Lecture Notes in Econometrics for Finance (PiF, introductory PhD course at UNISG)

Paul Söderlind[1]

17 September 2022

[1]University of St. Gallen. *Address:* s/bf-HSG, Unterer Graben 21, CH-9000 St. Gallen, Switzerland. *E-mail:* Paul.Soderlind@unisg.ch.

Contents

Warning: a few of the tables and figures are reused in later chapters. This can mess up the references, so that the text refers to a table/figure in another chapter. No worries: it is really the same table/figure. Still, I promise to fix this some day.

Chapter 1

Review of Statistics

Reference: Verbeek (2012) Appendix B
More advanced material is denoted by a star (*). It is not required reading.

1.1 Random Variables and Distributions

1.1.1 The Distribution of a Random Variable

A univariate distribution of a random variable x describes the probability of different values. If $f(x)$ is the probability density function (pdf), then the probability that x is between A and B is calculated as the area under the density function from A to B

$$\Pr(A < x \leq B) = \int_A^B f(x)dx. \tag{1.1}$$

See Figure 1.1 for illustrations of normal (gaussian) distributions.

Remark 1.1 *If $x \sim N(\mu, \sigma^2)$, then the probability density function is*

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}.$$

This is a bell-shaped curve centred on the mean μ and where the standard deviation σ determines the "width" of the curve.

The probability that $x \leq B$ (that is, $-\infty < x \leq B$) is measured by the *cumulative distribution function*, cdf(B). For instance, if x has a $N(0,1)$ distribution, then $\Pr(x \leq -1.645) = 0.05$ and $\Pr(x \leq 0) = 0.5$. Once you have the cdf, you can calculate the probability of $B < x$ as $1 - \text{cdf}(B)$. See Figure 1.2 for an illustration.

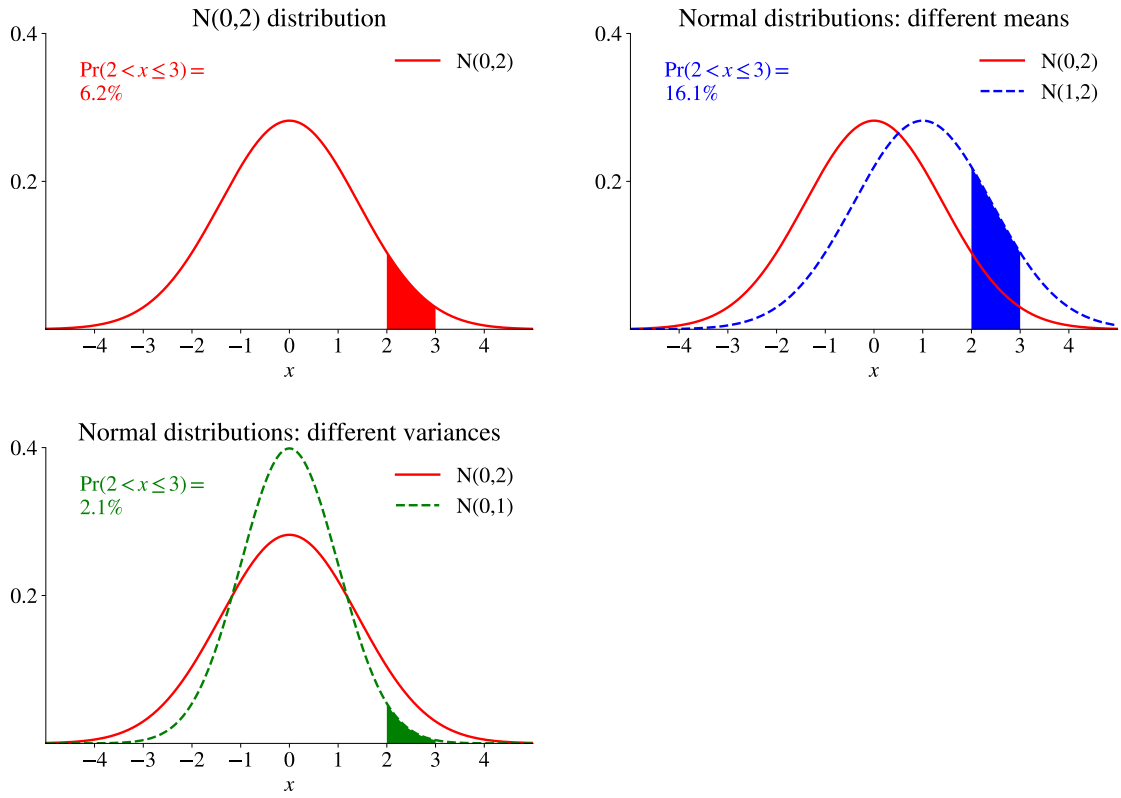


Figure 1.1: A few different normal distributions

If we invert the cdf, then we get the *quantiles* of the random variable. For instance, the 0.05th quantile of a $N(0,1)$ variable is -1.645, while the 0.5th quantile (also called the median) is 0.

1.1.2 The Joint Distribution of Several Random Variables

A bivariate distribution of the random variables x and y contains the same information as the two respective univariate distributions, but also information on how x and y are related. Let $h(x, y)$ be the joint density function, then the probability that x is between A and B and y is between C and D is calculated as the volume under the surface of the density function

$$\Pr(A < x \leq B \text{ and } C < y \leq D) = \int_A^B \int_C^D h(x, y) dy dx. \tag{1.2}$$

See Figure 1.3 for an example.

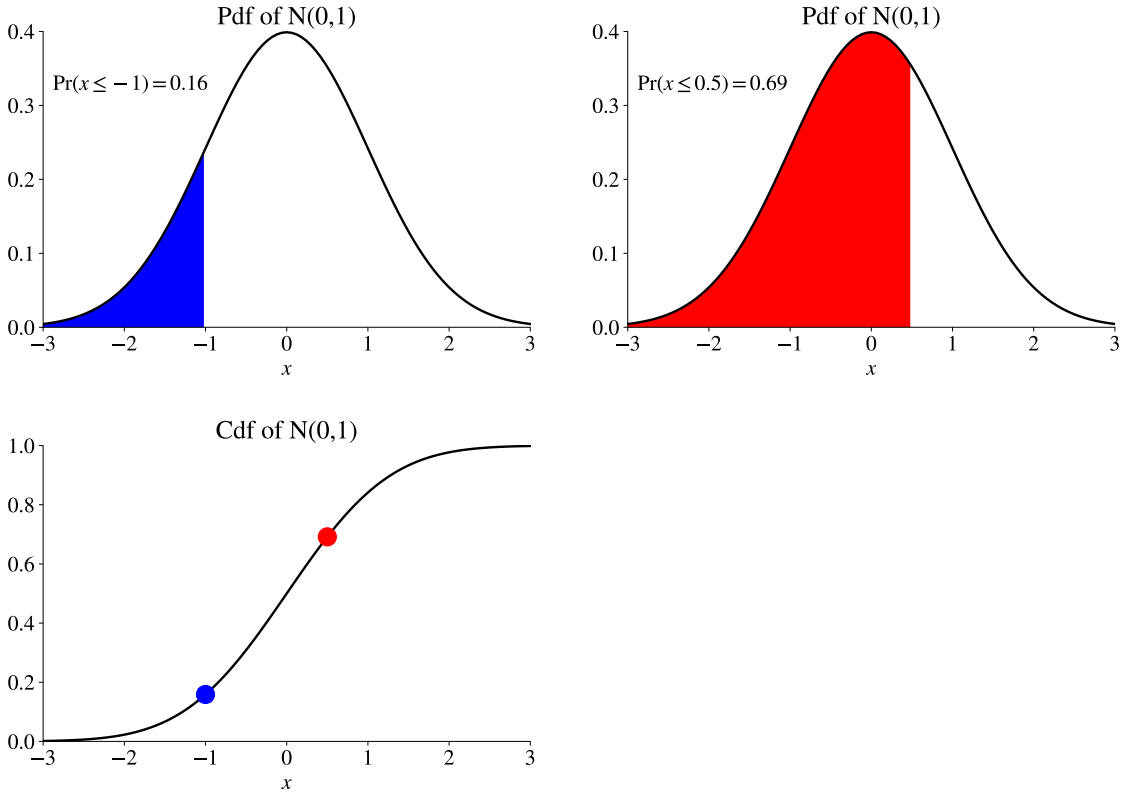


Figure 1.2: Pdf and cdf of N(0,1)

A joint normal distributions is completely described by the means and the covariance matrix

$$\begin{bmatrix} x \\ y \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}, \begin{bmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{bmatrix}\right), \tag{1.3}$$

where μ_x and μ_y denote means of x and y, σ_x^2 and σ_y^2 denote the variances of x and y and σ_{xy} denotes their covariance. Sometimes alternative notations are used: $\mathrm{E}\,x$ for the mean, $\mathrm{Std}(x)$ for the standard deviation, $\mathrm{Var}(x)$ for the variance and $\mathrm{Cov}(x, y)$ for the covariance. See Figure 31.2 for an example.

Clearly, if the covariance σ_{xy} is zero, then the variables are (linearly) unrelated to each other. Otherwise, information about x can help us to make a better guess of y. The correlation of x and y is defined as

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}. \tag{1.4}$$

See Figure 31.2 for an example.

If two random variables happen to be independent of each other, then the joint density function is just the product of the two univariate densities (here denoted $f(x)$ and $k(y)$)

$$h(x, y) = f(x)\, k(y) \text{ if } x \text{ and } y \text{ are independent.} \tag{1.5}$$

This is useful in many cases, for instance, when we construct likelihood functions for maximum likelihood estimation.

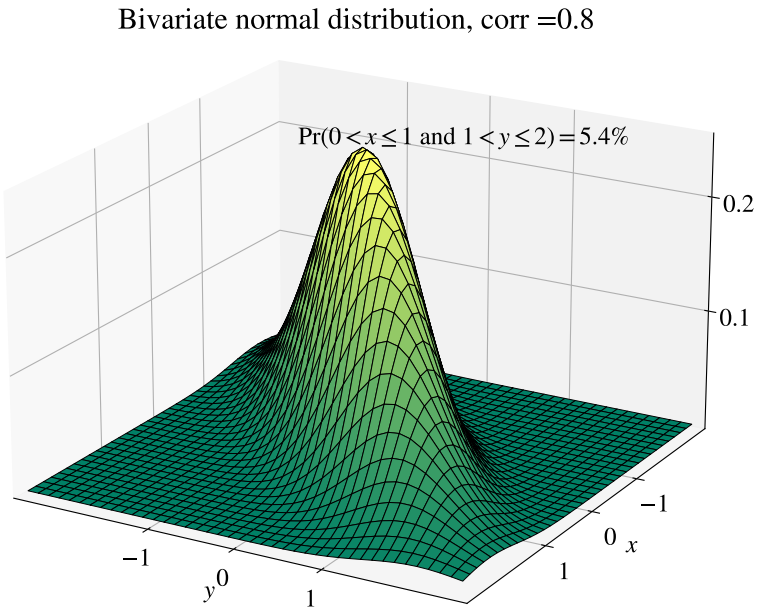


Figure 1.3: Density function bivariate normal distribution

1.1.3 Conditional Distributions*

If $h(x, y)$ is the joint density function and $f(x)$ the (marginal) density function of x, then the conditional density function is

$$g(y|x) = h(x, y)/f(x). \tag{1.6}$$

Notice that the conditional mean can be interpreted as the best guess of y given that we know x. Similarly, the conditional variance can be interpreted as the variance of the

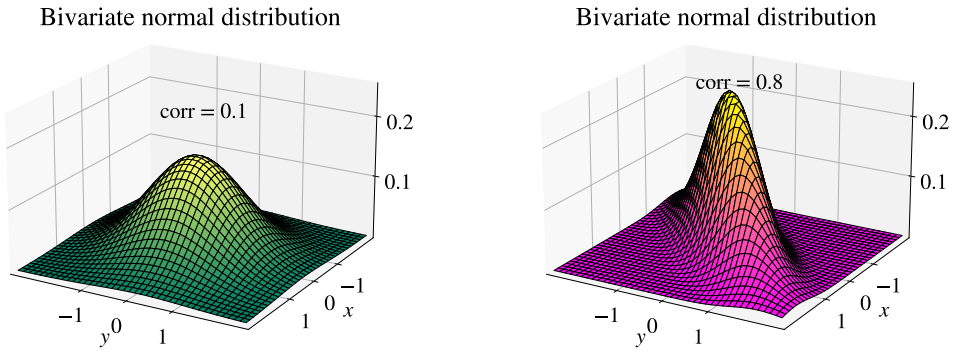


Figure 1.4: Density function bivariate normal distributions

forecast error (using the conditional mean as the forecast). The conditional and marginal distribution coincide if x and y are independent. (This follows directly from combining (1.5) and (1.6).)

For the bivariate normal distribution (1.3) we have the distribution of y conditional on a given value of x as

$$y|x \sim N\left(\mu_y + \frac{\sigma_{xy}}{\sigma_x^2}(x - \mu_x), \sigma_y^2 - \frac{\sigma_{xy}\sigma_{xy}}{\sigma_x^2}\right). \tag{1.7}$$

In this case, the mean depends on x, while the variance does not. Also notice that the variance is lower than in the unconditional distribution (we have more information). Independence of x and y would here mean a zero covariance: set $\sigma_{xy} = 0$ in (1.7) to see that the conditional and unconditional distributions coincide. See Figure 31.3 for an illustration: notice how the location and the width of the conditional distribution of y changes as a function of the correlation and the value of x.

Remark 1.2 *(Relation of (1.7) to a linear regression*) Suppose you regress* $y = a + bx + u$. *The mean in (1.7) is the same as* $a + bx$ *and the variance is the same as* $\text{Var}(u)$.

1.1.4 Illustrating a Distribution

If we know the type of distribution (uniform, normal, etc) a variable has, then the best way of illustrating the distribution is to estimate its parameters (mean, variance and whatever more—see below) and then draw the density function.

In case we are not sure about which distribution to use, the first step is typically to draw

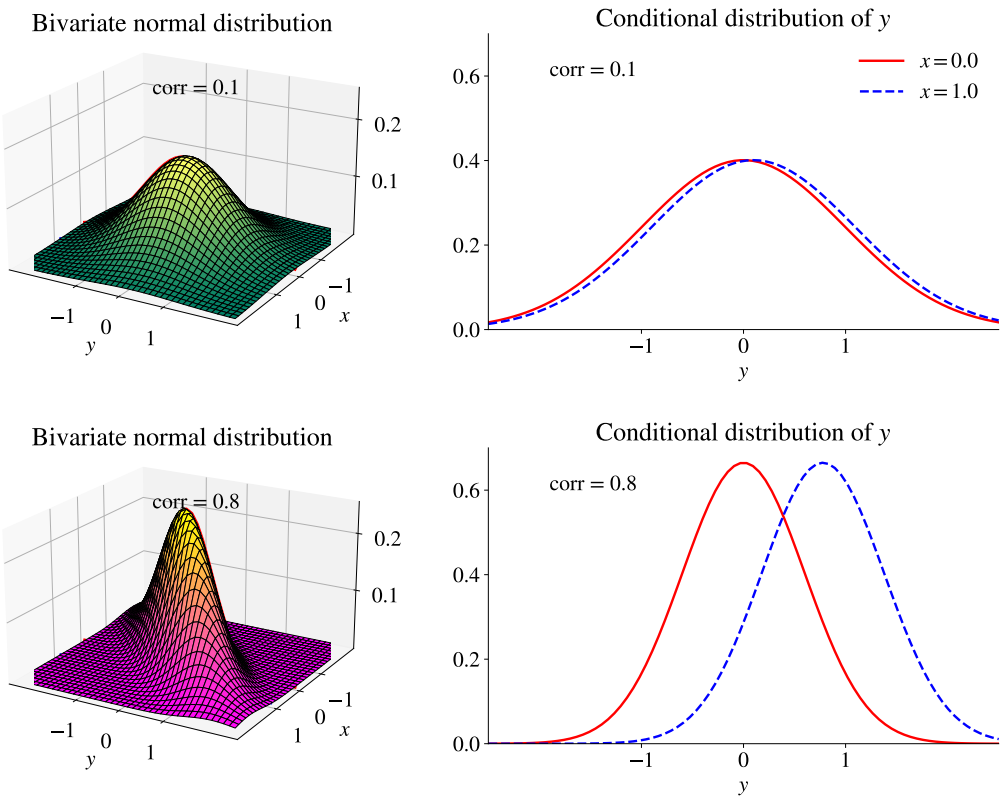


Figure 1.5: Density functions of normal distributions

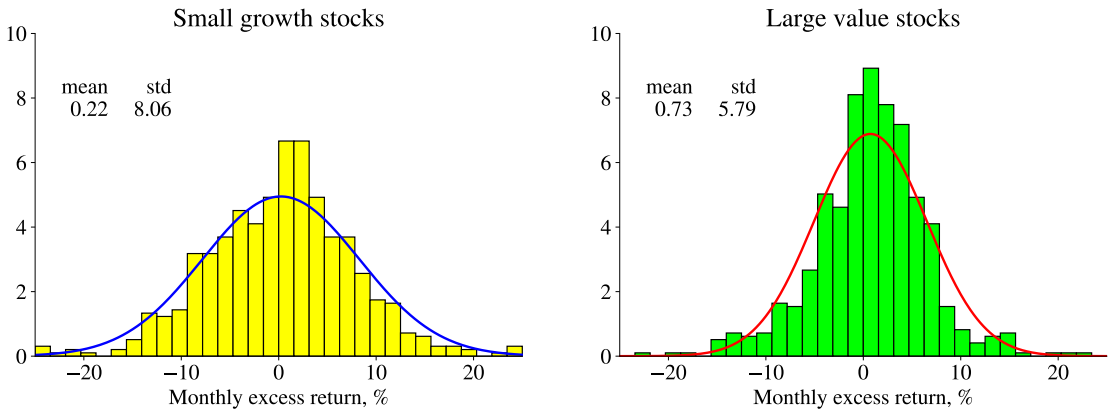
a histogram: it shows the relative frequencies for different bins (intervals). For instance, it could show the relative frequencies of a variable x_t being in each of the follow intervals: -0.5 to 0, 0 to 0.5 and 0.5 to 1.0. Clearly, the relative frequencies should sum to unity (or 100%), but they are sometimes normalized so the area under the histogram has an area of unity (as a probability density function).

Empirical Example 1.3 *(Histogram of equity returns) See Figure 1.6.*

1.1.5 Confidence Bands and t-tests

For a symmetric distribution, a 90% (two-sided) confidence band is constructed by finding a critical value c such that

$$\Pr(\mu - c < x \leq \mu + c) = 0.9. \tag{1.8}$$



Monthly data on two U.S. indices, 1970:01-2021:12
Sample size: 624
Solid line: estimated normal distribution

Figure 1.6: Histogram of returns, the curve is a normal distribution with the same mean and standard deviation as the return series

Replace 0.9 by 0.95 to get a 95% confidence band—and similarly for other confidence levels. In particular, if $x \sim N(\mu, \sigma^2)$, then

$$\begin{aligned} &\Pr(\mu - 1.64\sigma < x \leq \mu + 1.64\sigma) = 0.9 \text{ and} \\ &\Pr(\mu - 1.96\sigma < x \leq \mu + 1.96\sigma) = 0.95. \end{aligned} \tag{1.9}$$

As an example, suppose x is not a data series but a regression coefficient (denoted $\hat{\beta}$)—and we know that the standard error equals some number σ. We could then construct a 90% *confidence band around the point estimate* ($\hat{\beta}$) as

$$[\hat{\beta} - 1.64\sigma, \hat{\beta} + 1.64\sigma]. \tag{1.10}$$

In case this band does not include your null hypothesis $\beta = q$ ($q = 0$ is a commonly used special case), then we would be 90% that the (true) regression coefficient is different from q.

Alternatively, suppose we instead construct the 90% confidence band around q as

$$[q - 1.64\sigma, q + 1.64\sigma]. \tag{1.11}$$

If this band does not include the point estimate ($\hat{\beta}$), then we are also 90% sure that the (true) regression coefficient is different from q.

A third way to create a confidence band is to first create a standardized variable

$$t = \frac{\hat{\beta} - q}{\sigma}, \tag{1.12}$$

and then notice that we are 90% sure that t is in the interval

$$[-1.64, 1.64]. \tag{1.13}$$

(Provided the null hypothesis is true, that is, $\beta = q$.) This is a t-test. Testing the null hypothesis by using (1.10), (1.11) or (1.13) should give the same answer to the question: is there sufficient statistical evidence against the null hypothesis.

1.1.6 The Idea behind Confidence Bands and t-tests

Suppose we have estimated a parameter ($\hat{\beta}$) from a particular sample of data (observations 1 to T, say). The parameter could, for instance, be the mean or a regression coefficient. This estimate is actually a random variable so it makes sense to construct a confidence band as in (1.10). The reason for why it is a random variable is that another sample is most likely to produce a different estimate—and that if we could try all possible samples then the different estimates would have some sort of distribution. *If* we are willing to assume that data for those other samples would be similar (scattered around the same mean, showing the same degree of dispersion, etc) to the sample we actually study (observations 1 to T), then we can use our sample to guess how much other samples would differ. For instance, we can estimate the variance of the data (σ^2) and draw the conclusions about how different the sample averages would be in different samples (it would have a variance of σ^2/T as discussed in (1.16)).

1.1.7 Hypothesis Testing

We are here interested in testing the null hypothesis that $\beta = q$, where q is a number of interest (0.27, say). A null hypothesis is often denoted H_0. (Econometric programs often automatically report results for H_0: $\beta = 0$.) We here consider the alternative hypothesis (denoted H_1 or perhaps H_A) that $\beta \neq q$. This leads to a two-sided (or two-tailed) test.

Typically, we assume that the estimates are normally distributed. To be able to easily compare with printed tables of probabilities, transform to a $N(0,1)$ variable. In particular, if the true coefficient is really q, then $\hat{\beta} - q$ should have a zero mean (recall that $\mathrm{E}\,\hat{\beta}$ equals

the true value). Dividing by the standard error (deviation) of $\hat{\beta}$, we should have

$$t = \frac{\hat{\beta} - q}{\text{Std}(\hat{\beta})} \sim N(0, 1) \tag{1.14}$$

In case $|t|$ is very large (say, 1.64 or larger), then our estimate $\hat{\beta}$ is a very unlikely outcome if $\text{E}\,\hat{\beta}$ (which equals the true coefficient value, β) is indeed q. We therefore draw the conclusion that the true coefficient is not q, that is, we *reject the null hypothesis*.

1.2 Moments

1.2.1 Mean and Standard Deviation

The mean and variance of a series are estimated as

$$\bar{x} = \sum_{t=1}^{T} x_t / T \text{ and } \hat{\sigma}^2 = \sum_{t=1}^{T} (x_t - \bar{x})^2 / T. \tag{1.15}$$

The standard deviation (the square root of the variance) is the most common measure of volatility. (Sometimes we use $T - 1$ in the denominator of the sample variance instead T.) See Figure 1.6 for an illustration.

A sample mean is normally distributed if x_t is normally distributed, $x_t \sim N(\mu, \sigma^2)$. The reason is that a linear combination of normally distributed variables is (typically) also normally distributed. However, a sample average is often approximately normally distributed even if the variable is not (discussed below). If x_t is iid (independently and identically distributed), then the variance of a sample mean is

$$\text{Var}(\bar{x}) = \sigma^2 / T, \text{ if } x_t \text{ is iid.} \tag{1.16}$$

A sample average is (typically) *unbiased*, that is, the expected value of the sample average equals the population mean, that is,

$$\text{E}\,\bar{x} = \text{E}\,x_t = \mu. \tag{1.17}$$

Since sample averages are typically normally distributed in large samples, we thus have

$$\bar{x} \sim N(\mu, \sigma^2 / T), \tag{1.18}$$

so we can construct a *t-stat* as

$$t = \frac{\bar{x} - \mu}{\sigma/\sqrt{T}}, \tag{1.19}$$

which has an $N(0,1)$ distribution.

Proof. (of (1.16)–(1.17)) To prove (1.16), notice that

$$\begin{aligned}
\text{Var}(\bar{x}) &= \text{Var}\left(\textstyle\sum_{t=1}^{T} x_t/T\right) \\
&= \textstyle\sum_{t=1}^{T} \text{Var}\left(x_t/T\right) \\
&= T\,\text{Var}\left(x_t\right)/T^2 \\
&= \sigma^2/T.
\end{aligned}$$

The first equality is just a definition and the second equality follows from the assumption that x_t and x_s are independently distributed. This means, for instance, that $\text{Var}(x_2 + x_3) = \text{Var}(x_2) + \text{Var}(x_3)$ since the covariance is zero. The third equality follows from the assumption that x_t and x_s are identically distributed (so their variances are the same). The fourth equality is a trivial simplification.

To prove (1.17)

$$\begin{aligned}
\text{E}\,\bar{x} &= \text{E}\textstyle\sum_{t=1}^{T} x_t/T \\
&= \textstyle\sum_{t=1}^{T} \text{E}\,x_t/T \\
&= \text{E}\,x_t.
\end{aligned}$$

The first equality is just a definition and the second equality is always true (the expectation of a sum is the sum of expectations), and the third equality follows from the assumption of identical distributions which implies identical expectations. ■

1.2.2 Skewness and Kurtosis

The skewness, kurtosis and Bera-Jarque test for normality are useful diagnostic tools. They are

		Test statistic	Distribution
skewness	=	$\frac{1}{T}\sum_{t=1}^{T}\left(\frac{x_t-\mu}{\sigma}\right)^3$	$N\left(0, 6/T\right)$
kurtosis	=	$\frac{1}{T}\sum_{t=1}^{T}\left(\frac{x_t-\mu}{\sigma}\right)^4$	$N\left(3, 24/T\right)$
Bera-Jarque	=	$\frac{T}{6}\text{skewness}^2 + \frac{T}{24}\left(\text{kurtosis} - 3\right)^2$	χ^2_2.

(1.20)

This is implemented by using the estimated mean and standard deviation. See Figure 1.6 for an illustration.

The distributions stated on the right hand side of (1.20) are under the null hypothesis that x_t is iid $N(\mu, \sigma^2)$. The "excess kurtosis" is defined as the kurtosis minus 3. The test statistic for the normality test (Bera-Jarque) can be compared with 4.6 or 6.0, which are the 10% and 5% critical values of a χ^2_2 distribution.

Clearly, we can test the skewness and kurtosis by traditional t-stats as in

$$t = \frac{\text{skewness}}{\sqrt{6/T}} \text{ and } t = \frac{\text{kurtosis} - 3}{\sqrt{24/T}}, \tag{1.21}$$

which both have $N(0, 1)$ distribution under the null hypothesis of a normal distribution.

1.2.3 Covariance and Correlation

The covariance of two variables (here x and y) is typically estimated as

$$\hat{\sigma}_{xy} = \sum_{t=1}^{T} (x_t - \bar{x})(y_t - \bar{y})/T. \tag{1.22}$$

(Sometimes we use $T - 1$ in the denominator of the sample covariance instead of T.)

The correlation of two variables is then estimated as

$$\hat{\rho}_{xy} = \frac{\hat{\sigma}_{xy}}{\hat{\sigma}_x \hat{\sigma}_y}, \tag{1.23}$$

where $\hat{\sigma}_x$ and $\hat{\sigma}_y$ are the estimated standard deviations. A correlation must be between -1 and 1. Note that covariance and correlation measure the degree of *linear* relation only. This is illustrated in Figure 1.7.

Empirical Example 1.4 *(Scatter plot of equity returns) See Figure 1.8.*

Under the null hypothesis of no correlation—and if the data is approximately normally distributed, then

$$\frac{\hat{\rho}}{\sqrt{1 - \hat{\rho}^2}} \sim N(0, 1/T), \tag{1.24}$$

so we can form a t-stat as

$$t = \sqrt{T} \frac{\hat{\rho}}{\sqrt{1 - \hat{\rho}^2}}, \tag{1.25}$$

which has an $N(0, 1)$ distribution.

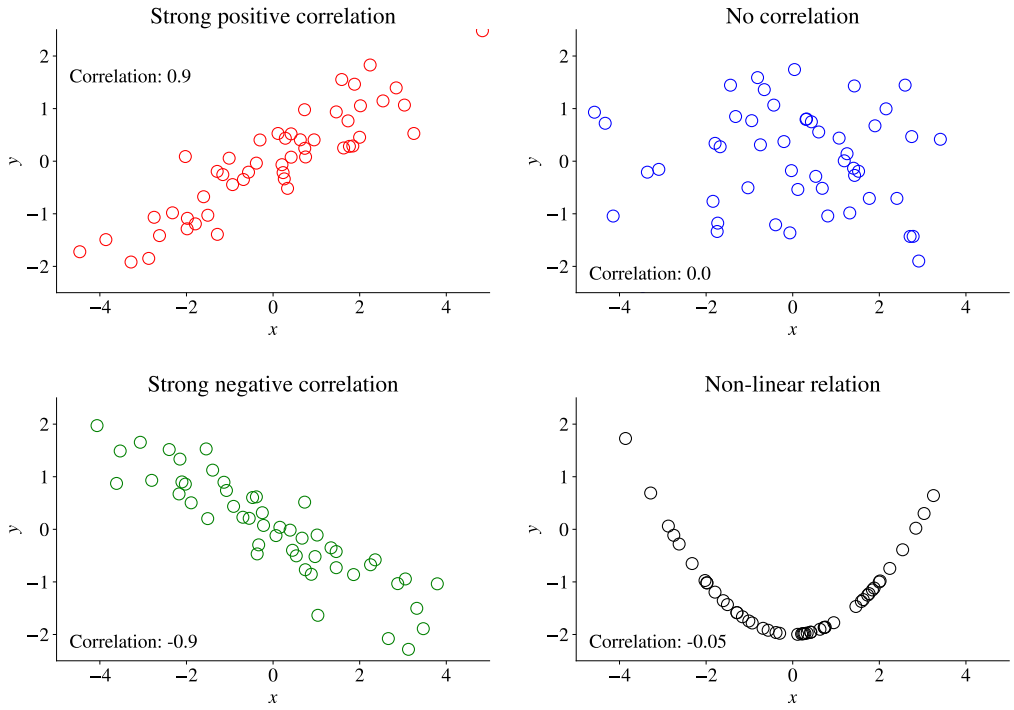


Figure 1.7: Example of correlations.

1.2.4 Correlations vs. Causality

Notice that a correlation between x and y does not say anything about causality. There are several possibilities, including

$$\begin{aligned} (x, \varepsilon) &\Rightarrow y \\ (y, u) &\Rightarrow x \\ (z, u) &\Rightarrow x \text{ and } (z, \varepsilon) \Rightarrow y \end{aligned} \tag{1.26}$$

In the first case, x and some other variables (here labelled ε) are indeed causing y, so changes in x are likely to be accompanied by changes in y. The second case shows the opposite: y is causing x. The third case is when some other variable z is driving the correlation between x and y. However, an independent move in x (due to u) will not lead to moves in y. This reasoning carries over to regression analysis too. In many regressions we would like to capture the causality, although forecasting models are more focused on the correlation per se.

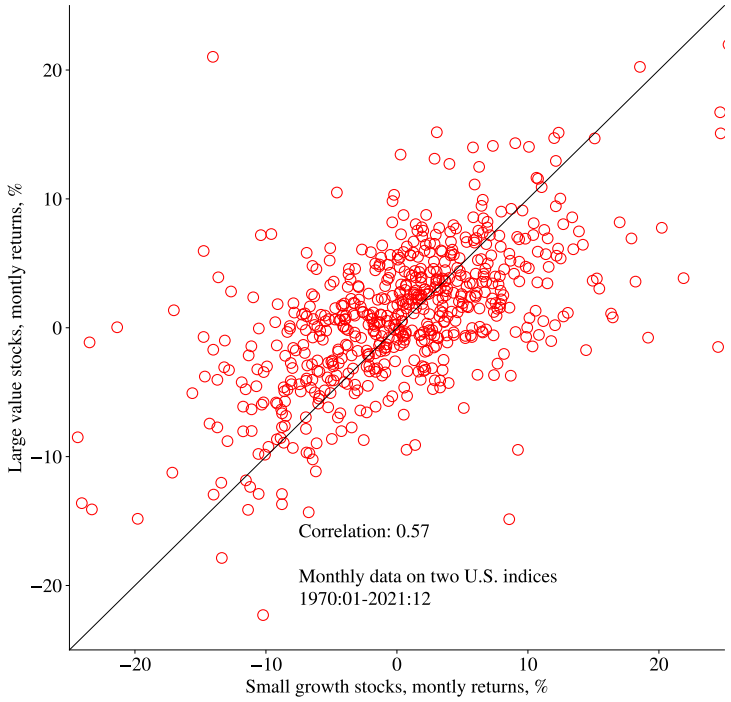


Figure 1.8: Scatter plot of two different portfolio returns

1.2.5 Correlations and the Variance of a Sample Average

The result in (1.16) that $\text{Var}(\bar{x}) = \sigma^2/T$ does not hold if x_t and x_{t-s} are correlated. To see that, consider the case when the sample has just two observations

$$\begin{aligned} \bar{x} &= (x_1 + x_2)/2 \text{ and} \\ \text{Var}(\bar{x}) &= (\sigma^2 + \sigma^2 + \sigma_{12} + \sigma_{21})/4. \end{aligned} \tag{1.27}$$

In the iid case we *assume* that $\sigma_{12} = \sigma_{21} = 0$, so $\text{Var}(\bar{x}) = \sigma^2/2$. In the other extreme case of perfect correlations, $\sigma_{12} = \sigma^2$ so the variance of the sample average is the same as for the data (no precision is gained by averaging), $\text{Var}(\bar{x}) = \sigma^2$.

More generally, with T observations, we have

$$\text{Var}(\bar{x}) = \textstyle\sum_{i=1}^{T}\sum_{j=1}^{T}\sigma_{ij}/T^2, \tag{1.28}$$

which is sum of all the elements of the covariance matrix, divided by T^2. This can be written as

$$\text{Var}(\bar{x}) = \frac{\bar{\sigma}^2 - \bar{\sigma}_{ij}}{T} + \bar{\sigma}_{ij}, \tag{1.29}$$

where $\bar{\sigma}^2$ is the average variance and $\bar{\sigma}_{ij}$ the average covariance of any two observations (x_t and x_{t-s}, say). This carries over to the variance of regression coefficients—when the residuals are correlated (over time or over cross sectional units).

Example 1.5 *(Covariance matrix with $T = 2$) The covariance matrix of x_1 and x_2 is*

$$\begin{bmatrix} \sigma^2 & \sigma_{12} \\ \sigma_{21} & \sigma^2 \end{bmatrix},$$

if we assume that x_1 and x_2 have the same variance (σ^2). Also, notice that $\sigma_{12} = \sigma_{21}$.

