as solving the moment conditions (2.05) with the given $\boldsymbol{J}$. Assuming that these moment conditions have a unique solution for $\boldsymbol{\beta}$, show that the matrix $\boldsymbol{J}\boldsymbol{J}^\top$ is of rank k, and hence positive semidefinite without being positive definite.

Construct a symmetric, positive definite, $l \times l$ weighting matrix $\boldsymbol{\Lambda}$ such that minimizing (2.12) with this $\boldsymbol{\Lambda}$ leads once more to the same estimator as that given by solving conditions (2.05). It is convenient to take $\boldsymbol{\Lambda}$ in the form

$\boldsymbol{J}\boldsymbol{J}^\top + \boldsymbol{N}\boldsymbol{N}^\top$. In the construction of $\boldsymbol{N}$, it may be useful to partition $\boldsymbol{W}$ as $[\boldsymbol{W}_1 \;\; \boldsymbol{W}_2]$, where the $n \times k$ matrix $\boldsymbol{W}_1$ is such that $\boldsymbol{W}_1^\top \boldsymbol{X}$ is nonsingular.

⋆2.5 Consider the linear regression model with serially correlated disturbances,

$$y_t = \beta_1 + \beta_2 x_t + u_t, \quad u_t = \rho u_{t-1} + \varepsilon_t, \tag{2.93}$$

where the ε_t are IID, and the autoregressive parameter ρ is assumed either to be known or to be estimated consistently. The explanatory variable x_t is assumed to be contemporaneously correlated with ε_t (see Part 1, Section 9.4 for the definition of contemporaneous correlation).

Recall from Chapter 8 that the covariance matrix $\boldsymbol{\Omega}$ of the vector $\boldsymbol{u}$ with typical element u_t is given by (F9.27), and that $\boldsymbol{\Omega}^{-1}$ can be expressed as $\boldsymbol{\Psi}\boldsymbol{\Psi}^\top$, where $\boldsymbol{\Psi}$ is defined in (F9.46). Express the model (2.93) in the form (2.20), without taking account of the first observation.

Let Ω_t be the information set for observation t with $\mathrm{E}(\varepsilon_t \,|\, \Omega_t) = 0$. Suppose that there exists a matrix $\boldsymbol{Z}$ of instrumental variables, with $\boldsymbol{Z}_t \in \Omega_t$, such that the explanatory vector $\boldsymbol{x}$ with typical element x_t is related to the instruments by the equation

$$\boldsymbol{x} = \boldsymbol{Z}\boldsymbol{\pi} + \boldsymbol{v}, \tag{2.94}$$

where $\mathrm{E}(v_t \,|\, \Omega_t) = 0$. Derive the explicit form of the expression $(\boldsymbol{\Psi}^\top \bar{\boldsymbol{X}})_t$ defined implicitly by equation (2.24) for the model (2.93). Find a matrix $\boldsymbol{W}$ of instruments that satisfy the predeterminedness condition in the form (2.30) and that lead to asymptotically efficient estimates of the parameters β_1 and β_2 computed on the basis of the theoretical moment conditions (2.31) with your choice of $\boldsymbol{W}$.

⋆2.6 Consider the model (2.20), where the matrix $\boldsymbol{\Psi}$ is chosen in such a way that the transformed disturbances, the $(\boldsymbol{\Psi}^\top \boldsymbol{u})_t$, are innovations with respect to the information sets Ω_t. In other words, $\mathrm{E}((\boldsymbol{\Psi}^\top \boldsymbol{u})_t \,|\, \Omega_t) = 0$. Suppose that the $n \times l$ matrix of instruments $\boldsymbol{W}$ is predetermined in the usual sense that $\boldsymbol{W}_t \in \Omega_t$. Show that these assumptions, along with the assumption that $\mathrm{E}((\boldsymbol{\Psi}^\top \boldsymbol{u})_t^2 \,|\, \Omega_t) = \mathrm{E}((\boldsymbol{\Psi}^\top \boldsymbol{u})_t^2) = 1$ for $t = 1, \dots, n$, are enough to prove the analog of (2.02), that is, that

$$\mathrm{Var}(n^{-1/2} \boldsymbol{W}^\top \boldsymbol{\Psi}^\top \boldsymbol{u}) = n^{-1} \mathrm{E}(\boldsymbol{W}^\top \boldsymbol{W}).$$

In order to perform just-identified estimation, let the $n \times k$ matrix $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{J}$, for an $l \times k$ matrix $\boldsymbol{J}$ of full column rank. Compute the asymptotic covariance matrix of the estimator obtained by solving the moment conditions

$$\boldsymbol{Z}^\top \boldsymbol{\Psi}^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) = \boldsymbol{J}^\top \boldsymbol{W}^\top \boldsymbol{\Psi}^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) = \boldsymbol{0}. \tag{2.95}$$

The covariance matrix you have found should be a sandwich. Find the choice of $\boldsymbol{J}$ that eliminates the sandwich, and show that this choice leads to an asymptotic covariance matrix that is smaller, in the usual sense, than the asymptotic covariance matrix for any other choice of $\boldsymbol{J}$.

Compute the GMM criterion function for model (2.20) with instruments $\boldsymbol{W}$, and show that the estimator found by minimizing this criterion function is just the estimator obtained using the optimal choice of $\boldsymbol{J}$.

2.7 Compare the asymptotic covariance matrix found in the preceding question for the estimator of the parameters of model (2.20), obtained by minimizing the GMM criterion function for the $n \times l$ matrix of predetermined instruments $\boldsymbol{W}$, with the covariance matrix (2.29) that corresponds to estimation with instruments $\boldsymbol{\Psi}^\top \bar{\boldsymbol{X}}$. In particular, show that the difference between the two is a positive semidefinite matrix.

2.8 Consider overidentified estimation based on the moment conditions

$$\mathrm{E}(\boldsymbol{W}^\top \boldsymbol{\Omega}^{-1} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})) = \boldsymbol{0},$$

which were given in (2.31), where the $n \times l$ matrix of instruments $\boldsymbol{W}$ satisfies the predeterminedness condition (2.30). Derive the GMM criterion function for these theoretical moment conditions, and show that the estimating equations that result from the minimization of this criterion function are

$$\boldsymbol{X}^\top \boldsymbol{\Omega}^{-1} \boldsymbol{W} (\boldsymbol{W}^\top \boldsymbol{\Omega}^{-1} \boldsymbol{W})^{-1} \boldsymbol{W}^\top \boldsymbol{\Omega}^{-1} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) = \boldsymbol{0}. \tag{2.96}$$

Suppose that $\mathcal{S}(\bar{\boldsymbol{X}})$, the span of the $n \times k$ matrix $\bar{\boldsymbol{X}}$ of optimal instruments defined by (2.24), is a linear subspace of $\mathcal{S}(\boldsymbol{W})$, the span of the transformed instruments. Show that, in this case, the estimating equations (2.96) are asymptotically equivalent to

$$\bar{\boldsymbol{X}}^\top \boldsymbol{\Omega}^{-1} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) = \boldsymbol{0},$$

of which the solution is the efficient estimator $\hat{\boldsymbol{\beta}}_{\mathrm{EGMM}}$ defined in (2.26).

2.9 Show that the asymptotic covariance matrix of the estimator obtained by solving the estimating equations (2.96) is

$$\operatorname*{plim}_{n \to \infty} \left(\frac{1}{n} \bar{\boldsymbol{X}}^\top \boldsymbol{\Omega}^{-1} \boldsymbol{W} (\boldsymbol{W}^\top \boldsymbol{\Omega}^{-1} \boldsymbol{W})^{-1} \boldsymbol{W}^\top \boldsymbol{\Omega}^{-1} \bar{\boldsymbol{X}} \right)^{-1}. \tag{2.97}$$

By expressing this asymptotic covariance matrix in terms of a matrix $\boldsymbol{\Psi}$ that satisfies the equation $\boldsymbol{\Omega}^{-1} = \boldsymbol{\Psi}\boldsymbol{\Psi}^\top$, show that the difference between it and the asymptotic covariance matrix of the efficient estimator $\hat{\boldsymbol{\beta}}_{\mathrm{EGMM}}$ of (2.26) is a positive semidefinite matrix.

⋆2.10 Give the explicit form of the $n \times n$ matrix $\boldsymbol{U}(j)$ for which $\hat{\boldsymbol{\Gamma}}(j)$, defined in (2.36), takes the form $n^{-1} \boldsymbol{W}^\top \boldsymbol{U}(j) \boldsymbol{W}$.

2.11 This question uses data on daily returns for the period 1989–1998 from the file **daily-crsp.data**. These data are made available by courtesy of the Center for Research in Security Prices (CRSP); see the comments at the bottom of the file. Let r_t denote the daily return on shares of Mobil Corporation, and let v_t denote the daily return for the CRSP value-weighted index. Using all but the first four observations (to allow for lags), run the regression

$$r_t = \beta_1 + \beta_2 v_t + u_t$$

by OLS. Report three different sets of standard errors: the usual OLS ones, ones based on the simplest HCCME, and ones based on a more advanced HCCME that corrects for the downward bias in the squared OLS residuals; see Section 3.x. Do the OLS standard errors appear to be reliable?

Assuming that the u_t are heteroskedastic but serially uncorrelated, obtain estimates of the β_i that are more efficient than the OLS ones. For this purpose, use r_{t-1}^2, v_t^2, v_{t-1}^2, and v_{t-2}^2 as additional instruments. Do these estimates appear to be more efficient than the OLS ones?

2.12 Using the data for consumption (C_t) and disposable income (Y_t) contained in the file **consumption.data**, construct the variables $c_t = \log C_t$, $\Delta c_t = c_t - c_{t-1}$, $y_t = \log Y_t$, and $\Delta y_t = y_t - y_{t-1}$. Then, for the period 1953:1 to 1996:4, run the regression

$$\Delta c_t = \beta_1 + \beta_2 \Delta y_t + \beta_3 \Delta y_{t-1} + u_t \tag{2.98}$$

by OLS, and test the hypothesis that the u_t are serially uncorrelated against the alternative that they follow an AR(1) process.

Calculate eight sets of HAC estimates of the standard errors of the OLS parameter estimates from regression (2.98), using the Newey-West estimator with the lag truncation parameter set to the values $p = 1, 2, 3, 4, 5, 6, 7, 8$.

2.13 Using the squares of Δy_t, Δy_{t-1}, and Δc_{t-1} as additional instruments, obtain feasible efficient GMM estimates of the parameters of (2.98) by minimizing the criterion function (2.42), with $\hat{\boldsymbol{\Sigma}}$ given by the HAC estimators computed in the previous exercise. For $p = 6$, carry out the iterative procedure described in Section 2.3 by which new parameter estimates are used to update the HAC estimator, which is then used to update the parameter estimates. **Warning:** It may be necessary to rescale the instruments so as to avoid numerical problems.

2.14 Suppose that $f_t = y_t - \boldsymbol{X}_t\boldsymbol{\beta}$. Show that, in this special case, the estimating equations (2.77) yield the generalized IV estimator.

2.15 Starting from the asymptotic covariance matrix (2.67), show that, when $\boldsymbol{\Omega}^{-1}\boldsymbol{F}_0$ is used in place of $\boldsymbol{Z}$, the covariance matrix of the resulting estimator is given by (2.83). Then show that, for the linear regression model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}$ with exogenous explanatory variables $\boldsymbol{X}$, this estimator is the GLS estimator.

⋆2.16 The minimization of the GMM criterion function (2.87) yields the estimating equations (2.89) with $\boldsymbol{A} = \boldsymbol{\Psi}^\top\boldsymbol{W}$. Assuming that the $n \times l$ instrument matrix $\boldsymbol{W}$ satisfies the predeterminedness condition in the form (2.30), show that these estimating equations are asymptotically equivalent to the equations

$$\bar{\boldsymbol{F}}_0^\top \boldsymbol{\Psi} \boldsymbol{P}_{\boldsymbol{\Psi}^\top\boldsymbol{W}} \boldsymbol{\Psi}^\top \boldsymbol{f}(\hat{\boldsymbol{\theta}}) = \boldsymbol{0}, \tag{2.99}$$

where, as usual, $\bar{\boldsymbol{F}}_0 \equiv \bar{\boldsymbol{F}}(\boldsymbol{\theta}_0)$, with $\boldsymbol{\theta}_0$ the true parameter vector. Next, derive the asymptotic covariance matrix of the estimator defined by these equations.

Show that the equations (2.99) are the optimal estimating equations for overidentified estimation based on the transformed zero functions $\boldsymbol{\Psi}^\top\boldsymbol{f}(\boldsymbol{\theta})$ and the transformed instruments $\boldsymbol{\Psi}^\top\boldsymbol{W}$. Show further that, if the condition $\mathcal{S}(\bar{\boldsymbol{F}}) \subseteq \mathcal{S}(\boldsymbol{W})$ is satisfied, the asymptotic covariance matrix of the estimator obtained by solving equations (2.99) coincides with the optimal asymptotic covariance matrix (2.83).

⋆2.17 Suppose the n–vector $\boldsymbol{f}(\boldsymbol{\theta})$ of elementary zero functions has a covariance matrix $\sigma^2\mathbf{I}$. Show that, if the instrumental variables used for GMM estimation are the columns of the $n \times l$ matrix $\boldsymbol{W}$, the GMM criterion function is

$$\frac{1}{\sigma^2}\boldsymbol{f}^\top(\boldsymbol{\theta})\boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{f}(\boldsymbol{\theta}). \tag{2.100}$$

Next, show that, whenever the instruments are predetermined, the artificial regression

$$\boldsymbol{f}(\boldsymbol{\theta}) = -\boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{F}(\boldsymbol{\theta})\boldsymbol{b} + \text{residuals}, \tag{2.101}$$

where $\boldsymbol{F}(\boldsymbol{\theta})$ is defined as usual by (2.63), satisfies all the requisite properties for hypothesis testing. These properties are that the regressand should be orthogonal to the regressors when they are evaluated at the GMM estimator obtained by minimizing (2.100); that the OLS covariance matrix from (2.101) should be a consistent estimate of the asymptotic variance of that estimator; and that (2.101) should admit one-step estimation.

⋆2.18 Derive a heteroskedasticity robust version of the artificial regression (2.101), assuming that the covariance matrix of the vector $\boldsymbol{f}(\boldsymbol{\theta})$ of zero functions is diagonal, but otherwise arbitrary.

⋆2.19 If the scalar random variable z is distributed according to the $\mathrm{N}(\mu, \sigma^2)$ distribution, show that

$$\mathrm{E}(e^z) = \exp(\mu + \tfrac{1}{2}\sigma^2).$$

⋆2.20 Let the components z_t of the n–vector $\boldsymbol{z}$ be IID drawings from the $\mathrm{N}(\mu, \sigma^2)$ distribution, and let s^2 be the OLS estimate of the disturbance variance from the regression of $\boldsymbol{z}$ on the constant vector $\boldsymbol{\iota}$. Show that the variance of s^2 is $2\sigma^4/(n-1)$.

Would this result still hold if the normality assumption were dropped? Without this assumption, what would you need to know about the distribution of the z_t in order to find the variance of s^2?

Chapter 3

The Method of Maximum Likelihood

3.1 Introduction

Estimating equations based on elementary zero functions and instrumental variables are not the only useful techniques of estimation, even though the estimation methods for regression models discussed up to this point (ordinary, nonlinear, and generalized least squares, instrumental variables, and GMM) can all be derived from them. In this chapter, we introduce another fundamental method of estimation, namely, the method of **maximum likelihood**. For regression models, if we make the assumption that the disturbances are normally distributed, the maximum likelihood, or **ML**, estimators coincide with the various least-squares estimators with which we are already familiar. But maximum likelihood can also be applied to an extremely wide variety of models other than regression models, and it generally yields estimators with excellent asymptotic properties. The major disadvantage of ML estimation is that it requires stronger distributional assumptions than other methods.

In the next section, we introduce the basic ideas of maximum likelihood estimation and discuss a few simple examples. Then, in Section 3.3, we explore the asymptotic properties of ML estimators. Ways of estimating the covariance matrix of an ML estimator will be discussed in Section 3.4. Some methods of hypothesis testing that are available for models estimated by ML will be introduced in Section 3.5 and discussed more formally in Section 3.6. The remainder of the chapter discusses some useful applications of maximum likelihood estimation. Section 3.7 deals with regression models with autoregressive disturbances, and Section 3.8 deals with models that involve transformations of the dependent variable.

3.2 Basic Concepts of Maximum Likelihood Estimation

Models that are estimated by maximum likelihood must be **fully specified parametric models**, in the sense of Part 1, Section 2.3. For such a model, once the parameter values are known, all necessary information is available to simulate the dependent variable(s). In Part 1, Section 2.2, we introduced the concept of the probability density function, or PDF, of a scalar random variable and of the joint density function, or joint density, of a set of random variables. If we can simulate the dependent variable, this means that its density must be known, both for each observation as a scalar r.v., and for the full sample as a vector r.v.

As usual, we denote the dependent variable by the n–vector $\boldsymbol{y}$. For a given k–vector $\boldsymbol{\theta}$ of parameters, let the joint density of $\boldsymbol{y}$ be written as $f(\boldsymbol{y}, \boldsymbol{\theta})$. This joint density constitutes the specification of the model. Since a density provides an unambiguous recipe for simulation, it suffices to specify the vector $\boldsymbol{\theta}$ in order to give a full characterization of a DGP in the model. Thus there is a one-to-one correspondence between the DGPs of the model and the admissible parameter vectors.

Maximum likelihood estimation is based on the specification of the model through the joint density $f(\boldsymbol{y}, \boldsymbol{\theta})$. When $\boldsymbol{\theta}$ is fixed, the function $f(\cdot, \boldsymbol{\theta})$ of $\boldsymbol{y}$ is interpreted as the density of $\boldsymbol{y}$. But if instead $f(\boldsymbol{y}, \boldsymbol{\theta})$ is evaluated at the n–vector $\boldsymbol{y}$ found in a given data set, then the function $f(\boldsymbol{y}, \cdot)$ of the model parameters can no longer be interpreted as a density. Instead, it is referred to as the **likelihood function** of the model for the given data set. ML estimation then amounts to maximizing the likelihood function with respect to the parameters. A parameter vector $\hat{\boldsymbol{\theta}}$ at which the likelihood takes on its maximum value is called a **maximum likelihood estimate**, or **MLE**, of the parameters.

In many cases, the successive observations in a sample are assumed to be statistically independent. In that case, the joint density of the entire sample is just the product of the densities of the individual observations. Let $f(y_t, \boldsymbol{\theta})$ denote the density of a typical observation, y_t. Then the joint density of the entire sample $\boldsymbol{y}$ is

$$f(\boldsymbol{y}, \boldsymbol{\theta}) = \prod_{t=1}^{n} f(y_t, \boldsymbol{\theta}). \tag{3.01}$$

Because (3.01) is a product, it is often a very large or very small number, perhaps so large or so small that it cannot easily be represented in a computer. For this and a number of other reasons, it is customary to work instead with the **loglikelihood function**

$$\ell(\boldsymbol{y}, \boldsymbol{\theta}) \equiv \log f(\boldsymbol{y}, \boldsymbol{\theta}) = \sum_{t=1}^{n} \ell_t(y_t, \boldsymbol{\theta}), \tag{3.02}$$

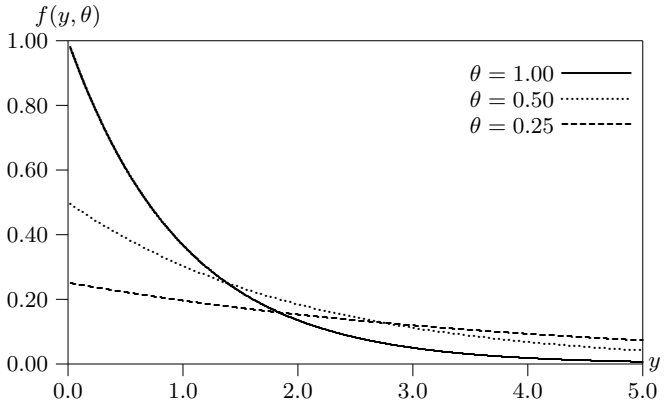


Figure 3.1 The exponential distribution

where $\ell_t(y_t, \boldsymbol{\theta})$, the **contribution** to the loglikelihood function made by observation t, is equal to $\log f_t(y_t, \boldsymbol{\theta})$. The t subscripts on f_t and ℓ_t have been added to allow for the possibility that the density of y_t may vary from observation to observation, perhaps because there are exogenous variables in the model. Whatever value of $\boldsymbol{\theta}$ maximizes the loglikelihood function (3.02) must also maximize the likelihood function (3.01), because $\ell(\boldsymbol{y}, \boldsymbol{\theta})$ is just a monotonic transformation of $f(\boldsymbol{y}, \boldsymbol{\theta})$.

The Exponential Distribution

As a simple example of ML estimation, suppose that each observation y_t is generated by the density

$$f(y_t, \theta) = \theta e^{-\theta y_t}, \quad y_t > 0, \;\; \theta > 0. \tag{3.03}$$

This is the density of what is called the **exponential distribution**.[1] This density is shown in Figure 9.1 for three values of the parameter θ, which is what we wish to estimate. There are assumed to be n independent observations from which to calculate the loglikelihood function.

Taking the logarithm of the density (3.03), we find that the contribution to the loglikelihood from observation t is $\ell_t(y_t, \theta) = \log \theta - \theta y_t$. Therefore,

$$\ell(\boldsymbol{y}, \theta) = \sum_{t=1}^{n} (\log \theta - \theta y_t) = n \log \theta - \theta \sum_{t=1}^{n} y_t. \tag{3.04}$$

[1] The exponential distribution is useful for analyzing dependent variables which must be positive, such as waiting times or the duration of unemployment. Models for duration data will be discussed in Chapter 4.

To maximize this loglikelihood function with respect to the single unknown parameter θ, we differentiate it with respect to θ and set the derivative equal to 0. The result is

$$\frac{n}{\theta} - \sum_{t=1}^{n} y_t = 0, \tag{3.05}$$

which can easily be solved to yield

$$\hat{\theta} = \frac{n}{\sum_{t=1}^{n} y_t}. \tag{3.06}$$

This solution is clearly unique, because the second derivative of (3.04), which is the first derivative of the left-hand side of (3.05), is always negative, which implies that the first derivative can vanish at most once. Since it is unique, the estimator $\hat{\theta}$ defined in (3.06) can be called *the* maximum likelihood estimator that corresponds to the loglikelihood function (3.04).

In this case, interestingly, the ML estimator $\hat{\theta}$ is exactly the same as a method-of-moments estimator. As we now show, the expected value of y_t is $1/\theta$. By definition, this expectation is

$$\mathrm{E}(y_t) = \int_0^\infty y_t \theta e^{-\theta y_t} dy_t.$$

Since $-\theta e^{-\theta y_t}$ is the derivative of $e^{-\theta y_t}$ with respect to y_t, we may integrate by parts to obtain

$$\int_0^\infty y_t \theta e^{-\theta y_t} dy_t = -\Big[y_t e^{-\theta y_t}\Big]_0^\infty + \int_0^\infty e^{-\theta y_t} dy_t = \Big[-\theta^{-1} e^{-\theta y_t}\Big]_0^\infty = \theta^{-1}.$$

The most natural estimator of θ is the one that matches θ^{-1} to the empirical analog of $\mathrm{E}(y_t)$, which is $\bar{y}$, the sample mean. This estimator of θ is therefore $1/\bar{y}$, which is identical to the ML estimator (3.06).

It is not uncommon for an ML estimator to coincide with a Z-estimator, as happens in this case. This may suggest that maximum likelihood is not a very useful addition to the econometrician's toolkit, but such an inference would be unwarranted. Even in this simple case, the ML estimator was considerably easier to obtain than the Z-estimator, because we did not need to calculate an expectation. In more complicated cases, this advantage of ML estimation is often much more substantial. Moreover, as we will see in the next three sections, the fact that an estimator is an MLE generally ensures that it has a number of desirable asymptotic properties and makes it easy to calculate standard errors and test statistics.[2]

[2] Notice that the abbreviation "MLE" here means "maximum likelihood estimator" rather than "maximum likelihood estimate." We will use "MLE" to mean either of these. Which of them it refers to in any given situation should generally be obvious from the context; see Part 1, Section 2.5.

Regression Models with Normal Disturbances

It is interesting to see what happens when we apply the method of maximum likelihood to the classical normal linear model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}, \quad \boldsymbol{u} \sim \mathrm{N}(\boldsymbol{0}, \sigma^2\mathbf{I}), \tag{3.07}$$

which was introduced in Part 1, Section 4.1. For this model, the explanatory variables in the matrix $\boldsymbol{X}$ are assumed to be exogenous. Consequently, in constructing the likelihood function, we may use the density of $\boldsymbol{y}$ conditional on $\boldsymbol{X}$. The elements u_t of the vector $\boldsymbol{u}$ are independently distributed as $\mathrm{N}(0, \sigma^2)$, and so y_t is distributed, conditionally on $\boldsymbol{X}$, as $\mathrm{N}(\boldsymbol{X}_t\boldsymbol{\beta}, \sigma^2)$. Thus the density of y_t is, from (F5.11),

$$f_t(y_t, \boldsymbol{\beta}, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y_t - \boldsymbol{X}_t\boldsymbol{\beta})^2}{2\sigma^2}\right). \tag{3.08}$$

The contribution to the loglikelihood function made by the t^{th} observation is the logarithm of (3.08). Since $\log\sigma = \frac{1}{2}\log\sigma^2$, this can be written as

$$\ell_t(y_t, \boldsymbol{\beta}, \sigma) = -\tfrac{1}{2}\log 2\pi - \tfrac{1}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(y_t - \boldsymbol{X}_t\boldsymbol{\beta})^2. \tag{3.09}$$

Since the observations are assumed to be independent, the loglikelihood function is just the sum of these contributions over all t, or

$$\begin{aligned} \ell(\boldsymbol{y}, \boldsymbol{\beta}, \sigma) &= -\tfrac{n}{2}\log 2\pi - \tfrac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{t=1}^{n}(y_t - \boldsymbol{X}_t\boldsymbol{\beta})^2 \\ &= -\tfrac{n}{2}\log 2\pi - \tfrac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}). \end{aligned} \tag{3.10}$$

In the second line, we rewrite the sum of squared residuals as the inner product of the residual vector with itself. To find the ML estimator, we need to maximize (3.10) with respect to the unknown parameters $\boldsymbol{\beta}$ and σ.

The first step in maximizing $\ell(\boldsymbol{y}, \boldsymbol{\beta}, \sigma)$ is to **concentrate** it with respect to the parameter σ. This means differentiating (3.10) with respect to σ, solving the resulting first-order condition for σ as a function of the data and the remaining parameters, and then substituting the result back into (3.10). This yields the **concentrated loglikelihood function**. The second step is to maximize this function with respect to $\boldsymbol{\beta}$. For models that involve variance parameters, it is very often convenient to concentrate the loglikelihood function in this way.

Differentiating the second line of (3.10) with respect to σ and equating the derivative to zero yields the first-order condition

$$\frac{\partial\ell(\boldsymbol{y}, \boldsymbol{\beta}, \sigma)}{\partial\sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) = 0,$$

and solving this yields the result that

$$\hat{\sigma}^2(\boldsymbol{\beta}) = \tfrac{1}{n}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}).$$

Here the notation $\hat{\sigma}^2(\boldsymbol{\beta})$ indicates that the value of σ^2 that maximizes (3.10) depends on $\boldsymbol{\beta}$.

Substituting $\hat{\sigma}^2(\boldsymbol{\beta})$ into the second line of (3.10) yields the concentrated loglikelihood function

$$\ell^c(\boldsymbol{y}, \boldsymbol{\beta}) = -\tfrac{n}{2}\log 2\pi - \tfrac{n}{2}\log\Big(\tfrac{1}{n}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\Big) - \tfrac{n}{2}. \tag{3.11}$$

The middle term here is minus $n/2$ times the logarithm of the sum of squared residuals, and the other two terms do not depend on $\boldsymbol{\beta}$. Thus we see that *maximizing* the concentrated loglikelihood function (3.11) is equivalent to *minimizing* the sum of squared residuals as a function of $\boldsymbol{\beta}$. Therefore, the ML estimator $\hat{\boldsymbol{\beta}}$ is identical to the OLS estimator.

Once $\hat{\boldsymbol{\beta}}$ has been found, the ML estimate $\hat{\sigma}^2$ of σ^2 is $\hat{\sigma}^2(\hat{\boldsymbol{\beta}})$, and the MLE of σ is the positive square root of $\hat{\sigma}^2$. Thus, as we saw in Part 1, Section 4.7, the MLE $\hat{\sigma}^2$ is biased downward.[3] The actual maximized value of the loglikelihood function can then be written in terms of the sum-of-squared residuals function SSR evaluated at $\hat{\boldsymbol{\beta}}$. From (3.11) we have

$$\ell(\boldsymbol{y}, \hat{\boldsymbol{\beta}}, \hat{\sigma}) = -\tfrac{n}{2}(1 + \log 2\pi - \log n) - \tfrac{n}{2}\log\mathrm{SSR}(\hat{\boldsymbol{\beta}}), \tag{3.12}$$

where $\mathrm{SSR}(\hat{\boldsymbol{\beta}})$ denotes the minimized sum of squared residuals.

Although it is convenient to concentrate (3.10) with respect to σ, as we have done, this is not the only way to proceed. In Exercise 3.1, readers are asked to show that the ML estimators of $\boldsymbol{\beta}$ and σ can be obtained equally well by concentrating the loglikelihood with respect to $\boldsymbol{\beta}$ rather than σ.

The fact that the ML and OLS estimators of $\boldsymbol{\beta}$ are identical depends critically on the assumption that the disturbances in (3.07) are normally distributed. If we had started with a different assumption about their distribution, we would have obtained a different ML estimator. The asymptotic efficiency result to be discussed in Section 3.4 would then imply that the least-squares estimator is asymptotically less efficient than the ML estimator whenever the two do not coincide.

[3] The bias arises because we evaluate $\mathrm{SSR}(\boldsymbol{\beta})$ at $\hat{\boldsymbol{\beta}}$ instead of at the true value $\boldsymbol{\beta}_0$. However, if one thinks of $\hat{\sigma}$ as an estimator of σ, rather than of $\hat{\sigma}^2$ as an estimator of σ^2, then it can be shown that both the OLS and the ML estimators are biased downward.

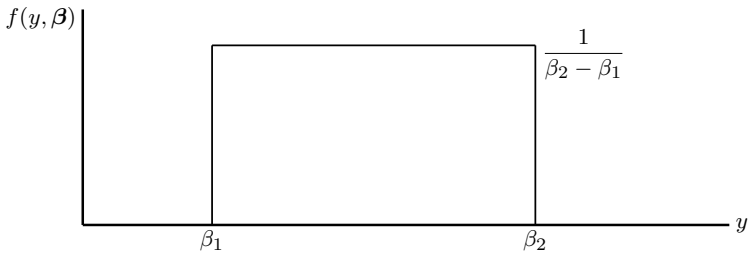


Figure 3.2 The uniform distribution

The Uniform Distribution

As a final example of ML estimation, we consider a somewhat pathological, but rather interesting, example. Suppose that the y_t are generated as independent realizations from the uniform distribution with parameters β_1 and β_2, which can be written as a vector $\boldsymbol{\beta}$; a special case of this distribution was introduced in Part 1, Section 2.2. The density function for y_t, which is graphed in Figure 3.2, is

$$f(y_t, \boldsymbol{\beta}) = 0 \text{ if } y_t < \beta_1,$$
$$f(y_t, \boldsymbol{\beta}) = \frac{1}{\beta_2 - \beta_1} \text{ if } \beta_1 \leq y_t \leq \beta_2,$$
$$f(y_t, \boldsymbol{\beta}) = 0 \text{ if } y_t > \beta_2.$$

Provided that $\beta_1 < y_t < \beta_2$ for all observations, the likelihood function is equal to $1/(\beta_2 - \beta_1)^n$, and the loglikelihood function is therefore

$$\ell(\boldsymbol{y}, \boldsymbol{\beta}) = -n \log(\beta_2 - \beta_1).$$

It is easy to verify that this function cannot be maximized by differentiating it with respect to the parameters and setting the partial derivatives to zero. Instead, the way to maximize $\ell(\boldsymbol{y}, \boldsymbol{\beta})$ is to make $\beta_2 - \beta_1$ as small as possible. But we clearly cannot make β_1 larger than the smallest observed y_t, and we cannot make β_2 smaller than the largest observed y_t. Otherwise, the likelihood function would be equal to 0. It follows that the ML estimators are

$$\hat{\beta}_1 = \min(y_t) \quad \text{and} \quad \hat{\beta}_2 = \max(y_t). \tag{3.13}$$

These estimators are rather unusual. For one thing, they always lie on one side of the true value. Because all the y_t must lie between β_1 and β_2, it must be the case that $\hat{\beta}_1 \geq \beta_{10}$ and $\hat{\beta}_2 \leq \beta_{20}$, where β_{10} and β_{20} denote the true parameter values. However, despite this, these estimators turn out to be consistent. Intuitively, this is because, as the sample size gets large, the observed values of y_t fill up the entire space between β_{10} and β_{20}.

The ML estimators defined in (3.13) are **super-consistent**, which means that they approach the true values of the parameters they are estimating at a rate faster than the usual rate of $n^{-1/2}$. Formally, $n^{1/2}(\hat{\beta}_1 - \beta_{10})$ tends to zero as $n \to \infty$, while $n(\hat{\beta}_1 - \beta_{10})$ tends to a finite limiting distribution; see Exercise 3.2 for more details. Now consider the parameter $\gamma \equiv \frac{1}{2}(\beta_1 + \beta_2)$. One way to estimate it is to use the ML estimator

$$\hat{\gamma} = \tfrac{1}{2}(\hat{\beta}_1 + \hat{\beta}_2).$$

Another approach would simply be to use the sample mean, say $\bar{\gamma}$, which is a least-squares estimator. But the ML estimator $\hat{\gamma}$ is super-consistent, while $\bar{\gamma}$ is only root-n consistent. This implies that, except perhaps for very small sample sizes, the ML estimator is very much more efficient than the least-squares estimator. In Exercise 3.3, readers are invited to perform a simulation experiment to illustrate this result.

Although economists rarely need to estimate the parameters of a uniform distribution directly, ML estimators with properties similar to those of (3.13) do occur from time to time. In particular, certain econometric models of auctions lead to super-consistent ML estimators; see Donald and Paarsch (1993, 1996). However, because these estimators violate standard regularity conditions, such as those given in Theorems 8.2 and 8.3 of Davidson and MacKinnon (1993), we will not consider them further.

Two Types of ML Estimator

There are two different ways of defining the ML estimator, although most MLEs actually satisfy both definitions. A **Type 1 ML estimator** maximizes the loglikelihood function over the set Θ, where Θ denotes the **parameter space** in which the parameter vector $\boldsymbol{\theta}$ lies, which is generally assumed to be a subset of $\mathbb{R}^k$. This is the natural meaning of an MLE, and all three of the ML estimators just discussed are Type 1 estimators.

If the loglikelihood function is differentiable and attains an *interior* maximum in the parameter space, then the MLE must satisfy the first-order conditions for a maximum. A **Type 2 ML estimator** is defined as a solution to the **likelihood equations**, which are just the following first-order conditions:

$$\boldsymbol{g}(\boldsymbol{y}, \hat{\boldsymbol{\theta}}) = \boldsymbol{0}, \tag{3.14}$$

where $\boldsymbol{g}(\boldsymbol{y}, \boldsymbol{\theta})$ is the **gradient vector**, or **score vector**, which has typical element

$$g_i(\boldsymbol{y}, \boldsymbol{\theta}) \equiv \frac{\partial \ell(\boldsymbol{y}, \boldsymbol{\theta})}{\partial \theta_i} = \sum_{t=1}^{n} \frac{\partial \ell_t(y_t, \boldsymbol{\theta})}{\partial \theta_i}. \tag{3.15}$$

Because there may be more than one value of $\boldsymbol{\theta}$ that satisfies the likelihood equations (3.14), the definition further requires the Type 2 estimator $\hat{\boldsymbol{\theta}}$ to be

associated with a local maximum of $\ell(\boldsymbol{y}, \boldsymbol{\theta})$ and, as $n \to \infty$, the value of the loglikelihood function associated with $\hat{\boldsymbol{\theta}}$ to be higher than the value associated with any other root of the likelihood equations.

The ML estimator (3.06) for the parameter of the exponential distribution and the OLS estimators of $\boldsymbol{\beta}$ and σ^2 in the regression model with normal disturbances, like most ML estimators, are both Type 1 and Type 2 MLEs. However, the MLEs for the parameters of the uniform distribution defined in (3.13) are Type 1 but not Type 2 MLEs, because they are not the solutions to any set of likelihood equations. In rare circumstances, there also exist MLEs that are Type 2 but not Type 1; see Kiefer (1978) for an example.

Computing ML Estimates

Maximum likelihood estimates are often quite easy to compute. Indeed, for the three examples considered above, we were able to obtain explicit expressions. When no such expressions are available, as is often the case, it is necessary to use some sort of nonlinear maximization procedure. Many such procedures are readily available.

The discussion of Newton's Method and quasi-Newton methods in Section 1.4 applies with very minor changes to ML estimation. Instead of minimizing the sum of squared residuals function $Q(\boldsymbol{\beta})$, we maximize the loglikelihood function $\ell(\boldsymbol{\theta})$. Since the maximization is done with respect to $\boldsymbol{\theta}$ for a given sample $\boldsymbol{y}$, we suppress the explicit dependence of ℓ on $\boldsymbol{y}$. As in the NLS case, Newton's Method makes use of the **Hessian**, which is now a $k \times k$ matrix $\boldsymbol{H}(\boldsymbol{\theta})$ with typical element $\partial^2 \ell(\boldsymbol{\theta}) / \partial \theta_i \partial \theta_j$. The Hessian is the matrix of second derivatives of the loglikelihood function, and thus also the matrix of first derivatives of the gradient.

Let $\boldsymbol{\theta}_{(j)}$ denote the value of the vector of estimates at step j of the algorithm, and let $\boldsymbol{g}_{(j)}$ and $\boldsymbol{H}_{(j)}$ denote, respectively, the gradient and the Hessian evaluated at $\boldsymbol{\theta}_{(j)}$. Then the fundamental equation for Newton's Method is

$$\boldsymbol{\theta}_{(j+1)} = \boldsymbol{\theta}_{(j)} - \boldsymbol{H}_{(j)}^{-1} \boldsymbol{g}_{(j)}. \tag{3.16}$$

This may be obtained in exactly the same way as equation (1.43). Because the loglikelihood function is to be maximized, the Hessian should be negative definite, at least when $\boldsymbol{\theta}_{(j)}$ is sufficiently near $\hat{\boldsymbol{\theta}}$. This ensures that the step defined by (3.16) is in an uphill direction.

For the reasons discussed in Section 1.4, Newton's Method usually does not work well, and often does not work at all, when the Hessian is not negative definite. In such cases, one popular way to obtain the MLE is to use some sort of quasi-Newton method, in which (3.16) is replaced by the formula

$$\boldsymbol{\theta}_{(j+1)} = \boldsymbol{\theta}_{(j)} + \alpha_{(j)} \boldsymbol{D}_{(j)}^{-1} \boldsymbol{g}_{(j)},$$

where $\alpha_{(j)}$ is a scalar which is determined at each step, and $\boldsymbol{D}_{(j)}$ is a matrix which approximates $-\boldsymbol{H}_{(j)}$ near the maximum but is constructed so that it

is always positive definite. Sometimes, as in the case of NLS estimation, an artificial regression can be used to compute the vector $\boldsymbol{D}_{(j)}^{-1} \boldsymbol{g}_{(j)}$. We will encounter one such artificial regression in Section 3.4, and another, more specialized, one in Section 4.3.

When the loglikelihood function is globally concave and not too flat, maximizing it is usually quite easy. At the other extreme, when the loglikelihood function has several local maxima, doing so can be very difficult. See the discussion in Section 1.4 following Figure 1.3. Everything that is said there about dealing with multiple minima in NLS estimation applies, with certain obvious modifications, to the problem of dealing with multiple maxima in ML estimation.

3.3 Asymptotic Properties of ML Estimators

One of the attractive features of maximum likelihood estimation is that ML estimators are consistent under quite weak regularity conditions and asymptotically normally distributed under somewhat stronger conditions. Therefore, if an estimator is an ML estimator and the regularity conditions are satisfied, it is not necessary to show that it is consistent or derive its asymptotic distribution. In this section, we sketch derivations of the principal asymptotic properties of ML estimators. A rigorous discussion is beyond the scope of this book; interested readers may consult, among other references, Davidson and MacKinnon (1993, Chapter 8) and Newey and McFadden (1994).

Consistency of the MLE

Since almost all maximum likelihood estimators are of Type 1, we will discuss consistency only for this type of MLE. We first show that the expectation of the loglikelihood function is greater when it is evaluated at the true values of the parameters than when it is evaluated at any other values. For consistency, we also need both a finite-sample identification condition and an asymptotic identification condition. The former requires the loglikelihood to be different for different sets of parameter values. If, contrary to this assumption, there were two distinct parameter vectors, $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, such that $\ell(\boldsymbol{y}, \boldsymbol{\theta}_1) = \ell(\boldsymbol{y}, \boldsymbol{\theta}_2)$ for all $\boldsymbol{y}$, then it would obviously be impossible to distinguish between $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$. Thus a finite-sample identification condition is necessary for the model to make sense. The role of the asymptotic identification condition will be discussed below.

Let $L(\boldsymbol{\theta}) = \exp\big(\ell(\boldsymbol{\theta})\big)$ denote the likelihood function, where the dependence on $\boldsymbol{y}$ of both L and ℓ has been suppressed for notational simplicity. We wish to apply a result known as **Jensen's Inequality** to the ratio $L(\boldsymbol{\theta}^*)/L(\boldsymbol{\theta}_0)$, where $\boldsymbol{\theta}_0$ is the true parameter vector and $\boldsymbol{\theta}^*$ is any other vector in the parameter space of the model. Jensen's Inequality tells us that, if X is a real-valued random

variable, then $\mathrm{E}\big(h(X)\big) \leq h\big(\mathrm{E}(X)\big)$ whenever $h(\cdot)$ is a concave function. The inequality is strict whenever h is strictly concave over at least part of the **support** of the random variable X, that is, the set of real numbers for which the density of X is nonzero, and the support contains more than one point. See Exercise 3.4 for the proof of a restricted version of Jensen's Inequality.

Since the logarithm is a strictly concave function over the nonnegative real line, and since likelihood functions are nonnegative, we can conclude from Jensen's Inequality that

$$\mathrm{E}_0 \log\left(\frac{L(\boldsymbol{\theta}^*)}{L(\boldsymbol{\theta}_0)}\right) < \log \mathrm{E}_0\left(\frac{L(\boldsymbol{\theta}^*)}{L(\boldsymbol{\theta}_0)}\right), \tag{3.17}$$

with strict inequality for all $\boldsymbol{\theta}^* \neq \boldsymbol{\theta}_0$, on account of the finite-sample identification condition. Here the notation E_0 means the expectation taken under the DGP characterized by the true parameter vector $\boldsymbol{\theta}_0$. Since the joint density of the sample is simply the likelihood function evaluated at $\boldsymbol{\theta}_0$, the expectation on the right-hand side of (3.17) can be expressed as an integral over the support of the vector random variable $\boldsymbol{y}$. We have

$$\mathrm{E}_0\left(\frac{L(\boldsymbol{\theta}^*)}{L(\boldsymbol{\theta}_0)}\right) = \int \frac{L(\boldsymbol{\theta}^*)}{L(\boldsymbol{\theta}_0)} L(\boldsymbol{\theta}_0)\, d\boldsymbol{y} = \int L(\boldsymbol{\theta}^*)\, d\boldsymbol{y} = 1,$$

where the last equality here holds because every density must integrate to 1. Therefore, because $\log 1 = 0$, the inequality (3.17) implies that

$$\mathrm{E}_0 \log\left(\frac{L(\boldsymbol{\theta}^*)}{L(\boldsymbol{\theta}_0)}\right) = \mathrm{E}_0\, \ell(\boldsymbol{\theta}^*) - \mathrm{E}_0\, \ell(\boldsymbol{\theta}_0) < 0. \tag{3.18}$$

In words, (3.18) says that the expectation of the loglikelihood function when evaluated at the true parameter vector, $\boldsymbol{\theta}_0$, is strictly greater than its expectation when evaluated at any other parameter vector, $\boldsymbol{\theta}^*$.

If we can apply a law of large numbers to the contributions to the loglikelihood function, then we can assert that $\operatorname{plim} n^{-1}\ell(\boldsymbol{\theta}) = \lim n^{-1}\mathrm{E}_0\, \ell(\boldsymbol{\theta})$. Then (3.18) implies that

$$\operatorname*{plim}_{n\to\infty} \frac{1}{n}\ell(\boldsymbol{\theta}^*) \leq \operatorname*{plim}_{n\to\infty} \frac{1}{n}\ell(\boldsymbol{\theta}_0), \tag{3.19}$$

for all $\boldsymbol{\theta}^* \neq \boldsymbol{\theta}_0$, where the inequality is not necessarily strict, because we have taken a limit. Since the MLE $\hat{\boldsymbol{\theta}}$ maximizes $\ell(\boldsymbol{\theta})$, it must be the case that

$$\operatorname*{plim}_{n\to\infty} \frac{1}{n}\ell(\hat{\boldsymbol{\theta}}) \geq \operatorname*{plim}_{n\to\infty} \frac{1}{n}\ell(\boldsymbol{\theta}_0). \tag{3.20}$$

The only way that (3.19) and (3.20) can both be true is if

$$\operatorname*{plim}_{n\to\infty} \frac{1}{n}\ell(\hat{\boldsymbol{\theta}}) = \operatorname*{plim}_{n\to\infty} \frac{1}{n}\ell(\boldsymbol{\theta}_0). \tag{3.21}$$

In words, (3.21) says that the plim of $1/n$ times the loglikelihood function must be the same when it is evaluated at the MLE $\hat{\boldsymbol{\theta}}$ as when it is evaluated at the true parameter vector $\boldsymbol{\theta}_0$.

By itself, the result (3.21) does not prove that $\hat{\boldsymbol{\theta}}$ is consistent, because the weak inequality does not rule out the possibility that there may be many values $\boldsymbol{\theta}^*$ for which $\operatorname{plim} n^{-1}\ell(\boldsymbol{\theta}^*) = \operatorname{plim} n^{-1}\ell(\boldsymbol{\theta}_0)$. We must therefore explicitly assume that $\operatorname{plim} n^{-1}\ell(\boldsymbol{\theta}^*) \neq \operatorname{plim} n^{-1}\ell(\boldsymbol{\theta}_0)$ for all $\boldsymbol{\theta}^* \neq \boldsymbol{\theta}_0$. This is a form of **asymptotic identification condition**; see Section 1.2. More primitive regularity conditions on the model and the DGP can be invoked to ensure that the MLE is asymptotically identified. For example, we need to rule out pathological cases like (F4.18), in which each new observation adds less and less information about one or more of the parameters.

Dependent Observations

Before we can discuss the asymptotic normality of the MLE, we need to introduce some notation and terminology, and we need to establish a few preliminary results. First, we consider the structure of the likelihood and loglikelihood functions for models in which the successive observations are not independent, as is the case, for instance, when a regression function involves lags of the dependent variable.

Recall the definition (F2.15) of the density of one random variable conditional on another. This definition can be rewritten so as to take the form of a factorization of the joint density:

$$f(y_1, y_2) = f(y_1) f(y_2 \,|\, y_1), \tag{3.22}$$

where we use y_1 and y_2 in place of the variables x_2 and x_1, respectively, that appear in (F2.15). It is permissible to apply (3.22) to situations in which y_1 and y_2 are really vectors of random variables. Accordingly, consider the joint density of three random variables, and group the first two together. Analogously to (3.22), we have

$$f(y_1, y_2, y_3) = f(y_1, y_2) f(y_3 \,|\, y_1, y_2). \tag{3.23}$$

Substituting (3.22) into (3.23) yields the following factorization of the joint density:

$$f(y_1, y_2, y_3) = f(y_1) f(y_2 \,|\, y_1) f(y_3 \,|\, y_1, y_2).$$

For a sample of size n, it is easy to see that this last result generalizes to

$$f(y_1, \dots, y_n) = f(y_1) f(y_2 \,|\, y_1) \cdots f(y_n \,|\, y_1, \dots, y_{n-1}).$$

This result can be written using a somewhat more convenient notation as follows:

$$f(\boldsymbol{y}^n) = \prod_{t=1}^{n} f(y_t \,|\, \boldsymbol{y}^{t-1}),$$

where the vector $\boldsymbol{y}^t$ is a t–vector with components $y_1, y_2, \ldots, y_t$. One can think of $\boldsymbol{y}^t$ as the subsample consisting of the first t observations of the full sample. For a model that is to be estimated by maximum likelihood, the density $f(\boldsymbol{y}^n)$ depends on a k–vector of parameters $\boldsymbol{\theta}$, and we can then write

$$f(\boldsymbol{y}^n, \boldsymbol{\theta}) = \prod_{t=1}^{n} f(y_t \,|\, \boldsymbol{y}^{t-1}; \boldsymbol{\theta}). \tag{3.24}$$

The structure of (3.24) is a straightforward generalization of that of (3.01), where the marginal densities of the successive observations are replaced by densities conditional on the preceding observations.

The loglikelihood function corresponding to (3.24) has an additive structure:

$$\ell(\boldsymbol{y}, \boldsymbol{\theta}) = \sum_{t=1}^{n} \ell_t(\boldsymbol{y}^t, \boldsymbol{\theta}), \tag{3.25}$$

where we omit the superscript n from $\boldsymbol{y}$ for the full sample. In addition, in the contributions $\ell_t(\cdot)$ to the loglikelihood, we do not distinguish between the current variable y_t and the lagged variables in the vector $\boldsymbol{y}^{t-1}$. In this way, (3.25) has exactly the same structure as (3.02).

The Gradient

The gradient, or score, vector $\boldsymbol{g}(\boldsymbol{y}, \boldsymbol{\theta})$ is a k–vector that was defined in (3.15). As that equation makes clear, each component of the gradient vector is itself a sum of n contributions, and this remains true when the observations are dependent; the partial derivative of ℓ_t with respect to θ_i now depends on $\boldsymbol{y}^t$ rather than just y_t. It is convenient to group these partial derivatives into a matrix. We define the $n \times k$ matrix $\boldsymbol{G}(\boldsymbol{y}, \boldsymbol{\theta})$ so as to have typical element

$$G_{ti}(\boldsymbol{y}^t, \boldsymbol{\theta}) \equiv \frac{\partial \ell_t(\boldsymbol{y}^t, \boldsymbol{\theta})}{\partial \theta_i}. \tag{3.26}$$

This matrix is called the **matrix of contributions to the gradient**, because

$$g_i(\boldsymbol{y}, \boldsymbol{\theta}) = \sum_{t=1}^{n} G_{ti}(\boldsymbol{y}^t, \boldsymbol{\theta}). \tag{3.27}$$

Thus each element of the gradient vector is the sum of the elements of one of the columns of the matrix $\boldsymbol{G}(\boldsymbol{y}, \boldsymbol{\theta})$.

A crucial property of the matrix $\boldsymbol{G}(\boldsymbol{y}, \boldsymbol{\theta})$ is that, if $\boldsymbol{y}$ is generated by the DGP characterized by $\boldsymbol{\theta}$, then the expectations of all the elements of the matrix, evaluated at $\boldsymbol{\theta}$, are zero. This result is a consequence of the fact that all

densities integrate to 1. Since ℓ_t is the log of the density f_t of y_t conditional on $\boldsymbol{y}^{t-1}$, we see that, for all t and for all $\boldsymbol{\theta}$,

$$\int \exp\bigl(\ell_t(\boldsymbol{y}^t, \boldsymbol{\theta})\bigr)\, dy_t = \int f_t(\boldsymbol{y}^t, \boldsymbol{\theta})\, dy_t = 1,$$

where the integral is over the support of y_t. Since this relation holds *identically* in $\boldsymbol{\theta}$, we can differentiate it with respect to the components of $\boldsymbol{\theta}$ and obtain a further set of identities. Under weak regularity conditions, it can be shown that the derivatives of the integral on the left-hand side are the integrals of the derivatives of the integrand. Thus, since the derivative of the constant 1 is 0, we have, identically in $\boldsymbol{\theta}$ and for $i = 1, \ldots, k$,

$$\int \exp\bigl(\ell_t(\boldsymbol{y}^t, \boldsymbol{\theta})\bigr) \frac{\partial \ell_t(\boldsymbol{y}^t, \boldsymbol{\theta})}{\partial \theta_i}\, dy_t = 0. \tag{3.28}$$

Since $\exp(\ell_t(\boldsymbol{y}^t, \boldsymbol{\theta}))$ is, for the DGP characterized by $\boldsymbol{\theta}$, the density of y_t conditional on $\boldsymbol{y}^{t-1}$, this last equation, along with the definition (3.26), gives

$$\mathrm{E}_{\boldsymbol{\theta}}\bigl(G_{ti}(\boldsymbol{y}^t, \boldsymbol{\theta}) \,|\, \boldsymbol{y}^{t-1}\bigr) = 0 \tag{3.29}$$

for all $t = 1, \ldots, n$ and $i = 1, \ldots, k$. The notation "$\mathrm{E}_{\boldsymbol{\theta}}$" here means that the expectation is being taken under the DGP characterized by $\boldsymbol{\theta}$. Taking unconditional expectations of (3.29) yields the desired result. Summing (3.29) over $t = 1, \ldots, n$ shows that $\mathrm{E}_{\boldsymbol{\theta}}(g_i(\boldsymbol{y}, \boldsymbol{\theta})) = 0$ for $i = 1, \ldots, k$, or, equivalently, that $\mathrm{E}_{\boldsymbol{\theta}}(\boldsymbol{g}(\boldsymbol{y}, \boldsymbol{\theta})) = \boldsymbol{0}$.

In addition to the conditional expectations of the elements of the matrix $\boldsymbol{G}(\boldsymbol{y}, \boldsymbol{\theta})$, we can compute the covariances of these elements. Let $t \neq s$, and suppose, without loss of generality, that $t < s$. Then the covariance under the DGP characterized by $\boldsymbol{\theta}$ of the ti^{th} and sj^{th} elements of $\boldsymbol{G}(\boldsymbol{y}, \boldsymbol{\theta})$ is

$$\begin{aligned} \mathrm{E}_{\boldsymbol{\theta}}\bigl(G_{ti}(\boldsymbol{y}^t, \boldsymbol{\theta}) G_{sj}(\boldsymbol{y}^s, \boldsymbol{\theta})\bigr) &= \mathrm{E}_{\boldsymbol{\theta}}\Bigl(\mathrm{E}_{\boldsymbol{\theta}}\bigl(G_{ti}(\boldsymbol{y}^t, \boldsymbol{\theta}) G_{sj}(\boldsymbol{y}^s, \boldsymbol{\theta})\bigr) \,|\, \boldsymbol{y}^t\Bigr) \\ &= \mathrm{E}_{\boldsymbol{\theta}}\Bigl(G_{ti}(\boldsymbol{y}^t, \boldsymbol{\theta}) \mathrm{E}_{\boldsymbol{\theta}}\bigl(G_{sj}(\boldsymbol{y}^s, \boldsymbol{\theta}) \,|\, \boldsymbol{y}^t\bigr)\Bigr) = 0. \end{aligned} \tag{3.30}$$

The step leading to the second line above follows because $G_{ti}(\cdot)$ is a deterministic function of $\boldsymbol{y}^t$, and the last step follows because the expectation of $G_{sj}(\cdot)$ is zero conditional on $\boldsymbol{y}^{s-1}$, by (3.29), and so also conditional on the subvector $\boldsymbol{y}^t$ of $\boldsymbol{y}^{s-1}$. The above proof shows that the covariance of the two matrix elements is also zero conditional on $\boldsymbol{y}^t$.

The Information Matrix and the Hessian

The covariance matrix of the elements of the t^{th} row $\boldsymbol{G}_t(\boldsymbol{y}^t, \boldsymbol{\theta})$ of $\boldsymbol{G}(\boldsymbol{y}, \boldsymbol{\theta})$ is the $k \times k$ matrix $\boldsymbol{I}_t(\boldsymbol{\theta})$, of which the ij^{th} element is $\mathrm{E}_{\boldsymbol{\theta}}(G_{ti}(\boldsymbol{y}^t, \boldsymbol{\theta}) G_{tj}(\boldsymbol{y}^t, \boldsymbol{\theta}))$.

As a covariance matrix, $\boldsymbol{I}_t(\boldsymbol{\theta})$ is normally positive definite. The sum of the matrices $\boldsymbol{I}_t(\boldsymbol{\theta})$ over all t is the $k \times k$ matrix

$$\boldsymbol{I}(\boldsymbol{\theta}) \equiv \sum_{t=1}^{n} \boldsymbol{I}_t(\boldsymbol{\theta}) = \sum_{t=1}^{n} \mathrm{E}_{\boldsymbol{\theta}}\big(\boldsymbol{G}_t^\top(\boldsymbol{y}^t, \boldsymbol{\theta})\boldsymbol{G}_t(\boldsymbol{y}^t, \boldsymbol{\theta})\big), \tag{3.31}$$

which is called the **information matrix**. The matrices $\boldsymbol{I}_t(\boldsymbol{\theta})$ are the **contributions** to the information matrix made by the successive observations.

An equivalent definition of the information matrix, as readers are invited to show in Exercise 3.5, is $\boldsymbol{I}(\boldsymbol{\theta}) \equiv \mathrm{E}_{\boldsymbol{\theta}}(\boldsymbol{g}(\boldsymbol{y}, \boldsymbol{\theta})\boldsymbol{g}^\top(\boldsymbol{y}, \boldsymbol{\theta}))$. In this second form, the information matrix is the expectation of the **outer product of the gradient** with itself; see Part 1, Section 2.4 for the definition of the outer product of two vectors. Less exotically, it is just the covariance matrix of the score vector. As the name suggests, and as we will see shortly, the information matrix is a measure of the total amount of information about the parameters in the sample. The requirement that it should be positive definite is a condition for **strong asymptotic identification** of those parameters, in the same sense as the strong asymptotic identification condition introduced in Section 1.2 for nonlinear regression models.

Closely related to (3.31) is the **asymptotic information matrix**

$$\mathcal{I}(\boldsymbol{\theta}) \equiv \operatorname*{plim}_{n\to\infty,\,\boldsymbol{\theta}} \frac{1}{n}\boldsymbol{I}(\boldsymbol{\theta}), \tag{3.32}$$

which measures the average amount of information about the parameters that is contained in the observations of the sample. As with the notation $\mathrm{E}_{\boldsymbol{\theta}}$, we use $\operatorname{plim}_{\boldsymbol{\theta}}$ to denote the plim under the DGP characterized by $\boldsymbol{\theta}$.

We have already defined the Hessian $\boldsymbol{H}(\boldsymbol{y}, \boldsymbol{\theta})$. For asymptotic analysis, we are generally more interested in the **asymptotic Hessian**,

$$\mathcal{H}(\boldsymbol{\theta}) \equiv \operatorname*{plim}_{n\to\infty,\,\boldsymbol{\theta}} \frac{1}{n}\boldsymbol{H}(\boldsymbol{y}, \boldsymbol{\theta}), \tag{3.33}$$

than in $\boldsymbol{H}(\boldsymbol{y}, \boldsymbol{\theta})$ itself. The asymptotic Hessian is related to the ordinary Hessian in exactly the same way as the asymptotic information matrix is related to the ordinary information matrix; compare (3.32) and (3.33).

There is a very important relationship between the asymptotic information matrix and the asymptotic Hessian. One version of this relationship, which is called the **information matrix equality**, is

$$\mathcal{I}(\boldsymbol{\theta}) = -\mathcal{H}(\boldsymbol{\theta}). \tag{3.34}$$

Both the Hessian and the information matrix measure the amount of curvature in the loglikelihood function. Although they are both measuring the same thing, the Hessian is negative definite, at least in the neighborhood of $\hat{\boldsymbol{\theta}}$, while the information matrix is always positive definite; that is why there is a minus sign in (3.34). The proof of (3.34) is the subject of Exercises 3.6 and 3.7. It depends critically on the assumption that the DGP is a special case of the model being estimated.

Asymptotic Normality of the MLE

In order for it to be asymptotically normally distributed, a maximum likelihood estimator must be a Type 2 MLE. In addition, it must satisfy certain regularity conditions, which are discussed in Davidson and MacKinnon (1993, Section 8.5). The Type 2 requirement arises because the proof of asymptotic normality is based on the likelihood equations (3.14), which apply only to Type 2 estimators.

The first step in the proof is to perform a Taylor expansion of the likelihood equations (3.14) around $\boldsymbol{\theta}_0$. This expansion yields

$$\boldsymbol{g}(\hat{\boldsymbol{\theta}}) = \boldsymbol{g}(\boldsymbol{\theta}_0) + \boldsymbol{H}(\bar{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = \boldsymbol{0}, \tag{3.35}$$

where we suppress the dependence on $\boldsymbol{y}$ for notational simplicity. The notation $\bar{\boldsymbol{\theta}}$ is our usual shorthand notation for Taylor expansions of vector expressions; see (1.21) and the subsequent discussion. We may therefore write

$$\big\|\bar{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\big\| \leq \big\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\big\|.$$

The fact that the ML estimator $\hat{\boldsymbol{\theta}}$ is consistent then implies that $\bar{\boldsymbol{\theta}}$ is also consistent.

If we solve (3.35) and insert the factors of powers of n that are needed for asymptotic analysis, we obtain the result that

$$n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = -\big(n^{-1}\boldsymbol{H}(\bar{\boldsymbol{\theta}})\big)^{-1}\big(n^{-1/2}\boldsymbol{g}(\boldsymbol{\theta}_0)\big). \tag{3.36}$$

Because $\bar{\boldsymbol{\theta}}$ is consistent, the matrix $n^{-1}\boldsymbol{H}(\bar{\boldsymbol{\theta}})$ which appears in (3.36) must tend to the same nonstochastic limiting matrix as $n^{-1}\boldsymbol{H}(\boldsymbol{\theta}_0)$, namely, $\mathcal{H}(\boldsymbol{\theta}_0)$. Therefore, equation (3.36) implies that

$$n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \stackrel{a}{=} -\mathcal{H}^{-1}(\boldsymbol{\theta}_0)n^{-1/2}\boldsymbol{g}(\boldsymbol{\theta}_0). \tag{3.37}$$

If the information matrix equality, equation (3.34), holds, then this result can equivalently be written as

$$n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \stackrel{a}{=} \mathcal{I}^{-1}(\boldsymbol{\theta}_0)n^{-1/2}\boldsymbol{g}(\boldsymbol{\theta}_0). \tag{3.38}$$

Since the information matrix equality holds only if the model is correctly specified, (3.38) is not in general valid for misspecified models.

The asymptotic normality of the Type 2 MLE follows immediately from the asymptotic equalities (3.37) or (3.38) if it can be shown that the vector $n^{-1/2}\boldsymbol{g}(\boldsymbol{\theta}_0)$ is asymptotically distributed as multivariate normal. As can be seen from (3.27), each element $n^{-1/2}g_i(\boldsymbol{\theta}_0)$ of this vector is $n^{-1/2}$ times a sum of n random variables, each of which has expectation 0, by (3.29). These random variables are mutually uncorrelated, by the result (3.30). Under standard regularity conditions, with which we will not concern ourselves,

a multivariate central limit theorem can therefore be applied to this vector. For finite n, the covariance matrix of the score vector is, by definition, the information matrix $\boldsymbol{I}(\boldsymbol{\theta}_0)$. Thus the covariance matrix of the vector $n^{-1/2}\boldsymbol{g}(\boldsymbol{\theta}_0)$ is $n^{-1}\boldsymbol{I}(\boldsymbol{\theta}_0)$, of which, by (3.32), the limit as $n \to \infty$ is the asymptotic information matrix $\mathcal{I}(\boldsymbol{\theta}_0)$. It follows that

$$n^{-1/2}\boldsymbol{g}(\boldsymbol{\theta}_0) \xrightarrow{d} \mathrm{N}\big(\mathbf{0}, \mathcal{I}(\boldsymbol{\theta}_0)\big). \tag{3.39}$$

This result, when combined with (3.37) or (3.38), implies that the Type 2 MLE is asymptotically normally distributed.

3.4 The Covariance Matrix of the ML Estimator

For Type 2 ML estimators, we can obtain the asymptotic distribution of the estimator by combining the result (3.39) for the asymptotic distribution of $n^{-1/2}\boldsymbol{g}(\boldsymbol{\theta}_0)$ with the result (3.37). The asymptotic distribution of the estimator is the distribution to which $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ converges in distribution. This distribution is normal, with expectation vector zero and covariance matrix

$$\mathcal{H}^{-1}(\boldsymbol{\theta}_0)\mathcal{I}(\boldsymbol{\theta}_0)\mathcal{H}^{-1}(\boldsymbol{\theta}_0), \tag{3.40}$$

which has the form of a sandwich covariance matrix. When the information matrix equality, equation (3.34), holds, the sandwich simplifies to $\mathcal{I}^{-1}(\boldsymbol{\theta}_0)$. Thus the asymptotic information matrix is seen to be the asymptotic precision matrix of a Type 2 ML estimator. This shows why the matrices $\boldsymbol{I}$ and $\mathcal{I}$ are called *information* matrices of various sorts.

Clearly, any method that allows us to estimate $\mathcal{I}(\boldsymbol{\theta}_0)$ consistently can be used to estimate the covariance matrix of the ML estimates. In fact, several different methods are widely used, because each has advantages in certain situations.

The first method is just to use minus the inverse of the Hessian, evaluated at the vector of ML estimates. Because these estimates are consistent, it is valid to evaluate the Hessian at $\hat{\boldsymbol{\theta}}$ rather than at $\boldsymbol{\theta}_0$. This yields the estimator

$$\widehat{\mathrm{Var}}_{\mathrm{H}}(\hat{\boldsymbol{\theta}}) = -\boldsymbol{H}^{-1}(\hat{\boldsymbol{\theta}}), \tag{3.41}$$

which is referred to as the **empirical Hessian** estimator. Notice that, since it is the covariance matrix of $\hat{\boldsymbol{\theta}}$ in which we are interested, the factor of $n^{1/2}$ is no longer present. This estimator is easy to obtain whenever Newton's Method, or some sort of quasi-Newton method that uses second derivatives, is used to maximize the loglikelihood function. In the case of quasi-Newton methods, $\boldsymbol{H}(\hat{\boldsymbol{\theta}})$ may sometimes be replaced by another matrix that approximates it. Provided that n^{-1} times the approximating matrix converges to $\mathcal{H}(\boldsymbol{\theta})$, this sort of replacement is asymptotically valid.

Although the empirical Hessian estimator often works well, it does not use all the information we have about the model. Especially for simpler models, we may actually be able to find an analytic expression for $\boldsymbol{I}(\boldsymbol{\theta})$. If so, we can use the inverse of $\boldsymbol{I}(\boldsymbol{\theta})$, evaluated at the ML estimates. This yields the **information matrix**, or **IM**, estimator

$$\widehat{\mathrm{Var}}_{\mathrm{IM}}(\hat{\boldsymbol{\theta}}) = \boldsymbol{I}^{-1}(\hat{\boldsymbol{\theta}}). \tag{3.42}$$

The advantage of this estimator is that it normally involves fewer random terms than does the empirical Hessian, and it may therefore be somewhat more efficient in finite samples. In the case of the classical normal linear model, to be discussed below, it is not at all difficult to obtain $\boldsymbol{I}(\boldsymbol{\theta})$, and the information matrix estimator is therefore the one that is normally used.

The third method is based on (3.31), from which we see that

$$\boldsymbol{I}(\boldsymbol{\theta}_0) = \mathrm{E}\big(\boldsymbol{G}^\top(\boldsymbol{\theta}_0)\boldsymbol{G}(\boldsymbol{\theta}_0)\big).$$

We can therefore estimate $n^{-1}\boldsymbol{I}(\boldsymbol{\theta}_0)$ consistently by $n^{-1}\boldsymbol{G}^\top(\hat{\boldsymbol{\theta}})\boldsymbol{G}(\hat{\boldsymbol{\theta}})$. The corresponding estimator of the covariance matrix, which is usually called the **outer-product-of-the-gradient**, or **OPG**, estimator, is

$$\widehat{\mathrm{Var}}_{\mathrm{OPG}}(\hat{\boldsymbol{\theta}}) = \big(\boldsymbol{G}^\top(\hat{\boldsymbol{\theta}})\boldsymbol{G}(\hat{\boldsymbol{\theta}})\big)^{-1}. \tag{3.43}$$

The OPG estimator has the advantage of being very easy to calculate. Unlike the empirical Hessian, it depends solely on first derivatives. Unlike the IM estimator, it requires no theoretical calculations. However, it tends to be less reliable in finite samples than either of the other two. The OPG estimator is sometimes called the **BHHH** estimator, because it was advocated by Berndt, Hall, Hall, and Hausman (1974) in a very well-known paper.

In practice, the estimators (3.41), (3.42), and (3.43) are all commonly used to estimate the covariance matrix of ML estimates, but many other estimators are available for particular models. Often, it may be difficult to obtain $\boldsymbol{I}(\boldsymbol{\theta})$, but not difficult to obtain another matrix that approximates it asymptotically, by starting either from the matrix $-\boldsymbol{H}(\boldsymbol{\theta})$ or from the matrix $\boldsymbol{G}^\top(\boldsymbol{\theta})\boldsymbol{G}(\boldsymbol{\theta})$ and taking expectations of some elements.

A fourth covariance matrix estimator, which follows directly from (3.40), is the **sandwich estimator**

$$\widehat{\mathrm{Var}}_{\mathrm{S}}(\hat{\boldsymbol{\theta}}) = \boldsymbol{H}^{-1}(\hat{\boldsymbol{\theta}})\boldsymbol{G}^\top(\hat{\boldsymbol{\theta}})\boldsymbol{G}(\hat{\boldsymbol{\theta}})\boldsymbol{H}^{-1}(\hat{\boldsymbol{\theta}}). \tag{3.44}$$

In normal circumstances, this estimator has little to recommend it. It is harder to compute than the OPG estimator and can be just as unreliable in finite samples. However, unlike the other three estimators, it is valid even when the information matrix equality does not hold. Since this equality generally fails

to hold when the model is misspecified, it may be desirable to compute (3.44) and compare it with the other estimators.

When an ML estimator is applied to a model which is misspecified in ways that do not affect the consistency of the estimator, it is said to be a **quasi-ML estimator**, or **QMLE**; see White (1982) and Gouriéroux, Monfort, and Trognon (1984). In general, the sandwich covariance matrix estimator (3.44) is valid for QML estimators, but the other covariance matrix estimators, which depend on the information matrix equality, are not valid. At least, they are not valid for all the parameters. We have seen that the ML estimator for a regression model with normal disturbances is just the OLS estimator. But we know that the latter is consistent under conditions which do not require normality. If the disturbances are not normal, therefore, the ML estimator is a QMLE. One consequence of this fact is explored in Exercise 3.8.

The Classical Normal Linear Model

It should help to make the theoretical results just discussed clearer if we apply them to the classical normal linear model. We will therefore discuss various ways of estimating the covariance matrix of the ML estimates $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}$ jointly. Of course, we saw in Section 3.4 how to estimate the covariance matrix of $\hat{\boldsymbol{\beta}}$ by itself, but we have not yet discussed how to estimate the variance of $\hat{\sigma}$.

For the classical normal linear model, the contribution to the loglikelihood function made by the t^{th} observation is given by expression (3.09). There are $k+1$ parameters. The first k of them are the elements of the vector $\boldsymbol{\beta}$, and the last one is σ. A typical element of any of the first k columns of the matrix $\boldsymbol{G}$, indexed by i, is

$$G_{ti}(\boldsymbol{\beta}, \sigma) = \frac{\partial \ell_t}{\partial \beta_i} = \frac{1}{\sigma^2}(y_t - \boldsymbol{X}_t\boldsymbol{\beta})x_{ti}, \quad i = 1, \dots, k, \tag{3.45}$$

and a typical element of the last column is

$$G_{t,k+1}(\boldsymbol{\beta}, \sigma) = \frac{\partial \ell_t}{\partial \sigma} = -\frac{1}{\sigma} + \frac{1}{\sigma^3}(y_t - \boldsymbol{X}_t\boldsymbol{\beta})^2. \tag{3.46}$$

These two equations give us everything we need to calculate the information matrix.

For $i, j = 1, \dots, k$, the ij^{th} element of $\boldsymbol{G}^\top\boldsymbol{G}$ is

$$\sum_{t=1}^{n} \frac{1}{\sigma^4}(y_t - \boldsymbol{X}_t\boldsymbol{\beta})^2 x_{ti}x_{tj}. \tag{3.47}$$

This is just the sum over all t of $G_{ti}(\boldsymbol{\beta}, \sigma)$ times $G_{tj}(\boldsymbol{\beta}, \sigma)$ as defined in (3.45). When we evaluate at the true values of $\boldsymbol{\beta}$ and σ, we have that $y_t - \boldsymbol{X}_t\boldsymbol{\beta} = u_t$

and $\mathrm{E}(u_t^2) = \sigma^2$, and so the expectation of this matrix element is easily seen to be

$$\sum_{t=1}^{n} \frac{1}{\sigma^2} x_{ti}x_{tj}. \tag{3.48}$$

In matrix notation, the whole $\boldsymbol{\beta}$-$\boldsymbol{\beta}$ block of $\boldsymbol{G}^\top\boldsymbol{G}$ has expectation $\boldsymbol{X}^\top\boldsymbol{X}/\sigma^2$. The $(i, k+1)^{\text{th}}$ element of $\boldsymbol{G}^\top\boldsymbol{G}$ is

$$\begin{aligned} &\sum_{t=1}^{n} \left(-\frac{1}{\sigma} + \frac{1}{\sigma^3}(y_t - \boldsymbol{X}_t\boldsymbol{\beta})^2\right)\left(\frac{1}{\sigma^2}(y_t - \boldsymbol{X}_t\boldsymbol{\beta})x_{ti}\right) \\ &= -\sum_{t=1}^{n} \frac{1}{\sigma^3}(y_t - \boldsymbol{X}_t\boldsymbol{\beta})x_{ti} + \sum_{t=1}^{n} \frac{1}{\sigma^5}(y_t - \boldsymbol{X}_t\boldsymbol{\beta})^3 x_{ti}. \end{aligned} \tag{3.49}$$

This is the sum over all t of the product of expressions (3.45) and (3.46). We know that $\mathrm{E}(u_t) = 0$, and, if the disturbances u_t are normal, we also know that $\mathrm{E}(u_t^3) = 0$. Consequently, the expectation of this sum is 0. This result depends critically on the assumption, following from normality, that the distribution of the disturbances is symmetric around zero. For a skewed distribution, the third moment would be nonzero, and (3.49) would therefore not have expectation 0.

Finally, the $(k+1), (k+1)^{\text{th}}$ element of $\boldsymbol{G}^\top\boldsymbol{G}$ is

$$\begin{aligned} &\sum_{t=1}^{n} \left(-\frac{1}{\sigma} + \frac{1}{\sigma^3}(y_t - \boldsymbol{X}_t\boldsymbol{\beta})^2\right)^2 \\ &= \frac{n}{\sigma^2} - \sum_{t=1}^{n} \frac{2}{\sigma^4}(y_t - \boldsymbol{X}_t\boldsymbol{\beta})^2 + \sum_{t=1}^{n} \frac{1}{\sigma^6}(y_t - \boldsymbol{X}_t\boldsymbol{\beta})^4. \end{aligned} \tag{3.50}$$

This is the sum over all t of the square of expression (3.46). To compute its expectation, we replace $y_t - \boldsymbol{X}_t\boldsymbol{\beta}$ by u_t and use the result that $\mathrm{E}(u_t^4) = 3\sigma^4$; see Exercise 4.F2. It is then not hard to see that expression (3.50) has expectation $2n/\sigma^2$. Once more, this result depends crucially on the normality assumption. If the kurtosis of the disturbances were greater (or less) than that of the normal distribution, the expectation of expression (3.50) would be larger (or smaller) than $2n/\sigma^2$.

Putting the results (3.48), (3.49), and (3.50) together, the asymptotic information matrix for $\boldsymbol{\beta}$ and σ jointly is seen to be

$$\mathcal{I}(\boldsymbol{\beta}, \sigma) = \operatorname*{plim}_{n\to\infty} \begin{bmatrix} n^{-1}\boldsymbol{X}^\top\boldsymbol{X}/\sigma^2 & \boldsymbol{0} \\ \boldsymbol{0}^\top & 2/\sigma^2 \end{bmatrix}. \tag{3.51}$$

Inverting this matrix, multiplying the inverse by n^{-1}, and replacing σ by $\hat{\sigma}$, we find that the IM estimator of the covariance matrix of all the parameter estimates is

$$\widehat{\mathrm{Var}}_{\mathrm{IM}}(\hat{\boldsymbol{\beta}}, \hat{\sigma}) = \begin{bmatrix} \hat{\sigma}^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1} & \boldsymbol{0} \\ \boldsymbol{0}^\top & \hat{\sigma}^2/2n \end{bmatrix}. \tag{3.52}$$

The upper left-hand block of this matrix would be the familiar OLS covariance matrix if we had used s instead of $\hat{\sigma}$ to estimate σ. The lower right-hand element is the approximate variance of $\hat{\sigma}$ under the assumption of normally distributed disturbances. If we had treated σ^2 instead of σ as a parameter, the lower right-hand element would have been different; see Exercise 3.10.

It is noteworthy that the information matrix (3.51), and therefore also the estimated covariance matrix (3.52), are block-diagonal. This implies that there is no covariance between $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}$. This is a property of all regression models, nonlinear as well as linear, with normal disturbances, and it is responsible for much of the simplicity of these models. The block-diagonality of the information matrix means that we can make inferences about $\boldsymbol{\beta}$ without taking account of the fact that σ has also been estimated, and we can make inferences about σ without taking account of the fact that $\boldsymbol{\beta}$ has also been estimated. If the information matrix were not block-diagonal, which in most other cases it is not, it would have been necessary to invert the entire matrix in order to obtain any block of the inverse.

Asymptotic Efficiency of the ML Estimator

A Type 2 ML estimator must be at least as asymptotically efficient as any other root-n consistent estimator that is asymptotically unbiased.[4] Therefore, at least in large samples, maximum likelihood estimation possesses an optimality property that is generally not shared by other estimation methods. We will not attempt to prove this result here; see Davidson and MacKinnon (1993, Section 8.8). However, we will discuss it briefly.

Let the MLE be denoted as $\tilde{\boldsymbol{\theta}}$ and consider any other root-n consistent and asymptotically unbiased estimator, say $\hat{\boldsymbol{\theta}}$. It can be shown that

$$n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = n^{1/2}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + \boldsymbol{v}, \tag{3.53}$$

where $\boldsymbol{v}$ is a random k–vector the expectation of which tends to zero asymptotically and which is asymptotically uncorrelated with the vector $n^{1/2}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$. We can rewrite (3.53) as

$$n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) - n^{1/2}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) - \boldsymbol{v} = \mathbf{0},$$

which shows that the distributions of $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ and $n^{1/2}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + \boldsymbol{v}$ are the same. This remains true of the normal limiting distributions as $n \to \infty$, and this lets us conclude that the asymptotic covariance matrix of $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ is equal to the asymptotic covariance matrix of $n^{1/2}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ plus the asymptotic

[4] All of the root-n consistent estimators that we have discussed are also asymptotically unbiased. However, as is discussed in Davidson and MacKinnon (1993, Section 4.5), it is possible for such an estimator to be asymptotically biased, and we must therefore rule out this possibility explicitly.

covariance matrix of $\boldsymbol{v}$, which is a positive semidefinite matrix. It follows that the asymptotic covariance matrix of the estimator $\hat{\boldsymbol{\theta}}$ must be larger than that of $\tilde{\boldsymbol{\theta}}$, in the usual sense.

The result (3.53) bears a strong, and by no means coincidental, resemblance to a result that we used in Part 1, Section 4.5 when proving the Gauss-Markov Theorem. This result says that, in the context of the linear regression model, any unbiased linear estimator can be written as the sum of the OLS estimator and a random component which has expectation zero and is uncorrelated with the OLS estimator. Asymptotically, equation (3.53) says essentially the same thing in the context of a very much broader class of models. The key property of (3.53) is that $\boldsymbol{v}$ is asymptotically uncorrelated with $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$. Therefore, the random vector $\boldsymbol{v}$ simply adds additional noise to the ML estimator.

The asymptotic efficiency result above is really an asymptotic version of the **Cramér-Rao lower bound**,[5] which actually applies to any unbiased estimator, regardless of sample size. It states that the covariance matrix of such an estimator can never be smaller than $\boldsymbol{I}^{-1}$, which, as we have seen, is asymptotically equal to the covariance matrix of the ML estimator. Readers are guided through the proof of this classical result in Exercise 3.12. However, since ML estimators are not in general unbiased, it is only the asymptotic version of the bound that is of interest in the context of ML estimation.

The fact that ML estimators attain the Cramér-Rao lower bound asymptotically is one of their many attractive features. However, like the Gauss-Markov Theorem, this result must be interpreted with caution. First of all, it is true only asymptotically. ML estimators may or may not perform well in samples of moderate size. Secondly, there may well exist an asymptotically biased estimator that is more efficient, in the sense of finite-sample mean squared error, than any given ML estimator. For example, the estimator obtained by imposing a restriction that is false, but not grossly incompatible with the data, may well be more efficient than the unrestricted ML estimator. The former cannot be more efficient asymptotically, because the variance of both estimators tends to zero as the sample size tends to infinity and the bias of the biased estimator does not, but it can be more efficient in finite samples.

3.5 Hypothesis Testing

Maximum likelihood estimation offers three different procedures for performing hypothesis tests, two of which usually have several different variants. These three procedures, which are collectively referred to as the **three classical tests**, are the **likelihood ratio**, **Wald**, and **Lagrange multiplier** tests. All three tests are asymptotically equivalent, in the sense that the differences between

[5] This bound was originally suggested by Fisher (1925) and later stated in its modern form by Cramér (1946) and Rao (1945).

any pair of the three tend in probability to zero (under the null hypothesis, and for DGPs that are "close" to the null hypothesis) as the sample size tends to infinity. If the number of equality restrictions is r, the limiting distribution is $\chi^2(r)$. We have already discussed Wald tests in Part 1, Sections 5.6, Part 1, 6.3, and and Part 1, 8.5, but we have not yet encountered the other two classical tests, at least not under their usual names.

As we remarked in Part 1, Section 5.2, a hypothesis in econometrics corresponds to a model. We let the model that corresponds to the alternative hypothesis be characterized by the loglikelihood function $\ell(\boldsymbol{\theta})$. Then the null hypothesis imposes r restrictions, which are in general nonlinear, on $\boldsymbol{\theta}$. We write these as $\boldsymbol{r}(\boldsymbol{\theta}) = \mathbf{0}$, where $\boldsymbol{r}(\boldsymbol{\theta})$ is an r–vector of smooth functions of the parameters. Thus the null hypothesis is represented by the model with loglikelihood $\ell(\boldsymbol{\theta})$, where the parameter space is restricted to those values of $\boldsymbol{\theta}$ that satisfy the restrictions $\boldsymbol{r}(\boldsymbol{\theta}) = \mathbf{0}$.

Likelihood Ratio Tests

The **likelihood ratio**, or **LR**, test is the simplest of the three classical tests. The test statistic is just twice the difference between the unconstrained maximum value of the loglikelihood function and the maximum subject to the restrictions. Thus the likelihood ratio statistic is just

$$\text{LR} = 2\big(\ell(\hat{\boldsymbol{\theta}}) - \ell(\tilde{\boldsymbol{\theta}})\big), \tag{3.54}$$

where $\tilde{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}$ denote, respectively, the restricted and unrestricted maximum likelihood estimates of $\boldsymbol{\theta}$. The LR statistic gets its name from the fact that the right-hand side of equation (3.54) is equal to

$$2 \log\left(\frac{L(\hat{\boldsymbol{\theta}})}{L(\tilde{\boldsymbol{\theta}})}\right),$$

or twice the logarithm of the ratio of the likelihood functions. One of its most attractive features is that the LR statistic is trivially easy to compute when both the restricted and unrestricted estimates are available. Whenever we impose, or relax, some restrictions on a model, twice the change in the value of the loglikelihood function provides immediate feedback on whether the restrictions are compatible with the data.

Precisely why the LR statistic is asymptotically distributed as $\chi^2(r)$ is not entirely obvious, and we will not attempt to explain it now. The asymptotic theory of the three classical tests will be discussed in detail in the next section. Some intuition can be gained by looking at the LR test for linear restrictions on the classical normal linear model. The LR statistic turns out to be closely related to the familiar F statistic, which can be written as

$$F = \frac{\big(\text{SSR}(\tilde{\boldsymbol{\beta}}) - \text{SSR}(\hat{\boldsymbol{\beta}})\big)/r}{\text{SSR}(\hat{\boldsymbol{\beta}})/(n-k)}, \tag{3.55}$$

where $\hat{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\beta}}$ are the unrestricted and restricted OLS (and hence also ML) estimates, respectively. The LR statistic can also be expressed in terms of the two sums of squared residuals, by use of the formula (3.12), which gives the maximized loglikelihood in terms of the minimized SSR. The statistic is

$$\begin{aligned} 2\big(\ell(\hat{\boldsymbol{\theta}}) - \ell(\tilde{\boldsymbol{\theta}})\big) &= 2\Big(\tfrac{n}{2}\log \text{SSR}(\tilde{\boldsymbol{\beta}}) - \tfrac{n}{2}\log \text{SSR}(\hat{\boldsymbol{\beta}})\Big) \\ &= n \log\left(\frac{\text{SSR}(\tilde{\boldsymbol{\beta}})}{\text{SSR}(\hat{\boldsymbol{\beta}})}\right). \end{aligned} \tag{3.56}$$

We can rewrite the last expression here as

$$n \log\left(1 + \frac{\text{SSR}(\tilde{\boldsymbol{\beta}}) - \text{SSR}(\hat{\boldsymbol{\beta}})}{\text{SSR}(\hat{\boldsymbol{\beta}})}\right) = n \log\left(1 + \frac{r}{n-k}F\right) \cong rF.$$

The approximate equality above follows from the facts that $n/(n-k) \stackrel{a}{=} 1$ and that $\log(1+a) \cong a$ whenever a is small. Under the null hypothesis, $\text{SSR}(\tilde{\boldsymbol{\beta}})$ should not be much larger than $\text{SSR}(\hat{\boldsymbol{\beta}})$, or, equivalently, $F/(n-k)$ should be a small quantity. Thus this approximation should generally be a good one. In fact, under the null hypothesis, the LR statistic (3.56) is asymptotically equal to r times the F statistic. Whether or not the null is true, the LR statistic is a deterministic, strictly increasing, function of the F statistic. As we will see later, this fact has important consequences if the statistics are bootstrapped. Without bootstrapping, it makes little sense to use an LR test rather than an F test in the context of the classical normal linear model, because the latter, but not the former, is exact in finite samples.

Wald Tests

Unlike LR tests, Wald tests depend only on the estimates of the unrestricted model. There is no real difference between Wald tests in models estimated by maximum likelihood and those in models estimated by other methods; see Section 1.7. As with the LR test, we wish to test the r restrictions $\boldsymbol{r}(\boldsymbol{\theta}) = \mathbf{0}$. The Wald test statistic is just a quadratic form in the vector $\boldsymbol{r}(\hat{\boldsymbol{\theta}})$ and the inverse of a matrix that estimates its covariance matrix.

By using the delta method (Part 1, Section 6.8), we find that

$$\widehat{\text{Var}}\big(\boldsymbol{r}(\hat{\boldsymbol{\theta}})\big) = \boldsymbol{R}(\boldsymbol{\theta}_0)\widehat{\text{Var}}(\hat{\boldsymbol{\theta}})\boldsymbol{R}^\top(\boldsymbol{\theta}_0), \tag{3.57}$$

where $\boldsymbol{R}(\boldsymbol{\theta})$ is an $r \times k$ matrix with typical element $\partial r_i(\boldsymbol{\theta})/\partial\theta_j$, and $\widehat{\text{Var}}(\hat{\boldsymbol{\theta}})$ is any one of the estimators of $\text{Var}(\hat{\boldsymbol{\theta}})$ that we looked at in the last section. Replacing the unknown $\boldsymbol{\theta}_0$ by $\hat{\boldsymbol{\theta}}$ in (3.57), we find that the Wald statistic is

$$\text{W} = \boldsymbol{r}^\top(\hat{\boldsymbol{\theta}})\big(\boldsymbol{R}(\hat{\boldsymbol{\theta}})\widehat{\text{Var}}(\hat{\boldsymbol{\theta}})\boldsymbol{R}^\top(\hat{\boldsymbol{\theta}})\big)^{-1}\boldsymbol{r}(\hat{\boldsymbol{\theta}}). \tag{3.58}$$

This is a quadratic form in the r–vector $\boldsymbol{r}(\hat{\boldsymbol{\theta}})$, which is asymptotically multivariate normal, and the inverse of an estimate of its covariance matrix. It is easy to see, using the first part of Part 1, Theorem 5.1, that (3.58) is asymptotically distributed as $\chi^2(r)$ under the null hypothesis. As readers are asked to show in Exercise 3.13, the Wald statistic (1.74) is just a special case of the one defined in (3.58). In the case of linear regression models subject to linear restrictions on the parameters, the Wald statistic is, like the LR statistic, a deterministic, strictly increasing, function of the F statistic if the information matrix estimator (3.42) of the covariance matrix of the parameters is used to construct the Wald statistic.

Wald tests are very widely used, in part because the square of every t statistic is really a Wald statistic. Nevertheless, they should be used with caution. Although Wald tests do not necessarily have poor finite-sample properties, and they do not necessarily perform less well in finite samples than the other classical tests, there is a good deal of evidence that they quite often do so. One reason for this is that Wald statistics are not invariant to reformulations of the restrictions. Some formulations may lead to Wald tests that are well-behaved, but others may lead to tests that severely overreject, or (much less commonly) underreject, in samples of moderate size.

As an example, consider the linear regression model

$$y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + u_t, \tag{3.59}$$

where we wish to test the hypothesis that the product of β_1 and β_2 is 1. To compute a Wald statistic, we need to estimate the covariance matrix of $\hat{\beta}_1$ and $\hat{\beta}_2$. If $\boldsymbol{X}$ denotes the $n \times 2$ matrix with typical element x_{ti}, $i = 1, 2$, and $\boldsymbol{M}_\iota$ is the matrix that takes deviations from the mean, then the IM estimator of this covariance matrix is

$$\widehat{\mathrm{Var}}(\hat{\beta}_1, \hat{\beta}_2) = \hat{\sigma}^2 (\boldsymbol{X}^\top \boldsymbol{M}_\iota \boldsymbol{X})^{-1}; \tag{3.60}$$

we could of course use s^2 instead of $\hat{\sigma}^2$. For notational convenience, we let V_{11}, V_{12} $(= V_{21})$, and V_{22} denote the three distinct elements of this matrix.

There are many ways to write the single restriction on (3.59) that we wish to test. Three formulations that seem particularly natural are

$$\begin{aligned} r_1(\beta_1, \beta_2) &\equiv \beta_1 - 1/\beta_2 = 0, \\ r_2(\beta_1, \beta_2) &\equiv \beta_2 - 1/\beta_1 = 0, \quad \text{and} \\ r_3(\beta_1, \beta_2) &\equiv \beta_1 \beta_2 - 1 = 0. \end{aligned}$$

Each of these ways of writing the restriction leads to a different Wald statistic. If the restriction is written in the form of r_1, then $\boldsymbol{R}(\beta_1, \beta_2) = [1 \quad 1/\beta_2^2]$. Combining this with (3.60), we find after a little algebra that the Wald statistic is

$$\mathrm{W}_1 = \frac{(\hat{\beta}_1 - 1/\hat{\beta}_2)^2}{V_{11} + 2V_{12}/\hat{\beta}_2^2 + V_{22}/\hat{\beta}_2^4}.$$

If instead the restriction is written in the form of r_2, then $\boldsymbol{R}(\beta_1, \beta_2) = [1/\beta_1^2 \quad 1]$, and the Wald statistic is

$$\mathrm{W}_2 = \frac{(\hat{\beta}_2 - 1/\hat{\beta}_1)^2}{V_{11}/\hat{\beta}_1^4 + 2V_{12}/\hat{\beta}_1^2 + V_{22}}.$$

Finally, if the restriction is written in the form of r_3, then $\boldsymbol{R}(\beta_1, \beta_2) = [\beta_2 \quad \beta_1]$, and the Wald statistic is

$$\mathrm{W}_3 = \frac{(\hat{\beta}_1 \hat{\beta}_2 - 1)^2}{\hat{\beta}_2^2 V_{11} + 2\hat{\beta}_1 \hat{\beta}_2 V_{12} + \hat{\beta}_1^2 V_{22}}.$$

In finite samples, these three Wald statistics can be quite different. Depending on the values of β_1 and β_2, any one of them may perform better or worse than the other two, and they can sometimes overreject severely. The performance of alternative Wald tests in models like (3.59) has been investigated by Gregory and Veall (1985, 1987). Other cases in which Wald tests perform very badly are discussed by Lafontaine and White (1986).

Because of their dubious finite-sample properties and their sensitivity to the way in which the restrictions are written, we recommend against using Wald tests when the outcome of a test is important, except when it would be very costly or inconvenient to estimate the restricted model. Asymptotic t statistics should also be used with great caution, since, as we saw in Section 1.7, every asymptotic t statistic is simply the signed square root of a Wald statistic. Because conventional confidence intervals are based on inverting asymptotic t statistics, they too should be used with caution.

Lagrange Multiplier Tests

The **Lagrange multiplier**, or **LM**, test is the third of the three classical tests. The name suggests that it is based on the vector of Lagrange multipliers from a constrained maximization problem. That can indeed be the case. In practice, however, LM tests are very rarely computed in this way. Instead, they are usually based on the gradient vector, or score vector, of the unrestricted loglikelihood function, evaluated at the restricted estimates. LM tests are very often computed by means of artificial regressions. In fact, as we will see, some of the GNR-based tests that we encountered in Section 1.7 and Section 2.7 are essentially Lagrange multiplier tests.

For simplicity, we begin our discussion of LM tests by considering the case in which the restrictions to be tested are zero restrictions, that is, restrictions according to which some of the model parameters are zero. In such cases, the r restrictions can be written as $\boldsymbol{\theta}_2 = \mathbf{0}$, where the parameter vector $\boldsymbol{\theta}$ is partitioned as $\boldsymbol{\theta} = [\boldsymbol{\theta}_1 \vdots \boldsymbol{\theta}_2]$, possibly after some reordering of the elements. The vector $\tilde{\boldsymbol{\theta}}$ of restricted estimates can then be expressed as $\tilde{\boldsymbol{\theta}} = [\tilde{\boldsymbol{\theta}}_1 \vdots \mathbf{0}]$.

The vector $\tilde{\boldsymbol{\theta}}_1$ maximizes the restricted loglikelihood function $\ell(\boldsymbol{\theta}_1, \mathbf{0})$, and so it satisfies the restricted likelihood equations

$$\boldsymbol{g}_1(\tilde{\boldsymbol{\theta}}_1, \mathbf{0}) = \mathbf{0}, \tag{3.61}$$

where $\boldsymbol{g}_1(\cdot)$ is the vector whose components are the $k - r$ partial derivatives of $\ell(\cdot)$ with respect to the elements of $\boldsymbol{\theta}_1$.

The formula (3.38), which gives the asymptotic form of an MLE, can be applied to the estimator $\tilde{\boldsymbol{\theta}}$ when $\boldsymbol{\theta}_2 = \mathbf{0}$. If we partition the true parameter vector $\boldsymbol{\theta}_0$ as $[\boldsymbol{\theta}_1^0 \vdots \mathbf{0}]$, we find that

$$n^{1/2}(\tilde{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1^0) \stackrel{a}{=} \big(\mathcal{I}_{11}(\boldsymbol{\theta}_0)\big)^{-1} n^{-1/2} \boldsymbol{g}_1(\boldsymbol{\theta}_0), \tag{3.62}$$

where $\mathcal{I}_{11}(\cdot)$ is the $(k-r) \times (k-r)$ top left block of the asymptotic information matrix $\mathcal{I}(\cdot)$ of the full unrestricted model. This block is, of course, just the asymptotic information matrix for the restricted model.

When the gradient vector of the unrestricted loglikelihood function is evaluated at the restricted estimates $\tilde{\boldsymbol{\theta}}$, the first $k - r$ elements, which are the elements of the vector $\boldsymbol{g}_1(\tilde{\boldsymbol{\theta}})$, are zero, by equation (3.61). However, the r–vector $\boldsymbol{g}_2(\tilde{\boldsymbol{\theta}})$, which contains the remaining r elements, is in general nonzero. In fact, a Taylor expansion gives

$$n^{-1/2}\boldsymbol{g}_2(\tilde{\boldsymbol{\theta}}) - n^{-1/2}\boldsymbol{g}_2(\boldsymbol{\theta}_0) - n^{-1}\boldsymbol{H}_{21}(\bar{\boldsymbol{\theta}})\, n^{1/2}(\tilde{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1^0) = \mathbf{0}, \tag{3.63}$$

where our usual shorthand notation $\bar{\boldsymbol{\theta}}$ is used for a vector that tends to $\boldsymbol{\theta}_0$ as $n \to \infty$, and $\boldsymbol{H}_{21}(\cdot)$ is the lower left block of the Hessian of the loglikelihood. We take the limit of the equation (3.63) as $n \to \infty$, making use of the asymptotic equality (3.62), and taking note that, according to the information matrix equality (3.34) for a correctly specified model, $\operatorname{plim}\big(n^{-1}\boldsymbol{H}_{21}(\boldsymbol{\theta}_0) + \mathcal{I}_{21}^0\big) = \mathbf{O}$ to see that

$$\operatorname*{plim}_{n\to\infty} \big[n^{-1/2}\boldsymbol{g}_2(\tilde{\boldsymbol{\theta}}) - n^{-1/2}\boldsymbol{g}_2(\boldsymbol{\theta}_0) + \mathcal{I}_{21}^0(\mathcal{I}_{11}^0)^{-1} n^{-1/2}\boldsymbol{g}_1(\boldsymbol{\theta}_0)\big] = \mathbf{O}.$$

where $\mathcal{I}^0 \equiv \mathcal{I}(\boldsymbol{\theta}_0)$. It follows that the asymptotic variance of $n^{-1/2}\boldsymbol{g}_2(\tilde{\boldsymbol{\theta}})$ is equal to that of $n^{-1/2}\boldsymbol{g}_2(\boldsymbol{\theta}_0) - \mathcal{I}_{21}^0(\mathcal{I}_{11}^0)^{-1}\boldsymbol{g}_1(\boldsymbol{\theta}_0)$. This last expression can be written as

$$\begin{bmatrix} -\mathcal{I}_{21}^0(\mathcal{I}_{11}^0)^{-1} & \mathbf{I} \end{bmatrix} \begin{bmatrix} n^{-1/2}\boldsymbol{g}_1(\boldsymbol{\theta}_0) \\ n^{-1/2}\boldsymbol{g}_2(\boldsymbol{\theta}_0) \end{bmatrix}, \tag{3.64}$$

where $\mathbf{I}$ is an $r \times r$ identity matrix. Then, since the asymptotic variance of the full gradient vector $n^{-1/2}\boldsymbol{g}(\boldsymbol{\theta}_0)$ is just $\mathcal{I}^0$, we see that the asymptotic variance of $n^{-1/2}\boldsymbol{g}_2(\tilde{\boldsymbol{\theta}})$ is

$$\begin{aligned} &\begin{bmatrix} -\mathcal{I}_{21}^0(\mathcal{I}_{11}^0)^{-1} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathcal{I}_{11}^0 & \mathcal{I}_{12}^0 \\ \mathcal{I}_{21}^0 & \mathcal{I}_{22}^0 \end{bmatrix} \begin{bmatrix} -(\mathcal{I}_{11}^0)^{-1}\mathcal{I}_{12}^0 \\ \mathbf{I} \end{bmatrix} \\ &\qquad = \mathcal{I}_{22}^0 - \mathcal{I}_{21}^0(\mathcal{I}_{11}^0)^{-1}\mathcal{I}_{12}^0. \end{aligned} \tag{3.65}$$

In Part 1, Exercise 9.F11, expressions were developed for the blocks of the inverses of partitioned matrices. It is easy to see from those expressions that the inverse of (3.65) is the 22 block of $\mathcal{I}^{-1}(\boldsymbol{\theta}_0)$. Thus, in order to obtain a statistic in asymptotically χ^2 form based on $\boldsymbol{g}_2(\tilde{\boldsymbol{\theta}})$, we can construct the quadratic form

$$\text{LM} = n^{-1/2}\boldsymbol{g}_2^\top(\tilde{\boldsymbol{\theta}})(\tilde{\mathcal{I}}^{-1})_{22}\, n^{-1/2}\boldsymbol{g}_2(\tilde{\boldsymbol{\theta}}) = \boldsymbol{g}_2^\top(\tilde{\boldsymbol{\theta}})(\tilde{\boldsymbol{I}}^{-1})_{22}\, \boldsymbol{g}_2(\tilde{\boldsymbol{\theta}}), \tag{3.66}$$

in which $\tilde{\mathcal{I}} = n^{-1}\boldsymbol{I}(\tilde{\boldsymbol{\theta}})$, and the notations $(\tilde{\mathcal{I}}^{-1})_{22}$ and $(\tilde{\boldsymbol{I}}^{-1})_{22}$ signify the 22 blocks of the inverses of $\tilde{\mathcal{I}}$ and $\boldsymbol{I}(\tilde{\boldsymbol{\theta}})$, respectively.

Since the statistic (3.66) is a quadratic form in an r–vector, which is asymptotically normally distributed with expectation $\mathbf{0}$, and the inverse of an $r \times r$ matrix that consistently estimates the covariance matrix of that vector, it is clear that the LM statistic is asymptotically distributed as $\chi^2(r)$ under the null. However, expression (3.66) is notationally awkward. Because the first-order conditions (3.61) imply that $\boldsymbol{g}_1(\tilde{\boldsymbol{\theta}}) = \mathbf{0}$, we can rewrite it as what appears to be a quadratic form with k rather than r degrees of freedom. We obtain

$$\text{LM} = \boldsymbol{g}^\top(\tilde{\boldsymbol{\theta}})\tilde{\boldsymbol{I}}^{-1}\boldsymbol{g}(\tilde{\boldsymbol{\theta}}), \tag{3.67}$$

where the notational awkwardness has disappeared. In addition, since (3.67) no longer depends on the partitioning of $\boldsymbol{\theta}$ that we used to express the zero restrictions, it is applicable quite generally, whether or not the restrictions are zero restrictions. This follows from the **invariance** of the LM test under reparametrizations of the model; see Exercise 3.15.

Expression (3.67) is the statistic associated with the **score form** of the LM test, often simply called the **score test**, since it is defined in terms of the score vector $\boldsymbol{g}(\boldsymbol{\theta})$ evaluated at the restricted estimates $\tilde{\boldsymbol{\theta}}$. It must, of course, be kept in mind that, despite its appearance, expression (3.67) has only r, and not k, degrees of freedom. This "using up" of $k - r$ degrees of freedom is due to the fact that the $k - r$ elements of $\boldsymbol{\theta}_1$ are estimated.

One way to maximize the loglikelihood function $\ell(\boldsymbol{\theta})$ subject to the restrictions $\boldsymbol{r}(\boldsymbol{\theta}) = \mathbf{0}$ is simultaneously to maximize the Lagrangian

$$\ell(\boldsymbol{\theta}) - \boldsymbol{r}^\top(\boldsymbol{\theta})\boldsymbol{\lambda}$$

with respect to $\boldsymbol{\theta}$ and minimize it with respect to the r–vector of Lagrange multipliers $\boldsymbol{\lambda}$. The Karush-Kuhn-Tucker (KKT) first-order conditions[6] that characterize the solution to this problem are the $k + r$ equations

$$\begin{aligned} \boldsymbol{g}(\tilde{\boldsymbol{\theta}}) - \boldsymbol{R}^\top(\tilde{\boldsymbol{\theta}})\tilde{\boldsymbol{\lambda}} &= \mathbf{0} \\ \boldsymbol{r}(\tilde{\boldsymbol{\theta}}) &= \mathbf{0}. \end{aligned}$$

[6] The canonical references for these conditions are Karush (1939) and Kuhn and Tucker (1951).

The first set of these equations allows us to rewrite the LM statistic (3.67) in terms of the Lagrange multipliers $\boldsymbol{\lambda}$, thereby obtaining the **LM form** of the test, which is

$$\text{LM} = \tilde{\boldsymbol{\lambda}}^\top \tilde{\boldsymbol{R}} \tilde{\boldsymbol{I}}^{-1} \tilde{\boldsymbol{R}}^\top \tilde{\boldsymbol{\lambda}}, \tag{3.68}$$

where $\tilde{\boldsymbol{R}} \equiv \boldsymbol{R}(\tilde{\boldsymbol{\theta}})$. The score form (3.67) is used much more often than the LM form (3.68), because $\boldsymbol{g}(\tilde{\boldsymbol{\theta}})$ is almost always available, no matter how the restricted estimates are obtained, whereas the vector $\tilde{\boldsymbol{\lambda}}$ is available only if they are obtained by using a Lagrangian.

LM Tests and Artificial Regressions

We have so far assumed that the information matrix estimator used to construct the LM statistic is $\tilde{\boldsymbol{I}} \equiv \boldsymbol{I}(\tilde{\boldsymbol{\theta}})$. Because this estimator is usually more efficient than other estimators of the information matrix, $\tilde{\boldsymbol{I}}$ is often referred to as the **efficient score** estimator of the information matrix. However, there are as many different ways to compute any given LM statistic as there are asymptotically valid ways to estimate the information matrix. In practice, $\tilde{\boldsymbol{I}}$ is often replaced by some other estimator, such as minus the empirical Hessian or the OPG estimator. For example, if the OPG estimator is used in (3.67), the statistic becomes

$$\tilde{\boldsymbol{g}}^\top (\tilde{\boldsymbol{G}}^\top \tilde{\boldsymbol{G}})^{-1} \tilde{\boldsymbol{g}}, \tag{3.69}$$

where $\tilde{\boldsymbol{g}} \equiv \boldsymbol{g}(\tilde{\boldsymbol{\theta}})$ and $\tilde{\boldsymbol{G}} \equiv \boldsymbol{G}(\tilde{\boldsymbol{\theta}})$. This **OPG variant** of the statistic is asymptotically, but not numerically, equivalent to the **efficient score variant** computed using $\tilde{\boldsymbol{I}}$. In contrast, the score and LM forms of the test are numerically equivalent provided both are computed using the same information matrix estimator.

The statistic (3.69) can readily be computed by use of an artificial regression called the **OPG regression**, which has the general form

$$\boldsymbol{\iota} = \boldsymbol{G}(\boldsymbol{\theta})\boldsymbol{c} + \text{residuals}, \tag{3.70}$$

where $\boldsymbol{\iota}$ is an n–vector of 1s. This regression can be constructed for any model for which the loglikelihood function can be written as the sum of n contributions. If we evaluate (3.70) at the vector of restricted estimates $\tilde{\boldsymbol{\theta}}$, it becomes

$$\boldsymbol{\iota} = \tilde{\boldsymbol{G}}\boldsymbol{c} + \text{residuals}, \tag{3.71}$$

and the explained sum of squares is

$$\boldsymbol{\iota}^\top \tilde{\boldsymbol{G}} (\tilde{\boldsymbol{G}}^\top \tilde{\boldsymbol{G}})^{-1} \tilde{\boldsymbol{G}}^\top \boldsymbol{\iota} = \tilde{\boldsymbol{g}}^\top (\tilde{\boldsymbol{G}}^\top \tilde{\boldsymbol{G}})^{-1} \tilde{\boldsymbol{g}},$$

by (3.27). The right-hand side above is equal to expression (3.69), and so the ESS from regression (3.71) is numerically equal to the OPG variant of the LM statistic.

In the case of regression (3.70), the total sum of squares is just n, the squared length of the vector $\boldsymbol{\iota}$. Therefore, ESS $= n -$ SSR. This result gives us a particularly easy way to calculate the LM test statistic, and it also puts an upper bound on it: The OPG variant of the LM statistic can never exceed the number of observations in the OPG regression.

Although the OPG variant of the LM statistic is easy to calculate for a very wide variety of models, it does not have particularly good finite-sample properties. In fact, there is a great deal of evidence to suggest that tests based on this form of the statistic are much more likely to overreject than tests based on any other form and that they can overreject very severely in some cases. Therefore, unless it is bootstrapped, the OPG form of the LM statistic should be used with great caution. See Davidson and MacKinnon (1993, Chapter 13) for references. Fortunately, in many circumstances, other artificial regressions with much better finite-sample properties can be used to compute LM statistics; see Davidson and MacKinnon (2001).

LM Tests and the GNR

Consider again the case of linear restrictions on the parameters of the classical normal linear model. By summing the contributions (3.45) to the gradient, we see that the gradient of the loglikelihood for this model with respect to $\boldsymbol{\beta}$ can be written as

$$\boldsymbol{g}(\boldsymbol{\beta}, \sigma) = \frac{1}{\sigma^2} \boldsymbol{X}^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}).$$

Since the information matrix (3.51) is block-diagonal, we need not bother with the gradient with respect to σ in order to compute the LM statistic (3.67). From (3.48), we know that the $\boldsymbol{\beta}$-$\boldsymbol{\beta}$ block of the information matrix is $\sigma^{-2}\boldsymbol{X}^\top\boldsymbol{X}$. Thus, if we write the restricted estimates of the parameters as $\tilde{\boldsymbol{\beta}}$ and $\tilde{\sigma}$, the statistic (3.67), computed with the efficient score estimator of the information matrix, takes the form

$$\frac{1}{\tilde{\sigma}^2} (\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}})^\top \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top (\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}}). \tag{3.72}$$

This variant of the LM statistic is, like the LR and some variants of the Wald statistic, a deterministic, strictly increasing, function of the F statistic (3.55); see Exercise 3.18.

More generally, for a nonlinear regression model subject to possibly nonlinear restrictions on the parameters, we see that, by analogy with (3.72), the LM statistic can be written as

$$\frac{1}{\tilde{\sigma}^2} (\boldsymbol{y} - \tilde{\boldsymbol{x}})^\top \tilde{\boldsymbol{X}} (\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}})^{-1} \tilde{\boldsymbol{X}}^\top (\boldsymbol{y} - \tilde{\boldsymbol{x}}), \tag{3.73}$$

where $\tilde{\boldsymbol{x}} \equiv \boldsymbol{x}(\tilde{\boldsymbol{\beta}})$ is the n–vector of nonlinear regression functions evaluated at the restricted ML estimates $\tilde{\boldsymbol{\beta}}$, and $\tilde{\boldsymbol{X}} \equiv \boldsymbol{X}(\tilde{\boldsymbol{\beta}})$ is the $n \times k$ matrix of

derivatives of the regression functions with respect to the components of $\boldsymbol{\beta}$. It is easy to show that (3.73) is just n times the uncentered R^2 from the GNR

$$\boldsymbol{y} - \tilde{\boldsymbol{x}} = \tilde{\boldsymbol{X}}\boldsymbol{b} + \text{residuals},$$

which corresponds to the unrestricted nonlinear regression, evaluated at the restricted estimates. As we saw in Section 1.7, this is one of the valid statistics that can be computed using a GNR.

Bootstrapping the Classical Tests

When two or more of the classical test statistics differ substantially in magnitude, or when we have any other reason to believe that asymptotic tests based on them may not be reliable, bootstrap tests provide an attractive alternative to asymptotic ones. Since maximum likelihood requires a fully specified model, it is generally appropriate to use a parametric bootstrap, rather than resampling. For any given parameter vector $\boldsymbol{\theta}$, the likelihood function is the density of the dependent variable. Therefore, parametric bootstrap samples $\boldsymbol{y}^*$ are simply realizations of vector random variables from the distribution characterized by that density, evaluated at consistent estimates of the model parameters. These estimates must, of course, satisfy the restrictions to be tested, and so the natural choice, and usually the best one, is the vector of restricted ML estimates.

The procedure we recommend for bootstrapping any of the classical tests is as follows. The model is estimated under the null to obtain the vector of restricted estimates $\tilde{\boldsymbol{\theta}}$, and the desired test statistic, $\hat{\tau}$, is computed. This step may, of course, entail the estimation of the unrestricted model. One then generates B bootstrap samples using the DGP characterized by $\tilde{\boldsymbol{\theta}}$. For each of them, a bootstrap statistic τ_j^*, $j = 1, \ldots, B$, is computed in the same way as was $\hat{\tau}$. A bootstrap P value can then be obtained in the usual way as the proportion of bootstrap statistics more extreme than $\hat{\tau}$ itself.

We strongly recommend use of the bootstrap whenever there is any reason to believe that classical tests based on asymptotic theory may not be reliable, unless calculating a moderate number of τ_j^* is computationally infeasible. When this calculation is expensive, methods that do not use a fixed value of B may be attractive; see Davidson and MacKinnon (2000).

It is important to note that, as we saw earlier in this section for some tests in linear regression models, certain classical test statistics may be deterministic, strictly increasing, functions of other statistics. The bootstrap P values must be identical for statistics related in this way, since a bootstrap P value depends only on the ordering of the statistic $\hat{\tau}$ and the bootstrap statistics τ_j^*, and this ordering is invariant under a deterministic, strictly increasing, function. If we can readily compute a number of test statistics that are not deterministically related, it is desirable to bootstrap all of them at once. This is usually much cheaper than bootstrapping them separately. In general, we would expect the

bootstrap P values from the various tests to be fairly similar, at least if the null hypothesis is true.

3.6 The Asymptotic Theory of the Three Classical Tests

In this section, much of which is fairly advanced, we show that the three classical test statistics are asymptotically equivalent. This is true both under the null hypothesis and under alternatives that are close to the null in a sense to be made precise later. The proof is however limited to the former case.

Theorem 3.1

Let the parametric model $\mathbb{M}_1$ represent the alternative hypothesis against which it is desired to test a null hypothesis represented by the model $\mathbb{M}_0 \subset \mathbb{M}_1$. Let the parameter vector for $\mathbb{M}_1$ be the k–vector $\boldsymbol{\theta}$, and let $\ell(\boldsymbol{\theta})$ be the corresponding loglikelihood function. Suppose that $\mathbb{M}_0$ is defined by imposing $r < k$ restrictions on $\boldsymbol{\theta}$. If the null hypothesis is true, so that the true DGP is contained in $\mathbb{M}_0$, then the three classical test statistics, LR, LM, and W, are asymptotically equivalent in the sense that the difference between any pair of them tends to zero in probability as the sample size n tends to infinity.

Proof:

Suppose that the true DGP is characterized by the parameter vector $\boldsymbol{\theta}_0$, which satisfies the r restrictions of the null hypothesis. Denote the asymptotic information matrix by $\mathcal{I}(\boldsymbol{\theta})$, and write $\mathcal{I} \equiv \mathcal{I}(\boldsymbol{\theta}_0)$. The gradient vector that contains the partial derivatives of $\ell(\boldsymbol{\theta})$ is denoted as $\boldsymbol{g}(\boldsymbol{\theta})$, and $\boldsymbol{s} \equiv \boldsymbol{g}(\boldsymbol{\theta}_0)$. Further, $\boldsymbol{H}(\boldsymbol{\theta})$ is the $k \times k$ Hessian matrix for $\ell(\boldsymbol{\theta})$, and $\boldsymbol{H} \equiv \boldsymbol{H}(\boldsymbol{\theta}_0)$. To avoid cluttering the notation, we omit zero subscripts throughout.

The results will be developed explicitly only for restrictions of the form $\boldsymbol{\theta}_2 = \boldsymbol{0}$, where $\boldsymbol{\theta} = [\boldsymbol{\theta}_1 \;\; \boldsymbol{\theta}_2]$, but they apply quite generally. As usual, let the unrestricted MLE be $\hat{\boldsymbol{\theta}}$ and the restricted MLE be $\tilde{\boldsymbol{\theta}} = [\tilde{\boldsymbol{\theta}}_1 \;\; \boldsymbol{0}]$.

By a second-order Taylor expansion of $\ell(\tilde{\boldsymbol{\theta}})$ around $\hat{\boldsymbol{\theta}}$, we obtain

$$\ell(\tilde{\boldsymbol{\theta}}) = \ell(\hat{\boldsymbol{\theta}}) + \tfrac{1}{2}(\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}})^\top \boldsymbol{H}(\bar{\boldsymbol{\theta}})(\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}),$$

where $\bar{\boldsymbol{\theta}}$ is defined as usual in such an expansion. The first-order term vanishes because of the likelihood equations $\boldsymbol{g}(\hat{\boldsymbol{\theta}}) = \boldsymbol{0}$. It follows that

$$\text{LR} = 2\big(\ell(\hat{\boldsymbol{\theta}}) - \ell(\tilde{\boldsymbol{\theta}})\big) = -(\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}})^\top \boldsymbol{H}(\bar{\boldsymbol{\theta}})(\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}).$$

The information matrix equality (3.34) and the consistency of $\hat{\boldsymbol{\theta}}$, which implies the consistency of $\bar{\boldsymbol{\theta}}$, then yield the result that

$$\text{LR} \stackrel{a}{=} n(\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}})^\top \mathcal{I}\,(\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}). \tag{3.74}$$

We can use the asymptotic equalities (3.38) and (3.62) to eliminate the estimators that appear in (3.74), replacing them by expressions that involve only information matrix $\mathcal{I}$ and the score vector $\boldsymbol{s}$, as follows:

$$n^{1/2}(\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}) = n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) - n^{1/2}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \stackrel{a}{=} \mathcal{I}^{-1}\boldsymbol{s} - \begin{bmatrix} \mathcal{I}_{11}^{-1}\boldsymbol{s}_1 \\ \mathbf{0} \end{bmatrix}. \tag{3.75}$$

Here $\mathcal{I}_{11}$ and $\boldsymbol{s}_1$ denote, respectively, the $(k-r) \times (k-r)$ block of $\mathcal{I}$ and the subvector of $\boldsymbol{s}$ that corresponds to $\boldsymbol{\theta}_1$. We rewrite the last expression in (3.75) as $\boldsymbol{J}\boldsymbol{s}$, where the $k \times k$ symmetric matrix $\boldsymbol{J}$ is defined as

$$\boldsymbol{J} \equiv \mathcal{I}^{-1} - \begin{bmatrix} \mathcal{I}_{11}^{-1} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \end{bmatrix}. \tag{3.76}$$

This means that (3.74) becomes

$$\text{LR} \stackrel{a}{=} \boldsymbol{s}^\top \boldsymbol{J}\,\mathcal{I}\,\boldsymbol{J}\boldsymbol{s}. \tag{3.77}$$

Moreover, from (3.76), we have that

$$\mathcal{I}\boldsymbol{J} = \mathbf{I}_k - \begin{bmatrix} \mathcal{I}_{11} & \mathcal{I}_{12} \\ \mathcal{I}_{21} & \mathcal{I}_{22} \end{bmatrix} \begin{bmatrix} \mathcal{I}_{11}^{-1} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \end{bmatrix} = \begin{bmatrix} \mathbf{O} & \mathbf{O} \\ -\mathcal{I}_{21}\mathcal{I}_{11}^{-1} & \mathbf{I}_{k_2} \end{bmatrix}, \tag{3.78}$$

where the suffixes on the two identity matrices above indicate their dimensions. If we denote the last $k \times k$ matrix in (3.78) by $\boldsymbol{Q}$, (3.78) can be written simply as $\mathcal{I}\boldsymbol{J} = \boldsymbol{Q}$. This in turn implies that $\mathcal{I}^{-1}\boldsymbol{Q} = \boldsymbol{J}$, and, since

$$\begin{bmatrix} \mathcal{I}_{11}^{-1} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \end{bmatrix} \boldsymbol{Q} = \begin{bmatrix} \mathcal{I}_{11}^{-1} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \end{bmatrix} \begin{bmatrix} \mathbf{O} & \mathbf{O} \\ -\mathcal{I}_{21}\mathcal{I}_{11}^{-1} & \mathbf{I}_{k_2} \end{bmatrix} = \mathbf{O},$$

it follows from (3.76) that $\boldsymbol{J}\boldsymbol{Q} = \boldsymbol{J}$. This implies that $\boldsymbol{J}\,\mathcal{I}\,\boldsymbol{J} = \boldsymbol{J}$, from which we conclude that (3.77) can be written as

$$\text{LR} \stackrel{a}{=} \boldsymbol{s}^\top \boldsymbol{J}\boldsymbol{s}. \tag{3.79}$$

This expression, together with the definition (3.76) of the matrix $\boldsymbol{J}$, shows clearly how $k - r$ of the k degrees of freedom of $\boldsymbol{s}^\top \mathcal{I}^{-1}\boldsymbol{s}$ are used up by the process of estimating $\boldsymbol{\theta}_1$ under the null hypothesis.

We now go through a similar exercise for the LM statistic, all variants of which are asymptotically equal to the statistic in (3.67). Expression (3.64) was shown to be asymptotically equal to $n^{-1/2}\boldsymbol{g}_2(\tilde{\theta})$. We can write this as

$$n^{-1/2}\boldsymbol{g}_2(\boldsymbol{\theta}) \stackrel{a}{=} \begin{bmatrix} -\mathcal{I}_{21}\mathcal{I}_{11}^{-1} & \mathbf{I}_{k_2} \end{bmatrix} \begin{bmatrix} \boldsymbol{s}_1 \\ \boldsymbol{s}_2 \end{bmatrix}. \tag{3.80}$$

By the restricted likelihood equations, $\boldsymbol{g}_1(\tilde{\boldsymbol{\theta}}) = \mathbf{0}$. Stacking this on top of (3.80) gives

$$n^{-1/2}\boldsymbol{g}(\tilde{\boldsymbol{\theta}}) = \begin{bmatrix} n^{-1/2}\boldsymbol{g}_1(\tilde{\boldsymbol{\theta}}) \\ n^{-1/2}\boldsymbol{g}_2(\tilde{\boldsymbol{\theta}}) \end{bmatrix} = \begin{bmatrix} \mathbf{O} & \mathbf{O} \\ -\mathcal{I}_{21}\mathcal{I}_{11}^{-1} & \mathbf{I}_{k_2} \end{bmatrix} \begin{bmatrix} \boldsymbol{s}_1 \\ \boldsymbol{s}_2 \end{bmatrix} = \boldsymbol{Q}\boldsymbol{s}.$$

We then see from (3.67) that

$$\text{LM} \stackrel{a}{=} \boldsymbol{s}^\top \boldsymbol{Q}^\top \mathcal{I}^{-1}\boldsymbol{Q}\boldsymbol{s} = \boldsymbol{s}^\top \boldsymbol{J}\boldsymbol{s}, \tag{3.81}$$

since $\mathcal{I}^{-1}\boldsymbol{Q} = \boldsymbol{J}$ and $\boldsymbol{Q}^\top \boldsymbol{J} = \boldsymbol{J}$ by our earlier results. The asymptotic equivalence of the LR and LM statistics follows from (3.79) and (3.81).

The Wald statistic (3.58), for the case of zero restrictions, can be written as

$$\text{W} = \hat{\boldsymbol{\theta}}_2^\top \big((\hat{\boldsymbol{I}}^{-1})_{22}\big)^{-1} \hat{\boldsymbol{\theta}}_2 = n^{1/2}\hat{\boldsymbol{\theta}}_2^\top \big([(n^{-1}\hat{\boldsymbol{I}})^{-1}]_{22}\big)^{-1} n^{1/2}\hat{\boldsymbol{\theta}}_2.$$

Thus

$$\text{W} \stackrel{a}{=} n^{1/2}\hat{\boldsymbol{\theta}}_2^\top \big((\mathcal{I}^{-1})_{22}\big)^{-1} n^{1/2}\hat{\boldsymbol{\theta}}_2. \tag{3.82}$$

Then, from (3.38) and the fact that, in the present case, $\boldsymbol{\theta}_0 = \mathbf{0}$, we see that

$$n^{1/2}\hat{\boldsymbol{\theta}}_2 = \begin{bmatrix} \mathbf{O} & \mathbf{I}_{k_2} \end{bmatrix} n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \stackrel{a}{=} \begin{bmatrix} \mathbf{O} & \mathbf{I}_{k_2} \end{bmatrix} \mathcal{I}^{-1}\boldsymbol{s}$$

Thus the right-hand side of (3.82) is asymptotically equal to

$$\boldsymbol{s}^\top \mathcal{I}^{-1} \begin{bmatrix} \mathbf{O} \\ \mathbf{I}_{k_2} \end{bmatrix} \big((\mathcal{I}^{-1})_{22}\big)^{-1} \begin{bmatrix} \mathbf{O} & \mathbf{I}_{k_2} \end{bmatrix} \mathcal{I}^{-1}\boldsymbol{s}. \tag{3.83}$$

Now

$$\begin{bmatrix} \mathbf{O} \\ \mathbf{I}_{k_2} \end{bmatrix} \big((\mathcal{I}^{-1})_{22}\big)^{-1} \begin{bmatrix} \mathbf{O} & \mathbf{I}_{k_2} \end{bmatrix} = \begin{bmatrix} \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \big((\mathcal{I}^{-1})_{22}\big)^{-1} \end{bmatrix}. \tag{3.84}$$

When we were developing the LM statistic in the previous section, we saw that the inverse of the 22 block of $\mathcal{I}^{-1}$ was equal to the right-hand side of (3.65), so that $\big((\mathcal{I}^{-1})_{22}\big)^{-1} = \mathcal{I}_{22} - \mathcal{I}_{21}\mathcal{I}_{11}^{-1}\mathcal{I}_{21}$. Now

$$\begin{aligned} \boldsymbol{Q}\,\mathcal{I}\,\boldsymbol{Q}^\top &= \begin{bmatrix} \mathbf{O} & \mathbf{O} \\ -\mathcal{I}_{21}(\mathcal{I}_{11})^{-1} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathcal{I}_{11} & \mathcal{I}_{12} \\ \mathcal{I}_{21} & \mathcal{I}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{O} & -(\mathcal{I}_{11})^{-1}\mathcal{I}_{12} \\ \mathbf{O} & \mathbf{I} \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathcal{I}_{22} - \mathcal{I}_{21}\mathcal{I}_{11}^{-1}\mathcal{I}_{21} \end{bmatrix} = \begin{bmatrix} \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \big((\mathcal{I}^{-1})_{22}\big)^{-1} \end{bmatrix}. \end{aligned}$$

Expression (3.83) is therefore equal to

$$\boldsymbol{s}^\top \mathcal{I}^{-1}\boldsymbol{Q}\,\mathcal{I}\,\boldsymbol{Q}^\top \mathcal{I}^{-1}\boldsymbol{s} = \boldsymbol{s}^\top \boldsymbol{J}\,\mathcal{I}\,\boldsymbol{J}\boldsymbol{s} = \boldsymbol{s}^\top \boldsymbol{J}\boldsymbol{s},$$

and so we can conclude that

$$\text{W} \stackrel{a}{=} \boldsymbol{s}^\top \boldsymbol{J}\boldsymbol{s}, \tag{3.85}$$

This, along with (3.79) and (3.81), shows that the three classical test statistics are all asymptotically equal to $\boldsymbol{s}^\top \boldsymbol{J}\boldsymbol{s}$, so that the differences between any pair of them tend to zero as $n \to \infty$, that is, they are all asymptotically equal. ■

Remark:

Because the score vector $\boldsymbol{s} \equiv n^{-1/2}\boldsymbol{g}(\boldsymbol{\theta}_0)$ does not converge in probability to a limiting random vector, but converges only in distribution, the proof shows only that the difference between any of the classical test statistics and $\boldsymbol{s}^\top \boldsymbol{J}\boldsymbol{s}$ tends to zero as $n \to \infty$. With the sort of asymptotic construction generally used in this book, neither the statistics nor $\boldsymbol{s}^\top \boldsymbol{J}\boldsymbol{s}$ converge in probability, but only in distribution. Intuitively, this is because, as the sample size grows, new information starts to outweigh the information in the first part of the sample. It can be shown that any limiting random variable would in fact be *independent* of the sample, and it is this contradiction that rules out convergence in probability.

Quadratic Approximations and Classical Test Statistics

The asymptotic equivalence of the three classical test statistics can be understood by a geometric argument based on quadratic approximations to the loglikelihood function. Consider first the case of a classical normal linear model with known disturbance variance. Then it can be seen directly from equation (3.10) that the loglikelihood function is a quadratic function of the parameter vector $\boldsymbol{\beta}$. Therefore, if σ^2 is known, the loglikelihood function is quadratic with respect to all the parameters that have to be estimated.

For simplicity, consider the special case in which there is just one regressor, in the form of a vector $\boldsymbol{x}$, and a single parameter, β. Then the loglikelihood function (3.10) can be written as

$$\ell(\beta) = a + b\beta - \tfrac{1}{2}h\beta^2, \tag{3.86}$$

where a is a numerical constant minus $\boldsymbol{y}^\top\boldsymbol{y}/2$, b is $\boldsymbol{x}^\top\boldsymbol{y}$, and h is $\boldsymbol{x}^\top\boldsymbol{x}$, independent of $\boldsymbol{y}$. We wish to test the restriction that $\beta = 0$.

The gradient of the loglikelihood function (3.86) is

$$g(\beta) \equiv \frac{\partial \ell(\beta)}{\partial \beta} = b - h\beta.$$

Setting this equal to 0, we find that the ML estimate is $\hat{\beta} = b/h$. Therefore, the LR statistic is

$$\mathrm{LR} = 2\big(\ell(\hat{\beta}) - \ell(0)\big) = 2a + 2b\hat{\beta} - h\hat{\beta}^2 - 2a = 2b^2/h - b^2/h = b^2/h.$$

For the loglikelihood function (3.86), $g(0) = b$. The Hessian, which in this simple case is the scalar $-h$, is independent both of $\boldsymbol{y}$ and β, and so the

information matrix $I(\beta)$ is the scalar quantity h for all β. It follows that the LM statistic is

$$\mathrm{LM} = g^\top(0)I^{-1}(0)g(0) = b^2/h.$$

Finally, for the Wald statistic, we use the fact that the inverse of the information matrix, h^{-1}, is the asymptotic variance of $\hat{\beta}$. Consequently,

$$\mathrm{W} = \hat{\beta}^2/V(\hat{\beta}) = h(b/h)^2 = b^2/h.$$

Thus we see that, in the special case of the quadratic loglikelihood function (3.86), the three classical test statistics are numerically equal. We would, of course, have obtained the same result if the null hypothesis had been that $\beta = \beta_0$ instead of $\beta = 0$.

In general, the loglikelihood function is not exactly quadratic. However, if we were to take a quadratic approximation to it, we could compute an LR statistic based on that approximation. Provided the approximation is made at a point that converges to the true parameter value under the null, the approximate LR statistic must have the same asymptotic distribution as the actual statistic. Thus the LM and Wald statistics can be thought of as approximate LR statistics that are computed using different quadratic approximations to the loglikelihood function.

This is illustrated in Figure 3.3 for the case in which there is a single parameter θ and the null hypothesis is that $\theta = \theta_0$. The solid line is the loglikelihood function. The dotted lines are two different quadratic approximations. One of these approximations, which is taken at θ_0, is the basis of the LM statistic. The other, which is taken at $\hat{\theta}$, is the basis of the Wald statistic. The LR statistic is twice the vertical distance CA in the figure. The LM statistic is twice the vertical distance CB, and the Wald statistic is twice the vertical distance DA.

The Three Classical Tests When the Null Is False

The asymptotic equivalence result established in Theorem 3.1 depends on the assumption that the DGP belongs to the null hypothesis. However, the three classical tests yield asymptotically equivalent inferences only if the equivalence holds more generally than just under the null hypothesis.

A test is said to be **consistent** against a DGP that does not belong to the null hypothesis if, under that DGP, the power of the test tends to 1 as the sample size tends to infinity. We saw in Part 1, Section 5.8 that, if the null and alternative hypotheses are classical normal linear models, power is determined by a noncentrality parameter that must tend to infinity for power to tend to 1. The three classical tests have a property similar to that of the exact tests of the classical normal linear model: Under DGPs in the alternative but not in the null, the classical test statistics tend to random variables that are distributed as noncentral chi-squared with r degrees of freedom, where the noncentrality parameters tend to infinity with the sample size.

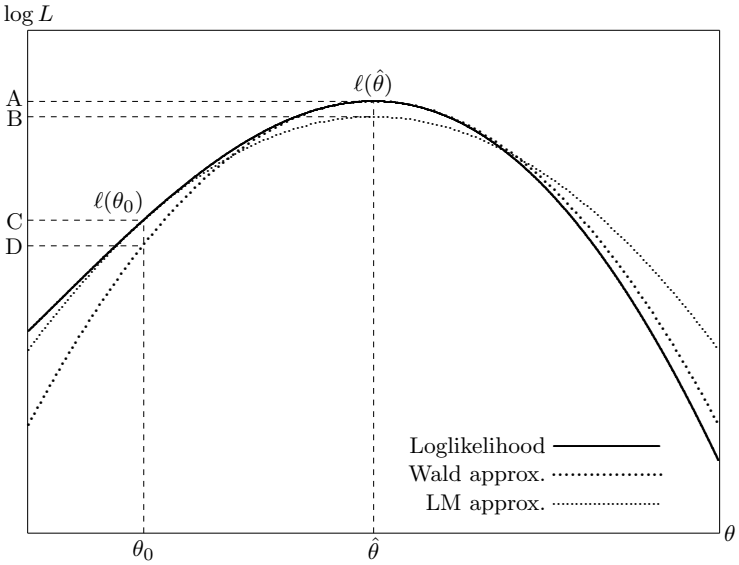


Figure 3.3 LR, LM, and Wald Tests

If all three classical tests can be shown to be consistent against a given DGP, then they are asymptotically equivalent under this DGP in the sense that, as $n \to \infty$, power tends to 1. But this does not rule out the possibility that, in finite samples, one of the tests may be much more powerful than the others. In order to investigate such a possibility, we want to develop a version of asymptotic theory in which the powers of different tests tend to different limits as $n \to \infty$ if they have very different powers in finite samples.

The simplest case we can study is that of the t statistic for the restriction $\beta_2 = 0$ in the linear regression model

$$\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{x}_2\beta_2 + \boldsymbol{u}.$$

The noncentrality parameter λ of the t statistic, in finite samples, is given as a function of β_2 and the disturbance variance σ^2:

$$\lambda = \frac{1}{\sigma}(\boldsymbol{x}_2^\top \boldsymbol{M}_1 \boldsymbol{x}_2)^{1/2}\beta_2.$$

For fixed β_2 and σ, λ tends to infinity as $n \to \infty$, since, under the regularity conditions for the classical normal linear model, $n^{-1}\boldsymbol{x}_2^\top \boldsymbol{M}_1 \boldsymbol{x}_2$ tends to a finite limit, which we denote by $S_{\boldsymbol{x}_2^\top \boldsymbol{M}_1 \boldsymbol{x}_2}$. It follows that $n^{-1/2}\lambda$, rather than λ itself, tends to a finite limit. But if, instead of keeping β_2 fixed, we subject it to what is called a **Pitman drift**, we can obtain a different result. Let δ be a fixed parameter, and, for each sample size n, let $\beta_2 = n^{-1/2}\delta$. Then

$$\lambda = n^{-1/2}\frac{1}{\sigma}(\boldsymbol{x}_2^\top \boldsymbol{M}_1 \boldsymbol{x}_2)^{1/2}\delta = \frac{1}{\sigma}(n^{-1}\boldsymbol{x}_2^\top \boldsymbol{M}_1 \boldsymbol{x}_2)^{1/2}\delta \to \frac{\delta}{\sigma}S_{\boldsymbol{x}_2^\top \boldsymbol{M}_1 \boldsymbol{x}_2}.$$

Since the limit of λ is no longer infinite, we can compare the possibly different limits obtained for different test statistics. A DGP for which the parameters depend explicitly on the sample size is called a **drifting DGP**.

If the model that corresponds to the alternative hypothesis is characterized by the loglikelihood function $\ell(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$, and the null hypothesis is the set of r zero restrictions $\boldsymbol{\theta}_2 = \boldsymbol{0}$, an appropriate drifting DGP for studying power is one for which $\boldsymbol{\theta}_1$ is fixed and $\boldsymbol{\theta}_2$ is given by $n^{-1/2}\boldsymbol{\delta}$ for a fixed r–vector $\boldsymbol{\delta}$. It can then be shown that, under this drifting DGP, just as under the null, the LR, LM, and Wald statistics converge in distribution as $n \to \infty$ to the same noncentral $\chi^2(r)$ distribution; see Exercise 3.19 for a very simple example. More generally, as discussed by Davidson and MacKinnon (1987), we can allow for drifting DGPs that do not lie within the alternative hypothesis, but that drift toward some fixed DGP in the null hypothesis. It then turns out that, for drifting DGPs that are, in an appropriate sense, equally distant from the null, the noncentrality parameter is maximized by those DGPs that do lie within the alternative hypothesis. This result justifies the intuition that, for a given number of degrees of freedom, tests against an alternative which happens to be true should have more power than tests against other alternatives.

3.7 ML Estimation of Models with Autoregressive Disturbances

In Section 2.8, we discussed several methods based on generalized or nonlinear least squares for estimating linear regression models with disturbances that follow an autoregressive process. An alternative approach is to use maximum likelihood. If it is assumed that the innovations are normally distributed, ML estimation is quite straightforward. With the normality assumption, the model (F9.36) considered in Part 1, Sections 9.7 and Part 1, 9.8 can be written as

$$y_t = \boldsymbol{X}_t\boldsymbol{\beta} + u_t, \quad u_t = \rho u_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim \text{NID}(0, \sigma_\varepsilon^2), \tag{3.87}$$

in which the disturbances follow an AR(1) process with parameter ρ that is assumed to be less than 1 in absolute value. If we omit the first observation, this model can be rewritten as in equation (F9.37). The result is just a nonlinear regression model, and so, as we saw in Section 3.2, the ML estimates of $\boldsymbol{\beta}$ and ρ must coincide with the NLS ones.

Maximum likelihood estimation of (3.87) is more interesting if we do not omit the first observation, because, in that case, the ML estimates no longer coincide with either the NLS or the (feasible) GLS estimates. For observations 2 through n, the contributions to the loglikelihood can be written as in (3.09):

$$\begin{aligned}\ell_t(\boldsymbol{y}^t, \boldsymbol{\beta}, \rho, \sigma_\varepsilon) = \\ -\tfrac{1}{2}\log 2\pi - \log\sigma_\varepsilon - \frac{1}{2\sigma_\varepsilon^2}(y_t - \rho y_{t-1} - \boldsymbol{X}_t\boldsymbol{\beta} + \rho\boldsymbol{X}_{t-1}\boldsymbol{\beta})^2.\end{aligned} \tag{3.88}$$

As required by (3.24), this expression is the log of the density of y_t conditional on the lagged dependent variable y_{t-1}.

For the first observation, the only information we have is that

$$y_1 = \boldsymbol{X}_1\boldsymbol{\beta} + u_1,$$

since the lagged dependent variable y_0 is not observed. However, with the normality assumption, we know from Part 1, Section 9.8 that the variance of u_1 is $\sigma_\varepsilon^2/(1-\rho^2)$. Thus the loglikelihood contribution from the first observation is the log of the density of the normal distribution with that variance, namely,

$$\begin{aligned}\ell_1(y_1, \boldsymbol{\beta}, \rho, \sigma_\varepsilon) = \\ -\tfrac{1}{2}\log 2\pi - \log\sigma_\varepsilon + \tfrac{1}{2}\log(1-\rho^2) - \frac{1-\rho^2}{2\sigma_\varepsilon^2}(y_1 - \boldsymbol{X}_1\boldsymbol{\beta})^2.\end{aligned} \tag{3.89}$$

Of course, we are assuming here that $\boldsymbol{X}_1$ is exogenous and therefore uncorrelated with u_1; see the discussion in Part 1, Section 9.8.