Example 1.6 *(*$\text{Var}(\bar{x})$*) Assume $\bar{\sigma}^2 = 1$, then* $\text{Var}(\bar{x})$ *is*

	$\bar{\sigma}_{ij} = 0$	$\bar{\sigma}_{ij} = 0.10$
$T = 10$:	0.1	0.19
$T = 100$:	0.01	0.109

1.3 Distributions Commonly Used in Tests

1.3.1 Standard Normal Distribution, $N(0, 1)$

Suppose the random variable x has a $N(\mu, \sigma^2)$ distribution. Then, the the *standardized variable* $(x - \mu)/\sigma$ has a standard normal distribution

$$t = \frac{x - \mu}{\sigma} \sim N(0, 1). \tag{1.30}$$

To see this, notice that $x - \mu$ has a mean of zero and that x/σ has a standard deviation of unity. (This result is the motivation for why the confidence band (1.13) gives the same result as (1.11).)

1.3.2 t-distribution

If we instead need to estimate σ to use in (1.30), then the test statistic has t_n-distribution

$$t = \frac{x - \mu}{\hat{\sigma}} \sim t_n, \tag{1.31}$$

where n denotes the "degrees of freedom," that is the number of observations minus the number of estimated parameters. For instance, if we have a sample with T data points and only estimate the mean, then $n = T - 1$.

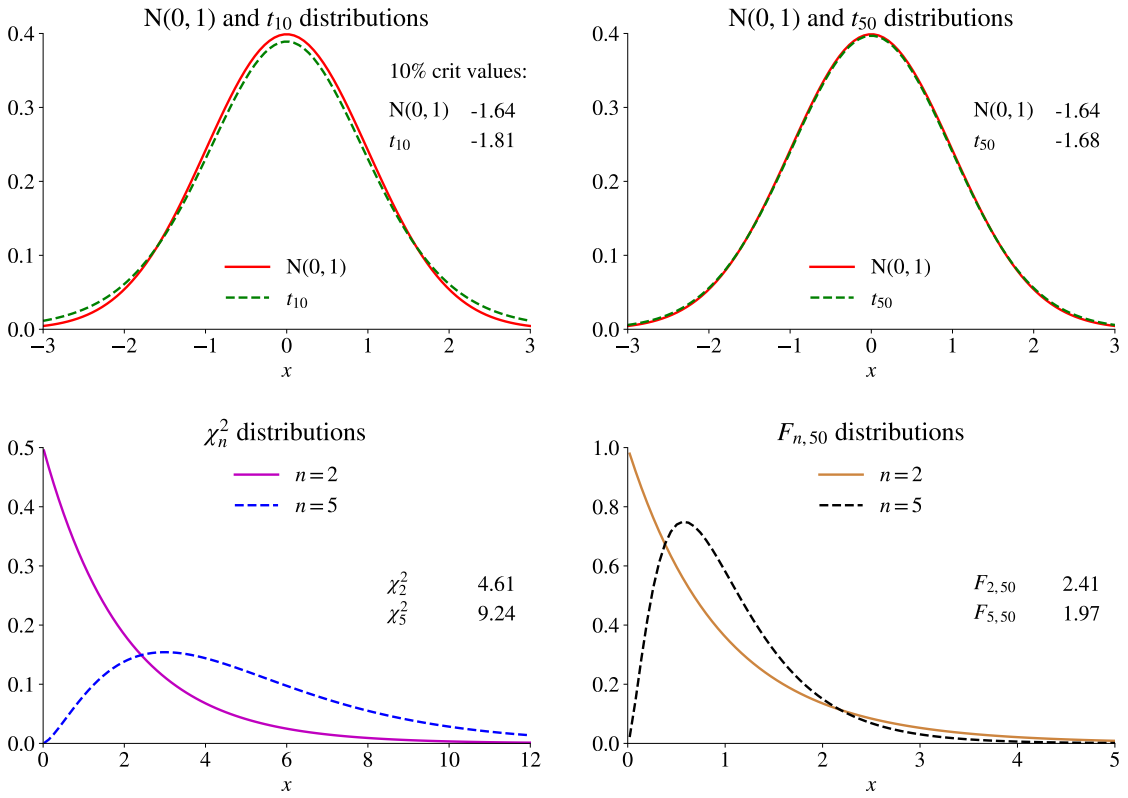


Figure 1.9: Probability density functions

The t-distribution has more probability mass in the tails than an $N(0,1)$ distribution:. It therefore gives a more "conservative" test (harder to reject the null hypothesis), but the difference vanishes as the degrees of freedom (sample size) increases. See Figure 31.5 for a comparison and Table 1.1 for critical values.

Example 1.7 *(t-distribution) If* $t = 2.0$ *and* $n = 50$*, then this is larger than the 10% critical value (but not the 5% critical value) for a 2-sided test in Table 1.1.*

1.3.3 Chi-square Distribution

If $z \sim N(0,1)$, then $z^2 \sim \chi^2_1$, that is, z^2 has a chi-square distribution with one degree of freedom. This can be generalized in several ways. For instance, if $x \sim N(\mu_x, \sigma_{xx})$ and $y \sim N(\mu_y, \sigma_{yy})$ and they are uncorrelated, then $[(x-\mu_x)/\sigma_x]^2 + [(y-\mu_y)/\sigma_y]^2 \sim \chi^2_2$.

More generally, we have

$$v'\Sigma^{-1}v \sim \chi^2_n, \text{ if the } n \times 1 \text{ vector } v \sim N(0,\Sigma). \tag{1.32}$$

See Figure 31.5 for an illustration and Table 1.2 for critical values.

Example 1.8 *(χ^2_2 distribution) Suppose x is a 2×1 vector*

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \sim N\left(\begin{bmatrix} 4 \\ 2 \end{bmatrix}, \begin{bmatrix} 5 & 3 \\ 3 & 4 \end{bmatrix} \right).$$

If $x_1 = 3$ and $x_2 = 5$, then

$$\begin{bmatrix} 3-4 \\ 5-2 \end{bmatrix}' \begin{bmatrix} 5 & 3 \\ 3 & 4 \end{bmatrix}^{-1} \begin{bmatrix} 3-4 \\ 5-2 \end{bmatrix} \approx 6.1$$

has a χ^2_2 distribution. Notice that 6.1 is higher than the 5% critical value (but not the 1% critical value) in Table 1.2.

1.3.4 F-distribution

If $x \sim \chi^2_{n_1}$ and $y \sim \chi^2_{n_2}$, then $(x/n_1)/(y/n_2)$ has an F_{n_1,n_2} distribution with (n_1, n_2) degrees of freedom. See Figure 31.5 for an illustration and Tables 1.3–1.4 for critical values.

1.4 Normal Distribution of the Sample Mean

In many cases, it is unreasonable to assume that a random variable x_t is normally distributed. The nice thing with a sample mean (or sample average), here denoted $\bar{x}$, is that it has very useful properties (in a reasonably large sample). This section gives a short summary of what happens to sample means as the sample size increases (often called "asymptotic theory").

The *law of large numbers* (LLN) says that the sample mean converges to the true population mean as the sample size goes to infinity. This holds for a very large class of random variables, but there are exceptions. A sufficient (but not necessary) condition for this convergence is that the sample average is unbiased (as in (1.17)) and that the variance goes to zero as the sample size goes to infinity (as in (1.16)). (This is also called convergence in mean square.) To see the LLN in action, see Figure 31.1.

The *central limit theorem* (CLT) says that $\sqrt{T}\bar{x}$ converges in distribution to a normal distribution as the sample size increases. See Figure 31.1 for an illustration. This also holds for a large class of random variables—and it is a very useful result since it allows

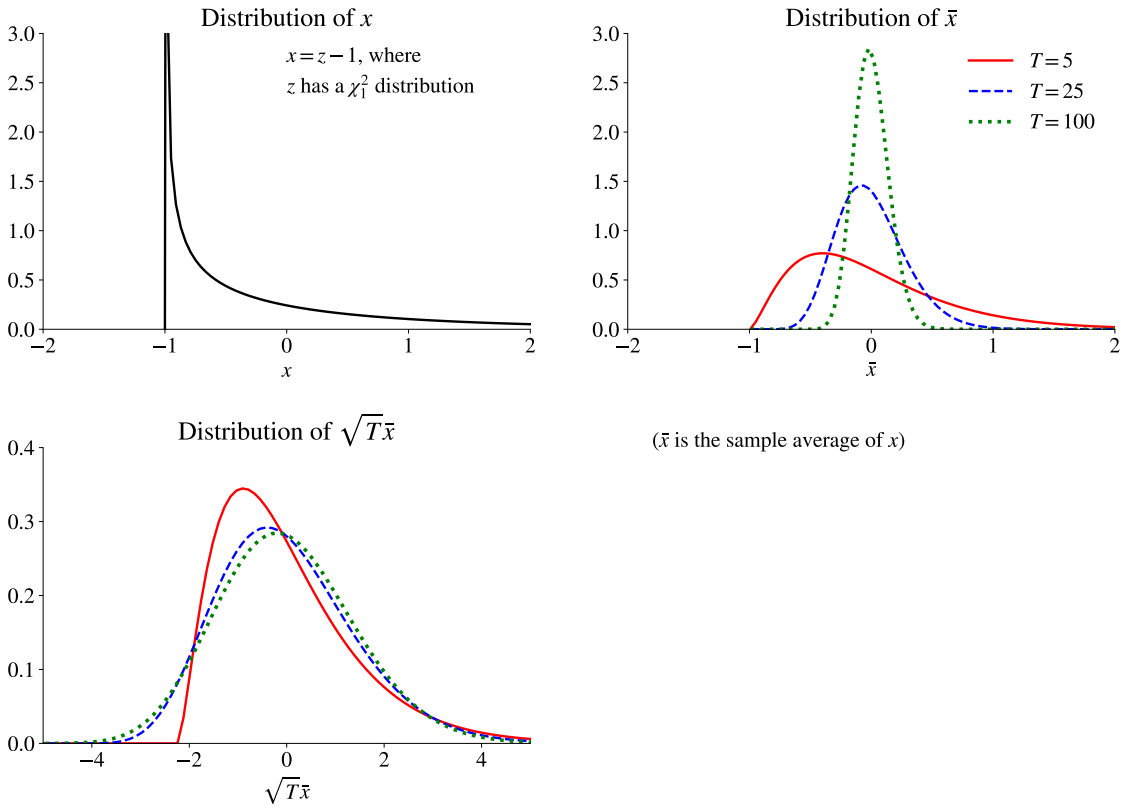


Figure 1.10: Sampling distributions

us to test hypotheses by assuming that $\sqrt{T}\bar{x}$ is normally distributed. Most estimators (including least squares and other methods) are effectively some kind of sample average, so the CLT can be applied.

1.5 Appendix: Statistical Tables

n	Significance level		
	10%	5%	1%
10	1.81	2.23	3.17
20	1.72	2.09	2.85
30	1.70	2.04	2.75
40	1.68	2.02	2.70
50	1.68	2.01	2.68
60	1.67	2.00	2.66
70	1.67	1.99	2.65
80	1.66	1.99	2.64
90	1.66	1.99	2.63
100	1.66	1.98	2.63
Normal	1.64	1.96	2.58

Table 1.1: Critical values (two-sided test) of t distribution (different degrees of freedom) and normal distribution.

1.6 Appendix: Data Sources

The data used in these lecture notes are from the following sources:

1. website of Kenneth French,
   `http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html`
2. Datastream
3. Federal Reserve Bank of St. Louis (FRED), `http://research.stlouisfed.org/fred2/`
4. website of Robert Shiller, `http://www.econ.yale.edu/~shiller/data.htm`
5. yahoo! finance, `http://finance.yahoo.com/`
6. OlsenData, `http://www.olsendata.com`

n	Significance level		
	10%	5%	1%
1	2.71	3.84	6.63
2	4.61	5.99	9.21
3	6.25	7.81	11.34
4	7.78	9.49	13.28
5	9.24	11.07	15.09
6	10.64	12.59	16.81
7	12.02	14.07	18.48
8	13.36	15.51	20.09
9	14.68	16.92	21.67
10	15.99	18.31	23.21

Table 1.2: Critical values of chisquare distribution (different degrees of freedom, n).

n_1	n_2					$\chi^2_{n_1}/n_1$
	10	30	50	100	300	
1	4.96	4.17	4.03	3.94	3.87	3.84
2	4.10	3.32	3.18	3.09	3.03	3.00
3	3.71	2.92	2.79	2.70	2.63	2.60
4	3.48	2.69	2.56	2.46	2.40	2.37
5	3.33	2.53	2.40	2.31	2.24	2.21
6	3.22	2.42	2.29	2.19	2.13	2.10
7	3.14	2.33	2.20	2.10	2.04	2.01
8	3.07	2.27	2.13	2.03	1.97	1.94
9	3.02	2.21	2.07	1.97	1.91	1.88
10	2.98	2.16	2.03	1.93	1.86	1.83

Table 1.3: 5% Critical values of F_{n_1,n_2} distribution (different degrees of freedom).

n_1			n_2			$\chi^2_{n_1}/n_1$
	10	30	50	100	300	
1	3.29	2.88	2.81	2.76	2.72	2.71
2	2.92	2.49	2.41	2.36	2.32	2.30
3	2.73	2.28	2.20	2.14	2.10	2.08
4	2.61	2.14	2.06	2.00	1.96	1.94
5	2.52	2.05	1.97	1.91	1.87	1.85
6	2.46	1.98	1.90	1.83	1.79	1.77
7	2.41	1.93	1.84	1.78	1.74	1.72
8	2.38	1.88	1.80	1.73	1.69	1.67
9	2.35	1.85	1.76	1.69	1.65	1.63
10	2.32	1.82	1.73	1.66	1.62	1.60

Table 1.4: 10% Critical values of F_{n_1,n_2} distribution (different degrees of freedom).

	0.0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.0	0.0013	0.0014	0.0014	0.0015	0.0015	0.0016	0.0016	0.0017	0.0018	0.0018
-2.9	0.0019	0.0019	0.0020	0.0021	0.0021	0.0022	0.0023	0.0023	0.0024	0.0025
-2.8	0.0026	0.0026	0.0027	0.0028	0.0029	0.0030	0.0031	0.0032	0.0033	0.0034
-2.7	0.0035	0.0036	0.0037	0.0038	0.0039	0.0040	0.0041	0.0043	0.0044	0.0045
-2.6	0.0047	0.0048	0.0049	0.0051	0.0052	0.0054	0.0055	0.0057	0.0059	0.0060
-2.5	0.0062	0.0064	0.0066	0.0068	0.0069	0.0071	0.0073	0.0075	0.0078	0.0080
-2.4	0.0082	0.0084	0.0087	0.0089	0.0091	0.0094	0.0096	0.0099	0.0102	0.0104
-2.3	0.0107	0.0110	0.0113	0.0116	0.0119	0.0122	0.0125	0.0129	0.0132	0.0136
-2.2	0.0139	0.0143	0.0146	0.0150	0.0154	0.0158	0.0162	0.0166	0.0170	0.0174
-2.1	0.0179	0.0183	0.0188	0.0192	0.0197	0.0202	0.0207	0.0212	0.0217	0.0222
-2.0	0.0228	0.0233	0.0239	0.0244	0.0250	0.0256	0.0262	0.0268	0.0274	0.0281
-1.9	0.0287	0.0294	0.0301	0.0307	0.0314	0.0322	0.0329	0.0336	0.0344	0.0351
-1.8	0.0359	0.0367	0.0375	0.0384	0.0392	0.0401	0.0409	0.0418	0.0427	0.0436
-1.7	0.0446	0.0455	0.0465	0.0475	0.0485	0.0495	0.0505	0.0516	0.0526	0.0537
-1.6	0.0548	0.0559	0.0571	0.0582	0.0594	0.0606	0.0618	0.0630	0.0643	0.0655
-1.5	0.0668	0.0681	0.0694	0.0708	0.0721	0.0735	0.0749	0.0764	0.0778	0.0793
-1.4	0.0808	0.0823	0.0838	0.0853	0.0869	0.0885	0.0901	0.0918	0.0934	0.0951
-1.3	0.0968	0.0985	0.1003	0.1020	0.1038	0.1056	0.1075	0.1093	0.1112	0.1131
-1.2	0.1151	0.1170	0.1190	0.1210	0.1230	0.1251	0.1271	0.1292	0.1314	0.1335
-1.1	0.1357	0.1379	0.1401	0.1423	0.1446	0.1469	0.1492	0.1515	0.1539	0.1562
-1.0	0.1587	0.1611	0.1635	0.1660	0.1685	0.1711	0.1736	0.1762	0.1788	0.1814
-0.9	0.1841	0.1867	0.1894	0.1922	0.1949	0.1977	0.2005	0.2033	0.2061	0.2090
-0.8	0.2119	0.2148	0.2177	0.2206	0.2236	0.2266	0.2296	0.2327	0.2358	0.2389
-0.7	0.2420	0.2451	0.2483	0.2514	0.2546	0.2578	0.2611	0.2643	0.2676	0.2709
-0.6	0.2743	0.2776	0.2810	0.2843	0.2877	0.2912	0.2946	0.2981	0.3015	0.3050
-0.5	0.3085	0.3121	0.3156	0.3192	0.3228	0.3264	0.3300	0.3336	0.3372	0.3409
-0.4	0.3446	0.3483	0.3520	0.3557	0.3594	0.3632	0.3669	0.3707	0.3745	0.3783
-0.3	0.3821	0.3859	0.3897	0.3936	0.3974	0.4013	0.4052	0.4090	0.4129	0.4168
-0.2	0.4207	0.4247	0.4286	0.4325	0.4364	0.4404	0.4443	0.4483	0.4522	0.4562
-0.1	0.4602	0.4641	0.4681	0.4721	0.4761	0.4801	0.4840	0.4880	0.4920	0.4960

Table 1.5: Values of the standard normal distribution function at x where x is the sum of the values in the first column and the first row.

	0.0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986

Table 1.6: Values of the standard normal distribution function at x where x is the sum of the values in the first column and the first row.

Chapter 2

Least Squares Estimation

Reference: Verbeek (2012) 2 and 4; Greene (2018) 2-4.
More advanced material is denoted by a star (*). It is not required reading.

2.1 Least Squares: The Optimization Problem and Its Solution

2.1.1 Simple Regression

The simplest regression model is

$$y_t = \beta_0 + \beta_1 x_t + u_t, \text{ where } \mathrm{E}\, u_t = 0 \text{ and } \mathrm{Cov}(x_t, u_t) = 0, \tag{2.1}$$

where we can observe (have data on) the dependent variable y_t and the regressor x_t but not the residual u_t. In principle, the residual should account for all the movements in y_t that we cannot explain by x_t. The subscript t refers to observation t, which could represent period t (when data is a time series) or investor t (when data is a cross-section). In the latter case, it is common to instead use i as subscript.

Remark 2.1 *(On notation) These notes sometimes use alternative notations for the regression equation, for instance, $y_t = \alpha + \beta x_t + u_t$ (as is typical in CAPM regressions) or $y_i = a + bx_i + u_i$.*

Notice the two very important assumptions: (*i*) the mean of the residual is zero; and (*ii*) the residual is not correlated with the regressor, x_t. This basically says that the residual is pure noise. In contrast, if the average of u_t was non-zero, then $\beta_0 + \beta_1 x_t$ would get the general level of y_t wrong. Also, if x_t and u_t were correlated, then the best guess of y_t based on x_t would not be $\beta_0 + \beta_1 x_t$.

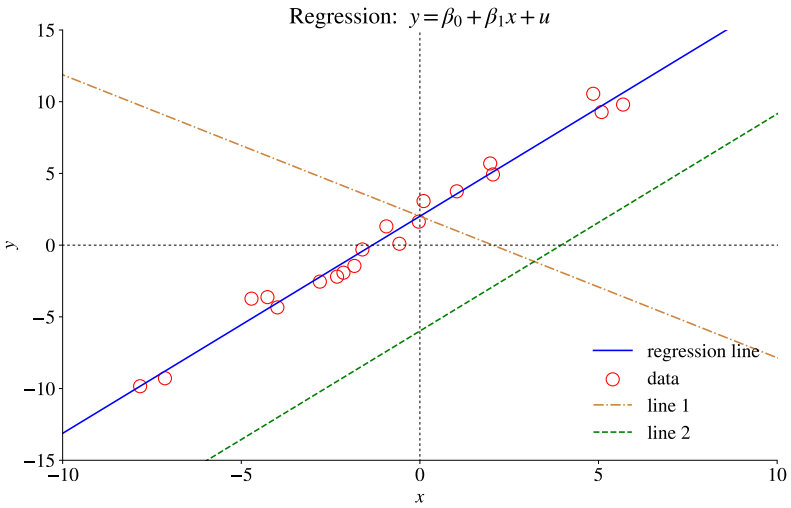


Figure 2.1: Example of OLS

Suppose you do not know β_0 or β_1, and that you have a sample of data: y_t and x_t for $t = 1, ..., T$. The LS estimator of β_0 and β_1 minimizes the loss function

$$\sum_{t=1}^{T}(y_t - b_0 - b_1 x_t)^2 = (y_1 - b_0 - b_1 x_1)^2 + (y_2 - b_0 - b_1 x_2)^2 + ... \tag{2.2}$$

by choosing b_0 and b_1 to make the loss function value as small as possible. The objective is thus to pick values of b_0 and b_1 in order to make the model fit the data as closely as possible—where close is taken to be a small variance of the unexplained part (the residual). See Figures 2.1–2.2 for illustrations. (Least squares is only one of many possible ways to estimate regression coefficients. We will discuss other methods later on.)

Remark 2.2 *Note that β_i is the true (unobservable) value which we estimate to be $\hat{\beta}_i$. Whereas β_i is an unknown (deterministic) number, $\hat{\beta}_i$ is a random variable since it is calculated as a function of the random sample of y_t and x_t. We use b_i as an argument in the loss function (so we contemplate different values of b_i) —and the optimal value is clearly $\hat{\beta}_i$.*

Remark 2.3 *(First order condition for minimizing a differentiable function). We want to find the value of b in the interval $b_{low} \leq b \leq b_{high}$, which makes the value of the differentiable function $f(b)$ as small as possible. The answer is b_{low}, b_{high}, or a value of b where $df(b)/db = 0$. See Figure 2.3.*

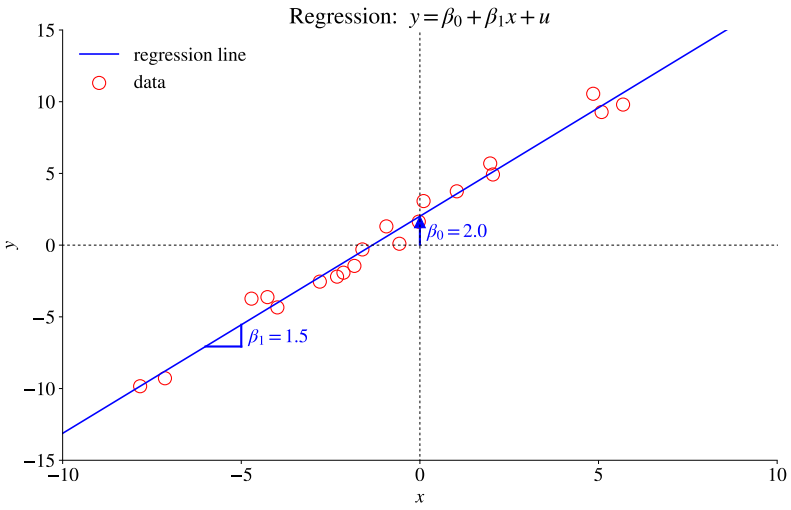


Figure 2.2: Example of OLS

The first order conditions for a minimum are that the derivatives of this loss function with respect to b_0 and b_1 should be zero. Notice that

$$\frac{\partial}{\partial b_0}(y_t - b_0 - b_1 x_t)^2 = -2(y_t - b_0 - b_1 x_t)1 \tag{2.3}$$

$$\frac{\partial}{\partial b_1}(y_t - b_0 - b_1 x_t)^2 = -2(y_t - b_0 - b_1 x_t)x_t. \tag{2.4}$$

Let $(\hat{\beta}_0, \hat{\beta}_1)$ be the values of (b_0, b_1) where the derivatives are zero (that is, $(\hat{\beta}_0, \hat{\beta}_1)$ are the optimal values)

$$\frac{\partial}{\partial \beta_0}\sum_{t=1}^{T}(y_t - \hat{\beta}_0 - \hat{\beta}_1 x_t)^2 = -2\sum_{t=1}^{T}1(y_t - \hat{\beta}_0 - \hat{\beta}_1 x_t) = 0 \tag{2.5}$$

$$\frac{\partial}{\partial \beta_1}\sum_{t=1}^{T}(y_t - \hat{\beta}_0 - \hat{\beta}_1 x_t)^2 = -2\sum_{t=1}^{T}x_t(y_t - \hat{\beta}_0 - \hat{\beta}_1 x_t) = 0, \tag{2.6}$$

which are two equations in two unknowns ($\hat{\beta}_0$ and $\hat{\beta}_1$), which must be solved simultaneously. These equations show that both the constant and x_t should be *orthogonal* to the fitted residuals, $\hat{u}_t = y_t - \hat{\beta}_0 - \hat{\beta}_1 x_t$. This is indeed a defining feature of LS and can be seen as the sample analogues of the assumptions in (2.1) that $\mathrm{E}\, u_t = 0$ and $\mathrm{Cov}(x_t, u_t) = 0$. To see this, note that (2.5) says that the sample average of $\hat{u}_t$ should be zero. Similarly, (2.6) says that the sample cross moment of $\hat{u}_t$ and x_t (that is, $\sum_{t=1}^{T}\hat{u}_t x_t/T$) should also be zero, which implies that the sample covariance is zero as well since $\hat{u}_t$ has a zero

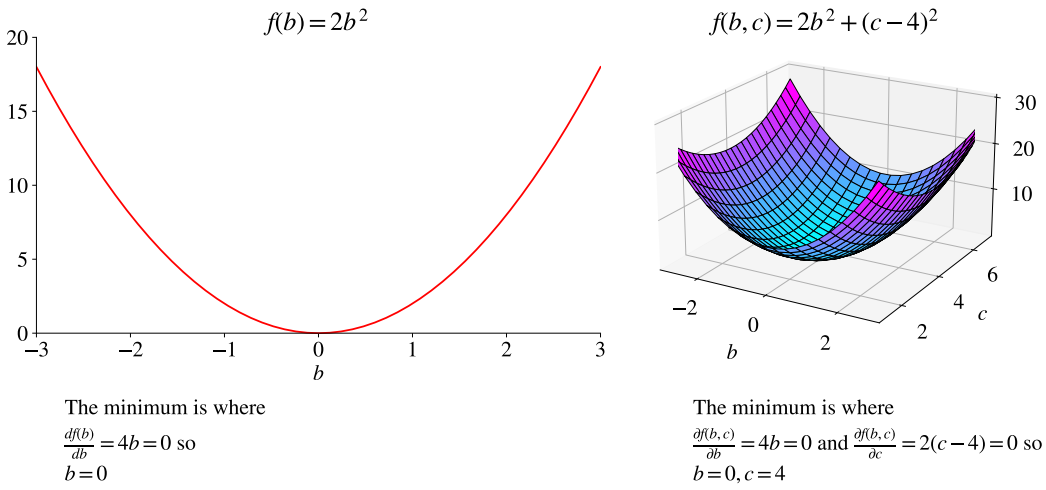


Figure 2.3: Quadratic loss function. Subfigure a: 1 coefficient; Subfigure b: 2 coefficients

sample mean (see Remark 2.4).

Remark 2.4 *(Cross moments and covariance) A covariance is defined as*

$$\begin{aligned}\operatorname{Cov}(x, y) &= \mathrm{E}[(x - \mathrm{E}\,x)(y - \mathrm{E}\,y)] \\ &= \mathrm{E}(xy - x\,\mathrm{E}\,y - y\,\mathrm{E}\,x + \mathrm{E}\,x\,\mathrm{E}\,y) \\ &= \mathrm{E}\,xy - \mathrm{E}\,x\,\mathrm{E}\,y - \mathrm{E}\,y\,\mathrm{E}\,x + \mathrm{E}\,x\,\mathrm{E}\,y \\ &= \mathrm{E}\,xy - \mathrm{E}\,x\,\mathrm{E}\,y.\end{aligned}$$

If $\mathrm{E}\,x = 0$ *or* $\mathrm{E}\,y = 0$*, then* $\operatorname{Cov}(x, y) = \mathrm{E}\,xy$*. When* $x = y$*, then we get* $\operatorname{Var}(x) = \mathrm{E}\,x^2 - (\mathrm{E}\,x)^2$*. These results hold for sample moments too.*

When the means of y and x are zero, then we know that intercept is zero ($\beta_0 = 0$). In this case, (2.6) with $\hat{\beta}_0 = 0$ immediately gives

$$\begin{aligned}\textstyle\sum_{t=1}^{T} x_t y_t &= \hat{\beta}_1 \textstyle\sum_{t=1}^{T} x_t x_t \text{ or} \\ \hat{\beta}_1 &= \frac{\sum_{t=1}^{T} x_t y_t / T}{\sum_{t=1}^{T} x_t x_t / T}.\end{aligned} \tag{2.7}$$

In this case, the coefficient estimator is the sample covariance (recall: means are zero) of y_t and x_t, divided by the sample variance of the regressor x_t (this statement is actually true even if the means are not zero and a constant is included on the right hand side—just more tedious to show it).

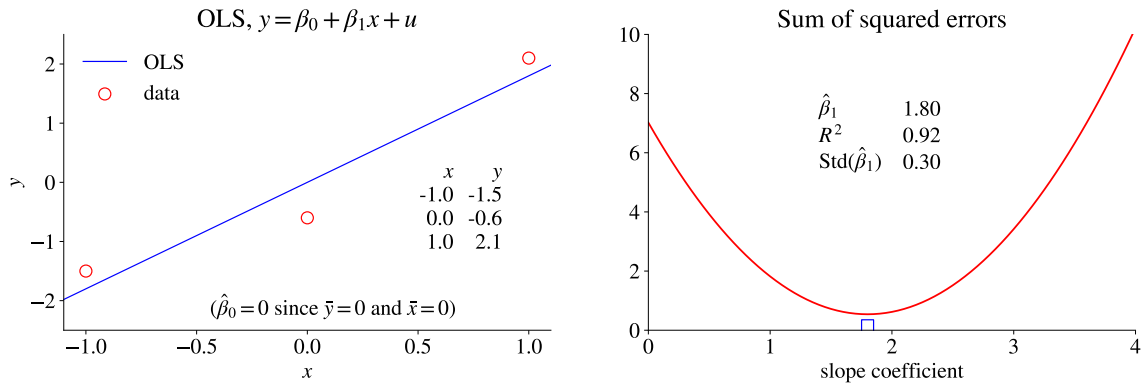


Figure 2.4: Example of OLS estimation

Empirical Example 2.5 *(CAPM regressions) See Table 2.1 and Figure 2.5 for CAPM regressions for two industry portfolios. The betas clearly differ.*

	HiTec	Utils
constant	−0.07	0.24
	(−0.54)	(1.75)
market return	1.24	0.51
	(36.79)	(13.53)
R^2	0.75	0.32
obs	624	624

Table 2.1: CAPM regressions, monthly returns, %, US data 1970:01-2021:12. Numbers in parentheses are t-stats.

Example 2.6 *(Simple regression) Consider the simple regression model (PSLS1). Suppose we have the following data*

t	x	y
1	−1	−1.5
2	0	−0.6
3	1	2.1

To calculate the LS estimate according to (2.7) we note that

$$\sum_{t=1}^{T} x_t x_t = (-1)^2 + 0^2 + 1^1 = 2 \text{ and}$$
$$\sum_{t=1}^{T} x_t y_t = (-1)(-1.5) + 0(-0.6) + 1 \times 2.1 = 3.6$$

This gives

$$\hat{\beta}_1 = \frac{3.6}{2} = 1.8.$$

The fitted residuals are

$$\begin{bmatrix} \hat{u}_1 \\ \hat{u}_2 \\ \hat{u}_3 \end{bmatrix} = \begin{bmatrix} -1.5 \\ -0.6 \\ 2.1 \end{bmatrix} - 1.8 \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 0.3 \\ -0.6 \\ 0.3 \end{bmatrix}.$$

The fitted residuals indeed obey the first order condition (2.6) since

$$\sum_{t=1}^{T} x_t \hat{u}_t = (-1) \times 0.3 + 0(-0.6) + 1 \times 0.3 = 0.$$

See Figure 2.4 for an illustration.

Example 2.7 *Using the same data as in Example 2.6 we can also calculate the sums of squared residuals for different values of the slope coefficient. With $\beta_1 = 1.6$ we get*

$\underline{t}$	$\underline{u_t}$	$\underline{u_1^2}$
1	$-1.5 - \mathbf{1.6} \times (-1) = 0.1$	0.01
2	$-0.6 - \mathbf{1.6} \times 0 = -0.6$	0.36
3	$2.1 - \mathbf{1.6} \times 1 = 0.5$	0.25
sum	0	0.62

With $\beta = 1.8$ and $\beta = 2.0$ we instead get

$\underline{t}$	$\underline{u_t}$	$\underline{u_1^2}$
1	$-1.5 - \mathbf{1.8} \times (-1) = 0.3$	0.09
2	$-0.6 - \mathbf{1.8} \times 0 = -0.6$	0.36
3	$2.1 - \mathbf{1.8} \times 1 = 0.3$	0.09
sum	0	0.54

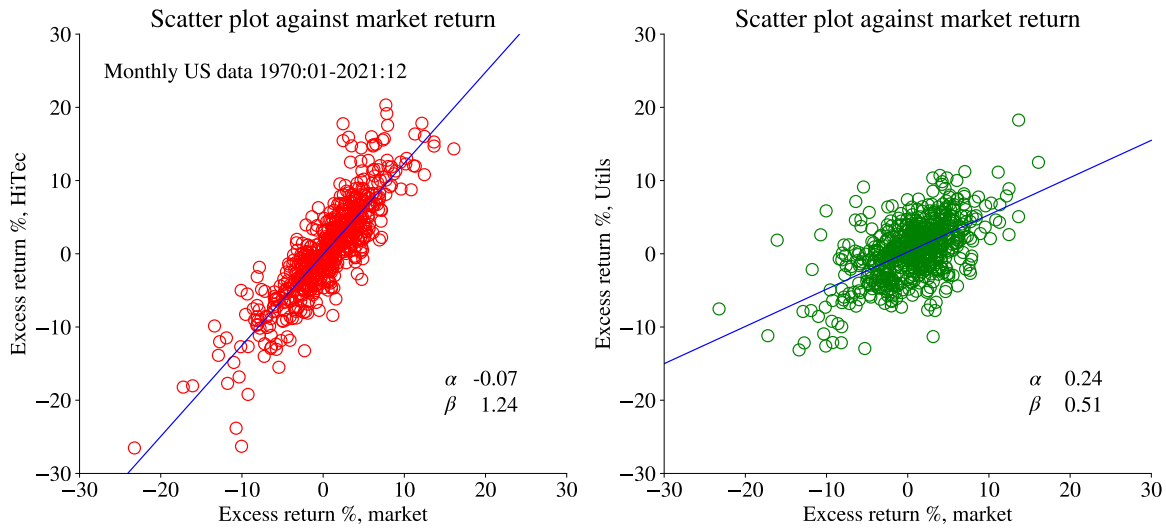


Figure 2.5: Scatter plot against market return

$\underline{t}$	$\underline{u_t}$	$\underline{u_1^2}$
1	$-1.5 - \mathbf{2.0} \times (-1) = 0.5$	0.25
2	$-0.6 - \mathbf{2.0} \times 0 = -0.6$	0.36
3	$2.1 - \mathbf{2.0} \times 1 = 0.1$	0.01
sum	0	0.62

Among these alternatives, $\beta = 1.8$ has the lowest sum of squared residuals (it is actually the optimum). See Figure 2.4.