The loglikelihood function for the model (3.87) based on the entire sample is obtained by adding the contribution (3.89) to the sum of the contributions (3.88), for $t = 2, \ldots, n$. The result is

$$\begin{aligned}\ell(\boldsymbol{y}, \boldsymbol{\beta}, \rho, \sigma_\varepsilon) = -\tfrac{n}{2}\log 2\pi - n\log\sigma_\varepsilon + \tfrac{1}{2}\log(1-\rho^2) \\ -\frac{1}{2\sigma_\varepsilon^2}\Big((1-\rho^2)(y_1 - \boldsymbol{X}_1\boldsymbol{\beta})^2 + \sum_{t=2}^{n}(y_t - \rho y_{t-1} - \boldsymbol{X}_t\boldsymbol{\beta} + \rho\boldsymbol{X}_{t-1}\boldsymbol{\beta})^2\Big).\end{aligned} \tag{3.90}$$

The term $\frac{1}{2}\log(1-\rho^2)$ that appears in (3.90) plays an extremely important role in ML estimation. Because it tends to minus infinity as ρ tends to ± 1, its presence in the loglikelihood function ensures that there must be a maximum within the **stationarity region** defined by $|\rho| < 1$. Therefore, maximum likelihood estimation using the full sample is guaranteed to yield an estimate of ρ for which the AR(1) process is stationary. This is not the case for any of the estimation techniques discussed in Part 1, Section 9.8.

Let us define $u_t(\boldsymbol{\beta})$ as $y_t - \boldsymbol{X}_t\boldsymbol{\beta}$ for $t = 1, \ldots, n$, and let $\hat{u}_t = u_t(\hat{\boldsymbol{\beta}})$. Then, from the first-order conditions for the maximization of (3.90), it can be seen that the ML estimators $\hat{\boldsymbol{\beta}}$, $\hat{\rho}$, and $\hat{\sigma}_\varepsilon^2$ satisfy the following equations:

$$\begin{aligned}(1-\hat{\rho}^2)\boldsymbol{X}_1^\top\hat{u}_1 + \sum_{t=2}^{n}(\boldsymbol{X}_t - \hat{\rho}\boldsymbol{X}_{t-1})^\top(\hat{u}_t - \hat{\rho}\hat{u}_{t-1}) = \boldsymbol{0}, \\ \hat{\rho}\hat{u}_1^2 - \frac{\hat{\rho}\hat{\sigma}_\varepsilon^2}{1-\hat{\rho}^2} + \sum_{t=2}^{n}\hat{u}_{t-1}(\hat{u}_t - \hat{\rho}\hat{u}_{t-1}) = 0, \text{ and} \\ \hat{\sigma}_\varepsilon^2 = \tfrac{1}{n}\Big((1-\hat{\rho}^2)\hat{u}_1^2 + \sum_{t=2}^{n}(\hat{u}_t - \hat{\rho}\hat{u}_{t-1})^2\Big).\end{aligned} \tag{3.91}$$

The first two of these equations are similar, but not identical, to the estimating equations developed in Part 1, section 9.8 for iterated feasible GLS or NLS with account taken of the first observation. In Exercise 3.21, an artificial regression is developed which makes it quite easy to solve equations (3.91). This approach is simpler than the better-known algorithm for finding ML estimates that was proposed by Beach and MacKinnon (1978).

3.8 Transformations of the Dependent Variable

Whenever we specify a regression model, one of the choices we implicitly have to make is whether, and how, to transform the dependent variable. For example, if y_t, a typical observation on the dependent variable, is always positive, it would be perfectly valid to use $\log y_t$, or $y_t^{1/2}$, or one of many other monotonically increasing nonlinear transformations, instead of y_t itself as the regressand.

For concreteness, let us suppose that there are just two alternative models, which we will refer to as Model 1 and Model 2:

$$\begin{aligned}y_t &= \boldsymbol{X}_{t1}\boldsymbol{\beta}_1 + u_t, \quad u_t \sim \text{NID}(0, \sigma_1^2), \text{ and} \\ \log y_t &= \boldsymbol{X}_{t2}\boldsymbol{\beta}_2 + v_t, \quad v_t \sim \text{NID}(0, \sigma_2^2).\end{aligned}$$

Precisely how the regressors of the two competing models are related need not concern us here. In many cases, some of the regressors for one model are transformations of some of the regressors for the other model. For example, $\boldsymbol{X}_{t1}$ might consist of a constant and z_t, and $\boldsymbol{X}_{t2}$ might consist of a constant and $\log z_t$. Model 2 is often called a **loglinear regression** model.

Although we may be able to specify plausible-looking regression models for a number of different transformations of the dependent variable, using any model except the correct one implies that, in general, the disturbances are neither normally nor identically distributed. For example, suppose that we estimate Model 1 when the data were actually generated by Model 2 with parameters $\boldsymbol{\beta}_{20}$ and σ_{20}^2. It follows that

$$\begin{aligned}y_t &= \exp(\boldsymbol{X}_{t2}\boldsymbol{\beta}_{20} + v_t) \\ &= \exp(\boldsymbol{X}_{t2}\boldsymbol{\beta}_{20})\exp(v_t) \\ &= \exp(\boldsymbol{X}_{t2}\boldsymbol{\beta}_{20})\exp(\tfrac{1}{2}\sigma_{20}^2) + \exp(\boldsymbol{X}_{t2}\boldsymbol{\beta}_{20})\big(\exp(v_t) - \exp(\tfrac{1}{2}\sigma_{20}^2)\big).\end{aligned} \tag{3.92}$$

The last line here uses the fact that $\exp(v_t)$ is a lognormal variable, of which the expectation is $\exp(\sigma_{20}^2/2)$. Thus the first term in the last line is the conditional expectation of y_t, and so the second term, which is y_t minus this conditional expectation, is the disturbance for Model 1.

Even if it should turn out that $\boldsymbol{X}_{t1}\boldsymbol{\beta}_1$, the regression function for Model 1, can provide a reasonably good approximation to the conditional expectation

in the last line of (3.92), the disturbances for that model cannot possibly have the properties we generally assume them to have. If the disturbances in Model 2 are normally and identically distributed, then the disturbances in Model 1 must be skewed to the right and heteroskedastic. Their skewness is a consequence of the fact that lognormal variables are always skewed to the right (see Exercise 3.20). Because their variance is proportional to the square of $\exp(\boldsymbol{X}_{t2}\boldsymbol{\beta}_{20})$, they are heteroskedastic.

As this example demonstrates, even when the disturbances in the DGP are normally, identically, and independently distributed, using the wrong transformation of the dependent variable as the regressand yields, in general, a regression with disturbances that are neither homoskedastic nor symmetric. Thus, when we encounter heteroskedasticity and skewness in the residuals of a regression, one possible way to eliminate them is to estimate a different regression model in which the dependent variable has been subjected to some sort of nonlinear transformation.

Comparing Alternative Models

It is perfectly easy to subject the dependent variable to various nonlinear transformations and estimate one or more regression models for each of them. However, least-squares estimation does not provide any way to compare the fits of competing models that involve different transformations. But maximum likelihood estimation under the assumption that the disturbances are normally distributed does provide a straightforward way to do so. The idea is to compare the loglikelihoods of the alternative models considered as models for the same dependent variable.

For Model 1, in which y_t is the regressand, the concentrated loglikelihood function is simply

$$-\frac{n}{2}\log 2\pi - \frac{n}{2} - \frac{n}{2}\log\left(\frac{1}{n}\sum_{t=1}^{n}(y_t - \boldsymbol{X}_{t1}\boldsymbol{\beta}_1)^2\right). \tag{3.93}$$

Expression (3.93) is just expression (3.11) specialized to Model 1. Most regression packages report the value of (3.93) evaluated at the OLS estimates as the maximized value of the loglikelihood function.

In order to construct the loglikelihood function for the loglinear Model 2, interpreted as a model for y_t rather than for $\log y_t$, we need the density of y_t as a function of the model parameters. This requires us to use a standard result about **transformations of variables**. Suppose that we wish to know the CDF of a random variable X, but that what we actually know is the CDF of a random variable Z defined as $Z = h(X)$, where $h(\cdot)$ is a strictly increasing deterministic function. Denote this known CDF by F_Z. Then we can obtain the CDF F_X of X as follows.

$$\begin{aligned} F_X(x) &= \Pr(X \le x) = \Pr\big(h(X) \le h(x)\big) \\ &= \Pr\big(Z \le h(x)\big) = F_Z\big(h(x)\big). \end{aligned} \tag{3.94}$$

The second equality above follows because $h(\cdot)$ is strictly increasing. The relation between the densities of the variables X and Z is obtained by differentiating the leftmost and rightmost quantities in (3.94) with respect to x. Denoting the densities by $f_X(\cdot)$ and $f_Z(\cdot)$, we obtain

$$f_X(x) = F'_X(x) = F'_Z\big(h(x)\big)h'(x) = f_Z\big(h(x)\big)h'(x).$$

If h is strictly decreasing, the above result must be modified so as to use the absolute value of the derivative. As readers are asked to show in Exercise 3.23, the result then becomes

$$f_X(x) = f_Z\big(h(x)\big)|h'(x)|. \tag{3.95}$$

It is not difficult to see that (3.95) is a perfectly general result which holds for any strictly monotonic function h.

The factor by which $f_Z(z)$ is multiplied in order to produce $f_X(x)$ is the absolute value of what is called the **Jacobian** of the transformation. For Model 2, X is replaced by y_t, and the transformation h is the logarithm, so that Z becomes $\log y_t$. The density of y_t is then given by (3.95) in terms of that of $\log y_t$:

$$f(y_t) = f(\log y_t)\left|\frac{d\log y_t}{dy_t}\right| = \frac{f(\log y_t)}{y_t}, \tag{3.96}$$

where we drop subscripts and denote the densities of y_t and $\log y_t$ by $f(y_t)$ and $f(\log y_t)$, respectively.

We can now compute the loglikelihood for Model 2 thought of as a model for the y_t. The concentrated loglikelihood for the $\log y_t$ is given by (3.11):

$$-\frac{n}{2}\log 2\pi - \frac{n}{2} - \frac{n}{2}\log\left(\frac{1}{n}\sum_{t=1}^{n}(\log y_t - \boldsymbol{X}_{t2}\boldsymbol{\beta}_2)^2\right). \tag{3.97}$$

This expression is the log of the product of the densities of the $\log y_t$. Since the density of y_t, by (3.96), is equal to $1/y_t$ times the density of $\log y_t$, the loglikelihood function we are seeking is

$$-\frac{n}{2}\log 2\pi - \frac{n}{2} - \frac{n}{2}\log\left(\frac{1}{n}\sum_{t=1}^{n}(\log y_t - \boldsymbol{X}_{t2}\boldsymbol{\beta}_2)^2\right) - \sum_{t=1}^{n}\log y_t. \tag{3.98}$$

The last term here is a **Jacobian term**. It is the sum over all t of the logarithm of the **Jacobian factor** $1/y_t$ in the density of y_t. This Jacobian term is absolutely critical. If it were omitted, Model 2 would be a model for $\log y_t$, and it would make no sense to compare the value of the loglikelihood for (3.97) with the value for Model 1, which is a model for y_t. But when the Jacobian term is included, the loglikelihoods for both models are expressed in terms of y_t, and it is perfectly valid to compare their values. We can say with confidence

that the model corresponding to whichever of (3.93) and (3.98) has the largest value is the model that better fits the data.

Most regression packages evaluate expression (3.97) at the OLS estimates for the loglinear model and report that as the maximized value of the loglikelihood function. In order to compute the loglikelihood (3.98), which is what we need if we are to compare the fits of the linear and loglinear models, we have to add the Jacobian term to the value reported by the package.

Of course, the logarithmic transformation is by no means the only one that we might employ in practice. For example, when the y_t are sharply skewed to the right, a transformation like $\sqrt{y_t}$ might make sense; see Exercise 3.28.

Weighted least squares also involves transforming the dependent variable. If we believe that the disturbance variance is proportional to w_t^2, the use of feasible GLS leads us to divide y_t and all the regressors by w_t. When this is done, the Jacobian of the transformation is just $1/w_t$, and the Jacobian term in the loglikelihood function is

$$-\sum_{t=1}^{n} \log w_t. \tag{3.99}$$

In order to compare a model that has y_t as the regressand with another model that has y_t/w_t as the regressand, we need to add (3.99) to the value of the loglikelihood reported for the second model. Doing this makes the loglikelihoods from the two models comparable. If it really is appropriate to use weighted least squares, then the loglikelihood function for the weighted model should be higher than the loglikelihood function for the original model.

The most common nonlinear transformation in econometrics is the logarithmic transformation. Very often, we may find ourselves estimating a number of models, some of which have y_t as the regressand and some of which have $\log y_t$ as the regressand. If we simply want to decide which model fits best, we already know how to do so. We just have to compute the loglikelihood function for each of the models, including the Jacobian term $-\sum_{t=1}^{n} \log y_t$ for models in which the regressand is $\log y_t$, and pick the model with the highest loglikelihood. But if we want to perform a formal statistical test, and perhaps reject one or more of the competing models as incompatible with the data, we must go beyond simply comparing loglikelihood values.

The Box-Cox Regression Model

Most procedures for testing linear and loglinear models make use of the **Box-Cox transformation**,

$$B(x,\lambda) = \begin{cases} \dfrac{x^\lambda - 1}{\lambda} & \text{when } \lambda \neq 0; \\ \log x & \text{when } \lambda = 0, \end{cases}$$

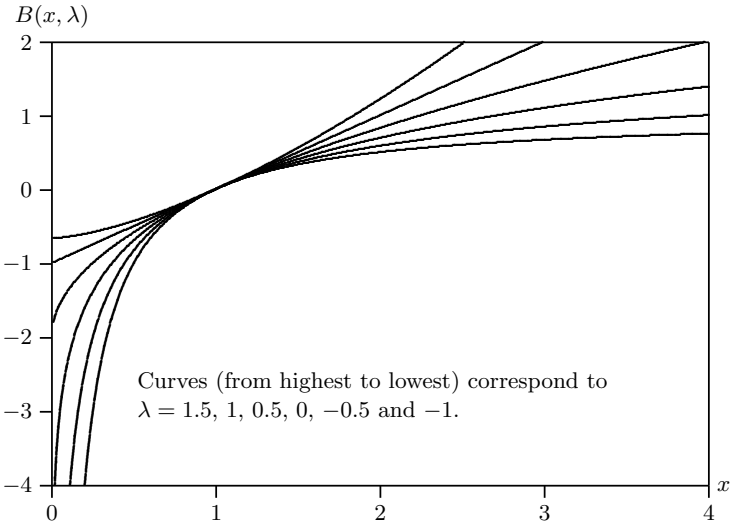


Figure 3.4 Box-Cox transformations for various values of λ

where λ is a parameter, which may be of either sign, and x, the argument of the transformation, must be positive. By l'Hôpital's Rule, $\log x$ is the limit of $(x^\lambda - 1)/\lambda$ as $\lambda \to 0$. Figure 3.4 shows the Box-Cox transformation for various values of λ. In practice, λ generally ranges from somewhat below 0 to somewhat above 1. It can be shown that $B(x,\lambda') \geq B(x,\lambda'')$ for $\lambda' \geq \lambda''$, and this inequality is evident in the figure. Thus the amount of curvature induced by the Box-Cox transformation increases as λ gets farther from 1 in either direction.

For the purposes of this section, the important thing about the Box-Cox transformation is that it allows us to formulate models which include both linear and loglinear regression models as special cases. In particular, consider the **Box-Cox regression model**

$$B(y_t,\lambda) = \sum_{i=1}^{k_1} \beta_i z_{ti} + \sum_{i=k_1+1}^{k} \beta_i B(x_{ti},\lambda) + u_t, \quad u_t \sim \text{NID}(0,\sigma^2), \tag{3.100}$$

in which there are k_1 regressors z_{ti} that are not subject to transformation and $k_2 = k - k_1$ nonconstant regressors x_{ti} that are always positive and are subject to transformation. The z_{ti} would include the constant term, if any, in addition to dummy variables and any other regressors that can take on nonpositive values. When $\lambda = 1$, this model reduces to the linear regression model

$$y_t - 1 = \sum_{i=1}^{k_1} \beta_i z_{ti} + \sum_{i=k_1+1}^{k} \beta_i (x_{ti} - 1) + u_t, \quad u_t \sim \text{NID}(0,\sigma^2).$$

Provided there is a constant term, or the equivalent of a constant term, among the z_{ti} regressors, this is equivalent to

$$y_t = \sum_{i=1}^{k_1} \beta_i z_{ti} + \sum_{i=k_1+1}^{k} \beta_i x_{ti} + u_t, \quad u_t \sim \text{NID}(0, \sigma^2), \tag{3.101}$$

with the β_i corresponding to the constant term redefined in the obvious way. When $\lambda = 0$, on the other hand, the Box-Cox model (3.100) reduces to the loglinear regression model

$$\log y_t = \sum_{i=1}^{k_1} \beta_i z_{ti} + \sum_{i=k_1+1}^{k} \beta_i \log x_{ti} + u_t, \quad u_t \sim \text{NID}(0, \sigma^2). \tag{3.102}$$

Thus it is clear that the linear regression model (3.101) and the loglinear regression model (3.102) can both be obtained as special cases of the Box-Cox regression model (3.100).

Testing Linear and Loglinear Regression Models

There are many ways in which we can test (3.101) and (3.102) against (3.100). Conceptually, the simplest is just to estimate all three models and perform two likelihood ratio tests. Let $\ell(\hat{\lambda})$ denote the maximum of the loglikelihood function for the unrestricted Box-Cox model (3.100), which readers are asked to derive in Exercise 3.29. Similarly, let $\ell(1)$ and $\ell(0)$ denote the maxima of the loglikelihood functions for the linear and loglinear models, respectively. Then the statistics for testing the linear and loglinear models against the Box-Cox regression model are

$$2\big(\ell(\hat{\lambda}) - \ell(1)\big) \quad \text{and} \quad 2\big(\ell(\hat{\lambda}) - \ell(0)\big),$$

respectively. If either of these statistics exceeds $\chi^2_{1-\alpha}(1)$, the $1-\alpha$ quantile of the $\chi^2(1)$ distribution, we may reject the model being tested at level α. In practice, this test tends to be quite powerful in samples of even moderate size, since it does not require a very large test statistic in order to reject the null hypothesis; the two most widely-used critical values are $\chi^2_{0.95}(1) = 3.84$ and $\chi^2_{0.99}(1) = 6.63$.

This procedure is conceptually very simple, but it requires us to estimate λ, which is a bit more work than simply running a linear regression. In some cases, however, we can avoid estimating λ. We know that $\ell(\hat{\lambda})$ must be larger than whichever of $\ell(1)$ and $\ell(0)$ is larger. Therefore, if

$$2\big(\ell(0) - \ell(1)\big) > \chi^2_{1-\alpha}(1), \tag{3.103}$$

we can certainly reject the linear model, even though we have not actually estimated the Box-Cox model or computed the LR test statistic. Similarly, if

$$2\big(\ell(1) - \ell(0)\big) > \chi^2_{1-\alpha}(1), \tag{3.104}$$

we can certainly reject the loglinear model. The quantities (3.103) and (3.104) provide lower bounds for the actual LR statistics. In practice, these lower bounds can often allow us to rule out models that are clearly incompatible with the data.

The fact that one can sometimes put a lower bound on the LR test statistic without actually estimating the unrestricted model is often very convenient. It was noted by Sargan (1964) in the context of choosing between linear and loglinear models, is widely used by applied workers, and has been proposed as a general basis for model selection by Pollak and Wales (1991). The procedure works in only one direction, of course. If, for example, (3.103) allows us to reject the linear model, then it tells us nothing about whether the loglinear model is acceptable to the data.

Lagrange Multiplier Tests

Since it is very easy to estimate linear and loglinear regression models, but somewhat harder to estimate the Box-Cox regression model, it is natural to use LM tests in this context. The first tests of this type were proposed by Godfrey and Wickens (1981). They are based on the OPG regression (3.70). However, as is often the case with tests based on the OPG regression, these tests tend to overreject quite severely in finite samples. Therefore, Davidson and MacKinnon (1985b) proposed Lagrange multiplier tests based on the **double-length artificial regression**, or **DLR**, that they had previously developed in Davidson and MacKinnon (1984a). This artificial regression is called "double-length" because it has $2n$ "observations," two for each of the actual observations in the sample.

For reasons of space, we will not write down the OPG or DLR test regressions here. Readers are asked to derive a special case of the former in Exercise 3.29. The latter, which are somewhat more complicated, are discussed in detail in Davidson and MacKinnon (1993, Chapter 14). If an LM test is to be used, we recommend use of the DLR rather than the OPG variant. There is a good deal of evidence that the DLR variant is much more reliable in finite samples; see Davidson and MacKinnon (1984b) and Godfrey, McAleer, and McKenzie (1988), among others. Of course, either variant of the test may easily be bootstrapped, as discussed in Section 3.5, and the OPG variant should perform acceptably when that is done. Because it is never necessary to estimate the unrestricted model, bootstrapping either of the LM tests is considerably less expensive than bootstrapping the LR test.

3.9 Final Remarks

Maximum likelihood estimation is widely used in many areas of econometrics, and we will encounter a number of important applications in the remainder of the book. Readers seeking a more advanced treatment of the theory than we

were able to give in this chapter may wish to consult Davidson and MacKinnon (1993), Cox and Hinkley (1974), or Stuart, Ord, and Arnold (1998).

As we have seen, ML estimation has many good properties, although these may be more apparent asymptotically than in finite samples. Its biggest limitation is the need for a fully specified parametric model. However, even if the dependent variable does not follow its assumed distribution, quasi-maximum likelihood estimators may still be consistent, even though they are not asymptotically efficient.

3.10 Exercises

3.1 Show that the ML estimator of the parameters $\boldsymbol{\beta}$ and σ of the classical normal linear model can be obtained by first concentrating the loglikelihood with respect to $\boldsymbol{\beta}$ and then maximizing the concentrated loglikelihood thereby obtained with respect to σ.

⋆**3.2** Let the n–vector $\boldsymbol{y}$ be a vector of mutually independent realizations from the uniform distribution on the interval $[\beta_1, \beta_2]$, usually denoted by $\mathrm{U}(\beta_1, \beta_2)$. Thus, $y_t \sim \mathrm{U}(\beta_1, \beta_2)$ for $t = 1, \ldots, n$. Let $\hat{\beta}_1$ be the ML estimator of β_1 given in (3.13), and suppose that the true values of the parameters are $\beta_1 = 0$ and $\beta_2 = 1$. Show that the CDF of $\hat{\beta}_1$ is

$$F(\beta) \equiv \Pr(\hat{\beta}_1 \leq \beta) = 1 - (1 - \beta)^n.$$

Use this result to show that $n(\hat{\beta}_1 - \beta_{10})$, which in this case is just $n\hat{\beta}_1$, is asymptotically exponentially distributed with $\theta = 1$. Note that the density of the exponential distribution was given in (3.03). (**Hint:** The limit as $n \to \infty$ of $(1 + x/n)^n$, for arbitrary real x, is e^x.)

Show that, for arbitrary given β_{10} and β_{20}, with $\beta_{20} > \beta_{10}$, the asymptotic distribution of $n(\hat{\beta}_1 - \beta_{10})$ is characterized by the density (3.03) with $\theta = (\beta_{20} - \beta_{10})^{-1}$.

3.3 Generate 10,000 random samples of sizes 20, 100, and 500 from the uniform $\mathrm{U}(0, 1)$ distribution. For each sample, compute $\bar{\gamma}$, the sample mean, and $\hat{\gamma}$, the average of the largest and smallest observations. Calculate the root mean squared error of each of these estimators for each of the three sample sizes. Do the results accord with what theory predicts?

3.4 Suppose that $h(\cdot)$ is a strictly concave, twice continuously differentiable, function on a possibly infinite interval of the real line. Let X be a random variable of which the support is contained in that interval. Suppose further that the first two moments of X exist. Prove Jensen's Inequality for the random variable X and the strictly concave function h by performing a Taylor expansion of h about $\mathrm{E}(X)$.

3.5 Prove that the definition (3.31) of the information matrix is equivalent to the definition

$$\boldsymbol{I}(\boldsymbol{\theta}) = \mathrm{E}_{\boldsymbol{\theta}}(\boldsymbol{g}(\boldsymbol{y}, \boldsymbol{\theta})\boldsymbol{g}^\top(\boldsymbol{y}, \boldsymbol{\theta})).$$

Hint: Use the result (3.30).

3.6 By differentiating the identity (3.28) with respect to θ_j, show that

$$\mathrm{E}_{\boldsymbol{\theta}}(G_{ti}(\boldsymbol{y}^t, \boldsymbol{\theta})G_{tj}(\boldsymbol{y}^t, \boldsymbol{\theta}) + (\boldsymbol{H}_t)_{ij}(\boldsymbol{y}^t, \boldsymbol{\theta})) = 0, \tag{3.105}$$

where the $k \times k$ matrix $\boldsymbol{H}_t(\boldsymbol{y}^t, \boldsymbol{\theta})$ is the Hessian of the contribution $\ell_t(\boldsymbol{y}^t, \boldsymbol{\theta})$ to the loglikelihood. The simplest way to proceed is to show first that (3.105) also holds if the left-hand side is the expectation conditional on $\boldsymbol{y}^{t-1}$.

3.7 Use the result (3.105) of the preceding exercise to prove the asymptotic information matrix equality (3.34).

3.8 Consider the linear regression model with exogenous explanatory variables,

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u},$$

where the only assumptions made regarding the disturbances are that they are uncorrelated and have expectation zero and finite variances that are, in general, different for each observation. The OLS estimator, which is consistent for this model, is equal to the ML estimator of the model under the assumption of homoskedastic normal disturbances. The ML estimator is therefore a QMLE for this model. Show that the $k \times k$ block of the sandwich covariance matrix estimator (3.44) that corresponds to $\hat{\boldsymbol{\beta}}$ is a version of the HCCME for the linear regression model.

3.9 Write out explicitly the empirical Hessian estimator of the covariance matrix of $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}$ for the classical normal linear model. How is it related to the IM estimator (3.52)?

How would your answer change if $\boldsymbol{X}\boldsymbol{\beta}$ in the classical normal linear model were replaced by $\boldsymbol{x}(\boldsymbol{\beta})$, a vector of nonlinear regression functions that implicitly depend on exogenous variables?

3.10 Suppose you treat σ^2 instead of σ as a parameter. Use arguments similar to the ones that led to equation (3.52) to derive the information matrix estimator of the covariance matrix of $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$. Then show that the same estimator can also be obtained by using the delta method.

⋆**3.11** Explain how to compute two different 95% confidence intervals for σ^2. One should be based on the covariance matrix estimator obtained in Exercise 9.10, and the other should be based on the original estimator (3.52). Are both of the intervals symmetric? Which seems more reasonable?

⋆**3.12** Let $\tilde{\boldsymbol{\theta}}$ denote any unbiased estimator of the k parameters of a parametric model fully specified by the loglikelihood function $\ell(\boldsymbol{\theta})$. The unbiasedness property can be expressed as the following identity:

$$\int L(\boldsymbol{y}, \boldsymbol{\theta})\tilde{\boldsymbol{\theta}}\, d\boldsymbol{y} = \boldsymbol{\theta}. \tag{3.106}$$

By using the relationship between $L(\boldsymbol{y}, \boldsymbol{\theta})$ and $\ell(\boldsymbol{y}, \boldsymbol{\theta})$ and differentiating this identity with respect to the components of $\boldsymbol{\theta}$, show that

$$\mathrm{Cov}_{\boldsymbol{\theta}}(\boldsymbol{g}(\boldsymbol{\theta}), (\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta})) = \mathbf{I},$$

where $\mathbf{I}$ is a $k \times k$ identity matrix, and the notation $\mathrm{Cov}_{\boldsymbol{\theta}}$ indicates that the covariance is to be calculated under the DGP characterized by $\boldsymbol{\theta}$.

Let $\boldsymbol{V}$ denote the $2k \times 2k$ covariance matrix of the $2k$–vector obtained by stacking the k components of $\boldsymbol{g}(\boldsymbol{\theta})$ above the k components of $\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}$. Partition this matrix into 4 $k \times k$ blocks as follows:

$$\boldsymbol{V} = \begin{bmatrix} \boldsymbol{V}_1 & \boldsymbol{C} \\ \boldsymbol{C}^\top & \boldsymbol{V}_2 \end{bmatrix},$$

where $\boldsymbol{V}_1$ and $\boldsymbol{V}_2$ are, respectively, the covariance matrices of the vectors $\boldsymbol{g}(\boldsymbol{\theta})$ and $\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}$ under the DGP characterized by $\boldsymbol{\theta}$. Then use the fact that $\boldsymbol{V}$ is positive semidefinite to show that the difference between $\boldsymbol{V}_2$ and $\boldsymbol{I}^{-1}(\boldsymbol{\theta})$, where $\boldsymbol{I}(\boldsymbol{\theta})$ is the (finite-sample) information matrix for the model, is a positive semidefinite matrix. **Hint:** Use the result of Part 1, Exercise 9.F11.

⋆**3.13** Consider the linear regression model

$$\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{u}, \quad \boldsymbol{u} \sim \mathrm{N}(\boldsymbol{0}, \sigma^2\mathbf{I}). \tag{3.107}$$

Derive the Wald statistic for the hypothesis that $\boldsymbol{\beta}_2 = \boldsymbol{0}$, as a function of the data, from the general formula (3.58). Show that it would be numerically identical to the Wald statistic (1.74) if the same estimate of σ^2 were used.

Show that, if the estimate of σ^2 is either the OLS or the ML estimator based on the unrestricted model (3.107), the Wald statistic is a deterministic, strictly increasing, function of the conventional F statistic. Give the explicit form of this deterministic function. Why can one reasonably expect that this result holds for tests of arbitrary linear restrictions on the parameters, and not only for zero restrictions of the type considered in this exercise?

⋆**3.14** Consider the Wald statistic W, the likelihood ratio statistic LR, and the Lagrange multiplier statistic LM for testing the hypothesis that $\boldsymbol{\beta}_2 = \boldsymbol{0}$ in the linear regression model (3.107). Since these are asymptotic tests, all the estimates of σ^2 are computed using the sample size n in the denominator. Express these three statistics as functions of the squared norms of the three components of the threefold decomposition (F5.34) of the dependent variable $\boldsymbol{y}$. By use of the inequalities

$$x > \log(1 + x) > \frac{x}{1 + x}, \quad x > 0,$$

show that W > LR > LM.

⋆**3.15** The model specified by the loglikelihood function $\ell(\boldsymbol{\theta})$ is said to be reparametrized if the parameter vector $\boldsymbol{\theta}$ is replaced by another parameter vector $\boldsymbol{\phi}$ related to $\boldsymbol{\theta}$ by a one to one relationship $\boldsymbol{\theta} = \boldsymbol{\Theta}(\boldsymbol{\phi})$ with inverse $\boldsymbol{\phi} = \boldsymbol{\Theta}^{-1}(\boldsymbol{\theta})$. The loglikelihood function for the reparametrized model is then defined as $\ell'(\boldsymbol{\phi}) \equiv \ell(\boldsymbol{\Theta}(\boldsymbol{\phi}))$. Explain why this definition makes sense.

Show that the maximum likelihood estimates $\hat{\boldsymbol{\phi}}$ of the reparametrized model are related to the estimates $\hat{\boldsymbol{\theta}}$ of the original model by the relation $\hat{\boldsymbol{\theta}} = \boldsymbol{\Theta}(\hat{\boldsymbol{\phi}})$. Specify the relationship between the gradients and information matrices of the two models in terms of the derivatives of the components of $\boldsymbol{\phi}$ with respect to those of $\boldsymbol{\theta}$.

Suppose that it is wished to test a set of r restrictions written as $\boldsymbol{r}(\boldsymbol{\theta}) = \boldsymbol{0}$. These restrictions can be applied to the reparametrized model in the form

$\boldsymbol{r}'(\boldsymbol{\phi}) \equiv \boldsymbol{r}(\boldsymbol{\Theta}(\boldsymbol{\phi})) = \boldsymbol{0}$. Show that the LR statistic is invariant to whether the restrictions are tested for the original or the reparametrized model. Show that the same is true for the LM statistic (3.67).

⋆**3.16** Show that the artificial OPG regression (3.71) possesses all the properties needed for hypothesis testing in the context of a model estimated by maximum likelihood. Specifically, show that

- the regressand $\boldsymbol{\iota}$ is orthogonal to the regressors $\boldsymbol{G}(\boldsymbol{\theta})$ when the latter are evaluated at the MLE $\hat{\boldsymbol{\theta}}$;
- the estimated OLS covariance matrix from (3.71) evaluated at $\hat{\boldsymbol{\theta}}$, when multiplied by n, consistently estimates the inverse of the asymptotic information matrix;
- the OPG regression (3.71) allows one-step estimation: If the OLS parameter estimates $\acute{\boldsymbol{c}}$ from (3.71) are evaluated at $\boldsymbol{\theta} = \acute{\boldsymbol{\theta}}$, where $\acute{\boldsymbol{\theta}}$ is any root-n consistent estimator of $\boldsymbol{\theta}$, then the one-step estimator $\grave{\boldsymbol{\theta}} \equiv \acute{\boldsymbol{\theta}} + \acute{\boldsymbol{c}}$ is asymptotically equivalent to $\hat{\boldsymbol{\theta}}$, in the sense that $n^{1/2}(\grave{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ and $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ tend to the same random variable as $n \to \infty$.

3.17 Show that the explained sum of squares from the artificial OPG regression (3.71) is equal to n times the uncentered R^2 from the same regression. Relate this fact to the use of test statistics that take the form of n times the R^2 of a GNR (Section 7.7).

3.18 Express the LM statistic (3.72) as a deterministic, strictly increasing, function of the F statistic (3.55).

⋆**3.19** Let the loglikelihood function $\ell(\theta)$ depend on one scalar parameter θ. For this special case, consider the distribution of the LM statistic (3.67) under the drifting DGP characterized by the parameter $\theta = n^{-1/2}\delta$ for a fixed δ. This DGP drifts toward the fixed DGP with $\theta = 0$, which we think of as representing the null hypothesis. Show first that $n^{-1}\boldsymbol{I}(n^{-1/2}\delta) \to \mathcal{I}(0)$ as $n \to \infty$. Here the asymptotic information matrix $\mathcal{I}(\theta)$ is just a scalar, since there is only one parameter.

Next, show that $n^{-1/2}$ times the gradient, evaluated at $\theta = 0$, which we may write as $n^{-1/2}g(0)$, is asymptotically normally distributed with expectation $\delta\,\mathcal{I}(0)$ and variance $\mathcal{I}(0)$. Finally, show that the LM statistic is asymptotically distributed as $\chi^2(1)$ with a finite noncentrality parameter, and give the value of that noncentrality parameter.

3.20 Let $z \sim \mathrm{N}(\mu, \sigma^2)$, and consider the lognormal random variable $x \equiv e^z$. Using the result that

$$\mathrm{E}(e^z) = \exp(\mu + \tfrac{1}{2}\sigma^2), \tag{3.108}$$

compute the second, third, and fourth central moments of x. Show that x is skewed to the right and has positive excess kurtosis.

Note: The **excess kurtosis** of a random variable is formally defined as the ratio of the fourth central moment to the square of the variance, minus 3.

⋆**3.21** The GNR proposed in Section 8.8 for NLS estimation of the model (3.87) can be written schematically as

$$\begin{bmatrix} (1-\rho^2)^{1/2}u_1(\boldsymbol{\beta}) \\ u_t(\boldsymbol{\beta}) - \rho u_{t-1}(\boldsymbol{\beta}) \end{bmatrix} = \begin{bmatrix} (1-\rho^2)^{1/2}\boldsymbol{X}_1 & 0 \\ \boldsymbol{X}_t - \rho\boldsymbol{X}_{t-1} & u_{t-1}(\boldsymbol{\beta}) \end{bmatrix} \begin{bmatrix} \boldsymbol{b} \\ b_\rho \end{bmatrix} + \text{residuals},$$

where $u_t(\boldsymbol{\beta}) \equiv y_t - \boldsymbol{X}_t\boldsymbol{\beta}$ for $t = 1, \ldots, n$, and the last $n-1$ rows of the artificial variables are indicated by their typical elements. Append one extra artificial observation to this artificial regression. For this observation, the regressand is $\big((1-\rho^2)u_1^2(\boldsymbol{\beta})/\sigma_\varepsilon - \sigma_\varepsilon\big)/\sqrt{2}$, the regressor in the column corresponding to ρ is $\rho\sigma_\varepsilon\sqrt{2}/(1-\rho^2)$, and the regressors in the columns corresponding to the elements of $\boldsymbol{\beta}$ are all 0. Show that, if at each iteration σ_ε^2 is updated by the formula

$$\sigma_\varepsilon^2 = \frac{1}{n}\Big((1-\rho^2)u_1^2(\boldsymbol{\beta}) + \sum_{t=2}^{n}\big(u_t(\boldsymbol{\beta}) - \rho u_{t-1}(\boldsymbol{\beta})\big)^2\Big),$$

then, if the iterations defined by the augmented artificial regression converge, the resulting parameter estimates satisfy the estimating equations (3.91) that define the ML estimator.

The odd-looking factors of $\sqrt{2}$ in the extra observation are there for a reason: Show that, when the artificial regression has converged, σ_ε^{-2} times the matrix of cross-products of the regressors is equivalent to the block of the information matrix that corresponds to $\boldsymbol{\beta}$ and ρ evaluated at the ML estimates. Explain why this means that we can use the OLS covariance matrix from the artificial regression to estimate the covariance matrix of $\hat{\boldsymbol{\beta}}$ and $\hat{\rho}$.

3.22 Using the artificial data in the file **ar1.data**, estimate the model

$$y_t = \beta_1 + \beta_2 x_t + u_t, \quad u_t = \rho u_{t-1} + \varepsilon_t, \quad t = 1, \ldots, 100,$$

which is correctly specified, in two different ways: ML omitting the first observation, and ML using all 100 observations. The second method should yield more efficient estimates of β_1 and β_2. For each of these two parameters, how large a sample of observations similar to the last 99 observations would be needed to obtain estimates as efficient as those obtained by using all 100 observations? Explain why your answer is greater than 100 in both cases.

3.23 Let the two random variables X and Z be related by the deterministic equation $Z = h(X)$, where h is strictly decreasing. Show that the densities of the two variables satisfy the equation

$$f_X(x) = -f_Z(h(x))h'(x).$$

Then show that (3.95) holds whenever h is a strictly monotonic function.

Let $X = Z^2$. Express the density of X in terms of that of Z, taking account of the possibility that the support of Z may include negative as well as positive numbers.

3.24 Suppose that a dependent variable y follows the exponential distribution given in (3.03), and let $x = y^2$. What is the density of x? Find the ML estimator of the parameter θ based on a sample of n observations x_t which are assumed to follow the distribution of which you have just obtained the density.

3.25 For a sample of n observations y_t generated from the exponential distribution, the loglikelihood function is (3.04), and the ML estimator is (3.06). Derive the asymptotic information matrix $\mathcal{I}(\theta)$, which is actually a scalar in this case, and use it to show how $n^{1/2}(\hat{\theta} - \theta_0)$ is distributed asymptotically. What is the empirical Hessian estimator of the variance of $\hat{\theta}$? What is the IM estimator?

There is an alternative parametrization of the exponential distribution, in which the parameter is $\phi \equiv 1/\theta$. Write down the loglikelihood function in terms of ϕ and obtain the asymptotic distribution of $n^{1/2}(\hat{\phi} - \phi_0)$. What is the empirical Hessian estimator of the variance of $\hat{\phi}$? What is the IM estimator?

⋆3.26 Consider the ML estimator $\hat{\theta}$ from the previous exercise. Explain how you could obtain an asymptotic confidence interval for θ in three different ways. The first should be based on inverting a Wald test in the θ parametrization, the second should be based on inverting a Wald test in the ϕ parametrization, and the third should be based on inverting an LR test.

Generate 100 observations from the exponential distribution with $\theta = 0.5$, find the ML estimate based on these artificial data, and calculate 95% confidence intervals for θ using the three methods just proposed. **Hint:** To generate the data, use uniformly distributed random numbers and the inverse of the exponential CDF.

3.27 Use the result (3.95) to derive the density of the $\mathrm{N}(\mu, \sigma^2)$ distribution from the density of the standard normal distribution.

In the classical normal linear model as specified in (3.07), it is the distribution of the disturbances $\boldsymbol{u}$ that is specified rather than that of the dependent variable $\boldsymbol{y}$. Reconstruct the loglikelihood function (3.10) starting from the densities of the disturbances u_t and using the Jacobians of the transformations that express the y_t in terms of the u_t.

3.28 Consider the model

$$y_t^{1/2} = \boldsymbol{X}_t\boldsymbol{\beta} + u_t, \quad u_t \sim \mathrm{NID}(0, \sigma^2),$$

in which it is assumed that all observations y_t on the dependent variable are positive. Write down the loglikelihood function for this model.

⋆3.29 Derive the loglikelihood function for the Box-Cox regression model (3.100). Then consider the following special case:

$$B(y_t, \lambda) = \beta_1 + \beta_2 B(x_t, \lambda) + u_t, \quad u_t \sim \mathrm{NID}(0, \sigma^2).$$

Derive the OPG regression for this model and explain precisely how to use it to test the hypotheses that the DGP is linear ($\lambda = 1$) and loglinear ($\lambda = 0$).

3.30 Consider the model XXX of the Canadian consumption function, with data from the file **consumption.data**, for the period 1953:1 to 1996:4. Compute the value of the maximized loglikelihood for this model regarded as a model for the *level* (not the log) of current consumption.

Formulate a model with the same algebraic form as XXX, but in levels of the income and consumption variables. Compute the maximized loglikelihood of this second model, and compare it with the value you obtained for the model in logs. Can you draw any conclusion about whether either model is misspecified?

Formulate a third model, using the variables in levels, but dividing them all by current income Y_t in order to account for heteroskedasticity. The result is a weighted least-squares model. Compute the maximized loglikelihood for this model as a model for the level of current consumption. Are there any more conclusions you can draw on the basis of your results?

3.31 Formulate a Box-Cox regression model which includes the first and second models of the previous exercise as special cases. Use the OPG regression to perform an LM test of the hypothesis that the Box-Cox parameter $\lambda = 0$, that is, that the loglinear model is correctly specified. Obtain both asymptotic and bootstrap P values.

3.32 The model XXX that was estimated in Exercise 3.30 can be written as

$$\Delta c_t = \beta_1 + \beta_2 \Delta y_t + \beta_3 \Delta y_{t-1} + \sigma \varepsilon_t,$$

where $\varepsilon_t \sim \text{NID}(0, 1)$. Suppose now that the ε_t, instead of being standard normal, follow the Cauchy distribution, with density $f(\varepsilon_t) = (\pi(1 + \varepsilon_t^2))^{-1}$. Estimate the resulting model by maximum likelihood, and compare the maximized value of the loglikelihood with the one obtained in Exercise 9.30.

3.33 Suppose that the dependent variable y_t is a proportion, so that $0 < y_t < 1$, $t = 1, \ldots, n$. An appropriate model for such a dependent variable is

$$\log\left(\frac{y_t}{1 - y_t}\right) = \boldsymbol{X}_t \boldsymbol{\beta} + u_t,$$

where $\boldsymbol{X}_t$ is a $k \times 1$ vector of exogenous variables, and $\boldsymbol{\beta}$ is a k–vector. Write down the loglikelihood function for this model under the assumption that $u_t \sim \text{NID}(0, \sigma^2)$. How would you maximize this loglikelihood function?

Chapter 4

Discrete and Limited Dependent Variables

4.1 Introduction

Although regression models are useful for modeling many types of data, they are not suitable for modeling every type. In particular, they should not be used when the dependent variable is discrete and can therefore take on only a countable number of values, or when it is continuous but is limited in the range of values it can take on. Since variables of these two types arise quite often, it is important to be able to deal with them, and many models have been proposed for doing so. In this chapter, we discuss some of the simplest and most commonly used models for discrete and limited dependent variables.

The most commonly encountered type of dependent variable that cannot be handled properly using a regression model is a **binary dependent variable**. Such a variable can take on only two values, which for practical reasons are almost always coded as 0 and 1. For example, a person may be in or out of the labor force, a commuter may drive to work or take public transit, a household may own or rent the home it resides in, and so on. In each case, the economic agent chooses between two alternatives, one of which is coded as 0 and one of which is coded as 1. A **binary response model** then tries to explain the probability that the agent chooses alternative 1 as a function of some observed explanatory variables. We discuss binary response models at some length in Section 4.2 and Section 4.3

A binary dependent variable is a special case of a **discrete dependent variable**. In Section 4.4, we briefly discuss several models for dealing with discrete dependent variables that can take on a fixed number of values. We consider two different cases, one in which the values have a natural ordering, and one in which they do not. Then, in Section 4.5, we discuss models for **count data**, in which the dependent variable can, in principle, take on any nonnegative, integer value.

Sometimes, a dependent variable is continuous but can take on only a limited range of values. For example, most types of consumer spending can be zero or positive but cannot be negative. If we have a sample that includes some

zero observations, we need to use a model that explicitly allows for this. By the same token, if the zero observations are excluded from the sample, we need to take account of this omission. Both types of model are discussed in Section 4.6. The related problem of **sample selectivity**, in which certain observations are omitted from the sample in a nonrandom way, is dealt with in Section 4.7. Finally, in Section 4.8, we discuss **duration models**, which attempt to explain how much time elapses before some event occurs or some state changes.

4.2 Binary Response Models: Estimation

In a binary response model, the value of the dependent variable y_t can take on only two values, 0 and 1. Let P_t denote the probability that $y_t = 1$ conditional on the information set Ω_t, which consists of exogenous and predetermined variables. A binary response model serves to model this conditional probability. Since the values are 0 or 1, it is clear that P_t is also the expectation of y_t conditional on Ω_t:

$$P_t \equiv \Pr(y_t = 1 \,|\, \Omega_t) = \mathrm{E}(y_t \,|\, \Omega_t),$$

Thus a binary response model can also be thought of as modeling a conditional expectation.

For many types of dependent variable, we can use a regression model to model conditional expectations, but that is not a sensible thing to do in this case. Suppose that $\boldsymbol{X}_t$ denotes a row vector of dimension k of variables that belong to the information set Ω_t, almost always including a constant term or the equivalent. Then a linear regression model would specify $\mathrm{E}(y_t \,|\, \Omega_t)$ as $\boldsymbol{X}_t\boldsymbol{\beta}$. But such a model fails to impose the condition that $0 \leq \mathrm{E}(y_t \,|\, \Omega_t) \leq 1$, which must hold because $\mathrm{E}(y_t \,|\, \Omega_t)$ is a probability. Even if this condition happened to hold for all observations in a particular sample, it would always be easy to find values of $\boldsymbol{X}_t$ for which the estimated probability $\boldsymbol{X}_t\hat{\boldsymbol{\beta}}$ would be less than 0 or greater than 1.

Since it makes no sense to have estimated probabilities that are negative or greater than 1, simply regressing y_t on $\boldsymbol{X}_t$ is not an acceptable way to model the conditional expectation of a binary variable. However, as we will see in the next section, such a regression can provide some useful information, and it is therefore not a completely useless thing to do in the early stages of an empirical investigation.

Any reasonable binary response model must ensure that $\mathrm{E}(y_t \,|\, \Omega_t)$ lies in the 0-1 interval. In principle, there are many ways to do this. In practice, however, two very similar models are widely used. Both of these models ensure that $0 < P_t < 1$ by specifying that

$$P_t \equiv \mathrm{E}(y_t \,|\, \Omega_t) = F(\boldsymbol{X}_t\boldsymbol{\beta}). \tag{4.01}$$

Here $\boldsymbol{X}_t\boldsymbol{\beta}$ is an **index function**, which maps from the vector $\boldsymbol{X}_t$ of explanatory variables and the vector $\boldsymbol{\beta}$ of parameters to a scalar index, and $F(x)$ is a **transformation function**, which has the properties that

$$F(-\infty) = 0, \quad F(\infty) = 1, \quad \text{and} \quad f(x) \equiv \frac{dF(x)}{dx} > 0. \tag{4.02}$$

These properties are, in fact, just the defining properties of the CDF of a probability distribution; recall Part 1, Section 1.2. They ensure that, although the index function $\boldsymbol{X}_t\boldsymbol{\beta}$ can take any value on the real line, the value of $F(\boldsymbol{X}_t\boldsymbol{\beta})$ must lie between 0 and 1.

The properties (4.02) also ensure that $F(x)$ is a nonlinear function. Consequently, changes in the values of the x_{ti}, which are the elements of $\boldsymbol{X}_t$, necessarily affect $\mathrm{E}(y_t \,|\, \Omega_t)$ in a nonlinear fashion. Specifically, when P_t is given by (4.01), its derivative with respect to x_{ti} is

$$\frac{\partial P_t}{\partial x_{ti}} = \frac{\partial F(\boldsymbol{X}_t\boldsymbol{\beta})}{\partial x_{ti}} = f(\boldsymbol{X}_t\boldsymbol{\beta})\beta_i, \tag{4.03}$$

where β_i is the i^{th} element of $\boldsymbol{\beta}$. Therefore, the magnitude of the derivative is proportional to $f(\boldsymbol{X}_t\boldsymbol{\beta})$. For the transformation functions that are almost always employed, $f(\boldsymbol{X}_t\boldsymbol{\beta})$ achieves a maximum at $\boldsymbol{X}_t\boldsymbol{\beta} = 0$ and then falls as $|\boldsymbol{X}_t\boldsymbol{\beta}|$ increases; see the CDFs plotted in Figure 4.1. Thus we see from (4.03) that the effect on P_t of a change in one of the independent variables is greatest when $P_t = .5$ and very small when P_t is close to 0 or 1.

The Probit Model

The first of the two widely-used choices for $F(x)$ is the cumulative standard normal distribution function,

$$\Phi(x) \equiv \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp\bigl(-\tfrac{1}{2}y^2\bigr)\,\mathrm{d}y.$$

When $F(\boldsymbol{X}_t\boldsymbol{\beta}) = \Phi(\boldsymbol{X}_t\boldsymbol{\beta})$, (4.01) is called the **probit model**. Although there exists no closed-form expression for $\Phi(x)$, it is easily evaluated numerically, and its first derivative is, of course, simply the standard normal density function, $\phi(x)$, which was defined in expression (F2.06).

One reason for the popularity of the probit model is that it can be derived from a model involving an unobserved, or **latent**, variable y_t°. Suppose that

$$y_t^\circ = \boldsymbol{X}_t\boldsymbol{\beta} + u_t, \quad u_t \sim \mathrm{NID}(0, 1). \tag{4.04}$$

We observe only the sign of y_t°, which determines the value of the observed binary variable y_t according to the relationship

$$y_t = 1 \text{ if } y_t^\circ > 0; \;\; y_t = 0 \text{ if } y_t^\circ \leq 0. \tag{4.05}$$

Together, equations (4.04) and (4.05) define what is called a **latent variable model**. One way to think of y_t° is as an index of the net utility associated with some action. If the action yields positive net utility, it is undertaken; otherwise, it is not undertaken. Because we observe only the sign of y_t°, we can normalize the variance of u_t to be unity. If the variance of u_t were some other value, say σ^2, we could divide $\boldsymbol{\beta}$, y_t°, and u_t by σ. Then u_t/σ would have variance 1, but the value of y_t would be unchanged. Another way to express this property is to say that the variance of u_t is not *identified* by the binary response model.

We can now compute P_t, the probability that $y_t = 1$. It is

$$\begin{aligned} \Pr(y_t = 1) &= \Pr(y_t^\circ > 0) = \Pr(\boldsymbol{X}_t\boldsymbol{\beta} + u_t > 0) \\ &= \Pr(u_t > -\boldsymbol{X}_t\boldsymbol{\beta}) = \Pr(u_t \leq \boldsymbol{X}_t\boldsymbol{\beta}) = \Phi(\boldsymbol{X}_t\boldsymbol{\beta}). \end{aligned} \tag{4.06}$$

The second-last equality in (4.06) makes use of the fact that the standard normal density function is symmetric around zero. The final result is just what we would get by letting $\Phi(\boldsymbol{X}_t\boldsymbol{\beta})$ play the role of the transformation function $F(\boldsymbol{X}_t\boldsymbol{\beta})$ in (4.01). Thus we have derived the probit model from the latent variable model that consists of (4.04) and (4.05).

The Logit Model

The logit model is very similar to the probit model. The only difference is that the function $F(x)$ is now the **logistic function**

$$\Lambda(x) \equiv \frac{1}{1 + \mathrm{e}^{-x}} = \frac{\mathrm{e}^x}{1 + \mathrm{e}^x}, \tag{4.07}$$

which has first derivative

$$\lambda(x) \equiv \frac{\mathrm{e}^x}{(1 + \mathrm{e}^x)^2} = \Lambda(x)\Lambda(-x). \tag{4.08}$$

This first derivative is evidently symmetric around zero, which implies that $\Lambda(-x) = 1 - \Lambda(x)$. A graph of the logistic function, as well as of the standard normal distribution function, is shown in Figure 4.1.

The logit model is most easily derived by assuming that

$$\log\left(\frac{P_t}{1 - P_t}\right) = \boldsymbol{X}_t\boldsymbol{\beta},$$

which says that the logarithm of the **odds** (that is, the ratio of the two probabilities) is equal to $\boldsymbol{X}_t\boldsymbol{\beta}$. Solving for P_t, we find that

$$P_t = \frac{\exp(\boldsymbol{X}_t\boldsymbol{\beta})}{1 + \exp(\boldsymbol{X}_t\boldsymbol{\beta})} = \frac{1}{1 + \exp(-\boldsymbol{X}_t\boldsymbol{\beta})} = \Lambda(\boldsymbol{X}_t\boldsymbol{\beta}).$$

This result is what we would get by letting $\Lambda(\boldsymbol{X}_t\boldsymbol{\beta})$ play the role of the transformation function $F(\boldsymbol{X}_t\boldsymbol{\beta})$ in (4.01).

Maximum Likelihood Estimation of Binary Response Models

By far the most common way to estimate binary response models is to use the method of maximum likelihood. Because the dependent variable is discrete, the likelihood function cannot be defined as a joint density function, as it was in Chapter 9 for models with a continuously distributed dependent variable. When the dependent variable can take on discrete values, the likelihood function for those values should be defined as the probability that the value is realized, rather than as the probability density at that value. With this redefinition, the *sum* of the possible values of the likelihood is equal to 1, just as the *integral* of the possible values of a likelihood based on a continuous distribution is equal to 1.

If, for observation t, the realized value of the dependent variable is y_t, then the likelihood for that observation if $y_t = 1$ is just the probability that $y_t = 1$, and if $y_t = 0$, it is the probability that $y_t = 0$. The logarithm of the appropriate probability is then the contribution to the loglikelihood made by observation t.

Since the probability that $y_t = 1$ is $F(\boldsymbol{X}_t\boldsymbol{\beta})$, the contribution to the loglikelihood function for observation t when $y_t = 1$ is $\log F(\boldsymbol{X}_t\boldsymbol{\beta})$. Similarly, the contribution to the loglikelihood function for observation t when $y_t = 0$ is $\log\big(1 - F(\boldsymbol{X}_t\boldsymbol{\beta})\big)$. Therefore, if $\boldsymbol{y}$ is an n–vector with typical element y_t, the loglikelihood function for $\boldsymbol{y}$ can be written as

$$\ell(\boldsymbol{y}, \boldsymbol{\beta}) = \sum_{t=1}^{n} \Big(y_t \log F(\boldsymbol{X}_t\boldsymbol{\beta}) + (1 - y_t) \log\big(1 - F(\boldsymbol{X}_t\boldsymbol{\beta})\big) \Big). \tag{4.09}$$

For each observation, one of the terms inside the large parentheses is always 0, and the other is always negative. The first term is 0 whenever $y_t = 0$, and the second term is 0 whenever $y_t = 1$. When either term is nonzero, it must be negative, because it is equal to the logarithm of a probability, and this probability must be less than 1 whenever $\boldsymbol{X}_t\boldsymbol{\beta}$ is finite. For the model to fit perfectly, $F(\boldsymbol{X}_t\boldsymbol{\beta})$ would have to equal 1 when $y_t = 1$ and 0 when $y_t = 0$, and the entire expression inside the parentheses would then equal 0. This could happen only if $\boldsymbol{X}_t\boldsymbol{\beta} = \infty$ whenever $y_t = 1$, and $\boldsymbol{X}_t\boldsymbol{\beta} = -\infty$ whenever $y_t = 0$. Therefore, we see that (4.09) is bounded above by 0.

Maximizing the loglikelihood function (4.09) is quite easy to do. For the logit and probit models, this function is globally concave with respect to $\boldsymbol{\beta}$ (see Pratt, 1981, and Exercise 4.1). This implies that the first-order conditions, or likelihood equations, uniquely define the ML estimator $\hat{\boldsymbol{\beta}}$, except for one special case that we consider in the subsection following the next one. These likelihood equations can be written as

$$\sum_{t=1}^{n} \frac{\big(y_t - F(\boldsymbol{X}_t\boldsymbol{\beta})\big) f(\boldsymbol{X}_t\boldsymbol{\beta}) x_{ti}}{F(\boldsymbol{X}_t\boldsymbol{\beta})\big(1 - F(\boldsymbol{X}_t\boldsymbol{\beta})\big)} = 0, \quad i = 1, \ldots, k. \tag{4.10}$$

There are many ways to find $\hat{\boldsymbol{\beta}}$ in practice. Because of the global concavity

of the loglikelihood function, Newton's Method generally works very well. Another approach, based on an artificial regression, will be discussed in the next section.

Conditions (4.10) look just like the first-order conditions for weighted least-squares estimation of the nonlinear regression model

$$y_t = F(\boldsymbol{X}_t\boldsymbol{\beta}) + v_t, \tag{4.11}$$

where the weight for observation t is

$$\Big(F(\boldsymbol{X}_t\boldsymbol{\beta})\big(1 - F(\boldsymbol{X}_t\boldsymbol{\beta})\big)\Big)^{-1/2}. \tag{4.12}$$

This weight is one over the square root of the variance of $v_t \equiv y_t - F(\boldsymbol{X}_t\boldsymbol{\beta})$, which is a binary random variable. By construction, v_t has mean 0, and its variance is

$$\begin{aligned} \mathrm{E}(v_t^2) &= \mathrm{E}\big(y_t - F(\boldsymbol{X}_t\boldsymbol{\beta})\big)^2 \\ &= F(\boldsymbol{X}_t\boldsymbol{\beta})\big(1 - F(\boldsymbol{X}_t\boldsymbol{\beta})\big)^2 + \big(1 - F(\boldsymbol{X}_t\boldsymbol{\beta})\big)\big(F(\boldsymbol{X}_t\boldsymbol{\beta})\big)^2 \\ &= F(\boldsymbol{X}_t\boldsymbol{\beta})\big(1 - F(\boldsymbol{X}_t\boldsymbol{\beta})\big). \end{aligned} \tag{4.13}$$

Notice how easy it is to take expectations in the case of a binary random variable. There are just two possible outcomes, and the probability of each of them is specified by the model.

Because the variance of v_t in regression (4.11) is not constant, applying nonlinear least squares to that regression would yield an inefficient estimator of the parameter vector $\boldsymbol{\beta}$. ML estimates could be obtained by applying iteratively reweighted nonlinear least squares. However, Newton's method, or a method based on the artificial regression to be discussed in the next section, is more direct and usually much faster.

Since the ML estimator is equivalent to weighted NLS, we can obtain it as an efficient GMM estimator. It is quite easy to construct elementary zero functions for a binary response model. The obvious function for observation t is $y_t - F(\boldsymbol{X}_t\boldsymbol{\beta})$. The covariance matrix of the n–vector of these zero functions is the diagonal matrix with typical element (4.13), and the row vector of derivatives of the zero function for observation t is $-f(\boldsymbol{X}_t\boldsymbol{\beta})\boldsymbol{X}_t$. With this information, we can set up the efficient estimating equations (2.82). As readers are asked to show in Exercise 4.3, these equations are equivalent to the likelihood equations (4.10).

Intuitively, efficient GMM and maximum likelihood give the same estimator because, once it is understood that the y_t are binary variables, the elementary zero functions serve to specify the probabilities $\Pr(y_t = 1)$, and they thus constitute a full specification of the model.

Comparing Probit and Logit Models

In practice, the probit and logit models generally yield very similar predicted probabilities, and the maximized values of the loglikelihood function (4.09) for the two models therefore tend to be very close. A formal comparison of these two values is possible. If twice the difference between them is greater than 3.84, the .05 critical value for the $\chi^2(1)$ distribution, then we can reject whichever model fits less well at the .05 level.[1] Such a procedure was discussed in Section 2.8 in the context of linear and loglinear models. In practice, however, experience shows that this sort of comparison rarely rejects either model unless the sample size is quite large.

In most cases, the only real difference between the probit and logit models is the way in which the elements of $\boldsymbol{\beta}$ are scaled. This difference in scaling occurs because the variance of the distribution for which the logistic function is the CDF can be shown to be $\pi^2/3$, while that of the standard normal distribution is, of course, unity. The logit estimates therefore all tend to be larger in absolute value than the probit estimates, although usually by a factor that is somewhat less than $\pi/\sqrt{3}$. Figure 4.1 plots the standard normal CDF, the logistic function, and the logistic function rescaled to have variance unity.

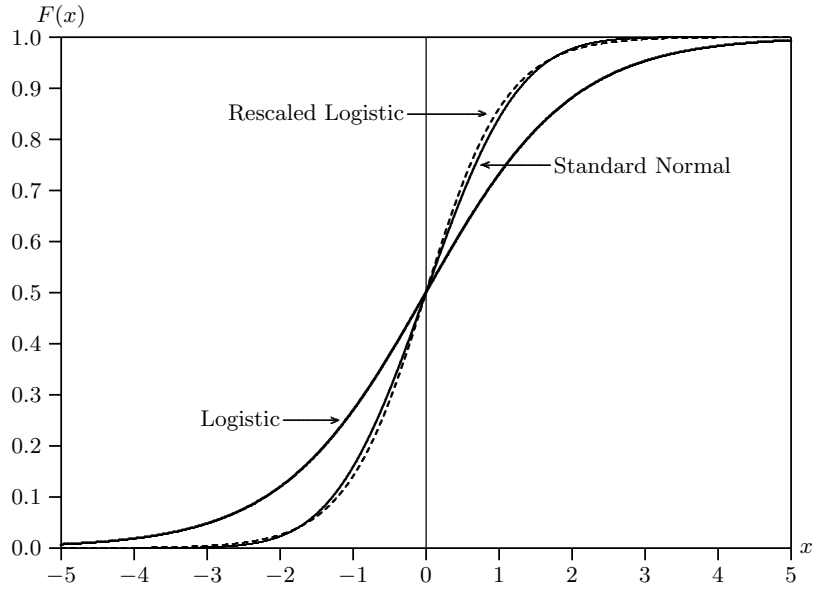


Figure 4.1 Alternative choices for $F(x)$

[1] This assumes that there exists a comprehensive model, with a single additional parameter, which includes the probit and logit models as special cases. It is not difficult to formulate such a model; see Exercise 4.4.

The resemblance between the standard normal CDF and the rescaled logistic function is striking. The main difference is that the rescaled logistic function puts more weight in the extreme tails.

The Perfect Classifier Problem

We have seen that the loglikelihood function (4.09) is bounded above by 0, and that it achieves this bound if $\boldsymbol{X}_t\boldsymbol{\beta} = -\infty$ whenever $y_t = 0$ and $\boldsymbol{X}_t\boldsymbol{\beta} = \infty$ whenever $y_t = 1$. Suppose there is some linear combination of the independent variables, say $\boldsymbol{X}_t\boldsymbol{\beta}^\bullet$, such that

$$\begin{aligned} y_t = 0 \text{ whenever } \boldsymbol{X}_t\boldsymbol{\beta}^\bullet < 0, \text{ and} \\ y_t = 1 \text{ whenever } \boldsymbol{X}_t\boldsymbol{\beta}^\bullet > 0. \end{aligned} \tag{4.14}$$

When this happens, there is said to be **complete separation** of the data. In this case, it is possible to make the value of $\ell(\boldsymbol{y}, \boldsymbol{\beta})$ arbitrarily close to 0 by setting $\boldsymbol{\beta} = \gamma\boldsymbol{\beta}^\bullet$ and letting $\gamma \to \infty$. This is precisely what any nonlinear maximization algorithm attempts to do if there exists a vector $\boldsymbol{\beta}^\bullet$ for which conditions (4.14) are satisfied. Because of the limitations of computer arithmetic, the algorithm must eventually terminate with some sort of numerical error at a value of the loglikelihood function that is slightly less than 0. If conditions (4.14) are satisfied, $\boldsymbol{X}_t\boldsymbol{\beta}^\bullet$ is said to be a **perfect classifier**, since it allows us to predict y_t with perfect accuracy for every observation.

The problem of perfect classifiers has a geometrical interpretation. In the k–dimensional space spanned by the columns of the matrix $\boldsymbol{X}$ formed from the row vectors $\boldsymbol{X}_t$, the vector $\boldsymbol{\beta}^\bullet$ defines a hyperplane that passes through the origin and that separates the observations for which $y_t = 1$ from those for which $y_t = 0$. Whenever one column of $\boldsymbol{X}$ is a constant, then the separating hyperplane can be represented in the $(k-1)$–dimensional space of the other explanatory variables. If we write

$$\boldsymbol{X}_t\boldsymbol{\beta}^\bullet = \alpha^\bullet + \boldsymbol{X}_{t2}\boldsymbol{\beta}_2^\bullet,$$

with $\boldsymbol{X}_{t2}$ a $1 \times (k-1)$ vector, then $\boldsymbol{X}_t\boldsymbol{\beta}^\bullet = 0$ is equivalent to $\boldsymbol{X}_{t2}\boldsymbol{\beta}_2^\bullet = -\alpha^\bullet$, which is the equation of a hyperplane in the space of the $\boldsymbol{X}_{t2}$ that in general does not pass through the origin. This is illustrated in Figure 4.2 for the case $k = 3$. The asterisks, which all lie to the northeast of the separating line for which $\boldsymbol{X}_t\boldsymbol{\beta}^\bullet = 0$, represent the $\boldsymbol{X}_{t2}$ for the observations with $y_t = 1$, and the circles to the southwest of the separating line represent them for the observations with $y_t = 0$.

It is clear from Figure 4.2 that, when a perfect classifier occurs, the separating hyperplane is not, in general, unique. One could move the intercept of the separating line in the figure up or down a little while maintaining the separating property. Likewise, one could swivel the line a little about the point of intersection with the vertical axis. Even if the separating hyperplane were

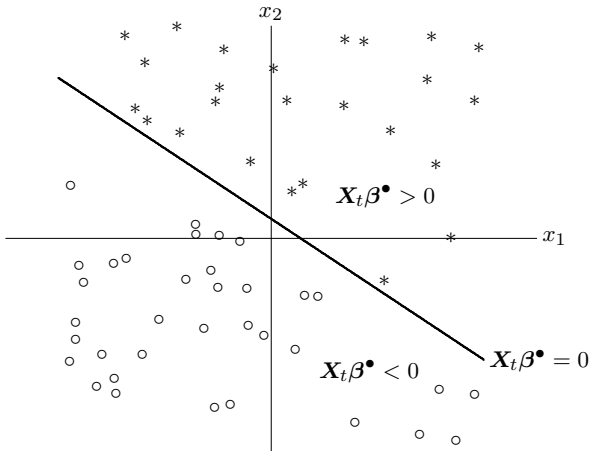


Figure 4.2 A perfect classifier yields a separating hyperplane

unique, we could not identify all the components of $\boldsymbol{\beta}$. This follows from the fact that the equation $\boldsymbol{X}_t\boldsymbol{\beta}^\bullet = 0$ is equivalent to the equation $\boldsymbol{X}_t(c\boldsymbol{\beta}^\bullet) = 0$ for any nonzero scalar c. The separating hyperplane is therefore defined equally well by any multiple of $\boldsymbol{\beta}^\bullet$. Although this suggests that we might be able to estimate $\boldsymbol{\beta}^\bullet$ up to a scalar factor by imposing a normalization on it, there is no question of estimating $\boldsymbol{\beta}^\bullet$ in the usual sense, and inference on it would require methods beyond the scope of this book.

Even when no parameter vector exists that satisfies the inequalities (4.14), there may exist a $\beta^\bullet$ that satisfies the corresponding nonstrict inequalities. There must then be at least one observation with $y_t = 0$ and $\boldsymbol{X}_t\boldsymbol{\beta}^\bullet = 0$, and at least one other observation with $y_t = 1$ and $\boldsymbol{X}_t\boldsymbol{\beta}^\bullet = 0$. In such a case, we speak of **quasi-complete separation** of the data. The separating hyperplane is then unique, and the upper bound of the loglikelihood is no longer zero, as readers are invited to verify in Exercise 4.6.

When there is either complete or quasi-complete separation, no finite ML estimator exists. This is likely to occur in practice when the sample is very small, when almost all of the y_t are equal to 0 or almost all of them are equal to 1, or when the model fits extremely well. Exercise 4.5 is designed to give readers a feel for the circumstances in which ML estimation is likely to fail because there is a perfect classifier.

If a perfect classifier exists, the loglikelihood should be close to its upper bound (which may be 0 or a small negative number) when the maximization algorithm quits. Thus, if the model seems to fit extremely well, or if the algorithm terminates in an unusual way, one should always check to see whether the parameter values imply the existence of a perfect classifier. For a detailed discussion of the perfect classifier problem, see Albert and Anderson (1984).

4.3 Binary Response Models: Inference

Inference about the parameters of binary response models is usually based on the standard results for ML estimation that were discussed in Chapter 3. It can be shown that

$$n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{d} \mathrm{N}\big(\mathbf{0}, \operatorname*{plim}_{n\to\infty}(n^{-1}\boldsymbol{X}^\top\boldsymbol{\Upsilon}(\boldsymbol{\beta}_0)\boldsymbol{X})^{-1}\big), \tag{4.15}$$

where $\boldsymbol{X}$ is an $n \times k$ matrix with typical row $\boldsymbol{X}_t$, $\boldsymbol{\beta}_0$ is the true value of $\boldsymbol{\beta}$, and $\boldsymbol{\Upsilon}(\boldsymbol{\beta})$ is an $n \times n$ diagonal matrix with typical diagonal element

$$\Upsilon_t(\boldsymbol{\beta}) \equiv \frac{f^2(\boldsymbol{X}_t\boldsymbol{\beta})}{F(\boldsymbol{X}_t\boldsymbol{\beta})\big(1 - F(\boldsymbol{X}_t\boldsymbol{\beta})\big)}. \tag{4.16}$$

Not surprisingly, the covariance matrix in expression (4.15) looks like the asymptotic covariance matrix for weighted least-squares estimation, with weights (4.12), of the GNR that corresponds to regression (4.11). This GNR is

$$y_t - F(\boldsymbol{X}_t\boldsymbol{\beta}) = f(\boldsymbol{X}_t\boldsymbol{\beta})\boldsymbol{X}_t\boldsymbol{b} + \text{residual}. \tag{4.17}$$

The factor of $f(\boldsymbol{X}_t\boldsymbol{\beta})$ that multiplies all the regressors of the GNR accounts for the numerator of (4.16). Its denominator is simply the variance of the disturbance in regression (4.11). Two ways to obtain the asymptotic covariance matrix (4.15) using general results for ML estimation are explored in Exercises 4.7 and 4.8.

In practice, the asymptotic result (4.15) is used to justify the covariance matrix estimator

$$\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}) = \big(\boldsymbol{X}^\top\boldsymbol{\Upsilon}(\hat{\boldsymbol{\beta}})\boldsymbol{X}\big)^{-1}, \tag{4.18}$$

in which the unknown $\boldsymbol{\beta}_0$ is replaced by $\hat{\boldsymbol{\beta}}$, and the factor of n^{-1}, which is needed only for asymptotic analysis, is omitted. This approximation may be used to obtain standard errors, t statistics, Wald statistics, and confidence intervals that are asymptotically valid. However, none of these is exact in finite samples.

It is clear from equations (4.15) and (4.18) that the ML estimator for the binary response model gives some observations more weight than others. In fact, the weight given to observation t is proportional to the square root of expression (4.16) evaluated at $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$. It can be shown that, for both the logit and probit models, the maximum weight is given to observations for which $\boldsymbol{X}_t\boldsymbol{\beta} = 0$, which implies that $P_t = .5$, while relatively little weight is given to observations for which P_t is close to 0 or 1; see Exercise 4.9. This makes sense, since when P_t is close to 0 or 1, a given change in $\boldsymbol{X}_t\boldsymbol{\beta}$ can have little effect on P_t, while when P_t is close to .5, such a change has a much larger effect. Thus we see that ML estimation, quite sensibly, gives more weight to observations that provide more information about the parameter values.

Likelihood Ratio Tests

It is straightforward to test restrictions on binary response models using LR tests. We simply estimate both the restricted and the unrestricted model and calculate twice the difference between the two maximized values of the loglikelihood function. As usual, the LR test statistic is asymptotically distributed as $\chi^2(r)$, where r is the number of restrictions.

One especially simple application of this procedure can be used to test whether the regressors in a binary response model have any explanatory power at all. The null hypothesis is that $\mathrm{E}(y_t \,|\, \Omega_t)$ is a constant, and the ML estimate of this constant is just $\bar{y}$, the unconditional sample mean of the dependent variable. It is not difficult to show that, under the null hypothesis, the loglikelihood function (4.09) reduces to

$$n\,\bar{y}\log(\bar{y}) + n(1 - \bar{y})\log(1 - \bar{y}), \tag{4.19}$$

which is very easy to calculate. Twice the difference between the unrestricted maximum of the loglikelihood function and the restricted maximum (4.19) is asymptotically distributed as $\chi^2(k-1)$. This statistic is analogous to the usual F test for all the slope coefficients in a linear regression model to equal zero, and many computer programs routinely compute it.

An Artificial Regression for Binary Choice Models

There is a convenient artificial regression for binary response models.[2]. Like the Gauss-Newton regression, to which it is closely related, the **binary response model regression**, or **BRMR**, can be used for a variety of purposes, including parameter estimation, covariance matrix estimation, and hypothesis testing.

The most intuitive way to think of the BRMR is as a modified version of the GNR. The ordinary GNR for the nonlinear regression model (4.11) is (4.17). However, it is inappropriate to use this GNR, because the disturbances are heteroskedastic, with variance given by (4.13). We need to divide the regressand and regressors of (4.17) by the square root of (4.13) in order to obtain an artificial regression with homoskedastic disturbances. The result is the BRMR,

$$V_t^{-1/2}(\boldsymbol{\beta})\big(y_t - F(\boldsymbol{X}_t\boldsymbol{\beta})\big) = V_t^{-1/2}(\boldsymbol{\beta})f(\boldsymbol{X}_t\boldsymbol{\beta})\boldsymbol{X}_t\boldsymbol{b} + \text{residual}, \tag{4.20}$$

where $V_t(\boldsymbol{\beta}) \equiv F(\boldsymbol{X}_t\boldsymbol{\beta})\big(1 - F(\boldsymbol{X}_t\boldsymbol{\beta})\big)$.

If the BRMR is evaluated at the vector of ML estimates $\hat{\boldsymbol{\beta}}$, it yields the covariance matrix

$$s^2\big(\boldsymbol{X}^\top\boldsymbol{\Upsilon}(\hat{\boldsymbol{\beta}})\boldsymbol{X}\big)^{-1}, \tag{4.21}$$

[2] This regression was originally proposed, independently in somewhat different forms, by Engle (1984) and Davidson and MacKinnon (1984b)

where s is the standard error of the artificial regression. Since (4.20) is a GLS regression, s tends to 1 asymptotically, and expression (4.21) is therefore a valid way to estimate $\text{Var}(\hat{\boldsymbol{\beta}})$. However, because there is no advantage to multiplying by a random variable that tends to 1, it is better simply to use (4.18), which may readily be obtained by dividing (4.21) by s^2.

Like other artificial regressions, the BRMR can be used as part of a numerical maximization algorithm, similar to the ones described in Section 1.4. The formula that determines $\boldsymbol{\beta}_{(j+1)}$, the value of $\boldsymbol{\beta}$ at step $j+1$, is

$$\boldsymbol{\beta}_{(j+1)} = \boldsymbol{\beta}_{(j)} + \alpha_{(j)} \boldsymbol{b}_{(j)},$$

where $\boldsymbol{b}_{(j)}$ is the vector of OLS estimates from the BRMR evaluated at $\boldsymbol{\beta}_{(j)}$, and $\alpha_{(j)}$ may be chosen in several ways. This procedure generally works very well, but a modified Newton procedure is usually even faster.

The BRMR is particularly useful for hypothesis testing. Suppose that $\boldsymbol{\beta}$ is partitioned as $[\boldsymbol{\beta}_1 \vdots \boldsymbol{\beta}_2]$, where $\boldsymbol{\beta}_1$ is a $(k-r)$–vector and $\boldsymbol{\beta}_2$ is an r–vector. If $\tilde{\boldsymbol{\beta}}$ denotes the vector of ML estimates subject to the restriction that $\boldsymbol{\beta}_2 = \boldsymbol{0}$, we can test that restriction by running the BRMR

$$\tilde{V}_t^{-1/2}(y_t - \tilde{F}_t) = \tilde{V}_t^{-1/2} \tilde{f}_t \boldsymbol{X}_{t1} \boldsymbol{b}_1 + \tilde{V}_t^{-1/2} \tilde{f}_t \boldsymbol{X}_{t2} \boldsymbol{b}_2 + \text{residual}, \tag{4.22}$$

where $\tilde{F}_t \equiv F(\boldsymbol{X}_t \tilde{\boldsymbol{\beta}})$, $\tilde{f}_t \equiv f(\boldsymbol{X}_t \tilde{\boldsymbol{\beta}})$, and $\tilde{V}_t \equiv V_t(\tilde{\boldsymbol{\beta}})$. Here $\boldsymbol{X}_t$ has been partitioned into two vectors, $\boldsymbol{X}_{t1}$ and $\boldsymbol{X}_{t2}$, corresponding to the partitioning of $\boldsymbol{\beta}$. The regressors that correspond to $\boldsymbol{\beta}_1$ are orthogonal to the regressand, while those that correspond to $\boldsymbol{\beta}_2$ are not. All the usual test statistics for $\boldsymbol{b}_2 = \boldsymbol{0}$ are valid. The best test statistic to use in finite samples is probably the explained sum of squares from regression (4.22). It is asymptotically distributed as $\chi^2(r)$ under the null hypothesis. An F statistic is also asymptotically valid, but since its denominator of s^2 is random, and there is no need to estimate the variance of (4.22), the explained sum of squares is preferable.

In the special case of the null hypothesis that all the slope coefficients are zero, regression (4.22) simplifies dramatically. In this case, $\boldsymbol{X}_{t1}$ is just unity, and $\tilde{V}_t$, $\tilde{F}_t$, and $\tilde{f}_t$ are all constants that do not depend on t. Since neither subtracting a constant from the regressand nor multiplying the regressand and regressors by a constant has any effect on the F statistic for $\boldsymbol{b}_2 = \boldsymbol{0}$, regression (4.22) is equivalent to the much simpler regression

$$\boldsymbol{y} = c_1 + \boldsymbol{X}_2 \boldsymbol{c}_2 + \text{residuals}. \tag{4.23}$$

The ordinary F statistic for $\boldsymbol{c}_2 = \boldsymbol{0}$ in regression (4.23) is an asymptotically valid test statistic for the hypothesis that $\boldsymbol{\beta}_2 = \boldsymbol{0}$. The fact that (4.23) is just an OLS regression of $\boldsymbol{y}$ on the constant and explanatory variables accounts for the claim we made in Section 4.2 that such a regression is not always completely useless!

Bootstrap Inference

Because binary response models are fully parametric, it is straightforward to bootstrap them using procedures similar to those discussed in Part 1, Section 7.4 For the model specified by (4.01), the bootstrap DGP is required to generate binary variables y_t^*, $t = 1, \ldots, n$, in such a way that

$$P_t^* \equiv \text{E}(y_t^* \mid \boldsymbol{X}_t) = F(\boldsymbol{X}_t \hat{\boldsymbol{\beta}}),$$

where $\hat{\boldsymbol{\beta}}$ is a vector of ML estimates. For a bootstrap test, this vector would be subject to whatever restrictions are being tested. In order to generate y_t^*, the easiest way to proceed is to draw u_t^* from the uniform distribution $\text{U}(0,1)$ and set $y_t^* = \text{I}(u_t^* \leq P_t^*)$, where, as usual, $\text{I}(\cdot)$ is an indicator function. Alternatively, in the case of the probit model, we can generate bootstrap samples by using (4.04) to generate latent variables and (4.05) to convert these to the binary dependent variables we actually need.

Bootstrap methods for binary response models may or may not yield more accurate inferences than asymptotic ones. In the case of test statistics, where the bootstrap samples must be generated under the null hypothesis, there seems to be evidence that bootstrap P values are generally more accurate than asymptotic ones. The value of bootstrapping appears to be particularly great when the number of restrictions is large and the sample size is moderate. However, in the case of confidence intervals, the evidence is rather mixed.

The bootstrap can also be used to reduce the bias of the ML estimates. As we saw in Part 1, Section 3.6, regression models tend to fit too well in finite samples, in the sense that the residuals tend to be smaller than the true disturbances. Binary response models also tend to fit too well, in the sense that the fitted probabilities, the $F(\boldsymbol{X}_t \hat{\boldsymbol{\beta}})$, tend to be closer to 0 and 1 than the true probabilities, the $F(\boldsymbol{X}_t \boldsymbol{\beta}_0)$. This overfitting causes the elements of $\hat{\boldsymbol{\beta}}$ to be biased away from zero.

If we generate B bootstrap samples using the parameter vector $\hat{\boldsymbol{\beta}}$, we can estimate the bias using

$$\text{Bias}^*(\hat{\boldsymbol{\beta}}) = \frac{1}{B} \sum_{j=1}^{B} \hat{\boldsymbol{\beta}}_j^* - \hat{\boldsymbol{\beta}},$$

where $\hat{\boldsymbol{\beta}}_j^*$ is the estimate of $\boldsymbol{\beta}$ using the j^{th} bootstrap sample. Therefore, a bias-corrected estimate is

$$\hat{\boldsymbol{\beta}}_{\text{bc}} \equiv \hat{\boldsymbol{\beta}} - \text{Bias}^*(\hat{\boldsymbol{\beta}}) = 2\hat{\boldsymbol{\beta}} - \frac{1}{B} \sum_{j=1}^{B} \hat{\boldsymbol{\beta}}_j^*.$$

Simulation results in MacKinnon and Smith (1998), which are by no means definitive, suggest that this estimator is less biased and has smaller mean squared error than the usual ML estimator.

The finite-sample bias of the ML estimator in binary response models can cause an important practical problem for the bootstrap. Since the probabilities associated with $\hat{\boldsymbol{\beta}}$ tend to be more extreme than the true ones, samples generated using $\hat{\boldsymbol{\beta}}$ are more prone to having a perfect classifier. Therefore, even though there is no perfect classifier for the original data, there may well be perfect classifiers for some of the bootstrap samples. The simplest way to deal with this problem is just to throw away any bootstrap samples for which a perfect classifier exists. However, if there is more than a handful of such samples, the bootstrap results must then be viewed with skepticism.

Specification Tests

Maximum likelihood estimation of binary response models almost always yields inconsistent estimates if the form of the transformation function, that is, $F(\boldsymbol{X}_t\boldsymbol{\beta})$, is misspecified. It is therefore very important to test whether this function has been specified correctly.

In Section 4.2, we derived the probit model by starting with the latent variable model (4.04), which has normally distributed, homoskedastic disturbances. A more general specification for a latent variable model, which allows for the disturbances to be heteroskedastic, is

$$y_t^\circ = \boldsymbol{X}_t\boldsymbol{\beta} + u_t, \quad u_t \sim \mathrm{N}\big(0, \exp(2\boldsymbol{Z}_t\boldsymbol{\gamma})\big), \tag{4.24}$$

where $\boldsymbol{Z}_t$ is a row vector of dimension r of observations on variables that belong to the information set Ω_t, and $\boldsymbol{\gamma}$ is an r–vector of parameters to be estimated along with $\boldsymbol{\beta}$. To ensure that both $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are identifiable, $\boldsymbol{Z}_t$ must not include a constant term or the equivalent. With this precaution, the model (4.04) is obtained by setting $\boldsymbol{\gamma} = \mathbf{0}$. Combining (4.24) with (4.05) yields the model

$$P_t \equiv \mathrm{E}(y_t \,|\, \Omega_t) = \Phi\left(\frac{\boldsymbol{X}_t\boldsymbol{\beta}}{\exp(\boldsymbol{Z}_t\boldsymbol{\gamma})}\right),$$

in which P_t depends on both the regression function $\boldsymbol{X}_t\boldsymbol{\beta}$ and the skedastic function $\exp(2\boldsymbol{Z}_t\boldsymbol{\gamma})$. Thus it is clear that heteroskedasticity of the u_t in a latent variable model affects the form of the transformation function.

Even when the binary response model being used is not the probit model, it still seems quite reasonable to consider the alternative hypothesis

$$P_t = F\left(\frac{\boldsymbol{X}_t\boldsymbol{\beta}}{\exp(\boldsymbol{Z}_t\boldsymbol{\gamma})}\right). \tag{4.25}$$

We can test against this alternative by using a BRMR to test the hypothesis that $\boldsymbol{\gamma} = \mathbf{0}$. The appropriate BRMR is

$$\tilde{V}_t^{-1/2}(y_t - \tilde{F}_t) = \tilde{V}_t^{-1/2}\tilde{f}_t\boldsymbol{X}_t\boldsymbol{b} - \tilde{V}_t^{-1/2}\boldsymbol{X}_t\tilde{\boldsymbol{\beta}}\tilde{f}_t\boldsymbol{Z}_t\boldsymbol{c} + \text{residual}, \tag{4.26}$$

where $\tilde{F}_t$, $\tilde{f}_t$, and $\tilde{V}_t$ are evaluated at the ML estimates $\tilde{\boldsymbol{\beta}}$ computed under the null hypothesis that $\boldsymbol{\gamma} = \mathbf{0}$ in (4.25). These are just the ordinary estimates for the binary response model defined by $P_t = F(\boldsymbol{X}_t\boldsymbol{\beta})$; they are usually probit or logit estimates. The explained sum of squares from (4.26) is asymptotically distributed as $\chi^2(r)$ under the null hypothesis.

Heteroskedasticity is not the only phenomenon that may lead the transformation function $F(\boldsymbol{X}_t\boldsymbol{\beta})$ to be specified incorrectly. Consider the family of models for which

$$P_t \equiv \mathrm{E}(y_t \,|\, \Omega_t) = F\left(\frac{\tau(\delta\boldsymbol{X}_t\boldsymbol{\beta})}{\delta}\right), \tag{4.27}$$

where δ is a scalar parameter, and $\tau(\cdot)$ may be any scalar function that is monotonically increasing in its argument and satisfies the conditions

$$\tau(0) = 0, \;\; \tau'(0) = 1, \;\; \text{and} \;\; \tau''(0) \neq 0, \tag{4.28}$$

where $\tau'(0)$ and $\tau''(0)$ are the first and second derivatives of $\tau(x)$, evaluated at $x = 0$. The family of models (4.27) allows for a wide range of transformation functions. It was considered by MacKinnon and Magee (1990), who showed, by using l'Hôpital's Rule, that

$$\lim_{\delta\to 0}\left(\frac{\tau(\delta x)}{\delta}\right) = x \quad \text{and} \quad \lim_{\delta\to 0}\left(\frac{\partial\big(\tau(\delta x)/\delta\big)}{\partial\delta}\right) = \frac{1}{2}x^2\tau''(0). \tag{4.29}$$

Hence the BRMR for testing the null hypothesis that $\delta = 0$ is

$$\tilde{V}_t^{-1/2}(y_t - \tilde{F}_t) = \tilde{V}_t^{-1/2}\tilde{f}_t\boldsymbol{X}_t\boldsymbol{b} + \tilde{V}_t^{-1/2}(\boldsymbol{X}_t\tilde{\boldsymbol{\beta}})^2\tilde{f}_t d + \text{residual}, \tag{4.30}$$

where everything is evaluated at the ML estimates $\tilde{\boldsymbol{\beta}}$ of the ordinary binary response model that (4.27) reduces to when $\delta = 0$. The constant factor $\tau''(0)/2$ that arises from (4.29) is irrelevant for testing and has been omitted. Thus regression (4.30) simply treats the squared values of the index function evaluated at $\tilde{\boldsymbol{\beta}}$ as if they were observations on a possibly omitted regressor, and the ordinary t statistic for $d = 0$ provides an asymptotically valid test.[3]

Tests based on the BRMRs (4.26) and (4.30) are valid only asymptotically. It is extremely likely that their finite-sample performance could be improved by using bootstrap P values instead of asymptotic ones. Since, in both cases, the null hypothesis is just an ordinary binary response model, computing bootstrap P values by using the procedures discussed in the previous subsection is quite straightforward.

[3] There is a strong resemblance between regression (4.30) and the test regression for the RESET test (Ramsey, 1969), in which squared fitted values are added to an OLS regression as a test for functional form. As MacKinnon and Magee (1990) showed, this resemblance is not coincidental.

4.4 Models for More Than Two Discrete Responses

Discrete dependent variables that can take on three or more different values are by no means uncommon in economics, and a large number of models has been devised to deal with such cases. These are sometimes referred to as **qualitative response models** and sometimes as **discrete choice models**. The binary response models we have already studied are special cases.

Discrete choice models can be divided into two types: ones designed to deal with **ordered responses**, and ones designed to deal with **unordered responses**. Surveys often produce ordered response data. For example, respondents might be asked whether they strongly agree, agree, neither agree nor disagree, disagree, or strongly disagree with some statement. Here there are five possible responses, which evidently can be ordered in a natural way. In many other cases, however, there is no natural way to order the various choices. A classic example is the choice of transportation mode. For intercity travel, people often have a choice among flying, driving, taking the train, and taking the bus. There is no natural way to order these four choices.

The Ordered Probit Model

The most widely-used model for ordered response data is the **ordered probit model**. This model can easily be derived from a latent variable model. The model for the latent variable is

$$y_t^\circ = \boldsymbol{X}_t\boldsymbol{\beta} + u_t, \quad u_t \sim \text{NID}(0,1), \tag{4.31}$$

which is identical to the latent variable model (4.04) that led to the ordinary probit model. As in the case of the latter, what we actually observe is a discrete variable y_t that can take on a limited, known, number of values. For simplicity, we assume that the number of values is just 3. It should be obvious how to extend the model to cases in which y_t can take on any known number of values.

The relation between the observed variable y_t and the latent variable y_t° is assumed to be given by

$$\begin{aligned} y_t &= 0 \text{ if } y_t^\circ < \gamma_1; \\ y_t &= 1 \text{ if } \gamma_1 \le y_t^\circ < \gamma_2; \\ y_t &= 2 \text{ if } y_t^\circ \ge \gamma_2. \end{aligned} \tag{4.32}$$

Thus $y_t = 0$ for small values of y_t°, $y_t = 1$ for intermediate values, and $y_t = 2$ for large values. The boundaries between the three cases are determined by the parameters γ_1 and γ_2. These **threshold parameters**, which usually must be estimated, determine how the values of y_t° get translated into the three possible values of y_t. It is essential that $\gamma_2 > \gamma_1$. Otherwise, the first and last lines of (4.32) would be incompatible, and we could never observe $y_t = 1$.

If $\boldsymbol{X}_t$ contains a constant term, it is impossible to identify the constant along with γ_1 and γ_2. To see this, suppose that the constant is equal to α. Then it is easy to check that y_t is unchanged if we replace the constant by $\alpha + \delta$ and replace γ_i by $\gamma_i + \delta$ for $i = 1, 2$. The easiest, but not the only, solution to this identification problem is just to set $\alpha = 0$. We adopt this solution here. In general, with no constant, the ordered probit model has as many threshold parameters as choices, less one. When there are just two choices, the single threshold parameter is equivalent to a constant, and the ordered probit model reduces to the ordinary probit model, with a constant.

In order to work out the loglikelihood function for this model, we need the probabilities of the three events $y_t = 0$, $y_t = 1$, and $y_t = 2$. The probability that $y_t = 0$ is

$$\begin{aligned} \Pr(y_t = 0) &= \Pr(y_t^\circ < \gamma_1) = \Pr(\boldsymbol{X}_t\boldsymbol{\beta} + u_t < \gamma_1) \\ &= \Pr(u_t < \gamma_1 - \boldsymbol{X}_t\boldsymbol{\beta}) = \Phi(\gamma_1 - \boldsymbol{X}_t\boldsymbol{\beta}). \end{aligned}$$

Similarly, the probability that $y_t = 2$ is

$$\begin{aligned} \Pr(y_t = 2) &= \Pr(y_t^\circ \ge \gamma_2) = \Pr(\boldsymbol{X}_t\boldsymbol{\beta} + u_t \ge \gamma_2) \\ &= \Pr(u_t \ge \gamma_2 - \boldsymbol{X}_t\boldsymbol{\beta}) = \Phi(\boldsymbol{X}_t\boldsymbol{\beta} - \gamma_2). \end{aligned}$$

Finally, the probability that $y_t = 1$ is

$$\begin{aligned} \Pr(y_t = 1) &= 1 - \Pr(y_t = 0) - \Pr(y_t = 2) \\ &= 1 - \Phi(\gamma_1 - \boldsymbol{X}_t\boldsymbol{\beta}) - \Phi(\boldsymbol{X}_t\boldsymbol{\beta} - \gamma_2) \\ &= \Phi(\gamma_2 - \boldsymbol{X}_t\boldsymbol{\beta}) - \Phi(\gamma_1 - \boldsymbol{X}_t\boldsymbol{\beta}). \end{aligned}$$

These probabilities depend solely on the value of the index function, $\boldsymbol{X}_t\boldsymbol{\beta}$, and on the two threshold parameters.

The loglikelihood function for the ordered probit model derived from (4.31) and (4.32) is

$$\begin{aligned} \ell(\boldsymbol{\beta}, \gamma_1, \gamma_2) = &\sum_{y_t=0} \log\big(\Phi(\gamma_1 - \boldsymbol{X}_t\boldsymbol{\beta})\big) + \sum_{y_t=2} \log\big(\Phi(\boldsymbol{X}_t\boldsymbol{\beta} - \gamma_2)\big) \\ &+ \sum_{y_t=1} \log\big(\Phi(\gamma_2 - \boldsymbol{X}_t\boldsymbol{\beta}) - \Phi(\gamma_1 - \boldsymbol{X}_t\boldsymbol{\beta})\big). \end{aligned} \tag{4.33}$$

Maximizing (4.33) numerically is generally not difficult to do, although steps may have to be taken to ensure that γ_2 is always greater than γ_1. Note that the function Φ in (4.33) may be replaced by any function F that satisfies the conditions (4.02), although it may then be harder to derive the probabilities from a latent variable model. Thus the ordered probit model is by no means the only qualitative response model for ordered data.

The ordered probit model is widely used in applied econometric work. A simple, graphical exposition of this model is provided by Becker and Kennedy (1992). Like the ordinary probit model, the ordered probit model can be generalized in a number of ways; see, for example, Terza (1985). An interesting application of a generalized version, which allows for heteroskedasticity, is Hausman, Lo, and MacKinlay (1992). They apply the model to price changes on the New York Stock Exchange at the level of individual trades. Because the price change from one trade to the next almost always takes on one of a small number of possible values, an ordered probit model is an appropriate way to model these changes.

The Multinomial Logit Model

The key feature of ordered qualitative response models like the ordered probit model is that all the choices depend on a single index function. This makes sense only when the responses have a natural ordering. A different sort of model is evidently necessary to deal with unordered responses. The most popular of these is the **multinomial logit model**, sometimes called the **multiple logit model**, which has been widely used in applied work.

The multinomial logit model is designed to handle $J+1$ responses, for $J \geq 1$. According to this model, the probability that any one of them is observed is

$$\Pr(y_t = l) = \frac{\exp(\boldsymbol{W}_{tl}\boldsymbol{\beta}^l)}{\sum_{j=0}^{J} \exp(\boldsymbol{W}_{tj}\boldsymbol{\beta}^j)} \quad \text{for } l = 0, \dots, J. \tag{4.34}$$

Here $\boldsymbol{W}_{tj}$ is a row vector of dimension k_j of observations on variables that belong to the information set of interest, and $\boldsymbol{\beta}^j$ is a k_j–vector of parameters, usually different for each $j = 0, \dots, J$.

Estimation of the multinomial logit model is reasonably straightforward. The loglikelihood function can be written as

$$\sum_{t=1}^{n} \Big(\sum_{j=0}^{J} \mathrm{I}(y_t = j) \boldsymbol{W}_{tj}\boldsymbol{\beta}^j - \log\Big(\sum_{j=0}^{J} \exp(\boldsymbol{W}_{tj}\boldsymbol{\beta}^j) \Big) \Big), \tag{4.35}$$

where $\mathrm{I}(\cdot)$ is the indicator function. Thus each observation contributes two terms to the loglikelihood function. The first is $\boldsymbol{W}_{tj}\boldsymbol{\beta}^j$, where $y_t = j$, and the second is minus the logarithm of the denominator that appears in (4.34). It is generally not difficult to maximize (4.35) by using some sort of modified Newton method, provided there are no perfect classifiers, since the loglikelihood function (4.35) is globally concave with respect to the entire vector of parameters, $[\boldsymbol{\beta}^0 \vdots \dots \vdots \boldsymbol{\beta}^J]$; see Exercise 4.16.

Some special cases of the multinomial logit model are of interest. One of these arises when the explanatory variables $\boldsymbol{W}_{tj}$ are the same for each choice j. If a model is intended to explain which of an unordered set of outcomes applies

to the different individuals in a sample, then the probabilities of all of these outcomes can be expected to depend on the same set of characteristics for each individual. For instance, a student wondering how to spend Saturday night may be able to choose among studying, partying, visiting parents, or going to the movies. In choosing, the student takes into account things like grades on the previous midterm, the length of time since the last visit home, the interest of what is being shown at the local movie theater, and so on. All these variables affect the probability of each possible outcome.

For models of this sort, it is not possible to identify $J+1$ parameter vectors $\boldsymbol{\beta}^j$, $j = 0, \dots, J$. To see this, let $\boldsymbol{X}_t$ denote the common set of explanatory variables for observation t, and define $\boldsymbol{\gamma}^j \equiv \boldsymbol{\beta}^j - \boldsymbol{\beta}^0$ for $j = 1, \dots, J$. On replacing the $\boldsymbol{W}_{tj}$ by $\boldsymbol{X}_t$ for all j, the probabilities defined in (4.34) become, for $l = 1, \dots, J$,

$$\Pr(y_t = l) = \frac{\exp(\boldsymbol{X}_t\boldsymbol{\beta}^l)}{\sum_{j=0}^{J} \exp(\boldsymbol{X}_t\boldsymbol{\beta}^j)} = \frac{\exp(\boldsymbol{X}_t\boldsymbol{\gamma}^l)}{1 + \sum_{j=1}^{J} \exp(\boldsymbol{X}_t\boldsymbol{\gamma}^j)},$$

where the second equality is obtained by dividing both the numerator and the denominator by $\exp(\boldsymbol{X}_t\boldsymbol{\beta}^0)$. For outcome 0, the probability is just

$$\Pr(y_t = 0) = \frac{1}{1 + \sum_{j=1}^{J} \exp(\boldsymbol{X}_t\boldsymbol{\gamma}^j)}.$$

It follows that all $J+1$ probabilities can be expressed in terms of the parameters $\boldsymbol{\gamma}^j$, $j = 1, \dots, J$, independently of $\boldsymbol{\beta}^0$. In practice, it is easiest to impose the restriction that $\boldsymbol{\beta}^0 = \mathbf{0}$, which is then enough to identify the parameters $\boldsymbol{\beta}^j$, $j = 1, \dots, J$. When $J = 1$, it is easy to see that this model reduces to the ordinary logit model with a single index function $\boldsymbol{X}_t\boldsymbol{\beta}^1$.

In certain cases, some but not all of the explanatory variables are common to all outcomes. In that event, for the common variables, a separate parameter cannot be identified for each outcome, for the same reason as above. In order to set up a model for which all the parameters are identified, it is necessary to set to zero those components of $\boldsymbol{\beta}^0$ that correspond to the common variables. Thus, for instance, at most J of the $\boldsymbol{W}_{tj}$ vectors can include a constant.

Another special case of interest is the so-called **conditional logit model**. For this model, the probability that agent t makes choice l is

$$\Pr(y_t = l) = \frac{\exp(\boldsymbol{W}_{tl}\boldsymbol{\beta})}{\sum_{j=0}^{J} \exp(\boldsymbol{W}_{tj}\boldsymbol{\beta})}. \tag{4.36}$$

where $\boldsymbol{W}_{tj}$ is a row vector with k components for each $j = 0, \dots, J$, and $\boldsymbol{\beta}$ is a k–vector of parameters, the same for each j. This model has been extensively used to model the choice among competing modes of transportation. The usual interpretation is that the elements of $\boldsymbol{W}_{tj}$ are the characteristics of

choice j for agent t, and agents make their choice by considering the weighted sums $\boldsymbol{W}_{tj}\boldsymbol{\beta}$ of these characteristics.

It is necessary that none of the explanatory variables in the $\boldsymbol{W}_{tj}$ vectors should be the same for all $J+1$ choices. In other words, no single variable should appear in each and every $\boldsymbol{W}_{tj}$. It is easy to see from (4.36) that, if there were such a variable, say w_{ti}, for some $i = 1, \ldots, k$, then this variable would be multiplied by the *same* parameter β_i for each choice. In consequence, the factor $\exp(w_{ti}\beta_i)$ would appear in the numerator and in every term of the denominator of (4.36) and could be cancelled out. This implies, in particular, that none of the explanatory variables can be constant for all $t = 1, \ldots, n$ and all $j = 0, \ldots, J$.

An important property of the general multinomial logit model defined by the set of probabilities (4.34) is that

$$\frac{\Pr(y_t = l)}{\Pr(y_t = j)} = \frac{\exp(\boldsymbol{W}_{tl}\boldsymbol{\beta}^l)}{\exp(\boldsymbol{W}_{tj}\boldsymbol{\beta}^j)}.$$

for any two responses l and j. Therefore, the ratio of the probabilities of any two responses depends solely on the explanatory variables $\boldsymbol{W}_{tl}$ and $\boldsymbol{W}_{tj}$ and the parameters $\boldsymbol{\beta}^l$ and $\boldsymbol{\beta}^j$ associated with those two responses. It does not depend on the explanatory variables or parameter vectors specific to any of the other responses. This property of the model is called the **independence of irrelevant alternatives**, or **IIA**, property.

The IIA property is often quite implausible. For example, suppose there are three modes of public transportation between a pair of cities: the bus, which is slow but cheap, the airplane, which is fast but expensive, and the train, which is a little faster than the bus and a lot cheaper than the airplane. Now consider what the model says would happen if the rail line were upgraded, causing the train to become much faster but considerably more expensive. Intuitively, we might expect a lot of people who previously flew to take the train instead, but relatively few to switch from the bus to the train. However, this is not what the model says. Instead, the IIA property implies that the ratio of travelers who fly to travelers who take the bus is the same whatever the characteristics of the train.

Although the IIA property is often not a plausible one, it can easily be tested; see Hausman and McFadden (1984), McFadden (1987), and Exercise 4.22. The simplicity of the multinomial logit model, despite the IIA property, makes this model very attractive for cases in which it does not appear to be incompatible with the data.

The Nested Logit Model

A discrete choice model that does not possess the IIA property is the **nested logit model**. For this model, the set of possible choices is decomposed into subsets. Let the set of outcomes $\{0, 1, \ldots, J\}$ be partitioned into m disjoint

subsets A_i, $i = 1, \ldots, m$. The model then supposes that, conditional on choosing an outcome in subset A_i, the choice among the members of A_i is governed by a standard multinomial logit model. We have, for $j \in A_i$, that

$$\Pr(y_t = j \,|\, y_t \in A_i) = \frac{\exp(\boldsymbol{W}_{tj}\boldsymbol{\beta}^j/\theta_i)}{\sum_{l\in A_i}\exp(\boldsymbol{W}_{tl}\boldsymbol{\beta}^l/\theta_i)}. \tag{4.37}$$

It is clear that the parameter θ_i, which can be thought of as a scale parameter for the parameter vectors $\boldsymbol{\beta}^j$, $j \in A_i$, is not identifiable on the basis of choice within the elements of subset A_i. However, it is what determines the probability of choosing some element in A_i. Specifically, we assume that

$$\Pr(y_t \in A_i) = \frac{\exp(\theta_i h_{ti})}{\sum_{k=1}^m \exp(\theta_k h_{tk})}, \tag{4.38}$$

where we have defined the **inclusive value** of subset A_i as:

$$h_{ti} = \log\Big(\sum_{j\in A_i}\exp(\boldsymbol{W}_{tj}\boldsymbol{\beta}^j/\theta_i)\Big). \tag{4.39}$$

Since it follows at once from (4.38) that $\sum_{i=1}^m \Pr(y_t \in A_i) = 1$, we can see that y_t must belong to one of the disjoint sets A_i.

By putting together (4.37) and (4.38), we obtain the $J+1$ probabilities for the different outcomes. For each $j = 0, \ldots, J$, let $i(j)$ be the subset containing j. In other words, $j \in A_{i(j)}$. Then we have that

$$\begin{aligned}\Pr(y_t = j) &= \Pr(y_t = j \,|\, y_t \in A_{i(j)})\Pr(y_t \in A_{i(j)}) \\ &= \frac{\exp(\boldsymbol{W}_{tj}\boldsymbol{\beta}^j/\theta_{i(j)})}{\sum_{l\in A_{i(j)}}\exp(\boldsymbol{W}_{tl}\boldsymbol{\beta}^l/\theta_{i(j)})}\,\frac{\exp(\theta_{i(j)}h_{ti(j)})}{\sum_{k=1}^m\exp(\theta_k h_{tk})}.\end{aligned} \tag{4.40}$$

It is not hard to check that, if $\theta_i = 1$ for all $i = 1, \ldots, m$, the probabilities (4.40) reduce to the probabilities (4.34) of the usual multinomial logit model; see Exercise 4.17. Thus the multinomial logit model is contained within the nested logit model as a special case. It follows, therefore, that testing the multinomial logit model against the alternative of the nested logit model, for some appropriate choice of the subsets A_i, is one way to test whether the IIA property is compatible with the data.

An Artificial Regression for Discrete Choice Models

In order to perform the test of the IIA property mentioned just above, and to perform inference generally in the context of discrete choice models, it is convenient to be able to make use of an artificial regression. The simplest such artificial regression was proposed by McFadden (1987) for multinomial

logit models. In this section, we present a generalized version that can be applied to any discrete choice model. We call this the **discrete choice artificial regression**, or **DCAR**.

As usual, we assume that there are $J+1$ possible outcomes, numbered from $j=0$ to $j=J$. Let the probability of choosing outcome j for observation t be given by the function $\Pi_{tj}(\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a k–vector of parameters. For the multinomial logit model, $\boldsymbol{\theta}$ would include all of the independent parameters in the set of parameter vectors $\boldsymbol{\beta}^j$, $j=0,\dots,J$. The function $\Pi_{tj}(\cdot)$ usually also depends on exogenous or predetermined explanatory variables that are not made explicit in the notation. We require that $\sum_{j=0}^{J}\Pi_{tj}(\boldsymbol{\theta})=1$ for all $t=1,\dots,n$ and for all admissible parameter vectors $\boldsymbol{\theta}$, in order that the set of $J+1$ outcomes should be exhaustive.

For each observation t, $t=1,\dots,n$, define the $J+1$ indicator variables d_{tj} as $d_{tj}=\mathrm{I}(y_t=j)$. Then the loglikelihood function of the discrete choice model is given by

$$\ell(\boldsymbol{y},\boldsymbol{\theta})=\sum_{t=1}^{n}\sum_{j=0}^{J}d_{tj}\log\Pi_{tj}(\boldsymbol{\theta}). \tag{4.41}$$

Just as for the loglikelihood functions (4.09) and (4.35), the contribution made by observation t is the logarithm of the probability that y_t should have taken on its observed value.

The DCAR has $n(J+1)$ "observations," $J+1$ for each real observation. For observation t, the $J+1$ components of the regressand, evaluated at $\boldsymbol{\theta}$, are given by $\Pi_{tj}^{-1/2}(\boldsymbol{\theta})\big(d_{tj}-\Pi_{tj}(\boldsymbol{\theta})\big)$, $j=0,\dots,J$. The components of the regressor corresponding to parameter θ_i, $i=1,\dots,k$, are given by $\Pi_{tj}^{-1/2}(\boldsymbol{\theta})\partial\Pi_{tj}(\boldsymbol{\theta})/\partial\theta_i$. Thus the DCAR may be written as

$$\Pi_{tj}^{-1/2}(\boldsymbol{\theta})\big(d_{tj}-\Pi_{tj}(\boldsymbol{\theta})\big)=\Pi_{tj}^{-1/2}(\boldsymbol{\theta})\boldsymbol{T}_{tj}(\boldsymbol{\theta})\boldsymbol{b}+\text{residual}, \tag{4.42}$$

for $t=1,\dots,n$ and $j=0,\dots,J$. Here $\boldsymbol{T}_{tj}(\boldsymbol{\theta})$ denotes the $1\times k$ vector of the partial derivatives of $\Pi_{tj}(\boldsymbol{\theta})$ with respect to the components of $\boldsymbol{\theta}$, and, as usual, $\boldsymbol{b}$ is a k–vector of artificial parameters. It is easy to see that the scalar product of the regressand and the regressor corresponding to θ_i is

$$\sum_{t=1}^{n}\sum_{j=0}^{J}\frac{\big(d_{tj}-\Pi_{tj}(\boldsymbol{\theta})\big)\partial\Pi_{tj}(\boldsymbol{\theta})/\partial\theta_i}{\Pi_{tj}(\boldsymbol{\theta})}. \tag{4.43}$$

The derivative of the loglikelihood function (4.41) with respect to θ_i is

$$\frac{\partial\ell(\boldsymbol{\theta})}{\partial\theta_i}=\sum_{t=1}^{n}\sum_{j=0}^{J}d_{tj}\frac{\partial\Pi_{tj}(\boldsymbol{\theta})/\partial\theta_i}{\Pi_{tj}(\boldsymbol{\theta})},$$

and we can see that this is equal to (4.43), because differentiating the identity $\sum_{j=0}^{J}\Pi_{tj}(\boldsymbol{\theta})=1$ with respect to θ_i shows that $\sum_{j=0}^{J}\partial\Pi_{tj}(\boldsymbol{\theta})/\partial\theta_i=0$.

It follows that the regressand is orthogonal to all the regressors when all the artificial variables are evaluated at the maximum likelihood estimates $\hat{\boldsymbol{\theta}}$.

In Exercises 4.18 and 4.19, readers are asked to show that regression (4.42), the DCAR, satisfies the other requirements for an artificial regression used for hypothesis testing. See also Exercise 4.22, in which readers are asked to implement by artificial regression the test of the IIA property discussed at the end of the previous subsection.

As with binary response models, it is easy to bootstrap discrete choice models, because they are fully parametrically specified. For the model characterized by the loglikelihood function (4.41), an easy way to implement the bootstrap DGP is, first, to construct the cumulative probabilities $P_{tj}(\hat{\boldsymbol{\theta}})\equiv\sum_{i=0}^{j}\Pi_{ti}(\hat{\boldsymbol{\theta}})$, for $j=0,\dots,J-1$, and then to draw a random number, u_t^* say for observation t, from the uniform distribution $\mathrm{U}(0,1)$. The bootstrap dependent variable y_t^* is then set equal to

$$y_t^*=\sum_{j=0}^{J-1}\mathrm{I}\big(u_t^*\geq P_{tj}(\hat{\boldsymbol{\theta}})\big).$$

All of the indicator functions in the above sum are zero if $u_t^*<P_{t0}(\hat{\boldsymbol{\theta}})=\Pi_{t0}(\hat{\boldsymbol{\theta}})$, an event which occurs with probability $\Pi_{t0}(\hat{\boldsymbol{\theta}})$, as desired. Similarly, $y_t^*=j$ for $j=1,\dots,J$ if and only if $P_{t(j-1)}(\hat{\boldsymbol{\theta}})\leq u_t^*<P_{tj}(\hat{\boldsymbol{\theta}})$, an event that occurs with probability $\Pi_{tj}(\hat{\boldsymbol{\theta}})=P_{tj}(\hat{\boldsymbol{\theta}})-P_{t(j-1)}(\hat{\boldsymbol{\theta}})$.

The Multinomial Probit Model

Another discrete choice model that can sometimes be used when the IIA property is unacceptable is the **multinomial probit model**. This model is theoretically attractive but computationally burdensome. The $J+1$ possible outcomes are generated by the latent variable model

$$y_{tj}^{\circ}=\boldsymbol{W}_{tj}\boldsymbol{\beta}^j+u_{tj},\qquad \boldsymbol{u}_t\sim\mathrm{N}(\boldsymbol{0},\boldsymbol{\Omega}), \tag{4.44}$$

where the y_{tj}° are not observed, and $\boldsymbol{u}_t$ is a $1\times(J+1)$ vector with typical element u_{tj}. What we observe are the binary variables y_{tj}, which are assumed to be determined as follows:

$$\begin{aligned}&y_{tj}=1 \text{ if } y_{tj}^{\circ}-y_{ti}^{\circ}\geq 0 \text{ for all } i=0,\dots,J,\\ &y_{tj}=0 \text{ otherwise.}\end{aligned} \tag{4.45}$$

As with the multinomial logit model, separate coefficients cannot be identified for all $J+1$ outcomes if an explanatory variable is common to all of the index functions $\boldsymbol{W}_{tj}\boldsymbol{\beta}^j$. The solution to this problem is the same as before: We set the components of $\boldsymbol{\beta}^0$ equal to 0 for all such variables.

It is clear from (4.45) that the observed y_{tj} depend only on the differences $y^\circ_{tj} - y^\circ_{t0}$, $j = 1, \dots, J$. Let z°_{tj} be equal to this difference. Then

$$\begin{aligned} &y_{tj} = 1 \text{ if } z^\circ_{tj} \geq z^\circ_{ti} \text{ for all } i = 1, \dots, J, \text{ and } z^\circ_{tj} \geq 0, \\ &y_{tj} = 0 \text{ otherwise.} \end{aligned} \tag{4.46}$$

Thus the probabilities $\Pr(y_{tj} = 1)$ are completely determined by the joint distribution of the z°_{tj}. We write the covariance matrix of this distribution as $\boldsymbol{\Sigma}$, where $\boldsymbol{\Sigma}$ is a $J \times J$ symmetric positive definite matrix, uniquely determined by the $(J+1) \times (J+1)$ matrix $\boldsymbol{\Omega}$ of (4.44), although $\boldsymbol{\Omega}$ is not uniquely determined by $\boldsymbol{\Sigma}$. It follows that the matrix $\boldsymbol{\Omega}$ cannot be identified on the basis of the observed variables y_t alone.

In fact, even $\boldsymbol{\Sigma}$ is identified only up to scale. This can be seen by observing that, if all the z°_{tj} in (4.46) are multiplied by the same positive constant, the values of the y_{tj} remain unchanged. In practice, it is customary to set the first diagonal element of $\boldsymbol{\Sigma}$ equal to 1 in order to set the scale of $\boldsymbol{\Sigma}$. Once the scale is fixed, then the only other restriction on $\boldsymbol{\Sigma}$ is that it must be symmetric and positive definite. In particular, it may well have nonzero off-diagonal elements, and these give the multinomial probit model a flexibility that is not shared by the multinomial logit model. In consequence, the multinomial probit model does not have the IIA property.

The latent variable model (4.44) can be interpreted as a model determining the utility levels yielded by the different outcomes. Then the correlation between z°_{tj} and z°_{ti}, for $i \neq j$, might measure the extent to which a preference for flying over driving, say, is correlated with a preference for taking the train over driving. In this example of transportation mode choice, we are assuming that driving is outcome 0. It seems fair to say that, although these correlations are what provides multinomial probit with greater flexibility than multinomial logit, they are a little difficult to interpret directly.

Unfortunately, the multinomial probit model is not at all easy to estimate. The event $y_{tj} = 1$ is observed if and only if

$$y^\circ_{tj} - y^\circ_{ti} \geq 0 \text{ for all } i = 1, \dots, J+1,$$

and the probability of this event is given by a J–dimensional integral. Therefore, in order to evaluate the loglikelihood function just once, the integral corresponding to whatever event occurred must be computed for every observation in the sample. This must generally be done a large number of times during the course of whatever nonlinear optimization procedure is used. Evaluating high-dimensional integrals of the normal distribution is analytically intractable. Consequently, except when J is very small, the multinomial probit model is usually estimated by simulation-based methods. See Hajivassiliou and Ruud (1994) and Gouriéroux and Monfort (1996) for discussions of some of the methods that have been proposed.

The treatment of qualitative response models in this section has necessarily been incomplete. Detailed surveys of the older literature include Amemiya (1985, Chapter 9) and McFadden (1984). For a more up-to-date survey, but one that is relatively superficial, see Maddala and Flores-Lagunes (2001).

4.5 Models for Count Data

Many economic variables are nonnegative integers. Examples include the number of patents granted to a firm and the number of visits to the hospital by an individual, where each is measured over some period of time. Data of this type are called **event count data** or, simply, **count data**. In many cases, the count is 0 for a substantial fraction of the observations.

One might think of using an ordered discrete choice model like the ordered probit model to handle data of this type. However, this is usually not appropriate, because such a model requires the number of possible outcomes to be fixed and known. Instead, we need a model for which any nonnegative integer value is a valid, although perhaps very unlikely, value. One way to obtain such a model is to start from a distribution which has this property. The most popular distribution of this type is the **Poisson distribution**. If a discrete random variable Y follows the Poisson distribution, then

$$\Pr(Y = y) = \frac{\mathrm{e}^{-\lambda}\lambda^y}{y!}, \quad y = 0, 1, 2, \dots. \tag{4.47}$$

This distribution is characterized by a single parameter, λ. It can be shown that the probabilities (4.47) sum to 1 over $y = 0, 1, 2, \dots$, and that the expectation and the variance of a Poisson random variable are both equal to λ, which must therefore take on only positive values; see Exercise 4.23.

The Poisson Regression Model

The simplest model for count data is the **Poisson regression model**, which is obtained by replacing the parameter λ in (4.47) by a nonnegative function of regressors and parameters. The most popular choice for this function is the **exponential mean function**

$$\lambda_t(\boldsymbol{\beta}) \equiv \exp(\boldsymbol{X}_t\boldsymbol{\beta}), \tag{4.48}$$

which makes use of the linear index function $\boldsymbol{X}_t\boldsymbol{\beta}$. Other specifications for the index function, possibly nonlinear, can also be used. Because the linear index function in (4.48) is the argument of an exponential, the model specified by (4.48) is sometimes called **loglinear**, since the log of $\lambda_t(\boldsymbol{\beta})$ is linear in $\boldsymbol{\beta}$. For any valid choice of $\lambda_t(\boldsymbol{\beta})$, we obtain the Poisson regression model

$$\Pr(Y_t = y) = \frac{\exp\big(-\lambda_t(\boldsymbol{\beta})\big)\big(\lambda_t(\boldsymbol{\beta})\big)^y}{y!}, \quad y = 0, 1, 2, \dots. \tag{4.49}$$

If the observed count value for observation t is y_t, then the contribution to the loglikelihood function is the logarithm of the right-hand side of (4.49), evaluated at $y = y_t$. Therefore, the entire loglikelihood function is

$$\ell(\boldsymbol{y}, \boldsymbol{\beta}) = \sum_{t=1}^{n} \big(- \exp(\boldsymbol{X}_t\boldsymbol{\beta}) + y_t \boldsymbol{X}_t\boldsymbol{\beta} - \log y_t! \big) \tag{4.50}$$

under the exponential mean specification (4.48).

Maximizing the function (4.50) is not difficult. The likelihood equations are

$$\frac{\partial \ell(\boldsymbol{y}, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{t=1}^{n} \big(y_t - \exp(\boldsymbol{X}_t\boldsymbol{\beta}) \big) \boldsymbol{X}_t = \boldsymbol{0}, \tag{4.51}$$

and the Hessian matrix is

$$\boldsymbol{H}(\boldsymbol{\beta}) = - \sum_{t=1}^{n} \exp(\boldsymbol{X}_t\boldsymbol{\beta}) \boldsymbol{X}_t^\top \boldsymbol{X}_t = -\boldsymbol{X}^\top \boldsymbol{\Upsilon}(\boldsymbol{\beta}) \boldsymbol{X}, \tag{4.52}$$

where $\boldsymbol{\Upsilon}(\boldsymbol{\beta})$ is an $n \times n$ diagonal matrix with typical diagonal element equal to $\Upsilon_t(\boldsymbol{\beta}) \equiv \exp(\boldsymbol{X}_t\boldsymbol{\beta})$. Since $\boldsymbol{H}(\boldsymbol{\beta})$ is negative definite, optimization techniques based on Newton's Method generally work very well. Inferences may be based on the standard asymptotic result that the asymptotic covariance matrix is equal to the inverse of the information matrix. This leads to the estimator

$$\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}^\top \hat{\boldsymbol{\Upsilon}} \boldsymbol{X})^{-1}, \tag{4.53}$$

where $\hat{\boldsymbol{\Upsilon}} \equiv \boldsymbol{\Upsilon}(\hat{\boldsymbol{\beta}})$. This estimated covariance matrix looks very much like the one for weighted least-squares estimation. In fact, if we were to run the nonlinear regression

$$y_t = \exp(\boldsymbol{X}_t\boldsymbol{\beta}) + u_t \tag{4.54}$$

by weighted least squares, using weights $\Upsilon_t^{-1/2}(\boldsymbol{\beta}) = \exp(-\frac{1}{2}\boldsymbol{X}_t\boldsymbol{\beta})$, the first-order conditions, treating the weights as fixed, would be equations (4.51). Regression (4.54) is the analog for the Poisson regression model of regression (4.11) for the binary response model. Thus ML estimation of the Poisson regression model specified by (4.49), where $\lambda_t(\boldsymbol{\beta})$ is given by an exponential mean function, is seen to be equivalent to weighted NLS estimation of the nonlinear regression model (4.54).

The weighted NLS interpretation suggests that an artificial regression must be available. This is indeed the case. Just as the BRMR (4.20) is the GNR that corresponds to the weighted version of (4.11), the artificial regression for the Poisson regression model is the GNR that corresponds to the weighted version of (4.54), namely,

$$\exp(-\tfrac{1}{2}\boldsymbol{X}_t\boldsymbol{\beta}) \big(y_t - \exp(\boldsymbol{X}_t\boldsymbol{\beta}) \big) = \exp(\tfrac{1}{2}\boldsymbol{X}_t\boldsymbol{\beta}) \boldsymbol{X}_t\boldsymbol{b} + \text{residual}. \tag{4.55}$$

Like the GNR and the BRMR, this regression may be used for a number of purposes, including estimating the covariance matrix of $\hat{\boldsymbol{\beta}}$. It is particularly useful for testing restrictions on $\boldsymbol{\beta}$ without having to estimate the model more than once; see Exercise 4.25.

Testing for Overdispersion in the Poisson Regression Model

Although its simplicity makes it attractive, the Poisson regression model is rarely entirely satisfactory. In practice, even though it may predict the mean event count accurately, it frequently tends to underpredict the frequency of zeros and large counts, because the variance of the actual data is larger than the variance predicted by the Poisson model. This failure of the model is called **overdispersion**. Before accepting a Poisson regression model, even tentatively, it is highly advisable to test it for overdispersion.

Several tests for overdispersion have been proposed. The simplest of these are based on the artificial OPG regression that we introduced in Section 3.5 for models estimated by maximum likelihood. The regressand of the OPG regression is equal to 1 for each observation, and the regressors are the partial derivatives of the loglikelihood contribution with respect to the parameters. Thus observation t of the OPG regression based on the loglikelihood function (4.50) can be written as

$$1 = \big(y_t - \exp(\boldsymbol{X}_t\boldsymbol{\beta}) \big) \boldsymbol{X}_t\boldsymbol{b} + \text{residual}. \tag{4.56}$$

When the regressors in (4.56) are evaluated at the ML estimates $\hat{\boldsymbol{\beta}}$, they are orthogonal to the regressand.

If the variance of y_t is indeed equal to $\exp(\boldsymbol{X}_t\boldsymbol{\beta})$, its expectation according to the loglinear Poisson regression model, then the quantity

$$z_t(\boldsymbol{\beta}) \equiv \big(y_t - \exp(\boldsymbol{X}_t\boldsymbol{\beta}) \big)^2 - y_t \tag{4.57}$$

has expectation 0.[4] We can test whether the expectation is really zero by running the OPG regression (4.56), adding an extra regressor with typical element $z_t(\hat{\boldsymbol{\beta}})$. Both n minus the sum of squared residuals from this augmented OPG regression and the t statistic associated with the extra regressor provide asymptotically valid test statistics; the former is asymptotically distributed as $\chi^2(1)$ under the null hypothesis, while the latter is asymptotically distributed as $\mathrm{N}(0, 1)$.

Testing can be made a little simpler if we note that the extra regressor (4.57) is uncorrelated with the regressors in (4.56) under the null. This is a simple consequence of the fact, which readers are asked to demonstrate in Exercise 4.24,

[4] The quantity $\big(y_t - \exp(\boldsymbol{X}_t\boldsymbol{\beta}) \big)^2 - \exp(\boldsymbol{X}_t\boldsymbol{\beta})$ also has expectation 0 and could be used in place of (4.57) in an OPG test regression. However, the simplifications that are discussed below would not be possible if the test regressor were redefined in this way.

that the third central moment of the Poisson distribution with parameter λ is equal to λ. We may write the testing OPG regression as

$$\boldsymbol{\iota} = \hat{\boldsymbol{G}}\boldsymbol{b} + c\hat{\boldsymbol{z}} + \text{residuals}, \tag{4.58}$$

where $\boldsymbol{\iota}$ is an n–vector of 1s, the matrix $\hat{\boldsymbol{G}} \equiv \boldsymbol{G}(\hat{\boldsymbol{\beta}})$ contains the regressors of (4.56) evaluated at $\hat{\boldsymbol{\beta}}$, and $\hat{\boldsymbol{z}} \equiv \boldsymbol{z}(\hat{\boldsymbol{\beta}})$ is the extra regressor, with typical element $z_t(\hat{\boldsymbol{\beta}})$. By the FWL Theorem, a test of the hypothesis that $c = 0$ can equally well be performed by running the FWL regression

$$\hat{\boldsymbol{M}}_{\boldsymbol{G}}\boldsymbol{\iota} = c\hat{\boldsymbol{M}}_{\boldsymbol{G}}\hat{\boldsymbol{z}} + \text{residuals}, \tag{4.59}$$

where $\hat{\boldsymbol{M}}_{\boldsymbol{G}}$ is the orthogonal projection matrix that projects on to the orthogonal complement of the span of the columns of $\hat{\boldsymbol{G}}$. But, since those columns are orthogonal to $\boldsymbol{\iota}$, the regressand of (4.59) is just $\boldsymbol{\iota}$. In addition, because $\boldsymbol{z}(\boldsymbol{\beta})$ is uncorrelated with the columns of $\boldsymbol{G}(\boldsymbol{\beta})$, the regressor is asymptotically equal to $\hat{\boldsymbol{z}}$. Therefore, regressions (4.58) and (4.59) are asymptotically equivalent to a regression of $\boldsymbol{\iota}$ on $\hat{\boldsymbol{z}}$. Once again, either the explained sum of squares or the t statistic for $c = 0$ yields an asymptotically valid test.

In Part 1, Exercise 5.F10, we saw that every t statistic is proportional to the cotangent of a certain angle, namely, the angle between the regressand and the regressor of the FWL regression that can be used to compute the statistic. Since this angle does not depend on which vector is the regressor and which vector is the regressand, this result implies that the t statistic from regressing $\boldsymbol{\iota}$ on $\hat{\boldsymbol{z}}$ is identical to the t statistic from regressing $\hat{\boldsymbol{z}}$ on $\boldsymbol{\iota}$. If we run the regression in this direction, however, we do not obtain the same ESS. Nevertheless, the ESS can be used as a valid statistic if the variables are scaled by estimates of the standard deviations of the elements of $\boldsymbol{z}(\boldsymbol{\beta})$. This rescaling yields the artificial regression that is most commonly used to test for overdispersion in the Poisson regression model.

Observe that, if Y is a random variable which follows the Poisson distribution with parameter λ, then

$$\begin{aligned}
\mathrm{E}\Big(\big((Y-\lambda)^2 - Y\big)^2\Big) &= \mathrm{E}\Big(\big((Y-\lambda)^2 - (Y-\lambda) - \lambda\big)^2\Big) \\
&= \mathrm{E}\big((Y-\lambda)^4\big) + \mathrm{E}\big((Y-\lambda)^2\big) + \lambda^2 \\
&\quad - 2\mathrm{E}\big((Y-\lambda)^3\big) - 2\lambda\mathrm{E}\big((Y-\lambda)^2\big) - 2\lambda\mathrm{E}(Y-\lambda) \\
&= \lambda + 3\lambda^2 + \lambda + \lambda^2 - 2\lambda - 2\lambda^2 = 2\lambda^2,
\end{aligned}$$

where we have used the result of Exercise 4.24 for both the third and fourth central moments of the Poisson distribution. A suitable testing regression with scaled variables can therefore be written as

$$\frac{1}{\sqrt{2}}\exp(-\boldsymbol{X}_t\hat{\boldsymbol{\beta}})z_t(\hat{\boldsymbol{\beta}}) = \frac{1}{\sqrt{2}}\exp(-\boldsymbol{X}_t\hat{\boldsymbol{\beta}})c + \text{residual}, \tag{4.60}$$

and both the t statistic and the explained sum of squares provide asymptotically valid test statistics.

The tests based on regression (4.60) were originally proposed by Cameron and Trivedi (1990). They also suggest tests based on regressions like (4.60), but with the regressor of (4.60) multiplied by various functions of the fitted values $\exp(\boldsymbol{X}_t\hat{\boldsymbol{\beta}})$. Common choices are the fitted values themselves or their squares. Cameron and Trivedi show that a test in which the regressor is multiplied by the function $g(\exp(\boldsymbol{X}_t\hat{\boldsymbol{\beta}}))$ of the fitted value has greatest power against DGPs for which the true variance of y_t is of the form $\exp(\boldsymbol{X}_t\boldsymbol{\beta}) + \alpha g(\exp(\boldsymbol{X}_t\boldsymbol{\beta}))$ for some scalar α. Tests with more than one degree of freedom can be performed by using several regressors constructed in this way. In all cases, an appropriate test statistic is the ESS. It is asymptotically distributed under the null as $\chi^2(r)$, where r is the number of regressors.

Other tests for overdispersion have been proposed by Cameron and Trivedi (1986), Lee (1986), and Mullahy (1997). Note that the finite-sample distributions of all these test statistics may differ substantially from their asymptotic ones. Better results may well be obtained by using bootstrap P values. A parametric bootstrap DGP is appropriate. It can easily be implemented by using a procedure for obtaining drawings from the Poisson distribution similar to the one we discussed for discrete choice models in the previous section.

Consequences of Overdispersion in the Poisson Regression Model

Finding evidence of overdispersion does not necessarily mean that we must abandon the Poisson regression model. Since the model is equivalent to weighted NLS, and weighted NLS is consistent even when the weights are incorrect, the ML estimator $\hat{\boldsymbol{\beta}}$ must be consistent whenever the exponential mean function $\lambda_t(\boldsymbol{\beta})$ is correctly specified. In this situation, $\hat{\boldsymbol{\beta}}$ is actually a quasi-ML estimator, or QMLE; see Section 3.4. However, as is generally the case for quasi-ML estimators, the covariance matrix estimator (4.53) is not valid if the entire model is not specified correctly.

To find the asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}$ when the model is not correctly specified, we may use the result (3.40), which is true for every quasi-ML estimator. If we replace the generic parameter vector $\boldsymbol{\theta}$ of that equation by $\boldsymbol{\beta}$, the limiting covariance matrix becomes

$$\mathcal{H}^{-1}(\boldsymbol{\beta}_0)\mathcal{I}(\boldsymbol{\beta}_0)\mathcal{H}^{-1}(\boldsymbol{\beta}_0). \tag{4.61}$$

For the Poisson regression model, we see from (4.52) that

$$\mathcal{H}(\boldsymbol{\beta}_0) = -\operatorname*{plim}_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}\exp(\boldsymbol{X}_t\boldsymbol{\beta}_0)\boldsymbol{X}_t^\top\boldsymbol{X}_t = -\operatorname*{plim}_{n\to\infty}\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{\Upsilon}(\boldsymbol{\beta}_0)\boldsymbol{X}. \tag{4.62}$$

From the definitions (3.31) and (3.32), and from the expression given in (4.51) for the gradient of the loglikelihood, it follows that the asymptotic information

matrix is

$$\mathcal{I}(\boldsymbol{\beta}_0) = \operatorname*{plim}_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} \omega_t^2(\boldsymbol{\beta}_0) \boldsymbol{X}_t^\top \boldsymbol{X}_t = \operatorname*{plim}_{n\to\infty} \frac{1}{n} \boldsymbol{X}^\top \boldsymbol{\Omega}(\boldsymbol{\beta}_0) \boldsymbol{X}, \tag{4.63}$$

where $\omega_t^2(\boldsymbol{\beta}_0) \equiv \mathrm{E}\big(y_t - \exp(\boldsymbol{X}_t\boldsymbol{\beta}_0)\big)^2$ is the conditional variance of y_t, and $\boldsymbol{\Omega}(\boldsymbol{\beta}_0)$ is the diagonal matrix with typical diagonal element $\omega_t^2(\boldsymbol{\beta}_0)$.

When the model is correctly specified, the conditional variance ω_t^2 is equal to the conditional mean $\exp(\boldsymbol{X}_t\boldsymbol{\beta}_0)$, and the asymptotic covariance matrix (4.61) simplifies to $\mathcal{I}^{-1}(\boldsymbol{\beta}_0) = -\mathcal{H}^{-1}(\boldsymbol{\beta}_0)$. When the model is not correctly specified, however, this simplification does not occur.

One quite plausible specification for the conditional variance of y_t is

$$\omega_t^2(\boldsymbol{\beta}) = \gamma^2 \exp(\boldsymbol{X}_t\boldsymbol{\beta}), \tag{4.64}$$

in which the conditional variance is proportional to the conditional expectation. Under this specification, the asymptotic covariance matrix (4.61) simplifies to γ^2 times $-\mathcal{H}^{-1}(\boldsymbol{\beta}_0)$. Since this is not a sandwich covariance matrix, it is clear that $\hat{\boldsymbol{\beta}}$ remains asymptotically efficient in this special case. An easy way to estimate this covariance matrix is simply to run the artificial regression (4.55), with $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$. Because s^2 provides a consistent estimator of γ^2, the OLS covariance matrix from this regression is asymptotically valid; see Exercise 4.26.

Even if we do not specify the conditional variance of y_t, we can obtain an asymptotically valid covariance matrix whenever the matrices (4.62) and (4.63) can be estimated consistently. To do this, we need to use a sandwich estimator similar to the HCCME. We can estimate (4.62) consistently if we replace $\boldsymbol{\beta}_0$ by $\hat{\boldsymbol{\beta}}$. In order to estimate (4.63) consistently, we replace the conditional variance $\omega_t^2(\boldsymbol{\beta}_0)$ by the squared residual $(y_t - \exp(\boldsymbol{X}_t\hat{\boldsymbol{\beta}}))^2$. Thus a valid estimator of $\mathrm{Var}(\hat{\boldsymbol{\beta}})$ when only the conditional mean part of the Poisson regression model is correctly specified is

$$\widehat{\mathrm{Var}}_{\mathrm{h}}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}^\top \hat{\boldsymbol{\Upsilon}} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \hat{\boldsymbol{\Omega}} \boldsymbol{X} (\boldsymbol{X}^\top \hat{\boldsymbol{\Upsilon}} \boldsymbol{X})^{-1}, \tag{4.65}$$

where $\hat{\boldsymbol{\Omega}}$ is the $n \times n$ diagonal matrix with diagonal element t given by $(y_t - \exp(\boldsymbol{X}_t\hat{\boldsymbol{\beta}}))^2$. As usual, the "h" subscript indicates that the matrix (4.65) is valid in the presence of heteroskedasticity of unknown form. Given the substantial risk of misspecification, it is strongly recommended to use the sandwich estimator (4.65) rather than (4.53) in practical applications. Notice that the sandwich estimator is very easy to calculate without any special software. If we run the artificial regression (4.55) and ask the regression package to compute an HCCME, it gives us either (4.65) or something that is asymptotically equal to (4.65); see Exercise 4.27.

Of course, except in the special case of (4.64), the ML estimator $\hat{\boldsymbol{\beta}}$ is not asymptotically efficient when the Poisson regression model is not correctly

specified. The fact that the covariance matrix has the sandwich form makes this clear. Moreover, $\hat{\boldsymbol{\beta}}$ is not even consistent if the conditional mean function $\exp(\boldsymbol{X}_t\boldsymbol{\beta})$ is not correctly specified. Many other models for count data have been suggested, and one or more of them may well fit better than the Poisson regression model does. Wooldridge (1999) and Cameron and Trivedi (2001) provide more advanced introductions to the topic of count data, and Cameron and Trivedi (1998) provides a detailed treatment of a large number of different models for data of this type.

4.6 Models for Censored and Truncated Data

Continuous dependent variables can sometimes take only a limited range of values. This may happen because they have been censored or truncated in some way. These two terms are easily confused. A sample is said to be **truncated** if some observations have been systematically excluded from the sample. For example, a sample of households with incomes under \$200,000 explicitly excludes households with incomes over that level. It is not a random sample of all households. If the dependent variable is income, or something correlated with income, results using the truncated sample could potentially be quite misleading.

On the other hand, a sample has been **censored** if no observations have been systematically excluded, but some of the information contained in them has been suppressed. Think of a "censor" who reads people's mail and blacks out certain parts of it. The recipients still get their mail, but parts of it are unreadable. To continue the previous example, suppose that households with all income levels are included in the sample, but for those with incomes in excess of \$200,000, the amount reported is always exactly \$200,000. This sort of censoring is often done in practice, presumably to protect the privacy of high-income respondents. In this case, the censored sample is still a random sample of all households, but the values reported for high-income households are not the true values.