2.1.2 Multiple Regression

All the previous results still hold in a multiple regression—with suitable reinterpretations of the notation.

Consider the linear model

$$\begin{aligned} y_t &= x_{1t}\beta_1 + x_{2t}\beta_2 + \cdots + x_{kt}\beta_k + u_t \\ &= x_t'\beta + u_t, \end{aligned} \tag{2.8}$$

where y_t and u_t are scalars, x_t a $k \times 1$ vector, and β is a $k \times 1$ vector of the true coefficients. In this expression, one of the elements of x_t is typically a constant equal to one (and the intercept is its coefficient).

Remark 2.8 *(On notation) These notes typically denote a vector of regression coefficients by β. The distinction from the $y_t = \alpha + \beta x_t + u_t$ notation sometimes used for simple regressions should be clear from the context.*

Least squares minimizes the sum of the squared fitted residuals

$$\sum_{t=1}^{T}(y_t - x_t'b)^2, \tag{2.9}$$

by choosing the vector b. The first order conditions (zero derivatives) hold at the (optimal) values $\hat{\beta}$, and can then be written

$$\mathbf{0}_{kx1} = \sum_{t=1}^{T} x_t(y_t - x_t'\hat{\beta}) \text{ or } \sum_{t=1}^{T} x_t y_t = \sum_{t=1}^{T} x_t x_t'\hat{\beta}. \tag{2.10}$$

Solve this as

$$\hat{\beta} = \left(\sum_{t=1}^{T} x_t x_t'\right)^{-1} \sum_{t=1}^{T} x_t y_t. \tag{2.11}$$

If the regressors are orthogonal (for instance, $\Sigma x_{1t}x_{2t} = 0$) then the results from the multiple regression (2.31) are the same as those from a series of simple regressions: y_t regressed on x_{1t}, y_t regressed on x_{2t}, etc. (This is easy to see since in this case $\Sigma x_t x_t'$ is a diagonal matrix which carries over to the inverse.) This is an unlikely case, unless the regressors have been pre-processed to indeed be orthogonal.

Remark 2.9 *(Matrix notation*) Let X be a $T \times k$ matrix where row t is filled with the elements of x_t and let Y be a $T \times 1$ where element t is y_t. Then, $X'X = \Sigma_{t=1}^{T} x_t x_t'$ and $X'Y = \Sigma_{t=1}^{T} x_t y_t$, so (2.11) can also be written $\hat{\beta} = (X'X)^{-1}X'Y$.*

Example 2.10 *(OLS with 2 regressors) With 2 regressors ($k = 2$) denoted x_{1t} and x_{2t},*

$$x_t y_t = \begin{bmatrix} x_{1t}y_t \\ x_{2t}y_t \end{bmatrix} \text{ and } x_t x_t' = \begin{bmatrix} x_{1t} \\ x_{2t} \end{bmatrix} \begin{bmatrix} x_{1t} & x_{2t} \end{bmatrix} = \begin{bmatrix} x_{1t}x_{1t} & x_{1t}x_{2t} \\ x_{2t}x_{1t} & x_{2t}x_{2t} \end{bmatrix}.$$

This means that (2.10) is

$$\begin{bmatrix} 0 \\ 0 \end{bmatrix} = \sum_{t=1}^{T} \begin{bmatrix} x_{1t}(y_t - x_{1t}\hat{\beta}_1 - x_{2t}\hat{\beta}_2) \\ x_{2t}(y_t - x_{1t}\hat{\beta}_1 - x_{2t}\hat{\beta}_2) \end{bmatrix}$$

and (2.11) is

$$\begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \left(\sum_{t=1}^{T} \begin{bmatrix} x_{1t}x_{1t} & x_{1t}x_{2t} \\ x_{2t}x_{1t} & x_{2t}x_{2t} \end{bmatrix}\right)^{-1} \sum_{t=1}^{T} \begin{bmatrix} x_{1t}y_t \\ x_{2t}y_t \end{bmatrix}.$$

Example 2.11 *(OLS with constant and one more regressor) In Example 2.10, let* $x_{1t} = 1$. *The first order conditions are then*

$$\begin{bmatrix} 0 \\ 0 \end{bmatrix} = \sum_{t=1}^{T} \begin{bmatrix} y_t - \hat{\beta}_1 - x_{2t}\hat{\beta}_2 \\ x_{2t}(y_t - \hat{\beta}_1 - x_{2t}\hat{\beta}_2) \end{bmatrix}.$$

The first line implies that $\hat{\beta}_1 = \bar{y}_t - \bar{x}_{2t}\hat{\beta}_2$ *(since dividing* 0 *by* T *is still* 0*). Using this in the second line to replace* $\hat{\beta}_1$ *and noticing that it does not matter if the term outside the parenthesis is* x_{2t} *or* $x_{2t} - \bar{x}_{2t}$ *(since the term in parenthesis is zero on average) gives* $\Sigma(x_{2t} - \bar{x}_{2t})[(y_t - \bar{y}_t) - (x_{2t} - \bar{x}_{2t})\hat{\beta}_2] = 0$. *We can then solve as* $\hat{\beta}_2 = \Sigma(x_{2t} - \bar{x}_{2t})(y_t - \bar{y}_t)/\Sigma(x_{2t} - \bar{x}_{2t})^2$, *which is the sample covariance of* x_{2t} *and* y_t *divided by the sample variance of* x_{2t} *(divide both numerator and denominator by* T *to see this).*

Example 2.12 *(Regression with an intercept and slope) Suppose we have the following data:*

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} -1.5 \\ -0.6 \\ 2.1 \end{bmatrix}, x_1 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}, x_2 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \text{and } x_3 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

This is clearly the same as in Example 2.6, except that we allow for an intercept (which turns out to be zero in this particular example). The notation we need to solve this problem is the same as for a general multiple regression. Therefore, calculate the following:

$$\begin{aligned} \sum_{t=1}^{T} x_t x_t' &= \begin{bmatrix} 1 \\ -1 \end{bmatrix}\begin{bmatrix} 1 & -1 \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \end{bmatrix}\begin{bmatrix} 1 & 0 \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \end{bmatrix}\begin{bmatrix} 1 & 1 \end{bmatrix} \\ &= \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \\ &= \begin{bmatrix} 3 & 0 \\ 0 & 2 \end{bmatrix} \end{aligned}$$

$$\begin{aligned} \sum_{t=1}^{T} x_t y_t &= \begin{bmatrix} 1 \\ -1 \end{bmatrix}(-1.5) + \begin{bmatrix} 1 \\ 0 \end{bmatrix}(-0.6) + \begin{bmatrix} 1 \\ 1 \end{bmatrix} 2.1 \\ &= \begin{bmatrix} -1.5 \\ 1.5 \end{bmatrix} + \begin{bmatrix} -0.6 \\ 0 \end{bmatrix} + \begin{bmatrix} 2.1 \\ 2.1 \end{bmatrix} \\ &= \begin{bmatrix} 0 \\ 3.6 \end{bmatrix} \end{aligned}$$

To calculate the LS estimate, notice that the inverse of the $\sum_{t=1}^{T} x_t x_t'$ *is*

$$\begin{bmatrix} 3 & 0 \\ 0 & 2 \end{bmatrix}^{-1} = \begin{bmatrix} 1/3 & 0 \\ 0 & 1/2 \end{bmatrix},$$

which can be verified by

$$\begin{bmatrix} 1/3 & 0 \\ 0 & 1/2 \end{bmatrix} \begin{bmatrix} 3 & 0 \\ 0 & 2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

The LS estimate is therefore

$$\begin{aligned} \hat{\beta} &= \left(\sum_{t=1}^{T} x_t x_t' \right)^{-1} \sum_{t=1}^{T} x_t y_t \\ &= \begin{bmatrix} 1/3 & 0 \\ 0 & 1/2 \end{bmatrix} \begin{bmatrix} 0 \\ 3.6 \end{bmatrix} \\ &= \begin{bmatrix} 0 \\ 1.8 \end{bmatrix}. \end{aligned}$$

Remark 2.13 *(The Frisch-Waugh-Lovell theorem*) Split up* x_t *into the* ***vectors*** x_{1t} *and* x_{2t} *and write (2.8) as* $y_t = x_{1t}'\beta_1 + x_{2t}'\beta_2 + u_t$*. First, regress* y_t *on* x_{1t} *and get the residuals* $\tilde{e}_1$*. Second, regress* x_{2t} *on* x_{1t} *and get the residuals* $\tilde{x}_{2t}$*. Third, regress* $\tilde{e}_1$ *on* $\tilde{x}_{2t}$*. This gives the same estimates as* β_2 *from the multiple regression of* y_t *on both* x_{1t} *and* x_{2t}*. (The proof is a straightforward reshuffling of the first order conditions, see, for instance, Greene (2018) 3.) The perhaps most common application of this is when* x_{1t} *contains various dummy variables (for instance, for different cross-sectional units) and* x_{2t} *are the variables of key interest. It can then be convenient to apply this 3-step approach. This is used in the fixed effects estimator for panel data.*

2.1.3 Least Squares: Goodness of Fit

The quality of a regression model is often measured in terms of its ability to explain the movements of the dependent variable.

Let $\hat{y}_t$ be the fitted (predicted) value of y_t. For instance, with (2.1) it would be $\hat{y}_t = \hat{\beta}_0 + \hat{\beta}_1 x_t$. If a constant is included in the regression (or the means of y and x are zero), then a check of the *goodness of fit* of the model is given by the fraction of the variation in

y_t that is explained by the model

$$R^2 = \frac{\text{Var}(\hat{y}_t)}{\text{Var}(y_t)} = 1 - \frac{\text{Var}(\hat{u}_t)}{\text{Var}(y_t)}, \tag{2.12}$$

which can also be rewritten as the squared correlation of the actual and fitted values

$$R^2 = \text{Corr}(y_t, \hat{y}_t)^2. \tag{2.13}$$

Notice that we must have constant in regression (unless both y_t and x_t have zero means) for R^2 to make sense.

A low variance of the residuals, $\text{Var}(\hat{u}_t)$, will be important for getting low standard errors of the estimates $(\hat{\beta}_0, \hat{\beta}_1, ...)$, since signal only little "noise" in the model. (The details are discussed in later sections.) Equation (2.12) shows that $\text{Var}(\hat{u}_t)$ and R^2 are negatively related, so it follows that a high R^2 will be associated with low standard errors of the estimates.

Example 2.14 *(R^2) From Example 2.6 we have* $\text{Var}(\hat{u}_t) = 0.18$ *and* $\text{Var}(y_t) = 2.34$*, so*

$$R^2 = 1 - 0.18/2.34 \approx 0.92.$$

See Figure 2.4.

Proof. (of (2.12)–(2.13)) Write the regression equation as

$$y_t = \hat{y}_t + \hat{u}_t,$$

where hats denote fitted values. Since $\hat{y}_t$ and $\hat{u}_t$ are uncorrelated (always true in OLS—provided the regression includes a constant), we have

$$\text{Var}(y_t) = \text{Var}(\hat{y}_t) + \text{Var}(\hat{u}_t).$$

R^2 is defined as the fraction of $\text{Var}(y_t)$ that is explained by the model

$$R^2 = \frac{\text{Var}(\hat{y}_t)}{\text{Var}(y_t)} = \frac{\text{Var}(y_t) - \text{Var}(\hat{u}_t)}{\text{Var}(y_t)} = 1 - \frac{\text{Var}(\hat{u}_t)}{\text{Var}(y_t)}.$$

Equivalently, we can rewrite R^2 by noting that

$$\text{Cov}\left(y_t, \hat{y}_t\right) = \text{Cov}\left(\hat{y}_t + \hat{u}_t, \hat{y}_t\right) = \text{Var}\left(\hat{y}_t\right).$$

Use this in the denominator of R^2 and multiply by $\text{Cov}(y_t, \hat{y}_t) / \text{Var}(\hat{y}_t) = 1$

$$R^2 = \frac{\text{Cov}(y_t, \hat{y}_t)^2}{\text{Var}(y_t)\,\text{Var}(\hat{y}_t)} = \text{Corr}(y_t, \hat{y}_t)^2 .$$

■

To understand this result, suppose that x_t has no explanatory power, so R^2 should be zero. How does that happen? Well, if x_t is uncorrelated with y_t, then $\hat{\beta}_1 = 0$. As a consequence $\hat{y}_t = \hat{\beta}_0$, which is a constant. This means that R^2 in (2.12) is zero, since the fitted residual has the same variance as the dependent variable ($\hat{y}_t$ captures nothing of the movements in y_t). Similarly, R^2 in (2.13) is also zero, since a constant is always uncorrelated with anything else (as correlations measure comovements around the means).

Remark 2.15 *(R^2 from simple regression*) Suppose $\hat{y}_t = \beta_0 + \beta_1 x_t$, so (2.13) becomes*

$$R^2 = \frac{\text{Cov}(y_t, \beta_0 + \beta_1 x_t)^2}{\text{Var}(y_t)\,\text{Var}(\beta_0 + \beta_1 x_t)} = \frac{\text{Cov}(y_t, x_t)^2}{\text{Var}(y_t)\,\text{Var}(x_t)} = \text{Corr}(y_t, x_t)^2.$$

The R^2 can never decrease as we add more regressors, which might make it attractive to add more and more regressors. To avoid that, some researchers advocate using an ad hoc punishment for many regressors, $\bar{R}^2 = 1 - (1 - R^2)(T - 1)/(T - k)$, where k is the number of regressors (including the constant). This measure can be negative.

Empirical Example 2.16 *(CAPM regressions) See Table 2.1 for CAPM regressions for two industry portfolios where the R^2 values clearly differ. This is seen also from the dispersion around the regression line in Figure 2.5.*

2.2 Missing Data

It is often the case that some data is missing For instance, we may not have data on regressor 3 for observation $t = 7$. If data is *missing in a random way*, then we can simply exclude (y_t, x_t) for the t with some missing data. In contrast, if data is missing in a non-random way (for instance, depending on the value of y_{it}), then we have to apply more sophisticated sample-selection models (not discussed in this chapter).

Remark 2.17 *(Replacing missing values with 0*) Instead of excluding (y_t, x_t) for the t with some missing data, we could set $(y_t, x_t) = (0, \mathbf{0}_k)$. This would not change the estimates, but it could lead to the wrong standard errors unless we are careful (see below for details).*

2.3 The Distribution of $\hat{\beta}$

Note that the estimated coefficients are random variables since they depend on which particular sample that has been "drawn." For instance, if our sample gives $\hat{\beta}_1 = 0.85$ and we know that the standard deviation across samples is 0.1 then we are pretty sure that the true value β_1 is not 0. In contrast, if the standard deviation across samples is 2.1, then our result is not such an unlikely outcome even if the true value β_1 is 0.

It is important to remember that we always assume that there are some true (but unknown) parameter values that would be the same across samples. The only reason why the estimates differ across samples is that the model is not perfect: there are residuals and they differ (randomly) across observations and thus also across different samples. See Figure 2.6 for an illustration from a computer simulation (Monte Carlo simulation).

We usually do not have several samples, so the variation across samples is not directly observable. However, we can (under some assumptions) use the *variation within our sample to figure out how the variation across samples ought to be*. This can help us testing hypotheses about the coefficients, for instance, that $\beta_1 = 0$.

To see where the uncertainty comes from, consider the simple case of only one regressor and a zero constant in (2.7). Use (2.1) to substitute for y_t (recall $\beta_0 = 0$)

$$\begin{aligned}\hat{\beta}_1 &= \frac{1}{\sum_{t=1}^{T} x_t x_t} \sum_{t=1}^{T} x_t \left(\beta_1 x_t + u_t\right) \\ &= \beta_1 + \frac{1}{\sum_{t=1}^{T} x_t x_t} \sum_{t=1}^{T} x_t u_t, \end{aligned} \tag{2.14}$$

so the OLS estimate, $\hat{\beta}_1$, equals the true value, β_1, plus the sample covariance of x_t and u_t divided by the sample variance of x_t. Since u_t is a random variable, $\hat{\beta}_1$ is too. Clearly, we do not know the true value β_1, so this decomposition is just conceptual.

When there are several regressors (x_t is a vector with k elements), then (2.14) becomes an expression for the vector

$$\hat{\beta} = \beta + \left(\sum_{t=1}^{T} x_t x_t'\right)^{-1} \sum_{t=1}^{T} x_t u_t, \tag{2.15}$$

where $\Sigma x_t x_t'$ is a $k \times k$ matrix and $\Sigma x_t u_t$ is a $k \times 1$ vector.

One of the basic assumptions in (2.1) is that the covariance of the regressor and the residual is zero. This should hold in a very large sample (or else OLS cannot be used to estimate β_1), but in a small sample it may be different from zero. Only as the sample gets very large can we be (almost) sure that the second term in (2.14) vanishes.

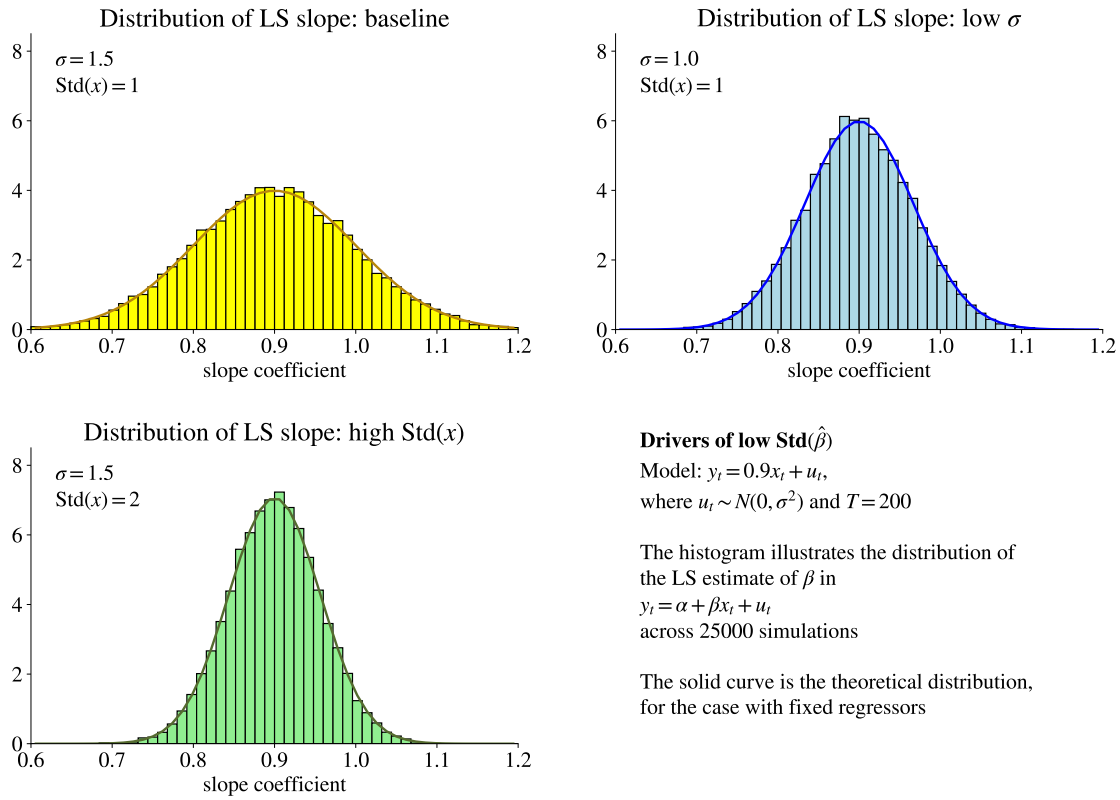


Figure 2.6: Distribution of OLS estimate, from simulation and theory

Equation (2.14) will give different values of $\hat{\beta}$ when we use different samples, that is different draws of the random variables x_t and y_t (and thus u_t). Since the true value, β, is a fixed constant, the distribution of these estimates across samples would describe the uncertainty we should have about the true value after having obtained a specific estimated value. However, we cannot observe this distribution directly (we do not have a lot of different samples). However, we can use the idea of this distribution to discuss the general properties of OLS—and we can (with some added assumptions) provide a good estimate of how that distribution could look like.

The first conclusion from (2.14) is that, with $u_t = 0$ the estimate would always be perfect. In contrast, with large movements in u_t we will see large movements in $\hat{\beta}$ (across samples). The second conclusion is that a small sample (small T) will also lead to large random movements in $\hat{\beta}_1$—in contrast to a large sample where the randomness in $\sum_{t=1}^{T} x_t u_t / T$ is averaged out more effectively (should be zero in a large sample).

There are three main routes to learn more about the distribution of $\hat{\beta}$: *(i)* set up a

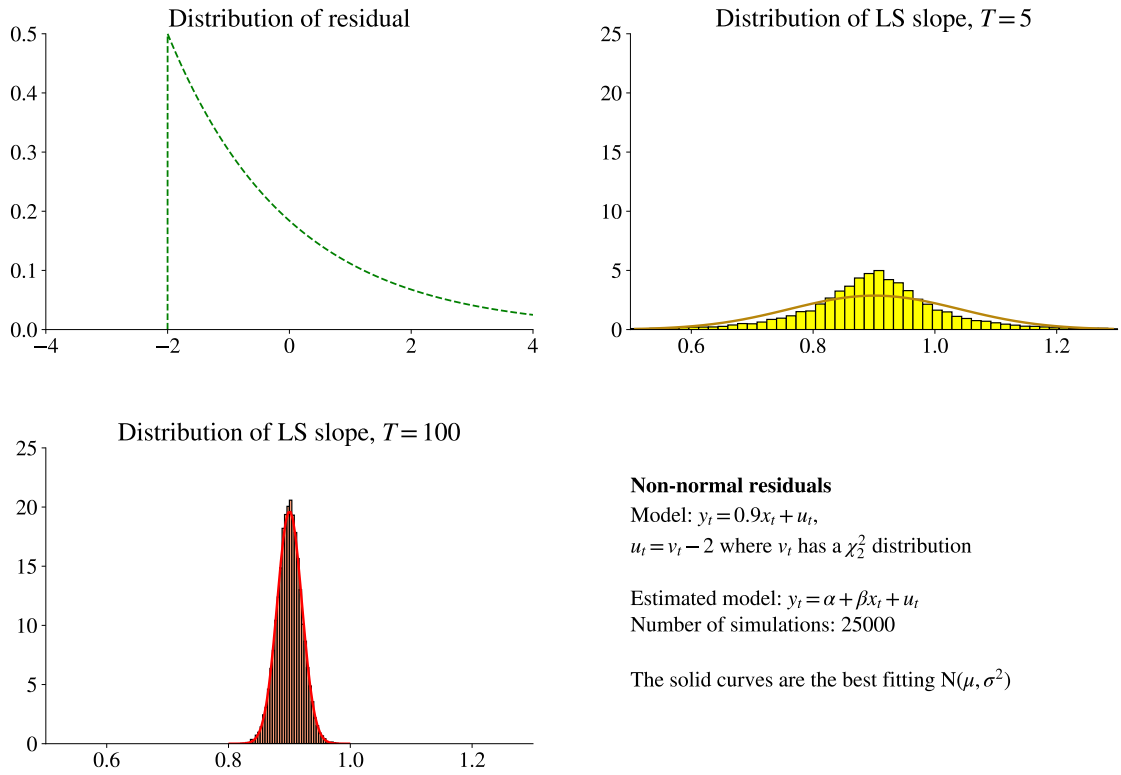


Figure 2.7: Distribution of OLS estimate, from simulations

small "experiment" in the computer and simulate the distribution (Monte Carlo or bootstrap simulations); *(ii)* pretend that the regressors can be treated as fixed numbers (or at least independent of the residuals in all periods) and then assume something about the distribution of the residuals; or *(iii)* use the asymptotic (large sample) distribution as an approximation. The asymptotic distribution can often be derived, in contrast to the exact distribution in a sample of a given size. If the actual sample is large, then the asymptotic distribution may be a good approximation.

The simulation approach has the advantage of giving a precise answer—but the disadvantage of requiring a very precise question (must write computer code that is tailor made for the particular model we are looking at, including the specific parameter values). See Figures 2.6, 2.9 and 2.7 for examples.

In contrast, asymptotic theory give more general results—but arriving there is hard. Treating the regressors as constants is easier—and is often good enough for illustrating the main properties of the estimation method.

The typical outcome of all three approaches will (under strong assumptions) be that

$$\hat{\beta} \sim N\left[\beta, \left(\sum_{t=1}^{T} x_t x_t'\right)^{-1} \sigma^2\right], \tag{2.16}$$

where σ^2 denoted the variance of the residuals, $\text{Var}(u_t)$.This expression allows for x_t to be a vector with k elements. Clearly, with $k = 1$, x_t is a scalar (and $x_t' = x_t$). In practice, we calculate/estimate both $\Sigma_{t=1}^{T} x_t x_t$ and σ^2 from the available data (the latter as the variance of the fitted residuals). See Table 2.1 for an empirical example and Figure 2.6 for an illustration of how the results depend on σ and the standard deviation of x_t.

Remark 2.18 *(Matrix notation*) Let X be a $T \times k$ matrix where row t is filled with the elements of x_t. Then, the variance-covariance matrix in (2.16) can also be written $(X'X)^{-1}\sigma^2$.*

Remark 2.19 *(Replacing missing values with 0*) If we set $(y_t, x_t) = (0, \mathbf{0}_k)$ is there is any missing value in (y_t, x_t) as suggested in Remark 2.17, then σ^2 in (2.16) should be multiplied by T_b/T, where T_b is the number of observations with data (not being missing values).*

Example 2.20 *(Applying (2.16)) When the regressor is just a constant (equal to one) $x_t = 1$, then we have*

$$\sum_{t=1}^{T} x_t x_t' = \sum_{t=1}^{T} 1 \times 1' = T \text{ so } \text{Var}(\hat{\beta}) = \sigma^2/T.$$

(This is the classical expression for the variance of a sample mean.)

Example 2.21 *(Applying (2.16)) When the regressor is a zero mean variable, then we have*

$$\sum_{t=1}^{T} x_t x_t' = \text{Var}(x_t)T \text{ so } \text{Var}(\hat{\beta}) = \sigma^2/\left[\text{Var}(x_t)T\right].$$

The variance is increasing in σ^2, but decreasing in both T and $\text{Var}(x_t)$.

Example 2.22 *(*Applying (2.16)) When the regressor is just a constant (equal to one) and one variable regressor with zero mean, f_t, so $x_t = [1, f_t]'$, then we have*

$$\sum_{t=1}^{T} x_t x_t' = \sum_{t=1}^{T} \begin{bmatrix} 1 & f_t \\ f_t & f_t^2 \end{bmatrix} = T \begin{bmatrix} 1 & 0 \\ 0 & \text{Var}(f_t) \end{bmatrix}, \text{ so}$$

$$\text{Var}\left(\begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix}\right) = \sigma^2 \left(\sum_{t=1}^{T} x_t x_t'\right)^{-1} = \begin{bmatrix} \sigma^2/T & 0 \\ 0 & \sigma^2/\left[\text{Var}(f_t)T\right] \end{bmatrix}.$$

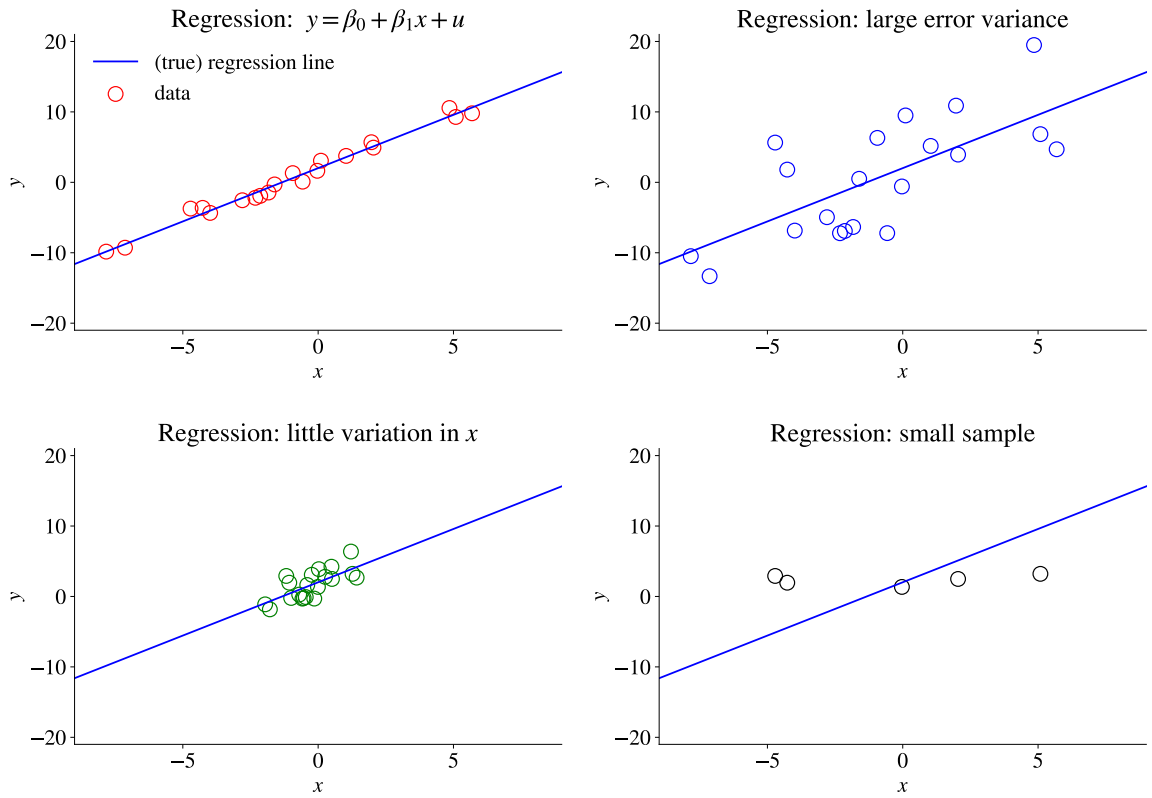


Figure 2.8: Regressions: importance of error variance and variation of regressor

This is combination of the two previous examples.

Example 2.23 *(Distribution of slope coefficient) From Example 2.6 we have* $\text{Var}(\hat{u}_t) = \sigma^2 = 0.18$ *and* $\sum_{t=1}^{T} x_t x_t = 2$, *so* $\text{Var}(\hat{\beta}_1) = 0.18/2 = 0.09$, *which gives* $\text{Std}(\hat{\beta}_1) = 0.3$.

Example 2.24 *(Covariance matrix of* b_1 *and* b_2*) From Example 2.12*

$$\sum_{t=1}^{T} x_t x_t' = \begin{bmatrix} 3 & 0 \\ 0 & 2 \end{bmatrix} \text{ and } \sigma^2 = 0.18, \text{ then}$$

$$\text{Var}\left(\begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix}\right) = \begin{bmatrix} \text{Var}(\hat{\beta}_1) & \text{Cov}(\hat{\beta}_1, \hat{\beta}_2) \\ \text{Cov}(\hat{\beta}_1, \hat{\beta}_2) & \text{Var}(\hat{\beta}_2) \end{bmatrix}$$

$$= \begin{bmatrix} 1/3 & 0 \\ 0 & 1/2 \end{bmatrix} 0.18 = \begin{bmatrix} 0.06 & 0 \\ 0 & 0.09 \end{bmatrix}.$$

The standard deviations (also called standard errors) are therefore

$$\begin{bmatrix} \text{Std}(\hat{\beta}_1) \\ \text{Std}(\hat{\beta}_2) \end{bmatrix} = \begin{bmatrix} 0.24 \\ 0.3 \end{bmatrix}.$$

An alternative way of expressing the distribution (often used in conjunction with asymptotic) theory is

$$\sqrt{T}(\hat{\beta} - \beta) \sim N\left[0, \left(\textstyle\sum_{t=1}^{T} x_t x_t'/T\right)^{-1} \sigma^2\right]. \tag{2.17}$$

This is the same as (2.16). (To see that, divide the LHS of (2.17) by $\sqrt{T}$. Then, the variance on the RHS must be divided by T, which gives the same variance as in (2.16). Then, add β to the LHS, which changes the mean on the RHS to β. We then have (2.16).)

2.3.1 The Distribution of $\hat{\beta}$ with Fixed Regressors

The assumption of fixed regressors makes a lot of sense in controlled experiments, where we actually can generate different samples with the same values of the regressors (the heat or whatever). It makes much less sense in econometrics. However, it is easy to derive results for this case—and those results happen to be very similar to what asymptotic theory gives.

The results we derive below are based on the *Gauss-Markov assumptions*: (a) the residuals have zero means, (b) have constant variances and (c) are not correlated across observations. In other words, the *residuals are zero mean iid variables*. (As an alternative to assuming fixed regressors (as we do here), it can instead be assumed that the residuals and regressors are independent. This delivers very similar results.) We will also assume that the residuals are normally distributed (not part of the typical Gauss-Markov assumptions).