Any dependent variable that has been either censored or truncated is said to be a **limited dependent variable**. Special methods are needed to deal with such variables because, if we simply use least squares, the consequences of truncation and censoring can be severe. Consider the regression model

$$y_t^\circ = \beta_1 + \beta_2 x_t + u_t, \quad u_t \sim \mathrm{NID}(0, \sigma^2), \tag{4.66}$$

where y_t° is a latent variable. We actually observe y_t, which differs from y_t° because it is either truncated or censored. For simplicity, suppose that censorship or truncation occurs whenever y_t° is less than 0. Clearly, the larger is the disturbance u_t, the larger is y_t°, and thus the greater must be the probability that $y_t^\circ \geq 0$. This probability must also depend on x_t. Thus, for

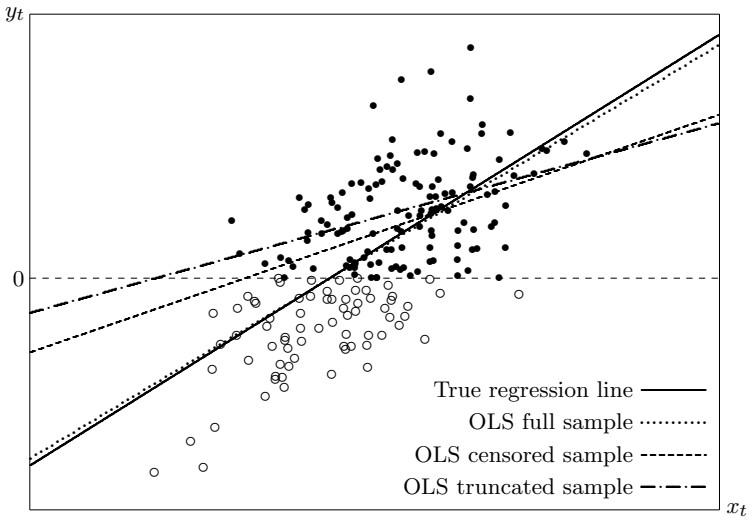


Figure 4.3 Effects of censoring and truncation

the sample we actually observe, u_t does not have conditional expectation 0 and is not uncorrelated with x_t. Since the disturbances no longer satisfy these key assumptions, it is not surprising that OLS estimation using truncated or censored samples yields estimators that are biased and inconsistent.

The consequences of censoring and truncation are illustrated in Figure 4.3. The figure shows 200 (x_t, y_t°) pairs generated from the model (4.66). The 71 observations with $y_t^\circ < 0$ are shown as circles, and the 129 observations with $y_t = y_t^\circ \geq 0$ are shown as black dots. The solid line is the true regression function, and the nearby dotted line is the regression function obtained by OLS estimation using all the observations. When the data are truncated, the observations with $y_t^\circ < 0$ are discarded. OLS estimation using this truncated sample yields the regression line shown in dots and dashes. When the data are censored, these 71 observations are retained, but y_t is set equal to 0 for all of them. OLS estimation using this censored sample yields the dashed regression line. Neither of these regression lines is at all close to the true one.

In this example, the consequences of either censoring or truncation are quite severe. Just how severe they are in any particular case depends on σ^2, the variance of the disturbances in the model (4.66), and on the extent of the censoring or truncation. If σ^2 is very small relative to the variation in the fitted values, so is the bias induced by limiting the dependent variable. This bias is also small if few observations are censored or truncated. Conversely, when σ^2 is large and many observations are censored or truncated, the bias can be extremely large.

Truncated Regression Models

It is quite simple to estimate a truncated regression model by maximum likelihood if the distribution of the disturbances in the latent variable model is assumed to be known. By far the most common assumption is that the disturbances are normally, independently, and identically distributed, as in (4.66). We restrict our attention to this special case.

If the regression function for the latent variable model is $\boldsymbol{X}_t\boldsymbol{\beta}$, the probability that y_t° is included in the sample is

$$\begin{aligned}\Pr(y_t^\circ \geq 0) &= \Pr(\boldsymbol{X}_t\boldsymbol{\beta} + u_t \geq 0)\\ &= 1 - \Pr(u_t < -\boldsymbol{X}_t\boldsymbol{\beta}) = 1 - \Pr(u_t/\sigma < -\boldsymbol{X}_t\boldsymbol{\beta}/\sigma)\\ &= 1 - \Phi(-\boldsymbol{X}_t\boldsymbol{\beta}/\sigma) = \Phi(\boldsymbol{X}_t\boldsymbol{\beta}/\sigma).\end{aligned}$$

When $y_t^\circ \geq 0$ and y_t is observed, the density of y_t is proportional to the density of y_t°. Otherwise, the density of y_t is 0. The factor of proportionality, which is needed to ensure that the density of y_t integrates to unity, is the inverse of the probability that $y_t^\circ \geq 0$. Therefore, the density of y_t can be written as

$$\frac{\sigma^{-1}\phi\big((y_t - \boldsymbol{X}_t\boldsymbol{\beta})/\sigma\big)}{\Phi(\boldsymbol{X}_t\boldsymbol{\beta}/\sigma)}.$$

This implies that the loglikelihood function, which is the sum over all t of the log of the density of y_t conditional on $y_t^\circ \geq 0$, is

$$\begin{aligned}\ell(\boldsymbol{y}, \boldsymbol{\beta}, \sigma) = &-\frac{n}{2}\log(2\pi) - n\log(\sigma) - \frac{1}{2\sigma^2}\sum_{t=1}^{n}(y_t - \boldsymbol{X}_t\boldsymbol{\beta})^2\\ &- \sum_{t=1}^{n}\log\Phi(\boldsymbol{X}_t\boldsymbol{\beta}/\sigma).\end{aligned} \tag{4.67}$$

Maximization of expression (4.67) is generally not difficult. Even though the loglikelihood function is not globally concave, there is a unique MLE; see Orme and Ruud (2002).

The first three terms in expression (4.67) comprise the loglikelihood function that corresponds to OLS regression; see equation (3.10). The last term is minus the summation over all t of the logarithms of the probabilities that an observation with regression function $\boldsymbol{X}_t\boldsymbol{\beta}$ belongs to the sample. Since these probabilities must be less than 1, this term must always be positive. It can be made larger by making the probabilities smaller. Thus the maximization algorithm chooses the parameters in such a way that these probabilities are smaller than they would be for the OLS estimates. The presence of this fourth term therefore causes the ML estimates of $\boldsymbol{\beta}$ and σ to differ, often substantially, from their least-squares counterparts, and it ensures that the ML estimates are consistent.

It is not difficult to modify this model to allow for other forms of truncation. The sample can be truncated from above, from below, or from both above and below. The truncation points must be known, but they can be fixed or they can vary across observations. See Exercises 4.29 and 4.30.

Censored Regression Models

The most popular model for censored data is the **tobit model**, which was first suggested in Tobin (1958), which is quite a famous paper. The simplest version of the tobit model is

$$y_t^\circ = \boldsymbol{X}_t\boldsymbol{\beta} + u_t, \quad u_t \sim \text{NID}(0, \sigma^2),$$
$$y_t = y_t^\circ \text{ if } y_t^\circ > 0; \quad y_t = 0 \text{ otherwise.}$$

Here y_t° is a latent variable that is observed whenever it is positive. However, when the latent variable is negative, the observation is censored, and we simply observe $y_t = 0$. The tobit model can readily be modified to allow for censoring from above instead of from below or for censoring from both above and below. It can also be modified to allow the point at which the censoring occurs to vary across observations in a deterministic way; see Exercise 4.31.

The loglikelihood function for the tobit model is a little unusual, but it is not difficult to derive. First, it is easy to see that

$$\begin{aligned}\Pr(y_t = 0) = \Pr(y_t^\circ \le 0) &= \Pr(\boldsymbol{X}_t\boldsymbol{\beta} + u_t \le 0)\\ &= \Pr\Big(\frac{u_t}{\sigma} \le \frac{-\boldsymbol{X}_t\boldsymbol{\beta}}{\sigma}\Big) = \Phi(-\boldsymbol{X}_t\boldsymbol{\beta}/\sigma).\end{aligned}$$

Therefore, since there is a positive probability that $y_t = 0$, the contribution to the loglikelihood function made by observations with $y_t = 0$ is not the log of the density, but the log of that positive probability, namely,

$$\ell_t(y_t, \boldsymbol{\beta}, \sigma) = \log \Phi(-\boldsymbol{X}_t\boldsymbol{\beta}/\sigma). \tag{4.68}$$

If y_t is positive, the density of y_t exists, and the contribution to the loglikelihood is its logarithm,

$$\log\Big(\tfrac{1}{\sigma}\phi\big((y_t - \boldsymbol{X}_t\boldsymbol{\beta})/\sigma\big)\Big), \tag{4.69}$$

which is the contribution to the loglikelihood function for an observation in a classical normal linear regression model without any censoring.

Combining expression (4.68), the contribution for the censored observations, with expression (4.69), the contribution for the uncensored ones, we find that the loglikelihood function for the tobit model is

$$\sum_{y_t=0} \log \Phi(-\boldsymbol{X}_t\boldsymbol{\beta}/\sigma) + \sum_{y_t>0} \log\Big(\tfrac{1}{\sigma}\phi\big((y_t - \boldsymbol{X}_t\boldsymbol{\beta})/\sigma\big)\Big). \tag{4.70}$$

This loglikelihood function is rather curious. The first term is the sum of the logs of probabilities, for the censored observations, while the second is the sum of the logs of densities, for the uncensored observations. This reflects the fact that the dependent variable in a tobit model has a distribution that is a mixture of discrete and continuous random variables. This fact does not, however, prevent the ML estimator for the tobit model from having the usual properties of consistency and asymptotic normality, as was shown explicitly by Amemiya (1973c).

It is generally somewhat easier to maximize the loglikelihood function (4.70) if the tobit model is reparametrized. The new parameters are $\boldsymbol{\gamma} \equiv \boldsymbol{\beta}/\sigma$ and $h \equiv 1/\sigma$. Since the loglikelihood function can be shown to be globally concave in the latter parametrization (Olsen, 1978), there must be a unique maximum no matter which parametrization is used. Even without any reparametrization, it is generally not at all difficult to maximize (4.70) by using a quasi-Newton algorithm.

The $(k+1) \times (k+1)$ covariance matrix of the ML estimates may, as usual, be estimated in several ways. Analytic expressions for the information matrix exist (Amemiya, 1973c), and at least two artificial regressions are available. One of these is the OPG regression that we discussed in Section 2.5, and the other is a double-length regression proposed by Orme (1995). The latter is substantially more complicated than the former, but it seems to work very much better. Since the tobit model is fully specified, it is straightforward to employ the parametric bootstrap. Simulation results in Davidson and MacKinnon (1999a) suggest that inferences based on it can be much more reliable than ones based only on asymptotic theory.

Testing the Tobit Model

There is an interesting relationship among the tobit, truncated regression, and probit models. If we both add and subtract the term $\sum_{y_t>0} \log\big(\Phi(\boldsymbol{X}_t\boldsymbol{\beta}/\sigma)\big)$ from the tobit loglikelihood function (4.70), it becomes

$$\begin{aligned}&\sum_{y_t>0} \log\Big(\tfrac{1}{\sigma}\phi\big((y_t - \boldsymbol{X}_t\boldsymbol{\beta})/\sigma\big)\Big) - \sum_{y_t>0} \log \Phi(\boldsymbol{X}_t\boldsymbol{\beta}/\sigma)\\ &+ \sum_{y_t=0} \log \Phi(-\boldsymbol{X}_t\boldsymbol{\beta}/\sigma) + \sum_{y_t>0} \log \Phi(\boldsymbol{X}_t\boldsymbol{\beta}/\sigma).\end{aligned} \tag{4.71}$$

The first line of (4.71) is the loglikelihood function for a truncated regression model estimated over all the observations for which $y_t > 0$; compare (4.67). The second line is the loglikelihood function for a probit model with index function $\boldsymbol{X}_t\boldsymbol{\beta}/\sigma$; compare (4.09). Of course, if all we had was the second line here, we could not identify $\boldsymbol{\beta}$ and σ separately, but since we also have the first line, that is not a problem.

Writing the tobit loglikelihood function in the form of (4.71) makes it clear that this model is really a probit model combined with a truncated regression

model, with the coefficient vectors in the two models restricted to be proportional to each other. This restriction can easily be tested by means of an LR test with k degrees of freedom. If this test leads to a rejection of the null hypothesis, then we probably should not be using a tobit model.

Of course, like all econometric models, the tobit model can and should be tested for a variety of types of possible misspecification. A large number of tests can be based on the OPG regression and on the double-length regression of Orme (1995). Tests based on the OPG regression are discussed by Pagan and Vella (1989) and Smith (1989). See also Chesher and Irish (1987).

4.7 Sample Selectivity

In the previous section, we considered samples truncated on the basis of the value of the dependent variable. Many samples are truncated on the basis of another variable that is correlated with the dependent variable. For example, people may choose to enter the labor force if their market wage exceeds their reservation wage and choose to stay out of it otherwise. Then a sample of people who are in the labor force must exclude those whose reservation wage exceeds their market wage. If the dependent variable, whatever it may be, is correlated with the difference between reservation and market wages, least squares yields inconsistent estimates. In this case, the sample is said to have been **selected** on the basis of this difference. The consequences of this type of sample selection are often said to be due to **sample selectivity**.

Let us consider a simple model that involves sample selectivity. Suppose that y_t° and z_t° are two latent variables, generated by the bivariate process

$$\begin{bmatrix} y_t^\circ \\ z_t^\circ \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}_t\boldsymbol{\beta} \\ \boldsymbol{W}_t\boldsymbol{\gamma} \end{bmatrix} + \begin{bmatrix} u_t \\ v_t \end{bmatrix}, \quad \begin{bmatrix} u_t \\ v_t \end{bmatrix} \sim \text{NID}\left(\mathbf{0}, \begin{bmatrix} \sigma^2 & \rho\sigma \\ \rho\sigma & 1 \end{bmatrix}\right), \tag{4.72}$$

where $\boldsymbol{X}_t$ and $\boldsymbol{W}_t$ are vectors of observations on exogenous or predetermined variables, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are unknown parameter vectors, σ is the standard deviation of u_t, and ρ is the correlation between u_t and v_t. The restriction that the variance of v_t is equal to 1 is imposed because only the sign of z_t° is observed. In fact, the variables that are actually observed are y_t and z_t, and they are related to y_t° and z_t° as follows:

$$\begin{aligned} &y_t = y_t^\circ \text{ if } z_t^\circ > 0; \ y_t \text{ unobserved otherwise;} \\ &z_t = 1 \text{ if } z_t^\circ > 0; \ z_t = 0 \text{ otherwise.} \end{aligned} \tag{4.73}$$

Thus there are two types of observations, those for which we observe $y_t = y_t^\circ$ and $z_t = 1$, along with both $\boldsymbol{X}_t$ and $\boldsymbol{W}_t$, and those for which we observe only $z_t = 0$ and $\boldsymbol{W}_t$.

Each observation contributes a factor to the likelihood function for this model that can be written as

$$\mathrm{I}(z_t = 0)\Pr(z_t = 0) + \mathrm{I}(z_t = 1)\Pr(z_t = 1)f(y_t^\circ \mid z_t = 1),$$

where $f(y_t^\circ \mid z_t = 1)$ denotes the density of y_t° conditional on $z_t = 1$. This is the appropriate way to specify the likelihood because, if we integrate with respect to y_t° and sum over the two possible values of z_t, the result is 1. Note also that the value of y_t° is needed only if it is observed, that is, if $z_t = 1$. The loglikelihood function is

$$\sum_{z_t=0} \log \Pr(z_t = 0) + \sum_{z_t=1} \log\big(\Pr(z_t = 1)f(y_t^\circ \mid z_t = 1)\big). \tag{4.74}$$

The first term of (4.74), which comes from the observations with $z_t = 0$, is exactly the same as the corresponding term in a probit model. The second term comes from the observations with $z_t = 1$. By using the fact that we can factor a joint density any way we please, it can also be written as

$$\sum_{z_t=1} \log\big(\Pr(z_t = 1 \mid y_t^\circ)f(y_t^\circ)\big),$$

where $f(y_t^\circ)$ is the density of y_t° conditional on predetermined or exogenous variables, which is just a normal density with expectation $\boldsymbol{X}_t\boldsymbol{\beta}$ and variance σ^2.

In order to write out the loglikelihood function (4.74) explicitly, we must calculate $\Pr(z_t = 1 \mid y_t^\circ)$. Since u_t and v_t are bivariate normal, we can write $v_t = \rho u_t/\sigma + \varepsilon_t$, where ε_t is a normally distributed random variable with expectation 0 and variance $1 - \rho^2$. Thus

$$z_t^\circ = \boldsymbol{W}_t\boldsymbol{\gamma} + \rho(y_t^\circ - \boldsymbol{X}_t\boldsymbol{\beta})/\sigma + \varepsilon_t, \quad \varepsilon_t \sim \text{NID}(0, 1 - \rho^2).$$

Because $y_t = y_t^\circ$ when $z_t = 1$, it follows that

$$\Pr(z_t = 1 \mid y_t^\circ) = \Phi\left(\frac{\boldsymbol{W}_t\boldsymbol{\gamma} + \rho(y_t - \boldsymbol{X}_t\boldsymbol{\beta})/\sigma}{(1 - \rho^2)^{1/2}}\right).$$

Thus the loglikelihood function (4.74) becomes

$$\begin{aligned} &\sum_{z_t=0} \log \Phi(-\boldsymbol{W}_t\boldsymbol{\gamma}) + \sum_{z_t=1} \log\Big(\tfrac{1}{\sigma}\phi\big((y_t - \boldsymbol{X}_t\boldsymbol{\beta})/\sigma\big)\Big) \\ &+ \sum_{z_t=1} \log \Phi\left(\frac{\boldsymbol{W}_t\boldsymbol{\gamma} + \rho(y_t - \boldsymbol{X}_t\boldsymbol{\beta})/\sigma}{(1 - \rho^2)^{1/2}}\right). \end{aligned} \tag{4.75}$$

The first term looks like the corresponding term for a standard probit model in which z_t is explained by $\boldsymbol{W}_t$, the second term looks like the loglikelihood function for a linear regression of y_t on $\boldsymbol{X}_t$, with normal disturbances, and the third

term is one that we have not seen before. If $\rho = 0$, this term would collapse to the term corresponding to observations with $z_t = 1$ in the probit model for z_t, and we could estimate the probit model and the regression model separately. In general, however, this term forces us to estimate both equations together by making the probability that $z_t = 1$ depend on $y_t - \boldsymbol{X}_t\boldsymbol{\beta}$.

Heckman's Two-Step Method

From the point of view of asymptotic efficiency, the best way to estimate the model characterized by (4.72) and (4.73) is simply to maximize the loglikelihood function (4.75). With modern computing equipment and appropriate software, this is not unreasonably difficult to do, although numerical problems can be encountered when ρ approaches ± 1. Instead of ML estimation, however, it is popular to use a computationally simpler technique, which is known as **Heckman's two-step method**; see Heckman (1976, 1979). Although we do not recommend that practitioners rely solely on this method, it can be useful for preliminary work, and it yields insights into the nature of sample selectivity. In addition, it provides a good starting point for the nonlinear algorithm used to obtain the MLE.

Heckman's two-step method is based on the fact that the first equation of (4.72), for observations where y_t is observed, can be rewritten as

$$y_t = \boldsymbol{X}_t\boldsymbol{\beta} + \rho\sigma v_t + e_t. \tag{4.76}$$

Here the disturbance u_t is divided into two parts, one perfectly correlated with v_t, the disturbance in the equation for the latent variable z_t°, and one independent of v_t. The idea is to replace the unobserved disturbance v_t in (4.76) by its expectation conditional on $z_t = 1$ and on the explanatory variables $\boldsymbol{W}_t$. This conditional expectation is

$$\mathrm{E}(v_t \,|\, z_t = 1, \boldsymbol{W}_t) = \mathrm{E}(v_t \,|\, v_t > -\boldsymbol{W}_t\boldsymbol{\gamma}, \boldsymbol{W}_t) = \frac{\phi(\boldsymbol{W}_t\boldsymbol{\gamma})}{\Phi(\boldsymbol{W}_t\boldsymbol{\gamma})}, \tag{4.77}$$

where readers are asked to prove the last equality in Exercise 4.32. The quantity $\phi(x)/\Phi(x)$ is known as the **inverse Mills ratio**; see Johnson, Kotz, and Balakrishnan (1994). In the first step of Heckman's two-step method, an ordinary probit model is used to obtain consistent estimates $\hat{\boldsymbol{\gamma}}$ of the parameters of the selection equation. In the second step, the unobserved v_t in regression (4.76) is replaced by the **selectivity regressor** $\phi(\boldsymbol{W}_t\hat{\boldsymbol{\gamma}})/\Phi(\boldsymbol{W}_t\hat{\boldsymbol{\gamma}})$, and regression (4.76) becomes

$$y_t = \boldsymbol{X}_t\boldsymbol{\beta} + \rho\sigma \frac{\phi(\boldsymbol{W}_t\hat{\boldsymbol{\gamma}})}{\Phi(\boldsymbol{W}_t\hat{\boldsymbol{\gamma}})} + \text{residual}. \tag{4.78}$$

This **Heckman regression**, as it is often called, is easy to estimate by OLS and yields consistent estimates of $\boldsymbol{\beta}$.

Regression (4.78) provides a test for sample selectivity as well as an estimation technique. The coefficient of the selectivity regressor is $\rho\sigma$. Since $\sigma \neq 0$, the

ordinary t statistic for this coefficient to be zero can be used to test the hypothesis that $\rho = 0$, and it is asymptotically distributed as $\mathrm{N}(0, 1)$ under the null hypothesis. If this coefficient is not significantly different from zero, the investigator may reasonably decide that selectivity is not a problem and proceed to use least squares as usual.

Although the Heckman regression (4.78) yields consistent estimates of $\boldsymbol{\beta}$, the OLS covariance matrix is valid only when $\rho = 0$. The problem is that the selectivity regressor is being treated like any other regressor, when it is in fact part of the disturbance. It is possible to obtain a valid covariance matrix estimate to go along with the two-step estimates of $\boldsymbol{\beta}$ from (4.78), but the calculation is quite cumbersome, and the estimated covariance matrix is not always positive definite. See Greene (1981) and Lee (1982) for details.

It should be stressed that the consistency of this two-step estimator, like that of the ML estimator, depends critically on the assumption of bivariate normality. This can be seen from the specification of the selectivity regressor as the inverse Mills ratio (4.77). When the elements of $\boldsymbol{W}_t$ are the same as the elements of $\boldsymbol{X}_t$, as is often the case in practice, it is only the nonlinearity of the inverse Mills ratio as a function of $\boldsymbol{W}_t\boldsymbol{\gamma}$ that makes the parameters of the second-step regression identifiable. The form of the nonlinear relationship would be different if the disturbances did not follow the normal distribution.

4.8 Duration Models

Economists are sometimes interested in how much time elapses before some event occurs. For example, they may be interested in the length of labor disputes (that is, strike duration), the age of first marriage for men and women (that is, the duration of the state of being single), the duration of unemployment spells, the duration between trades on a stock exchange, or the length of time people wait before trading in a car. In this section, we will discuss some simple econometric models for duration data of this type.

In many cases, each observation in the sample consists of a measured duration, denoted t_i, and a $1 \times k$ vector of exogenous variables, denoted $\boldsymbol{X}_i$. In adopting this formulation, we have implicitly ruled out the possibility, which more complicated models can allow for, that the exogenous variables may change as time passes. To avoid notational confusion, we use i to index observations. In theory, duration is a nonnegative, continuous random variable. In practice, however, t_i is often reported as an integer number of weeks or months. When it is always a small integer, a count data model like the ones discussed in Section 4.5 may be appropriate. However, when t_i can take on a large number of integer values, it is conventional to model duration as being continuous. Almost all of the literature deals with the continuous case.

Survivor Functions and Hazard Functions

In practice, interest often centers not so much on how t_i is related to $\boldsymbol{X}_i$ but rather on how the probability that a state will endure varies over the duration of the state. For example, we may be interested in seeing how the probability that someone finds a job changes as the length of time they have been unemployed increases. Before we can answer this sort of question, we need to discuss a few fundamental concepts.

Suppose that how long a state endures is measured by T, a nonnegative, continuous random variable with PDF $f(t)$ and CDF $F(t)$, where t is a realization of T. Then the **survivor function** is defined as

$$S(t) \equiv 1 - F(t).$$

This is the probability that a state which started at time $t = 0$ is still going on at time t. The probability that it ends in any short period of time, say the period from time t to time $t + \Delta t$, is

$$\Pr(t < T \leq t + \Delta t) = F(t + \Delta t) - F(t). \tag{4.79}$$

This probability is unconditional. For many purposes, we may be interested in the probability that a state ends between time t and time $t + \Delta t$, conditional on having reached time t in the first place. This probability is

$$\Pr(t < T \leq t + \Delta t \,|\, T \geq t) = \frac{F(t + \Delta t) - F(t)}{S(t)}. \tag{4.80}$$

Since we are dealing with continuous time, it is natural to divide (4.79) and (4.80) by Δt and consider what happens as $\Delta t \to 0$. The limit of $1/\Delta t$ times (4.79) as $\Delta t \to 0$ is simply the PDF $f(t)$, and the limit of $1/\Delta t$ times the right-hand side of equation (4.80) is

$$h(t) \equiv \frac{f(t)}{S(t)} = \frac{f(t)}{1 - F(t)}. \tag{4.81}$$

The function $h(t)$ defined in (4.81) is called the **hazard function**. For many purposes, it is more interesting to model the hazard function than to model the survivor function directly.

Functional Forms

For a parametric model of duration, we need to specify a functional form for one of the functions $F(t)$, $S(t)$, $f(t)$, or $h(t)$, which then implies functional forms for the others. One of the simplest possible choices is the exponential distribution, which was discussed in Section 3.2. For this distribution,

$$f(t, \theta) = \theta e^{-\theta t}, \quad \text{and} \quad F(t, \theta) = 1 - e^{-\theta t}, \quad \theta > 0.$$

Therefore, the hazard function is

$$h(t) = \frac{f(t)}{S(t)} = \frac{\theta e^{-\theta t}}{e^{-\theta t}} = \theta.$$

Thus, if duration follows an exponential distribution, the hazard function is simply a constant.

Since the restriction that the hazard function is a constant is a very strong one, the exponential distribution is rarely used in applied work. A much more flexible functional form is provided by the **Weibull distribution**, which has two parameters, θ and α. For this distribution,

$$F(t, \theta, \alpha) = 1 - \exp\bigl(-(\theta t)^\alpha\bigr). \tag{4.82}$$

As readers are asked to show in Exercise 4.33, the survivor, density, and hazard functions for the Weibull distribution are as follows:

$$\begin{aligned} S(t) &= \exp\bigl(-(\theta t)^\alpha\bigr); \\ f(t) &= \alpha \theta^\alpha t^{\alpha-1} \exp\bigl(-(\theta t)^\alpha\bigr); \\ h(t) &= \alpha \theta^\alpha t^{\alpha-1}. \end{aligned} \tag{4.83}$$

When $\alpha = 1$, it is easy to see that the Weibull distribution collapses to the exponential, and the hazard is just a constant. For $\alpha < 1$, the hazard is decreasing over time, and for $\alpha > 1$, the hazard is increasing. Hazard functions of the former type are said to exhibit **negative duration dependence**, while those of the latter type are said to exhibit **positive duration dependence**. In the same way, a constant hazard is said to be **duration independent**.

Although the Weibull distribution is not nearly as restrictive as the exponential, it does not allow for the possibility that the hazard may first increase and then decrease over time, which is something that is frequently observed in practice. Various other distributions do allow for this type of behavior. A particularly simple one is the **lognormal distribution**, which was discussed in Section 3.8. Suppose that $\log t$ is distributed as $\mathrm{N}(\mu, \sigma^2)$. Then we have

$$\begin{aligned} F(t) &= \Phi\Bigl(\tfrac{1}{\sigma}(\log t - \mu)\Bigr), \\ S(t) &= 1 - \Phi\Bigl(\tfrac{1}{\sigma}(\log t - \mu)\Bigr) = \Phi\Bigl(-\tfrac{1}{\sigma}(\log t - \mu)\Bigr), \\ f(t) &= \frac{1}{\sigma t}\phi\Bigl(\tfrac{1}{\sigma}(\log t - \mu)\Bigr), \text{ and} \\ h(t) &= \frac{1}{\sigma t}\frac{\phi\bigl((\log t - \mu)/\sigma\bigr)}{\Phi\bigl(-(\log t - \mu)/\sigma\bigr)}. \end{aligned}$$

For this distribution, the hazard rises quite rapidly and then falls rather slowly. This behavior can be observed in Figure 4.4, which shows several hazard functions based on the exponential, Weibull, and lognormal distributions.

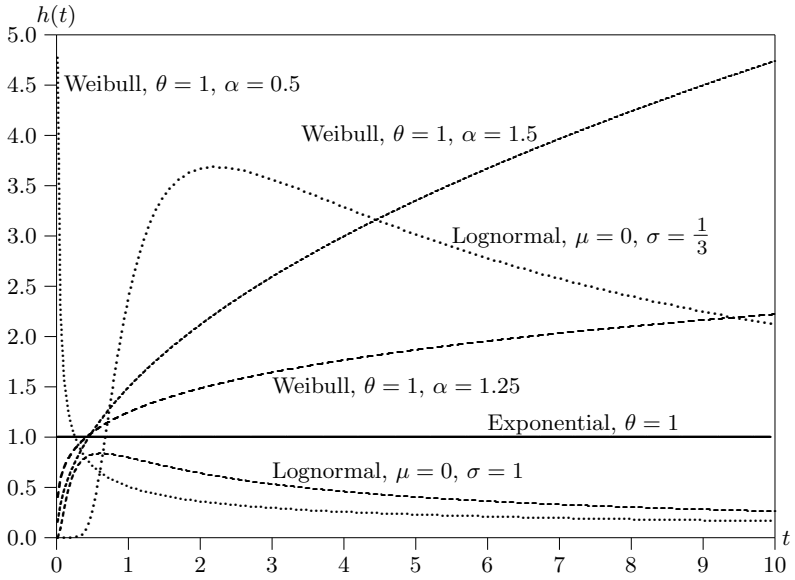


Figure 4.3 Various hazard functions

Maximum Likelihood Estimation

It is reasonably straightforward to estimate many duration models by maximum likelihood. In the simplest case, the data consist of n independent observations t_i on observed durations, each with an associated regressor vector $\boldsymbol{X}_i$. The loglikelihood function for $\boldsymbol{t}$, the vector of observations with typical element t_i, is just

$$\ell(\boldsymbol{t}, \boldsymbol{\theta}) = \sum_{i=1}^{n} \log f(t_i \,|\, \boldsymbol{X}_i, \boldsymbol{\theta}), \tag{4.84}$$

where $f(t_i \,|\, \boldsymbol{X}_i, \boldsymbol{\theta})$ denotes the density of t_i conditional on the data vector $\boldsymbol{X}_i$ for the parameter vector $\boldsymbol{\theta}$. In many cases, it may be easier to write the loglikelihood function as

$$\ell(\boldsymbol{t}, \boldsymbol{\theta}) = \sum_{i=1}^{n} \log h(t_i \,|\, \boldsymbol{X}_i, \boldsymbol{\theta}) + \sum_{i=1}^{n} \log S(t_i \,|\, \boldsymbol{X}_i, \boldsymbol{\theta}), \tag{4.85}$$

where $h(t_i \,|\, \boldsymbol{X}_i, \boldsymbol{\theta})$ is the hazard function and $S(t_i \,|\, \boldsymbol{X}_i, \boldsymbol{\theta})$ is the survivor function. The equivalence of (4.84) and (4.85) is ensured by (4.81), in which the hazard function was defined.

As with other models we have looked at in this chapter, it is convenient to let the loglikelihood depend on explanatory variables through an index function. As an example, suppose that duration follows a Weibull distribution, with

a parameter θ_i for observation i that has the form of the exponential mean function (4.48), so that $\theta_i = \exp(\boldsymbol{X}_i\boldsymbol{\beta}) > 0$. From (4.83) we see that the hazard and survivor functions for observation i are

$$\alpha \exp(\alpha \boldsymbol{X}_i\boldsymbol{\beta}) t^{\alpha-1} \quad \text{and} \quad \exp\bigl(-t^{\alpha} \exp(\alpha \boldsymbol{X}_i\boldsymbol{\beta})\bigr),$$

respectively. In practice, it is simpler to absorb the factor of α into the parameter vector $\boldsymbol{\beta}$, so as to yield an exponent of just $\boldsymbol{X}_i\boldsymbol{\beta}$ in these expressions. Then the loglikelihood function (4.85) becomes

$$\ell(\boldsymbol{t}, \boldsymbol{\beta}, \alpha) = n \log \alpha + \sum_{i=1}^{n} \boldsymbol{X}_i\boldsymbol{\beta} + (\alpha - 1) \sum_{i=1}^{n} \log t_i - \sum_{i=1}^{n} t_i^{\alpha} \exp(\boldsymbol{X}_i\boldsymbol{\beta}),$$

and ML estimates of the parameters α and $\boldsymbol{\beta}$ are obtained by maximizing this function in the usual way.

In practice, many data sets contain observations for which t_i is not actually observed. For example, if we have a sample of people who entered unemployment at various points in time, it is extremely likely that some people in the sample were still unemployed when data collection ended. If we omit such observations, we are effectively using a truncated data set, and we therefore obtain inconsistent estimates. However, if we include them but treat the observed t_i as if they were the lengths of completed spells of unemployment, we also obtain inconsistent estimates. In both cases, the inconsistency occurs for essentially the same reasons as it does when we apply OLS to a sample that has been truncated or censored; see Section 4.6.

If we are using ML estimation, it is easy enough to deal with duration data that have been censored in this way, provided we know that censorship has occurred. For ordinary, uncensored observations, the contribution to the loglikelihood function is a contribution like those in (4.84) or (4.85). For censored observations, where the observed t_i is the duration of an **incomplete spell**, it is the logarithm of the probability of censoring, which is the probability that the duration exceeds t_i, that is, the log of the survivor function. Therefore, if U denotes the set of uncensored observations, the loglikelihood function for the entire sample can be written as

$$\ell(\boldsymbol{t}, \boldsymbol{\theta}) = \sum_{i \in U} \log h(t_i \,|\, \boldsymbol{X}_i, \boldsymbol{\theta}) + \sum_{i=1}^{n} \log S(t_i \,|\, \boldsymbol{X}_i, \boldsymbol{\theta}). \tag{4.86}$$

Notice that uncensored observations contribute to both terms in equation (4.86), while censored observations contribute only to the second term. When there is no censoring, the same observations contribute to both terms, and the loglikelihood function (4.86) reduces to (4.85).

Proportional Hazard Models

One class of models that is quite widely used is the class of **proportional hazard models**, originally proposed by Cox (1972), in which the hazard function for the i^{th} economic agent is given by

$$h(\boldsymbol{X}_i, t) = g_1(\boldsymbol{X}_i)\, g_2(t), \tag{4.87}$$

for various specifications of the functions $g_1(\boldsymbol{X}_i)$ and $g_2(t)$. The latter is called the **baseline hazard function**. An implication of (4.87) is that the ratio of the hazards for any two agents, say the ones indexed by i and j, depends on the regressors but does not depend on t. This ratio is

$$\frac{h(\boldsymbol{X}_i, t)}{h(\boldsymbol{X}_j, t)} = \frac{g_1(\boldsymbol{X}_i)\, g_2(t)}{g_1(\boldsymbol{X}_j)\, g_2(t)} = \frac{g_1(\boldsymbol{X}_i)}{g_1(\boldsymbol{X}_j)}.$$

Thus the ratio of the conditional probability that agent i exits the state to the probability that agent j does so is constrained to be the same for all t. This makes proportional hazard models econometrically convenient, but they do impose fairly strong restrictions on behavior.

Both the exponential and Weibull distributions lead to proportional hazard models. As we have already seen, a natural specification of $g_1(\boldsymbol{X}_i)$ for these models is $\exp(\boldsymbol{X}_i\boldsymbol{\beta})$. For the exponential distribution, the baseline hazard function is just 1, and for the Weibull distribution it is $\alpha t^{\alpha-1}$.

One attractive feature of proportional hazards models is that it is possible to obtain consistent estimates of the parameters of the function $g_1(\boldsymbol{X}_i)$, without estimating those of $g_2(t)$ at all, by using a method called **partial likelihood** which we will not attempt to describe; see Cox and Oakes (1984) or Lancaster (1990). The baseline hazard function $g_2(t)$ can then be estimated in various ways, some of which do not require us to specify its functional form.

Complications

The class of duration models that we have discussed is quite limited. It does not allow the exogenous variables to change over time, and it does not allow for any **individual heterogeneity**, that is, variation in the hazard function across agents. The latter has serious implications for econometric inference. Suppose, for simplicity, that there are two types of agent, each with a constant hazard, which is twice as high for agents of type H as for those of type L. If we estimate a duration model for all agents together, we must observe negative duration dependence, because the type H agents exit the state more rapidly than the type L agents, and the ratio of type H to type L agents declines as duration increases.

There has been a great deal of work on duration models during the past two decades, and there are now numerous models that allow for time-varying

explanatory variables and/or individual heterogeneity. Classic references are Heckman and Singer (1984), Kiefer (1988), and Lancaster (1990). More recent work is discussed in Neumann (1999), Gouriéroux and Jasiak (2001), and van den Berg (2001).

4.9 Final Remarks

This chapter has dealt with a large number of types of dependent variable for which ordinary regression models are not appropriate: binary dependent variables (Section 4.2 and Section 4.3); discrete dependent variables that can take on more than two values, which may or may not be ordered (Section 4.4); count data (Section 4.5); limited dependent variables, which may be either censored or truncated (Section 4.6); dependent variables where the observations included in the sample have been determined endogenously (Section 4.7); and duration data (Section 4.8). In most cases, we have made strong distributional assumptions and relied on maximum likelihood estimation. This is generally the easiest way to proceed, but it can lead to seriously misleading results if the assumptions are false. It is therefore important to test the specification of these models carefully.

4.10 Exercises

⋆**4.1** Consider the contribution made by observation t to the loglikelihood function (4.09) for a binary response model. Show that this contribution is globally concave with respect to $\boldsymbol{\beta}$ if the function F is such that $F(-x) = 1-F(x)$, and if it, its derivative f, and its second derivative f' satisfy the condition

$$f'(x)F(x) - f^2(x) < 0 \tag{4.88}$$

for all real finite x.

Show that condition (4.88) is satisfied by both the logistic function $\Lambda(\cdot)$, defined in (4.07), and the standard normal CDF $\Phi(\cdot)$.

4.2 Prove that, for the logit model, the likelihood equations (4.10) reduce to

$$\sum_{t=1}^{n} x_{ti}(y_t - \Lambda(\boldsymbol{X}_t\boldsymbol{\beta})) = 0, \quad i = 1, \dots, k.$$

4.3 Show that the efficient GMM estimating equations (2.82), when applied to the binary response model specified by (4.01), are equivalent to the likelihood equations (4.10).

4.4 If $F_1(\cdot)$ and $F_2(\cdot)$ are two CDFs defined on the real line, show that any convex combination $(1-\alpha)F_1(\cdot) + \alpha F_2(\cdot)$ of them is also a properly defined CDF. Use this fact to construct a model that nests the logit model for which $\Pr(y_t = 1) = \Lambda(\boldsymbol{X}_t\boldsymbol{\beta})$ and the probit model for which $\Pr(y_t = 1) = \Phi(\boldsymbol{X}_t\boldsymbol{\beta})$ with just one additional parameter.

⋆**4.5** Consider the latent variable model

$$y_t^\circ = \beta_1 + \beta_2 x_t + u_t, \quad u_t \sim \mathrm{N}(0,1),$$
$$y_t = 1 \text{ if } y_t^\circ > 0, \;\; y_t = 0 \text{ if } y_t^\circ \leq 0.$$

Suppose that $x_t \sim \mathrm{N}(0,1)$. Generate 500 samples of 20 observations on (x_t, y_t) pairs, 100 assuming that $\beta_1 = 0$ and $\beta_2 = 1$, 100 assuming that $\beta_1 = 1$ and $\beta_2 = 1$, 100 assuming that $\beta_1 = -1$ and $\beta_2 = 1$, 100 assuming that $\beta_1 = 0$ and $\beta_1 = 2$, and 100 assuming that $\beta_1 = 0$ and $\beta_2 = 3$. For each of the 500 samples, attempt to estimate a probit model. In each of the five cases, what proportion of the time does the estimation fail because of perfect classifiers? Explain why there were more failures in some cases than in others.

Repeat this exercise for five sets of 100 samples of size 40, with the same parameter values. What do you conclude about the effect of sample size on the perfect classifier problem?

4.6 Suppose that there is quasi-complete separation of the data used to estimate the binary response model (4.01), with a transformation function F such that $F(-x) = 1 - F(x)$ for all real x, and a separating hyperplane defined by the parameter vector $\boldsymbol{\beta}^\bullet$. Show that the upper bound of the loglikelihood function (4.09) is equal to $-n_b \log 2$, where n_b is the number of observations for which $\boldsymbol{X}_t\boldsymbol{\beta}^\bullet = 0$.

4.7 The contribution to the loglikelihood function (4.09) made by observation t is $y_t \log F(\boldsymbol{X}_t\boldsymbol{\beta}) + (1-y_t)\log(1 - F(\boldsymbol{X}_t\boldsymbol{\beta}))$. First, find G_{ti}, the derivative of this contribution with respect to β_i. Next, show that the expectation of G_{ti} is zero when it is evaluated at the true $\boldsymbol{\beta}$. Then obtain a typical element of the asymptotic information matrix by using the fact that it is equal to $\lim_{n\to\infty} n^{-1}\sum_{t=1}^{n} \mathrm{E}(G_{ti}G_{tj})$. Finally, show that the asymptotic covariance matrix (4.15) is equal to the inverse of this asymptotic information matrix.

4.8 Calculate the Hessian matrix corresponding to the loglikelihood function (4.09). Then use the fact that minus the expectation of the asymptotic Hessian is equal to the asymptotic information matrix to obtain the same result for the latter that you obtained in the previous exercise.

⋆**4.9** Plot $\Upsilon_t(\boldsymbol{\beta})$, which is defined in equation (4.16), as a function of $\boldsymbol{X}_t\boldsymbol{\beta}$ for both the logit and probit models. For the logit model only, prove that $\Upsilon_t(\boldsymbol{\beta})$ achieves its maximum value when $\boldsymbol{X}_t\boldsymbol{\beta} = 0$ and declines monotonically as $|\boldsymbol{X}_t\boldsymbol{\beta}|$ increases.

4.10 The file **participation.data**, which is taken from Gerfin (1996), contains data for 872 Swiss women who may or may not participate in the labor force. The variables in the file are:

y_t	Labor force participation variable (0 or 1).
I_t	Log of nonlabor income.
A_t	Age in decades (years divided by 10).
E_t	Education in years.
nu_t	Number of children under 7 years of age.
no_t	Number of children over 7 years of age.
F_t	Citizenship dummy variable (1 if not Swiss).

The dependent variable is y_t. For the standard specification, the regressors are all of the other variables, plus A_t^2. Estimate the standard specification as both a probit and a logit model. Is there any reason to prefer one of these two models?

4.11 For the probit model estimated in Exercise 4.10, obtain at least three sensible sets of standard error estimates. If possible, these should include ones based on the Hessian, ones based on the OPG estimator (3.43), and ones based on the information matrix estimator (4.18). You may make use of the BRMR, regression (4.20), and/or the OPG regression (3.70), if appropriate.

4.12 Test the hypothesis that the probit model estimated in Exercise 4.10 should include two additional regressors, namely, the squares of nu_t and no_t. Do this in three different ways, by calculating an LR statistic and two LM statistics based on the OPG and BRMR regressions.

4.13 Use the BRMR (4.30) to test the specification of the probit model estimated in Exercise 4.10. Then use the BRMR (4.26) to test for heteroskedasticity, where $\boldsymbol{Z}_t$ consists of all the regressors except the constant term.

⋆**4.14** Show, by use of l'Hôpital's Rule or otherwise, that the two results in (4.29) hold for all functions $\tau(\cdot)$ which satisfy conditions (4.28).

4.15 For the probit model estimated in Exercise 4.10, the estimated probability that $y_t = 1$ for observation t is $\Phi(\boldsymbol{X}_t\hat{\boldsymbol{\beta}})$. Compute this estimated probability for every observation, and also compute two confidence intervals at the .95

level for the actual probabilities. Both confidence intervals should be based on the covariance matrix estimator (4.18). One of them should use the delta method (Part 1, Section 6.6), and the other should be obtained by transforming the end points of a confidence interval for the index function. Compare the two intervals for the observations numbered 2, 63, and 311 in the sample. Are both intervals symmetric about the estimated probability? Which of them provides more reasonable answers?

⋆**4.16** Consider the expression

$$-\log\left(\sum_{j=0}^{J}\exp(\boldsymbol{W}_{tj}\boldsymbol{\beta}^j)\right), \tag{4.89}$$

which appears in the loglikelihood function (4.35) of the multinomial logit model. Let the vector $\boldsymbol{\beta}^j$ have k_j components, let $k \equiv k_0 + \ldots + k_J$, and let $\boldsymbol{\beta} \equiv [\boldsymbol{\beta}^0 \vdots \ldots \vdots \boldsymbol{\beta}^J]$. The $k \times k$ Hessian matrix $\boldsymbol{H}$ of (4.89) with respect to $\boldsymbol{\beta}$ can be partitioned into blocks of dimension $k_i \times k_j$, $i = 0, \ldots, J$, $j = 0, \ldots, J$, containing the second-order partial derivatives of (4.89) with respect to an element of $\boldsymbol{\beta}^i$ and an element of $\boldsymbol{\beta}^j$. Show that, for $i \neq j$, the (i, j) block can be written as

$$p_i p_j \boldsymbol{W}_{ti}^\top \boldsymbol{W}_{tj},$$

where $p_i \equiv \exp(\boldsymbol{W}_{ti}\boldsymbol{\beta}^i)/(\sum_{j=0}^{J}\exp(\boldsymbol{W}_{tj}\boldsymbol{\beta}^j))$ is the probability ascribed to choice i by the multinomial logit model. Then show that the diagonal (i, i) block can be written as

$$-p_i(1 - p_i)\boldsymbol{W}_{ti}^\top \boldsymbol{W}_{ti}.$$

Let the k–vector $\boldsymbol{a}$ be partitioned conformably with the above partitioning of the Hessian $\boldsymbol{H}$, so that we can write $\boldsymbol{a} = [\boldsymbol{a}_0 \vdots \ldots \vdots \boldsymbol{a}_J]$, where each of the vectors $\boldsymbol{a}_j$ has k_j components for $j = 0, \ldots, J$. Show that the quadratic form $\boldsymbol{a}^\top \boldsymbol{H}\boldsymbol{a}$ is equal to

$$\left(\sum_{j=0}^{J} p_j w_j\right)^2 - \sum_{j=0}^{J} p_j w_j^2, \tag{4.90}$$

where the scalar product w_j is defined as $\boldsymbol{W}_{tj}\boldsymbol{a}_j$.

Show that expression (4.90) is nonpositive, and explain why this result shows that the multinomial logit loglikelihood function (4.35) is globally concave.

4.17 Show that the nested logit model reduces to the multinomial logit model if $\theta_i = 1$ for all $i = 1, \ldots, m$. Then show that it also does so if all the subsets A_i used to define the former model are singletons.

⋆**4.18** Show that the expectation of the Hessian of the loglikelihood function (4.41), evaluated at the parameter vector $\boldsymbol{\theta}$, is equal to the negative of the $k \times k$ matrix

$$\boldsymbol{I}(\boldsymbol{\theta}) \equiv \sum_{t=1}^{n}\sum_{j=0}^{J}\frac{1}{\Pi_{tj}(\boldsymbol{\theta})}\boldsymbol{T}_{tj}^\top(\boldsymbol{\theta})\boldsymbol{T}_{tj}(\boldsymbol{\theta}), \tag{4.91}$$

where $\boldsymbol{T}_{tj}(\boldsymbol{\theta})$ is the $1 \times k$ vector of partial derivatives of $\Pi_{tj}(\boldsymbol{\theta})$ with respect to the components of $\boldsymbol{\theta}$. Demonstrate that (4.91) can also be computed using the outer product of the gradient definition of the information matrix.

Use the above result to show that the matrix of sums of squares and cross-products of the regressors of the DCAR, regression (4.42), evaluated at $\boldsymbol{\theta}$, is $\boldsymbol{I}(\boldsymbol{\theta})$. Show further that $1/s^2$ times the estimated OLS covariance matrix from (4.42) is an asymptotically valid estimate of the covariance matrix of the MLE $\hat{\boldsymbol{\theta}}$ if the artificial variables are evaluated at $\hat{\boldsymbol{\theta}}$.

⋆**4.19** Let the one-step estimator $\grave{\boldsymbol{\theta}}$ be defined as usual for the discrete choice artificial regression (4.42) evaluated at a root-n consistent estimator $\acute{\boldsymbol{\theta}}$ as $\grave{\boldsymbol{\theta}} = \acute{\boldsymbol{\theta}} + \acute{\boldsymbol{b}}$, where $\acute{\boldsymbol{b}}$ is the vector of OLS parameter estimates from (4.42). Show that $\grave{\boldsymbol{\theta}}$ is asymptotically equivalent to the MLE $\hat{\boldsymbol{\theta}}$.

4.20 Consider the binary choice model characterized by the probabilities (4.01). Both the BRMR (4.20) and the DCAR (4.42) with $J = 1$ apply to this model, but the two artificial regressions are obviously different, since the BRMR has n artificial observations when the sample size is n, while the DCAR has $2n$. Show that the two artificial regressions are nevertheless equivalent, in the sense that all scalar products of corresponding pairs of artificial variables, regressand or regressor, are identical for the two regressions.

⋆**4.21** In terms of the notation of the DCAR, regression (4.42), the probability Π_{tj} that $y_t = j$, $j = 0, \ldots, J$, for the nested logit model is given by expression (4.40). Show that, if the index $i(j)$ is such that $j \in A_{i(j)}$, the partial derivative of Π_{tj} with respect to θ_i, evaluated at $\theta_k = 1$ for $k = 1, \ldots, m$, where m is the number of subsets A_k, is

$$\frac{\partial \Pi_{tj}}{\partial \theta_i} = \Pi_{tj}(\delta_{i(j)i}v_{tj} - \sum_{l \in A_i}\Pi_{tl}v_{tl}).$$

Here $v_{tj} \equiv -\boldsymbol{W}_{tj}\boldsymbol{\beta}^j + h_{ti(j)}$, where h_{ti} denotes the inclusive value (4.39) of subset A_i, and δ_{ij} is the Kronecker delta.

When $\theta_k = 1$, $k = 1, \ldots, m$, the nested logit probabilities reduce to the multinomial logit probabilities (4.34). Show that, if the Π_{tj} are given by (4.34), then the vector of partial derivatives of Π_{tj} with respect to the components of $\boldsymbol{\beta}^l$ is $\Pi_{tj}\boldsymbol{W}_{tl}(\delta_{jl} - \Pi_{tl})$.

⋆**4.22** Explain how to use the DCAR (4.42) to test the IIA assumption for the conditional logit model (4.36). This involves testing it against the nested logit model (4.40) with the $\boldsymbol{\beta}^j$ constrained to be the same. Do this for the special case in which $J = 2$, $A_1 = \{0, 1\}$, $A_2 = \{2\}$. **Hint:** Use the results proved in the preceding exercise.

4.23 Using the fact that the infinite series expansion of the exponential function, convergent for all real z, is

$$\exp z = \sum_{n=0}^{\infty}\frac{z^n}{n!},$$

where by convention we define $0! = 1$, show that $\sum_{y=0}^{\infty} e^{-\lambda}\lambda^y/y! = 1$, and that therefore the Poisson distribution defined by (4.47) is well defined on the nonnegative integers. Then show that the expectation and variance of a random variable Y that follows the Poisson distribution are both equal to λ.

4.24 Let the n^{th} uncentered moment of the Poisson distribution with parameter λ be denoted by $M_n(\lambda)$. Show that these moments can be generated by the

recurrence $M_{n+1}(\lambda) = \lambda(M_n(\lambda) + M_n'(\lambda))$, where $M_n'(\lambda)$ is the derivative of $M_n(\lambda)$. Using this result, show that the third and fourth *central* moments of the Poisson distribution are λ and $\lambda + 3\lambda^2$, respectively.

4.25 Explain precisely how you would use the artificial regression (4.55) to test the hypothesis that $\boldsymbol{\beta}_2 = \mathbf{0}$ in the Poisson regression model for which $\lambda_t(\boldsymbol{\beta}) = \exp(\boldsymbol{X}_{t1}\boldsymbol{\beta}_1 + \boldsymbol{X}_{t2}\boldsymbol{\beta}_2)$. Here $\boldsymbol{\beta}_1$ is a k_1–vector and $\boldsymbol{\beta}_2$ is a k_2–vector, with $k = k_1 + k_2$. Consider two cases, one in which the model is estimated subject to the restriction and one in which it is estimated unrestrictedly.

⋆**4.26** Suppose that y_t is a count variable, with conditional expectation $\mathrm{E}(y_t) = \exp(\boldsymbol{X}_t\boldsymbol{\beta})$ and conditional variance $\mathrm{E}(y_t - \exp(\boldsymbol{X}_t\boldsymbol{\beta}))^2 = \gamma^2 \exp(\boldsymbol{X}_t\boldsymbol{\beta})$. Show that ML estimates of $\boldsymbol{\beta}$ under the incorrect assumption that y_t is generated by a Poisson regression model with mean $\exp(\boldsymbol{X}_t\boldsymbol{\beta})$ are asymptotically efficient in this case. Also show that the OLS covariance matrix from the artificial regression (4.55) is asymptotically valid.

4.27 Suppose that y_t is a count variable with conditional mean $\mathrm{E}(y_t) = \exp(\boldsymbol{X}_t\boldsymbol{\beta})$ and unknown conditional variance. Show that, if the artificial regression (4.55) is evaluated at the ML estimates for a Poisson regression model which specifies the conditional mean correctly, the HCCME HC_0 for that artificial regression is numerically equal to expression (4.65), which is an asymptotically valid covariance matrix estimator in this case.

4.28 The file **count.data**, which is taken from Gurmu (1997), contains data for 485 household heads who may or may not have visited a doctor during a certain period of time. The variables in the file are:

y_t	Number of doctor visits (a nonnegative integer).
C_t	Number of children in the household.
A_t	A measure of access to health care.
H_t	A measure of health status.

Using these data, obtain ML estimates of a Poisson regression model to explain the variable y_t, where

$$\lambda_t(\boldsymbol{\beta}) = \exp(\beta_1 + \beta_2 C_t + \beta_3 A_t + \beta_4 H_t).$$

In addition to the estimates of the parameters, report three different standard errors. One of these should be based on the inverse of the information matrix, which is valid only when the model is correctly specified. The other two should be computed using the artificial regression (4.55). One of them should be valid under the assumption that the conditional variance is proportional to $\lambda_t(\boldsymbol{\beta})$, and the other should be valid whenever the conditional mean is specified correctly. Can you explain the differences among the three sets of standard errors?

Test the model for overdispersion in two different ways. One test should be based on the OPG regression, and the other should be based on the testing regression (4.60). Note that this model is *not* the one actually estimated in Gurmu (1997).

4.29 Consider the latent variable model

$$y_t^\circ = \boldsymbol{X}_t\boldsymbol{\beta} + u_t, \quad u_t \sim \mathrm{NID}(0, \sigma^2), \tag{4.92}$$

where $y_t = y_t^\circ$ whenever $y_t^\circ \leq y^{\max}$ and is not observed otherwise. Write down the loglikelihood function for a sample of n observations on y_t.

4.30 As in the previous question, suppose that y_t° is given by (4.92). Assume that $y_t = y_t^\circ$ whenever $y^{\min} \leq y_t^\circ \leq y^{\max}$ and is not observed otherwise. Write down the loglikelihood function for a sample of n observations on y_t.

4.31 Suppose that $y_t^\circ = \boldsymbol{X}_t\boldsymbol{\beta} + u_t$ with $u_t \sim \mathrm{NID}(0, \sigma^2)$. Suppose further that $y_t = y_t^\circ$ if $y_t^\circ < y_t^c$, and $y_t = y_t^c$ otherwise, where y_t^c is the known value at which censoring occurs for observation t. Write down the loglikelihood function for this model.

⋆**4.32** Let z be distributed as $\mathrm{N}(0,1)$. Show that $\mathrm{E}(z \,|\, z < x) = -\phi(x)/\Phi(x)$, where Φ and ϕ are, respectively, the CDF and PDF of the standard normal distribution. Then show that $\mathrm{E}(z \,|\, z > x) = \phi(x)/\Phi(-x) = \phi(-x)/\Phi(-x)$. The second result explains why the inverse Mills ratio appears in (4.77).

4.33 Starting from expression (4.82) for the CDF of the Weibull distribution, show that the survivor function, the PDF, and the hazard function are as given in (4.83).

Chapter 5

Multivariate Models

5.1 Introduction

Up to this point, almost all the models we have discussed have involved just one equation. In most cases, there has been only one equation because there has been only one dependent variable. Even in the few cases in which there were several dependent variables, interest centered on just one of them. For example, in the case of the simultaneous equations model that was discussed in Chapter 10, we chose to estimate just one structural equation at a time.

In this chapter, we discuss models which jointly determine the values of two or more dependent variables using two or more equations. Such models are called **multivariate** because they attempt to explain multiple dependent variables. As we will see, the class of multivariate models is considerably larger than the class of simultaneous equations models. Every simultaneous equations model is a multivariate model, but many interesting multivariate models are not simultaneous equations models.

In the next section, which is quite long, we provide a detailed discussion of GLS, feasible GLS, and ML estimation of systems of linear regressions. Then, in Section 5.3, we discuss the estimation of systems of nonlinear equations which may involve cross-equation restrictions but do not involve simultaneity. Next, in Section 5.4, we provide a detailed treatment of the linear simultaneous equations model. We approach it from the point of view of GMM estimation, which leads to the well-known 3SLS estimator. In Section 5.5, we discuss the application of maximum likelihood to this model. Finally, in Section 5.6, we briefly discuss some of the methods for estimating nonlinear simultaneous equations models.

5.2 Seemingly Unrelated Linear Regressions

The **multivariate linear regression model** was investigated by Zellner (1962), who called it the **seemingly unrelated regressions model**. An **SUR system**, as such a model is often called, involves n observations on each of g dependent variables. In principle, these could be any set of variables measured at the same points in time or for the same cross-section. In practice, however, the

dependent variables are often quite similar to each other. For example, in the time-series context, each of them might be the output of a different industry or the inflation rate for a different country. In view of this, it might seem more appropriate to speak of "seemingly related regressions," but the terminology is too well-established to change.

We suppose that there are g dependent variables indexed by i. Let $\boldsymbol{y}_i$ denote the n–vector of observations on the i^{th} dependent variable, $\boldsymbol{X}_i$ denote the $n \times k_i$ matrix of regressors for the i^{th} equation, $\boldsymbol{\beta}_i$ denote the k_i–vector of parameters, and $\boldsymbol{u}_i$ denote the n–vector of disturbances. Then the i^{th} equation of a multivariate linear regression model may be written as

$$\boldsymbol{y}_i = \boldsymbol{X}_i \boldsymbol{\beta}_i + \boldsymbol{u}_i, \quad \mathrm{E}(\boldsymbol{u}_i \boldsymbol{u}_i^\top) = \sigma_{ii} \mathbf{I}_n, \tag{5.01}$$

where $\mathbf{I}_n$ is the $n \times n$ identity matrix. The reason we use σ_{ii} to denote the variance of the disturbances will become apparent shortly. In most cases, some columns are common to two or more of the matrices $\boldsymbol{X}_i$. For instance, if every equation has a constant term, each of the $\boldsymbol{X}_i$ must contain a column of 1s.

Since equation (5.01) is just a linear regression model with IID disturbances, we can perfectly well estimate it by ordinary least squares if we assume that all the columns of $\boldsymbol{X}_i$ are either exogenous or predetermined. If we do this, however, we ignore the possibility that the disturbances may be correlated across the equations of the system. In many cases, it is plausible that u_{ti}, the disturbance for observation t of equation i, should be correlated with u_{tj}, the disturbance for observation t of equation j. For example, we might expect that a macroeconomic shock which affects the inflation rate in one country would simultaneously affect the inflation rate in other countries as well.

To allow for this possibility, the assumption that is usually made about the disturbances in the model (5.01) is

$$\mathrm{E}(u_{ti} u_{tj}) = \sigma_{ij} \text{ for all } t, \quad \mathrm{E}(u_{ti} u_{sj}) = 0 \text{ for all } t \neq s, \tag{5.02}$$

where σ_{ij} is the ij^{th} element of the $g \times g$ positive definite matrix $\boldsymbol{\Sigma}$. This assumption allows all the u_{ti} for a given t to be correlated, but it specifies that they are homoskedastic and independent across t. The matrix $\boldsymbol{\Sigma}$ is called the **contemporaneous covariance matrix**, a term inspired by the time-series context. The disturbances u_{ti} may be arranged into an $n \times g$ matrix $\boldsymbol{U}$, of which a typical row is the $1 \times g$ vector $\boldsymbol{U}_t$. It then follows from (5.02) that

$$\mathrm{E}(\boldsymbol{U}_t^\top \boldsymbol{U}_t) = \frac{1}{n} \mathrm{E}(\boldsymbol{U}^\top \boldsymbol{U}) = \boldsymbol{\Sigma}. \tag{5.03}$$

If we combine equations (5.01), for $i = 1, \dots, g$, with assumption (5.02), we obtain the classical SUR model.

We have not yet made any sort of exogeneity or predeterminedness assumption. A rather strong assumption is that $\mathrm{E}(\boldsymbol{U} \mid \boldsymbol{X}) = \mathbf{O}$, where $\boldsymbol{X}$ is an $n \times l$

matrix with full rank, the set of columns of which is the union of all the linearly independent columns of all the matrices $\boldsymbol{X}_i$. Thus l is the total number of variables that appear in any of the $\boldsymbol{X}_i$ matrices. This exogeneity assumption, which is the analog of assumption (F4.11) for univariate regression models, is undoubtedly too strong in many cases. A considerably weaker assumption is that $\mathrm{E}(\boldsymbol{U}_t \mid \boldsymbol{X}_t) = \boldsymbol{0}$, where $\boldsymbol{X}_t$ is the t^{th} row of $\boldsymbol{X}$. This is the analog of the predeterminedness assumption (F4.13) for univariate regression models. The results that we will state are valid under either of these assumptions.

Precisely how we want to estimate a linear SUR system depends on what further assumptions we make about the matrix $\boldsymbol{\Sigma}$ and the distribution of the disturbances. In the simplest case, $\boldsymbol{\Sigma}$ is assumed to be known, at least up to a scalar factor, and the distribution of the disturbances is unspecified. The appropriate estimation method is then generalized least squares. If we relax the assumption that $\boldsymbol{\Sigma}$ is known, then we need to use feasible GLS. If we continue to assume that $\boldsymbol{\Sigma}$ is unknown but impose the assumption that the disturbances are normally distributed, then we may want to use maximum likelihood, which is generally consistent even when the normality assumption is false. In practice, both feasible GLS and ML are widely used.

GLS Estimation with a Known Covariance Matrix

Even though it is rarely a realistic assumption, we begin by assuming that the contemporaneous covariance matrix $\boldsymbol{\Sigma}$ of a linear SUR system is known, and we consider how to estimate the model by GLS. Once we have seen how to do so, it will be easy to see how to estimate such a model by other methods. The trick is to convert a system of g linear equations and n observations into what looks like a single equation with gn observations and a known $gn \times gn$ covariance matrix that depends on $\boldsymbol{\Sigma}$.

By making appropriate definitions, we can write the entire SUR system of which a typical equation is (5.01) as

$$\boldsymbol{y}_\bullet = \boldsymbol{X}_\bullet \boldsymbol{\beta}_\bullet + \boldsymbol{u}_\bullet. \tag{5.04}$$

Here $\boldsymbol{y}_\bullet$ is a gn–vector consisting of the n–vectors $\boldsymbol{y}_1$ through $\boldsymbol{y}_g$ stacked vertically, and $\boldsymbol{u}_\bullet$ is similarly the vector of $\boldsymbol{u}_1$ through $\boldsymbol{u}_g$ stacked vertically. The matrix $\boldsymbol{X}_\bullet$ is a $gn \times k$ block-diagonal matrix, where k is equal to $\sum_{i=1}^{g} k_i$. The diagonal blocks are the matrices $\boldsymbol{X}_1$ through $\boldsymbol{X}_g$. Thus we have

$$\boldsymbol{X}_\bullet \equiv \begin{bmatrix} \boldsymbol{X}_1 & \boldsymbol{O} & \cdots & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{X}_2 & \cdots & \boldsymbol{O} \\ \vdots & \vdots & \vdots & \vdots \\ \boldsymbol{O} & \boldsymbol{O} & \cdots & \boldsymbol{X}_g \end{bmatrix}, \tag{5.05}$$

where each of the $\boldsymbol{O}$ blocks has n rows and as many columns as the $\boldsymbol{X}_i$ block that it shares those columns with. To be conformable with $\boldsymbol{X}_\bullet$, the vector $\boldsymbol{\beta}_\bullet$ is a k–vector consisting of the vectors $\boldsymbol{\beta}_1$ through $\boldsymbol{\beta}_g$ stacked vertically.

From the above definitions and the rules for matrix multiplication, it is not difficult to see that

$$\begin{bmatrix} \boldsymbol{y}_1 \\ \vdots \\ \boldsymbol{y}_g \end{bmatrix} \equiv \boldsymbol{y}_\bullet = \boldsymbol{X}_\bullet \boldsymbol{\beta}_\bullet + \boldsymbol{u}_\bullet = \begin{bmatrix} \boldsymbol{X}_1 \boldsymbol{\beta}_1 \\ \vdots \\ \boldsymbol{X}_g \boldsymbol{\beta}_g \end{bmatrix} + \begin{bmatrix} \boldsymbol{u}_1 \\ \vdots \\ \boldsymbol{u}_g \end{bmatrix}.$$

Thus it is apparent that the single equation (5.04) is precisely what we obtain by stacking the equations (5.01) vertically, for $i = 1, \ldots, g$. Using the notation of (5.04), we can write the OLS estimator for the entire system very compactly as

$$\hat{\boldsymbol{\beta}}_\bullet^{\text{OLS}} = (\boldsymbol{X}_\bullet^\top \boldsymbol{X}_\bullet)^{-1} \boldsymbol{X}_\bullet^\top \boldsymbol{y}_\bullet, \tag{5.06}$$

as readers are asked to verify in Exercise 5.4. But the assumptions we have made about $\boldsymbol{u}_\bullet$ imply that this estimator is not efficient.

The next step is to figure out the covariance matrix of the vector $\boldsymbol{u}_\bullet$. Since the disturbances are assumed to have mean zero, this matrix is just the expectation of the matrix $\boldsymbol{u}_\bullet \boldsymbol{u}_\bullet^\top$. Under assumption (5.02), we find that

$$\begin{aligned} \mathrm{E}(\boldsymbol{u}_\bullet \boldsymbol{u}_\bullet^\top) &= \begin{bmatrix} \mathrm{E}(\boldsymbol{u}_1 \boldsymbol{u}_1^\top) & \cdots & \mathrm{E}(\boldsymbol{u}_1 \boldsymbol{u}_g^\top) \\ \vdots & \vdots & \vdots \\ \mathrm{E}(\boldsymbol{u}_g \boldsymbol{u}_1^\top) & \cdots & \mathrm{E}(\boldsymbol{u}_g \boldsymbol{u}_g^\top) \end{bmatrix} \\ &= \begin{bmatrix} \sigma_{11} \boldsymbol{I}_n & \cdots & \sigma_{1g} \boldsymbol{I}_n \\ \vdots & \vdots & \vdots \\ \sigma_{g1} \boldsymbol{I}_n & \cdots & \sigma_{gg} \boldsymbol{I}_n \end{bmatrix} \equiv \boldsymbol{\Sigma}_\bullet. \end{aligned} \tag{5.07}$$

Here, $\boldsymbol{\Sigma}_\bullet$ is a symmetric $gn \times gn$ covariance matrix. In Exercise 5.1, readers are asked to show that $\boldsymbol{\Sigma}_\bullet$ is positive definite whenever $\boldsymbol{\Sigma}$ is.

The matrix $\boldsymbol{\Sigma}_\bullet$ can be written more compactly as $\boldsymbol{\Sigma}_\bullet \equiv \boldsymbol{\Sigma} \otimes \boldsymbol{I}_n$ if we use the **Kronecker product** symbol $\otimes$. The Kronecker product $\boldsymbol{A} \otimes \boldsymbol{B}$ of a $p \times q$ matrix $\boldsymbol{A}$ and an $r \times s$ matrix $\boldsymbol{B}$ is a $pr \times qs$ matrix consisting of pq blocks, laid out in the pattern of the elements of $\boldsymbol{A}$. For $i = 1, \ldots, p$ and $j = 1, \ldots, q$, the ij^{th} block of the Kronecker product is the $r \times s$ matrix $a_{ij}\boldsymbol{B}$, where a_{ij} is the ij^{th} element of $\boldsymbol{A}$. As can be seen from (5.07), that is exactly how the blocks of $\boldsymbol{\Sigma}_\bullet$ are defined in terms of $\boldsymbol{I}_n$ and the elements of $\boldsymbol{\Sigma}$.

Kronecker products have a number of useful properties. In particular, if $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$, and $\boldsymbol{D}$ are conformable matrices, then the following relationships hold:

$$\begin{aligned} (\boldsymbol{A} \otimes \boldsymbol{B})^\top &= \boldsymbol{A}^\top \otimes \boldsymbol{B}^\top, \\ (\boldsymbol{A} \otimes \boldsymbol{B})(\boldsymbol{C} \otimes \boldsymbol{D}) &= (\boldsymbol{AC}) \otimes (\boldsymbol{BD}), \text{ and} \\ (\boldsymbol{A} \otimes \boldsymbol{B})^{-1} &= \boldsymbol{A}^{-1} \otimes \boldsymbol{B}^{-1}. \end{aligned} \tag{5.08}$$

Of course, the last line of (5.08) can be true only for nonsingular, square matrices $\boldsymbol{A}$ and $\boldsymbol{B}$. The Kronecker product is not commutative, by which we mean that $\boldsymbol{A} \otimes \boldsymbol{B}$ and $\boldsymbol{B} \otimes \boldsymbol{A}$ are different matrices. However, the elements of these two products are the same; they are just laid out differently. In fact, it can be shown that $\boldsymbol{B} \otimes \boldsymbol{A}$ can be obtained from $\boldsymbol{A} \otimes \boldsymbol{B}$ by a sequence of interchanges of rows and columns. Exercise 5.2 asks readers to prove these properties of Kronecker products. For an exceedingly detailed discussion of the properties of Kronecker products, see Magnus and Neudecker (1988).

As we have seen, the system of equations defined by (5.01) and (5.02) is equivalent to the single equation (5.04), with gn observations and disturbances that have covariance matrix $\boldsymbol{\Sigma}_\bullet$. Therefore, when the matrix $\boldsymbol{\Sigma}$ is known, we can obtain consistent and efficient estimates of the $\boldsymbol{\beta}_i$, or equivalently of $\boldsymbol{\beta}_\bullet$, simply by using the classical GLS estimator (F9.04). We find that

$$\begin{aligned}\hat{\boldsymbol{\beta}}_\bullet^{\text{GLS}} &= (\boldsymbol{X}_\bullet^\top \boldsymbol{\Sigma}_\bullet^{-1} \boldsymbol{X}_\bullet)^{-1} \boldsymbol{X}_\bullet^\top \boldsymbol{\Sigma}_\bullet^{-1} \boldsymbol{y}_\bullet \\ &= \big(\boldsymbol{X}_\bullet^\top (\boldsymbol{\Sigma}^{-1} \otimes \mathbf{I}_n) \boldsymbol{X}_\bullet\big)^{-1} \boldsymbol{X}_\bullet^\top (\boldsymbol{\Sigma}^{-1} \otimes \mathbf{I}_n) \boldsymbol{y}_\bullet, \end{aligned} \tag{5.09}$$

where, to obtain the second line, we have used the last of equations (5.08). This GLS estimator is sometimes called the **SUR estimator**. From the result (F9.05) for GLS estimation, its covariance matrix is

$$\text{Var}(\hat{\boldsymbol{\beta}}_\bullet^{\text{GLS}}) = \big(\boldsymbol{X}_\bullet^\top (\boldsymbol{\Sigma}^{-1} \otimes \mathbf{I}_n) \boldsymbol{X}_\bullet\big)^{-1}. \tag{5.10}$$

Since $\boldsymbol{\Sigma}$ is assumed to be known, we can use this covariance matrix directly, because there are no variance parameters to estimate.

As in the univariate case, there is a criterion function associated with the GLS estimator (F9.04). This criterion function is simply expression (F9.06) adapted to the model (5.04), namely,

$$(\boldsymbol{y}_\bullet - \boldsymbol{X}_\bullet \boldsymbol{\beta}_\bullet)^\top (\boldsymbol{\Sigma}^{-1} \otimes \mathbf{I}_n)(\boldsymbol{y}_\bullet - \boldsymbol{X}_\bullet \boldsymbol{\beta}_\bullet). \tag{5.11}$$

The first-order conditions for the minimization of (5.11) with respect to $\boldsymbol{\beta}_\bullet$ can be written as

$$\boldsymbol{X}_\bullet^\top (\boldsymbol{\Sigma}^{-1} \otimes \mathbf{I}_n)(\boldsymbol{y}_\bullet - \boldsymbol{X}_\bullet \boldsymbol{\beta}_\bullet) = \mathbf{0}. \tag{5.12}$$

These moment conditions, which are analogous to conditions (F9.07) for the case of univariate GLS estimation, can be interpreted as a set of estimating equations that define the GLS estimator (5.09).

In the slightly less unrealistic situation in which $\boldsymbol{\Sigma}$ is assumed to be known only up to a scalar factor, so that $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{\Delta}$, the form of (5.09) would be unchanged, but with $\boldsymbol{\Delta}$ replacing $\boldsymbol{\Sigma}$, and the covariance matrix (5.10) would become

$$\text{Var}(\hat{\boldsymbol{\beta}}_\bullet^{\text{GLS}}) = \sigma^2 \big(\boldsymbol{X}_\bullet^\top (\boldsymbol{\Delta}^{-1} \otimes \mathbf{I}_n) \boldsymbol{X}_\bullet\big)^{-1}.$$

In practice, to estimate $\text{Var}(\hat{\boldsymbol{\beta}}_\bullet^{\text{GLS}})$, we replace σ^2 by something that estimates it consistently. Two natural estimators are

$$\hat{\sigma}^2 \equiv \frac{1}{gn} \hat{\boldsymbol{u}}_\bullet^\top (\boldsymbol{\Delta}^{-1} \otimes \mathbf{I}_n) \hat{\boldsymbol{u}}_\bullet, \quad \text{and}$$
$$s^2 \equiv \frac{1}{(gn-k)} \hat{\boldsymbol{u}}_\bullet^\top (\boldsymbol{\Delta}^{-1} \otimes \mathbf{I}_n) \hat{\boldsymbol{u}}_\bullet,$$

where $\hat{\boldsymbol{u}}_\bullet$ denotes the vector of disturbances from GLS estimation of (5.04). The first of these estimators is analogous to the ML estimator of σ^2 in the linear regression model, and the second is analogous to the GLS estimator.

At this point, a word of warning is in order. Although the GLS estimator (5.09) has quite a simple form, it can be expensive to compute when gn is large. In consequence, no sensible regression package would actually use this formula. We can proceed more efficiently by working directly with the estimating equations (5.12). Writing them out explicitly, we obtain

$$\begin{aligned}&\boldsymbol{X}_\bullet^\top (\boldsymbol{\Sigma}^{-1} \otimes \mathbf{I}_n)(\boldsymbol{y}_\bullet - \boldsymbol{X}_\bullet \hat{\boldsymbol{\beta}}_\bullet) \\ &\quad = \begin{bmatrix} \boldsymbol{X}_1^\top & \cdots & \mathbf{O} \\ \vdots & \ddots & \vdots \\ \mathbf{O} & \cdots & \boldsymbol{X}_g^\top \end{bmatrix} \begin{bmatrix} \sigma^{11}\mathbf{I}_n & \cdots & \sigma^{1g}\mathbf{I}_n \\ \vdots & \ddots & \vdots \\ \sigma^{g1}\mathbf{I}_n & \cdots & \sigma^{gg}\mathbf{I}_n \end{bmatrix} \begin{bmatrix} \boldsymbol{y}_1 - \boldsymbol{X}_1 \hat{\boldsymbol{\beta}}_1^{\text{GLS}} \\ \vdots \\ \boldsymbol{y}_g - \boldsymbol{X}_g \hat{\boldsymbol{\beta}}_g^{\text{GLS}} \end{bmatrix} \\ &\quad = \begin{bmatrix} \sigma^{11}\boldsymbol{X}_1^\top & \cdots & \sigma^{1g}\boldsymbol{X}_1^\top \\ \vdots & \ddots & \vdots \\ \sigma^{g1}\boldsymbol{X}_g^\top & \cdots & \sigma^{gg}\boldsymbol{X}_g^\top \end{bmatrix} \begin{bmatrix} \boldsymbol{y}_1 - \boldsymbol{X}_1 \hat{\boldsymbol{\beta}}_1^{\text{GLS}} \\ \vdots \\ \boldsymbol{y}_g - \boldsymbol{X}_g \hat{\boldsymbol{\beta}}_g^{\text{GLS}} \end{bmatrix} = \mathbf{0}, \end{aligned} \tag{5.13}$$

where σ^{ij} denotes the ij^{th} element of the matrix $\boldsymbol{\Sigma}^{-1}$. By solving the k equations (5.13) for the $\hat{\boldsymbol{\beta}}_i$, we find easily enough (see Exercise 5.5) that

$$\hat{\boldsymbol{\beta}}_\bullet^{\text{GLS}} = \begin{bmatrix} \sigma^{11}\boldsymbol{X}_1^\top \boldsymbol{X}_1 & \cdots & \sigma^{1g}\boldsymbol{X}_1^\top \boldsymbol{X}_g \\ \vdots & \ddots & \vdots \\ \sigma^{g1}\boldsymbol{X}_g^\top \boldsymbol{X}_1 & \cdots & \sigma^{gg}\boldsymbol{X}_g^\top \boldsymbol{X}_g \end{bmatrix}^{-1} \begin{bmatrix} \sum_{j=1}^g \sigma^{1j}\boldsymbol{X}_1^\top \boldsymbol{y}_j \\ \vdots \\ \sum_{j=1}^g \sigma^{gj}\boldsymbol{X}_g^\top \boldsymbol{y}_j \end{bmatrix}. \tag{5.14}$$

Although this expression may look more complicated than (5.09), it is much less costly to compute. Recall that we grouped all the linearly independent explanatory variables of the entire SUR system into the $n \times l$ matrix $\boldsymbol{X}$. By computing the matrix product $\boldsymbol{X}^\top \boldsymbol{X}$, we may obtain all the blocks of the form $\boldsymbol{X}_i^\top \boldsymbol{X}_j$ merely by selecting the appropriate rows and corresponding columns of this product. Similarly, if we form the $n \times g$ matrix $\boldsymbol{Y}$ by stacking the g dependent variables horizontally rather than vertically, so that

$$\boldsymbol{Y} \equiv [\, \boldsymbol{y}_1 \quad \cdots \quad \boldsymbol{y}_g \,],$$

then all the vectors of the form $\boldsymbol{X}_i^\top \boldsymbol{y}_j$ needed on the right-hand side of (5.14) can be extracted as a selection of the elements of the j^{th} column of the product $\boldsymbol{X}^\top \boldsymbol{Y}$.

The covariance matrix (5.10) can also be expressed in a form more suitable for computation. By a calculation just like the one that gave us (5.13), we see that (5.10) can be expressed as

$$\operatorname{Var}(\hat{\boldsymbol{\beta}}_\bullet^{\text{GLS}}) = \begin{bmatrix} \sigma^{11}\boldsymbol{X}_1^\top \boldsymbol{X}_1 & \cdots & \sigma^{1g}\boldsymbol{X}_1^\top \boldsymbol{X}_g \\ \vdots & \ddots & \vdots \\ \sigma^{g1}\boldsymbol{X}_g^\top \boldsymbol{X}_1 & \cdots & \sigma^{gg}\boldsymbol{X}_g^\top \boldsymbol{X}_g \end{bmatrix}^{-1}. \tag{5.15}$$

Again, all the blocks here are selections of rows and columns of $\boldsymbol{X}^\top \boldsymbol{X}$.

For the purposes of further analysis, the estimating equations (5.13) can be expressed more concisely by writing out the i^{th} row as follows:

$$\sum_{j=1}^{g} \sigma^{ij} \boldsymbol{X}_i^\top (\boldsymbol{y}_j - \boldsymbol{X}_j \hat{\boldsymbol{\beta}}_j^{\text{GLS}}) = \boldsymbol{0}. \tag{5.16}$$

The matrix equation (5.13) is clearly equivalent to the set of equations (5.16) for $i = 1, \ldots, g$.

Feasible GLS Estimation

In practice, the contemporaneous covariance matrix $\boldsymbol{\Sigma}$ is very rarely known. When it is not, the easiest approach is simply to replace $\boldsymbol{\Sigma}$ in (5.09) by a matrix that estimates it consistently. In principle, there are many ways to do so, but the most natural approach is to base the estimate on OLS residuals. This leads to the following feasible GLS procedure, which is probably the most commonly-used procedure for estimating linear SUR systems.

The first step is to estimate each of the equations by OLS. This yields consistent, but inefficient, estimates of the $\boldsymbol{\beta}_i$, along with g vectors of least-squares residuals $\hat{\boldsymbol{u}}_i$. The natural estimator of $\boldsymbol{\Sigma}$ is then

$$\hat{\boldsymbol{\Sigma}} \equiv \frac{1}{n} \hat{\boldsymbol{U}}^\top \hat{\boldsymbol{U}}, \tag{5.17}$$

where $\hat{\boldsymbol{U}}$ is an $n \times g$ matrix with i^{th} column $\hat{\boldsymbol{u}}_i$. By construction, the matrix $\hat{\boldsymbol{\Sigma}}$ is symmetric, and it is positive definite whenever the columns of $\hat{\boldsymbol{U}}$ are not linearly dependent. The feasible GLS estimator is given by

$$\hat{\boldsymbol{\beta}}_\bullet^{\text{F}} = \big(\boldsymbol{X}_\bullet^\top (\hat{\boldsymbol{\Sigma}}^{-1} \otimes \mathbf{I}_n) \boldsymbol{X}_\bullet\big)^{-1} \boldsymbol{X}_\bullet^\top (\hat{\boldsymbol{\Sigma}}^{-1} \otimes \mathbf{I}_n) \boldsymbol{y}_\bullet, \tag{5.18}$$

and the natural way to estimate its covariance matrix is

$$\widehat{\operatorname{Var}}(\hat{\boldsymbol{\beta}}_\bullet^{\text{F}}) = \big(\boldsymbol{X}_\bullet^\top (\hat{\boldsymbol{\Sigma}}^{-1} \otimes \mathbf{I}_n) \boldsymbol{X}_\bullet\big)^{-1}. \tag{5.19}$$

As expected, the feasible GLS estimator (5.18) and the estimated covariance matrix (5.19) have precisely the same forms as their full GLS counterparts, which are (5.09) and (5.10), respectively.

Because we divided by n in (5.17), $\hat{\boldsymbol{\Sigma}}$ must be a biased estimator of $\boldsymbol{\Sigma}$. If k_i is the same for all i, then it would seem natural to divide by $n - k_i$ instead, and this would produce unbiased estimates of the diagonal elements if $\boldsymbol{X}_\bullet$ were exogenous. But we cannot do that when k_i is not the same in all equations. If we were to divide different elements of $\hat{\boldsymbol{U}}^\top \hat{\boldsymbol{U}}$ by different quantities, the resulting estimate of $\boldsymbol{\Sigma}$ would not necessarily be positive definite.

Replacing $\boldsymbol{\Sigma}$ with an estimator $\hat{\boldsymbol{\Sigma}}$ based on OLS estimates, or indeed any other estimator, inevitably degrades the finite-sample properties of the GLS estimator. In general, we would expect the performance of the feasible GLS estimator, relative to that of the GLS estimator, to be especially poor when the sample size is small and the number of equations is large. Under the strong assumption that all the regressors are exogenous, exact inference based on the normal and χ^2 distributions is possible whenever the disturbances are normally distributed and $\boldsymbol{\Sigma}$ (or $\boldsymbol{\Delta}$) is known, but this is not the case when $\boldsymbol{\Sigma}$ has to be estimated. Not surprisingly, there is evidence that bootstrapping can yield more reliable inferences than using asymptotic theory for SUR models; see, among others, Rilstone and Veall (1996) and Fiebig and Kim (2000).

Cases in Which OLS Estimation Is Efficient

The SUR estimator (5.09) is efficient under the assumptions we have made, because it is just a special case of the GLS estimator (F9.04), the efficiency of which was proved in Section 7.2. In contrast, the OLS estimator (5.06) is, in general, inefficient. The reason is that, unless the matrix $\boldsymbol{\Sigma}$ is proportional to an identity matrix, the disturbances of equation (5.04) are not IID. Nevertheless, there are two important special cases in which the OLS estimator is numerically identical to the SUR estimator, and therefore just as efficient.

In the first case, the matrix $\boldsymbol{\Sigma}$ is diagonal, although the diagonal elements need not be the same. This implies that the disturbances of equation (5.04) are heteroskedastic but serially independent. It might seem that this heteroskedasticity would cause inefficiency, but that turns out not to be the case. If $\boldsymbol{\Sigma}$ is diagonal, then so is $\boldsymbol{\Sigma}^{-1}$, which means that $\sigma^{ij} = 0$ for $i \neq j$. In that case, the estimating equations (5.16) simplify to

$$\sigma^{ii} \boldsymbol{X}_i^\top (\boldsymbol{y}_i - \boldsymbol{X}_i \hat{\boldsymbol{\beta}}_i^{\text{GLS}}) = \boldsymbol{0}, \quad i = 1, \ldots, g.$$

The factors σ^{ii}, which must be nonzero, have no influence on the solutions to the above equations, which are therefore the same as the solutions to the g independent sets of equations $\boldsymbol{X}_i^\top (\boldsymbol{y}_i - \boldsymbol{X}_i \hat{\boldsymbol{\beta}}_i) = \boldsymbol{0}$ which define the equation-by-equation OLS estimator (5.06). Thus, if the disturbances are uncorrelated across equations, the GLS and OLS estimators are numerically identical. The "seemingly" unrelated equations are indeed unrelated in this case.

In the second case, the matrix $\boldsymbol{\Sigma}$ is not diagonal, but all the regressor matrices $\boldsymbol{X}_1$ through $\boldsymbol{X}_g$ are the same, and are thus all equal to the matrix $\boldsymbol{X}$ that contains all the explanatory variables. Thus the estimating equations (5.16) become

$$\sum_{j=1}^{g} \sigma^{ij} \boldsymbol{X}^\top (\boldsymbol{y}_j - \boldsymbol{X}\hat{\boldsymbol{\beta}}_j^{\text{GLS}}) = \boldsymbol{0}, \quad i = 1, \dots, g.$$

If we multiply these equations by σ_{mi}, for any m between 1 and g, and sum over i from 1 to g, we obtain

$$\sum_{i=1}^{g} \sum_{j=1}^{g} \sigma_{mi} \sigma^{ij} \boldsymbol{X}^\top (\boldsymbol{y}_j - \boldsymbol{X}\hat{\boldsymbol{\beta}}_j^{\text{GLS}}) = \boldsymbol{0}. \tag{5.20}$$

Since the σ_{mi} are elements of $\boldsymbol{\Sigma}$ and the σ^{ij} are elements of its inverse, it follows that the sum $\sum_{i=1}^{g} \sigma_{mi}\sigma^{ij}$ is equal to δ_{mj}, the Kronecker delta, which is equal to 1 if $m = j$ and to 0 otherwise. Thus, for each $m = 1, \dots, g$, there is just one nonzero term on the left-hand side of (5.20) after the sum over i is performed, namely, that for which $j = m$. In consequence, equations (5.20) collapse to

$$\boldsymbol{X}^\top (\boldsymbol{y}_m - \boldsymbol{X}\hat{\boldsymbol{\beta}}_m^{\text{GLS}}) = \boldsymbol{0}.$$

Since these are the estimating equations that define the OLS estimator of the m^{th} equation, we conclude that $\hat{\boldsymbol{\beta}}_m^{\text{GLS}} = \hat{\boldsymbol{\beta}}_m^{\text{OLS}}$ for all m.

A GMM Interpretation

The above proof is straightforward enough, but it is not particularly intuitive. A much more intuitive way to see why the SUR estimator is identical to the OLS estimator in this special case is to interpret all of the estimators we have been studying as GMM estimators. This interpretation also provides a number of other insights and suggests a simple way of testing the overidentifying restrictions that are implicitly present whenever the SUR and OLS estimators are not identical.

Consider the gl theoretical moment conditions

$$\mathrm{E}\big(\boldsymbol{X}^\top (\boldsymbol{y}_i - \boldsymbol{X}_i \boldsymbol{\beta}_i)\big) = \boldsymbol{0}, \text{ for } i = 1, \dots, g, \tag{5.21}$$

which state that every regressor, whether or not it appears in a particular equation, must be uncorrelated with the disturbances for every equation. In the general case, these moment conditions are used to estimate k parameters, where $k = \sum_{i=1}^{g} k_i$. Since, in general, $k < gl$, we have more moment conditions than parameters, and we can choose a set of linear combinations of the conditions that minimizes the covariance matrix of the estimator. As is clear from the estimating equations (5.12), that is precisely what the SUR estimator (5.09) does. Although these estimating equations were derived from the

principles of GLS, they are evidently the empirical counterpart of the optimal moment conditions (2.18) given in Section 11.2 in the context of GMM for the case of a known covariance matrix and exogenous regressors. Therefore, the SUR estimator is, in general, an efficient GMM estimator.

In the special case in which every equation has the same regressors, the number of parameters is also equal to gl. Therefore, we have just as many parameters as moment conditions, and the empirical counterpart of (5.21) collapses to

$$\boldsymbol{X}^\top (\boldsymbol{y}_i - \boldsymbol{X}\boldsymbol{\beta}_i) = \boldsymbol{0}, \text{ for } i = 1, \dots, g,$$

which are just the moment conditions that define the equation-by-equation OLS estimator. Each of these g sets of equations can be solved for the l parameters in $\boldsymbol{\beta}_i$, and the unique solution is $\hat{\boldsymbol{\beta}}_i^{\text{OLS}}$.

We can now see that the two cases in which OLS is efficient arise for two quite different reasons. Clearly, no efficiency gain relative to OLS is possible unless there are more moment conditions than the OLS estimator utilizes. In other words, there can be no efficiency gain unless $gl > k$. In the second case, OLS is efficient because $gl = k$. In the first case, there are in general additional moment conditions, but, because there is no contemporaneous correlation, they are not informative about the model parameters.

We now derive the efficient GMM estimator from first principles and show that it is identical to the SUR estimator. We start from the set of gl sample moments

$$(\mathbf{I}_g \otimes \boldsymbol{X})^\top (\boldsymbol{\Sigma}^{-1} \otimes \mathbf{I}_n)(\boldsymbol{y}_\bullet - \boldsymbol{X}_\bullet \boldsymbol{\beta}_\bullet). \tag{5.22}$$

These provide the sample analog, for the linear SUR model, of the left-hand side of the theoretical moment conditions (2.18). The matrix in the middle is the inverse of the covariance matrix of the stacked vector of disturbances. Using the second result in (5.08), expression (5.22) can be rewritten as

$$(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{X}^\top)(\boldsymbol{y}_\bullet - \boldsymbol{X}_\bullet \boldsymbol{\beta}_\bullet). \tag{5.23}$$

The covariance matrix of this gl–vector is

$$(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{X}^\top)(\boldsymbol{\Sigma} \otimes \mathbf{I}_n)(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{X}) = \boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{X}^\top \boldsymbol{X}, \tag{5.24}$$

where we have made repeated use of the second result in (5.08). Combining (5.23) and (5.24) to construct the appropriate quadratic form, we find that the criterion function for fully efficient GMM estimation is

$$\begin{aligned}(\boldsymbol{y}_\bullet - \boldsymbol{X}_\bullet \boldsymbol{\beta}_\bullet)^\top (\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{X}) \big(\boldsymbol{\Sigma} \otimes (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\big) (\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{X}^\top)(\boldsymbol{y}_\bullet - \boldsymbol{X}_\bullet \boldsymbol{\beta}_\bullet) \\ = (\boldsymbol{y}_\bullet - \boldsymbol{X}_\bullet \boldsymbol{\beta}_\bullet)^\top (\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{P}_{\boldsymbol{X}})(\boldsymbol{y}_\bullet - \boldsymbol{X}_\bullet \boldsymbol{\beta}_\bullet),\end{aligned} \tag{5.25}$$

where, as usual, $\boldsymbol{P}_{\boldsymbol{X}}$ is the hat matrix, which projects orthogonally on to the subspace spanned by the columns of $\boldsymbol{X}$.

It is not hard to see that the vector $\hat{\boldsymbol{\beta}}_\bullet^{\text{GMM}}$ which minimizes expression (5.25) must be identical to $\hat{\boldsymbol{\beta}}_\bullet^{\text{GLS}}$. The first-order conditions may be written as

$$\sum_{j=1}^{g} \sigma^{ij} \boldsymbol{X}_i^\top \boldsymbol{P}_{\boldsymbol{X}} (\boldsymbol{y}_j - \boldsymbol{X}_j \hat{\boldsymbol{\beta}}_j^{\text{GMM}}) = \boldsymbol{0}. \tag{5.26}$$

But since each of the matrices $\boldsymbol{X}_i$ lies in $\mathcal{S}(\boldsymbol{X})$, it must be the case that $\boldsymbol{P}_{\boldsymbol{X}} \boldsymbol{X}_i = \boldsymbol{X}_i$, and so conditions (5.26) are actually identical to conditions (5.16), which define the GLS estimator.

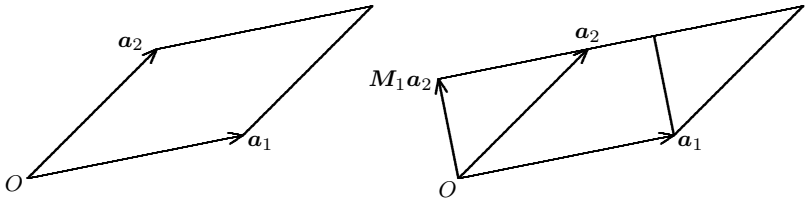
Since the GLS, and equally the feasible GLS, estimator can be interpreted as efficient GMM estimators, it is natural to test the overidentifying restrictions that these estimators depend on. These are the restrictions that certain columns of $\boldsymbol{X}$ do not appear in certain equations. The usual Hansen-Sargan statistic, which is just the minimized value of the criterion function (5.25), is asymptotically distributed as $\chi^2(gl - k)$ under the null hypothesis. As usual, the degrees of freedom for the test is equal to the number of moment conditions minus the number of estimated parameters. Investigators should always report the Hansen-Sargan statistic whenever they estimate a multivariate regression model using feasible GLS.

Since feasible GLS is really a feasible efficient GMM estimator, we might prefer to use the continuously updated GMM estimator, which was introduced in Section 11.2. Although the latter estimator is asymptotically equivalent to the one-step feasible GMM estimator, it may have better properties in finite samples. In this case, the continuously updated estimator is simply iterated feasible GLS, and it works as follows. After obtaining the feasible GLS estimator (5.18), we use it to recompute the residuals. These are then used in the formula (5.17) to obtain an updated estimate of the contemporaneous covariance matrix $\boldsymbol{\Sigma}$, which is then plugged back into the formula (5.18) to obtain an updated estimate of $\boldsymbol{\beta}_\bullet$. This procedure may be repeated as many times as desired. If the procedure converges, then, as we will see shortly, the estimator that results is equal to the ML estimator computed under the assumption of normal disturbances.

Determinants of Square Matrices

The most popular alternative to feasible GLS estimation is maximum likelihood estimation under the assumption that the disturbances are normally distributed. We will discuss this estimation method in the next subsection. However, in order to develop the theory of ML estimation for systems of equations, we must first say a few words about **determinants**.

A $p \times p$ square matrix $\boldsymbol{A}$ defines a mapping from Euclidean p–dimensional space, E^p, into itself, by which a vector $\boldsymbol{x} \in E^p$ is mapped into the p–vector $\boldsymbol{A}\boldsymbol{x}$. The determinant of $\boldsymbol{A}$ is a scalar quantity which measures the extent to which this mapping expands or contracts p–dimensional volumes in E^p.



(a) The parallelogram defined by $\boldsymbol{a}_1$ and $\boldsymbol{a}_2$

(b) Rectangle of equal area formed with $\boldsymbol{a}_1$ and $\boldsymbol{M}_1\boldsymbol{a}_2$

Figure 5.1 Determinants in two dimensions

Consider a simple example in E^2. Volume in 2–dimensional space is just area. The simplest area to consider is the unit square, which can be defined as the parallelogram defined by the two unit basis vectors $\boldsymbol{e}_1$ and $\boldsymbol{e}_2$, where $\boldsymbol{e}_i$ has only one nonzero component, in position i. The area of the unit square is, by definition, 1. The image of the unit square under the mapping defined by a 2×2 matrix $\boldsymbol{A}$ is the parallelogram defined by the two columns of the matrix

$$\boldsymbol{A}[\,\boldsymbol{e}_1 \quad \boldsymbol{e}_2\,] = \boldsymbol{A}\mathbf{I} = \boldsymbol{A} \equiv [\,\boldsymbol{a}_1 \quad \boldsymbol{a}_2\,],$$

where $\boldsymbol{a}_1$ and $\boldsymbol{a}_2$ are the two columns of $\boldsymbol{A}$. The area of a parallelogram in Euclidean geometry is given by base times height, where the length of either one of the two defining vectors can be taken as the base, and the height is then the perpendicular distance between the two parallel sides that correspond to this choice of base. This is illustrated in Figure 5.1.

If we choose $\boldsymbol{a}_1$ as the base, then, as we can see from the figure, the height is the length of the vector $\boldsymbol{M}_1\boldsymbol{a}_2$, where $\boldsymbol{M}_1$ is the orthogonal projection on to the orthogonal complement of $\boldsymbol{a}_1$. Thus the area of the parallelogram defined by $\boldsymbol{a}_1$ and $\boldsymbol{a}_2$ is $\|\boldsymbol{a}_1\|\|\boldsymbol{M}_1\boldsymbol{a}_2\|$. By use of Pythagoras' Theorem and a little algebra (see Exercise 5.6), it can be seen that

$$\|\boldsymbol{a}_1\|\|\boldsymbol{M}_1\boldsymbol{a}_2\| = |a_{11}a_{22} - a_{12}a_{21}|, \tag{5.27}$$

where a_{ij} is the ij^{th} element of $\boldsymbol{A}$. This quantity is the absolute value of the determinant of $\boldsymbol{A}$, which we write as $|\det \boldsymbol{A}|$. The determinant itself, which is defined as $a_{11}a_{22} - a_{12}a_{21}$, can be of either sign. Its signed value can be written as "$\det \boldsymbol{A}$", but it is more commonly, and perhaps somewhat confusingly, written as $|\boldsymbol{A}|$.

Algebraic expressions for determinants of square matrices of dimension higher than 2 can be found easily enough, but we will have no need of them. We will, however, need to make use of some of the properties of determinants. The principal properties that will matter to us are as follows:

- The determinant of the transpose of a matrix is equal to the determinant of the matrix itself. That is, $|\boldsymbol{A}^\top| = |\boldsymbol{A}|$.

- The determinant of a triangular matrix is the product of its diagonal elements.
- Since a diagonal matrix can be regarded as a special triangular matrix, its determinant is also the product of its diagonal elements.
- Since an identity matrix is a diagonal matrix with all diagonal elements equal to unity, the determinant of an identity matrix is 1.
- If a matrix can be partitioned so as to be block-diagonal, then its determinant is the product of the determinants of the diagonal blocks.
- Interchanging two rows, or two columns, of a matrix leaves the absolute value of the determinant unchanged but changes its sign.
- The determinant of the product of two square matrices of the same dimensions is the product of their determinants, from which it follows that the determinant of $\boldsymbol{A}^{-1}$ is the reciprocal of the determinant of $\boldsymbol{A}$.
- If a matrix can be inverted, its determinant must be nonzero. Conversely, if a matrix is singular, its determinant is 0.
- The derivative of $\log|\boldsymbol{A}|$ with respect to the ij^{th} element a_{ij} of $\boldsymbol{A}$ is the ji^{th} element of $\boldsymbol{A}^{-1}$.

Maximum Likelihood Estimation

If we assume that the disturbances of an SUR system are normally distributed, the system can be estimated by maximum likelihood. The model to be estimated can be written as

$$\boldsymbol{y}_\bullet = \boldsymbol{X}_\bullet\boldsymbol{\beta}_\bullet + \boldsymbol{u}_\bullet, \quad \boldsymbol{u}_\bullet \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{\Sigma}\otimes\mathbf{I}_n). \tag{5.28}$$

The loglikelihood function for this model is the logarithm of the joint density of the components of the vector $\boldsymbol{y}_\bullet$. In order to derive that density, we must start with the density of the vector $\boldsymbol{u}_\bullet$.

Up to this point, we have not actually written down the density of a random vector that follows the multivariate normal distribution. We will do so in a moment. But first, we state a more fundamental result, which extends the result (3.95) that was proved in Section 10.8 for univariate densities of transformations of variables to the case of multivariate densities.

Let $\boldsymbol{z}$ be a random m–vector with known density $f_{\boldsymbol{z}}(\boldsymbol{z})$, and let $\boldsymbol{x}$ be another random m–vector such that $\boldsymbol{z} = \boldsymbol{h}(\boldsymbol{x})$, where the deterministic function $\boldsymbol{h}(\cdot)$ is a one to one mapping of the support of the random vector $\boldsymbol{x}$, which is a subset of $\mathbb{R}^m$, into the support of $\boldsymbol{z}$. Then the multivariate analog of the result (3.95) is

$$f_{\boldsymbol{x}}(\boldsymbol{x}) = f_{\boldsymbol{z}}\big(\boldsymbol{h}(\boldsymbol{x})\big)\big|\det \boldsymbol{J}(\boldsymbol{x})\big|, \tag{5.29}$$

where $\boldsymbol{J}(\boldsymbol{x}) \equiv \partial\boldsymbol{h}(\boldsymbol{x})/\partial\boldsymbol{x}$ is the Jacobian matrix of the transformation, that is, the $m\times m$ matrix containing the derivatives of the components of $\boldsymbol{h}(\boldsymbol{x})$ with

respect to those of $\boldsymbol{x}$, and we have written $|\det \boldsymbol{J}(\boldsymbol{x})|$ to signify the absolute value of the determinant.

Using (5.29), it is not difficult to show that, if the $m\times 1$ vector $\boldsymbol{z}$ follows the multivariate normal distribution with mean vector $\boldsymbol{0}$ and covariance matrix $\boldsymbol{\Omega}$, then its density is equal to

$$(2\pi)^{-m/2}|\boldsymbol{\Omega}|^{-1/2}\exp\big(-\tfrac{1}{2}\boldsymbol{z}^\top\boldsymbol{\Omega}^{-1}\boldsymbol{z}\big). \tag{5.30}$$

Readers are asked to prove a slightly more general result in Exercise 5.8.

For the system (5.28), the function $\boldsymbol{h}(\cdot)$ that gives $\boldsymbol{u}_\bullet$ as a function of $\boldsymbol{y}_\bullet$ is the right-hand side of the equation

$$\boldsymbol{u}_\bullet = \boldsymbol{y}_\bullet - \boldsymbol{X}_\bullet\boldsymbol{\beta}_\bullet. \tag{5.31}$$

Thus we see that, if there are no lagged dependent variables in the matrix $\boldsymbol{X}_\bullet$, then the Jacobian of the transformation is just the identity matrix, of which the determinant is 1.

The Jacobian is, in general, much more complicated if there are lagged dependent variables, because the elements of $\boldsymbol{X}_\bullet$ depend on the elements of $\boldsymbol{y}_\bullet$. However, as readers are invited to check in Exercise 5.10, even though the Jacobian is not equal to the identity matrix in such a case, its determinant is still 1. Therefore, we can ignore the Jacobian when we compute the density of $\boldsymbol{y}_\bullet$. When we substitute (5.31) into (5.30), as the result (5.29) tells us to do, we find that the density of $\boldsymbol{y}_\bullet$ is $(2\pi)^{-gn/2}$ times

$$|\boldsymbol{\Sigma}\otimes\mathbf{I}_n|^{-1/2}\exp\big(-\tfrac{1}{2}(\boldsymbol{y}_\bullet - \boldsymbol{X}_\bullet\boldsymbol{\beta}_\bullet)^\top(\boldsymbol{\Sigma}^{-1}\otimes\mathbf{I}_n)(\boldsymbol{y}_\bullet - \boldsymbol{X}_\bullet\boldsymbol{\beta}_\bullet)\big). \tag{5.32}$$

Jointly maximizing the logarithm of this function with respect to $\boldsymbol{\beta}_\bullet$ and the elements of $\boldsymbol{\Sigma}$ gives the ML estimator of the SUR system.

The argument of the exponential function in (5.32) plays the same role for a multivariate linear regression model as the sum of squares term plays in the loglikelihood function (3.10) for a linear regression model with IID normal disturbances. In fact, it is clear from (5.32) that maximizing the loglikelihood with respect to $\boldsymbol{\beta}_\bullet$ for a given $\boldsymbol{\Sigma}$ is equivalent to minimizing the function

$$(\boldsymbol{y}_\bullet - \boldsymbol{X}_\bullet\boldsymbol{\beta}_\bullet)^\top(\boldsymbol{\Sigma}^{-1}\otimes\mathbf{I}_n)(\boldsymbol{y}_\bullet - \boldsymbol{X}_\bullet\boldsymbol{\beta}_\bullet)$$

with respect to $\boldsymbol{\beta}_\bullet$. This expression is just the criterion function (5.11) that is minimized in order to obtain the GLS estimator (5.09). Therefore, the ML estimator $\hat{\boldsymbol{\beta}}_\bullet^{\text{ML}}$ must have exactly the same form as (5.09), with the matrix $\boldsymbol{\Sigma}$ replaced by its ML estimator $\hat{\boldsymbol{\Sigma}}_{\text{ML}}$, which we will derive shortly.

It follows from (5.32) that the loglikelihood function $\ell(\boldsymbol{\Sigma}, \boldsymbol{\beta}_\bullet)$ for the model (5.28) can be written as

$$-\tfrac{gn}{2}\log 2\pi - \tfrac{1}{2}\log|\boldsymbol{\Sigma}\otimes\mathbf{I}_n| - \tfrac{1}{2}(\boldsymbol{y}_\bullet - \boldsymbol{X}_\bullet\boldsymbol{\beta}_\bullet)^\top(\boldsymbol{\Sigma}^{-1}\otimes\mathbf{I}_n)(\boldsymbol{y}_\bullet - \boldsymbol{X}_\bullet\boldsymbol{\beta}_\bullet).$$

The properties of determinants set out in the previous subsection can be used to show that the determinant of $\boldsymbol{\Sigma} \otimes \mathbf{I}_n$ is $|\boldsymbol{\Sigma}|^n$; see Exercise 5.11. Thus this loglikelihood function simplifies to

$$-\frac{gn}{2}\log 2\pi - \frac{n}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}(\boldsymbol{y}_\bullet - \boldsymbol{X}_\bullet\boldsymbol{\beta}_\bullet)^\top(\boldsymbol{\Sigma}^{-1}\otimes\mathbf{I}_n)(\boldsymbol{y}_\bullet - \boldsymbol{X}_\bullet\boldsymbol{\beta}_\bullet). \tag{5.33}$$

We have already seen how to maximize the function (5.33) with respect to $\boldsymbol{\beta}_\bullet$ conditional on $\boldsymbol{\Sigma}$. Now we want to maximize it with respect to $\boldsymbol{\Sigma}$.

Maximizing $\ell(\boldsymbol{\Sigma}, \boldsymbol{\beta}_\bullet)$ with respect to $\boldsymbol{\Sigma}$ is of course equivalent to maximizing it with respect to $\boldsymbol{\Sigma}^{-1}$, and it turns out to be technically simpler to differentiate with respect to the elements of the latter matrix. Note first that, since the determinant of the inverse of a matrix is the reciprocal of the determinant of the matrix itself, we have $-\log|\boldsymbol{\Sigma}| = \log|\boldsymbol{\Sigma}^{-1}|$, so that we can readily express all of (5.33) in terms of $\boldsymbol{\Sigma}^{-1}$ rather than $\boldsymbol{\Sigma}$.

It is obvious that the derivative of any $p \times q$ matrix $\boldsymbol{A}$ with respect to its ij^{th} element is the $p \times q$ matrix $\boldsymbol{E}_{ij}$, all the elements of which are 0, except for the ij^{th}, which is 1. Recall that we write the ij^{th} element of $\boldsymbol{\Sigma}^{-1}$ as σ^{ij}. We therefore find that

$$\frac{\partial\boldsymbol{\Sigma}^{-1}}{\partial\sigma^{ij}} = \boldsymbol{E}_{ij}, \tag{5.34}$$

where in this case $\boldsymbol{E}_{ij}$ is a $g \times g$ matrix. We remarked in our earlier discussion of determinants that the derivative of $\log|\boldsymbol{A}|$ with respect to a_{ij} is the ji^{th} element of $\boldsymbol{A}^{-1}$. Armed with this result and (5.34), we see that the derivative of the loglikelihood function $\ell(\boldsymbol{\Sigma}, \boldsymbol{\beta}_\bullet)$ with respect to the element σ^{ij} is

$$\frac{\partial\ell(\boldsymbol{\Sigma}, \boldsymbol{\beta}_\bullet)}{\partial\sigma^{ij}} = \frac{n}{2}\sigma_{ij} - \frac{1}{2}(\boldsymbol{y}_\bullet - \boldsymbol{X}_\bullet\boldsymbol{\beta}_\bullet)^\top(\boldsymbol{E}_{ij}\otimes\mathbf{I}_n)(\boldsymbol{y}_\bullet - \boldsymbol{X}_\bullet\boldsymbol{\beta}_\bullet). \tag{5.35}$$

The Kronecker product $\boldsymbol{E}_{ij} \otimes \mathbf{I}_n$ has only one nonzero block containing $\mathbf{I}_n$. It is easy to conclude from this that

$$(\boldsymbol{y}_\bullet - \boldsymbol{X}_\bullet\boldsymbol{\beta}_\bullet)^\top(\boldsymbol{E}_{ij}\otimes\mathbf{I}_n)(\boldsymbol{y}_\bullet - \boldsymbol{X}_\bullet\boldsymbol{\beta}_\bullet) = (\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}_i)^\top(\boldsymbol{y}_j - \boldsymbol{X}_j\boldsymbol{\beta}_j).$$

By equating the partial derivative (5.35) to zero, we find that the ML estimator $\hat{\sigma}_{ij}^{\text{ML}}$ is

$$\hat{\sigma}_{ij}^{\text{ML}} = \frac{1}{n}(\boldsymbol{y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\beta}}_i^{\text{ML}})^\top(\boldsymbol{y}_j - \boldsymbol{X}_j\hat{\boldsymbol{\beta}}_j^{\text{ML}}).$$

If we define the $n \times g$ matrix $\boldsymbol{U}(\boldsymbol{\beta}_\bullet)$ to have i^{th} column $\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}_i$, then we can conveniently write the ML estimator of $\boldsymbol{\Sigma}$ as follows:

$$\hat{\boldsymbol{\Sigma}}_{\text{ML}} = \frac{1}{n}\boldsymbol{U}^\top(\hat{\boldsymbol{\beta}}_\bullet^{\text{ML}})\boldsymbol{U}(\hat{\boldsymbol{\beta}}_\bullet^{\text{ML}}). \tag{5.36}$$

This looks like equation (5.17), which defines the covariance matrix used in feasible GLS estimation. Equations (5.36) and (5.17) have exactly the same

form, but they are based on different matrices of residuals. Equation (5.36) and equation (5.09) evaluated at $\hat{\boldsymbol{\Sigma}}^{\text{ML}}$, that is

$$\hat{\boldsymbol{\beta}}_\bullet^{\text{ML}} = \big(\boldsymbol{X}_\bullet^\top(\hat{\boldsymbol{\Sigma}}_{\text{ML}}^{-1}\otimes\mathbf{I}_n)\boldsymbol{X}_\bullet\big)^{-1}\boldsymbol{X}_\bullet^\top(\hat{\boldsymbol{\Sigma}}_{\text{ML}}^{-1}\otimes\mathbf{I}_n)\boldsymbol{y}_\bullet, \tag{5.37}$$

together define the ML estimator for the model (5.28).

Equations (5.36) and (5.37) are exactly the ones that are used by the continuously updated GMM estimator to update the estimates of $\boldsymbol{\Sigma}$ and $\boldsymbol{\beta}_\bullet$, respectively. It follows that, if the continuous updating procedure converges, it converges to the ML estimator. Consequently, we can estimate the covariance matrix of $\hat{\boldsymbol{\beta}}_\bullet^{\text{ML}}$ in the same way as for the GLS or GMM estimator, by the formula

$$\widehat{\text{Var}}(\hat{\boldsymbol{\beta}}_\bullet^{\text{ML}}) = \big(\boldsymbol{X}_\bullet^\top(\hat{\boldsymbol{\Sigma}}_{\text{ML}}^{-1}\otimes\mathbf{I}_n)\boldsymbol{X}_\bullet\big)^{-1}. \tag{5.38}$$

It is also possible to estimate the covariance matrix of the estimated contemporaneous covariance matrix, $\hat{\boldsymbol{\Sigma}}^{\text{ML}}$, although this is rarely done. If the elements of $\boldsymbol{\Sigma}$ are stacked in a vector of dimension g^2, a suitable estimator is

$$\widehat{\text{Var}}\big(\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}}_\bullet^{\text{ML}})\big) = \frac{2}{n}\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}}_\bullet^{\text{ML}})\otimes\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}}_\bullet^{\text{ML}}). \tag{5.39}$$

Notice that the estimated variance of any diagonal element of $\boldsymbol{\Sigma}$ is just twice the square of that element, divided by n. This is precisely what is obtained for the univariate case in Exercise 10.10. As with that result, the asymptotic validity of (5.39) depends critically on the assumption that the disturbances are multivariate normal.

As we saw in Chapter 9, ML estimators are consistent and asymptotically efficient *if* the underlying model is correctly specified. It may therefore seem that the asymptotic efficiency of the ML estimator (5.37) depends critically on the multivariate normality assumption. However, the fact that the ML estimator is identical to the continuously updated efficient GMM estimator means that it is in fact efficient in the same sense as the latter. When the disturbances are not normal, the estimator is more properly termed a QMLE (see Section 9.4). As such, it is consistent, but not necessarily efficient, under assumptions about the disturbances that are no stronger than those needed for feasible GLS to be consistent. Moreover, if the stronger assumptions made in (5.02) hold, even without normality, then the estimator (5.38) of $\text{Var}(\hat{\boldsymbol{\beta}}_\bullet^{\text{ML}})$ is asymptotically valid. If the disturbances are not normal, it would be necessary to have information about their actual distribution in order to derive an estimator with a smaller asymptotic variance than (5.37).

It is of considerable theoretical interest to concentrate the loglikelihood function (5.33) with respect to $\boldsymbol{\Sigma}$. In order to do so, we use the first-order conditions that led to (5.36) to define $\boldsymbol{\Sigma}(\boldsymbol{\beta}_\bullet)$ as the matrix that maximizes (5.33) for given $\boldsymbol{\beta}_\bullet$. We find that

$$\boldsymbol{\Sigma}(\boldsymbol{\beta}_\bullet) \equiv \frac{1}{n}\boldsymbol{U}^\top(\boldsymbol{\beta}_\bullet)\boldsymbol{U}(\boldsymbol{\beta}_\bullet).$$

A calculation of a type that should now be familiar then shows that

$$\begin{aligned}(\boldsymbol{y}_\bullet - \boldsymbol{X}_\bullet\boldsymbol{\beta}_\bullet)^\top(\boldsymbol{\Sigma}^{-1}\otimes\mathbf{I}_n)(\boldsymbol{y}_\bullet - \boldsymbol{X}_\bullet\boldsymbol{\beta}_\bullet)\\ = \sum_{i=1}^{g}\sum_{j=1}^{g}\sigma^{ij}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}_i)^\top(\boldsymbol{y}_j - \boldsymbol{X}_j\boldsymbol{\beta}_j).\end{aligned} \tag{5.40}$$

When $\sigma^{ij} = \sigma^{ij}(\boldsymbol{\beta}_\bullet)$, which denotes the ij^{th} element of $\boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}_\bullet)$, the right-hand side of equation (5.40) is

$$\begin{aligned}\sum_{i=1}^{g}\sum_{j=1}^{g}\sigma^{ij}(\boldsymbol{\beta}_\bullet)\big(\boldsymbol{U}^\top(\boldsymbol{\beta}_\bullet)\boldsymbol{U}(\boldsymbol{\beta}_\bullet)\big)_{ij} &= n\sum_{i=1}^{g}\sum_{j=1}^{g}\sigma^{ij}(\boldsymbol{\beta}_\bullet)\sigma_{ij}(\boldsymbol{\beta}_\bullet)\\ &= n\sum_{i=1}^{g}(\mathbf{I}_g)_{ii} = n\,\mathrm{Tr}(\mathbf{I}_g) = gn,\end{aligned}$$

where we have made use of the trace operator, which sums the diagonal elements of a square matrix; see Section 2.6. By substituting this result into expression (5.33), we see that the concentrated loglikelihood function can be written as

$$-\frac{gn}{2}(\log 2\pi + 1) - \frac{n}{2}\log\left|\frac{1}{n}\boldsymbol{U}^\top(\boldsymbol{\beta}_\bullet)\boldsymbol{U}(\boldsymbol{\beta}_\bullet)\right|. \tag{5.41}$$

This expression depends on the data only through the determinant of the covariance matrix of the residuals. It is the multivariate generalization of the concentrated loglikelihood function (3.11) that we obtained in Section 9.2 in the univariate case. We saw there that the concentrated function depends on the data only through the sum of squared residuals.

It is quite possible to minimize the determinant in (5.41) with respect to $\boldsymbol{\beta}_\bullet$ directly. It may or may not be numerically simpler to do so than to solve the coupled equations (5.37) and (5.36).

We saw in Section 3.6 that the squared residuals of a univariate regression model tend to be smaller than the squared disturbances, because least-squares estimates make the sum of squared residuals as small as possible. For a similar reason, the residuals from ML estimation of a multivariate regression model tend to be too small and too highly correlated with each other. We observe both effects, because the determinant of $\boldsymbol{\Sigma}$ can be made smaller either by reducing the sums of squared residuals associated with the individual equations or by increasing the correlations among the residuals. This is likely to be most noticeable when g and/or the k_i are large relative to n.

Although feasible GLS and ML with the assumption of normally distributed disturbances are by far the most commonly used methods of estimating linear SUR systems, they are by no means the only ones that have been proposed. For fuller treatments, a classic reference on linear SUR systems is Srivastava and Giles (1987), and a useful recent survey paper is Fiebig (2001).

5.3 Systems of Nonlinear Regressions

Many multivariate regression models are nonlinear. For example, economists routinely estimate **demand systems**, in which the shares of consumer expenditure on various classes of goods and services are explained by incomes, prices, and perhaps other explanatory variables. Demand systems may be estimated using aggregate time-series data, cross-section data, or mixed time-series/cross-section (panel) data on households.[1]

The **multivariate nonlinear regression model** is a system of nonlinear regressions which can be written as

$$y_{ti} = x_{ti}(\boldsymbol{\beta}) + u_{ti}, \quad t = 1,\dots,n, \;\; i = 1,\dots,g. \tag{5.42}$$

Here y_{ti} is the t^{th} observation on the i^{th} dependent variable, $x_{ti}(\boldsymbol{\beta})$ is the t^{th} observation on the regression function which determines the conditional mean of that dependent variable, $\boldsymbol{\beta}$ is a k–vector of parameters to be estimated, and u_{ti} is a disturbance which is assumed to have expectation zero conditional on all the explanatory variables that implicitly appear in all the regression functions $x_{tj}(\boldsymbol{\beta})$, $j = 1,\dots,g$. In the demand system case, y_{ti} would be the share of expenditure on commodity i for observation t, and the explanatory variables would include prices and income. We assume that the disturbances in (5.42), like those in (5.01), satisfy assumption (5.02). They are serially uncorrelated, homoskedastic within each equation, and have contemporaneous covariance matrix $\boldsymbol{\Sigma}$ with typical element σ_{ij}.

The equations of the system (5.42) can also be written using essentially the same notation as we used for univariate nonlinear regression models in Chapter 7. If, for each $i = 1,\dots,g$, the n–vectors $\boldsymbol{y}_i$, $\boldsymbol{x}_i(\boldsymbol{\beta})$, and $\boldsymbol{u}_i$ are defined to have typical elements y_{ti}, u_{ti}, and $x_{ti}(\boldsymbol{\beta})$, respectively, then the entire system can be expressed as

$$\boldsymbol{y}_i = \boldsymbol{x}_i(\boldsymbol{\beta}) + \boldsymbol{u}_i, \quad \mathrm{E}(\boldsymbol{u}_i\boldsymbol{u}_j^\top) = \sigma_{ij}\mathbf{I}_n, \quad i,j = 1,\dots,g. \tag{5.43}$$

We have written (5.42) and (5.43) in such a way that there is just a single vector of parameters, denoted $\boldsymbol{\beta}$. Every individual parameter may, at least in principle, appear in every equation, although that is rare in practice. In the demand systems case, however, some but not all of the parameters typically do appear in every equation of the system. Thus systems of nonlinear regressions very often involve **cross-equation restrictions**.

Multivariate nonlinear regression models can be estimated in essentially the same way as the multivariate linear regression model (5.01). Feasible GLS

[1] The literature on demand systems is vast; see, among many others, Christensen, Jorgenson, and Lau (1975), Barten (1977), Deaton and Muellbauer (1980), Pollak and Wales (1981, 1987), Browning and Meghir (1991), Lewbel (1991), and Blundell, Browning, and Meghir (1994).

and maximum likelihood are both commonly used. The results we obtained in the previous section still apply, provided they are modified to allow for the nonlinearity of the regression functions and for cross-equation restrictions. Our discussion will therefore be quite brief.

Estimation

Nonlinear GLS estimates can be obtained either by minimizing the criterion function

$$\big(\boldsymbol{y} - \boldsymbol{x}(\boldsymbol{\beta})\big)^\top \boldsymbol{\Omega}^{-1} \big(\boldsymbol{y} - \boldsymbol{x}(\boldsymbol{\beta})\big), \tag{5.44}$$

or, equivalently, by solving the set of first-order conditions

$$\boldsymbol{X}^\top(\boldsymbol{\beta}) \boldsymbol{\Omega}^{-1} \big(\boldsymbol{y} - \boldsymbol{x}(\boldsymbol{\beta})\big) = \boldsymbol{0}. \tag{5.45}$$

For the multivariate nonlinear regression model (5.42), the criterion function can be written so that it looks very much like expression (5.11). Let $\boldsymbol{y}_\bullet$ once again denote a gn–vector of the $\boldsymbol{y}_i$ stacked vertically, and let $\boldsymbol{x}_\bullet(\boldsymbol{\beta})$ denote a gn–vector of the $\boldsymbol{x}_i(\boldsymbol{\beta})$ stacked in the same way. The criterion function (5.44) then becomes

$$\big(\boldsymbol{y}_\bullet - \boldsymbol{x}_\bullet(\boldsymbol{\beta})\big)^\top (\boldsymbol{\Sigma}^{-1} \otimes \mathbf{I}_n) \big(\boldsymbol{y}_\bullet - \boldsymbol{x}_\bullet(\boldsymbol{\beta})\big). \tag{5.46}$$

Minimizing (5.46) with respect to $\boldsymbol{\beta}$ yields nonlinear GLS estimates which, by the results of Section 8.2, are consistent and asymptotically efficient under standard regularity conditions.

The first-order conditions for the minimization of (5.46) give rise to the following moment conditions, which have a very similar form to the moment conditions (5.12) that we found for the linear case:

$$\boldsymbol{X}_\bullet^\top(\boldsymbol{\beta}) (\boldsymbol{\Sigma}^{-1} \otimes \mathbf{I}_n) \big(\boldsymbol{y}_\bullet - \boldsymbol{x}_\bullet(\boldsymbol{\beta})\big) = \boldsymbol{0}. \tag{5.47}$$

Here, the $gn \times k$ matrix $\boldsymbol{X}_\bullet(\boldsymbol{\beta})$ is a matrix of partial derivatives of the $x_{ti}(\boldsymbol{\beta})$. If the $n \times k$ matrices $\boldsymbol{X}_i(\boldsymbol{\beta})$ are defined, just as in the univariate case, so that the tj^{th} element of $\boldsymbol{X}_i(\boldsymbol{\beta})$ is $\partial x_{ti}(\boldsymbol{\beta})/\partial \beta_j$, for $t = 1, \dots, n$, $j = 1, \dots, k$, then $\boldsymbol{X}_\bullet(\boldsymbol{\beta})$ is the matrix formed by stacking the $\boldsymbol{X}_i(\boldsymbol{\beta})$ vertically. Except in the special case in which each parameter appears in only one equation of the system, $\boldsymbol{X}_\bullet(\boldsymbol{\beta})$ does not have the block-diagonal structure of $\boldsymbol{X}_\bullet$ in (5.05).

Despite this fact, it is not hard to show that the moment conditions (5.47) can be expressed in a compact form like (5.16), but with a double sum. As readers are asked to check in Exercise 5.12, we obtain estimating equations of the form

$$\sum_{i=1}^{g} \sum_{j=1}^{g} \sigma^{ij} \boldsymbol{X}_i^\top(\boldsymbol{\beta}) \big(\boldsymbol{y}_j - \boldsymbol{x}_j(\boldsymbol{\beta})\big) = \boldsymbol{0}. \tag{5.48}$$

The vector $\hat{\boldsymbol{\beta}}^{\text{GLS}}$ that solves these equations is the nonlinear GLS estimator.

Adapting expression (F9.05) to the model (5.43) gives the standard estimate of the covariance matrix of the nonlinear GLS estimator, namely,

$$\widehat{\text{Var}}(\hat{\boldsymbol{\beta}}^{\text{GLS}}) = \big(\boldsymbol{X}_\bullet^\top(\hat{\boldsymbol{\beta}}^{\text{GLS}}) (\boldsymbol{\Sigma}^{-1} \otimes \mathbf{I}_n) \boldsymbol{X}_\bullet(\hat{\boldsymbol{\beta}}^{\text{GLS}})\big)^{-1}. \tag{5.49}$$

This can also be written (see Exercise 5.12 again) as

$$\widehat{\text{Var}}(\hat{\boldsymbol{\beta}}^{\text{GLS}}) = \left(\sum_{i=1}^{g} \sum_{j=1}^{g} \sigma^{ij} \boldsymbol{X}_i^\top(\hat{\boldsymbol{\beta}}^{\text{GLS}}) \boldsymbol{X}_j(\hat{\boldsymbol{\beta}}^{\text{GLS}}) \right)^{-1}. \tag{5.50}$$

Feasible GLS estimation works in essentially the same way for nonlinear multivariate regression models as it does for linear ones. The individual equations of the system are first estimated separately by either ordinary or nonlinear least squares, as appropriate. The residuals are then grouped into an $n \times g$ matrix $\hat{\boldsymbol{U}}$, and equation (5.17) is used to obtain the estimate $\hat{\boldsymbol{\Sigma}}$. We can then replace $\boldsymbol{\Sigma}$ by $\hat{\boldsymbol{\Sigma}}$ in the GLS criterion function (5.46) or in the moment conditions (5.47) to obtain the feasible GLS estimator $\hat{\boldsymbol{\beta}}^{\text{F}}$. We may also use a continuously updated estimator, alternately updating our estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$. If this iterated feasible GLS procedure converges, then we have obtained ML estimates, although there may well be more computationally attractive ways to do so.

Maximum likelihood estimation under the assumption of normality is very popular for multivariate nonlinear regression models. For the system (5.42), the loglikelihood function can be written as

$$-\frac{gn}{2} \log 2\pi - \frac{n}{2} \log |\boldsymbol{\Sigma}| - \frac{1}{2} \big(\boldsymbol{y}_\bullet - \boldsymbol{x}_\bullet(\boldsymbol{\beta})\big)^\top (\boldsymbol{\Sigma}^{-1} \otimes \mathbf{I}_n) \big(\boldsymbol{y}_\bullet - \boldsymbol{x}_\bullet(\boldsymbol{\beta})\big). \tag{5.51}$$

This is the analog of the loglikelihood function (5.33) for the linear case. Maximizing (5.51) with respect to $\boldsymbol{\beta}$ for given $\boldsymbol{\Sigma}$ is equivalent to minimizing the criterion function (5.46) with respect to $\boldsymbol{\beta}$, and so the first-order conditions are equations (5.47). Maximizing (5.51) with respect to $\boldsymbol{\Sigma}$ for given $\boldsymbol{\beta}$ leads to first-order conditions that can be written as

$$\boldsymbol{\Sigma}(\boldsymbol{\beta}) = \frac{1}{n} \boldsymbol{U}^\top(\boldsymbol{\beta}) \boldsymbol{U}(\boldsymbol{\beta}),$$

in exactly the same way as the maximization of (5.33) with respect to $\boldsymbol{\Sigma}$ led to equation (5.36). Here the $n \times g$ matrix $\boldsymbol{U}(\boldsymbol{\beta})$ is defined so that its i^{th} column is $\boldsymbol{y}_i - \boldsymbol{x}_i(\boldsymbol{\beta})$.

Thus the estimating equations that define the ML estimator are

$$\begin{aligned} &\boldsymbol{X}_\bullet^\top(\hat{\boldsymbol{\beta}}^{\text{ML}}) (\hat{\boldsymbol{\Sigma}}_{\text{ML}}^{-1} \otimes \mathbf{I}_n) \big(\boldsymbol{y}_\bullet - \boldsymbol{x}_\bullet(\hat{\boldsymbol{\beta}}^{\text{ML}})\big) = \boldsymbol{0}, \text{ and} \\ &\hat{\boldsymbol{\Sigma}}_{\text{ML}} = \frac{1}{n} \boldsymbol{U}^\top(\hat{\boldsymbol{\beta}}^{\text{ML}}) \boldsymbol{U}(\hat{\boldsymbol{\beta}}^{\text{ML}}). \end{aligned} \tag{5.52}$$

As in the linear case, these are also the estimating equations for the continuously updated GMM estimator. The covariance matrix of $\hat{\boldsymbol{\beta}}^{\text{ML}}$ is, of course, given by either of the formulas (5.49) or (5.50) evaluated at $\hat{\boldsymbol{\beta}}^{\text{ML}}$ and $\hat{\boldsymbol{\Sigma}}_{\text{ML}}$. The loglikelihood function concentrated with respect to $\boldsymbol{\Sigma}$ can be written,

just like expression (5.41), as

$$-\tfrac{gn}{2}(\log 2\pi + 1) - \tfrac{n}{2}\log\Big|\tfrac{1}{n}\boldsymbol{U}^\top(\boldsymbol{\beta})\boldsymbol{U}(\boldsymbol{\beta})\Big|. \tag{5.53}$$

As in the linear case, it may or may not be numerically easier to maximize the concentrated function directly than to solve the estimating equations (5.52).

The Gauss-Newton Regression

The Gauss-Newton regression can be very useful in the context of multivariate regression models, both linear and nonlinear. The starting point for setting up the GNR for both types of multivariate model is the GNR for the standard univariate model $\boldsymbol{y} = \boldsymbol{x}(\boldsymbol{\beta}) + \boldsymbol{u}$, with $\mathrm{Var}(\boldsymbol{u}) = \boldsymbol{\Omega}$. This GNR takes the form

$$\boldsymbol{\Psi}^\top\big(\boldsymbol{y} - \boldsymbol{x}(\boldsymbol{\beta})\big) = \boldsymbol{\Psi}^\top\boldsymbol{X}(\boldsymbol{\beta})\boldsymbol{b} + \text{residuals}, \tag{5.54}$$

where, as usual, $\boldsymbol{X}(\boldsymbol{\beta})$ is the matrix of partial derivatives of the regression functions, and $\boldsymbol{\Psi}$ is such that $\boldsymbol{\Psi}\boldsymbol{\Psi}^\top = \boldsymbol{\Omega}^{-1}$.

Expressed as a univariate regression, the multivariate model (5.43) becomes

$$\boldsymbol{y}_\bullet = \boldsymbol{x}_\bullet(\boldsymbol{\beta}) + \boldsymbol{u}_\bullet, \quad \mathrm{Var}(\boldsymbol{u}_\bullet) = \boldsymbol{\Sigma} \otimes \mathbf{I}_n. \tag{5.55}$$

If we now define the $g \times g$ matrix $\boldsymbol{\Psi}$ such that $\boldsymbol{\Psi}\boldsymbol{\Psi}^\top = \boldsymbol{\Sigma}^{-1}$, it is clear that

$$(\boldsymbol{\Psi} \otimes \mathbf{I}_n)(\boldsymbol{\Psi} \otimes \mathbf{I}_n)^\top = (\boldsymbol{\Psi} \otimes \mathbf{I}_n)(\boldsymbol{\Psi}^\top \otimes \mathbf{I}_n) = (\boldsymbol{\Psi}\boldsymbol{\Psi}^\top \otimes \mathbf{I}_n) = \boldsymbol{\Sigma}^{-1} \otimes \mathbf{I}_n,$$

where the last expression is the inverse of the covariance matrix of $\boldsymbol{u}_\bullet$. From (5.54), the GNR corresponding to (5.55) is therefore

$$(\boldsymbol{\Psi}^\top \otimes \mathbf{I}_n)\big(\boldsymbol{y}_\bullet - \boldsymbol{x}_\bullet(\boldsymbol{\beta})\big) = (\boldsymbol{\Psi}^\top \otimes \mathbf{I}_n)\boldsymbol{X}_\bullet(\boldsymbol{\beta})\boldsymbol{b} + \text{residuals}. \tag{5.56}$$

The $gn \times k$ matrix $\boldsymbol{X}_\bullet(\boldsymbol{\beta})$ is the matrix of partial derivatives that we already defined for use in the moment conditions (5.47). Observe that, as required for a properly defined artificial regression, the inner product of the regressand with the matrix of regressors yields the left-hand side of the moment conditions (5.47), and the inverse of the inner product of the regressor matrix with itself has the same form as the covariance matrix (5.49).

The Gauss-Newton regression (5.56) can be useful in a number of contexts. It provides a convenient way to solve the estimating equations (5.47) in order to obtain an estimate of $\boldsymbol{\beta}$ for given $\boldsymbol{\Sigma}$, and it automatically computes the covariance matrix estimate (5.49) as well. Because feasible GLS and ML estimation are algebraically identical as regards the estimation of the parameter vector $\boldsymbol{\beta}$, the GNR is useful in both contexts. In practice, it is frequently used to calculate test statistics for restrictions on $\boldsymbol{\beta}$; see Section 7.7. Another important use is to impose cross-equation restrictions after equation-by-equation estimation. For this purpose, the multivariate GNR is just as useful for linear systems as for nonlinear ones; see Exercise 5.13.

5.4 Linear Simultaneous Equations Models

In Chapter 10, we dealt with instrumental variables estimation of a single equation in which some of the explanatory variables are endogenous. As we noted there, it is necessary to have information about the data-generating process for all of the endogenous variables in order to determine the optimal instruments. However, we actually dealt with only one equation, or at least only one equation at a time. The model that we consider in this section and the next, namely, the **linear simultaneous equations model**, extends what we did in Chapter 10 to a model in which all of the endogenous variables have the same status. Our objective is to obtain efficient estimates of the full set of parameters that appear in all of the simultaneous equations.

The Model

The i^{th} equation of a linear simultaneous system can be written as

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta}_i + \boldsymbol{u}_i = \boldsymbol{Z}_i\boldsymbol{\beta}_{1i} + \boldsymbol{Y}_i\boldsymbol{\beta}_{2i} + \boldsymbol{u}_i, \tag{5.57}$$

where $\boldsymbol{X}_i$ is an $n \times k_i$ matrix of explanatory variables that can be partitioned as $\boldsymbol{X}_i = [\,\boldsymbol{Z}_i \quad \boldsymbol{Y}_i\,]$. Here $\boldsymbol{Z}_i$ is an $n \times k_{1i}$ matrix of variables that are assumed to be exogenous or predetermined, and $\boldsymbol{Y}_i$ is an $n \times k_{2i}$ matrix of endogenous variables, with $k_{1i} + k_{2i} = k_i$. The k_i–vector $\boldsymbol{\beta}_i$ of parameters can be partitioned as $[\boldsymbol{\beta}_{1i} \vdots \boldsymbol{\beta}_{2i}]$ to conform with the partitioning of $\boldsymbol{X}$. The g endogenous variables $\boldsymbol{y}_1$ through $\boldsymbol{y}_g$ are assumed to be jointly generated by g equations of the form (5.57). The number of exogenous or predetermined variables that appear anywhere in the system is l. This implies that $k_{1i} \leq l$ for all i.[2]

We make the standard assumption (5.02) about the disturbances. Thus we allow for contemporaneous correlation, but not for heteroskedasticity or serial correlation. It is, of course, quite possible to allow for these extra complications, but they are not admitted in the context of the model currently under discussion, which thus has a distinctly classical flavor, as befits a model that has inspired a long and distinguished literature.

Except for the explicit distinction between endogenous and predetermined explanatory variables, equation (5.57) looks very much like the typical equation (5.01) of an SUR system. However, there is one important difference, which is concealed by the notation. It is that, as with the simple demand-supply

[2] Readers should be warned that the notation we have introduced in equation (5.57) is not universal. In particular, some authors reverse the definitions of $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ and then define $\boldsymbol{X}$ to be the $n \times l$ matrix of all the exogenous and predetermined variables, which we will denote below by $\boldsymbol{W}$. Our notation emphasizes the similarities between the linear simultaneous equations model (5.57) and the linear SUR system (5.01), as well as making it clear that $\boldsymbol{W}$ plays the role of a matrix of instruments.

model of Section 10.2, the dependent variables $\boldsymbol{y}_i$ are not necessarily distinct. Since equations (5.57) form a simultaneous system, it is arbitrary which one of the endogenous variables is put on the left-hand side with a coefficient of 1, at least in any equation in which more than one endogenous variable appears. It is a matter of simple algebra to select one of the variables in the matrix $\boldsymbol{Y}_i$, take it over to the left-hand side while taking $\boldsymbol{y}_i$ over to the right, and then rescale the coefficients so that the selected variable has a coefficient of 1. This point can be important in practice.

Just as we did with the linear SUR model, we can convert the system of equations (5.57) to a single equation by stacking them vertically. As before, the gn–vectors $\boldsymbol{y}_\bullet$ and $\boldsymbol{u}_\bullet$ consist of the $\boldsymbol{y}_i$ and the $\boldsymbol{u}_i$, respectively, stacked vertically. The $gn \times k$ matrix $\boldsymbol{X}_\bullet$, where $k = k_1 + \ldots + k_g$, is defined to be a block-diagonal matrix with diagonal blocks $\boldsymbol{X}_i$, just as in equation (5.05). The full system can then be written as

$$\boldsymbol{y}_\bullet = \boldsymbol{X}_\bullet \boldsymbol{\beta}_\bullet + \boldsymbol{u}_\bullet, \quad \mathrm{E}(\boldsymbol{u}_\bullet \boldsymbol{u}_\bullet^\top) = \boldsymbol{\Sigma} \otimes \mathbf{I}_n, \tag{5.58}$$

where the k–vector $\boldsymbol{\beta}_\bullet$ is formed by stacking the $\boldsymbol{\beta}_i$ vertically. As before, the $g \times g$ matrix $\boldsymbol{\Sigma}$ is the contemporaneous covariance matrix of the disturbances. The true value of $\boldsymbol{\beta}_\bullet$ will be denoted $\boldsymbol{\beta}_\bullet^0$.