For notational convenience, write (2.15) as

$$\hat{\beta} = \beta + S_{xx}^{-1}\left(x_1 u_1 + x_2 u_2 + \ldots x_T u_T\right), \text{ where } S_{xx} = \textstyle\sum_{t=1}^{T} x_t x_t'. \tag{2.18}$$

Since x_t is assumed to be non-random, the expected value of this expression is

$$\mathrm{E}\,\hat{\beta} = \beta + S_{xx}^{-1}\left(x_1 \,\mathrm{E}\, u_1 + x_2 \,\mathrm{E}\, u_2 + \ldots x_T \,\mathrm{E}\, u_T\right) = \beta \tag{2.19}$$

since we always assume that the residuals have zero means (see (2.1)). This says that OLS is unbiased when the regressors are fixed (which does not always carry over to the

case of stochastic regressors). The interpretation is that we can expect OLS to give (on average) a correct answer. That is, if we could draw many different samples and estimate the slope coefficient in each of them, then the average of those estimates would be the correct number (β). Clearly, this is something we want from an estimation method (a method that was systematically wrong would not be very attractive).

Remark 2.25 *(Linear combination of normally distributed variables.) If the random variables z_t and v_t are normally distributed and independent of each other, then $a + bz_t + cv_t$ is normally distributed with a mean of $a + b\mu_z + c\mu_v$ and a variance of $b^2\sigma_z^2 + c^2\sigma_v^2$.*

Suppose $u_t \sim N(0, \sigma^2)$ and the residuals are independent of each other, then (2.18) shows that $\hat{\beta}$ is normally distributed. The reason is that $\hat{\beta}$ is just a constant (β) plus a linear combination of independent normally distributed residuals (with fixed regressors x_t and S_{xx}^{-1} can be treated as constants). It is straightforward to see that the mean of this normal distribution is β (the true value), since the rest is a linear combination of the residuals—and they all have a zero mean.

Finding the variance-covariance matrix of $\hat{\beta}$ is just slightly more complicated. Remember that we treat x_t as fixed numbers ("constants") and assume that the residuals are iid: they are uncorrelated with each other (follows from independently distributed) and have the same variances (follows from identically distributed). We also notice that the variance (-covariance) matrix of $x_1 u_1$ equals

$$\text{Var}(x_t u_t) = x_t x_t' \sigma_t^2. \tag{2.20}$$

where $\sigma_t^2 = \text{Var}(u_t)$ and where we use the fact that the vector x_t is non-random.

Example 2.26 *(of (2.20)) With*

$$x_t = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \text{ and } \sigma_t^2 = 0.18, \text{ we get}$$

$$\text{Var}(x_t u_t) = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \begin{bmatrix} 1 & 2 \end{bmatrix} \times 0.18 = \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix} \times 0.18.$$

The variance of (2.18) can then be written

$$\begin{aligned}
\text{Var}(\hat{\beta}) &= S_{xx}^{-1} \text{Var}\left(x_1 u_1 + x_2 u_2 + \ldots x_T u_T\right) S_{xx}^{-1} \\
&= S_{xx}^{-1} \left(x_1 x_1' \sigma_1^2 + x_2 x_2' \sigma_2^2 + \ldots x_T x_T' \sigma_T^2\right) S_{xx}^{-1} \\
&= S_{xx}^{-1} \left(x_1 x_1' \sigma^2 + x_2 x_2' \sigma^2 + \ldots x_T x_T' \sigma^2\right) S_{xx}^{-1} \\
&= S_{xx}^{-1} \left(\sum_{t=1}^{T} x_t x_t'\right) \sigma^2 S_{xx}^{-1} \\
&= S_{xx}^{-1} \sigma^2. \qquad (2.21)
\end{aligned}$$

The first line follows directly from (2.18), since β is a constant. The second line follows from assuming that the residuals are uncorrelated with each other ($\text{Cov}(u_i, u_j) = 0$ if $i \neq j$), so all cross terms ($x_i x_j \text{Cov}(u_i, u_j)$) are zero. The third line follows from assuming that the variances are the same across observations ($\sigma_i^2 = \sigma_j^2 = \sigma^2$). The fourth and fifth lines are just algebraic simplifications which use the definition of S_{xx}.

There are three main ways of getting a low uncertainty (low $\text{Var}(\hat{\beta})$). For simplicity, focus on the case with just one regressor. We then have the following results. First, a large sample (T is large), decreases the S_{xx}^{-1} factor (since $S_{xx} = \Sigma_{t=1}^{T} x_t x_t'$ increases with T) while σ^2 stays constant: a larger sample gives a smaller uncertainty about the estimate. Second, large movements in the regressors (large value of $S_{xx} = \Sigma_{t=1}^{T} x_t x_t'$) should help us estimate the link between x and y since the movements in y driven by x should then dominate over the movements in y driven by the residual. Third, a lower volatility of the residuals (lower σ^2) also gives a lower uncertainty about the estimate. See Figures 2.6 and 2.8.

A key assumption in regression analysis is that our sample is "representative" of the population. In practice, this means that *we can estimate* both S_{xx} and σ^2 in (2.21) *from the data in the sample*. This is the main "trick" behind using our (one and only) sample to inform us about how the distribution of $\hat{\beta}$ (across samples) looks like. This is a plausible assumption when our sample is a random draw from the population (say, 700 out of a total of 10,000 firms). It is perhaps a stronger assumption when the sample is a time series of data. Then we basically assume that the past (before the sample) and the future (after the sample) will have the same structure. In case you are not willing to accept those assumptions, the t-stats are useless for you.

2.3.2 Multicollinearity

When the regressors in a multiple regression are highly correlated, then we have a practical problem: the standard errors of individual coefficients tend to be large, even if the R^2 suggests that the regression does fairly well.

As a simple example, consider the regression

$$y_t = \beta_1 x_{1t} + \beta_2 x_{2t} + u_t, \tag{2.22}$$

where (for simplicity) the dependent variable and the regressors have zero means. In this case, the variance (assuming iid errors) is

$$\text{Var}(\hat{\beta}_2) = \frac{\sigma^2}{T \text{ Var}(x_{2t})} \frac{1}{1 - \text{Corr}(x_{1t}, x_{2t})^2}, \tag{2.23}$$

where the new term is the (squared) correlation. If the regressors are highly correlated, then the uncertainty about the slope coefficient is high. The basic reason is that we see that the regressors have an effect on y_t, but it is hard to tell if that effect is from regressor one or two (since they are so similar). This can well lead to a situation where the R^2 is high and a joint test easily rejects the null hypothesis that all slopes are zero—but each individual slope coefficient is insignificant.

More generally, in the multiple regression

$$y_t = x_t' \beta + u_t, \tag{2.24}$$

it is straightforward to show that for all slope coefficients (not the intercept)

$$\text{Var}(\hat{\beta}_i) = \frac{\sigma^2}{T \text{ Var}(x_{it})} \frac{1}{1 - R_i^2}, \tag{2.25}$$

where R_i^2 is the R^2 value obtained from regressing x_{it} on the other regressors (including a constant). The last term $(1/(1 - R_i^2))$ is often called the *variance inflation factor* and some regression packages report the maximum across the regressors, and a value of 10 or larger ($R_i^2 \geq 0.9$) is considered highly problematic. (The name variance inflation factor is meant to indicate how much the variance increases compared to a simple regression, assuming σ^2 is unchanged. In practice, the estimated σ^2 often change considerably.)

Proof. (of 2.23*). Recall that for a 2×2 matrix we have

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}.$$

For the regression (2.22) we get

$$\begin{bmatrix} \sum_{t=1}^{T} x_{1t}^2 & \sum_{t=1}^{T} x_{1t}x_{2t} \\ \sum_{t=1}^{T} x_{1t}x_{2t} & \sum_{t=1}^{T} x_{2t}^2 \end{bmatrix}^{-1} = \frac{1}{\sum_{t=1}^{T} x_{1t}^2 \sum_{t=1}^{T} x_{2t}^2 - \left(\sum_{t=1}^{T} x_{1t}x_{2t}\right)^2} \begin{bmatrix} \sum_{t=1}^{T} x_{2t}^2 & -\sum_{t=1}^{T} x_{1t}x_{2t} \\ -\sum_{t=1}^{T} x_{1t}x_{2t} & \sum_{t=1}^{T} x_{1t}^2 \end{bmatrix}.$$

The variance of the second slope coefficient is σ^2 time the lower right element of this matrix. Multiply and divide by T to get

$$\begin{aligned} \operatorname{Var}(\hat{\beta}_2) &= \frac{\sigma^2}{T} \frac{\sum_{t=1}^{T} x_{1t}^2/T}{\sum_{t=1}^{T} \frac{1}{T} x_{1t}^2 \sum_{t=1}^{T} \frac{1}{T} x_{2t}^2 - \left(\sum_{t=1}^{T} \frac{1}{T} x_{1t}x_{2t}\right)^2} \\ &= \frac{\sigma^2}{T} \frac{\operatorname{Var}(x_{1t})}{\operatorname{Var}(x_{1t}) \operatorname{Var}(x_{2t}) - \operatorname{Cov}(x_{1t}, x_{2t})^2} \\ &= \frac{\sigma^2}{T} \frac{1/\operatorname{Var}(x_{2t})}{1 - \frac{\operatorname{Cov}(x_{1t}, x_{2t})^2}{\operatorname{Var}(x_{1t}) \operatorname{Var}(x_{2t})}}, \end{aligned}$$

which is the same as (2.23). ■

2.4 The Distribution of $\hat{\beta}$: More General Results

2.4.1 Problems with the Gauss-Markov (iid) and Normality Assumptions

The previous results on the distribution of $\hat{\beta}$ have several weak points—which will be briefly discussed here.

First, the Gauss-Markov assumptions of iid residuals (constant volatility and no correlation across observations) are likely to be false in many cases. These issues (heteroskedasticity and autocorrelation) are therefore discussed at length later on.

Second, the idea of fixed regressor is clearly just a simplifying assumption—and unlikely to be relevant for economics and financial data. If the regressors are random variables then we typically not rule out that u_t and x_{t+s} are correlated, for instance, when the regressors include the lagged dependent variable. This can make OLS biased in small

samples, although the OLS estimate might converge to the true values (so OLS is "consistent") as the sample size increases.

Third, there are no particularly strong reasons for why the residuals should be normally distributed. If not, the estimates are unlikely to be normally distributed in small samples, but may well be in large samples (due to the central limit theorem). This is discussed in some detail below.

The next few sections introduce these issues, but later chapters will discuss them in more detail.

2.4.2 Failure of the Gauss-Markov Assumptions

If the residuals are not iid, then we have to stop at the first line of (2.21), so

$$\text{Var}(\hat{\beta}) = S_{xx}^{-1} S S_{xx}^{-1}, \text{ where } S = \text{Var}\left(\Sigma_{t=1}^{T} x_t u_t\right). \tag{2.26}$$

The S matrix is estimated in different ways (for instance, using White's or Newey-West's methods) depending on the properties of the residuals (heteroskedasticity or autocorrelation).

2.4.3 Bias

If an estimation method is *biased*, then it produces systematically wrong (say, too low) coefficients.

Figure 2.9 illustrates some simulation results from estimating an AR(1)

$$y_t = \beta_0 + \beta_1 y_{t-1} + u_t \tag{2.27}$$

on artificially generated samples where "data" follow

$$y_t = 0.9 y_{t-1} + u_t, \text{ where } u_t \text{ is iid.} \tag{2.28}$$

In this case, the regressor is a (stochastic) random variable (not fixed). Figure 2.9 suggests that the estimates are biased (not centered on the true value) in small samples.

To understand these results, recall that (2.15) says that

$$\hat{\beta} = \beta + \left(\sum_{t=1}^{T} x_t x_t'\right)^{-1} \sum_{t=1}^{T} x_t u_t \tag{2.29}$$

where u_t are the true residuals. We will never observe the true residuals, so (2.29) can

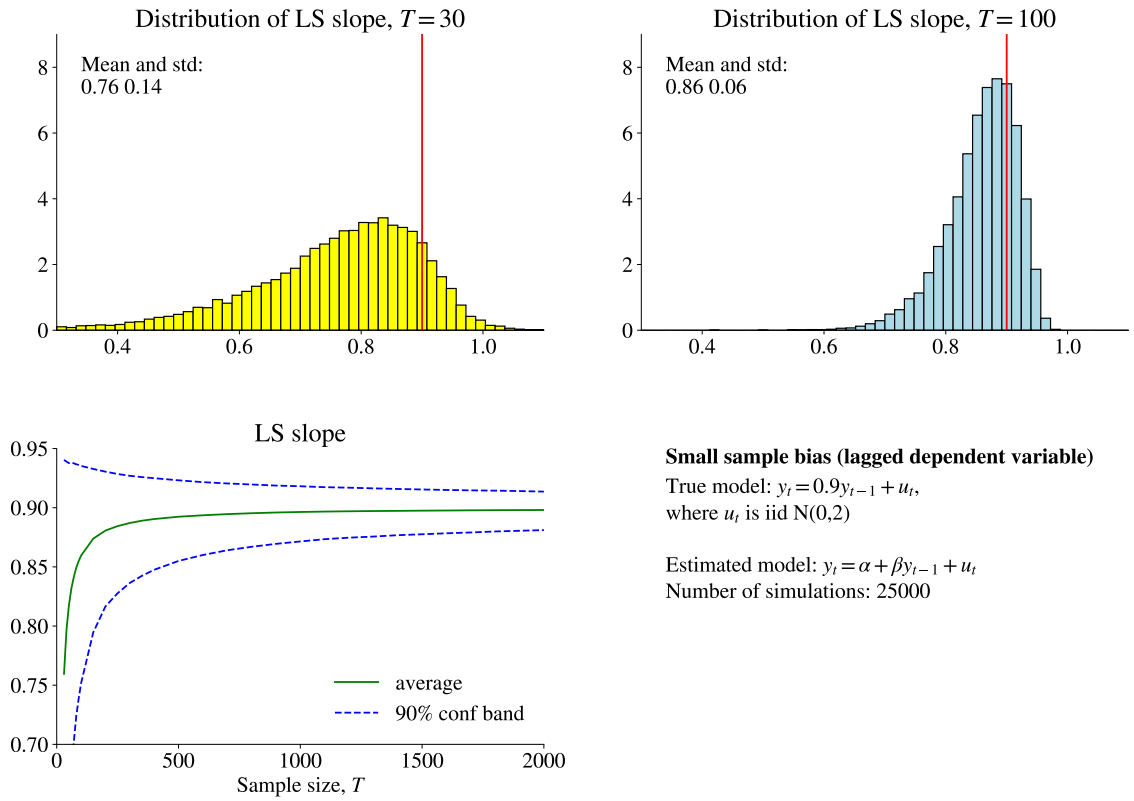


Figure 2.9: Results from a Monte Carlo experiment of LS estimation of the AR coefficient.

only be used for a conceptual discussion.

To get *unbiased estimates* ($\mathrm{E}\,\hat{\beta} = \beta$), the second term of the right hand side of (2.29) should have an expectation of zero. This would happen when u_t and x_{t+s} (for all s) are independent. This is hard to guarantee when the regressors are random variables. For instance, in the AR(1) example, then u_t affects x_{t+1} so there is an interaction between the numerator and denominator. This is probably most easily investigated by (Monte Carlo) simulations. In many cases, the bias decreases rapidly as the sample size increases (see the discussion of "consistency").

Remark 2.27 *(*Bias of AR(1)) It can be shown (see, for instance, Pesaran (2015) 14) that a bias corrected estimate of the AR(1) coefficient can be calculated as* $(1+T\hat{\beta}_1)/(T-3)$.

Remark 2.28 *(Unbiasedness with stochastic regressors*)* $\hat{\beta}_1$ *in (2.29) is unbiased if* x_τ *(all* τ *in the sample,* $1-T$*) does not help predict* u_t*. If we we can write use the law of*

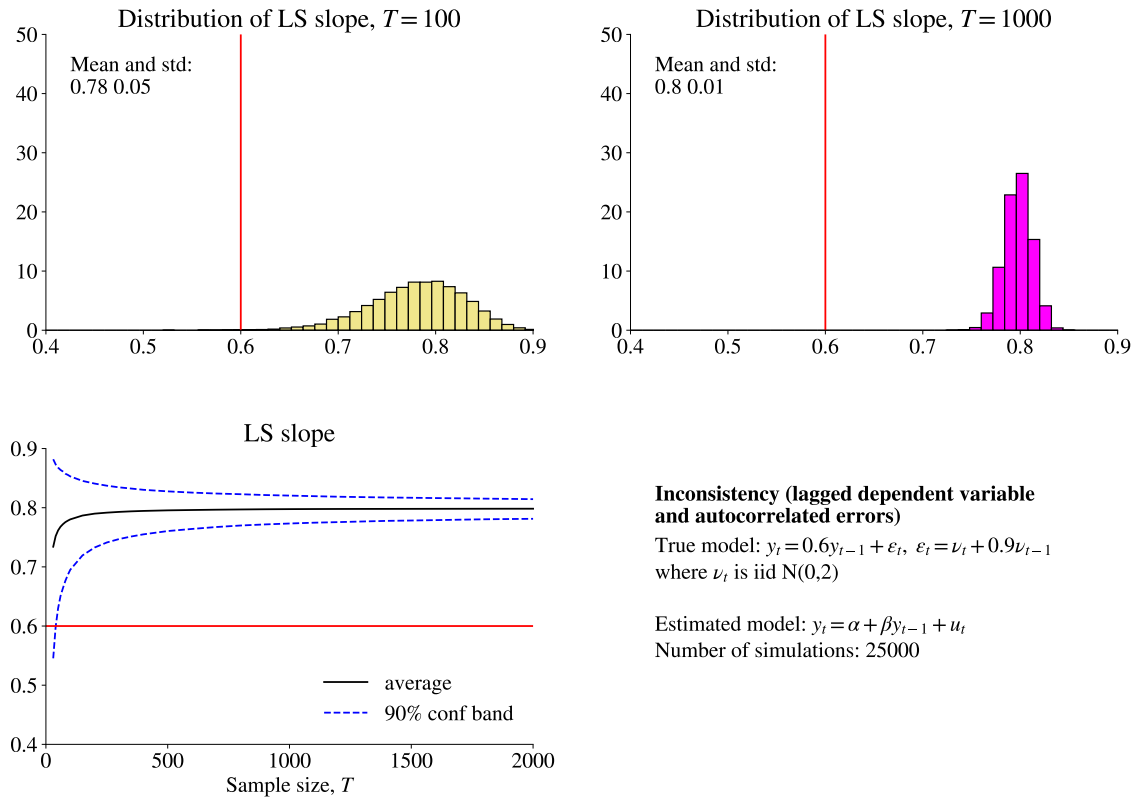


Figure 2.10: Results from a Monte Carlo experiment of LS estimation of the AR coefficient when data is from an ARMA process.

iterated expectations ($\mathrm{E}_x(\mathrm{E}\,u|x) = \mathrm{E}\,u$) *to write*

$$\mathrm{E}\,\hat{\beta} = \beta + \mathrm{E}_x\left(\sum_{t=1}^{T} x_t x_t'\right)^{-1} \sum_{t=1}^{T} x_t\,\mathrm{E}(u_t|x),$$

where x *here denotes the full sample of* x_t. *If* $\mathrm{E}(u_t|x) = 0$ *then* $\mathrm{E}\,\hat{\beta} = \beta$. *Clearly, for an AR(1) this does* not *hold since* u_t *affects* y_t *which is the regressor in* $t+1$ (x_{t+1}).

2.4.4 Consistency

If an estimation method is *inconsistent*, then it produces systematically wrong (say, too low) coefficients also in very large samples (actually, in the limit as $T \to \infty$).

Figure 2.9 suggests that the problem with the AR(1) estimation vanishes as the sample size increases. This suggests the importance of doing simulations (to understand the properties of the estimation method)—and of using large data sets.

To get *consistent estimates* (which is defined as the bias and the variance of $\hat{\beta}_1$ go to

zero as $T \to \infty$), then it is enough if x_t and u_t (in the same period) are uncorrelated. This is indeed the case in the AR(1) simulations discussed before. To see this from (2.29), notice that a "law of large numbers" makes the numerator ($\Sigma_{t=1}^{T} x_t u_t / T$) converge to the population covariance of u_t and x_t. (Also, the denominator converges to a fixed number, so we can focus on the numerator.)

This means that *if* we knew that $\text{Cov}(x_t, u_t) = 0$ (in the population), then we would also know that OLS is consistent. However, since the true errors are never observed, this cannot be shown by empirical methods. (Recall OLS always construct fitted errors so they are uncorrelated with the regressors.) Instead, we have to rely on theoretical arguments that make it plausible to *believe or not* in consistency.

To make matters worse, it is often the case that $\hat{\beta}_1$ converges (as T increases), but perhaps not to what you hoped for. As an illustration of how tricky this can be, consider the case in Figure 2.10. It estimates the same AR(1) as in (2.27) but where the simulated "data" now follows

$$y_t = \rho y_{t-1} + \varepsilon_t, \text{ where } \varepsilon_t = v_t + \theta v_{t-1} \text{ where } v_t \text{ is iid.} \tag{2.30}$$

In this case, the residuals (here called ε_t) are themselves autocorrelated. The figure clearly shows that the OLS estimate of the slope β_1 in (2.27) does *not* converge to the true value ρ as the sample sizes increases: OLS is inconsistent. The reason in this case is that ε_t and y_{t-1} (the regressor) both depend on v_{t-1} so they are correlated.

An a priori argument for why OLS should be able to estimate a model consistently thus require a careful discussion of the model properties: how can we explain that the residuals are uncorrelated with the regressors? (Alternatively, we use an instrumental variables technique, which is discussed later on.) This typically involves a discussion of the following.

1. Have we excluded (omitted) some relevant regressors? If so, their effect is captured by the residual. If these excluded regressors are correlated to some of the included regressors, then we have a problem.

2. Do we use a lagged dependent variable as regressor at the same time as the residual is autocorrelated? (This is the previous example.)

3. Does y_t affect x_t? If so a shock to the equation that explains y_t also drives x_t and we get a correlation between the regressor (x_t) and the residual. A classical case is when we try to estimate how the demand for a product depends on its price.

In fact, such an equation actually estimates a mix between the demand and supply elasticities.

4. Is the regressor measured without (important) errors? If not, we again have a correlation between (the used) regressor and the residual.

2.4.5 Normality

If the regressors x_t are fixed numbers and u_t is normally distributed, then the second term in (2.29) shows that the normality carries over to $\hat{\beta}$ also in small samples. Actually, we can relax the assumption about the regressors (to allow them to be random) as long as we assume that x_t and u_t are independent (the same assumption as needed for unbiasedness). We can test the assumption of normally distributed residuals by using a Bera-Jarque test

$$\text{BJ} = \frac{T}{6}\text{skewness}^2 + \frac{T}{24}\left(\text{kurtosis} - 3\right)^2, \tag{2.31}$$

which has χ^2_2 distribution under the null hypothesis that both the skewness and excess kurtosis (that is, kurtosis-3) are zero.

Remark 2.29 *(Small sample distribution with stochastic regressors*) With stochastic regressors, the small-sample distribution of $\hat{\beta}$ is typically unknown. Even in the most restrictive case where u_t is iid $N(0,\sigma^2)$ and $\text{E}(u_t|x_\tau) = 0$ for all τ, we can only get that (2.16) holds* ***conditional*** *on the sample. More generally, $\hat{\beta}$ in (2.15) is a product of two random variables, $(\Sigma_{t=1}^T x_t x_t')^{-1}$ and $\Sigma_{t=1}^T x_t u_t$ and there is no strong reason to assume that this product is normally distributed just because u_t is.*

Even if the normality test fails, we can often still hope for a (close to) normal distribution of $\hat{\beta}_1$ if the sample is large—due to the central limit theorem. This is illustrated in Figure 2.7. It is based on simulations where the residual is drawn from a very non-normal distribution. For a small sample, this carries over to $\hat{\beta}_1$ and the t-stat for the hypothesis that $\beta_1 = 0$. However, already a moderately sized sample tend to give an almost normal distribution.

To understand the theory of this rewrite (2.29) by subtracting β_1 from both sides and then multiply both sides by $\sqrt{T}$ (2.15) says that

$$\sqrt{T}(\hat{\beta}_1 - \beta_1) = \left(\frac{1}{T}\sum_{t=1}^T x_t x_t'\right)^{-1} \sqrt{T}\frac{1}{T}\sum_{t=1}^T x_t u_t. \tag{2.32}$$

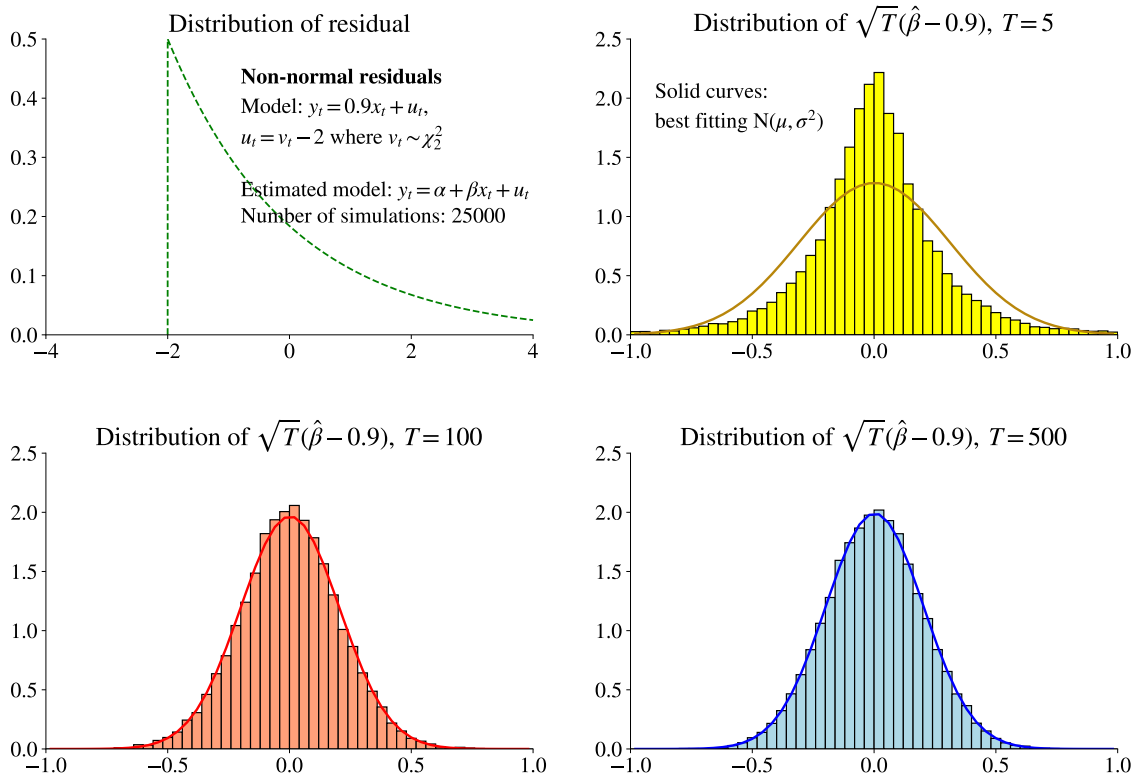


Figure 2.11: Distribution of OLS estimate, from simulations

The inverted term is the sample average of $x_t x_t'$ which will converge to a matrix of fixed numbers (the population mean of $x_t x_t'$) as $T \to \infty$. We can therefore focus on what happens to the numerator. It is $\sqrt{T}$ times the sample average of $x_t u_t$ (a vector). Under weak conditions a central limit theorem applies to $\sqrt{T}\times$a sample average: it typically converges to a normal distribution.

This shows that $\sqrt{T}\hat{\beta}_1$ has an *asymptotic normal distribution*. This often holds as a reasonable approximation also in moderately sized samples. See Figure 2.11 for an illustration.

Actually, it turns out that this is a property of many estimators (not just OLS), basically because most estimators are some kind of sample average. The properties of this distribution are quite similar to those that we derived by assuming that the regressors were fixed numbers.

Based on (2.32), the key result is that the distribution of $\sqrt{T}(\hat{\beta}_1 - \beta_1)$ converges to a

normal distribution (as T increases)

$$\sqrt{T}(\hat{\beta} - \beta) \xrightarrow{d} N\left(0, \Sigma_{xx}^{-1} \Sigma \Sigma_{xx}^{-1}\right), \tag{2.33}$$

where is the probability limit of $\Sigma_{t=1}^{T} x_t x_t'/T$ (essentially, the limit of the matrix as T increases) and $\Sigma = \text{Var}(\Sigma_{t=1}^{T} x_t u_t/\sqrt{T})$. Cancelling T terms and making a somewhat liberal interpretation gives

$$\text{“}\hat{\beta} \xrightarrow{d} \text{”} N\left(\beta, S_{xx}^{-1} S S_{xx}^{-1}\right), \tag{2.34}$$

where the covariance matrix is same as in (2.26).

Chapter 3

Least Squares: Testing

Reference: Verbeek (2012) 2 and 4; Greene (2018) 4-5 and parts of 9 and 20.
More advanced material is denoted by a star (*). It is not required reading.

3.1 Hypothesis Testing

3.1.1 Testing a Single Coefficient: A t-test

We are interested in testing the null hypothesis (H_0) that $\beta = q$, where q is a number of interest. (Econometric programs typically report results for H_0: $\beta = 0$.) Here, the alternative hypothesis is that $\beta \neq q$, so this is a two-sided (also called "two-tailed") test.

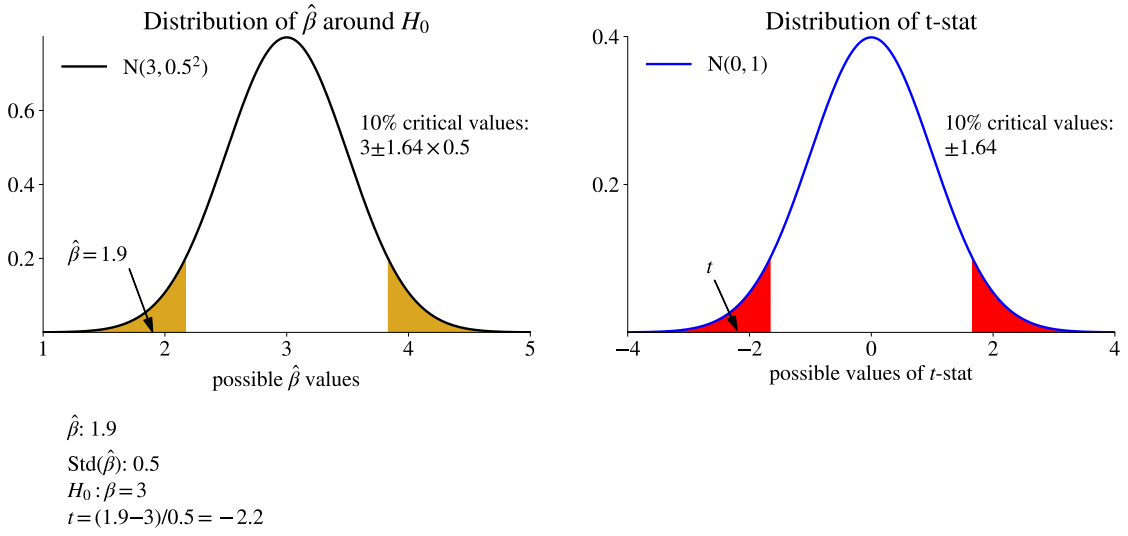


Figure 3.1: Distribution of $\hat{\beta}$ and t-stat

We assume that the estimates are normally distributed, which may be a good approximation when the sample is large (because of the central limit theorem). If the null hypothesis is true, then

$$\hat{\beta} \sim N(q, \text{Var}(\hat{\beta})). \tag{3.1}$$

To be able to easily compare with printed tables of probabilities, we transform to a $N(0,1)$ variable. In particular, if the true coefficient is really q, then $\hat{\beta} - q$ should have a zero mean. Dividing by the standard error (deviation) of $\hat{\beta}$, we should have

$$t = \frac{\hat{\beta} - q}{\text{Std}(\hat{\beta})} \sim N(0,1) \tag{3.2}$$

We reject the null hypothesis when $|t|$ is very large, for instance, if $|t| > 1.64$.

This decision is driven by *(a)* how far $\hat{\beta}$ is from q; *(b)* how uncertain $\hat{\beta}$ is (as measured by $\text{Std}(\hat{\beta})$); *(c)* and how we define the cut off (here 1.64). The latter is typically done by first choosing a *significance level* (for instance, 10%) which defines a *critical value* (1.64 for the 10% significance level): reject the null hypothesis if $|t|$ is larger that the critical value (1.64 on the 10% level, 1.96 on the 5% level). The significance level represents the probability, in a random sample, of falsely rejecting a null hypothesis that is actually true. See Figure 3.1 for an illustration. A lower significance level (5%, giving a critical value of 1.96) is therefore a more conservative test (we require stronger evidence to reject the null hypothesis), and thus run a lower risk of a false rejection. The significance level is thus a trade-off between actually being able of rejecting the null hypothesis and sometime doing it falsely. See Figure 3.2 for an illustration of the probabilities according to an N(0,1) distribution.

Otherwise, when $|t|$ is not very large (for instance, $|t|$), then evidence is not sufficient to reject the null hypothesis. (You may compare with a court of law where the null hypothesis is that the accused is not guilty.)

Example 3.1 *(t-test) Let* $\hat{\beta} = 1.9$, $\text{Std}(\hat{\beta}) = 0.5$ *and* $q = 3$. *Then,* $t = (1.9 - 3)/0.5 = -2.2$ *so* $|t| > 1.64$ *and also* $|t| > 1.96$. *The null hypothesis is thus rejected at both the 10% and the 5% significance levels.*

Empirical Example 3.2 *(CAPM regressions for industry portfolios) See Table 3.1.*

Empirical Example 3.3 *(Multi-factor regressions for industry portfolios) See Table 3.2.*

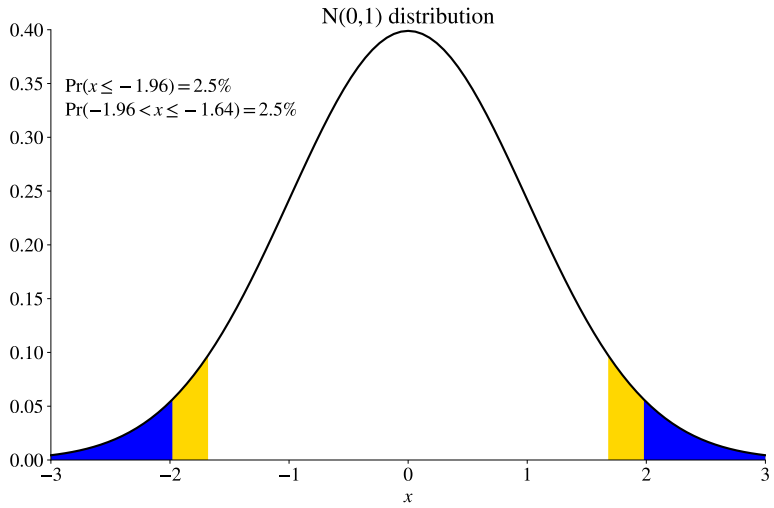


Figure 3.2: Density function of a standard normal distribution

The *p-value* is a related concept. It is the lowest significance level at which we can reject the null hypothesis: a lower number is a stronger rejection. See Figure 3.3 for an illustration.