Efficient GMM Estimation

One of the main reasons for estimating a full system of equations is to obtain an efficiency gain relative to single-equation estimation. In Section 11.2, we saw how to obtain the most efficient possible estimator for a single equation in the context of efficient GMM estimation. The theoretical moment conditions that lead to such an estimator are given in equation (2.18), which we rewrite here for easy reference:

$$\mathrm{E}\big(\bar{\boldsymbol{X}}^\top \boldsymbol{\Omega}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\big) = \mathbf{0}. \tag{2.18}$$

Because we are assuming that there is no serial correlation, these moment conditions are also valid for the linear simultaneous equations model (5.57). We simply need to reinterpret them in terms of that model.

In reinterpreting the moment conditions (2.18), it is clear that $\boldsymbol{y}_\bullet$ replaces the vector $\boldsymbol{y}$, $\boldsymbol{X}_\bullet \boldsymbol{\beta}_\bullet$ replaces the vector $\boldsymbol{X}\boldsymbol{\beta}$, and $\boldsymbol{\Sigma}^{-1} \otimes \mathbf{I}_n$ replaces the matrix $\boldsymbol{\Omega}^{-1}$. What is not quite so clear is what replaces the matrix $\bar{\boldsymbol{X}}$. Recall that $\bar{\boldsymbol{X}}$ in (2.18) is the matrix defined row by row so as to contain the expectations of the explanatory variables for each observation conditional on the information that is predetermined for that observation. We need to obtain the matrix that corresponds to $\bar{\boldsymbol{X}}$ in equation (2.18) for the model (5.58).

Let $\boldsymbol{W}$ denote an $n \times l$ matrix of exogenous and predetermined variables, the columns of which are all of the linearly independent columns of the $\boldsymbol{Z}_i$. For these variables, the expectations conditional on predetermined information are just the variables themselves. Thus we only need worry about the

endogenous explanatory variables. Because their joint DGP is given by the system of linear equations (5.57), it must be possible to solve these equations for the endogenous variables as functions of the predetermined variables and the disturbances. Since these equations are linear and have the same form for all observations, the solution must have the form

$$\boldsymbol{y}_i = \boldsymbol{W}\boldsymbol{\pi}_i + \text{disturbances}, \tag{5.59}$$

where $\boldsymbol{\pi}_i$ is an l–vector of parameters that are, in general, nonlinear functions of the parameters $\boldsymbol{\beta}_\bullet$. As the notation indicates, the variables contained in the matrix $\boldsymbol{W}$ serve as instrumental variables for the estimation of the model parameters. Later, we will investigate more fully the nature of the $\boldsymbol{\pi}_i$. We pay little attention to the disturbances, because our objective is to compute the conditional expectations of the elements of the $\boldsymbol{y}_i$, and we know that each of the disturbances must have expectation 0 conditional on all the exogenous and predetermined variables.

The vector of conditional expectations of the elements of $\boldsymbol{y}_i$ is just $\boldsymbol{W}\boldsymbol{\pi}_i$. Since equations (5.59) take the form of linear regressions with exogenous and predetermined explanatory variables, OLS estimates of the $\boldsymbol{\pi}_i$ are consistent. As we saw in Section 5.2, they are also efficient, even though the disturbances generally display contemporaneous correlation, because the same regressors appear in every equation. Thus we can replace the unknown $\boldsymbol{\pi}_i$ by their OLS estimates based on equations (5.59). This means that the conditional expectations of the vectors $\boldsymbol{y}_i$ are estimated by the OLS fitted values, that is, the vectors $\boldsymbol{W}\hat{\boldsymbol{\pi}}_i = \boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{y}_i$. When this is done, the matrices that contain the estimates of the conditional expectations of the elements of the $\boldsymbol{X}_i$ can be written as

$$\hat{\boldsymbol{X}}_i \equiv [\, \boldsymbol{Z}_i \quad \boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{Y}_i \,] = \boldsymbol{P}_{\boldsymbol{W}} [\, \boldsymbol{Z}_i \quad \boldsymbol{Y}_i \,] = \boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{X}_i. \tag{5.60}$$

We write $\hat{\boldsymbol{X}}_i$ rather than $\bar{\boldsymbol{X}}_i$ because the unknown conditional expectations are estimated. The step from the second to the third expression in (5.60) is possible because all the columns of all the $\boldsymbol{Z}_i$ are, by construction, contained in the span of the columns of $\boldsymbol{W}$.

We are now ready to construct the matrix to be used in place of $\bar{\boldsymbol{X}}$ in (2.18). It is the block-diagonal $gn \times k$ matrix $\hat{\boldsymbol{X}}_\bullet$, with diagonal blocks the $\hat{\boldsymbol{X}}_i$. This allows us to write the estimating equations for efficient GMM estimation as

$$\hat{\boldsymbol{X}}_\bullet^\top (\boldsymbol{\Sigma}^{-1} \otimes \mathbf{I}_n)(\boldsymbol{y}_\bullet - \boldsymbol{X}_\bullet \boldsymbol{\beta}_\bullet) = \mathbf{0}. \tag{5.61}$$

These equations, which are the empirical versions of the theoretical moment conditions (2.18), can be rewritten in several other ways. In particular, they can be written in the form

$$\begin{bmatrix} \sigma^{11}\boldsymbol{X}_1^\top \boldsymbol{P}_{\boldsymbol{W}} & \cdots & \sigma^{1g}\boldsymbol{X}_1^\top \boldsymbol{P}_{\boldsymbol{W}} \\ \vdots & \ddots & \vdots \\ \sigma^{g1}\boldsymbol{X}_g^\top \boldsymbol{P}_{\boldsymbol{W}} & \cdots & \sigma^{gg}\boldsymbol{X}_g^\top \boldsymbol{P}_{\boldsymbol{W}} \end{bmatrix} \begin{bmatrix} \boldsymbol{y}_1 - \boldsymbol{X}_1\boldsymbol{\beta}_1 \\ \vdots \\ \boldsymbol{y}_g - \boldsymbol{X}_g\boldsymbol{\beta}_g \end{bmatrix} = \mathbf{0},$$

by analogy with equation (5.13), and in the form

$$\sum_{j=1}^{g} \sigma^{ij} \boldsymbol{X}_i^\top \boldsymbol{P}_{\boldsymbol{W}} (\boldsymbol{y}_j - \boldsymbol{X}_j \boldsymbol{\beta}_j) = \boldsymbol{0}, \quad i = 1, \ldots, g, \tag{5.62}$$

by analogy with equation (5.16). It is also straightforward to check (see Exercise 5.14) that they can be written as

$$\boldsymbol{X}_\bullet^\top (\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{P}_{\boldsymbol{W}})(\boldsymbol{y}_\bullet - \boldsymbol{X}_\bullet \boldsymbol{\beta}_\bullet) = \boldsymbol{0}, \tag{5.63}$$

from which it follows immediately that equations (5.61) are equivalent to the first-order conditions for the minimization of the criterion function

$$(\boldsymbol{y}_\bullet - \boldsymbol{X}_\bullet \boldsymbol{\beta}_\bullet)^\top (\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{P}_{\boldsymbol{W}})(\boldsymbol{y}_\bullet - \boldsymbol{X}_\bullet \boldsymbol{\beta}_\bullet). \tag{5.64}$$

The efficient GMM estimator $\hat{\boldsymbol{\beta}}_\bullet^{\text{GMM}}$ defined by (5.63) is the analog for a linear simultaneous equations system of the GLS estimator (5.09) for an SUR system.

The asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}_\bullet^{\text{GMM}}$ can readily be obtained from expression (2.29). In the notation of (5.61), we find that

$$\operatorname{Var}\Big(\operatorname*{plim}_{n\to\infty} n^{1/2}(\hat{\boldsymbol{\beta}}_\bullet^{\text{GMM}} - \boldsymbol{\beta}_\bullet^0)\Big) = \operatorname*{plim}_{n\to\infty} \Big(\frac{1}{n} \hat{\boldsymbol{X}}_\bullet^\top (\boldsymbol{\Sigma}^{-1} \otimes \mathbf{I}_n) \hat{\boldsymbol{X}}_\bullet\Big)^{-1}. \tag{5.65}$$

This covariance matrix can also be written, in the notation of (5.63), as

$$\operatorname*{plim}_{n\to\infty} \Big(\frac{1}{n} \boldsymbol{X}_\bullet^\top (\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{P}_{\boldsymbol{W}}) \boldsymbol{X}_\bullet\Big)^{-1}. \tag{5.66}$$

Of course, the estimator $\hat{\boldsymbol{\beta}}_\bullet^{\text{GMM}}$ is not feasible if, as is almost always the case, the matrix $\boldsymbol{\Sigma}$ is unknown. However, it is obvious that we can deal with this problem by using a procedure analogous to feasible GLS estimation of an SUR system. We will return to this issue at the end of this section.

Two Special Cases

If the matrix $\boldsymbol{\Sigma}$ is diagonal, then equations (5.62) simplify to

$$\sigma^{ii} \boldsymbol{X}_i^\top \boldsymbol{P}_{\boldsymbol{W}} (\boldsymbol{y}_i - \boldsymbol{X}_i \boldsymbol{\beta}_i) = \boldsymbol{0}, \quad i = 1, \ldots, g. \tag{5.67}$$

The factors of σ^{ii} have no influence on the solutions to these equations, which are therefore just the generalized IV, or 2SLS, estimators for each of the equations of the system treated individually, with a common matrix $\boldsymbol{W}$ of instrumental variables. This result is the analog of what we found for an SUR system with diagonal $\boldsymbol{\Sigma}$. Here it is the equation-by-equation IV estimator that takes the place of the equation-by-equation OLS estimator.

Just as single-equation OLS estimation is consistent but in general inefficient for an SUR system, so is single-equation IV estimation consistent but in general inefficient for the linear simultaneous equations model. As readers are asked to verify in Exercise 5.15, the estimating equations (5.67), without the factors of σ^{ii}, can be rewritten for the entire system as

$$\boldsymbol{X}_\bullet^\top (\mathbf{I}_g \otimes \boldsymbol{P}_{\boldsymbol{W}})(\boldsymbol{y}_\bullet - \boldsymbol{X}_\bullet \boldsymbol{\beta}_\bullet) = \boldsymbol{0}. \tag{5.68}$$

In general, solving equations (5.68) yields an inefficient estimator unless the true contemporaneous covariance matrix $\boldsymbol{\Sigma}$ is diagonal.

There is, however, another case in which the estimating equations (5.68) yield an asymptotically efficient estimator. This case is analogous to the case of an SUR system with the same explanatory variables in each equation, but it takes a rather different form in this context. What we require is that each of the equations in the system should be just identified.

When we say that a single equation is just identified by an IV estimator, part of what we mean is that the number of instruments is equal to the number of explanatory variables, or, equivalently for a linear regression, to the number of parameters. If equation i is just identified, therefore, the two matrices $\boldsymbol{W}$ and $\boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{X}_i$ have the same dimensions. In fact, they span the same linear subspace provided that $\boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{X}_i$ is of full column rank. Consequently, there exists an $l \times l$ matrix $\boldsymbol{J}_i$ such that $\boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{X}_i \boldsymbol{J}_i = \boldsymbol{W}$. Premultiplying the i^{th} equation of (5.62) by $\boldsymbol{J}_i^\top$ thus gives

$$\sum_{j=1}^{g} \sigma^{ij} \boldsymbol{W}^\top (\boldsymbol{y}_j - \boldsymbol{X}_j \boldsymbol{\beta}_j) = \boldsymbol{0}.$$

If all the equations of a simultaneous equations system are just identified, then the above relation holds for each $i = 1, \ldots, g$. We can then multiply equation i by σ_{mi} and sum over i, as in equation (5.20). This yields the decoupled estimating equations

$$\boldsymbol{W}^\top (\boldsymbol{y}_m - \boldsymbol{X}_m \boldsymbol{\beta}_m) = \boldsymbol{0}, \quad m = 1, \ldots, g,$$

which define the single-equation (simple) IV estimators in the just-identified case. Therefore, as with the SUR model, there is no advantage to system estimation rather than equation-by-equation estimation when every equation is just identified, because the estimating equations use up all of the available moment conditions.

Identification

In order to be able to solve the estimating equations (5.63) for $\boldsymbol{\beta}_\bullet$, it must be possible to invert the matrix

$$\boldsymbol{X}_\bullet^\top (\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{P}_{\boldsymbol{W}}) \boldsymbol{X}_\bullet. \tag{5.69}$$

Thus, in finite samples, the parameters of the model (5.58) are identified if this matrix is nonsingular. Although this statement is accurate, it is neither complete nor transparent. In particular, even if the matrix (5.69) is singular, it may still be possible to identify some of the parameters.

Whenever the contemporaneous covariance matrix $\boldsymbol{\Sigma}$ is nonsingular, it can be shown, as spelled out in Exercise 5.16, that the matrix (5.69) is singular if and only if at least one of the matrices $\boldsymbol{P_W X}_i$ does not have full column rank. In other words, the system of equations is unidentified if and only if at least one of its component equations is unidentified. The result of the exercise also shows that the parameters of those equations for which $\boldsymbol{P_W X}_i$ does have full column rank can be identified uniquely by the estimating equations (5.63). In consequence, provided that $\boldsymbol{\Sigma}$ is nonsingular, we can study identification equation by equation without loss of generality.

A necessary condition for $\boldsymbol{P_W X}_i$ to have full column rank is that l, the number of instruments contained in the matrix $\boldsymbol{W}$, should be no less than k_i, the number of explanatory variables contained in $\boldsymbol{X}_i$. This condition is called the **order condition** for identification of equation i. It is an accounting condition, and, as such, can be expressed in more than one way. Recall that we defined k_{1i} as the number of exogenous or predetermined explanatory variables in $\boldsymbol{X}_i$, that is, the dimension of the matrix $\boldsymbol{Z}_i$. Since the total number of exogenous or predetermined variables in the full system is l, the number of such variables *excluded* from equation i is $l - k_{1i}$. The number of endogenous explanatory variables *included* in equation i is, by definition k_{2i}, which is the dimension of the matrix $\boldsymbol{Y}_i$. Therefore, the inequality $l \geq k_i$ is equivalent to

$$l \geq k_{1i} + k_{2i} \quad \text{or} \quad l - k_{1i} \geq k_{2i}. \tag{5.70}$$

The second inequality here says that the number of predetermined variables excluded from an equation must be at least as great as the number of endogenous explanatory variables in that equation.

The necessary and sufficient condition for the identification of the parameters of equation i is that $\boldsymbol{P_W X}_i$ should have full column rank of k_i. This condition, which is, not surprisingly, called the **rank condition** for identification, holds whenever the $k_i \times k_i$ matrix $\boldsymbol{X}_i^\top \boldsymbol{P_W X}_i$ is nonsingular. It is easy to check whether the rank condition holds for any given data set. However, it is not so easy to check whether it holds asymptotically. The problem is that, because some of the columns of $\boldsymbol{X}_i$ are endogenous, $\operatorname{plim} n^{-1} \boldsymbol{X}_i^\top \boldsymbol{P_W X}_i$ depends on the parameters of the DGP. This point is important, and we will discuss it at some length below.

Structural and Reduced Forms

When the equations of a linear simultaneous equations model are written in the form (5.57), it is normally the case that each equation has a direct economic interpretation. In the model of Section 8.2, for instance, the two

equations are intended to correspond to demand and supply functions. It is for this reason that these are called **structural equations**. The full system of equations constitutes what is called the **structural form** of the model.

It is convenient for our subsequent analysis to stack the equations (5.57) horizontally, instead of vertically as in the system (5.58). We thus define the $n \times g$ matrix $\boldsymbol{Y}$ as $[\boldsymbol{y}_1 \;\; \boldsymbol{y}_2 \;\; \cdots \;\; \boldsymbol{y}_g]$. Similarly, the vectors $\boldsymbol{u}_i$ of disturbances can be stacked side by side to form the $n \times g$ matrix $\boldsymbol{U}$. In this notation, the entire set of equations (5.57) can be represented as

$$\boldsymbol{Y\Gamma} = \boldsymbol{WB} + \boldsymbol{U}, \tag{5.71}$$

where the $g \times g$ matrix $\boldsymbol{\Gamma}$ and the $l \times g$ matrix $\boldsymbol{B}$ are defined in such a way as to make (5.71) equivalent to (5.57). Each equation of the system (5.57) contributes one column to (5.71). This can be seen by writing equation i of (5.57) in the form

$$[\,\boldsymbol{y}_i \quad \boldsymbol{Y}_i\,] \begin{bmatrix} 1 \\ -\boldsymbol{\beta}_{2i} \end{bmatrix} = \boldsymbol{Z}_i \boldsymbol{\beta}_{1i} + \boldsymbol{u}_i. \tag{5.72}$$

All of the columns of $\boldsymbol{Y}_i$ are also columns of $\boldsymbol{Y}$, as is $\boldsymbol{y}_i$ itself, and so column i of the matrix $\boldsymbol{\Gamma}$ has 1 for element i, and the elements of the vector $-\boldsymbol{\beta}_{2i}$ for the other nonzero elements. The endogenous variables that are excluded from equation i contribute zero elements to the column. Similarly, all the columns of $\boldsymbol{Z}_i$ are also columns of $\boldsymbol{W}$, and so the nonzero elements of column i of $\boldsymbol{B}$ are the elements of $\boldsymbol{\beta}_{1i}$, in appropriate positions. The "structure" of the structural equations is embodied in the structure of the matrices $\boldsymbol{\Gamma}$ and $\boldsymbol{B}$.

If (5.71) is to represent a model by which the g endogenous variables are generated, it is necessary for $\boldsymbol{\Gamma}$ to be nonsingular. We can thus postmultiply both sides of equation (5.71) by $\boldsymbol{\Gamma}^{-1}$ to obtain

$$\boldsymbol{Y} = \boldsymbol{WB\Gamma}^{-1} + \boldsymbol{V}, \tag{5.73}$$

where $\boldsymbol{V} \equiv \boldsymbol{U\Gamma}^{-1}$. The representation (5.73) is called the **reduced form** of the model, and its component equations (the columns of the matrix equation) are the **reduced form equations**. These reduced form equations are regressions, which in general are nonlinear in the parameters. Because they have only exogenous or predetermined regressors, they can be estimated consistently by nonlinear least squares.

Unless all the equations of the system are just identified, (5.73) is in fact what is called the **restricted reduced form** or **RRF**. This is in contrast to the **unrestricted reduced form**, or **URF**, which can be written as

$$\boldsymbol{Y} = \boldsymbol{W\Pi} + \boldsymbol{V}, \tag{5.74}$$

where $\boldsymbol{\Pi}$ is an unrestricted $l \times g$ matrix. Notice that equation (5.59) is simply the i^{th} equation of this system, with $\boldsymbol{y}_i$ the i^{th} column of the matrix $\boldsymbol{Y}$ and $\boldsymbol{\pi}_i$ the i^{th} column of the matrix $\boldsymbol{\Pi}$.

It may at first sight seem odd to refer to (5.73) as the *restricted* reduced form and to (5.74) as the *unrestricted* one. The URF (5.74) has gl regression coefficients, since $\boldsymbol{\Pi}$ is an $l \times g$ matrix, while the RRF (5.73) appears to have $gl + g^2$ parameters, since $\boldsymbol{B}$ is $l \times g$ and $\boldsymbol{\Gamma}$ is $g \times g$. But remember that $\boldsymbol{\Gamma}$ has g elements which are constrained to equal 1, and both $\boldsymbol{\Gamma}$ and $\boldsymbol{B}$ have many zero elements corresponding to excluded endogenous and predetermined explanatory variables, respectively. As readers are invited to show in Exercise 5.18, if all the equations of the system are just identified, so that the order condition (5.70) is satisfied with equality for each $i = 1, \ldots, g$, then there are exactly as many parameters in the RRF as in the URF. When some of the order conditions are inequalities, there are fewer parameters in the RRF than in the URF.

Asymptotic Identification

Whether or not the parameters of a linear simultaneous system are identified by a given data set depends only on the order condition and the properties of the actual data, but this is not true of asymptotic identification. Since the parameters must be asymptotically identified if the parameter estimates are to be consistent, it is worth studying in some detail the conditions for asymptotic identification in such a system.

We assume that the probability limit of $n^{-1}\boldsymbol{W}^\top\boldsymbol{U}$ is a zero matrix and that the $l \times l$ matrix

$$\boldsymbol{S}_{\boldsymbol{W}^\top\boldsymbol{W}} \equiv \operatorname*{plim}_{n\to\infty} \frac{1}{n}\boldsymbol{W}^\top\boldsymbol{W}$$

is positive definite and, consequently, nonsingular. The nonsingularity of the matrix $\boldsymbol{W}^\top\boldsymbol{W}$ is not necessary for identification by a given data set, since, if there are enough instruments, it is quite possible that each of the matrices $\boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{X}_i$, $i = 1, \ldots, g$, should have full column rank even though some of the instruments are linearly dependent. Similarly, it is not *necessary* that $\boldsymbol{S}_{\boldsymbol{W}^\top\boldsymbol{W}}$ should be nonsingular for asymptotic identification. However, since it is always possible to eliminate linearly dependent instruments, it is convenient to make the nonsingularity assumption. By doing so, we make it clearer how asymptotic identification depends on the actual parameter values.

For simplicity of notation, we focus on the asymptotic identification of the first equation of the system, which can be written as

$$\boldsymbol{y}_1 = \boldsymbol{Z}_1\boldsymbol{\beta}_{11} + \boldsymbol{Y}_1\boldsymbol{\beta}_{21} + \boldsymbol{u}_1. \tag{5.75}$$

Since identification can be treated equation by equation without loss of generality, and since the ordering of the equations is quite arbitrary, our results will be perfectly general. The matrix $\boldsymbol{X}_1$ of explanatory variables for the first equation is $\boldsymbol{X}_1 = [\,\boldsymbol{Z}_1 \quad \boldsymbol{Y}_1\,]$. Recall that the $n \times l$ matrix $\boldsymbol{W}$ contains all the linearly independent columns of the $\boldsymbol{Z}_i$, and in particular those of $\boldsymbol{Z}_1$. Let us order the columns of $\boldsymbol{W}$ so that the k_{11} columns of $\boldsymbol{Z}_1$ come first.

The $n \times k_{21}$ matrix $\boldsymbol{Y}_1$ is given by a selection of the columns of the matrix $\boldsymbol{Y}$. The first column of $\boldsymbol{Y}$, which corresponds to the equation we are studying, is not among these, because $\boldsymbol{y}_1$ appears only on the left-hand side of that equation. However, we can freely reorder the remaining columns of $\boldsymbol{Y}$ so that the k_{21} columns of $\boldsymbol{Y}_1$ are the columns 2 through $k_{21} + 1$ of $\boldsymbol{Y}$. This done, we can express the first $k_{21} + 1$ columns of the URF (5.74), in partitioned form, as

$$[\,\boldsymbol{y}_1 \quad \boldsymbol{Y}_1\,] = [\,\boldsymbol{Z}_1 \quad \boldsymbol{W}_1\,]\begin{bmatrix}\boldsymbol{\pi}_{11} & \boldsymbol{\Pi}_{11}\\ \boldsymbol{\pi}_{21} & \boldsymbol{\Pi}_{21}\end{bmatrix} + [\,\boldsymbol{v}_1 \quad \boldsymbol{V}_1\,], \tag{5.76}$$

where we have introduced some further convenient notation. First, the $n \times (l - k_{11})$ matrix $\boldsymbol{W}_1$ contains all the columns of $\boldsymbol{W}$ that are not in $\boldsymbol{Z}_1$. Then, for the ordering that we have chosen for the columns of $\boldsymbol{Y}$ and $\boldsymbol{W}$, $\boldsymbol{\pi}_{11}$ is the $k_{11} \times 1$ vector of parameters in the first reduced form equation (that is, the equation that defines $\boldsymbol{y}_1$) associated with the instruments in the matrix $\boldsymbol{Z}_1$, while the $(l - k_{11}) \times 1$ vector $\boldsymbol{\pi}_{21}$ contains the parameters of the first reduced form equation associated with the instruments in $\boldsymbol{W}_1$. Finally, the matrices $\boldsymbol{\Pi}_{11}$ and $\boldsymbol{\Pi}_{21}$ are, respectively, of dimensions $k_{11} \times k_{21}$ and $(l - k_{11}) \times k_{21}$. They contain the parameters of the reduced form equations numbered 2 through $k_{21} + 1$ and associated with the instruments in $\boldsymbol{Z}_1$ and $\boldsymbol{W}_1$, respectively. The matrix $[\,\boldsymbol{v}_1 \quad \boldsymbol{V}_1\,]$ of disturbances is partitioned in the same way as the left-hand side of (5.76).

We can write the matrix $\boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{X}_1$ as

$$\boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{X}_1 = \boldsymbol{P}_{\boldsymbol{W}}[\,\boldsymbol{Z}_1 \quad \boldsymbol{Y}_1\,] = [\,\boldsymbol{Z}_1 \quad \boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{Y}_1\,], \tag{5.77}$$

because $\boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{Z}_1 = \boldsymbol{Z}_1$. With the help of (5.76), the second block of the rightmost expression above becomes

$$\boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{Y}_1 = [\,\boldsymbol{Z}_1 \quad \boldsymbol{W}_1\,]\begin{bmatrix}\boldsymbol{\Pi}_{11}\\ \boldsymbol{\Pi}_{21}\end{bmatrix} + \boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{V}_1, \tag{5.78}$$

where we again use the fact that $\boldsymbol{P}_{\boldsymbol{W}}[\,\boldsymbol{Z}_1 \quad \boldsymbol{W}_1\,] = [\,\boldsymbol{Z}_1 \quad \boldsymbol{W}_1\,]$, and $\boldsymbol{\Pi}_{11}$ and $\boldsymbol{\Pi}_{21}$ contain the true parameter values. Reorganizing equations (5.77) and (5.78) gives

$$\boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{X}_1 = \boldsymbol{W}\begin{bmatrix}\mathbf{I}_{k_{11}} & \boldsymbol{\Pi}_{11}\\ \mathbf{O} & \boldsymbol{\Pi}_{21}\end{bmatrix} + [\,\mathbf{O} \quad \boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{V}_1\,]. \tag{5.79}$$

The necessary and sufficient condition for the asymptotic identification of the parameters of the first equation is the nonsingularity of the probability limit as $n \to \infty$ of the matrix $n^{-1}\boldsymbol{X}_1^\top\boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{X}_1$. It is easy to see from (5.79) that this limit is

$$\operatorname*{plim}_{n\to\infty} \frac{1}{n}\boldsymbol{X}_1^\top\boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{X}_1 = \begin{bmatrix}\mathbf{I}_{k_{11}} & \mathbf{O}\\ \boldsymbol{\Pi}_{11}^\top & \boldsymbol{\Pi}_{21}^\top\end{bmatrix}\boldsymbol{S}_{\boldsymbol{W}^\top\boldsymbol{W}}\begin{bmatrix}\mathbf{I}_{k_{11}} & \boldsymbol{\Pi}_{11}\\ \mathbf{O} & \boldsymbol{\Pi}_{21}\end{bmatrix}.$$

In Exercise 5.19, readers are invited to check that everything that depends on the matrix $\boldsymbol{V}$ does indeed tend to zero in the above limit. Since we assumed that $\boldsymbol{S}_{\boldsymbol{W}^\top \boldsymbol{W}}$ is positive definite, it follows that equation (5.75) is asymptotically identified if and only if the matrix

$$\begin{bmatrix} \mathbf{I}_{k_{11}} & \boldsymbol{\Pi}_{11} \\ \mathbf{O} & \boldsymbol{\Pi}_{21} \end{bmatrix} \tag{5.80}$$

is of full column rank $k_1 = k_{11} + k_{21}$. Because this matrix has l rows, this is not possible unless $l \geq k_1$, that is, unless the order condition is satisfied. However, even if the order condition is satisfied, there can perfectly well exist parameter values for which (5.80) does not have full column rank. The important conclusion of this analysis is that asymptotic identification of an equation in a linear simultaneous system depends not only on the properties of the instrumental variables $\boldsymbol{W}$, but also on the specific parameter values of the DGP.

In Exercise 5.20, readers are asked to show that the matrix (5.80) has full column rank if and only if the $(l - k_{11}) \times k_{21}$ submatrix $\boldsymbol{\Pi}_{21}$ has full column rank. While this is a simple enough condition, it is expressed in terms of the reduced form parameters, which are usually not subject to a simple interpretation. It is therefore desirable to have a characterization of the asymptotic identification condition in terms of the structural parameters. In Exercise 5.21, notation that is suitable for deriving such a characterization is proposed, and readers are asked to develop it in Exercise 5.22.

The numerical condition that the matrix (5.69) be nonsingular is satisfied by almost all data sets, even when the rank condition for asymptotic identification is not satisfied. When this happens, the failure of that condition manifests itself as the phenomenon of **weak instruments** that we discussed in Section 10.4. In such a case, we might be tempted to add additional instruments, such as lags of the instruments themselves or other predetermined variables that may be correlated with them. But doing this cannot lead to asymptotic identification, because it would simply append columns of zeros to the matrix $\boldsymbol{\Pi}$ of reduced form coefficients, and it is obvious that such an operation cannot convert a matrix of deficient rank into one of full rank.

A discussion of asymptotic identification that is more detailed than the present one, but still reasonably compact, is provided by Davidson and MacKinnon (1993, Section 18.3). Much fuller treatments may be found in Fisher (1976) and Hsiao (1983).

Three-Stage Least Squares

The efficient GMM estimator defined by the estimating equations (5.63) is not feasible unless $\boldsymbol{\Sigma}$ is known. However, we can compute a feasible GMM estimator if we can obtain a consistent estimate of $\boldsymbol{\Sigma}$, and this is easy to do. We first estimate the individual equations of the system by generalized IV,

or two-stage least squares, to use the traditional terminology. This inefficient equation-by-equation estimator is characterized formally by the estimating equations (5.68). After computing it, we then use the 2SLS residuals to compute the matrix $\hat{\boldsymbol{\Sigma}}_{\text{2SLS}}$, as in (5.17). Using $\hat{\boldsymbol{\Sigma}}_{\text{2SLS}}$ in place of $\boldsymbol{\Sigma}$ in equations (5.63) yields the popular **three-stage least-squares**, or **3SLS**, estimator, which was originally proposed by Zellner and Theil (1962). This estimator can be written as

$$\hat{\boldsymbol{\beta}}_\bullet^{\text{3SLS}} = \big(\boldsymbol{X}_\bullet^\top (\hat{\boldsymbol{\Sigma}}_{\text{2SLS}}^{-1} \otimes \boldsymbol{P}_{\boldsymbol{W}}) \boldsymbol{X}_\bullet\big)^{-1} \boldsymbol{X}_\bullet^\top (\hat{\boldsymbol{\Sigma}}_{\text{2SLS}}^{-1} \otimes \boldsymbol{P}_{\boldsymbol{W}}) \boldsymbol{y}_\bullet. \tag{5.81}$$

The relationship between this 3SLS estimator and the 2SLS estimator for the entire system is essentially the same as the relationship between the feasible GLS estimator (5.18) for an SUR system and the OLS estimator (5.06). As with (5.18), we may wish to compute the continuously updated version of the 3SLS estimator (5.81), in which case we iteratively update the estimates of $\boldsymbol{\beta}_\bullet$ and $\boldsymbol{\Sigma}$ by using equations (5.81) and (5.17), respectively.

From the results (5.65) and (5.66), it is clear that we can estimate the covariance matrix of the classical 3SLS estimator (5.81) by

$$\widehat{\text{Var}}(\hat{\boldsymbol{\beta}}_\bullet^{\text{3SLS}}) = \big(\boldsymbol{X}_\bullet^\top (\hat{\boldsymbol{\Sigma}}_{\text{2SLS}}^{-1} \otimes \boldsymbol{P}_{\boldsymbol{W}}) \boldsymbol{X}_\bullet\big)^{-1}, \tag{5.82}$$

which is analogous to (5.19) for the SUR case. Asymptotically valid inferences can then be made in the usual way. As with the SUR estimator, we can perform a Hansen-Sargan test of the overidentifying restrictions by using the fact that, under the null hypothesis, the criterion function (5.64) evaluated at $\hat{\boldsymbol{\beta}}_\bullet^{\text{3SLS}}$ and $\hat{\boldsymbol{\Sigma}}_{\text{2SLS}}$ is asymptotically distributed as $\chi^2(gl - k)$. Of course, this is also true if the procedure has been iterated one or more times.

5.5 Maximum Likelihood Estimation

Like the SUR model, the linear simultaneous equations model can be estimated by maximum likelihood under the assumption that the disturbances, in addition to satisfying the requirements (5.02), are normally distributed. In contrast to the situation with an SUR system, where the ML estimator is numerically identical to the continuously updated feasible GLS estimator, the ML estimator of a linear simultaneous equations model is, in general, different from the continuously updated 3SLS estimator. The ML and 3SLS estimators are, however, asymptotically equivalent, whether or not the latter is continuously updated.

Because the algebra of ML estimation is quite complicated, we have divided our treatment of the subject between this section and a technical appendix, which appears at the end of the chapter, just prior to the exercises. All of the principal results are stated and discussed in this section, but many of them are derived in the appendix.

Full-Information Maximum Likelihood

The maximum likelihood estimator of a linear simultaneous system is called the **full-information maximum likelihood**, or **FIML**, estimator. It is so called because it uses information about all the equations in the system, unlike the limited-information maximum likelihood estimator (LIML) that will be discussed later in this section.

The loglikelihood function that must be maximized to obtain the FIML estimator can be written in several different ways. In terms of the notation used in equation (5.58), it is

$$\begin{aligned} -\frac{gn}{2}\log 2\pi - \frac{n}{2}\log|\boldsymbol{\Sigma}| + n\log|\det\boldsymbol{\Gamma}| \\ -\frac{1}{2}(\boldsymbol{y}_\bullet - \boldsymbol{X}_\bullet\boldsymbol{\beta}_\bullet)^\top(\boldsymbol{\Sigma}^{-1}\otimes\mathbf{I}_n)(\boldsymbol{y}_\bullet - \boldsymbol{X}_\bullet\boldsymbol{\beta}_\bullet). \end{aligned} \tag{5.83}$$

This looks very much like the loglikelihood function (5.51) for a multivariate nonlinear regression model with normally distributed disturbances. The principal difference is the third term, $n\log|\det\boldsymbol{\Gamma}|$, which is a Jacobian term. This term is the logarithm of the absolute value of the Jacobian of the transformation from $\boldsymbol{u}_\bullet$ to $\boldsymbol{y}_\bullet$. As we will see in the appendix, the loglikelihood function can also be written without an explicit Jacobian term if we start from the restricted reduced form (5.73).

Maximizing the loglikelihood function (5.83) with respect to $\boldsymbol{\Sigma}$ is exactly the same as maximizing the loglikelihood function (5.33) with respect to it. If we had ML estimates of $\boldsymbol{\beta}_\bullet$, or, equivalently, of $\boldsymbol{B}$ and $\boldsymbol{\Gamma}$, the ML estimate of $\boldsymbol{\Sigma}$ would be

$$\hat{\boldsymbol{\Sigma}}_{\mathrm{ML}} = \frac{1}{n}(\boldsymbol{Y}\hat{\boldsymbol{\Gamma}}_{\mathrm{ML}} - \boldsymbol{W}\hat{\boldsymbol{B}}_{\mathrm{ML}})^\top(\boldsymbol{Y}\hat{\boldsymbol{\Gamma}}_{\mathrm{ML}} - \boldsymbol{W}\hat{\boldsymbol{B}}_{\mathrm{ML}}), \tag{5.84}$$

which is just the sample covariance matrix of the structural-form disturbances; compare equation (5.36).

Recall from (5.57) that the parameter vector $\boldsymbol{\beta}_i$ of equation i contains both the vector $\boldsymbol{\beta}_{1i}$, which is associated with the predetermined explanatory variables, and the vector $\boldsymbol{\beta}_{2i}$, which is associated with the endogenous explanatory variables. As is clear from equation (5.72), the matrix $\boldsymbol{B}$ is determined by the $\boldsymbol{\beta}_{1i}$ alone, and the matrix $\boldsymbol{\Gamma}$ by the $\boldsymbol{\beta}_{2i}$ alone. We can obtain the first-order conditions for maximizing the loglikelihood function (5.83) with respect to the $\boldsymbol{\beta}_{1i}$ in exactly the same way as we obtained conditions (5.12) from the criterion function (5.11) for an SUR system. The first-order conditions that we seek can be written as

$$\boldsymbol{Z}_\bullet^\top(\boldsymbol{\Sigma}^{-1}\otimes\mathbf{I}_n)(\boldsymbol{y}_\bullet - \boldsymbol{X}_\bullet\boldsymbol{\beta}_\bullet) = \mathbf{0}, \tag{5.85}$$

where the $gn\times\sum_i k_{1i}$ matrix $\boldsymbol{Z}_\bullet$ is defined, similarly to $\boldsymbol{X}_\bullet$, as a matrix with diagonal blocks $\boldsymbol{Z}_i$. The number of equations in (5.85) is $\sum_i k_{1i}$, since there is one equation for each of the $\boldsymbol{\beta}_{1i}$.

Since it is rather complicated to work out the first-order conditions for the maximization of (5.83) with respect to the $\boldsymbol{\beta}_{2i}$, we leave this derivation to the appendix. These conditions can be expressed as

$$\boldsymbol{Y}_\bullet^\top(\boldsymbol{B},\boldsymbol{\Gamma})(\boldsymbol{\Sigma}^{-1}\otimes\mathbf{I}_n)(\boldsymbol{y}_\bullet - \boldsymbol{X}_\bullet\boldsymbol{\beta}_\bullet) = \mathbf{0}, \tag{5.86}$$

where the $gn\times\sum_i k_{2i}$ matrix $\boldsymbol{Y}_\bullet(\boldsymbol{B},\boldsymbol{\Gamma})$ is again defined in terms of diagonal blocks. Block i is the $n\times k_{2i}$ matrix $\boldsymbol{Y}_i(\boldsymbol{B},\boldsymbol{\Gamma})$, which is the submatrix of $\boldsymbol{W}\boldsymbol{B}\boldsymbol{\Gamma}^{-1}$ formed by selecting the columns that correspond to the columns of the matrix $\boldsymbol{Y}_i$ of included endogenous explanatory variables in equation i. The conditions (5.85) and (5.86) can be grouped together as

$$\boldsymbol{X}_\bullet^\top(\boldsymbol{B},\boldsymbol{\Gamma})(\boldsymbol{\Sigma}^{-1}\otimes\mathbf{I}_n)(\boldsymbol{y}_\bullet - \boldsymbol{X}_\bullet\boldsymbol{\beta}_\bullet) = \mathbf{0}, \tag{5.87}$$

where the i^{th} diagonal block of $\boldsymbol{X}_\bullet(\boldsymbol{B},\boldsymbol{\Gamma})$ is the $n\times k_i$ matrix $[\boldsymbol{Z}_i \;\; \boldsymbol{Y}_i(\boldsymbol{B},\boldsymbol{\Gamma})]$. There are $k = \sum_i k_{1i} + \sum_i k_{2i}$ equations in (5.87).

With (5.84) and (5.87), we have assembled all of the first-order conditions that define the FIML estimator. We write them here as a set of estimating equations:

$$\begin{aligned} &\boldsymbol{X}_\bullet^\top(\hat{\boldsymbol{B}}_{\mathrm{ML}},\hat{\boldsymbol{\Gamma}}_{\mathrm{ML}})(\hat{\boldsymbol{\Sigma}}_{\mathrm{ML}}^{-1}\otimes\mathbf{I}_n)(\boldsymbol{y}_\bullet - \boldsymbol{X}_\bullet\hat{\boldsymbol{\beta}}_\bullet^{\mathrm{ML}}) = \mathbf{0}, \text{ and} \\ &\hat{\boldsymbol{\Sigma}}_{\mathrm{ML}} = \frac{1}{n}(\boldsymbol{Y}\hat{\boldsymbol{\Gamma}}_{\mathrm{ML}} - \boldsymbol{W}\hat{\boldsymbol{B}}_{\mathrm{ML}})^\top(\boldsymbol{Y}\hat{\boldsymbol{\Gamma}}_{\mathrm{ML}} - \boldsymbol{W}\hat{\boldsymbol{B}}_{\mathrm{ML}}). \end{aligned} \tag{5.88}$$

Solving these equations, which must of course be done numerically, yields the FIML estimator.

There are many numerical methods for obtaining FIML estimates. One of them is to make use of the artificial regression

$$(\boldsymbol{\Psi}^\top\otimes\mathbf{I}_n)(\boldsymbol{y}_\bullet - \boldsymbol{X}_\bullet\boldsymbol{\beta}_\bullet) = (\boldsymbol{\Psi}^\top\otimes\mathbf{I}_n)\boldsymbol{X}_\bullet(\boldsymbol{B},\boldsymbol{\Gamma})\boldsymbol{b} + \text{residuals}, \tag{5.89}$$

where, as usual, $\boldsymbol{\Psi}\boldsymbol{\Psi}^\top = \boldsymbol{\Sigma}^{-1}$. This is analogous to the multivariate GNR (5.56). If we start from initial consistent estimates, this artificial regression can be used to update the estimates of $\boldsymbol{B}$ and $\boldsymbol{\Gamma}$, and equation (5.84) can be used to update the estimate of $\boldsymbol{\Sigma}$. Like other artificial regressions, (5.89) can also be used to compute test statistics and covariance matrices.

Another approach is to concentrate the loglikelihood function with respect to $\boldsymbol{\Sigma}$. As readers are asked to show in Exercise 5.24, the concentrated loglikelihood function can be written as

$$-\frac{gn}{2}(\log 2\pi + 1) + n\log|\det\boldsymbol{\Gamma}| - \frac{n}{2}\log\left|\frac{1}{n}(\boldsymbol{Y}\boldsymbol{\Gamma} - \boldsymbol{W}\boldsymbol{B})^\top(\boldsymbol{Y}\boldsymbol{\Gamma} - \boldsymbol{W}\boldsymbol{B})\right|, \tag{5.90}$$

which is the analog of (5.41) and (5.53). Expression (5.90) may be maximized directly with respect to $\boldsymbol{B}$ and $\boldsymbol{\Gamma}$ to yield $\hat{\boldsymbol{B}}_{\mathrm{ML}}$ and $\hat{\boldsymbol{\Gamma}}_{\mathrm{ML}}$. This approach may or may not be easier numerically than solving equations (5.88).

The FIML estimator is not defined if the matrix $(\boldsymbol{Y\Gamma} - \boldsymbol{WB})^\top(\boldsymbol{Y\Gamma} - \boldsymbol{WB})$ that appears in (5.90) does not have full rank for all admissible values of $\boldsymbol{B}$ and $\boldsymbol{\Gamma}$, and this requires that $n \geq g + k$. This result suggests that n may have to be substantially greater than $g + k$ if FIML is to have good finite-sample properties; see Sargan (1975) and Brown (1981).

Comparison with Three-Stage Least Squares

Even though the FIML and 3SLS estimators are asymptotically equivalent, the FIML estimator is not, in general, equal to the continuously updated 3SLS estimator. In order to study the relationship between the two estimators, we write out explicitly the estimating equations for 3SLS and compare them with the estimating equations (5.88) for FIML. Equations (5.61) and (5.17) imply that the continuously updated version of the 3SLS estimator is defined by the equations

$$\begin{aligned} &\hat{\boldsymbol{X}}_\bullet^\top(\hat{\boldsymbol{\Sigma}}_{\text{3SLS}}^{-1} \otimes \mathbf{I}_n)(\boldsymbol{y}_\bullet - \boldsymbol{X}_\bullet\hat{\boldsymbol{\beta}}_\bullet^{\text{3SLS}}) = \mathbf{0}, \text{ and} \\ &\hat{\boldsymbol{\Sigma}}_{\text{3SLS}} = \frac{1}{n}(\boldsymbol{Y}\hat{\boldsymbol{\Gamma}}_{\text{3SLS}} - \boldsymbol{W}\hat{\boldsymbol{B}}_{\text{3SLS}})^\top(\boldsymbol{Y}\hat{\boldsymbol{\Gamma}}_{\text{3SLS}} - \boldsymbol{W}\hat{\boldsymbol{B}}_{\text{3SLS}}). \end{aligned} \tag{5.91}$$

The second of these equations has exactly the same form as the second equation of (5.88). The first equation is also very similar to the first equation of (5.88), but there is one difference. In (5.88), the leftmost matrix on the left-hand side of the first equation is the transpose of $\boldsymbol{X}_\bullet(\hat{\boldsymbol{B}}_{\text{ML}}, \hat{\boldsymbol{\Gamma}}_{\text{ML}})$, of which the typical diagonal block is $[\,\boldsymbol{Z}_i \quad \boldsymbol{Y}_i(\hat{\boldsymbol{B}}_{\text{ML}}, \hat{\boldsymbol{\Gamma}}_{\text{ML}})\,]$. In contrast, the corresponding matrix in the first equation of (5.91) is the transpose of $\hat{\boldsymbol{X}}_\bullet$, of which the typical diagonal block is, from (5.60), $[\,\boldsymbol{Z}_i \quad \boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{Y}_i\,]$.

In both cases, the matrix is an estimate of the matrix of optimal instruments for equation i, that is, the matrix of the expectations of the explanatory variables conditional on all predetermined information. It is clear from the RRF (5.73) that this matrix is $[\,\boldsymbol{Z}_i \quad \boldsymbol{Y}_i(\boldsymbol{B}, \boldsymbol{\Gamma})\,]$, where $\boldsymbol{B}$ and $\boldsymbol{\Gamma}$ are the true parameters of the DGP. FIML uses the FIML estimates of $\boldsymbol{B}$ and $\boldsymbol{\Gamma}$ in place of the true values, while 3SLS estimates $\boldsymbol{Y}_i(\boldsymbol{B}, \boldsymbol{\Gamma})$ by $\boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{Y}_i$, that is, by the fitted values from estimation of the *unrestricted* reduced form (5.74). The latter is, in general, less efficient than the former.

If the restricted and unrestricted reduced forms are equivalent, as they must be if all the equations of the system are just identified, then the estimating equations (5.91) and (5.88) are also equivalent, and the 3SLS and FIML estimators must coincide. In this case, as we saw in the last section, 3SLS is also the same as 2SLS, that is, equation-by-equation IV estimation. Thus all the estimators we have considered are identical in the just-identified case. When there are overidentifying restrictions, and 3SLS is used without continuous updating, then the 3SLS estimators of $\boldsymbol{B}$ and $\boldsymbol{\Gamma}$ are replaced by the 2SLS ones in the second equation of (5.91). Solving this equation yields the classical 3SLS estimator (5.81), which is evidently much easier to compute than the FIML estimator.

Our treatment of the relationship between 3SLS and FIML has been quite brief. For much fuller treatments, see Hausman (1975) and Hendry (1976).

Inference Based on FIML Estimates

Since the first equation of (5.88) is just an estimating equation for efficient GMM, we can estimate the covariance matrix of $\hat{\boldsymbol{\beta}}_\bullet^{\text{ML}}$ by the obvious estimate of n^{-1} times the asymptotic covariance matrix (5.66), namely,

$$\widehat{\text{Var}}(\hat{\boldsymbol{\beta}}_\bullet^{\text{ML}}) = \big(\boldsymbol{X}_\bullet^\top(\hat{\boldsymbol{B}}_{\text{ML}}, \hat{\boldsymbol{\Gamma}}_{\text{ML}})(\hat{\boldsymbol{\Sigma}}_{\text{ML}}^{-1} \otimes \mathbf{I}_n)\boldsymbol{X}_\bullet(\hat{\boldsymbol{B}}_{\text{ML}}, \hat{\boldsymbol{\Gamma}}_{\text{ML}})\big)^{-1}. \tag{5.92}$$

Notice that, if we evaluate the artificial regression (5.89) at the ML estimates, then $1/s^2$ times the OLS covariance matrix is equal to this matrix.

There are two differences between the estimated covariance matrix for FIML given in equation (5.92) and the estimated covariance matrix for the classical 3SLS estimator given in equation (5.82). The first is that they use different estimates of $\boldsymbol{\Sigma}$. The second is that, in (5.92), the endogenous variables in $\boldsymbol{X}_\bullet$ are replaced by their fitted values, based on the FIML estimates, while in (5.82) they are replaced by their projections on to $\mathcal{S}(\boldsymbol{W})$.

If the model (5.57) is correctly specified, and the disturbances really do satisfy the assumptions we have made about them, then each row $\boldsymbol{V}_t$ of the matrix of disturbances $\boldsymbol{V}$ in the URF (5.74) must have properties like those of the structural disturbances $\boldsymbol{U}_t$ in (5.03). This implies that the disturbances in every equation of the URF must be homoskedastic and serially independent. This suggests that the first step in testing the statistical assumptions on which FIML estimation is based should *always* be to perform tests for heteroskedasticity and serial correlation on the equations of the unrestricted reduced form; suitable testing procedures were discussed in Sections 8.5 and 8.7. If there is strong evidence that the $\boldsymbol{V}_t$ are not IID, then either at least one of the structural equations is misspecified, or we need to make more complicated assumptions about the disturbances.

It is also important to test any overidentifying restrictions. In the case of FIML, it is natural to use a likelihood ratio test rather than a Hansen-Sargan test, as we suggested for 3SLS and SUR estimation. The number of restrictions is, once again, $gl - k$, the difference between the number of coefficients in the URF and the number in the structural model. The restricted value of the loglikelihood function is the maximized value of either the loglikelihood function (5.83) or the concentrated loglikelihood function (5.90), and the unrestricted value is

$$-\tfrac{gn}{2}\big(\log 2\pi + 1\big) - \tfrac{n}{2}\log\Big|\tfrac{1}{n}(\boldsymbol{Y} - \boldsymbol{W}\hat{\boldsymbol{\Pi}})^\top(\boldsymbol{Y} - \boldsymbol{W}\hat{\boldsymbol{\Pi}})\Big|,$$

where $\hat{\boldsymbol{\Pi}}$ denotes the matrix of OLS estimates of the parameters of the URF. Twice the difference between the unrestricted and restricted values of the

loglikelihood function is asymptotically distributed as $\chi^2(gl-k)$ if the model is correctly specified and the overidentifying restrictions are satisfied.

Limited-Information Maximum Likelihood

When a system of equations consists of just one structural equation, together with one or more reduced-form equations, the FIML estimator of the structural equation reduces to a single-equation estimator. We can write the single structural equation as

$$\boldsymbol{y} = \boldsymbol{Z}\boldsymbol{\beta}_1 + \boldsymbol{Y}\boldsymbol{\beta}_2 + \boldsymbol{u}, \tag{5.93}$$

where we use a notation similar to that of (5.57), but without indices on the variables and parameters. There are k_1 elements in $\boldsymbol{\beta}_1$ and k_2 in $\boldsymbol{\beta}_2$, with $k = k_1 + k_2$. A complete simultaneous system can be formed by combining (5.93) with the equations of the unrestricted reduced form for the endogenous variables in the matrix $\boldsymbol{Y}$. We write these equations as

$$\boldsymbol{Y} = \boldsymbol{W}\boldsymbol{\Pi} + \boldsymbol{V} = \boldsymbol{Z}\boldsymbol{\Pi}_1 + \boldsymbol{W}_1\boldsymbol{\Pi}_2 + \boldsymbol{V}, \tag{5.94}$$

where the matrix $\boldsymbol{W}_1$ contains all the predetermined instruments that are excluded from the matrix $\boldsymbol{Z}$.

Since the equations of the unrestricted reduced form are just identified by construction, the only equation of the system consisting of (5.93) and (5.94) that can be overidentified is (5.93) itself. If it is also just identified, then, as we have seen, 3SLS and FIML estimation both give exactly the same results as IV estimation of (5.93) by itself. If equation (5.93) is overidentified, then it turns out that 3SLS, without continuous updating, also gives the same estimates of the parameters of (5.93) as IV estimation. Readers are asked to prove this result in Exercise 5.27. However, continuously updated 3SLS and ML give different, and possibly better, estimates in this case.

Maximum likelihood estimation of equation (5.93), implicitly treating it as part of a system with (5.94), is called **limited-information maximum likelihood**, or **LIML**. The terminology "limited-information" refers to the fact that no use is made of any overidentifying or cross-equation restrictions that may apply to the parameters of the matrix $\boldsymbol{\Pi}$ of reduced-form coefficients. Formally, LIML is FIML applied to a system in which only one equation is overidentified. However, as we will see, LIML is in fact a single-equation estimation method, in the same sense that 2SLS applied to (5.93) alone is a single-equation method. The calculations necessary to see this are rather complicated, and so here we will simply state the principal result, which dates back as far as Anderson and Rubin (1949). A derivation of this result may be found in Davidson and MacKinnon (1993, Chapter 18).

The Anderson-Rubin result is that the LIML estimate of $\boldsymbol{\beta}_2$ in equation (5.93) is given by minimizing the ratio

$$\kappa \equiv \frac{(\boldsymbol{y} - \boldsymbol{Y}\boldsymbol{\beta}_2)^\top \boldsymbol{M}_{\boldsymbol{Z}} (\boldsymbol{y} - \boldsymbol{Y}\boldsymbol{\beta}_2)}{(\boldsymbol{y} - \boldsymbol{Y}\boldsymbol{\beta}_2)^\top \boldsymbol{M}_{\boldsymbol{W}} (\boldsymbol{y} - \boldsymbol{Y}\boldsymbol{\beta}_2)}, \tag{5.95}$$

where $\boldsymbol{M}_{\boldsymbol{Z}}$ projects off the predetermined variables included in (5.93), and $\boldsymbol{M}_{\boldsymbol{W}}$ projects off all the instruments, both those in $\boldsymbol{Z}$ and those in $\boldsymbol{W}_1$. The value $\hat{\kappa}$ that minimizes (5.95) may be found by a non-iterative procedure that is discussed in the appendix. The maximized value of the loglikelihood function is then

$$-\frac{gn}{2}\log 2\pi - \frac{n}{2}\log\hat{\kappa} - \frac{n}{2}\log|\boldsymbol{Y}_*^\top \boldsymbol{M}_{\boldsymbol{W}} \boldsymbol{Y}_*|, \tag{5.96}$$

where $\boldsymbol{Y}_* \equiv [\boldsymbol{y} \;\; \boldsymbol{Y}]$.

If we write equation (5.93) as $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}$, then the LIML estimator of $\boldsymbol{\beta}$ is defined by the estimating equations

$$\boldsymbol{X}^\top(\mathbf{I} - \hat{\kappa}\boldsymbol{M}_{\boldsymbol{W}})(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}^{\text{LIML}}) = \mathbf{0}, \tag{5.97}$$

which can be solved explicitly once $\hat{\kappa}$ has been computed. We find that

$$\hat{\boldsymbol{\beta}}^{\text{LIML}} = \big(\boldsymbol{X}^\top(\mathbf{I} - \hat{\kappa}\boldsymbol{M}_{\boldsymbol{W}})\boldsymbol{X}\big)^{-1}\boldsymbol{X}^\top(\mathbf{I} - \hat{\kappa}\boldsymbol{M}_{\boldsymbol{W}})\boldsymbol{y}. \tag{5.98}$$

A suitable estimate of the covariance matrix of the LIML estimator is

$$\widehat{\text{Var}}(\hat{\boldsymbol{\beta}}^{\text{LIML}}) = \hat{\sigma}^2\big(\boldsymbol{X}^\top(\mathbf{I} - \hat{\kappa}\boldsymbol{M}_{\boldsymbol{W}})\boldsymbol{X}\big)^{-1}, \tag{5.99}$$

where

$$\hat{\sigma}^2 \equiv \frac{1}{n}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}^{\text{LIML}})^\top(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}^{\text{LIML}}).$$

Given (5.99), confidence intervals, asymptotic t tests, and Wald tests can readily be computed in the usual way.

Since $\boldsymbol{W} = [\,\boldsymbol{Z} \;\; \boldsymbol{W}_1\,]$ is the matrix containing all the instruments, we can decompose $\boldsymbol{M}_{\boldsymbol{W}}$ as $\boldsymbol{M}_{\boldsymbol{Z}} - \boldsymbol{P}_{\boldsymbol{M}_{\boldsymbol{Z}}\boldsymbol{W}_1}$. This makes it clear that $\kappa \geq 1$, since the numerator of (5.95) cannot be smaller than the denominator. If equation (5.93) is just identified, then, by the order condition, $\boldsymbol{Y}$ and $\boldsymbol{W}_1$ have the same number of columns. In this case, it can be shown that the minimized value of κ is actually equal to 1; see Exercise 5.28.

In the context of 2SLS estimation, we saw in Section 10.6 that the Hansen-Sargan test can be used to test overidentifying restrictions. In the case of LIML estimation, it is easier to test these restrictions by a likelihood ratio test. As shown in Exercise 5.28, the maximized loglikelihood of the unconstrained model for which the overidentifying restrictions of (5.93) are relaxed is the same as expression (5.96) for the constrained model, but with $\kappa = 1$. Thus the LR statistic for testing the overidentifying restrictions, which is twice the difference between the unconstrained and constrained maxima, is simply equal to $n\log\hat{\kappa}$. This test statistic was first proposed by Anderson and Rubin (1950). Since there are $l-k$ overidentifying restrictions, the LR statistic is asymptotically distributed as $\chi^2(l-k)$.

K-Class Estimators

In equation (5.97), we have written the LIML estimating equations in the form of the estimating equations for a ***K*-class** estimator, following Theil (1961). The K-class is the set of estimators defined by the estimating equations (5.97) with an arbitrary scalar K replacing $\hat{\kappa}$. The LIML estimator is thus a K-class estimator with $K = \hat{\kappa}$. Similarly, the 2SLS estimator (5.63) is a K-class estimator with $K = 1$, and the OLS estimator is a K-class estimator with $K = 0$.

Numerous other K-class estimators have been proposed. It can be shown that, under standard regularity conditions, these estimators are consistent whenever the plim of K is 1. Thus 2SLS is consistent, and OLS is inconsistent. Since $n \log \hat{\kappa}$ is asymptotically distributed as $\chi^2(l-k)$ when the overidentifying restrictions are satisfied, it must be the case that $\operatorname{plim} \log \hat{\kappa} = 0$, which implies that $\operatorname{plim} \hat{\kappa} = 1$. It follows that LIML is asymptotically equivalent to 2SLS. In finite samples, however, the properties of LIML may be quite different from those of 2SLS. The strangest feature of the LIML estimator is that it has no finite moments. This implies that its density tends to have very thick tails, as readers are asked to illustrate in Exercise 5.32. However, if we measure bias by comparing the median of the estimator with the true value, the LIML estimator is generally much less biased than the 2SLS estimator.

Fuller (1977) has proposed a modified LIML estimator that sets K equal to $\hat{\kappa} - \alpha/(n-k)$, where α is a positive constant that must be chosen by the investigator. One good choice is $\alpha = 1$, since it yields estimates that are approximately unbiased. In contrast to the LIML estimator, which has no finite moments, Fuller's modified estimator has all moments finite provided the sample size is large enough. Mariano (2001) provides a recent summary of the finite-sample properties of LIML, 2SLS, and other K-class estimators.

Invariance of ML Estimators

One important feature of the FIML and LIML estimators is that they are invariant to any reparametrization of the model. This is actually a general property of all ML estimators, which was explored in Exercise 9.15. Since simultaneous equations systems can be parametrized in many different ways, this is a useful property for these estimators to have. It means that two investigators using the same data set must obtain the same estimates even if they employ different parametrizations.

As an example, consider the two-equation demand-supply model that was first discussed in Section 10.2:

$$q_t = \gamma_d p_t + \boldsymbol{X}_t^d \boldsymbol{\beta}_d + u_t^d \tag{5.100}$$

$$q_t = \gamma_s p_t + \boldsymbol{X}_t^s \boldsymbol{\beta}_s + u_t^s. \tag{5.101}$$

As the notation indicates, equation (5.100) is a demand function, and equation (5.101) is a supply function. In this system, p_t and q_t denote the price and

quantity of some commodity in period t, which may well be in logarithms, $\boldsymbol{X}_t^d$ and $\boldsymbol{X}_t^s$ are row vectors of exogenous or predetermined variables, $\boldsymbol{\beta}_d$ and $\boldsymbol{\beta}_s$ are the corresponding vectors of parameters, and γ_d and γ_s are the slopes of the demand and supply functions, which can be interpreted as elasticities if p_t and q_t are in logarithms.

Now suppose that we reparametrize the supply function as

$$p_t = \gamma_s' q_t + \boldsymbol{X}_t^s \boldsymbol{\beta}_s' + u_t'^s, \tag{5.102}$$

where $\gamma_s' = 1/\gamma_s$ and $\boldsymbol{\beta}_s' = -\boldsymbol{\beta}_s/\gamma_s$. The invariance property of maximum likelihood implies that, if we first use FIML to estimate the system consisting of equations (5.100) and (5.101) and then use it to estimate the system consisting of equations (5.100) and (5.102), we obtain exactly the same estimates of the parameters of equation (5.100). Moreover, the estimated parameters of equations (5.101) and (5.102) bear precisely the same relationship as the true parameters. That is,

$$\hat{\gamma}_s' = 1/\hat{\gamma}_s \quad \text{and} \quad \hat{\boldsymbol{\beta}}_s' = -\hat{\boldsymbol{\beta}}_s/\hat{\gamma}_s. \tag{5.103}$$

If we use LIML to estimate equations (5.101) and (5.102), the two sets of LIML estimates likewise satisfy conditions (5.103).

The invariance property of LIML and FIML is not shared by 2SLS, 3SLS, or any other GMM estimator. If, for example, we use 3SLS to estimate the two versions of this system of equations, the two sets of estimates do not satisfy conditions (5.103); see Exercise 5.31.

5.6 Nonlinear Simultaneous Equations Models

As we saw in Section 5.3, it is fairly straightforward to extend the SUR model so as to allow for the possibility of nonlinearity. However, additional complications can arise with nonlinear simultaneous equations models. With an SUR system, the right-hand sides of the several regressions do not depend on current endogenous variables, but this is not true of a simultaneous system. If endogenous variables enter nonlinearly in such a system, then, since it is not always possible to find solutions to nonlinear equations in closed form, it may be infeasible to set up a reduced form in which each endogenous variable is expressed as a function only of predetermined variables and parameters.

Feasible Efficient GMM

The easiest way to take account of all interesting cases is to work in terms of zero functions and treat the nonlinear simultaneous system by the methods we developed in Section 11.5 for nonlinear GMM. The main extension needed

for a simultaneous system is just that each elementary zero function depends, in general, on a vector of endogenous variables, rather than on just one.

Suppose that there are g equations that, for each observation, simultaneously determine g endogenous variables, and suppose further that these equations can be written as

$$f_{ti}(\boldsymbol{Y}_t, \boldsymbol{\theta}) = u_{ti}, \quad t = 1, \dots, n, \quad i = 1, \dots, g.$$

The functions $f_{ti}(\cdot)$ depend implicitly on predetermined explanatory variables. They are, in general, nonlinear functions of both the $1 \times g$ vector $\boldsymbol{Y}_t$ that contains the endogenous variables for observation t and the k–vector $\boldsymbol{\theta}$ of model parameters. The u_{ti} are disturbances with expectation zero. In some cases, we may be ready to assume that the u_{ti} satisfy the conditions (5.02) that we have imposed on the other models considered in this chapter.

It is clear that the f_{ti} are elementary zero functions. We may stack them in the way we stacked the dependent variables of an SUR system. First, we define the n–vectors $\boldsymbol{f}_i(\boldsymbol{Y}, \boldsymbol{\theta})$, $i = 1, \dots, g$, so that the t^{th} element of $\boldsymbol{f}_i(\boldsymbol{Y}, \boldsymbol{\theta})$ is $f_{ti}(\boldsymbol{Y}_t, \boldsymbol{\theta})$, where $\boldsymbol{Y}$ is the $n \times g$ matrix of which the t^{th} row is $\boldsymbol{Y}_t$. Then we stack the $\boldsymbol{f}_i$ vertically to construct the $gn \times 1$ vector $\boldsymbol{f}_\bullet(\boldsymbol{Y}, \boldsymbol{\theta})$. Under assumptions (5.02), the covariance matrix of this stacked vector is $\boldsymbol{\Sigma} \otimes \mathbf{I}_n$.

According to the theory developed in Section 11.5, the optimal instruments for efficient GMM are given in terms of the matrix $\bar{\boldsymbol{F}}(\boldsymbol{\theta})$ defined in equation (2.85). If, as before, we define the $g \times g$ matrix $\boldsymbol{\Psi}$ such that $\boldsymbol{\Psi}\boldsymbol{\Psi}^\top = \boldsymbol{\Sigma}^{-1}$, then the matrix $\boldsymbol{\Psi}$ of (2.85) becomes $\boldsymbol{\Psi} \otimes \mathbf{I}_n$ in the present case. The matrix $\boldsymbol{F}(\boldsymbol{\theta})$ of that equation becomes a $gn \times k$ matrix $\boldsymbol{F}_\bullet(\boldsymbol{Y}, \boldsymbol{\theta})$, of which the ti^{th} element is the derivative of the t^{th} element of $\boldsymbol{f}_\bullet(\boldsymbol{Y}, \boldsymbol{\theta})$ with respect to θ_i, the i^{th} element of $\boldsymbol{\theta}$. Under assumptions (5.02), the matrix $\bar{\boldsymbol{F}}_\bullet$ needed for the optimal estimating equations is just the $gn \times k$ matrix of which the t^{th} row is the expectation of the t^{th} row of $\boldsymbol{F}_\bullet$ conditional on all information predetermined at time t. The estimating equations we need correspond to equations (2.82). However, as discussed in the paragraph following (2.82), we must use $\bar{\boldsymbol{F}}_\bullet(\boldsymbol{\theta})$ instead of $\boldsymbol{F}_\bullet(\boldsymbol{\theta})$ in formulating the optimal instruments. We obtain

$$\bar{\boldsymbol{F}}_\bullet^\top(\boldsymbol{\theta})(\boldsymbol{\Sigma}^{-1} \otimes \mathbf{I}_n)\boldsymbol{f}_\bullet(\boldsymbol{Y}, \boldsymbol{\theta}) = \mathbf{0}. \tag{5.104}$$

Although the notation differs slightly, the only important difference between (2.82) and (5.104) is that the latter equations involve $\bar{\boldsymbol{F}}_\bullet(\boldsymbol{\theta})$ instead of $\boldsymbol{F}_\bullet(\boldsymbol{\theta})$. There is also no factor of n^{-1} in (5.104), an omission that evidently has no effect on the solution.

It is precisely in the construction of the matrix $\bar{\boldsymbol{F}}_\bullet$ that difficulties may arise. Since there may be no analytical expression for some or all of the endogenous variables, there may be no direct way of computing or even estimating $\bar{\boldsymbol{F}}_\bullet$. In that case, we may proceed as in Section 11.5 by selecting a set of $l \geq k$ instruments, that we group into the $n \times l$ matrix $\boldsymbol{W}$. We then replace the estimating equations (5.104) by

$$\boldsymbol{F}_\bullet^\top(\boldsymbol{Y}, \boldsymbol{\theta})(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{P}_{\boldsymbol{W}})\boldsymbol{f}_\bullet(\boldsymbol{Y}, \boldsymbol{\theta}) = \mathbf{0}, \tag{5.105}$$

which closely resemble equations (5.63) for the linear case. Equivalently, we may minimize the criterion function

$$\boldsymbol{f}_\bullet^\top(\boldsymbol{Y}, \boldsymbol{\theta})(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{P}_{\boldsymbol{W}})\boldsymbol{f}_\bullet(\boldsymbol{Y}, \boldsymbol{\theta}), \tag{5.106}$$

which is comparable to expression (5.64) for the linear case. The first-order conditions for minimizing (5.106) with respect to $\boldsymbol{\theta}$ are equivalent to the estimating equations (5.105).

If, as is usually the case, the matrix $\boldsymbol{\Sigma}$ is not known, then we must first obtain preliminary consistent estimates, say $\acute{\boldsymbol{\theta}}$. We might do this by solving the estimating equations (5.105) or minimizing the criterion function (5.106) with $\boldsymbol{\Sigma}$ replaced by an identity matrix. Alternatively, if cross-equation restrictions are not needed for identification, we might estimate each equation separately by the methods of Section 9.5. We can then use these preliminary estimates to form an estimate of $\boldsymbol{\Sigma}$ by the formula

$$\acute{\boldsymbol{\Sigma}} = \frac{1}{n} \begin{bmatrix} \boldsymbol{f}_1^\top(\boldsymbol{Y}, \acute{\boldsymbol{\theta}}) \\ \vdots \\ \boldsymbol{f}_g^\top(\boldsymbol{Y}, \acute{\boldsymbol{\theta}}) \end{bmatrix} [\, \boldsymbol{f}_1(\boldsymbol{Y}, \acute{\boldsymbol{\theta}}) \quad \cdots \quad \boldsymbol{f}_g(\boldsymbol{Y}, \acute{\boldsymbol{\theta}}) \,].$$

This estimate can then be used in either (5.105) or (5.106) to obtain more efficient estimates. We can either stop after one round or iterate to obtain continuously updated estimates.

The one-round procedure yields a generalization of the **nonlinear instrumental variables**, or **NLIV**, estimator $\hat{\boldsymbol{\theta}}_{\text{NLIV}}$, which we first encountered in Section 10.9. It was originally proposed by Jorgenson and Laffont (1974). In Exercise 5.33, readers are asked to write down the first-order conditions that define the estimator $\hat{\boldsymbol{\theta}}_{\text{NLIV}}$, along with the usual estimate of its covariance matrix.

The NLIV estimator is sometimes called **nonlinear three-stage least squares**, or **NL3SLS**. We prefer not to do so, because that name is quite misleading. For the reasons discussed in Section 8.9 in connection with nonlinear two-stage least squares, we never actually replace endogenous variables by their fitted values from reduced-form regressions. Moreover, there are really just two stages, the first in which preliminary consistent estimates are obtained, the second in which (5.105) or (5.106) is used with the estimated $\boldsymbol{\Sigma}$.