Example 3.4 *(p-value) Continuing Example 3.1, notice that according to a N(0,1) distribution, the probability of* -2.2 *or lower is 1.4%, so the p-value is 2.8%. We thus reject the null hypothesis at the 10% significance level and also at the 5% significance level.*

We sometimes compare with a t-distribution instead of a $N(0,1)$, especially when the sample is short. For instance, with 22 data points and two estimated coefficients (so there are 20 degrees of freedom), the 10% critical value of a t-distribution is 1.72 (while it is 1.64 for the standard normal distribution). However, for samples of more than 30–40 data points, the difference is trivial.

Remark 3.5 *(One-sided test*) As an examples of a one-sided test let* $H_0 : \beta \leq q$ *and* $H_1 : \beta > q$*. Sometimes the null hypothesis is written* $\beta = q$*, but that makes little practical difference. We then reject the null hypothesis at the 10% significance level if* $t >$ 1.28 *which is the 0.90 quantile of a* $N(0,1)$ *distribution. Conversely, when* $H_0 : \beta \geq q$ *and* $H_1 : \beta < q$*, then we reject the null hypothesis if* $t < -1.28$*. Since 1.28 is the 20% critical value in a two-sided test, we can actually use a two sided test (for instance, from a regression package) to also do a one-sided test: (a) if we reject the null hypothesis on the 20% level in a double sided test; (b) and the sign is right; then (c) this is a rejection on the 10% level in a one-sided test.*

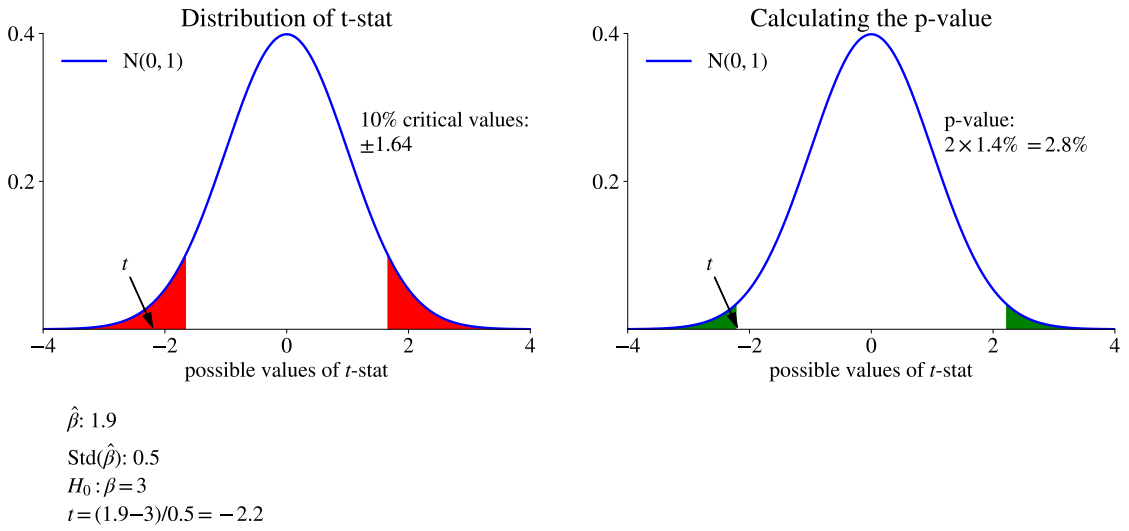


Figure 3.3: Calculating the p-value

3.1.2 Confidence Bands

A significance level of 10% means that there is (if the null hypothesis is true) a 90% probability that the t value in (3.2) is within the interval (band) $(-1.64, 1.64)$, that is,

$$\Pr(-1.64 \leq t \leq 1.64) = 90\%. \tag{3.3}$$

The t-test discussed above rejects the null hypothesis ($\beta = q$) when t is outside this confidence band. Notice that

$$t \text{ is outside } [-1.64, 1, 64] \iff \tag{3.4}$$

$$\hat{\beta} \text{ is outside } [q - 1.64\,\text{Std}(\hat{\beta}), q + 1.64\,\text{Std}(\hat{\beta})] \text{ and} \tag{3.5}$$

$$q \text{ is outside } [\hat{\beta} - 1.64\,\text{Std}(\hat{\beta}), \hat{\beta} + 1.64\,\text{Std}(\hat{\beta})]. \tag{3.6}$$

The interval in (3.5) is a 90% confidence band of β *centered on the null hypothesis*, while the confidence band in (3.6) is *centered on the point estimate*. These are alternative ways of doing a hypothesis test—and are often used to provide more information than just a reject/no reject decision. For instance, we are 90% sure that the true β value is within the band centered around the point estimate. See Figures 3.1 and 3.4.

Proof. (that t and $\hat{\beta}$ are outside their confidence bands at the same time) For $\hat{\beta}$ to be

	HiTec	Utils
constant	−0.07	0.24
	(−0.54)	(1.75)
market return	1.24	0.51
	(36.79)	(13.53)
R^2	0.75	0.32
Autocorr	0.38	0.97
White	0.03	0.00
All slopes	0.00	0.00
obs	624	624

Table 3.1: CAPM regressions, monthly returns, %, US data 1970:01-2021:12. Numbers in parentheses are t-stats. Autocorr is the p-value for no autocorrelation; White is the p-value for homoskedasticity; All slopes is the p-value for all slope coefficients being zero.

outside the band we must have

$$\hat{\beta} < q - 1.64\,\text{Std}(\hat{\beta}) \text{ or } \hat{\beta} > q + 1.64\,\text{Std}(\hat{\beta}).$$

Rearrange this by subtracting q from both sides of the inequalities and then divide both sides by $\text{Std}(\hat{\beta})$

$$\frac{\hat{\beta} - q}{\text{Std}(\hat{\beta})} < -1.64 \text{ or } \frac{\hat{\beta} - q}{\text{Std}(\hat{\beta})} > 1.64.$$

■

Example 3.6 *(t-test and confidence band around q) With* $\text{Std}(\hat{\beta}) = 0.5$ *and* $q = 3$*, the 90% confidence band is* $3 \pm 1.64 \times 0.5$*, that is,* $[2.18, 3.82]$*. Notice that* $\hat{\beta} = 1.90$ *is outside this band, so we reject the null hypothesis. Equivalently,* $t = (1.9 - 3)/0.5 = -2.2$ *is outside the band* $[-1.64, 1.64]$.

Example 3.7 *(t-test and confidence band around $\hat{\beta}$) With* $\text{Std}(\hat{\beta}) = 0.5$ *and* $\hat{\beta} = 1.9$*, the 90% confidence band is* $1.9 \pm 1.64 \times 0.5$*, that is,* $[1.08, 2.72]$*. Notice that* $q = 3$ *is outside this band, so we reject the null hypothesis.*

3.1.3 Power and Size*

The *size is the probability of rejecting a true* H_0. It should be low. Provided you use a valid test (correct standard error, etc), *the size is the significance level* you have cho-

	HiTec	Utils
constant	0.12	0.13
	(1.06)	(1.00)
market return	1.13	0.59
	(35.97)	(16.64)
SMB	0.21	−0.20
	(4.15)	(−3.97)
HML	−0.52	0.31
	(−10.75)	(5.35)
R^2	0.82	0.40
Autocorr	0.66	0.63
White	0.00	0.00
All slopes	0.00	0.00
obs	624	624

Table 3.2: Fama-French regressions, monthly returns, %, US data 1970:01-2021:12. Numbers in parentheses are t-stats. Autocorr the p-value for no autocorrelation; White is the p-value for homoskedasticity; All slopes is the p-value for all slope coefficients being zero.

sen (which defines the critical values). For instance, with a t-test with critical values $(-1.64, 1.64)$, the size is 10%. (The size is sometime called the type I error.) This means that we run a 10% chance of wrongly rejecting a true null hypothesis. See Table 3.3.

	H_0 not rejected	H_0 rejected
H_0 is true	1− size	size
H_0 is false	1− power	power

Table 3.3: Size and power

The *power is the probability of rejecting a false* H_0. It should be high. Typically, it cannot be controlled (but some tests are better than others...). This power depends on how false H_0 is, which we will never know. All we can do is to create (artificial) examples to get an idea of what the power would be for different tests and for different values of the true parameter β. For instance, with a t-test using the critical values -1.64 and 1.64, the power would be

$$\text{power} = \Pr(t \leq -1.64) + \Pr(t \geq 1.64). \tag{3.7}$$

(1−power is sometimes called the type II error. This is the probability of not rejecting a

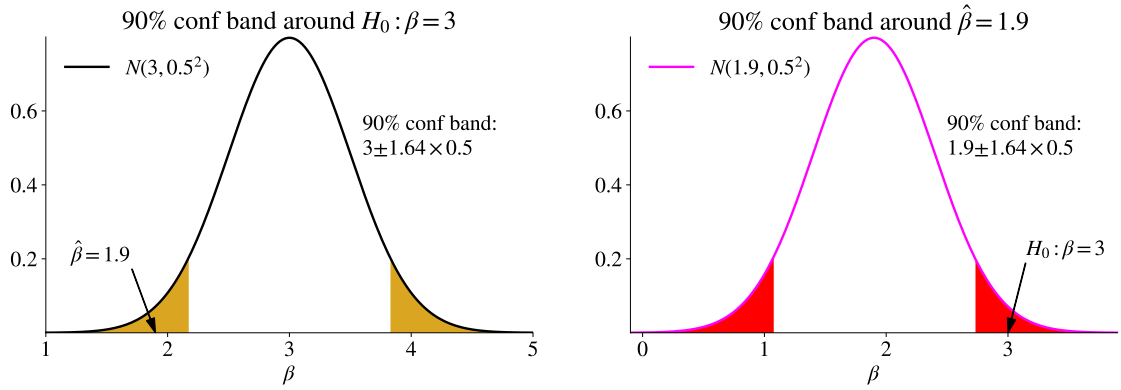


Figure 3.4: Confidence band around the null hypothesis or around the point estimate

false H_0.)

To make this more concrete, suppose we test the null hypothesis that the coefficient is equal to q, but the true value happens to be β. Since the OLS estimate, $\hat{\beta}$ is distributed as $N[\beta, \text{Std}(\hat{\beta})]$, it must be the case that the t-stat is distributed as

$$t = \frac{\hat{\beta} - q}{\text{Std}(\hat{\beta})} \sim N\left(\frac{\beta - q}{\text{Std}(\hat{\beta})}, 1\right). \tag{3.8}$$

We can then calculate the power as the probability that $t \leq -1.64$ or $t \geq 1.64$, when t has the distribution on the RHS in (3.8). Clearly, the results depend on what the true value β really is. See Figure 3.5.

Example 3.8 *If $\beta = 3.6$, $q = 3$ and $\text{Std}(\hat{\beta}) = 0.5$, then the power is 0.33.*

3.1.4 Testing A Linear Combination

We can form linear combinations of the regressions coefficients and apply a t-test.

Let R be a $1 \times k$ (row) vector that defines our linear combination and suppose we want to test $R\beta = q$. This is easily by nothing that

$$\text{Var}(R\hat{\beta}) = RV(\hat{\beta})R', \tag{3.9}$$

is a scalar, so the t-test becomes

$$t = \frac{R\hat{\beta} - q}{\sqrt{RV(\hat{\beta})R'}} \sim N(0, 1). \tag{3.10}$$

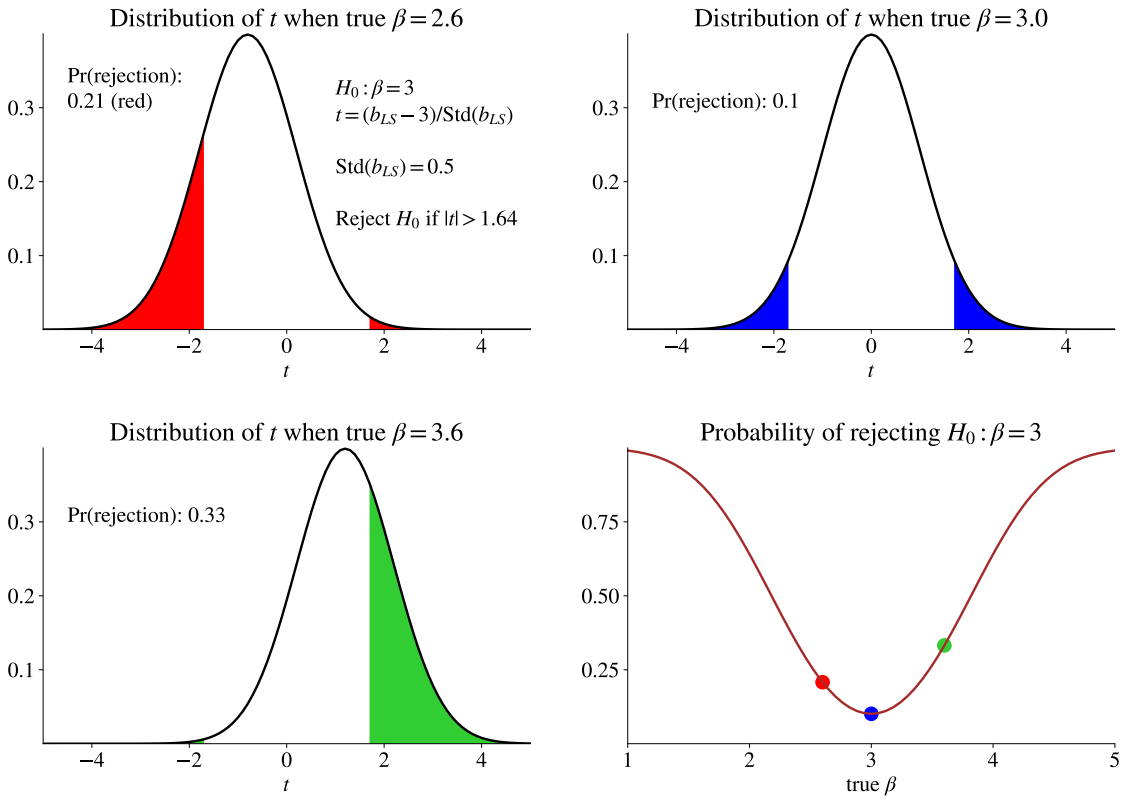


Figure 3.5: Power of t-test, assuming different true parameter values

Example 3.9 *(Testing a difference) For simplicity, suppose we have only two coefficients and want to test the difference. Then,* $R = [\ 1 \quad -1\ ]$. *Suppose (again for simplicity) that* $V(\hat{\beta}) = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$, *where* $\rho = \text{Cov}(\hat{\beta}_1, \hat{\beta}_2)$. *Clearly,* $RV(\hat{\beta})R'$ *equals* $\text{Var}(\hat{\beta}_1) + \text{Var}(\hat{\beta}_2) - 2\,\text{Cov}(\hat{\beta}_1, \hat{\beta}_2)$ *which here is* $2(1-\rho)$. *A higher covariance means that* $\hat{\beta}_1$ *and* $\hat{\beta}_2$ *tend to move in the same direction so the difference has a small uncertainty. It is then easy to test the difference. The opposite is true when testing a sum (then* $R = [\ 1 \quad 1\ ]$*).*

3.1.5 Joint Test of Several Coefficients: Chi-Square Test

A joint test of several coefficients is different from testing the coefficients one at a time. For instance, suppose your economic hypothesis is that $\beta_1 = 1$ and $\beta_3 = 0$. You could clearly test each coefficient individually (by a t-test), but that may give conflicting results. In addition, it does not use the information in the sample as effectively as possible. It might well be the case that we cannot reject any of the hypotheses (that $\beta_1 = 1$ and

$\beta_3 = 0$), but that a joint test might be able to reject it. Intuitively, a joint test is like exploiting the power of repeated sampling.

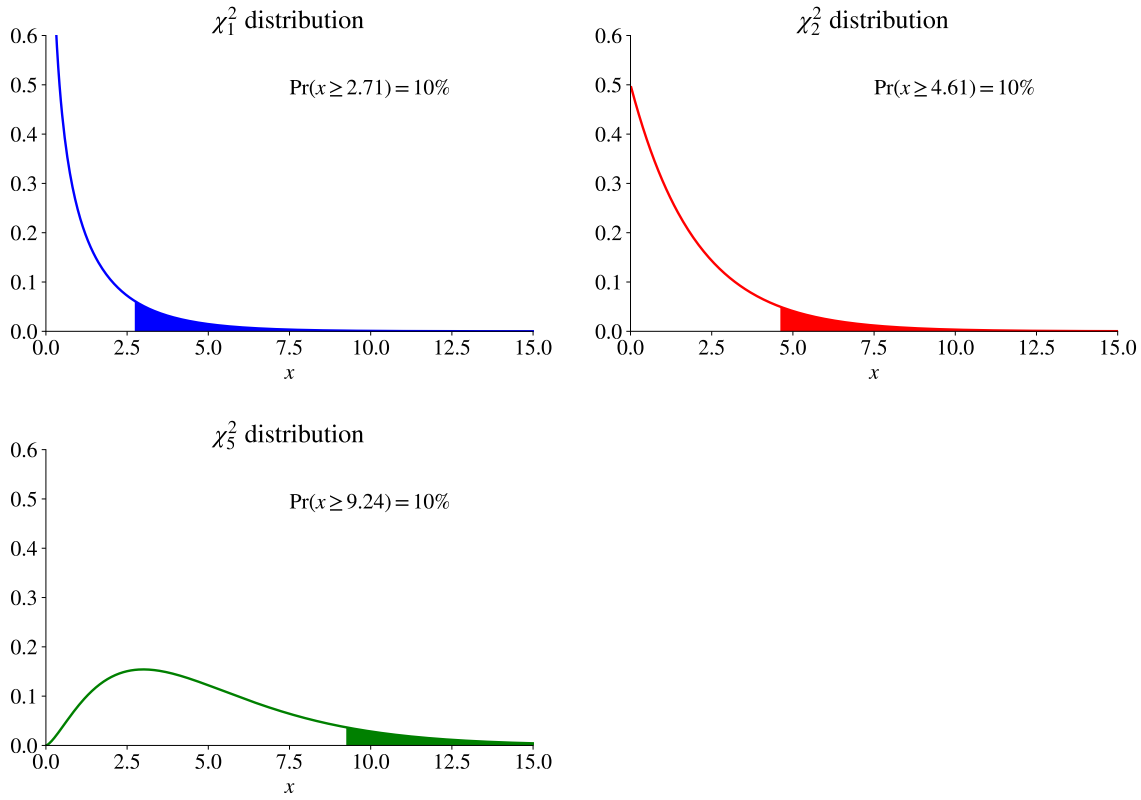


Figure 3.6: Density functions of χ^2 distributions with different degrees of freedom

A joint test makes use of the following remark.

Remark 3.10 *(Chi-square distribution) If v is a zero mean vector with n elements which are jointly normally distributed ($v \sim N(0, \Sigma)$), then*

$$v' \Sigma^{-1} v \sim \chi^2_n.$$

As a special case, suppose the vector only has one element. In this case, the quadratic form can be written $[v/\operatorname{Std}(v)]^2$, which is the square of a t-statistic.

Example 3.11 *(Quadratic form with a chi-square distribution) If the 2×1 vector v has the following normal distribution*

$$\begin{bmatrix} v_1 \\ v_2 \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix} \right),$$

then the quadratic form

$$\begin{bmatrix} v_1 \\ v_2 \end{bmatrix}' \begin{bmatrix} 1 & 0 \\ 0 & 1/2 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = v_1^2 + v_2^2/2$$

has a χ_2^2 distribution. (In a more general example, the variables could be correlated.)

For instance, suppose we have estimated a model with three coefficients and the null hypothesis is

$$H_0 : \beta_1 = 1 \text{ and } \beta_3 = 0. \tag{3.11}$$

It is convenient to write this on matrix form as

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \text{ or more generally} \tag{3.12}$$

$$R\beta = q, \tag{3.13}$$

where q has J (here 2) rows. Notice that the covariance matrix of these linear combinations is then

$$\text{Var}(R\hat{\beta}) = RV(\hat{\beta})R', \tag{3.14}$$

where $V(\hat{\beta})$ denotes the covariance matrix of the coefficients. Putting together these results we have the test static (a scalar)

$$(R\hat{\beta} - q)'[RV(\hat{\beta})R']^{-1}(R\hat{\beta} - q) \sim \chi_J^2. \tag{3.15}$$

This test statistic is compared to the critical values of a χ_J^2 distribution . (Alternatively, it can be put in the form of an F statistic, which is a small sample refinement.)

A particularly important case is the test of the joint hypothesis that all $k - 1$ slope coefficients in the regression (that is, excluding the intercept) are zero. It can be shown that the test statistic for this hypothesis is (assuming your regression also contains an intercept)

$$TR^2/(1 - R^2) \sim \chi_{k-1}^2. \tag{3.16}$$

Empirical Example 3.12 *(Test of all slopes) See Tables 3.1 and 3.2.*

Example 3.13 *(Joint test) Suppose H_0: $\beta_1 = 0$ and $\beta_3 = 0$; $(\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3) = (2, 777, 3)$*

and

$$R = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \text{ and } V(\hat{\beta}) = \begin{bmatrix} 4 & 0 & 0 \\ 0 & 33 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \text{ so}$$

$$RV(\hat{\beta})R' = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 4 & 0 & 0 \\ 0 & 33 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 4 & 0 \\ 0 & 1 \end{bmatrix}.$$

(We assume $V(\hat{\beta})$ is diagonal just because it makes it easier to invert.) Then, (3.15) is

$$\left(\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 2 \\ 777 \\ 3 \end{bmatrix} - \begin{bmatrix} 0 \\ 0 \end{bmatrix} \right)' \begin{bmatrix} 4 & 0 \\ 0 & 1 \end{bmatrix}^{-1} \left(\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 2 \\ 777 \\ 3 \end{bmatrix} - \begin{bmatrix} 0 \\ 0 \end{bmatrix} \right)$$

$$\begin{bmatrix} 2 & 3 \end{bmatrix} \begin{bmatrix} 0.25 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 2 \\ 3 \end{bmatrix} = 10,$$

which is higher than the 10% critical value of the χ^2_2 distribution (which is 4.61).

Remark 3.14 *(*An alternative form of (3.15)) Define the standardised values $z = (R\hat{\beta} - q)/\operatorname{Std}(R\hat{\beta} - q)$. Then, (3.15) can be also be written $z' \operatorname{Corr}(z)^{-1} z$.*

Remark 3.15 *(Power and size of a joint test*) Suppose $v \sim N(v_0, \Sigma)$, where v_0 might be non-zero. Then $v'\Sigma^{-1}v \sim \chi^2_n(\lambda)$ with $\lambda = v_0'\Sigma^{-1}v_0$ and where $\chi^2_n(\lambda)$ is a non-central chi-square distribution with non-centrality parameter λ. (This distribution coincides with the traditional chi-square when $\lambda = 0$.) In particular, if $R\beta - q = q_0$ (instead of zero), then the test static in (3.15) would have a $\chi^2_J(\lambda)$ distribution with $\lambda = q_0'[RV(\hat{\beta})R']^{-1}q_0$. We could then calculate the power of the test in (3.15) for different values of q_0.*

Proof. (of (3.16)) Recall that $R^2 = \operatorname{Var}(\hat{y}_t) / \operatorname{Var}(y_t) = 1 - \operatorname{Var}(\hat{u}_t) / \operatorname{Var}(y_t)$, where $\hat{y}_t = x_t'\hat{\beta}$ and $\hat{u}_t$ are the fitted value and residual respectively. We therefore get $TR^2/(1 - R^2) = T \operatorname{Var}(\hat{y}_t) / \operatorname{Var}(\hat{u}_t)$. To simplify the algebra, assume that both y_t and x_t are demeaned and that no intercept is used. (We get the same results, but after more work, if we relax this assumption.) In this case we can rewrite as $TR^2/(1 - R^2) = T\hat{\beta}' \operatorname{Var}(x_t)\hat{\beta}/\sigma^2$, where $\sigma^2 = \operatorname{Var}(\hat{u}_t)$. If the iid assumptions are correct, then the

variance-covariance matrix of $\hat{\beta}$ is $V(\hat{\beta}) = [T \text{ Var}(x_t)]^{-1}\sigma^2$, so we get

$$\begin{aligned} TR^2/(1-R^2) &= \hat{\beta}' T \text{ Var}(x_t)/\sigma^2 \hat{\beta} \\ &= \hat{\beta}' V(\hat{\beta})^{-1} \hat{\beta}. \end{aligned}$$

This has the same form as (3.15) with $R = I$ and $q = \mathbf{0}$ and J equal to the number of slope coefficients. ■

3.1.6 Confidence Bands around a Forecast and a Forecast Error*

Suppose we have estimated the linear model

$$y_t = x_t'\beta + u_t. \tag{3.17}$$

For a given (known) vector x_s, our *forecast* of y_s is

$$\text{E}(y_s|x_s) = x_s'\hat{\beta}.$$

For a given x_s, this is just a linear combination of the estimated coefficients, so the result in (3.14) holds, but with x_s' replacing R

$$\text{Var}[\text{E}(y_s|x_s)] = x_s' V(\hat{\beta}) x_s. \tag{3.18}$$

Instead, if we want the uncertainty about the *forecast error*

$$y_s - \text{E}(y_s|x_s) = x_s'(\beta - \hat{\beta}) + u_s, \tag{3.19}$$

then we have to add the uncertainty of u_s

$$\text{Var}[y_s - \text{E}(y_s|x_s)] = x_s' V(\hat{\beta}) x_s + \sigma^2. \tag{3.20}$$

(To show this last result, notice that x_s is not random and that u_s is not correlated with $\hat{\beta}$ if the latter is estimated from a sample that does not contain period s.)

3.1.7 A Joint Test of Several Coefficients: F-test*

The joint test can also be cast in *terms of the F distribution* (which may have better small sample properties).

Divide (3.15) by J and replace $V(\hat{\beta})$ by the estimated covariance matrix $\hat{V}(\hat{\beta})$. This is, for instance, $\hat{V}(\hat{\beta}) = \hat{\sigma}^2 \left(\sum_{t=1}^{T} x_t x_t'\right)^{-1}$, but where we (as in reality) have to estimate

the variance of the residuals by the sample variance of the fitted residuals, $\hat{\sigma}^2$. This gives

$$\frac{\left(R\hat{\beta}-q\right)'\left[R\hat{V}(\hat{\beta})R'\right]^{-1}\left(R\hat{\beta}-q\right)}{J} \sim F_{J,T-k}, \text{ where} \tag{3.21}$$
$$\hat{V}(\hat{\beta}) = \hat{\sigma}^2\left(\sum_{t=1}^{T} x_t x_t'\right)^{-1}.$$

The test of the joint hypothesis that all $k-1$ slope coefficients in the regression (that is, excluding the intercept) are zero can be written (assuming your regression also contains an intercept)

$$\frac{R^2/(k-1)}{(1-R^2)/(T-k)} \sim F_{k-1,T-k}. \tag{3.22}$$

Proof. (of (3.21)) Equation (3.21) can also be written

$$\frac{\left(R\hat{\beta}-q\right)'\left[R\sigma^2\left(\sum_{t=1}^{T} x_t x_t'\right)^{-1}R'\right]^{-1}\left(R\hat{\beta}-q\right)/J}{\hat{\sigma}^2/\sigma^2}.$$

The numerator is a χ^2_J variable divided by J. The denominator can be written $\Sigma_{t=1}^{T}(\hat{u}_t/\sigma)^2/(T-k)$. If the residuals are normally distributed (and independent across time), then this is a χ^2_{T-k} variable (not χ^2_T since we have estimated k parameters which influence $\hat{u}_t$) divided by $T-k$. In addition, if the numerator and denominator are independent (which requires that the residuals are independent of the regressors), then the ratio has an $F_{J,T-k}$ distribution. ■

Example 3.16 *(Joint F test) Continuing Example 3.13, and assuming that $\hat{V}(\hat{\beta}) = V(\hat{\beta})$, we have a test statistic of $10/2 = 5$. Assume $T-k = 50$, then the 10% critical value (from an $F_{2,50}$ distribution) is 2.4, so the null hypothesis is rejected at the 10% level.*

3.1.8 Testing (Nonlinear) Joint Hypotheses: The Delta Method*

Consider an estimator $\hat{\beta}_{k\times 1}$ which satisfies

$$\sqrt{T}(\hat{\beta}-\beta) \xrightarrow{d} N\left(0, V_{k\times k}\right), \tag{3.23}$$

and suppose we want the asymptotic distribution of a transformation of β

$$\gamma_{q\times 1} = f(\beta), \tag{3.24}$$

where $f(.)$ has continuous first derivatives. Under that null hypothesis (that the true value is γ)

$$\begin{aligned}\sqrt{T}(f(\hat{\beta})-\gamma) &\xrightarrow{d} N\left(0, \Lambda_{q\times q}\right), \text{ where}\\ \Lambda &= \frac{\partial f(\beta)}{\partial \beta'} V \frac{\partial f(\beta)'}{\partial \beta}, \text{ where}\end{aligned} \tag{3.25}$$

$\frac{\partial f(\beta)}{\partial \beta'}$ is the $q \times k$ matrix of partial derivatives (the Jacobian)

$$\frac{\partial f(\beta)}{\partial \beta'} = \begin{bmatrix} \frac{\partial f_1(\beta)}{\partial \beta_1} & \cdots & \frac{\partial f_1(\beta)}{\partial \beta_k} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_q(\beta)}{\partial \beta_1} & \cdots & \frac{\partial f_q(\beta)}{\partial \beta_k} \end{bmatrix}_{q\times k} \tag{3.26}$$

The derivatives can sometimes be found analytically, otherwise numerical differentiation can be used. Now, a test can be done as in the same way as in the linear case.

Example 3.17 *(Testing a Sharpe ratio) Stack the mean ($\mu = \mathrm{E}\, x_t$) and second moment ($\mu_2 = \mathrm{E}\, x_t^2$) as $\beta = [\mu, \mu_2]'$. The Sharpe ratio is calculated as a function of β*

$$\frac{\mathrm{E}(x)}{\sigma(x)} = f(\beta) = \frac{\mu}{(\mu_2 - \mu^2)^{1/2}}, \text{ so } \frac{\partial f(\beta)}{\partial \beta'} = \begin{bmatrix} \frac{\mu_2}{(\mu_2-\mu^2)^{3/2}} & \frac{-\mu}{2(\mu_2-\mu^2)^{3/2}} \end{bmatrix}.$$

If $\hat{\beta}$ is distributed as in (3.23), then (3.25) is straightforward to apply.

Example 3.18 *(Linear function) When $f(\beta) = R\beta$, then the Jacobian is $\frac{\partial f(\beta)}{\partial \beta'} = R$, so $\Lambda = RVR'$, just like in (3.14).*

Example 3.19 *(Testing a correlation of x_t and y_t) Suppose you have estimated the variances of (x_t, y_t) and also their covariance. Stack the parameters in the vector $\beta = [\sigma_{xx}, \sigma_{yy}, \sigma_{xy}]'$. The correlation and the Jacobian is then*

$$\rho(x,y) = f(\beta) = \frac{\sigma_{xy}}{\sigma_{xx}^{1/2}\sigma_{yy}^{1/2}}, \text{ so } \frac{\partial f(\beta)}{\partial \beta'} = \begin{bmatrix} -\frac{1}{2}\frac{\sigma_{xy}}{\sigma_{xx}^{3/2}\sigma_{yy}^{1/2}} & -\frac{1}{2}\frac{\sigma_{xy}}{\sigma_{xx}^{1/2}\sigma_{yy}^{3/2}} & \frac{1}{\sigma_{xx}^{1/2}\sigma_{yy}^{1/2}} \end{bmatrix}.$$

Proof. (Sketch of a proof of (3.25), requiring some asymptotics*) By the mean value theorem we have

$$f(\hat{\beta}) = f(\beta) + \frac{\partial f(\beta^*)}{\partial \beta'}(\hat{\beta} - \beta),$$

where the derivatives are evaluated at β^* which is (weakly) between $\hat{\beta}$ and β. Premultiply

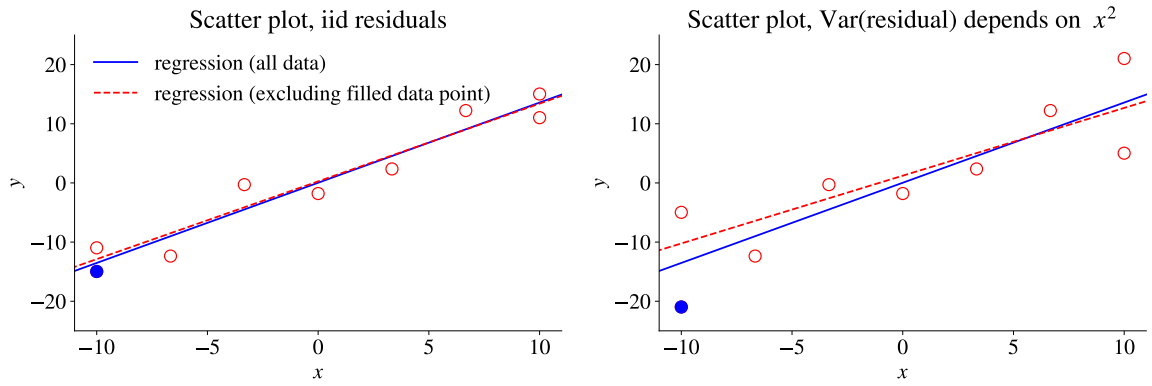


Figure 3.7: Effect of heteroskedasticity on uncertainty about regression line

by $\sqrt{T}$ and rearrange as

$$\sqrt{T}[f(\hat{\beta}) - f(\beta)] = \frac{\partial f(\beta^*)}{\partial \beta'} \sqrt{T}(\hat{\beta} - \beta_0).$$

If $\hat{\beta}$ is consistent (plim $\hat{\beta} = \beta$) and $\partial f\ (\beta^*)\ /\partial \beta'$ is continuous, then (by Slutsky's theorem) the probability limit of the derivatives is $\partial g\ (\beta)\ /\partial \beta'$ (that is, evaluated at the true β —and thus a constant). If $\sqrt{T}(\hat{\beta} - \beta_0)$ is asymptotically normally distributed, then (by the continuous mapping theorem) this carries over to the left hand side. ■

3.2 Heteroskedasticity

Suppose we have a regression model

$$y_t = x_t' b + u_t, \text{ where } \mathrm{E}\, u_t = 0 \text{ and } \mathrm{Cov}(x_{it}, u_t) = 0. \tag{3.27}$$

In the standard case we assume that u_t is iid (independently and identically distributed), which rules out variation in the volatility of the residual (heteroskedasticity).

In case the residuals actually are heteroskedastic, least squares (LS) is nevertheless a useful estimator: it is still consistent (we get the correct values as the sample becomes really large). However, the standard expression for the standard errors of the coefficients is, except in a special case, not correct. This is illustrated in Table 6.4, which shows results from simulations.

To test for heteroskedasticity, we can use *White's test of heteroskedasticity*. The test assumes that the fitted residuals are from consistent estimates (there is little point in testing

residuals from false models...) and that the regressors may be stochastic variables.

The null hypothesis is homoskedasticity, and the alternative hypothesis is the kind of heteroskedasticity which can be explained by the levels, squares, and cross products of the regressors—clearly a special form of heteroskedasticity. The reason for this specification is that if the squared residuals are uncorrelated with the squared regressors, then the usual LS covariance matrix applies—even if the residuals have some other sort of heteroskedasticity (this is the special case mentioned before).

To implement White's test, let w_t be a vector of the squares and cross products of the regressors (be sure to have a constant among the regressors). The test is then to run a regression of squared fitted residuals on w_t

$$\hat{u}_t^2 = w_t'\gamma + v_t, \tag{3.28}$$

and to test if all the slope coefficients (not the intercept) in γ are zero. This can be done be using the fact that $TR^2/(1-R^2) \sim \chi_p^2$, $p = \dim(w_t) - 1$. (Some authors prefer to use TR^2 instead, but the difference is likely to be small.)

There are several versions of White's test: (a) using only the linear terms (also called the Breusch-Pagan test); (b) using only the linear and quadratic terms (not the cross products); (c) using only a subset of the regressors.

Example 3.20 *(White's test) If the regressors include* $(1, x_{1t}, x_{2t})$ *then* w_t *in (3.28) is the vector* $(1, x_{1t}, x_{2t}, x_{1t}^2, x_{1t}x_{2t}, x_{2t}^2)$.

Remark 3.21 *(*Duplicate variables in* w_t*. If* x_t *contains a dummy variable, then its square will be the same. You can still use the same test statistic, but* p *should be the number of linearly independent variables in* w_t *minus 1.)*

Empirical Example 3.22 *(Test of heteroskedasticity) See Tables 3.1 and 3.2.*

There are two ways to handle heteroskedasticity in the residuals. First, we could use some other estimation method than LS that incorporates the structure of the heteroskedasticity. For instance, combining the regression model (3.27) with an ARCH structure of the residuals—and estimating the whole thing with maximum likelihood (MLE). As a by-product we get the correct standard errors—provided the assumed distribution (in the likelihood function) is correct. Second, we could stick to OLS, but use another expression for the variance of the coefficients: a heteroskedasticity consistent covariance matrix,

	$\gamma = 0$		$\gamma = 1$	
α :	0	1	0	1
Simulated	7.1	18.9	13.4	25.1
OLS formula	7.1	13.3	13.4	19.2
White's	7.0	18.5	13.3	24.3
Bootstrap	7.1	18.5	13.4	24.4
Bootstrap 2	7.0	18.5	13.3	24.3
FGLS	7.5	17.3	14.0	24.1

Table 3.4: Standard error of OLS slope (%) under heteroskedasticity (simulation evidence). Model: $y_t = 1+0.9x_t+\epsilon_t$, where $\epsilon_t \sim N(0, \sigma_t^2)$, with $\sigma_t^2 = (1+\gamma|z_t|+\alpha|x_t|)^2$, where z_t is iid N(0,1) and independent of x_t. Sample length: 200. Number of simulations: 25000. The bootstrap draws pairs (y_s, x_s) with replacement while bootstrap 2 is a wild bootstrap.

among which "*White's covariance matrix*" is the most common. (There is also a third possible solution: using GLS, but that is often a non-robust approach.)

To understand the construction of White's covariance matrix, recall that the variance of $\hat{\beta}$ is found from

$$\hat{\beta} = \beta + S_{xx}^{-1}(x_1u_1 + x_2u_2 + \dots x_Tu_T), \tag{3.29}$$

where $S_{xx} = \Sigma_{t=1}^{T} x_t x_t'$. If we assume that the residuals are uncorrelated with each other, then

$$\begin{aligned} \text{Var}(\hat{\beta}) &= S_{xx}^{-1}\left(x_1x_1'\sigma_1^2 + x_2x_2'\sigma_2^2 + \dots x_Tx_T'\sigma_T^2\right)S_{xx}^{-1} \\ &= S_{xx}^{-1}\underbrace{\textstyle\sum_{t=1}^{T} x_t x_t'\sigma_t^2}_{S} S_{xx}^{-1}. \end{aligned} \tag{3.30}$$

(Notice that S_{xx} and S denote very different things.) This expression cannot be simplified further since σ_t is not constant—and also related to x_t^2. The idea of White's estimator is to estimate S by

$$\hat{S} = \textstyle\sum_{t=1}^{T} x_t x_t' \hat{u}_t^2. \tag{3.31}$$

It is straightforward to show that the standard expression for the variance underestimates the true variance when there is a positive relation between x_t^2 and σ_t^2 (and vice versa). The intuition is that much of the precision (low variance of the estimates) of OLS comes from data points with extreme values of the regressors: think of a scatter plot and

notice that the slope depends a lot on fitting the data points with very low and very high values of the regressor. This nice property is destroyed if the data points with extreme values of the regressor also have lots of noise (high variance of the residual). See Figure 3.7 and Table 6.4.

White's covariance matrix should be applied when White's test (3.28) indicates problems, otherwise perhaps not. While White's covariance estimator provides safety against heteroskedasticity, it also comes at a cost: estimating the S matrix as in (3.31) risks introducing more noise.

Remark 3.23 *(Standard OLS vs White's variance*) For simplicity, consider the case of only one regressor. If x_t^2 is not related to σ_t^2, then we could write the last term in (3.30) as*

$$\begin{aligned}\textstyle\sum_{t=1}^{T} x_t^2 \sigma_t^2 &= \frac{1}{T}\textstyle\sum_{t=1}^{T}\sigma_t^2 \sum_{t=1}^{T} x_t^2 \\ &= \overline{\sigma^2}\textstyle\sum_{t=1}^{T} x_t^2\end{aligned}$$

where $\overline{\sigma^2}$ is the average variance, typically estimated as $\sum_{t=1}^{T} u_t^2/T$. That is, it is the same as for standard OLS. In addition, notice that

$$\textstyle\sum_{t=1}^{T} x_t^2 \sigma_t^2 > \frac{1}{T}\sum_{t=1}^{T}\sigma_t^2 \sum_{t=1}^{T} x_t^2$$

if x_t^2 is positively related to σ_t^2 (and vice versa). For instance, with $(x_1^2, x_2^2) = (10, 1)$ and $(\sigma_1^2, \sigma_2^2) = (5, 2)$, $\sum_{t=1}^{T} x_t^2 \sigma_t^2 = 10 \times 5 + 1 \times 2 = 52$ while $\frac{1}{T}\sum_{t=1}^{T}\sigma_t^2 \sum_{t=1}^{T} x_t^2 = \frac{1}{2}(5+2)(10+1) = 38.5$.

Remark 3.24 *(GLS*) With heteroskedasticity and/or autocorrelation, OLS is still consistent and we can adjust the covariance matrix of the coefficients. However, OLS is less efficient (higher uncertainty of the coefficients) than GLS (Generalized Least Squares) is. The basic idea of GLS is transform regression equation so*

$$y_t^* = x_t^{*\prime}\beta + \varepsilon_t^*,$$

have iid residuals. Estimating β with LS on this transformation is efficient (called GLS) and the traditional expressions of the covariance matrix of the coefficients can be used. For instance, with heteroskedasticity*, the transformation is*

$$\frac{y_t}{\sigma_t} = \frac{x_t'}{\sigma_t}\beta + \frac{\varepsilon_t}{\sigma_t}.$$

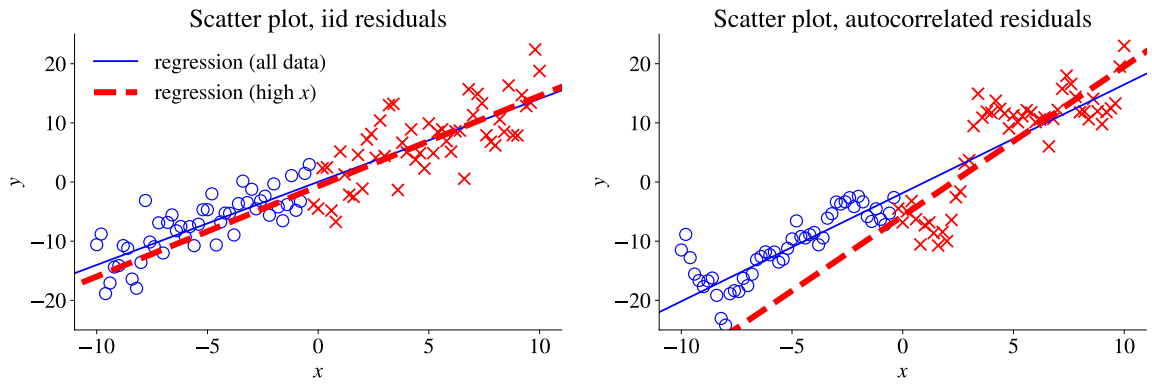


Figure 3.8: Effect of autocorrelation on uncertainty about regression line

(Yes, also the constant is divided by σ_t.) Notice that ε_t/σ_t has a constant variance (equal to one). In practice we don't know σ_t, so we first estimate it (the method is then called "feasible" GLS, FGLS.) FGLS may improve the efficiency, but can be unstable if we model the heteroskedasticity wrongly. A commonly applied approach is the following (a) *let $\hat{\varepsilon}_t$ be the residual from the OLS regression;* (b) *regress* $\ln(\hat{\varepsilon}_t^2)$ *on the regressors and all the squares (and cross-products of them);* (c) *let z_t be the fitted values from the regression in (b) and set* $\sigma_t = \sqrt{\exp(z_t)}$.

3.3 Autocorrelation

Autocorrelation of the residuals ($\text{Cov}(u_t u_{t-s}) \neq 0$) is also a violation of the iid assumptions underlying the standard expressions for the variance of $\hat{\beta}$. In this case, LS is typically still consistent (exceptions: when the lagged dependent variable is a regressor), but the variances are again wrong.

The typical effect of positively autocorrelated residuals is to increase the uncertainty about the OLS estimates—above what is indicated by the traditional standard errors based on the iid assumption. This is perhaps easiest to understand in the case of estimating the mean of a data series, that is, when regressing a data series on a constant only. If the residual is positively autocorrelated (have long swings), then the sample mean can deviate from the true mean for an extended period of time—perhaps for most of a sample: the estimate is imprecise. See Figure 3.8 for an illustration.

There are several straightforward tests of autocorrelation—all based on using the fitted residuals. The tests all assume that the fitted residuals are from consistent estimates. The

null hypothesis is no autocorrelation. First, estimate the autocorrelations of the fitted residuals as

$$\rho_s = \text{Corr}(\hat{u}_t, \hat{u}_{t-s}), s = 1, ..., L. \tag{3.32}$$

Second, test autocorrelation s by using the fact that $\sqrt{T}\hat{\rho}_s$ has a standard normal distribution (in large samples)

$$\sqrt{T}\hat{\rho}_s \sim N(0, 1). \tag{3.33}$$

To extend (3.33) to higher-order autocorrelation, use the Box-Pierce test

$$Q_L = T \sum_{s=1}^{L} \hat{\rho}_s^2 \rightarrow^d \chi_L^2. \tag{3.34}$$

Empirical Example 3.25 *(Test of autocorrelation) See Tables 3.1 and 3.2.*

An alternative for testing the first autocorrelation coefficient is the Durbin-Watson. The test statistic is (approximately)

$$DW \approx 2 - 2\hat{\rho}_1, \tag{3.35}$$

and the null hypothesis is rejected in favour of positive autocorrelation if $DW < 1.5$ or so (depending on sample size and the number of regressors).

These tests can also be applies to each of the elements in $x_t u_t$ (instead of just u_t), since it is actually autocorrelations of these cross terms that matter most (see the discussion below).

ρ :	0.0	0.75
Simulated	5.8	23.0
OLS formula	5.8	8.7
Newey-West	5.7	16.3
VARHAC	5.7	22.4
Bootstrapped	5.5	19.6
FGLS	5.9	23.1

Table 3.5: Standard error of OLS intercept (%) under autocorrelation (simulation evidence). Model: $y_t = 1 + 0.9x_t + \epsilon_t$, where $\epsilon_t = \rho\epsilon_{t-1} + \xi_t, \xi_t$ is iid N(). NW uses 5 lags. VARHAC uses 5 lags and a VAR(1). The bootstrap uses blocks of size 20. Sample length: 300. Number of simulations: 25000.

If there is autocorrelation, then we can choose to estimate a fully specified model (including how the autocorrelation is generated) by MLE or we can stick to OLS but apply an

	$\kappa = 0.0$		$\kappa = 0.75$	
ρ :	0.0	0.75	0.0	0.75
Simulated	5.8	8.7	3.9	10.9
OLS formula	5.8	8.6	3.9	5.8
Newey-West	5.7	8.4	3.8	8.9
VARHAC	5.7	8.5	3.8	10.5
Bootstrapped	5.8	8.5	3.8	10.1
FGLS	5.8	4.7	3.9	5.9

Table 3.6: Standard error of OLS slope (%) under autocorrelation (simulation evidence). Model: $y_t = 1 + 0.9x_t + \epsilon_t$, where $\epsilon_t = \rho\epsilon_{t-1} + \xi_t$, ξ_t is iid N(). $x_t = \kappa x_{t-1} + \eta_t$, η_t is iid N(). NW uses 5 lags. VARHAC uses 5 lags and a VAR(1). The bootstrap uses blocks of size 20. Sample length: 300. Number of simulations: 25000.

autocorrelation consistent covariance matrix—for instance, the "*Newey-West covariance matrix*."

To understand the Newey-West covariance matrix, notice that the the variance of $\text{Var}(\hat{\beta})$ is

$$\text{Var}(\hat{\beta}) = S_{xx}^{-1}\underbrace{\text{Var}\left(x_1u_1 + x_2u_2 + \ldots x_Tu_T\right)}_{S}S_{xx}^{-1}, \tag{3.36}$$

where the regressors may be stochastic variables.

However, there might be correlation across time periods, so the S term in the middle needs to account for terms like $\text{Cov}(x_tu_t, x_{t-s}u_{t-s})$. For instance, for $T = 3$ the S term is

$$\begin{aligned} S = &\text{Var}(x_1u_1) + \text{Var}(x_2u_2) + \text{Var}(x_3u_3) + \\ &2\,\text{Cov}(x_2u_2, x_1u_1) + 2\,\text{Cov}(x_3u_3, x_2u_2) + 2\,\text{Cov}(x_3u_3, x_1u_1). \end{aligned} \tag{3.37}$$

When data is uncorrelated across time (observations), then all the covariance terms are zero. With autocorrelation, they may not be. For a general T, the S term is

$$S = \sum_{t=1}^{T}\text{Var}\left(x_tu_t\right) + 2\sum_{s=1}^{m}\sum_{t=s+1}^{T}\text{Cov}\left(x_tu_t, x_{t-s}u_{t-s}\right), \tag{3.38}$$

where m denotes the number of covariance terms that might be non-zero (at most, $m = T - 1$).

It is clear from (3.38) that what really counts is not so much the autocorrelation in u_t per se, but the autocorrelation of x_tu_t. If this is positive, then the standard expression un-

derestimates the true variance of the estimated coefficients (and vice versa). For instance, the autocorrelation of $x_t u_t$ is likely to be positive when both the residual and the regressor are positively autocorrelated. (Notice that a constant, $x_t = 1$ is extremely positively autocorrelated, so autocorrelation of the residual along is enough to cause problems with the intercept.) In contrast, when the regressor has no autocorrelation, then the product does not either. This is illustrated in Tables 3.5–18.3.

The idea of the Newey-West estimator is to estimate S by (with several regressors)

$$\hat{S} = \Lambda_0 + \sum_{s=1}^{m}(1 - \frac{s}{m+1})(\Lambda_s + \Lambda_s'), \text{ where} \tag{3.39}$$

$$\Lambda_s = \sum_{t=s+1}^{T} x_t x_{t-s}' \hat{u}_t \hat{u}_{t-s}. \tag{3.40}$$

The weights $1-s/(m+1)$ are close to 1 for small lags (s values), but decline linearly (tent shaped weights) to zero. The point of using such weights is to make sure that the $\hat{S}$ matrix remains invertible (to show this is somewhat involved). This suggests that m should be somewhat larger than last lag with significant autocorrelation. However, a common rule of thumb is to use round $m = \text{floor}(0.75T^{1/3})$, where floor() means rounding down to nearest integer (and alternative rule is $m = \text{floor}(4(T/100)^{2/9})$).

For instance, with only one lag ($m = 1$) the calculation is (with several regressors)

$$\hat{S} = \sum_{t=1}^{T} x_t x_t' \hat{u}_t^2 + \sum_{t=2}^{T}(1 - \frac{1}{2})\left(x_t x_{t-1}' + x_{t-1} x_t'\right) \hat{u}_t \hat{u}_{t-1}, \tag{3.41}$$

and by excluding all lags (setting $m = 0$), the Newey-West estimator coincides with White's estimator

$$\hat{S} = \sum_{t=1}^{T} x_t x_t' \hat{u}_t^2. \tag{3.42}$$

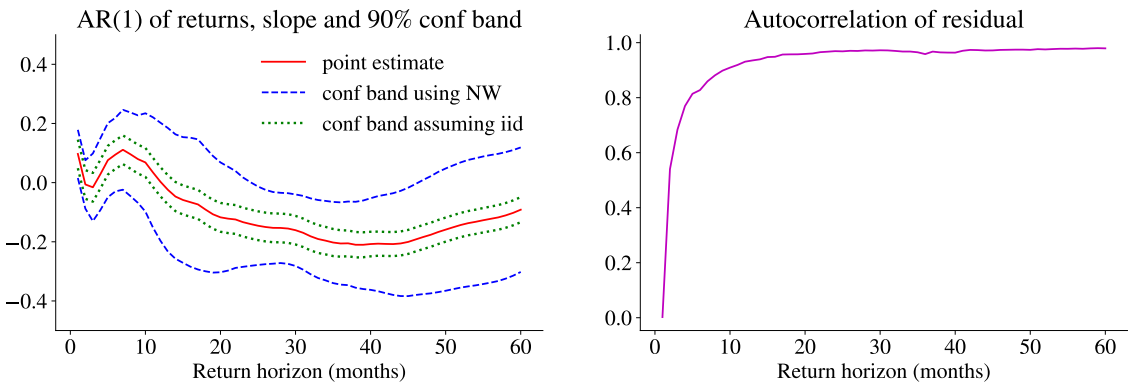
Hence, Newey-West estimator handles also heteroskedasticity.

Remark 3.26 *(VARHAC*) The VARHAC estimator of the covariance matrix (see Andrews and Monahan (1992)) is to first fit a VAR(p) to $z_t = x_t \hat{u}_t$*

$$z_t = A_0 + \sum_{i=1}^{p} A_i z_{t-i} + \varepsilon_t$$

and then calculate $D = I - \sum_{i=1}^{p} A_i$. Then, $\hat{S} = D^{-1}\hat{S}^{\varepsilon}D^{-1}$, where $\hat{S}^{\varepsilon}$ is Newey-West estimate applied to $\hat{\varepsilon}_t$ only (use $\hat{\varepsilon}_t$ instead of $x_t \hat{u}_t$ in (3.39)).

Empirical Example 3.27 *(Autocorrelation from overlapping return periods) Figures 18.7–3.10 are empirical examples of the importance of using the Newey-West method rather*



Slope with two different 90% conf bands (assuming iid or using NW)

Monthly US stock excess returns 1926:01-2021:12, overlapping data

Figure 3.9: Slope coefficient, LS vs Newey-West standard errors

than relying of the iid assumptions. In both cases, the residuals have strong positive autocorrelation.

The Newey-West approach should be applied when the tests of the residuals indicate autocorrelation, otherwise probably not. The method involves estimating lots of parameters in the S matrix—and this can in itself introduce noise and uncertainty.

Remark 3.28 *(GLS*) With first-order* autocorrelation, *($\varepsilon_t = \rho\varepsilon_{t-1} + v_t$, where v_t is iid), we can implement FGLS by doing a "quasi-difference" of the regression equation*

$$y_t - \rho y_{t-1} = (x_t - \rho x_{t-1})' \beta + (\varepsilon_t - \rho\varepsilon_{t-1}) .$$

This new residual, $\varepsilon_t - \rho\varepsilon_{t-1}$, is iid. In practice we don't know ρ, so we first estimate it.

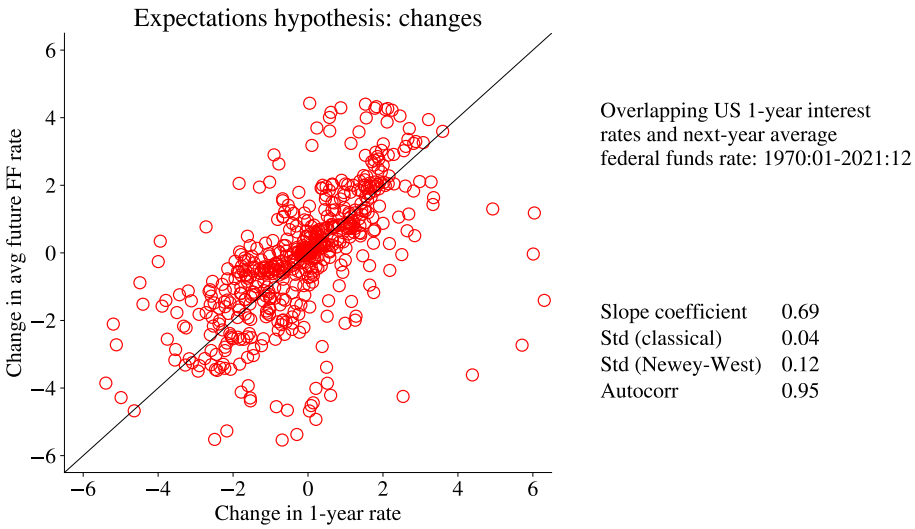


Figure 3.10: US 12-month interest and average federal funds rate (next 12 months)

Chapter 4

The Variance of a (Time Series) Sample Average

Reference: Hayashi (2000) 6.5
Additional references: Hamilton (1994) 14; Verbeek (2004) 4.10; Harris and Matyas (1999); and Pindyck and Rubinfeld (1998) Appendix 10.1; Cochrane (2001) 11.7

4.1 The Variance of a Sample Average

Many estimators (including OLS, MLE and GMM) are based on some sort of sample average. Unless we are sure that the series in the average is iid, we need an estimator of the variance (of the sample average) that takes serial correlation into account. For a time series average, the Newey and West (1987) estimator is probably the most popular.

To illustrate the idea, consider a time series sample mean, $\bar{x}$, of a $K \times 1$ vector x_t

$$\bar{x} = \frac{1}{T}\sum_{t=1}^{T} x_t. \tag{4.1}$$

If x_t is iid, then

$$\text{Cov}(\sqrt{T}\bar{x}) = \text{Cov}(x_t), \tag{4.2}$$

which is a $K \times K$ matrix. This clearly is the same as saying that $\text{Cov}(\bar{x}) = \text{Cov}(x_t)/T$. Instead, if x_t is autocorrelated, then

$$\text{Cov}(\sqrt{T}\bar{x}) = \sum_{s=-(T-1)}^{T-1} \left(1 - \frac{|s|}{T}\right) \Gamma(s), \text{ where } \Gamma(s) = \text{Cov}(x_t, x_{t-s}), \tag{4.3}$$

where $\Gamma(s)$ is a $K \times K$ matrix, where cell (i, j) is the covariance between element i of x_t and element j of x_{t-s}. Here we assume that $\text{Cov}(x_t, x_{t-s})$ only depends on s (not on t), but this can be relaxed.

Example 4.1 *($\Gamma(s)$ for a vector with two variables) If $x_t = [x_{1t}, x_{2t}]'$ where x_{1t} is one variable and x_{2t} is another, then*

$$\Gamma(s) = \begin{bmatrix} \text{Cov}(x_{1,t}, x_{1,t-s}) & \text{Cov}(x_{1,t}, x_{2,t-s}) \\ \text{Cov}(x_{2,t}, x_{1,t-s}) & \text{Cov}(x_{2,t}, x_{2,t-s}) \end{bmatrix}.$$

Proof. (of (4.3)) Notice that for $T = 3$, we have

$$\begin{aligned} \text{Var}(x_1 + x_2 + x_3) = &\underbrace{\text{Cov}(x_1, x_3)}_{\Gamma(-2)} + \underbrace{\text{Cov}(x_1, x_2) + \text{Cov}(x_2, x_3)}_{2\Gamma(-1)} + \\ &\underbrace{\text{Var}(x_1) + \text{Var}(x_2) + \text{Var}(x_3)}_{3\Gamma(0)} + \underbrace{\text{Cov}(x_2, x_1) + \text{Cov}(x_3, x_2)}_{2\Gamma(1)} + \underbrace{\text{Cov}(x_3, x_1)}_{\Gamma(2)}. \end{aligned}$$

The general pattern is

$$\text{Var}\left(\sum_{t=1}^{T} x_t\right) = \sum_{s=-(T-1)}^{T-1} (T - |s|)\,\Gamma(s).$$

Divide both sides by T to get (4.3). ■

Remark 4.2 *(Cross-sectional averages) The insight that correlations matter for an average applies also to a cross-sectional average. The only difference is that it is harder to motivate why the variances should be the same across observations. As an example, consider the cross-sectional average return (in period t) across n assets, $\bar{R}_t = \Sigma_{i=1}^{n} R_{i,t}/n$. It is clear that $\text{Var}(\bar{R}_t) = \mathbf{1}'\Sigma\mathbf{1}/n^2$, where $\mathbf{1}$ is an $n \times 1$ vector of ones and Σ is the covariance matrix of the n assets. This is just the sum of all elements, divided by n^2, which is very similar to (4.3), although we are here studying a cross-section, not a time series.*

Taking the limit of (4.3) as $T \to \infty$, we get

$$\lim_{T\to\infty} \text{Cov}(\sqrt{T}\bar{x}) = \sum_{s=-\infty}^{\infty} \Gamma(s). \tag{4.4}$$

Example 4.3 *(Variance of sample mean of AR(1).) Let $x_t = \rho x_{t-1} + u_t$, where $\text{Var}(u_t) = \sigma^2$. Let $\Gamma(s)$ denote the sth autocovariance and notice that $\Gamma(s) = \rho^{|s|}\sigma^2/(1-\rho^2)$. The asymptotic (as $T \to \infty$ so $|s|/T \to 0$ in (4.3)) variance can be written*

$$\text{Var}(\sqrt{T}\bar{x}) = \sum_{s=-\infty}^{\infty} \Gamma(s) = \frac{\sigma^2}{1-\rho^2} \sum_{s=-\infty}^{\infty} \rho^{|s|} = \frac{\sigma^2}{1-\rho^2}\frac{1+\rho}{1-\rho},$$

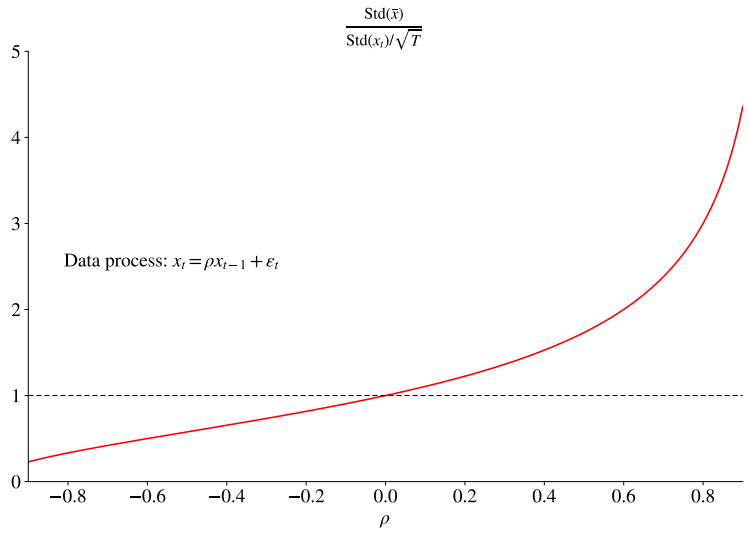


Figure 4.1: Variance of $\sqrt{T}\times$ sample average of an AR(1) series

which is increasing in ρ (provided $|\rho| < 1$, as required for stationarity). The variance of $\sqrt{T}\bar{x}$ is much larger for ρ close to one than for ρ close to zero: the high autocorrelation create long swings, so the mean cannot be estimated with good precision in a small sample. If we disregard all autocovariances, then we would conclude that the variance of $\sqrt{T}\bar{x}$ is $\sigma^2/\left(1-\rho^2\right)$, that is, the variance of x_t. This is much smaller (larger) than the true value when $\rho > 0$ ($\rho < 0$). For instance, with $\rho = 0.9$, it is 19 times too small. See Figure 4.1 for an illustration. Notice that $\text{Var}(\sqrt{T}\bar{x})/\text{Var}(x_t) = \text{Var}(\bar{x})/[\text{Var}(x_t)/T]$*, so the ratio also shows the relation between the true variance of $\bar{x}$ and the classical estimator of it (based of the iid assumption).*

4.2 The Newey-West Estimator

The Newey-West estimator of the variance-covariance matrix of $\sqrt{T}\bar{x}$ is

$$\widehat{\text{Cov}}(\sqrt{T}\bar{x}) = \sum_{s=-n}^{n}\left(1-\frac{|s|}{n+1}\right)\widehat{\text{Cov}}\left(x_t, x_{t-s}\right), \tag{4.5}$$

where n is a finite "bandwidth" parameter. The "weights," $1-|s|/(n+1)$, are clearly tent-shaped: 1 at the zero lag—and lower as the lags become longer. Figure 4.2 illustrates the weights (the term in parentheses in (4.5)) for different choices of the bandwidth (n). This is similar to (4.3), but the weights decrease quicker (assuming $n < T-1$). This

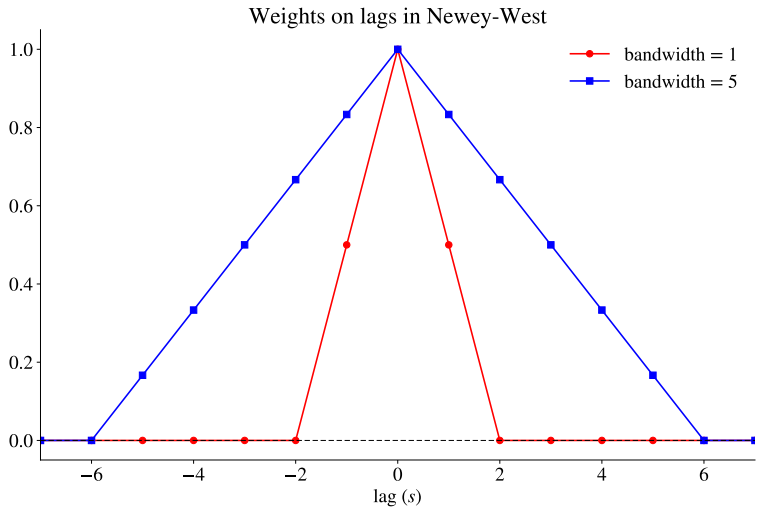


Figure 4.2: Weights in Newey-West

suggests that n should be somewhat larger than last lag with significant autocorrelation. Alternatively, a common rule of thumb is $n = \text{floor}(0.75T^{1/3})$, where floor() means rounding down to nearest integer (sometimes $n = \text{floor}(4(T/100)^{2/9})$ is used instead).

Example 4.4 *(Newey-West estimator) With $n = 1$ in (4.5) the Newey-West estimator becomes*

$$\widehat{\text{Cov}}(\sqrt{T}\bar{x}) = \frac{1}{2}\widehat{\text{Cov}}\left(x_t, x_{t+1}\right) + \widehat{\text{Cov}}\left(x_t, x_t\right) + \frac{1}{2}\widehat{\text{Cov}}\left(x_t, x_{t-1}\right).$$

Remark 4.5 *(VARHAC*) The VARHAC estimator of the covariance matrix (see Andrews and Monahan (1992)) is as follows. First, fit a VAR(p) to x_t*

$$x_t = A_0 + \sum_{i=1}^{p} A_i x_{t-i} + \varepsilon_t$$

and calculate $D = I - \sum_{i=1}^{p} A_i$. Then, use $\text{Cov}(\sqrt{T}\bar{x}) = D^{-1}S^*D^{-1}$, where S^* is Newey-West estimate of $\text{Cov}(\sqrt{T}\bar{\varepsilon})$. As an example, let x_t be a scalar that follows an AR(1) process, $x_t = \rho x_{t-1} + \varepsilon_t$. If ε_t is iid, then $\text{Cov}(\sqrt{T}\bar{\varepsilon}) = \sigma^2$ where σ^2 is the variance of ε_t. $D = 1 - \rho$, so $\text{Cov}(\sqrt{T}\bar{x}) = \sigma^2/(1-\rho)^2$ which is the same as the variance in Example 4.3 (since $(1-\rho^2)/(1+\rho) = 1-\rho$).

Chapter 5

Asymptotic Results on OLS*

Reference: Verbeek (2012) 2 and 5

5.1 Motivation of Asymptotics

There are several problems when the standard assumptions about linear regressions are wrong. First, the result that $\mathrm{E}\,\hat{\beta} = \beta$ (unbiased) relies on the assumption that the regressors are fixed or alternatively that $\{u_1, ..., u_T\}$ and $\{x_1, ..., x_T\}$ are independent. Otherwise, it is not true (in a finite sample)—see Figure 5.1. Second, the result that $\hat{\beta}$ is normally distributed relies on the assumption that residuals are normally distributed. Otherwise it is not true (in a finite sample). See Figure 5.2.

What *is* true when the standard assumptions are not satisfied? How should we test hypotheses? Two ways to find answers: *(a)* do computer (Monte Carlo or bootstrap) simulations; *(b)* find results for $T \to \infty$ ("asymptotic properties") and use them as approximations for large samples.