Nonlinear FIML Estimation

The other full-system estimation method that is widely used is **nonlinear FIML**. In order to derive the loglikelihood function, it is convenient to stack the vectors $\boldsymbol{f}_i(\boldsymbol{Y}, \boldsymbol{\theta})$ horizontally. Let $\boldsymbol{h}_t(\boldsymbol{Y}_t, \boldsymbol{\theta})$ be a $1 \times g$ row vector containing the elements $f_{t1}, \dots, f_{tg}$. Then the model to be estimated can be written as

$$\boldsymbol{h}_t(\boldsymbol{Y}_t, \boldsymbol{\theta}) = \boldsymbol{U}_t, \quad \boldsymbol{U}_t \sim \text{NID}(\mathbf{0}, \boldsymbol{\Sigma}). \tag{5.107}$$

The row vector $\boldsymbol{U}_t$ contains the disturbances u_{ti}, $i = 1, \ldots, g$, which are now assumed to be multivariate normal. In order to obtain the density of $\boldsymbol{Y}_t$, we start from the density of $\boldsymbol{U}_t$, replace $\boldsymbol{U}_t$ by $\boldsymbol{h}_t(\boldsymbol{Y}_t, \boldsymbol{\theta})$, and multiply by the Jacobian factor $|\det \boldsymbol{J}_t|$, where $\boldsymbol{J}_t \equiv \partial \boldsymbol{h}_t(\boldsymbol{\theta})/\partial \boldsymbol{Y}_t$ is the $g \times g$ matrix of derivatives of $\boldsymbol{h}_t$ with respect to the elements of $\boldsymbol{Y}_t$. The result is

$$(2\pi)^{-g/2} |\det \boldsymbol{J}_t| |\boldsymbol{\Sigma}|^{-1/2} \exp\Bigl(-\frac{1}{2} \boldsymbol{h}_t(\boldsymbol{Y}_t, \boldsymbol{\theta}) \boldsymbol{\Sigma}^{-1} \boldsymbol{h}_t^\top(\boldsymbol{Y}_t, \boldsymbol{\theta})\Bigr).$$

Taking the logarithm of this, summing it over all observations, and then concentrating the result with respect to $\boldsymbol{\Sigma}$, yields the concentrated loglikelihood function for the model (5.107):

$$-\frac{gn}{2}(\log 2\pi + 1) + \sum_{t=1}^{n} \log |\det \boldsymbol{J}_t| - \frac{n}{2} \log \Bigl| \frac{1}{n} \sum_{t=1}^{n} \boldsymbol{h}_t^\top(\boldsymbol{Y}_t, \boldsymbol{\theta}) \boldsymbol{h}_t(\boldsymbol{Y}_t, \boldsymbol{\theta}) \Bigr|.$$

The main difference between this function and its counterpart for the linear case, expression (5.90), is that the Jacobian matrices $\boldsymbol{J}_t$ are in general different for each observation. Evaluating all these determinants could well be expensive when n is large and g is not very small.

Another difference between the linear and nonlinear cases is that, in the latter, FIML and NLIV are not even asymptotically equivalent in general. In fact, if the disturbances are not normally distributed, the FIML estimator may actually be inconsistent; see Phillips (1982). If the disturbances are indeed normal, then, for the usual reasons, the FIML estimator is more efficient asymptotically than the NLIV estimator, although its efficiency may come at a price in terms of computational complexity. More detailed treatments of nonlinear FIML estimation may be found in Amemiya (1985, Chapter 8) and Gallant (1987, Chapter 6).

5.7 Final Remarks

Notation is a bugbear with multivariate regression models. These models can be written in many equivalent ways, and notation that is well suited to one estimation method may not be convenient for another. Once the notational hurdle has been crossed, we have seen that it is not excessively difficult to estimate multivariate regression models, including simultaneous equations models, using a variety of familiar techniques. All the procedures we have discussed use some combination of (feasible) generalized least squares, instrumental variables, GMM, and maximum likelihood. Except in the case of nonlinear simultaneous equations models, there is always a technique based on feasible GLS and/or instrumental variables that is asymptotically equivalent to maximum likelihood.

5.8 Appendix: Detailed Results on FIML

This appendix derives several results on FIML estimation that were too technical to include in the main text.

First-Order Conditions for FIML

For the purpose of obtaining the first-order conditions (5.86), it is convenient to write the loglikelihood function (5.83) in terms of the restricted reduced form (5.73). In the RRF, the $\boldsymbol{y}_i$ are stacked horizontally. However, if we are to use the same approach as for the SUR model, we must stack them vertically. The i^{th} column of (5.73) can be written as

$$\boldsymbol{y}_i = \boldsymbol{W}\boldsymbol{B}\boldsymbol{\gamma}^i + \boldsymbol{v}_i, \tag{5.108}$$

where the g–vector $\boldsymbol{\gamma}^i$ is the i^{th} column of $\boldsymbol{\Gamma}^{-1}$, and $\boldsymbol{v}_i$ is the i^{th} column of $\boldsymbol{V}$. Then equations (5.108) can be written as

$$\begin{aligned} \boldsymbol{y}_\bullet &= (\mathbf{I}_g \otimes \boldsymbol{W}\boldsymbol{B})\boldsymbol{\gamma}^\bullet + \boldsymbol{v}_\bullet \\ &= (\mathbf{I}_g \otimes \boldsymbol{W})\boldsymbol{\pi}_\bullet + \boldsymbol{v}_\bullet \\ &= \boldsymbol{W}_\bullet \boldsymbol{\pi}_\bullet + \boldsymbol{v}_\bullet. \end{aligned} \tag{5.109}$$

Here the g^2–vector $\boldsymbol{\gamma}^\bullet$ contains the $\boldsymbol{\gamma}^i$ stacked vertically, the gn–vector $\boldsymbol{v}_\bullet$ contains the $\boldsymbol{v}_i$ stacked vertically, the $gl \times gn$ matrix $\boldsymbol{W}_\bullet$ denotes $\mathbf{I}_g \otimes \boldsymbol{W}$, and the gl–vector $\boldsymbol{\pi}_\bullet$ contains the $\boldsymbol{\pi}_i$ stacked vertically. The $\boldsymbol{\pi}_i$ are the columns of the matrix $\boldsymbol{\Pi}$, defined here as $\boldsymbol{B}\boldsymbol{\Gamma}^{-1}$, as in the restricted reduced form.

By rewriting the last equation in (5.109) so that $\boldsymbol{v}_\bullet$ is a function of $\boldsymbol{y}_\bullet$, we obtain the transformation that gives $\boldsymbol{v}_\bullet$ in terms of $\boldsymbol{y}_\bullet$. Exactly as with the transformation (5.31), the determinant of the Jacobian of this transformation is 1. Thus, in order to obtain the joint density of $\boldsymbol{y}_\bullet$, we simply have to find the density of the vector $\boldsymbol{v}_\bullet$ and then replace $\boldsymbol{v}_\bullet$ by $\boldsymbol{y}_\bullet - \boldsymbol{W}_\bullet \boldsymbol{\pi}_\bullet$.

Since we have assumed that $\boldsymbol{v}_\bullet$ is multivariate normal, and we know that its expectation is a zero vector, the only thing we need to write down its density is its covariance matrix. Recall that $\boldsymbol{V} = \boldsymbol{U}\boldsymbol{\Gamma}^{-1}$, where $\boldsymbol{U}$ is the matrix of structural form disturbances. Thus

$$\boldsymbol{v}_i = \boldsymbol{U}\boldsymbol{\gamma}^i = \sum_{j=1}^{g} \boldsymbol{u}_j \gamma^{ji}, \quad i = 1, \ldots, g,$$

where γ^{ji} is the ji^{th} element of $\boldsymbol{\Gamma}^{-1}$. By stacking these equations vertically, it is not hard to see that

$$\boldsymbol{v}_\bullet = \bigl((\boldsymbol{\Gamma}^\top)^{-1} \otimes \mathbf{I}_n\bigr)\boldsymbol{u}_\bullet.$$

Since the covariance matrix of $\boldsymbol{u}_\bullet$ is assumed to be $\boldsymbol{\Sigma} \otimes \mathbf{I}_n$, it follows that the covariance matrix of $\boldsymbol{v}_\bullet$ can be written as

$$\begin{aligned}\operatorname{Var}(\boldsymbol{v}_\bullet) = \operatorname{E}(\boldsymbol{v}_\bullet \boldsymbol{v}_\bullet^\top) &= \big((\boldsymbol{\Gamma}^\top)^{-1} \otimes \mathbf{I}_n\big)(\boldsymbol{\Sigma} \otimes \mathbf{I}_n)(\boldsymbol{\Gamma}^{-1} \otimes \mathbf{I}_n) \\ &= (\boldsymbol{\Gamma}^\top)^{-1} \boldsymbol{\Sigma} \boldsymbol{\Gamma}^{-1} \otimes \mathbf{I}_n.\end{aligned}$$

For some of the following calculations, it will be convenient to denote the matrix $(\boldsymbol{\Gamma}^\top)^{-1}\boldsymbol{\Sigma}\boldsymbol{\Gamma}^{-1}$ by $\boldsymbol{\Omega}$.

Using this notation, the density of $\boldsymbol{y}_\bullet$ is $(2\pi)^{-gn/2}$ times

$$|\boldsymbol{\Omega} \otimes \mathbf{I}_n|^{-1/2} \exp\big(-\tfrac{1}{2}(\boldsymbol{y}_\bullet - \boldsymbol{W}_\bullet \boldsymbol{\pi}_\bullet)^\top (\boldsymbol{\Omega}^{-1} \otimes \mathbf{I}_n)(\boldsymbol{y}_\bullet - \boldsymbol{W}_\bullet \boldsymbol{\pi}_\bullet)\big).$$

This may be compared with (5.32), the analogous expression for a linear SUR system. It follows that the loglikelihood function for the linear simultaneous equations model can be written as

$$-\tfrac{gn}{2} \log 2\pi - \tfrac{n}{2} \log |\boldsymbol{\Omega}| - \tfrac{1}{2}(\boldsymbol{y}_\bullet - \boldsymbol{W}_\bullet \boldsymbol{\pi}_\bullet)^\top (\boldsymbol{\Omega}^{-1} \otimes \mathbf{I}_n)(\boldsymbol{y}_\bullet - \boldsymbol{W}_\bullet \boldsymbol{\pi}_\bullet). \tag{5.110}$$

This expression is deceptively simple, because the vector $\boldsymbol{\pi}_\bullet$ depends in a complicated way on the vector of structural parameters $\boldsymbol{\beta}_\bullet$. However, since (5.110) depends on $\boldsymbol{\Omega}$ in precisely the same way in which expression (5.33), the loglikelihood function for a linear SUR system, depends on $\boldsymbol{\Sigma}$, the ML estimator of $\boldsymbol{\Omega}$ must have exactly the same form as (5.36).

It is of interest to compare the loglikelihood functions (5.110) and (5.83). A little algebra, which is detailed in Exercise 5.23, shows that

$$(\boldsymbol{\Gamma}^\top \otimes \mathbf{I}_n)(\boldsymbol{y}_\bullet - \boldsymbol{W}_\bullet \boldsymbol{\pi}_\bullet) = \boldsymbol{y}_\bullet - \boldsymbol{X}_\bullet \boldsymbol{\beta}_\bullet, \tag{5.111}$$

which is the vector of residuals from the structural form expressed as in (5.58) in stacked form. Thus the quadratic form that appears in (5.110) can also be written as

$$(\boldsymbol{y}_\bullet - \boldsymbol{X}_\bullet \boldsymbol{\beta}_\bullet)^\top (\boldsymbol{\Sigma}^{-1} \otimes \mathbf{I}_n)(\boldsymbol{y}_\bullet - \boldsymbol{X}_\bullet \boldsymbol{\beta}_\bullet). \tag{5.112}$$

Now consider the second term in (5.110). By the definition of $\boldsymbol{\Omega}$ and the properties of determinants, this term is

$$-\tfrac{n}{2} \log |\boldsymbol{\Omega}| = -\tfrac{n}{2} \log\big(|\det \boldsymbol{\Gamma}|^{-2} |\boldsymbol{\Sigma}|\big) = n \log |\det \boldsymbol{\Gamma}| - \tfrac{n}{2} \log |\boldsymbol{\Sigma}|. \tag{5.113}$$

If we start with (5.110) and replace the quadratic form by expression (5.112) and the second term by the rightmost expression in (5.113), we obtain the loglikelihood function (5.83). Thus we see that these two ways of writing the loglikelihood function are indeed equivalent.

In order to write down the ML estimator of $\boldsymbol{\Omega}$, we define the $n \times g$ matrix $\boldsymbol{V}(\boldsymbol{\beta}_\bullet)$ to have i^{th} column $\boldsymbol{y}_i - \boldsymbol{W}\boldsymbol{\beta}\boldsymbol{\gamma}^i$, which is just the i^{th} block of the vector $\boldsymbol{y}_\bullet - \boldsymbol{W}_\bullet \boldsymbol{\pi}_\bullet$. It follows that $\boldsymbol{V}(\boldsymbol{\beta}_\bullet) = \boldsymbol{Y} - \boldsymbol{W}\boldsymbol{B}\boldsymbol{\Gamma}^{-1}$. When evaluated at the

ML estimator $\hat{\boldsymbol{\beta}}_\bullet^{\text{ML}}$, this is just the ML estimator of the disturbances of the RRF (5.73). By analogy with (5.36), we find that

$$\hat{\boldsymbol{\Omega}}_{\text{ML}} = \tfrac{1}{n} \boldsymbol{V}^\top(\hat{\boldsymbol{\beta}}_\bullet^{\text{ML}}) \boldsymbol{V}(\hat{\boldsymbol{\beta}}_\bullet^{\text{ML}}).$$

We are entitled to write $\boldsymbol{V}$ as a function of $\boldsymbol{\beta}_\bullet$ here because, as we saw when defining the RRF, the matrices $\boldsymbol{B}$ and $\boldsymbol{\Gamma}$ on which (5.110) depends through the vector $\boldsymbol{W}_\bullet \boldsymbol{\pi}_\bullet$ are uniquely determined by the structural parameters in the vector $\boldsymbol{\beta}_\bullet$. Conversely, if we obtain ML estimators of the matrices $\boldsymbol{B}$ and $\boldsymbol{\Gamma}$, these uniquely determine the ML estimator of $\boldsymbol{\beta}_\bullet$.

Only the last term of the loglikelihood function (5.110) depends on $\boldsymbol{B}$ and $\boldsymbol{\Gamma}$. Therefore, conditional on $\boldsymbol{\Sigma}$, the maximization of (5.110) reduces as usual to the minimization of a quadratic form, which in this case is

$$(\boldsymbol{y}_\bullet - \boldsymbol{W}_\bullet \boldsymbol{\pi}_\bullet)^\top (\boldsymbol{\Omega}^{-1} \otimes \mathbf{I}_n)(\boldsymbol{y}_\bullet - \boldsymbol{W}_\bullet \boldsymbol{\pi}_\bullet). \tag{5.114}$$

From the definition of $\boldsymbol{\Omega}$ and the properties (5.08) of Kronecker products, we observe that $\boldsymbol{\Omega}^{-1} \otimes \mathbf{I}_n = (\boldsymbol{\Gamma} \otimes \mathbf{I}_n)(\boldsymbol{\Sigma}^{-1} \otimes \mathbf{I}_n)(\boldsymbol{\Gamma}^\top \otimes \mathbf{I}_n)$.

From the first equation in (5.109), we can see that the quadratic form (5.114) can also be written as

$$\big(\boldsymbol{y}_\bullet - (\mathbf{I}_g \otimes \boldsymbol{W}\boldsymbol{B})\boldsymbol{\gamma}^\bullet\big)^\top (\boldsymbol{\Omega}^{-1} \otimes \mathbf{I}_n)\big(\boldsymbol{y}_\bullet - (\mathbf{I}_g \otimes \boldsymbol{W}\boldsymbol{B})\boldsymbol{\gamma}^\bullet\big).$$

From this expression, we see that the partial derivatives of (5.110) with respect to the g^2 elements of $\boldsymbol{\gamma}^\bullet$ are the g^2 elements of the vector

$$(\mathbf{I}_g \otimes \boldsymbol{B}^\top \boldsymbol{W}^\top)(\boldsymbol{\Omega}^{-1} \otimes \mathbf{I}_n)\big(\boldsymbol{y}_\bullet - (\mathbf{I}_g \otimes \boldsymbol{W}\boldsymbol{B})\boldsymbol{\gamma}^\bullet\big). \tag{5.115}$$

The conditions we seek are not given by simply equating the elements of this vector to zero, because many elements of the matrix $\boldsymbol{\Gamma}$ are restricted to be equal to 0 or 1. The restrictions translate into complicated conditions on the elements of $\boldsymbol{\gamma}^\bullet$ which, fortunately, we need not concern ourselves with. Rather, we compute the derivatives of $\boldsymbol{\gamma}^\bullet$ with respect to any element γ_{ij} of $\boldsymbol{\Gamma}$ which is *not* restricted, and then use the chain rule to obtain the derivative of (5.110) with respect to γ_{ij}. We can then quite properly equate the resulting derivative to zero in order to obtain a first-order condition.

The vectors that are stacked in $\boldsymbol{\gamma}^\bullet$ are the columns of $\boldsymbol{\Gamma}^{-1}$, and it is therefore not hard to see that $(\boldsymbol{\Gamma}^\top \otimes \mathbf{I}_g)\boldsymbol{\gamma}^\bullet$ is a vector of g^2 components that are all either 0 or 1, and thus independent of the elements of $\boldsymbol{\Gamma}$. Differentiating this relation with respect to γ_{ij} thus gives

$$(\boldsymbol{E}_{ji} \otimes \mathbf{I}_g)\boldsymbol{\gamma}^\bullet + (\boldsymbol{\Gamma}^\top \otimes \mathbf{I}_g)\frac{\partial \boldsymbol{\gamma}^\bullet}{\partial \gamma_{ij}} = \mathbf{0},$$

where $\boldsymbol{E}_{ji}$ is a $g \times g$ matrix of which the ji^{th} element is 1 and the other elements are 0. Consequently, the derivative of $\boldsymbol{\gamma}^\bullet$ with respect to γ_{ij} is the g^2–vector

$$-\big((\boldsymbol{\Gamma}^\top)^{-1} \otimes \mathbf{I}_g\big)(\boldsymbol{E}_{ji} \otimes \mathbf{I}_g)\boldsymbol{\gamma}^\bullet.$$

The derivative of expression (5.110) with respect to γ_{ij} is the scalar product of this vector with the vector (5.115), that is, the negative of

$$\begin{aligned}&\boldsymbol{\gamma}^{\bullet\top}(\boldsymbol{E}_{ij}\otimes\mathbf{I}_g)(\boldsymbol{\Gamma}^{-1}\otimes\mathbf{I}_g)(\mathbf{I}_g\otimes\boldsymbol{B}^\top\boldsymbol{W}^\top)(\boldsymbol{\Omega}^{-1}\otimes\mathbf{I}_n)(\boldsymbol{y}_\bullet-\boldsymbol{W}_\bullet\boldsymbol{\pi}_\bullet)\\&\quad=\boldsymbol{\gamma}^{\bullet\top}(\boldsymbol{E}_{ij}\otimes\mathbf{I}_g)(\boldsymbol{\Gamma}^{-1}\otimes\boldsymbol{B}^\top\boldsymbol{W}^\top)(\boldsymbol{\Gamma}\otimes\mathbf{I}_n)(\boldsymbol{\Sigma}^{-1}\otimes\mathbf{I}_n)(\boldsymbol{y}_\bullet-\boldsymbol{X}_\bullet\boldsymbol{\beta}_\bullet)\\&\quad=\boldsymbol{\gamma}^{\bullet\top}(\boldsymbol{E}_{ij}\otimes\boldsymbol{B}^\top\boldsymbol{W}^\top)(\boldsymbol{\Sigma}^{-1}\otimes\mathbf{I}_n)(\boldsymbol{y}_\bullet-\boldsymbol{X}_\bullet\boldsymbol{\beta}_\bullet).\end{aligned}\tag{5.116}$$

The second line above makes use of the expression of $\boldsymbol{\Omega}$ in terms of $\boldsymbol{\Gamma}$ and $\boldsymbol{\Sigma}$, and of the result (5.111). It is straightforward to see that (5.116) is one row of the left-hand side of (5.86), which therefore contains all the first-order conditions with respect to the unrestricted elements of $\boldsymbol{\Gamma}$.

Eigenvalues and Eigenvectors

Before we can discuss LIML estimation, we need to introduce a few more concepts of matrix algebra. A scalar λ is said to be an **eigenvalue** (also called a **characteristic root** or a **latent root**) of a matrix $\boldsymbol{A}$ if there exists a nonzero vector $\boldsymbol{\xi}$ such that

$$\boldsymbol{A}\boldsymbol{\xi}=\lambda\boldsymbol{\xi}.\tag{5.117}$$

Thus the action of $\boldsymbol{A}$ on $\boldsymbol{\xi}$ produces a vector with the same direction as $\boldsymbol{\xi}$, but a different length unless $\lambda=1$. The vector $\boldsymbol{\xi}$ is called the **eigenvector** that corresponds to the eigenvalue λ. Although these concepts are defined quite generally, we will restrict our attention to the eigenvalues and eigenvectors of real symmetric matrices.

Equation (5.117) implies that

$$(\boldsymbol{A}-\lambda\mathbf{I})\boldsymbol{\xi}=\mathbf{0},\tag{5.118}$$

from which we conclude that the matrix $\boldsymbol{A}-\lambda\mathbf{I}$ is singular. Its determinant, $|\boldsymbol{A}-\lambda\mathbf{I}|$, is therefore equal to zero. It can be shown that this determinant is a polynomial in λ. The degree of the polynomial is m if $\boldsymbol{A}$ is $m\times m$. The fundamental theorem of algebra tells us that such a polynomial has m complex roots, say $\lambda_1,\dots,\lambda_m$. To each λ_i there must correspond an eigenvector $\boldsymbol{\xi}_i$. This eigenvector is determined only up to a scale factor, because if $\boldsymbol{\xi}_i$ is an eigenvector corresponding to λ_i, then so is $\alpha\boldsymbol{\xi}_i$ for any nonzero scalar α. The eigenvector $\boldsymbol{\xi}_i$ does not necessarily have real elements if λ_i itself is not real.

If $\boldsymbol{A}$ is a real symmetric matrix, it can be shown that the eigenvalues λ_i are all real and that the eigenvectors can be chosen to be real as well. If $\boldsymbol{A}$ is also a positive definite matrix, then all its eigenvalues are positive. This follows from the facts that $\boldsymbol{\xi}^\top\boldsymbol{A}\boldsymbol{\xi}=\lambda\boldsymbol{\xi}^\top\boldsymbol{\xi}$ and that both $\boldsymbol{\xi}^\top\boldsymbol{\xi}$ and $\boldsymbol{\xi}^\top\boldsymbol{A}\boldsymbol{\xi}$ must be positive scalars when $\boldsymbol{A}$ is positive definite.

The eigenvectors of a real symmetric matrix can be chosen to be mutually orthogonal. Consider any two eigenvectors $\boldsymbol{\xi}_i$ and $\boldsymbol{\xi}_j$ that correspond to two distinct eigenvalues λ_i and λ_j. We see that

$$\lambda_i\boldsymbol{\xi}_j^\top\boldsymbol{\xi}_i=\boldsymbol{\xi}_j^\top\boldsymbol{A}\boldsymbol{\xi}_i=(\boldsymbol{A}\boldsymbol{\xi}_j)^\top\boldsymbol{\xi}_i=\lambda_j\boldsymbol{\xi}_j^\top\boldsymbol{\xi}_i.\tag{5.119}$$

But this is impossible unless $\boldsymbol{\xi}_j^\top\boldsymbol{\xi}_i=0$. Thus we conclude that $\boldsymbol{\xi}_i$ and $\boldsymbol{\xi}_j$ are necessarily orthogonal. If not all the eigenvalues are distinct, then two (or more) eigenvectors may correspond to one and the same eigenvalue. When that happens, these two eigenvectors span a space that is orthogonal to all other eigenvalues by the reasoning just given. Since any linear combination of the two eigenvectors is also an eigenvector corresponding to the one eigenvalue, we may choose an orthogonal set of them. Thus, whether or not all the eigenvalues are distinct, eigenvectors may be chosen to be **orthonormal**, by which we mean that they are mutually orthogonal and each has norm equal to 1. When the eigenvectors of a real symmetric matrix $\boldsymbol{A}$ are chosen in this way, they provide an **orthonormal basis** for $\mathcal{S}(\boldsymbol{A})$.

Let $\boldsymbol{\Xi}\equiv[\,\boldsymbol{\xi}_1\ \cdots\ \boldsymbol{\xi}_m\,]$ be a matrix the columns of which are an orthonormal set of eigenvectors of $\boldsymbol{A}$, corresponding to the eigenvalues λ_i, $i=1,\dots,m$. Then we can write the eigenvalue relationship (5.117) for all the eigenvalues at once as

$$\boldsymbol{A}\boldsymbol{\Xi}=\boldsymbol{\Xi}\boldsymbol{\Lambda},\tag{5.120}$$

where $\boldsymbol{\Lambda}$ is a diagonal matrix with λ_i as its i^{th} diagonal element. The i^{th} column of $\boldsymbol{A}\boldsymbol{\Xi}$ is $\boldsymbol{A}\boldsymbol{\xi}_i$, and the i^{th} column of $\boldsymbol{\Xi}\boldsymbol{\Lambda}$ is $\lambda_i\boldsymbol{\xi}_i$. Since the columns of the matrix $\boldsymbol{\Xi}$ are orthonormal, we find that $\boldsymbol{\Xi}^\top\boldsymbol{\Xi}=\mathbf{I}$, which implies that $\boldsymbol{\Xi}^\top=\boldsymbol{\Xi}^{-1}$. A matrix with this property is said to be an **orthogonal matrix**. Postmultiplying (5.120) by $\boldsymbol{\Xi}^\top$ gives

$$\boldsymbol{A}=\boldsymbol{\Xi}\boldsymbol{\Lambda}\boldsymbol{\Xi}^\top.\tag{5.121}$$

Taking determinants of both sides of (5.121), we obtain

$$|\boldsymbol{A}|=|\boldsymbol{\Xi}||\boldsymbol{\Xi}^\top||\boldsymbol{\Lambda}|=|\boldsymbol{\Xi}||\boldsymbol{\Xi}^{-1}||\boldsymbol{\Lambda}|=|\boldsymbol{\Lambda}|=\prod_{i=1}^m\lambda_i,$$

from which we may deduce the important result that the determinant of a symmetric matrix is the product of its eigenvalues. In fact, this result holds for nonsymmetric matrices as well.

LIML Estimation

Consider the system of equations consisting of the structural equation (5.93) and the reduced form equations (5.94). The matrix of coefficients of the endogenous variables in this system of equations is

$$\begin{bmatrix}1&\mathbf{0}\\-\boldsymbol{\beta}_2&\mathbf{I}\end{bmatrix}.$$

Because this matrix is triangular, its determinant is simply the product of the elements on the principal diagonal, which is 1. Therefore, there is no Jacobian term in the loglikelihood function (5.83) for such a system, and the ML estimates may be obtained by minimizing the determinant

$$\left|(\boldsymbol{Y} - \boldsymbol{W}\boldsymbol{B}\boldsymbol{\Gamma}^{-1})^\top(\boldsymbol{Y} - \boldsymbol{W}\boldsymbol{B}\boldsymbol{\Gamma}^{-1})\right| = \left|(\boldsymbol{Y}\boldsymbol{\Gamma} - \boldsymbol{W}\boldsymbol{B})^\top(\boldsymbol{Y}\boldsymbol{\Gamma} - \boldsymbol{W}\boldsymbol{B})\right|.$$

It can, with considerable effort, be shown that minimizing this determinant is equivalent to minimizing the ratio

$$\kappa \equiv \frac{(\boldsymbol{y} - \boldsymbol{Y}\boldsymbol{\beta}_2)^\top \boldsymbol{M}_{\boldsymbol{Z}}(\boldsymbol{y} - \boldsymbol{Y}\boldsymbol{\beta}_2)}{(\boldsymbol{y} - \boldsymbol{Y}\boldsymbol{\beta}_2)^\top \boldsymbol{M}_{\boldsymbol{W}}(\boldsymbol{y} - \boldsymbol{Y}\boldsymbol{\beta}_2)} = \frac{\boldsymbol{\gamma}^\top \boldsymbol{Y}_*^\top \boldsymbol{M}_{\boldsymbol{Z}} \boldsymbol{Y}_* \boldsymbol{\gamma}}{\boldsymbol{\gamma}^\top \boldsymbol{Y}_*^\top \boldsymbol{M}_{\boldsymbol{W}} \boldsymbol{Y}_* \boldsymbol{\gamma}} \tag{5.122}$$

with respect to $\boldsymbol{\beta}_2$, where $\boldsymbol{Y}_* \equiv [\boldsymbol{y} \;\; \boldsymbol{Y}]$ and $\boldsymbol{\gamma} = [1 \vdots -\boldsymbol{\beta}_2]$; see Davidson and MacKinnon (1993, Chapter 18).

It is possible to minimize κ without doing any sort of nonlinear optimization. The first-order conditions obtained by differentiating the middle expression in (5.122) with respect to $\boldsymbol{\beta}_2$ can be rearranged as

$$\boldsymbol{Y}^\top(\boldsymbol{M}_{\boldsymbol{Z}} - \hat{\kappa}\boldsymbol{M}_{\boldsymbol{W}})\boldsymbol{Y}_*\boldsymbol{\gamma} = \boldsymbol{0}, \tag{5.123}$$

where $\hat{\kappa}$ is defined by (5.122) with the minimizing value of $\boldsymbol{\beta}_2$. From (5.122), we see that the expression

$$\boldsymbol{\gamma}^\top \boldsymbol{Y}_*^\top(\boldsymbol{M}_{\boldsymbol{Z}} - \hat{\kappa}\boldsymbol{M}_{\boldsymbol{W}})\boldsymbol{Y}_*\boldsymbol{\gamma} = \boldsymbol{y}^\top(\boldsymbol{M}_{\boldsymbol{Z}} - \hat{\kappa}\boldsymbol{M}_{\boldsymbol{W}})\boldsymbol{Y}_*\boldsymbol{\gamma} - \boldsymbol{\beta}_2^\top \boldsymbol{Y}^\top(\boldsymbol{M}_{\boldsymbol{Z}} - \hat{\kappa}\boldsymbol{M}_{\boldsymbol{W}})\boldsymbol{Y}_*\boldsymbol{\gamma}$$

is equal to zero. By (5.123), the second term on the right-hand side is zero for any $\boldsymbol{\beta}_2$. Therefore, the first term must also be zero, which implies that

$$\boldsymbol{Y}_*^\top(\boldsymbol{M}_{\boldsymbol{Z}} - \hat{\kappa}\boldsymbol{M}_{\boldsymbol{W}})\boldsymbol{Y}_*\boldsymbol{\gamma} = \boldsymbol{0}.$$

If we premultiply this equation by $(\boldsymbol{Y}_*^\top \boldsymbol{M}_{\boldsymbol{W}} \boldsymbol{Y}_*)^{-1/2}$ and insert that factor multiplied by its inverse before $\boldsymbol{\gamma}$, we see, after some rearrangement, that

$$\big((\boldsymbol{Y}_*^\top \boldsymbol{M}_{\boldsymbol{W}} \boldsymbol{Y}_*)^{-1/2} \boldsymbol{Y}_*^\top \boldsymbol{M}_{\boldsymbol{Z}} \boldsymbol{Y}_* (\boldsymbol{Y}_*^\top \boldsymbol{M}_{\boldsymbol{W}} \boldsymbol{Y}_*)^{-1/2} - \hat{\kappa}\boldsymbol{\mathrm{I}}\big)\boldsymbol{\gamma}^* = \boldsymbol{0},$$

where $\boldsymbol{\gamma}^* \equiv (\boldsymbol{Y}_*^\top \boldsymbol{M}_{\boldsymbol{W}} \boldsymbol{Y}_*)^{1/2}\boldsymbol{\gamma}$. This set of first-order conditions now has the form of a standard eigenvalue-eigenvector problem for a real symmetric matrix; see equation (5.118). Thus it is clear that $\hat{\kappa}$ is an eigenvalue of the matrix

$$(\boldsymbol{Y}_*^\top \boldsymbol{M}_{\boldsymbol{W}} \boldsymbol{Y}_*)^{-1/2} \boldsymbol{Y}_*^\top \boldsymbol{M}_{\boldsymbol{Z}} \boldsymbol{Y}_* (\boldsymbol{Y}_*^\top \boldsymbol{M}_{\boldsymbol{W}} \boldsymbol{Y}_*)^{-1/2}, \tag{5.124}$$

which depends only on observable data, and not on unknown parameters. In fact, $\hat{\kappa}$ must be the smallest eigenvalue, because it is the smallest possible value of the ratio (5.122). Given $\hat{\kappa}$, we can use equations (5.98) to compute the LIML estimates. It is worthy of note that, if there is only one endogenous variable in the matrix $\boldsymbol{Y}$, then the determinantal equation that determines the eigenvalues of (5.124) is just a quadratic equation, of which the smaller root is $\hat{\kappa}$, which can therefore be expressed in this case as a closed-form function of the data.

5.9 Exercises

⋆5.1 Show that the $gn \times gn$ covariance matrix $\boldsymbol{\Sigma}_\bullet$ defined in equation (5.07) is positive definite if and only if the $g \times g$ matrix $\boldsymbol{\Sigma}$ used to define it is positive definite.

⋆5.2 Prove the first result of equations (5.08) for an arbitrary $p \times q$ matrix $\boldsymbol{A}$ and an arbitrary $r \times s$ matrix $\boldsymbol{B}$. Prove the second result for $\boldsymbol{A}$ and $\boldsymbol{B}$ as above, and for $\boldsymbol{C}$ and $\boldsymbol{D}$ arbitrary $q \times t$ and $s \times u$ matrices, respectively. Prove the third result in (5.08) for an arbitrary nonsingular $p \times p$ matrix $\boldsymbol{A}$ and nonsingular $r \times r$ matrix $\boldsymbol{B}$.

Give details of the interchanges of rows and columns needed to convert $\boldsymbol{A} \otimes \boldsymbol{B}$ into $\boldsymbol{B} \otimes \boldsymbol{A}$, where $\boldsymbol{A}$ is $p \times q$ and $\boldsymbol{B}$ is $r \times s$.

⋆5.3 If $\boldsymbol{B}$ is positive definite, show that $\mathbf{I} \otimes \boldsymbol{B}$ is also positive definite, where $\mathbf{I}$ is an identity matrix of arbitrary dimension. What about $\boldsymbol{B} \otimes \mathbf{I}$? If $\boldsymbol{A}$ is another positive definite matrix, is it the case that $\boldsymbol{B} \otimes \boldsymbol{A}$ is positive definite?

5.4 Show explicitly that expression (5.06) provides the OLS estimates of the parameters of all the equations of the SUR system.

5.5 Show explicitly that expression (5.14) for the GLS estimator of the parameters of an SUR system follows from the estimating equations (5.13).

5.6 Show that, for any two vectors $\boldsymbol{a}_1$ and $\boldsymbol{a}_2$ in E^2, the quantity $\|\boldsymbol{a}_1\|^2\|\boldsymbol{M}_1\boldsymbol{a}_2\|^2$, where $\boldsymbol{M}_1$ is the orthogonal projection on to the orthogonal complement of $\boldsymbol{a}_1$ in E^2, is equal to the square of $a_{11}a_{22} - a_{12}a_{21}$, where a_{ij} denotes the i^{th} element of $\boldsymbol{a}_j$, for $i, j = 1, 2$.

5.7 Using only the properties of determinants listed at the end of the subsection on determinants in Section 5.2, show that the determinant of a positive definite matrix $\boldsymbol{B}$ is positive. (**Hint:** write $\boldsymbol{B} = \boldsymbol{A}\boldsymbol{A}^\top$.) Show further that, if $\boldsymbol{B}$ is positive semidefinite, without being positive definite, then its determinant must be zero.

⋆5.8 Suppose that m independent random variables, z_i, each of which is distributed as $\mathrm{N}(0, 1)$, are grouped into an m–vector $\boldsymbol{z}$. Let $\boldsymbol{x} = \boldsymbol{\mu} + \boldsymbol{A}\boldsymbol{z}$, where $\boldsymbol{\mu}$ is an m–vector and $\boldsymbol{A}$ is a nonsingular $m \times m$ matrix, and let $\boldsymbol{\Omega} \equiv \boldsymbol{A}\boldsymbol{A}^\top$. Show that the mean of the vector $\boldsymbol{x}$ is $\boldsymbol{\mu}$ and its covariance matrix is $\boldsymbol{\Omega}$. Then show that the density of $\boldsymbol{x}$ is

$$(2\pi)^{-m/2}|\boldsymbol{\Omega}|^{-1/2}\exp\big(-\tfrac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\Omega}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})\big). \tag{5.125}$$

This extends the result of Exercise 4.F5 for the bivariate normal density to the multivariate normal density. **Hints:** Remember that the joint density of m independent random variables is equal to the product of their densities, and use the result (5.29).

5.9 Consider a univariate linear regression model in which the regressors may include lags of the dependent variable. Let $\boldsymbol{y}$ and $\boldsymbol{u}$ denote, respectively, the vectors of observations on the dependent variable and the disturbances, and assume that $\boldsymbol{u} \sim \mathrm{N}(\boldsymbol{0}, \sigma^2\mathbf{I}_n)$. Show that, even though the Jacobian matrix of the transformation (5.31) is not an identity matrix, the determinant of the Jacobian is unity. Then write down the loglikelihood function for this model.

For simplicity, assume that any lagged values of the dependent variable prior to the sample period are observed.

⋆**5.10** Consider a multivariate linear regression model of the form (5.28) in which the regressors may include lags of the dependent variables and the disturbances are normally distributed. By ordering the data appropriately, show that the determinant of the Jacobian of the transformation (5.31) is equal to unity. Then explain why this implies that the loglikelihood function, conditional on pre-sample observations, can be written as (5.33).

5.11 Let $\boldsymbol{A}$ and $\boldsymbol{B}$ be square matrices, of dimensions $p \times p$ and $q \times q$, respectively. Use the properties of determinants given in Section 5.2 to show that the determinant of $\boldsymbol{A} \otimes \boldsymbol{B}$ is equal to that of $\boldsymbol{B} \otimes \boldsymbol{A}$.

Use this result, along with any other needed properties of determinants given in Section 5.2, to show that the determinant of $\boldsymbol{\Sigma} \otimes \mathbf{I}_n$ is $|\boldsymbol{\Sigma}|^n$.

5.12 Verify that the moment conditions (5.47) and the estimating equations (5.48) are equivalent. Show also that expressions (5.49) and (5.50) for the covariance matrix estimator for the nonlinear SUR model are equivalent. Explain how (5.50) is related to the covariance matrix estimator (5.15) that corresponds to it in the linear case.

⋆**5.13** The **linear expenditure system** is a system of demand equations that can be written as

$$s_i = \frac{\gamma_i p_i}{E} + \alpha_i \Big(\frac{E - \sum_{j=1}^{m+1} p_j \gamma_j}{E} \Big). \tag{5.126}$$

Here, s_i, for $i = 1, \dots, m$, is the share of total expenditure E spent on commodity i conditional on E and the prices p_i, for $i = 1, \dots, m+1$. The equation indexed by $i = m+1$ is omitted as redundant, because the sum of the expenditure shares spent on all commodities is necessarily equal to 1. The model parameters are the α_i, $i = 1, \dots, m$, the γ_i, $i = 1, \dots, m+1$, and the $m \times m$ contemporaneous covariance matrix $\boldsymbol{\Sigma}$.

Express the system (5.126) as a linear SUR system by use of a suitable nonlinear reparametrization. The equations of the resulting system must be subject to a set of cross-equation restrictions. Express these restrictions in terms of the new parameters, and then set up a GNR in the manner of Section 5.3 that allows one to obtain restricted estimates of the α_i and γ_i.

5.14 Show that the estimating equations (5.63) are equivalent to the estimating equations (5.61).

5.15 Show that the estimating equations (5.68) are equivalent to the equations that correspond to the equation-by-equation IV (or 2SLS) estimator for all the equations of the system jointly.

⋆**5.16** The $k \times k$ matrix $\boldsymbol{X}_\bullet^\top (\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{P}_{\boldsymbol{W}}) \boldsymbol{X}_\bullet$ given in expression (5.69) is positive semidefinite by construction. Show this property explicitly by expressing the matrix in the form $\boldsymbol{A}^\top \boldsymbol{A}$, where $\boldsymbol{A}$ is a matrix with k columns and at least k rows that should depend on a $g \times g$ nonsingular matrix $\boldsymbol{\Psi}$ which satisfies the relation $\boldsymbol{\Psi}\boldsymbol{\Psi}^\top = \boldsymbol{\Sigma}^{-1}$.

Show that a positive semidefinite matrix expressed in the form $\boldsymbol{A}^\top \boldsymbol{A}$ is positive definite if and only if $\boldsymbol{A}$ has full column rank. In the present case, the matrix $\boldsymbol{A}$ fails to have full column rank if and only if there exists a k–vector $\boldsymbol{\beta}$, different from zero, such that $\boldsymbol{A}\boldsymbol{\beta} = \mathbf{0}$. Since $k = \sum_{i=1}^g k_i$, we may write the vector

$\boldsymbol{\beta}$ as $[\boldsymbol{\beta}_1 \vdots \dots \vdots \boldsymbol{\beta}_g]$, where $\boldsymbol{\beta}_i$ is a k_i–vector for $i = 1, \dots, g$. Show that there exists a nonzero $\boldsymbol{\beta}$ such that $\boldsymbol{A}\boldsymbol{\beta} = \mathbf{0}$ if and only if, for at least one i, there is a nonzero $\boldsymbol{\beta}_i$ such that $\boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{X}_i \boldsymbol{\beta}_i = \mathbf{0}$, that is, if $\boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{X}_i$ does not have full column rank.

Show that, if $\boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{X}_i$ has full column rank, then there exists a unique solution of the estimating equations (5.63) for the parameters $\boldsymbol{\beta}_i$ of equation i.

5.17 Consider the linear simultaneous equations model

$$\begin{aligned} y_{t1} &= \beta_{11} + \beta_{21} z_{t2} + \beta_{31} z_{t3} + \gamma_{21} y_{t2} + u_{t1} \\ y_{t2} &= \beta_{12} + \beta_{22} z_{t2} + \beta_{42} z_{t4} + \beta_{52} z_{t5} + \gamma_{12} y_{t1} + u_{t2}. \end{aligned} \tag{5.127}$$

If this model is written in the matrix notation of (5.71), precisely what are the matrices $\boldsymbol{B}$ and $\boldsymbol{\Gamma}$ equal to?

5.18 Demonstrate that, if each equation in the linear simultaneous equations model (5.57) is just identified, in the sense that the order condition for identification is satisfied as an equality, then the number of restrictions on the elements of the matrices $\boldsymbol{\Gamma}$ and $\boldsymbol{B}$ of the restricted reduced form (5.73) is exactly g^2. In other words, demonstrate that the restricted and unrestricted reduced forms have the same number of parameters in this case.

5.19 Show that all terms that depend on the matrix $\boldsymbol{V}$ of disturbances in the finite-sample expression for $n^{-1} \boldsymbol{X}_1^\top \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{X}_1$ obtained from equation (5.79) tend to zero as $n \to \infty$.

5.20 Consider the following $p \times q$ partitioned matrix

$$\boldsymbol{A} = \begin{bmatrix} \mathbf{I}_m & \boldsymbol{A}_{12} \\ \mathbf{O} & \boldsymbol{A}_{22} \end{bmatrix},$$

where $m < \min(p, q)$. Show that $\boldsymbol{A}$ has full column rank if and only if $\boldsymbol{A}_{22}$ has full column rank. **Hint:** In order to do so, one can show that the existence of a nonzero q–vector $\boldsymbol{x}$ such that $\boldsymbol{A}\boldsymbol{x} = \mathbf{0}$ implies the existence of a nonzero $(q-m)$–vector $\boldsymbol{x}_2$ such that $\boldsymbol{A}_{22}\boldsymbol{x}_2 = \mathbf{0}$, and vice versa.

⋆**5.21** Consider equation (5.75), the first structural equation of the linear simultaneous system (5.71), with the variables ordered as described in the discussion of the asymptotic identification of this equation. Let the matrices $\boldsymbol{\Gamma}$ and $\boldsymbol{B}$ of the full system (5.71) be partitioned as follows:

$$\boldsymbol{B} = \begin{bmatrix} \boldsymbol{\beta}_{11} & \boldsymbol{B}_{12} \\ \mathbf{0} & \boldsymbol{B}_{22} \end{bmatrix} \quad \text{and} \quad \boldsymbol{\Gamma} = \begin{bmatrix} 1 & \boldsymbol{\Gamma}_{02} \\ -\boldsymbol{\beta}_{21} & \boldsymbol{\Gamma}_{12} \\ \mathbf{0} & \boldsymbol{\Gamma}_{22} \end{bmatrix},$$

where $\boldsymbol{\beta}_{11}$ is a k_{11}–vector, $\boldsymbol{B}_{12}$ and $\boldsymbol{B}_{22}$ are, respectively, $k_{11} \times (g-1)$ and $(l - k_{11}) \times (g-1)$ matrices, $\boldsymbol{\beta}_{21}$ is a k_{21}–vector, and $\boldsymbol{\Gamma}_{02}$, $\boldsymbol{\Gamma}_{12}$, and $\boldsymbol{\Gamma}_{22}$ are, respectively, $1 \times (g-1)$, $k_{21} \times (g-1)$, and $(g - k_{21} - 1) \times (g-1)$ matrices. Check that the restrictions imposed in this partitioning correspond correctly to the structure of (5.75).

Let $\boldsymbol{\Gamma}^{-1}$ be partitioned as

$$\boldsymbol{\Gamma}^{-1} = \begin{bmatrix} \gamma^{00} & \boldsymbol{\Gamma}^{01} & \boldsymbol{\Gamma}^{02} \\ \boldsymbol{\gamma}^{10} & \boldsymbol{\Gamma}^{11} & \boldsymbol{\Gamma}^{12} \end{bmatrix},$$

where the rows of $\boldsymbol{\Gamma}^{-1}$ are partitioned in the same pattern as the columns of $\boldsymbol{\Gamma}$, and vice versa. Show that $\boldsymbol{\Gamma}_{22}\boldsymbol{\Gamma}^{12}$ is an identity matrix, and that $\boldsymbol{\Gamma}_{22}\boldsymbol{\Gamma}^{11}$ is a zero matrix, and specify the dimensions of these matrices. Show also that the matrix $[\boldsymbol{\Gamma}^{11} \;\; \boldsymbol{\Gamma}^{12}]$ is square and nonsingular.

⋆**5.22** It was shown in Section 5.4 that the rank condition for the asymptotic identification of equation (5.75) is that the $(l - k_{11}) \times k_{21}$ matrix $\boldsymbol{\Pi}_{21}$ of the unrestricted reduced form (5.76) should have full column rank. Show that, in terms of the structural parameters, $\boldsymbol{\Pi}_{21}$ is equal to $\boldsymbol{B}_{22}\boldsymbol{\Gamma}^{11}$. Then consider the matrix

$$\begin{bmatrix} \boldsymbol{\Gamma}_{22} \\ \boldsymbol{B}_{22} \end{bmatrix}, \tag{5.128}$$

and show, by postmultiplying it by the nonsingular matrix $[\boldsymbol{\Gamma}^{11} \;\; \boldsymbol{\Gamma}^{12}]$, that it is of full column rank $g - 1$ if and only if $\boldsymbol{B}_{22}\boldsymbol{\Gamma}^{11}$ is of full column rank. Conclude that the rank condition for the asymptotic identification of (5.75) is that (5.128) should have full column rank.

⋆**5.23** Consider the expression $(\boldsymbol{\Gamma}^\top \otimes \mathbf{I}_n)\boldsymbol{y}_\bullet$, in the notation of Section 5.5. Show that it is equal to a gn–vector that can be written as

$$\begin{bmatrix} \boldsymbol{Y}\boldsymbol{\gamma}_1 \\ \vdots \\ \boldsymbol{Y}\boldsymbol{\gamma}_m \end{bmatrix},$$

where $\boldsymbol{\gamma}_i$, $i = 1, \ldots, g$, is the i^{th} column of $\boldsymbol{\Gamma}$.

Show similarly that $(\boldsymbol{\Gamma}^\top \otimes \mathbf{I}_n)(\mathbf{I}_g \otimes \boldsymbol{W}\boldsymbol{B})\boldsymbol{\gamma}^\bullet$ is equal to a gn–vector that can be written as

$$\begin{bmatrix} \boldsymbol{W}\boldsymbol{b}_1 \\ \vdots \\ \boldsymbol{W}\boldsymbol{b}_m \end{bmatrix},$$

where $\boldsymbol{b}_i$ is the i^{th} column of $\boldsymbol{B}$.

Using these results, demonstrate that $(\boldsymbol{\Gamma}^\top \otimes \mathbf{I}_n)(\boldsymbol{y}_\bullet - (\mathbf{I}_g \otimes \boldsymbol{W}\boldsymbol{B})\boldsymbol{\gamma}^\bullet)$ is equal to $\boldsymbol{y}_\bullet - \boldsymbol{X}_\bullet\boldsymbol{\beta}_\bullet$. Explain why this proves the result (5.111).

5.24 By expressing the loglikelihood function (5.110) for the linear simultaneous equations model in terms of $\boldsymbol{\Sigma}$ rather than $\boldsymbol{\Omega}$, show that concentrating the resulting function with respect to $\boldsymbol{\Sigma}$ yields the concentrated loglikelihood function (5.90).

5.25 Write down the concentrated loglikelihood function for the restricted reduced form (5.73) as a special case of (5.53). Then show that this concentrated loglikelihood function is identical to expression (5.90).

5.26 In the model (5.127), what is the identification status of each of the two equations? How would your answer change if an additional regressor, x_{t6}, were added to the first equation only, to the second equation only, or to both equations?

⋆**5.27** Consider the linear simultaneous system of equations (5.93) and (5.94). Write down the estimating equations for the 3SLS estimator for the system, and show that they define the same estimator of the parameters of (5.93) as the IV estimator applied to that equation alone with instruments $\boldsymbol{W}$.

State and prove the analogous result for an SUR system in which only one equation is overidentified.

⋆**5.28** In the just-identified case of LIML estimation, for which, in the notation of (5.94), the number of excluded instruments in the matrix $\boldsymbol{W}_1$ is equal to the number of included endogenous variables in the matrix $\boldsymbol{Y}$, show that the minimized value of the ratio κ given by (5.95) is equal to the global minimum of 1. Show further that the vector of estimates $\hat{\boldsymbol{\beta}}_2$ that attains this minimum is the IV, or 2SLS, estimator of $\boldsymbol{\beta}_2$ for equation (5.93) with instruments $\boldsymbol{W}$.

In the overidentified case of LIML estimation, explicitly formulate a model containing the model consisting of (5.93) and (5.94) as a special case, with the overidentifying restrictions relaxed. Show that the maximized loglikelihood for this unconstrained model is the same function of the data as for the constrained model, but with $\hat{\kappa}$ replaced by 1.

5.29 Consider the demand-supply model

$$\begin{aligned} q_t &= \beta_{11} + \beta_{21}x_{t2} + \beta_{31}x_{t3} + \gamma_{21}p_t + u_{t1} \\ q_t &= \beta_{12} + \beta_{42}x_{t4} + \beta_{52}x_{t5} + \gamma_{22}p_t + u_{t2}, \end{aligned} \tag{5.129}$$

where q_t is the log of quantity, p_t is the log of price, x_{t2} is the log of income, x_{t3} is a dummy variable that accounts for regular demand shifts, and x_{t4} and x_{t5} are the prices of inputs. Thus the first equation of (5.129) is a demand function and the second equation is a supply function.

For this model, precisely what is the vector $\boldsymbol{\beta}_\bullet$ that was introduced in equation (5.58)? What are the matrices $\boldsymbol{B}$ and $\boldsymbol{\Gamma}$ that were introduced in equation (5.71)? How many overidentifying restrictions are there?

5.30 The file **demand-supply.data** contains 120 observations generated by the model (5.129). Estimate this model by 2SLS, LIML, 3SLS, and FIML. In each case, test the overidentifying restrictions, either for each equation individually or for the whole system, as appropriate.

5.31 The second equation of (5.129) can be rewritten as

$$p_t = \beta'_{12} + \beta'_{42}x_{t4} + \beta'_{52}x_{t5} + \gamma'_{12}q_t + u'_{t2}. \tag{5.130}$$

Estimate the system that consists of the first equation of (5.129) and equation (5.130) by 3SLS and FIML. What is the relationship between the FIML estimates of this system and the FIML estimates of (5.129)? What is the relationship between the two sets of 3SLS estimates?

5.32 Consider the system

$$\boldsymbol{y}_1 = \beta + \gamma\boldsymbol{y}_2 + \boldsymbol{u}, \quad \boldsymbol{y}_2 = \boldsymbol{W}\boldsymbol{\pi}_1 + \boldsymbol{v}, \tag{5.131}$$

in which the first equation is the only structural equation and the first column of $\boldsymbol{W}$ is a vector of 1s. For sample size $n = 25$, and for $l = 2, 4, 6, 8$, generate $l - 1$ additional instrumental variables as independent drawings from $\mathrm{N}(0, 1)$. Generate the endogenous variables $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ using the DGP given by (5.131) with $\beta = 1$ and $\gamma = 1$, $\boldsymbol{\pi}_1$ an l–vector with every element equal to 1, and the 2×2 contemporaneous covariance matrix $\boldsymbol{\Sigma}$ such that the diagonal elements are equal to 4, and the off-diagonal elements to 2. Estimate the parameters β and γ using both IV (2SLS) and LIML.

Repeat the exercise many times and plot the empirical distributions of the two estimators of γ. How do their properties vary with the degree of overidentification?

5.33 What are the first-order conditions for minimizing expression (5.106), the NLIV criterion function? What is the usual estimate of the covariance matrix of the NLIV estimator?

References

Albert, A., and J. A. Anderson (1984). "On the existence of maximum likelihood estimates in logistic regression models," *Biometrika*, **71**, 1–10.

Amemiya, T. (1973c). "Regression analysis when the dependent variable is truncated normal," *Econometrica*, **41**, 997–1016.

Amemiya, T. (1985). *Advanced Econometrics*, Cambridge, Mass., Harvard University Press.

Andrews, D. W. K. (1991). "Heteroskedasticity and autocorrelation consistent covariance matrix estimation," *Econometrica*, **59**, 817–58.

Andrews, D. W. K., and J. C. Monahan (1992). "An improved heteroskedasticity and autocorrelation consistent covariance matrix estimator," *Econometrica*, **60**, 953–66.

Andrews, D. W. K. (1997). "A stopping rule for the computation of generalized method of moments estimators," *Econometrica*, **65**, 913–31.

Bard, Y. (1974). *Nonlinear Parameter Estimation*, New York, Academic Press.

Bates, D. M., and D. G. Watts (1988). *Nonlinear Regression Analysis and Its Applications*, New York, John Wiley & Sons.

Becker, W. E., and P. E. Kennedy (1992). "A graphical exposition of the ordered probit," *Econometric Theory*, **8**, 127–31.

Beach, C. M., and J. G. MacKinnon (1978). "A maximum likelihood procedure for regression with autocorrelated errors," *Econometrica*, **46**, 51–58.

Berndt, E. R., B. H. Hall, R. E. Hall, and J. A. Hausman (1974). "Estimation and inference in nonlinear structural models," *Annals of Economic and Social Measurement*, **3**, 653–65.

Cameron, A. C., and P. K. Trivedi (1986). "Econometric models based on count data: Comparisons and applications of some estimators and tests," (3.06), **1**, 29–53.

Cameron, A. C., and P. K. Trivedi (1998). *Regression Analysis of Count Data*, Cambridge, Cambridge University Press.

Cameron, A. C., and P. K. Trivedi (2001). "Essentials of count data regression," Ch. 15 in *A Companion to Theoretical Econometrics*, ed. B. Baltagi, Oxford, Blackwell Publishers, 331–48.

Chesher, A., and M. Irish (1987). "Residual analysis in the grouped and censored normal linear model," *Journal of Econometrics*, **34**, 33–61.

Cox, D. R. (1972). "Regression models and life tables" (with discussion), *Journal of the Royal Statistical Society*, Series B, **34**, 187–220.

Cox, D. R., and D. V. Hinkley (1974). *Theoretical Statistics*, London, Chapman and Hall.

Cox, D. R., and D. Oakes (1984). *Analysis of Survival Data*, London, Chapman and Hall.

Cragg, J. G. (1983). "More efficient estimation in the presence of heteroskedasticity of unknown form," *Econometrica*, **51**, 751–63.

Cramér, H. (1946). *Mathematical Methods of Statistics*, Princeton, Princeton University Press.

Davidson, R., and J. G. MacKinnon (1984a). "Model specification tests based on artificial linear regressions," *International Economic Review*, **25**, 485–502.

Davidson, R., and J. G. MacKinnon (1984b). "Convenient specification tests for logit and probit models," *Journal of Econometrics*, **25**, 241–62.

Davidson, R., and J. G. MacKinnon (1985b). "Testing linear and loglinear regressions against Box-Cox alternatives," *Canadian Journal of Economics*, **18**, 499–517.

Davidson, R., and J. G. MacKinnon (1987). "Implicit alternatives and the local power of test statistics," *Econometrica*, **55**, 1305–29.

Davidson, R., and J. G. MacKinnon (1993). *Estimation and Inference in Econometrics*, New York, Oxford University Press.

Davidson, R., and J. G. MacKinnon (1999a). "Bootstrap testing in nonlinear models," (2.63), **40**, 487–508.

Davidson, R., and J. G. MacKinnon (2000). "Bootstrap tests: How many bootstraps?" *Econometric Reviews*, **19**, 55–68.

Davidson, R., and J. G. MacKinnon (2004). *Econometric Theory and Methods*, New York, Oxford University Press. (ETM)

Donald, S. G., and H. J. Paarsch (1993). "Piecewise pseudo-maximum likelihood estimation in empirical models of auctions," (2.63), **34**, 121–48.

Donald, S. G., and H. J. Paarsch (1996). "Identification, estimation, and testing in parametric empirical models of auctions within the independent private values paradigm," *Econometric Theory*, **12**, 517–67.

Dorsey, R. E., and W. J. Mayer (1995). "Genetic algorithms for estimation problems with multiple optima, nondifferentiability, and other irregular features," *Journal of Business and Economic Statistics*, **13**, 53–66.

Engle, R. F. (1984). "Wald, likelihood ratio and Lagrange multiplier tests in econometrics," Ch. 13 in *Handbook of Econometrics*, Vol. 2, ed. Z. Griliches and M. D. Intriligator, Amsterdam, North-Holland, 775–826.

Durbin, J. (1960). "Estimation of parameters in time-series regression models," *Journal of the Royal Statistical Society*, Series B, **22**, 139–53.

Fisher, R. A. (1925). "The theory of statistical estimation," *Proceedings of the Cambridge Philosophical Society*, **22**, 700–25.

Gill, P. E., W. Murray, and M. H. Wright (1981). *Practical Optimization*, New York, Academic Press.

Godambe, V. P. (1960). "An optimum property of regular maximum likelihood estimation," *Annals of Mathematical Statistics* **31**, 1208–11.

Godambe, V. P., and M. E. Thompson (1978). "Some aspects of the theory of estimating equations," *Journal of Statistical Planning and Inference*, **2**, 95–104.

Godfrey, L. G., M. McAleer, and C. R. McKenzie (1988). "Variable addition and Lagrange Multiplier tests for linear and logarithmic regression models," *Review of Economics and Statistics*, **70**, 492–503.

Godfrey, L. G., and M. R. Wickens (1981). "Testing linear and log-linear regressions for functional form," *Review of Economic Studies*, **48**, 487–96.

Goffe, W. L., G. D. Ferrier, and J. Rogers (1994). "Global optimization of statistical functions with simulated annealing," *Journal of Econometrics*, **60**, 65–99.

Gouriéroux, C., and J. Jasiak (2001). "Durations," Ch. 21 in *A Companion to Theoretical Econometrics*, ed. B. Baltagi, Oxford, Blackwell Publishers, 444–65.

Gouriéroux, C., and A. Monfort (1996). *Simulation Based Econometric Methods*, Oxford, Oxford University Press.

Gouriéroux, C., A. Monfort, and A. Trognon (1984). "Pseudo-maximum likelihood methods: Theory," *Econometrica*, **52**, 681–700.

Greene, W. H. (1981). "Sample selection bias as a specification error: Comment," *Econometrica*, **49**, 795–98.

Gregory, A. W., and M. R. Veall (1985). "On formulating Wald tests for nonlinear restrictions," *Econometrica*, **53**, 1465–68.

Gregory, A. W., and M. R. Veall (1987). "Formulating Wald tests of the restrictions implied by the rational expectations hypothesis," (3.06), **2**, 61–68.

Hajivassiliou, V. A., and P. A. Ruud (1994). "Classical estimation methods for LDV models using simulation," Ch. 40 in *Handbook of Econometrics*, Vol. 4, ed. R. F. Engle and D. L. McFadden, Amsterdam, Elsevier, 2383–441.

Hamilton, J. D. (1994). *Time Series Analysis*, Princeton, Princeton University Press.

Hansen, L. P. (1982). "Large sample properties of generalized method of moments estimators," *Econometrica*, **50**, 1029–54.

Hansen, L. P., J. Heaton, and A. Yaron (1996). "Finite-sample properties of some alternative GMM estimators," *Journal of Business and Economic Statistics*, **14** 262–80.

Hausman, J. A., A. W. Lo, and A. C. MacKinlay (1992). "An ordered probit analysis of transaction stock prices," *Journal of Financial Economics*, **31**, 319–79.

Hausman, J. A., and D. L. McFadden (1984). "A specification test for the multinomial logit model," *Econometrica*, **52**, 1219–40.

Heckman, J. J. (1976). "The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models," *Annals of Economic and Social Measurement*, **5**, 475–92.

Heckman, J. J. (1979). "Sample selection bias as a specification error," *Econometrica*, **47**, 153–61.

Heckman, J. J., and B. Singer (1984). "A method for minimizing the impact of distributional assumptions in econometric models for duration data," *Econometrica*, **52**, 271–320.

Johnson, N. L., S. Kotz, and N. Balakrishnan (1994). *Continuous Univariate Distributions*, Vol. 1, second edition, New York, John Wiley & Sons.

Karush, W. (1939). "Minima of Functions of Several Variables with Inequalities as Side Constraints", (M.Sc. thesis), Dept. of Mathematics, Univ. of Chicago, Chicago, Illinois.

Kiefer, N. M. (1978). "Discrete parameter variation: Efficient estimation of a switching regression model," *Econometrica*, **46**, 427–34.

Kiefer, N. M. (1988). "Economic duration data and hazard functions," *Journal of Economic Literature*, **26**, 646–79.

Kiviet, J. F. (1986). "On the rigour of some misspecification tests for modelling dynamic relationships," *Review of Economic Studies*, **53**, 241–61.

Kuhn, H. W. and Tucker, A. W. (1951). "Nonlinear programming", in *Proceedings of 2nd Berkeley Symposium*, Berkeley: University of California Press. pp. 481-92.

Lafontaine, F., and K. J. White (1986). "Obtaining any Wald statistic you want," *Economics Letters*, **21**, 35–40.

Lancaster, T. (1990). *The Econometric Analysis of Transition Data*, Cambridge, Cambridge University Press.

Lee, L.-F. (1982). "Some approaches to the correction of selectivity bias," *Review of Economic Studies*, **49**, 355–72.

Lee, L.-F. (1986). "Specification test for Poisson regression models," (2.63), **27**, 689–706.

MacKinnon, J. G., and L. Magee (1990). "Transforming the dependent variable in regression models," (2.63), **31**, 315–39.

MacKinnon, J. G., and A. A. Smith, Jr. (1998). "Approximate bias correction in econometrics," *Journal of Econometrics*, **85**, 205–30.

Maddala, G. S., and A. Flores-Lagunes (2001). "Qualitative response models," Ch. 17 in *A Companion to Theoretical Econometrics*, ed. B. Baltagi, Oxford, Blackwell Publishers, 366–82.

McCullough, B. D. (1999). "Econometric software reliability: EViews, LIMDEP, SHAZAM and TSP," (3.06), **14**, 191–202.

McCullough, B. D. (2003). "Some details of nonlinear estimation," Ch. 8 in *Numerical Methods in Statistical Computing for the Social Sciences*, ed. M. Altman, J. Gill, and M. P. McDonald, New York, Wiley, 245–67.

McCullough, B. D., and H. D. Vinod (2003). "Verifying the solution from a nonlinear solver: A case study," *American Economic Review*, **93**, 873–92.

McFadden, D. L. (1984). "Econometric analysis of qualitative response models," Ch. 24 in *Handbook of Econometrics*, Vol. 2, ed. Z. Griliches and M. D. Intriligator, Amsterdam, North-Holland, 1395–457.

McFadden, D. L. (1987). "Regression-based specification tests for the multinomial logit model," *Journal of Econometrics*, **34**, 63–82.

Mullahy, J. (1997). "Heterogeneity, excess zeros, and the structure of count data models," *Journal of Applied Econometrics*, **12**, 337–50.

Neumann, G. R. (1999). "Search models and duration data," Ch. 7 in *Handbook of Applied Econometrics*, Vol. II, ed. H. M. Pesaran and P. Schmidt, Oxford, Blackwell.

Newey, W. K., and D. L. McFadden (1994). "Large sample estimation and hypothesis testing," Ch. 36 in *Handbook of Econometrics*, Vol. 4, ed. F. F. Engle and D. L. McFadden, Amsterdam, North-Holland, 2111–245.

Newey, W. K., and K. D. West (1987). "A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix," *Econometrica*, **55**, 703–8.

Newey, W. K., and K. D. West (1994). "Automatic lag selection in covariance matrix estimation," *Review of Economic Studies*, **61**, 631–53.

Olsen, R. J. (1978). "Note on the uniqueness of the maximum likelihood estimator of the tobit model," *Econometrica*, **46**, 1211–15.

Orme, C. D. (1995). "On the use of artificial regressions in certain microeconometric models," *Econometric Theory*, **11**, 290–305.

Orme, C. D., and P. A. Ruud (2002). "On the uniqueness of the maximum likelihood estimator," *Economics Letters*, **75**, 209–17.

Pagan, A. R., and F. Vella (1989). "Diagnostic tests for models based on individual data: A survey," *Journal of Applied Econometrics*, **4**, S29–59.

Pollak, R. A., and T. J. Wales (1991). "The likelihood dominance criterion: A new approach to model selection," *Journal of Econometrics*, **47**, 227–42.

Press, W. H., S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery (1992a). *Numerical Recipes in C*, second edition, Cambridge, Cambridge University Press.

Press, W. H., S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery (1992b). *Numerical Recipes in Fortran*, second edition, Cambridge, Cambridge University Press.

Quandt, R. E. (1983). "Computational problems and methods," Ch. 12 in *Handbook of Econometrics*, Vol. 1, ed. Z. Griliches and M. D. Intriligator, Amsterdam, North-Holland, 699–771.

Ramsey, J. B. (1969). "Tests for specification errors in classical linear least-squares regression analysis," *Journal of the Royal Statistical Society*, Series B, **31**, 350–71.

Rao, C. R. (1945). "Information and accuracy attainable in estimation of statistical parameters," *Bulletin of the Calcutta Mathematical Society*, **37**, 81–91.

Sargan, J. D. (1964). "Wages and prices in the United Kingdom: A study in econometric methodology," in *Econometric Analysis for National Economic Planning*, ed. P. E. Hart, G. Mills, and J. K. Whitaker, London, Butterworths; reprinted in *Quantitative Economics and Econometric Analysis*, ed. K. F. Wallis and D. F. Hendry (1984), Oxford, Basil Blackwell.

Seber, G. A. F., and C. J. Wild (1989). *Nonlinear Regression*, New York, John Wiley & Sons.

Smith, R. J. (1989). "On the use of distributional mis-specification checks in limited dependent variable models," *Economic Journal*, **99**, 178–92.

Stuart, A., J. K. Ord, and S. Arnold (1998). *Kendall's Advanced Theory of Statistics*, Vol. 2A: *Classical Inference and the Linear Model*, sixth edition, London, Edward Arnold.

Terza, J. (1985). "Ordinal probit: A generalization," *Communications in Statistics*, **14**, 1–12.

Tobin, J. (1958). "Estimation of relationships for limited dependent variables," *Econometrica*, **26**, 24–36.

van den Berg, G. J. (2001). "Duration models: Specification, identification, and multiple durations," Ch. 55 in *Handbook of Econometrics*, Vol. 5, ed. J. J. Heckman and E. E. Leamer, Amsterdam, North-Holland.

Veall, M. R. (1990). "Testing for a global maximum in an econometric context," *Econometrica*, **58**, 1459–65.

Wald, A. (1943). "Tests of statistical hypotheses concerning several parameters when the number of observations is large," *Transactions of the American Mathematical Society*, **54**, 426–82.

West, K. D., and D. M. Wilcox (1996). "A comparison of alternative instrumental variables estimators of a dynamic linear model," *Journal of Business and Economic Statistics*, **14**, 281–93.

White, H. (1982). "Maximum likelihood estimation of misspecified models," *Econometrica*, **50**, 1–26.

White, H. (2000). *Asymptotic Theory for Econometricians*, revised edition, Orlando, Academic Press.

White, H., and I. Domowitz (1984). "Nonlinear regression with dependent observations," *Econometrica*, **52**, 143–61.

Wooldridge, J. M. (1999). "Quasi-likelihood methods for count data," Ch. 8 in *Handbook of Applied Econometrics*, Vol. II, ed. H. M. Pesaran and P. Schmidt, Oxford, Blackwell.

Author Index

Subject Index