The results from asymptotic theory are more general (and perhaps prettier) than simulations—and can be used as approximations if the sample is large. The basic reason for why this works this is that most estimators are sample averages and sample averages often have nice properties as $T \to \infty$. In particular, we can make use of the law of large numbers (LLN) and the central limit theorem (CLT). See Figure 5.3.

However, the asymptotic results are unlikely to be good approximations in small samples. In those cases we need simulations.

5.2 Asymptotics: Consistency

Reference: Greene (2018) 4.4; Hamilton (1994) 8.2; Davidson (2000) 3

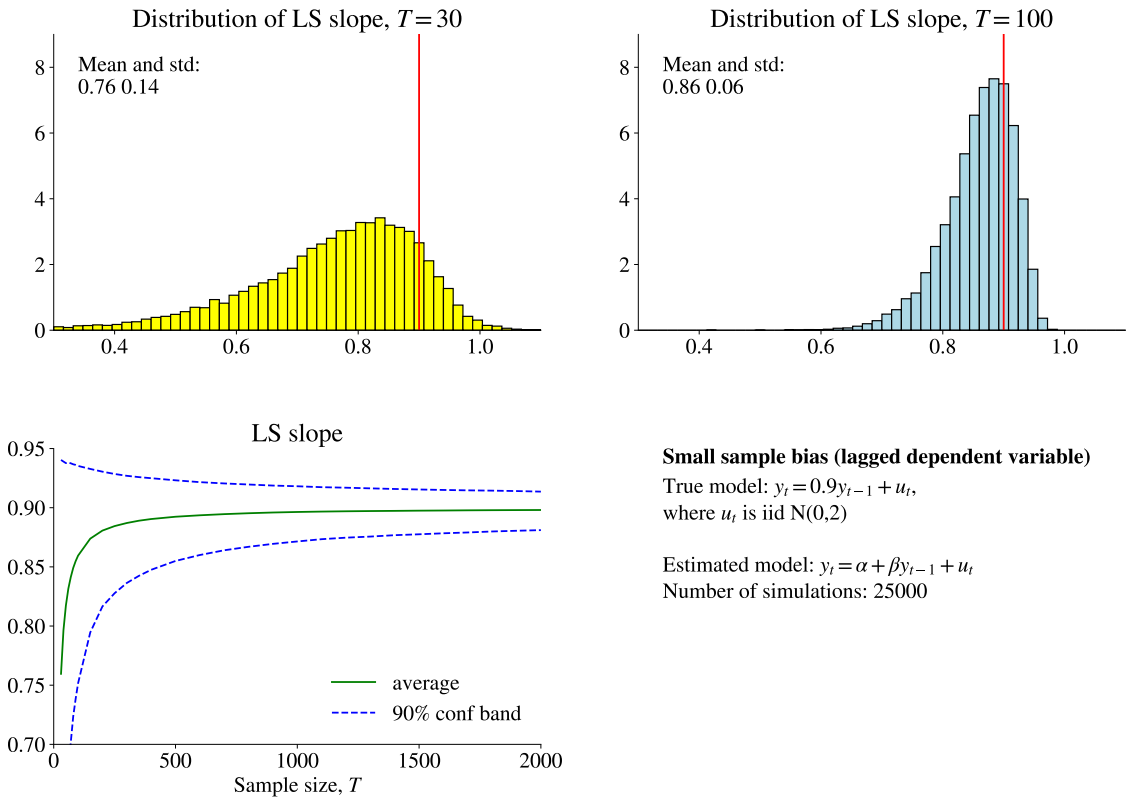


Figure 5.1: Distribution of LS estimator of autoregressive parameter

Issue: will our estimator come closer to the truth as the sample size increases? If not, use another estimator (method).

5.2.1 Probability Limits

We need some basic facts about statistics (probability limits) for the discussion of consistency.

Remark 5.1 *(Convergence in probability)* $\hat{\beta}$ *(which depends on the sample size* T*) converges in probability to* b *if for every* $\varepsilon > 0$

$$\lim_{T\to\infty} \Pr(|\hat{\beta} - b| < \varepsilon) = 1.$$

Notation: $\operatorname{plim} \hat{\beta} = b$ *or* $\hat{\beta} \to^p b$ *where* plim *stands for the probability limit.*

Remark 5.2 *(Probability limits of a product and of a function) If* $\operatorname{plim} \hat{\alpha} = a$ *and* $\operatorname{plim} \hat{\beta} = b$*, then* $\operatorname{plim} \hat{\alpha}\hat{\beta} = ab$*. (In contrast, this does not not hold for expectations:* $\mathrm{E}\,\hat{\alpha}\hat{\beta} \neq$

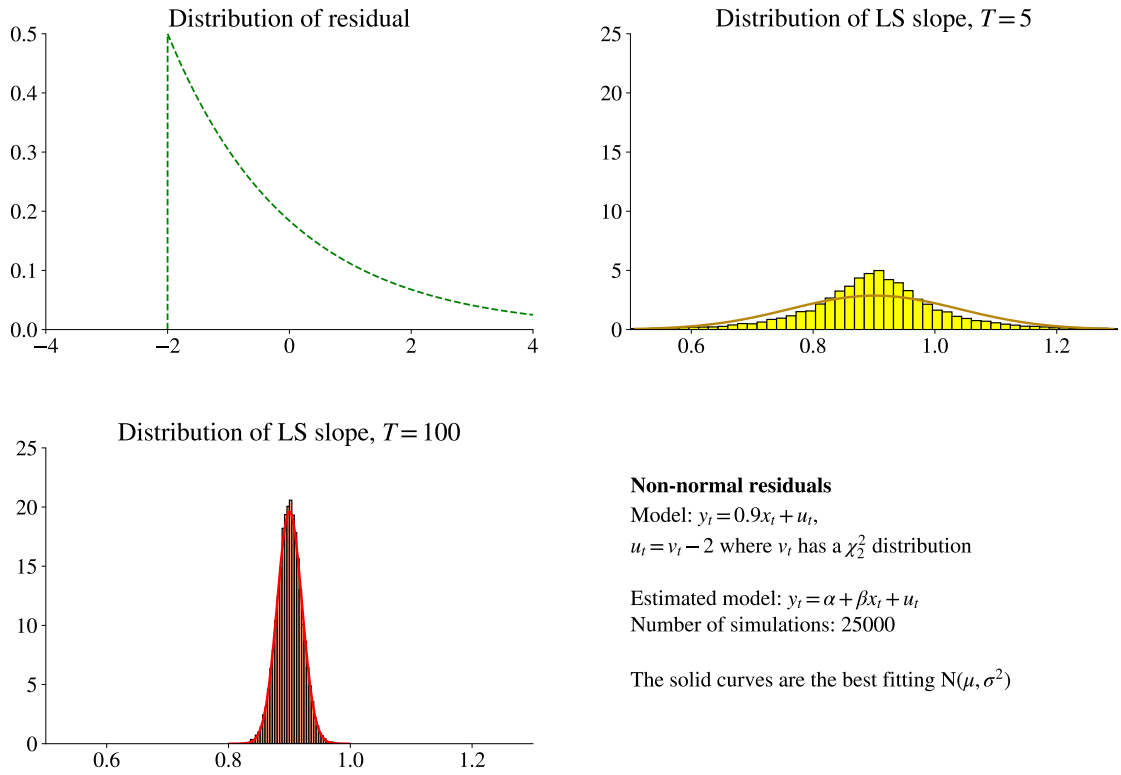


Figure 5.2: Results from a Monte Carlo experiment with thick-tailed errors.

$\mathrm{E}\,\hat{\alpha}\,\mathrm{E}\,\hat{\beta}$ *unless* $\hat{\alpha}$ *and* $\hat{\beta}$ *are uncorrelated.) More generally, Slutsky's theorem says that if* $g()$ *is a continuous function, then* $\operatorname{plim} g(\hat{\alpha}) = g(\operatorname{plim} \hat{\alpha})$.

Remark 5.3 *(Law of large numbers, simple version) A LLN says that the sample average converges to the population mean as the sample size increases (to infinity). Clearly, this means that the sample average is a consistent estimator of the population mean. Notation:* $\operatorname{plim}(\bar{x}) = \mathrm{E}(x)$.

5.2.2 Consistency of OLS

Remark 5.4 *(Consistency) Consistency means that the estimate* $\hat{\beta}$ *converges in probability to the true value as the sample size increases (to infinity).*

The OLS estimate of a slope coefficient is (after dividing and multiplying by T)

$$\hat{\beta} = \beta + \underbrace{\left(\frac{1}{T}\sum_{t=1}^{T} x_t x_t'\right)^{-1}}_{\to \Sigma_{xx}^{-1}} \underbrace{\frac{1}{T}\sum_{t=1}^{T} x_t u_t}_{\to \mathrm{E}(x,u)} \tag{5.1}$$

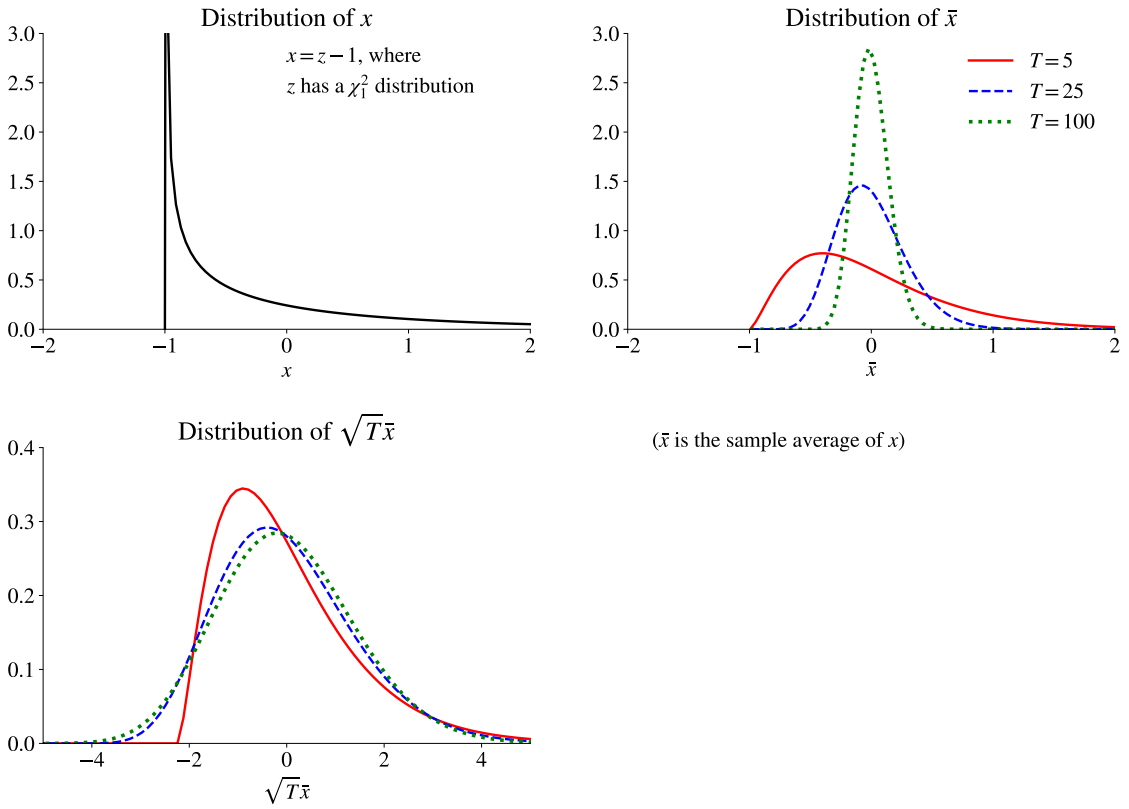


Figure 5.3: Distribution of sample averages

where u_t are the residuals we could calculate if we knew the true slope coefficient (denoted β), that is, the true residuals. The symbols below the equation indicate what the different terms converge to (according to a LLN) as the sample size increases. In particular, the inverse is a continuous function so the first term converges to the inverse of the second moment matrix of x_t ($\mathrm{E}\, x_t x_t'$) which is denoted Σ_{xx}. (This clearly assumes that x_t is such that the expectation is well defined.) Also, the two terms form a product so we can apply the rule that the probability limit is the product of the two (individual) probability limits.

In short, the probability limit is

$$\operatorname{plim} \hat{\beta} = \beta + \Sigma_{xx}^{-1} \mathrm{E}(x_t u_t), \tag{5.2}$$

where Σ_{xx}^{-1} is (asymptotically) a matrix of constants: there is nothing random about it. Clearly, for the estimate $\hat{\beta}$ to converge to the true values (β), $\mathrm{E}(x_t u_t) = 0$ is needed. If $\mathrm{E}\, u_t = 0$ (which is a basic assumption in most regression analysis) $\mathrm{E}(x_t u_t) = \mathrm{Cov}(x_t, u_t)$,

so consistency of $\hat{\beta}$ requires the regressors and the (true) residuals to be be uncorrelated.

Some observations:

1. We can not (easily) test this, since OLS *creates* $\hat{\beta}$ and the fitted residuals $\hat{u}_t$ such that $\sum_{t=1}^{T} x_t \hat{u}_t / T = 0$.

2. The standard regression assumption that u_t and x_t are independent implies that $\mathrm{E}(x_t u_t) = 0$. This means that the standard regression assumptions take it for granted that OLS is consistent.

3. OLS can be biased, but still be consistent. This means that OLS is systematically wrong in any small sample, but the problem vanishes in large samples. See Figure 5.1. In these figures, $\mathrm{Cov}(u_{t-1}, x_t) \neq 0$ so OLS is biased since x_t is not independent of *all* residuals, but $\mathrm{Cov}(u_t, x_t) = 0$ so it is consistent since x_t is not correlated with the *contemporaneous* residual.

4. There are cases when $\mathrm{E}(x_t u_t) = 0$ doesn't make sense. Then OLS is inconsistent. More on this later.

5. See Figure 5.1 for an example of where OLS is consistent, and Figure 5.4 when it is not.

What have we learned? Well,...under what conditions ($\mathrm{E}(x_t u_t) = 0$) OLS comes closer to the correct value as T increases.

5.3 When OLS Is Inconsistent

Q. When do we have $\mathrm{E}(x_t, u_t) \neq 0$?

A. Need to think hard...

But the usual suspects are *(i)* omitted variables; *(ii)* autocorrelated errors combined with lagged dependent variable; *(iii)* measurement errors in regressors; and *(iv)* endogenous regressors.

5.3.1 Omitted Variables

Reference: Greene (2018) 4.3

Consider the regression

$$y_t = x_t' \beta + h_t' \gamma + \varepsilon_t, \tag{5.3}$$

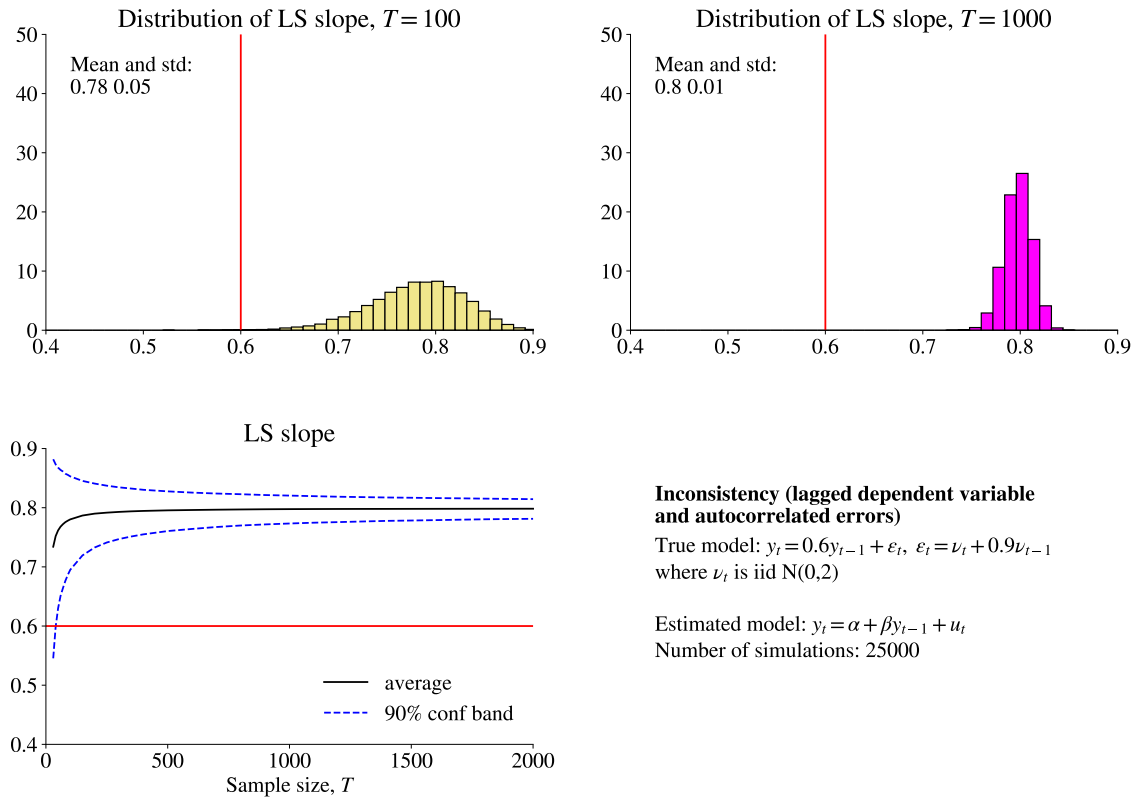


Figure 5.4: Results from a Monte Carlo experiment of LS estimation of the AR coefficient when data is from an ARMA process.

where $\mathrm{E}(x_t\varepsilon_{tt}) = 0$.

Suppose we omit (exclude) the h_t variables and instead estimate

$$y_t = x_t'\beta + u_t. \tag{5.4}$$

This means that the residual from in the regression (5.4) is $u_t = h_t'\gamma + \varepsilon_t$, that is, it incorporates the effect of both the omitted variables and the "true" residual.

Recall that the OLS estimates are

$$\hat{\beta} = \beta + S_{xx}^{-1}\Sigma_{t=1}^{T} x_t u_t, \tag{5.5}$$

where $S_{xx} = \Sigma_{t=1}^{T} x_t x_t'$. Since $u_t = h_t'\gamma + \varepsilon_t$, we can write this as

$$\hat{\beta} = \beta + S_{xx}^{-1}\textstyle\sum_{t=1}^{T} x_t h_t'\gamma + S_{xx}^{-1}\sum_{t=1}^{T} x_t\varepsilon_t. \tag{5.6}$$

The last term should vanish as the sample size increases (the residual in (5.3) should not

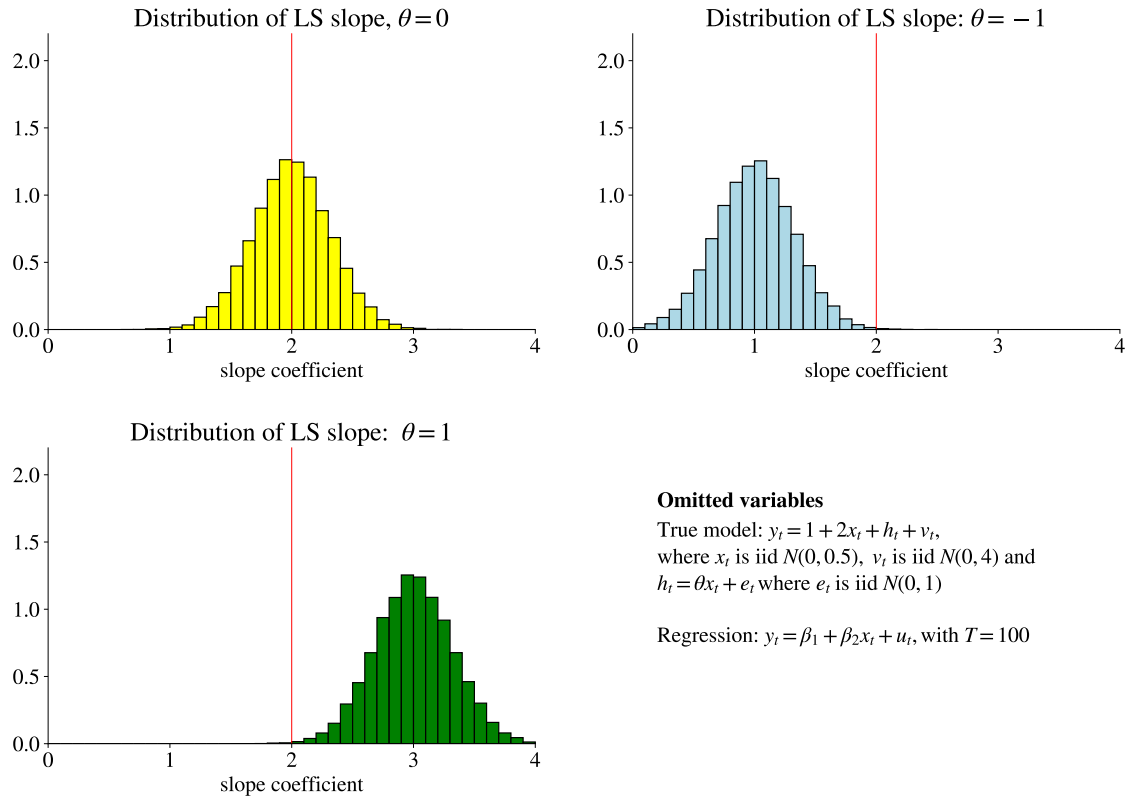


Figure 5.5: Effect of omitted variables

be correlated with any of the regressors), while the middle term can be written

$$[\ \hat{\theta}_1 \quad \dots \quad \hat{\theta}_L\ ]\gamma \tag{5.7}$$

where $\hat{\theta}_i$ is the (column) vector of coefficients obtained by regressing h_{it} on x_t

$$\hat{\theta}_i = S_{xx}^{-1} \sum_{t=1}^{T} x_t h_{it}. \tag{5.8}$$

See Figure 5.5.

Together this shows that the probability limit of $\hat{\beta}$ is

$$\operatorname{plim} \hat{\beta} = \beta + [\ \operatorname{plim} \hat{\theta}_1 \quad \dots \quad \operatorname{plim} \hat{\theta}_L\ ]\gamma. \tag{5.9}$$

This analysis shows that $\hat{\beta}$ incorporates how x_t comoves with the h_t. In case they are uncorrelated ($\theta_i = \mathbf{0}$), then omitting the h_t variables does not affect the point estimates of β. However, if they are correlated, then the point estimates $\hat{\beta}$ are inconsistent (and

biased) in the sense of being systematically different from the true β values in (5.3).

Notice the following:

- $\hat{\beta}$ from (5.4) *is* actually the right number to use if we want to predict: "given x_t, what is the best guess of y_t?" The reason is that $\hat{\beta}$ factors in also how x_t predicts h_t (which affects y_t).
- $\hat{\beta}$ from (5.4) *is not* the right number to use if we want to understand an economic mechanism: "if we increase x_{it}, by one unit (but holding all other variables constant), what is the likely effect on y_t?" The reason is that we here need a consistent estimate of β. Even in this case, (5.9) and an economic theory might be useful in assessing (guessing) the sign of the bias.

5.3.2 Autocorrelated Errors Combined with Lagged Dependent Variable

As an example of how autocorrelated errors combined with a lagged dependent variable as regressor leads to inconsistent OLS estimates, consider

$$y_t = \beta_1 + \beta_2 x_t + \beta_3 y_{t-1} + u_t, \text{ where} \tag{5.10}$$

$$u_t = \upsilon_t + \theta \upsilon_{t-1}, \upsilon_t \text{ iid.} \tag{5.11}$$

As a special case, $\beta_2 = 0$ gives an ARMA(1,1) model, which is a well known case which cannot be estimated by OLS. See Figure 5.4.

The issue is that y_{t-1} is correlated with the lagged shock (υ_{t-1}) and hence with the OLS residuals u_t: $\text{Cov}(y_{t-1}, u_t) \neq 0$. This is a common problem in dynamic models.

Remark 5.5 *(AR(1) with autocorrelated errors, theoretical result*) Consider the case in (5.10)–(5.11) but where $\beta_2 = 0$ so the regression is an AR(1) but the errors follow an MA(1) process. In the limit, the OLS estimate is $\hat{\beta}_3 = \text{Cov}(y_t, y_{t-1})/\text{Var}(y_{t-1})$. Using (5.10) to replace y_t gives $\hat{\beta}_3 = \beta_3 + \text{Cov}(\upsilon_t + \theta\upsilon_{t-1}, y_{t-1})/\text{Var}(y_{t-1})$. Since the 2nd term is non-zero (it is $\theta\,\text{Var}(\upsilon_{t-1})/\text{Var}(y_{t-1})$), this is not equal to β_3.*

5.3.3 Measurement Errors in a Regressor

As an example of how measurement errors in a regressor gives inconsistent OLS estimates, consider a simple (true) model like

$$y_t = \beta_1 + \beta_2 w_t + \upsilon_t. \tag{5.12}$$

However, we estimate with a proxy x_t for the regressor w_t

$$y_t = \beta_1 + \beta_2 x_t + u_t, \text{ with} \tag{5.13}$$

$$x_t = w_t + e_t, \tag{5.14}$$

where e_t is a measurement error. This is a common problem in micro data, including corporate finance. This leads to $\text{Cov}(x_t, u_t) \neq 0$ (since both x_t and u_t depend on the measurement error e_t) so OLS is inconsistent for estimating β_2. See Figure 5.6.

To see the precise source of the inconsistency, solve for $w_t = x_t - e_t$, use in correct model (5.12) to get

$$\begin{aligned} y_t &= \beta_1 + \beta_2 (x_t - e_t) + \upsilon_t \\ &= \beta_1 + \beta_2 x_t \underbrace{-\beta_2 e_t + \upsilon_t}_{u_t}. \end{aligned} \tag{5.15}$$

From (5.14) we know that x_t is correlated with the measurement error (e_t), which gives $\text{Cov}(x_t, u_t) \neq 0$. In fact, it can be shown that

$$\text{plim}\, \hat{\beta}_2 = \beta_2 \left(1 - \frac{\text{Var}(e_t)}{\text{Var}(w_t) + \text{Var}(e_t)} \right). \tag{5.16}$$

Notice that $\hat{\beta}_2 \to 0$ if the measurement error dominates ($\text{Var}(e_t) \to \infty$), since y_t is not related to the measurement error. In contrast, $\hat{\beta}_2 \to \beta_2$ as measurement vanishes ($\text{Var}(e_t) \to 0$): no measurement error. Measurement errors will thus bias the coefficient towards zero. Any significant coefficient can therefore be seen as a conservative estimate.

Proof. (of (5.16)) To simplify, assume that x_t has a zero mean. From (5.2), we then have $\text{plim}\, \hat{\beta}_2 = \beta_2 + \Sigma_{xx}^{-1} \text{E}(x_t u_t)$. Here, $\Sigma_{xx}^{-1} = 1/\text{Var}(x_t)$, but notice from (5.14) that $\text{Var}(x_t) = \text{Var}(w_t) + \text{Var}(e_t)$ if w_t and e_t are uncorrelated. We also have $\text{E}(x_t u_t) = \text{Cov}(x_t, u_t)$, which from the definition of x_t in (5.14) and of u_t in (5.15) gives

$$\text{Cov}(x_t, u_t) = \text{Cov}(w_t + e_t, -\beta_2 e_t + \upsilon_t) = -\beta_2 \text{Var}(e_t).$$

Together we get

$$\text{plim}\, \hat{\beta}_2 = \beta_2 + \Sigma_{xx}^{-1} \text{E}(x_t u_t) = \beta_2 - \beta_2 \frac{\text{Var}(e_t)}{\text{Var}(w_t) + \text{Var}(e_t)},$$

which is (5.16). ■

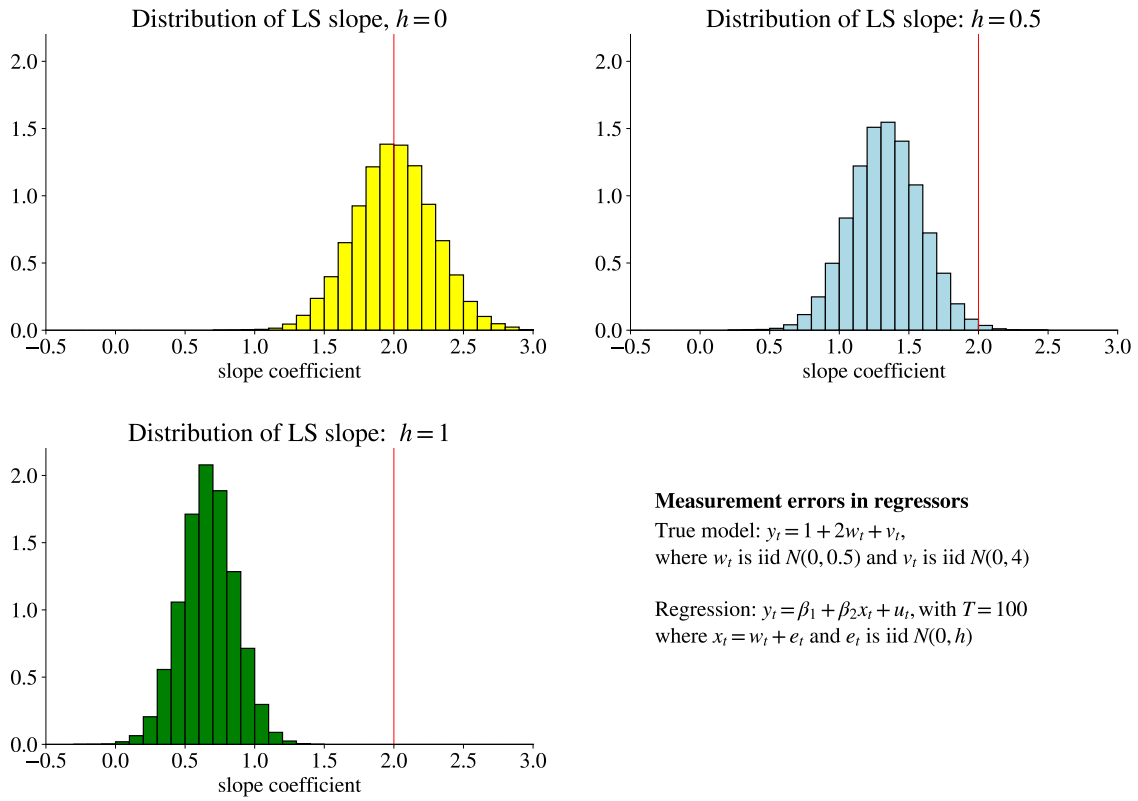


Figure 5.6: Effect of measurement error in regressor

5.3.4 Endogenous Regressors (System of Simultaneous Equations)

Consider the simplest simultaneous equations model for supply and demand on a market. Supply is

$$q_t = \gamma p_t + u_t^s, \ \gamma > 0, \tag{5.17}$$

and demand is

$$q_t = \beta p_t + \alpha A_t + u_t^d, \ \beta < 0, \tag{5.18}$$

where A_t is an observable demand shock (perhaps income).

Example 5.6 *(Supply and Demand*) The system (the "structural form") is therefore*

$$\begin{bmatrix} 1 & -\gamma \\ 1 & -\beta \end{bmatrix} \begin{bmatrix} q_t \\ p_t \end{bmatrix} + \begin{bmatrix} 0 \\ -\alpha \end{bmatrix} A_t = \begin{bmatrix} u_t^s \\ u_t^d \end{bmatrix}.$$

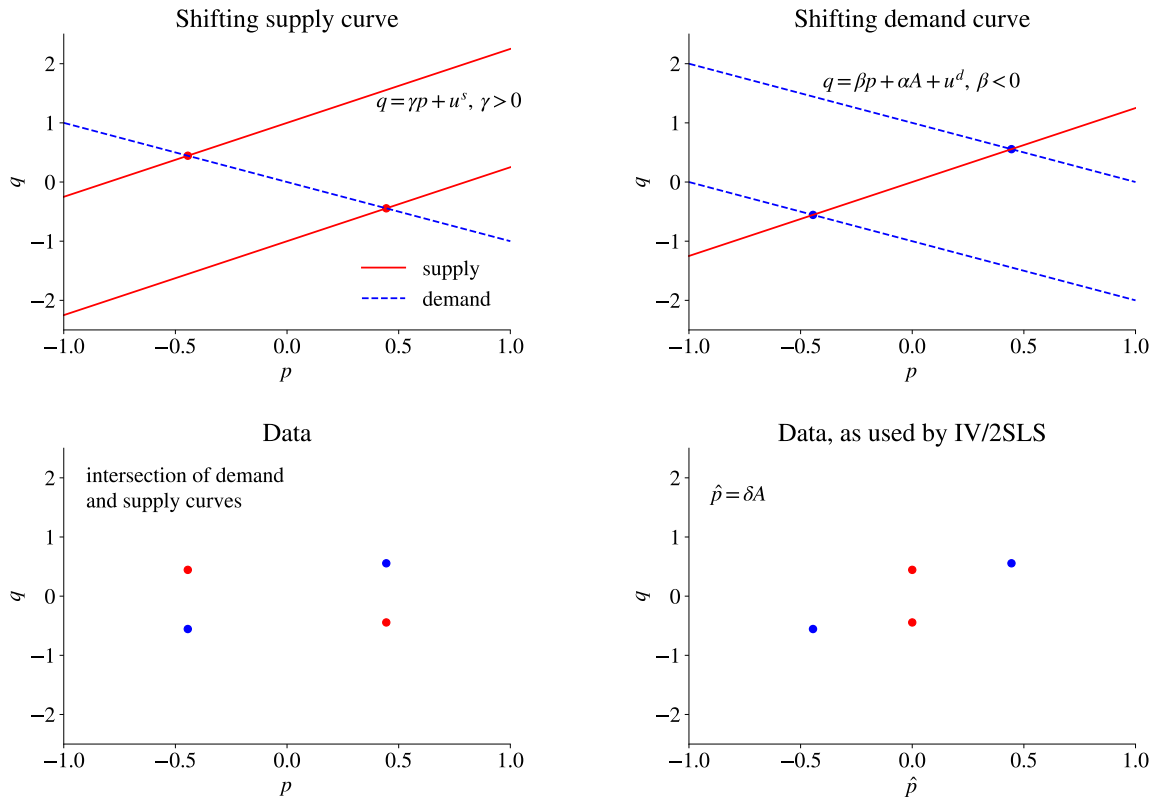


Figure 5.7: Illustration of demand and supply curves

This can be solved in terms of the exogenous variables (the "reduced form") as

$$\begin{bmatrix} q_t \\ p_t \end{bmatrix} = \begin{bmatrix} -\frac{\gamma}{\beta-\gamma}\alpha \\ -\frac{1}{\beta-\gamma}\alpha \end{bmatrix} A_t + \begin{bmatrix} \frac{\beta}{\beta-\gamma} & -\frac{\gamma}{\beta-\gamma} \\ \frac{1}{\beta-\gamma} & -\frac{1}{\beta-\gamma} \end{bmatrix} \begin{bmatrix} u_t^s \\ u_t^d \end{bmatrix}.$$

Suppose we try to estimate the supply equation (5.17) by LS. However, p_t is correlated with u_t^s (since $u_t^s \to q_t \to p_t$), so we cannot hope that LS will be consistent. See 5.7 for an illustration (disregard the IV/2SLS subfigure for now). It is clear that the OLS estimate $\hat{\gamma}_{OLS}$ will be a mixture of the true γ and β values (and other things), see Example 5.7 and Figure 5.8. It is sometimes possible to use economic theory to assess the sign of the bias. In some such cases, it can be argued that any significant coefficient is a conservative estimate.

Example 5.7 *(Supply equation with LS*) Using the reduced form from Example 12.4, it is straightforward to show that the probability limit of the OLS estimate of γ is (assuming*

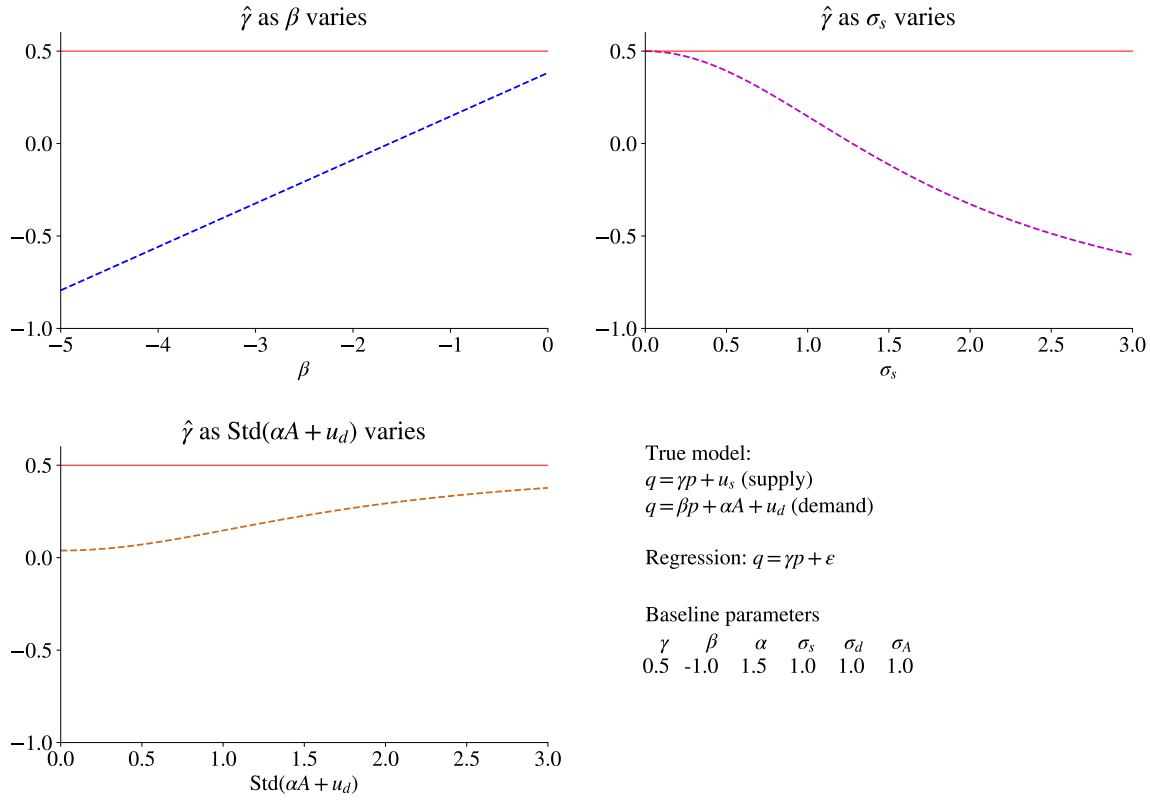


Figure 5.8: OLS estimate of γ in supply equation

that the supply and demand shocks are uncorrelated)

$$\begin{aligned}\operatorname{plim}\hat{\gamma}_{OLS} &= \frac{\operatorname{Cov}(q_t, p_t)}{\operatorname{Var}(p_t)} \\ &= \frac{\gamma\alpha^2\operatorname{Var}(A_t) + \gamma\operatorname{Var}(u_t^d) + \beta\operatorname{Var}(u_t^s)}{\alpha^2\operatorname{Var}(A_t) + \operatorname{Var}(u_t^d) + \operatorname{Var}(u_t^s)}.\end{aligned}$$

First, suppose the supply shocks are zero, $\operatorname{Var}(u_t^s) = 0$*, then* $\operatorname{plim}\hat{\gamma} = \gamma$*, so we indeed estimate the supply elasticity, as we wanted. Think of a fixed supply curve, and a demand curve which moves around. These point of* p_t *and* q_t *should trace out the supply curve. It is clearly* u_t^s *that causes a simultaneous equations problem in estimating the supply curve:* u_t^s *affects both* q_t *and* p_t *and the latter is the regressor in the supply equation. With no movements in* u_t^s *there is no correlation between the shock and the regressor. Second, now suppose instead that the both demand shocks are zero (both* $A_t = 0$ *and* $\operatorname{Var}(u_t^d) = 0$*). Then* $\operatorname{plim}\hat{\gamma} = \beta$*, so the estimated value is not the supply, but the demand elasticity. Not good. This time, think of a fixed demand curve, and a supply curve which moves around.*

5.4 Asymptotic Normality

Reference: Greene (2018) 4.4; Hamilton (1994) 8.2; Davidson (2000) 3

Issue: what is the distribution of your estimator in large samples?

5.4.1 Central Limit Theorems

We need some basic facts about statistics (central limit theorems) for the discussion of asymptotic normality.

Remark 5.8 *(Convergence in distribution) Let $\hat{z}$ be a random variable (which depends on the sample size T) and let Z be another random variable that does not. If $\lim_{T\to\infty} \Pr(\hat{z} < c) = \Pr(Z < c)$ for every c, then $\hat{z}$ converges in distribution to the random variable Z. Notation $\hat{z} \xrightarrow{d} Z$.*

Remark 5.9 *(Central limit theorem, simple version) A CLT says that $\sqrt{T}\bar{x} \xrightarrow{d} N()$, that is, becomes normally distributed when T becomes really large. This holds for many random variables (although exceptions exist). Notice that the distribution of $\bar{x}$ converges to a spike as T increases (LLN), but the distribution of $\sqrt{T}\bar{x}$ converges to a normal distribution. See Figure 5.3.*

Remark 5.10 *(Continuous mapping theorem.) Let the random variables $\hat{z}$ and $\hat{q}$ and the non-random a_T be such that $\hat{z} \xrightarrow{d} Z$, $\hat{q} \xrightarrow{p} Q$ (a finite and positive definite matrix) and $a_T \to a$ (a traditional limit). Also, let $g(z, y, a)$ be a continuous function. Then $g(\hat{z}, \hat{q}, a_T) \xrightarrow{d} g(Z, Q, a)$.*

Example 5.11 *For instance, the sequences in Remark 5.10 could be $\hat{z} = \sqrt{T}\Sigma_{t=}^{T} w_t/T$ (the scaled sample average of a random variable w_t); $\hat{q} = \Sigma_{t=}^{T} w_t^2/T$ (the sample second moment); and $a = \Sigma_{t=1}^{T} 0.7^t$ (which converges to 2.333).*

5.4.2 Asymptotic Normality of OLS

Subtract β from both sides of (5.1), and multiply both sides by $\sqrt{T}$ to get

$$\sqrt{T}(\hat{\beta} - \beta) = \underbrace{\left(\frac{1}{T}\sum_{t=1}^{T} x_t x_t'\right)^{-1}}_{\to \Sigma_{xx}^{-1}} \underbrace{\sqrt{T}\frac{1}{T}\sum_{t=1}^{T} x_t u_t}_{\sqrt{T}\times\text{sample average}} \tag{5.19}$$

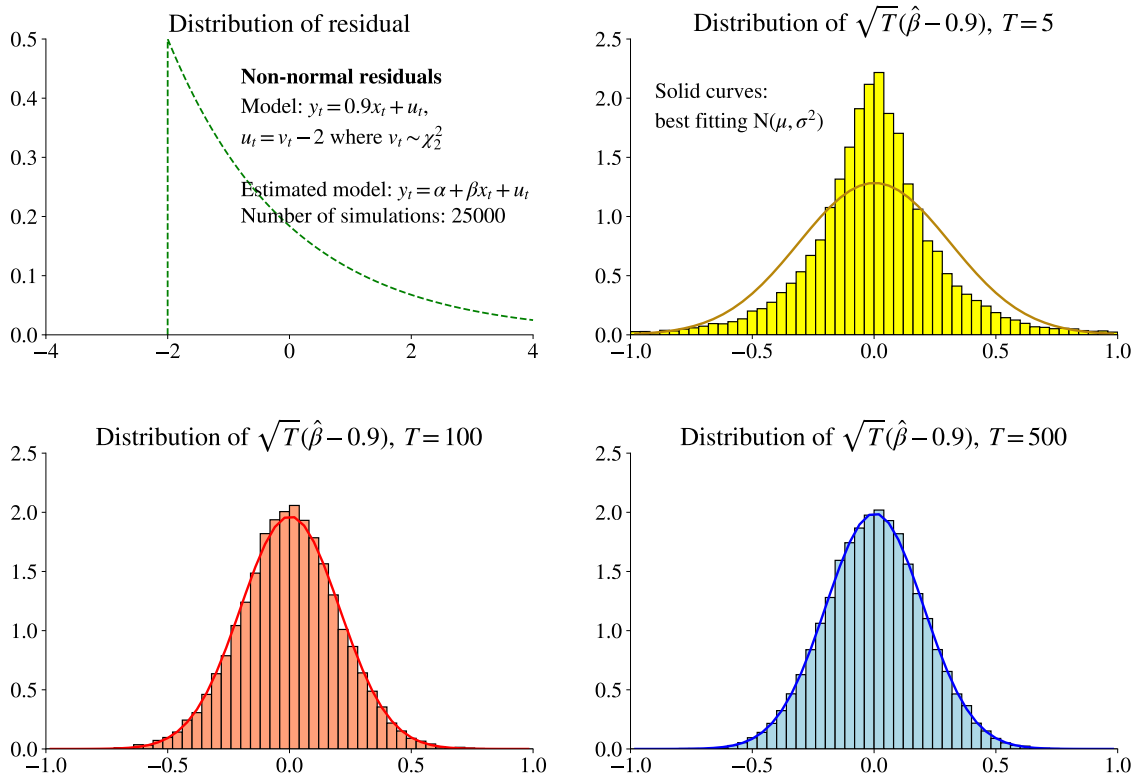


Figure 5.9: Results from a Monte Carlo experiment with thick-tailed errors.

The first term converges (by a LLN) in probability limit to Σ_{xx}^{-1}, assuming the x_t is such that the limit is well defined. (This is like the $\hat{q} \xrightarrow{p} Q$ variables in Remark 5.10.) The second term is $\sqrt{T}\times$sample average (of $x_t u_t$), which (by a CLT) will (typically) converge in distribution to a normally distributed variable. According to Remark 5.10, we should therefore expect $\sqrt{T}\ \hat{\beta}$ to be normally distributed in *large* samples—even if the residual doesn't have a normal distribution. See Figure 5.9 for an example.

According to Remark 5.10 and (5.19) $\sqrt{T}(\hat{\beta} - \beta)$ converges in distribution to Σ_{xx}^{-1} times a normally distributed variable (vector). If OLS is consistent, then the normal distribution has a zero mean ($\mathrm{E}\, x_t u_t = 0$). Let Σ denote the variance-covariance matrix of $\Sigma_{t=1}^{T} x_t u_t / \sqrt{T}$

$$\Sigma = \mathrm{Var}\left(\sum_{t=1}^{T} x_t u_t / \sqrt{T}\right). \tag{5.20}$$

Together, we then have

$$\sqrt{T}(\hat{\beta} - \beta) \xrightarrow{d} N\left(0, \Sigma_{xx}^{-1} \Sigma \Sigma_{xx}^{-1}\right) \tag{5.21}$$

Remark 5.12 *(*Relation to expression in earlier chapters) We have previously shown that*

$$\text{Var}(\hat{\beta}) = S_{xx}^{-1} S S_{xx}^{-1}, \text{ where } S_{xx} = \Sigma_{t=1}^{T} x_t x_t' \text{ and } S = \text{Var}(\Sigma_{t=1}^{T} x_t u_t).$$

To see that this is really the same as (5.21), notice two things. First, if $\text{Var}(\sqrt{T}\hat{\beta}) = A$, *then* $\text{Var}(\hat{\beta}) = A/T$. *This transforms the covariance matrix in (5.21) to*

$$\text{Var}(\hat{\beta}) = \Sigma_{xx}^{-1} \Sigma \Sigma_{xx}^{-1} / T.$$

Second, notice that $\Sigma_{xx} = S_{xx}/T$ *(in probability limits) and that* $\Sigma = S/T$. *Use this in the previous equation to get*

$$\text{Var}(\hat{\beta}) = (TS_{xx}^{-1})(S/T)(TS_{xx}^{-1})/T$$

and cancel the T *terms to get* $\text{Var}(\hat{\beta}) = S_{xx}^{-1} S S_{xx}^{-1}$.

5.5 Spurious Regressions

Strong trends often causes problems in econometric models where y_t is regressed on x_t. In essence, if no trend is included in the regression, then x_t will appear to be significant, just because it is a proxy for that trend. The same holds for non-stationary processes, even if they have no deterministic trends. The reason is that the innovations accumulate and the series therefore tend to be trending in small samples. Asymptotic results are typically of *little use* here, since the non-stationarity means that the asymptotic results are degenerate (for instance, infinite variance). A warning sign of a spurious regression is when $R^2 > DW$ statistic.

Empirical Example 5.13 *(Regressing the price level on GDP) See Figure 5.10 for results from regressing the U.S. price level (GDP deflator) on output (GDP level). The results indicate a very significant regression slope, but extreme autocorrelation. This is likely to be a spurious regression. Also, economics would suggest that nominal (price level) and real variables (output) variables are driven by completely different factors.*

See Figures 5.11–5.14 for a Monte Carlo simulation.

For trend-stationary data, this problem is easily solved by detrending with a linear trend (before estimating or just adding a trend to the regression).

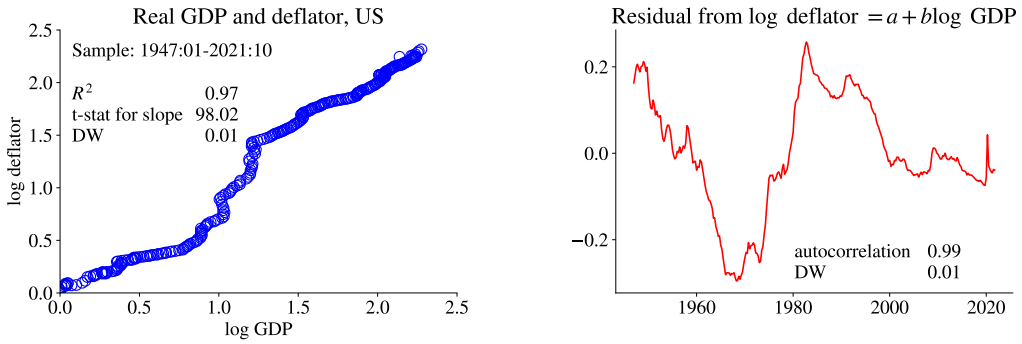


Figure 5.10: Example of a spurious regression

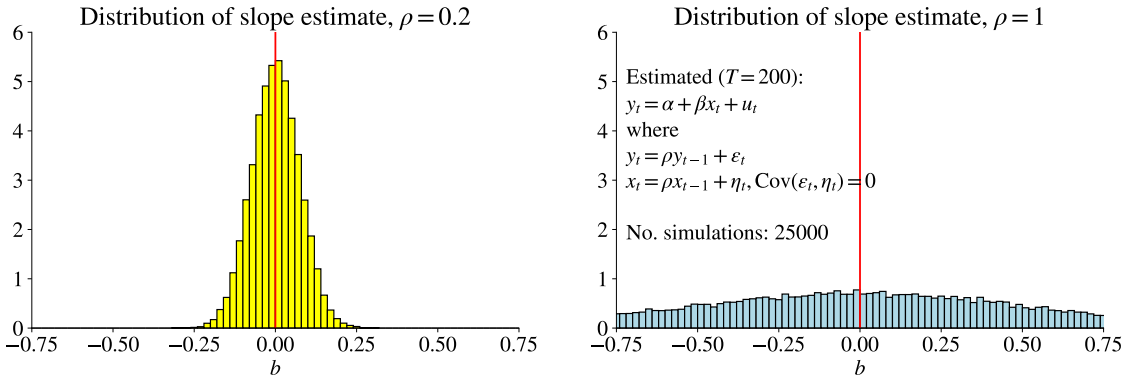


Figure 5.11: Distribution of slope coefficient when y_t and x_t are independent AR(1) processes

However, this is usually a poor method for a unit root processes. What is needed is a first difference. For instance, a first difference of the random walk with drift is

$$\begin{aligned} \Delta y_t &= y_t - y_{t-1} \\ &= \mu + \varepsilon_t, \end{aligned} \tag{5.22}$$

which is white noise (any finite difference, like $y_t - y_{t-s}$, will give a stationary series), so we could proceed by applying standard econometric tools to Δy_t.

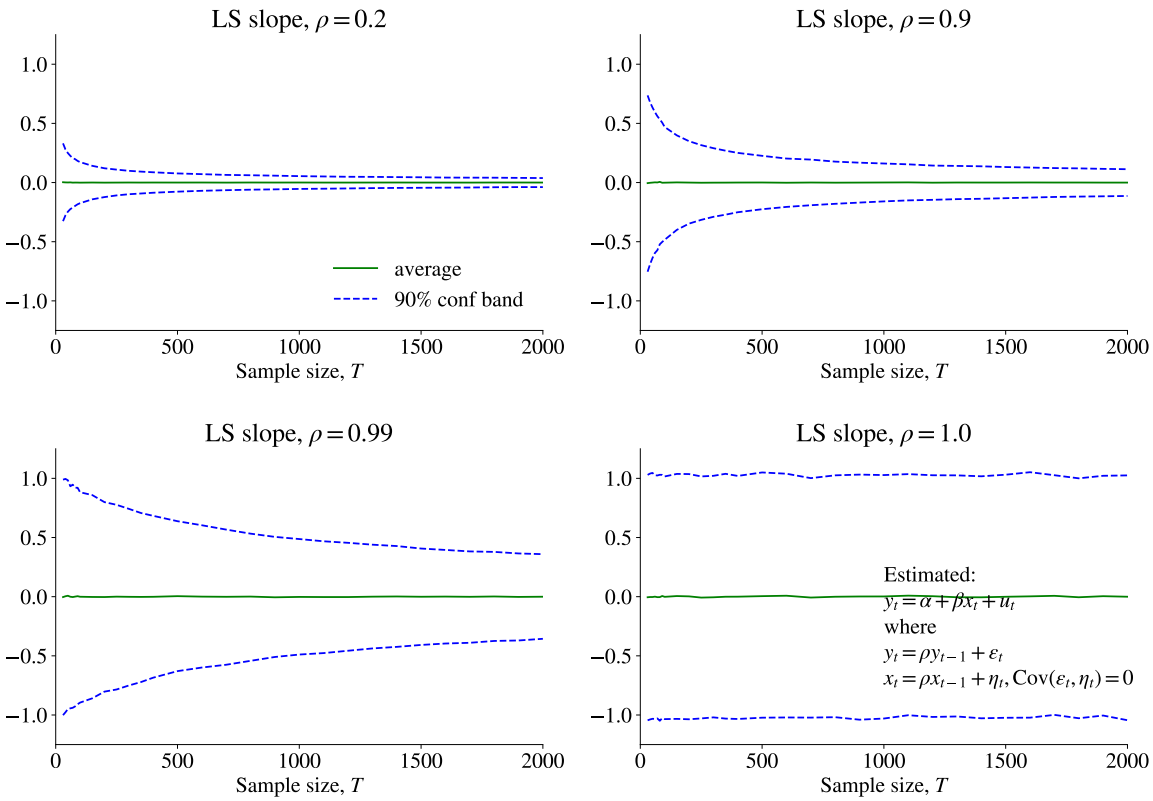


Figure 5.12: Distribution of slope coefficient when y_t and x_t are independent AR(1) processes

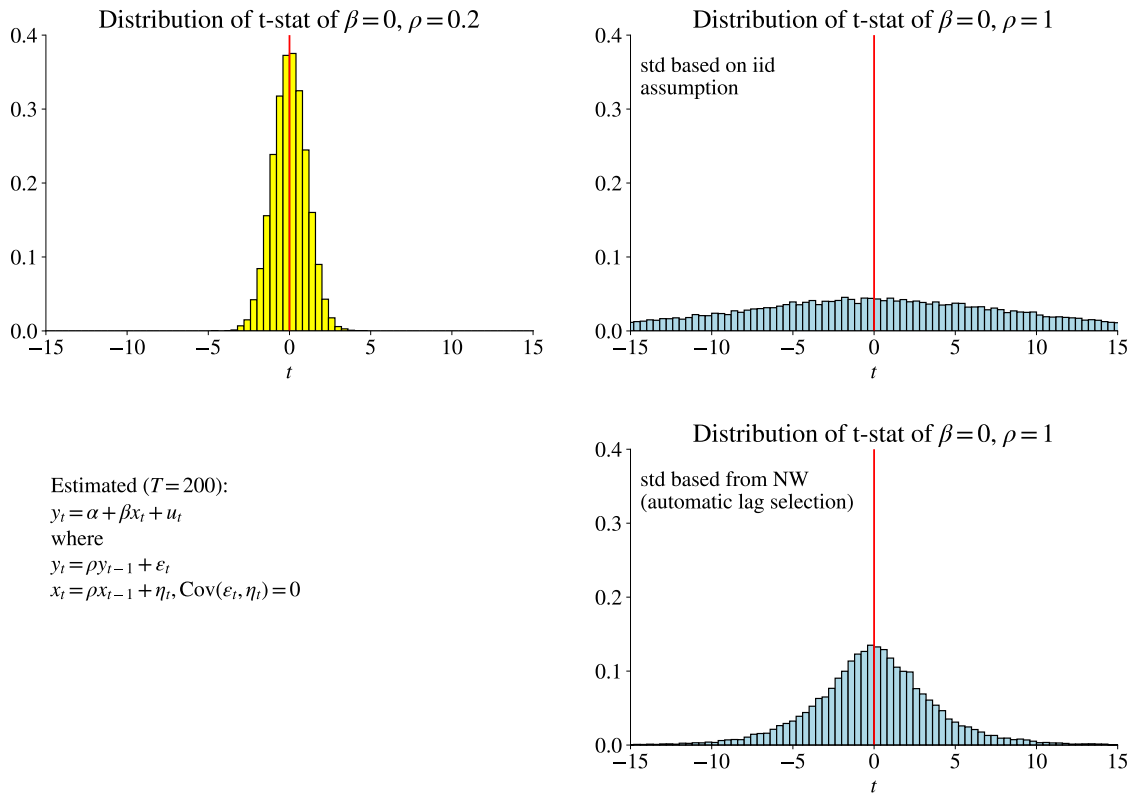


Figure 5.13: Distribution of the t-statistic when y_t and x_t are independent AR(1) processes. See Figure 5.11.

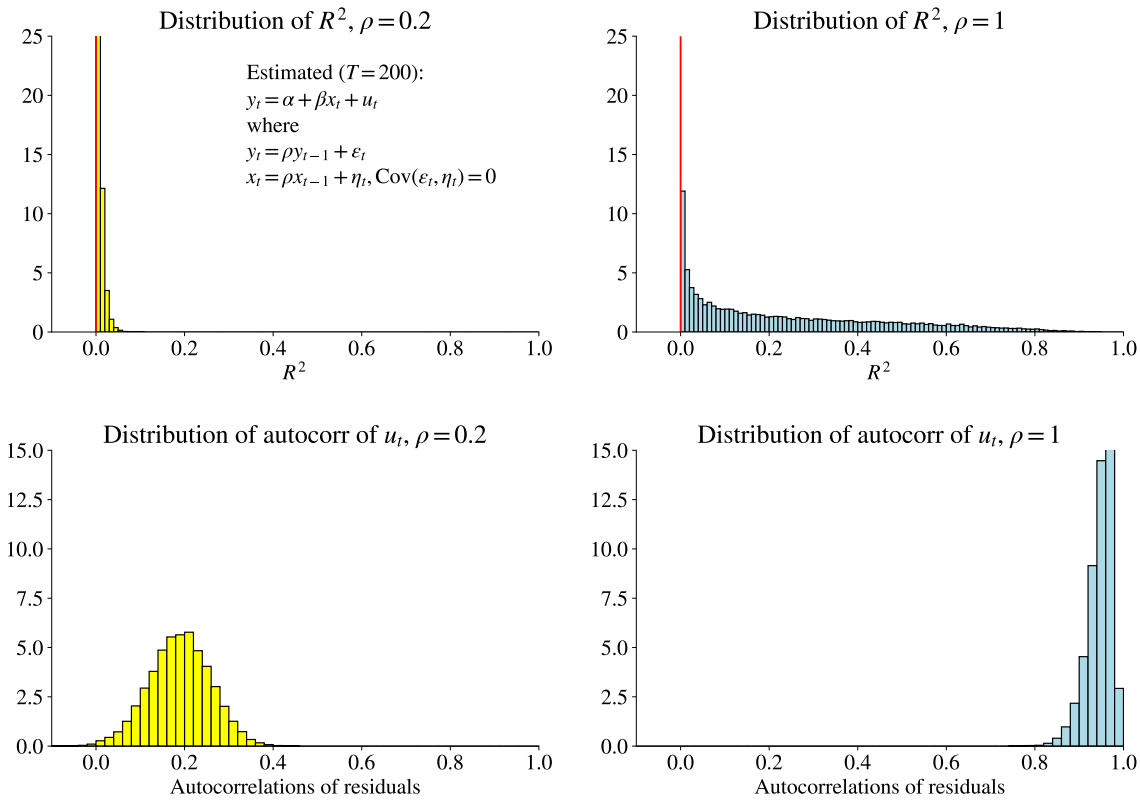


Figure 5.14: Distribution of R^2 and autocorrelation of residuals. See Figure 5.11.

Chapter 6

Simulating the Finite Sample Properties

Reference: Greene (2000) 5.3 and Horowitz (2001)
Additional references: Cochrane (2001) 15.2; Davidson and MacKinnon (1993) 21; Davison and Hinkley (1997); Efron and Tibshirani (1993) (bootstrapping, chap 9 in particular); and Berkowitz and Kilian (2000) (bootstrapping in time series models)

6.1 Introduction

We know the small sample properties of regression coefficients in linear models with fixed regressors and iid normal error terms. When these conditions are not satisfied, then we must rely on asymptotic results or apply Monte Carlo/bootstrap simulations to approximate the small sample properties. For instance, if the regression residuals have autocorrelation and/or heteroskedasticity, then we may either use a consistent estimator of the covariance matrix (Newey-West, White, etc) and apply the usual test by comparing with the asymptotically correct $N(0,1)$ or χ^2_q distributions. Alternatively, we can compare the test statistic (based on either the classical covariance matrix or a consistent one) with a simulated distribution. The advantage of the simulations is that they might provide better approximations of the small sample properties than the asymptotic distribution does.

The results from the simulations can be used to study, for instance, (a) the distribution of a point estimate (to create confidence bands or a standard deviation) or (b) the distribution of a test statistic (to generate appropriate critical values).

How these simulations should be implemented depends crucially on the properties of the model and data: if the residuals are autocorrelated, heteroskedastic, or perhaps correlated across regressions equations. These notes summarize a few typical cases.

Empirical Example 6.1 *(The empirical importance of simulated standard errors) The*

need for using Monte Carlos or bootstraps varies across applications and data sets. For a case where it does not matter much, see Table 6.1, and for a case where it matters, compare the traditional and bootstrapped t-stats in Tables 6.2–6.3.

	α	t (LS)	t (NW)	t (boot)
A (NoDur)	2.77	2.31	2.14	1.90
B (Durbl)	−0.19	−0.09	−0.10	−0.09
C (Manuf)	0.19	0.22	0.21	0.19
D (Enrgy)	1.44	0.64	0.62	0.60
E (HiTec)	−0.86	−0.55	−0.55	−0.49
F (Telcm)	1.08	0.72	0.69	0.60
G (Shops)	1.38	1.09	1.03	1.02
H (Hlth)	2.23	1.46	1.48	1.47
I (Utils)	2.93	1.79	1.75	1.84
J (Other)	−0.92	−0.96	−0.91	−0.75

Table 6.1: Estimates of CAPM on US industry portfolios 1970:01-2021:12. NW uses 1 lag. The bootstrap samples (y_t, x_t) pairs, in blocks of 10 observations and has 3000 simulations.

	2y	3y	4y	5y
factor	1.00	1.85	2.64	3.39
	(6.44)	(6.45)	(6.53)	(6.63)
constant	−0.00	−0.00	−0.00	−0.00
	(−0.00)	(−0.22)	(−0.48)	(−0.77)
R^2	0.12	0.12	0.13	0.13
obs	684	684	684	684

Table 6.2: Regression of different excess (1-year) holding period returns (in columns, indicating the maturity of the respective bond) on a single forecasting factor and a constant. Numbers in parentheses are t-stats. U.S. data for 1964:01-2021:12.

6.2 Monte Carlo Simulations

6.2.1 Monte Carlo Simulations in the Simplest Case

Monte Carlo simulations is essentially a way to generate many artificial (small) samples from a parameterised model and then estimate the statistic (for instance, a slope coef-

	2y	3y	4y	5y
factor	1.00	1.85	2.64	3.39
	(3.74)	(3.80)	(3.91)	(4.02)
constant	−0.00	−0.00	−0.00	−0.00
	(−0.00)	(−0.10)	(−0.23)	(−0.36)
R^2	0.12	0.12	0.13	0.13
obs	684	684	684	684

Table 6.3: Regression of different excess (1-year) holding period returns (in columns, indicating the maturity of the respective bond) on a single forecasting factor and a constant. U.S. data for 1964:01-2021:12. Numbers in parentheses are t-stats. Bootstrapped standard errors, with blocks of 10 observations.

ficient) on each of those samples. The distribution (across the artificial samples) of the statistic is then used as an approximation of the small sample distribution of the estimator.

The following is an example of how Monte Carlo simulations could be done in the special case of a linear model with a scalar dependent variable

$$y_t = x_t'\beta + u_t, \tag{6.1}$$

where u_t is iid $N(0, \sigma^2)$ and x_t is stochastic but independent of $u_{t\pm s}$ for all s. (This means that x_t cannot include lags of y_t.)

Suppose we want to find the small sample distribution of a function of the estimate, $g(\hat{\beta})$. To do a Monte Carlo experiment, we need information on *(i)* the coefficients β; *(ii)* the variance of u_t, σ^2; *(iii)* and a process for x_t.

The process for x_t is typically estimated from the data on x_t (for instance, a VAR system $x_t = A_1 x_{t-1} + A_2 x_{t-2} + e_t$). Alternatively, we could simply use the actual sample of x_t and repeat it.

The values of β and σ^2 are often a mix of estimation results and theory. In some case, we simply take the point estimates. In other cases, we adjust the point estimates so that $g(\beta) = 0$ holds, that is, so you *simulate the model under the null hypothesis* in order to study the size of tests and to find valid critical values for small samples. Alternatively, you may *simulate the model under an alternative hypothesis* in order to study the power of the test using either critical values from either the asymptotic distribution or from a (perhaps simulated) small sample distribution.

To make this discussion a bit more concrete, suppose you want to use these simulations to get a 5% critical value for testing the null hypothesis $g(\beta) = 0$. The Monte Carlo

experiment follows these steps.

1. Construct an artificial sample of the regressors (see above), $\tilde{x}_t$ for $t = 1, \ldots, T$. Draw random numbers $\tilde{u}_t$ for $t = 1, \ldots, T$ from a prespecified distribution (for instance, $N(0, \sigma^2)$) and use those together with the artificial sample of $\tilde{x}_t$ to calculate an artificial sample $\tilde{y}_t$ for $t = 1, \ldots, T$ from

$$\tilde{y}_t = \tilde{x}_t'\beta + \tilde{u}_t, \tag{6.2}$$

   by using the prespecified values of the coefficients β (perhaps your point estimates).

2. Calculate an estimate $\tilde{\beta}$ and record it along with the value of $g(\tilde{\beta})$ and perhaps also the test statistic of the hypothesis that $g(\beta) = 0$.

3. Repeat the previous steps N (3000, say) times. The more times you repeat, the better is the approximation of the small sample distribution.

4. Sort your simulated $\tilde{\beta}$, $g(\tilde{\beta})$, and the test statistic in ascending order. For a one-sided test (for instance, a chi-square test), take the $(0.95N)$th observations in this sorted vector as your 5% critical value. For a two-sided test (for instance, a t-test), take the $(0.025N)$th and $(0.975N)$th observations as the 5% critical values. You could also record how many times the 5% critical values from the asymptotic distribution would reject a true null hypothesis.

5. You may also want to plot a histogram of $\tilde{\beta}$, $g(\tilde{\beta})$, and the test statistic to see if there is a small sample bias, and how the distribution looks like. Is it close to normal? How wide is it? You could also estimate the variance-covariance matrix of $\tilde{\beta}$ by treating each estimate (from each simulation) as an observation—and then estimate the covariance matrix across these observations.

We use the same basic procedure when y_t is a vector, except that we have to consider correlations across the elements of the vector of residuals u_t. For instance, we could generate the vector $\tilde{u}_t$ from a $N(\mathbf{0}, \Sigma)$ distribution—where Σ is the variance-covariance matrix of u_t.

Remark 6.2 *(Generating $N(\mu, \Sigma)$ random numbers*) Suppose you want to draw an $n \times 1$ vector ε_t of $N(\mu, \Sigma)$ variables. Use the Cholesky decomposition of Σ to calculate the lower triangular P such that $\Sigma = PP'$. Draw u_t from an $N(0, I_n)$ distribution, and*

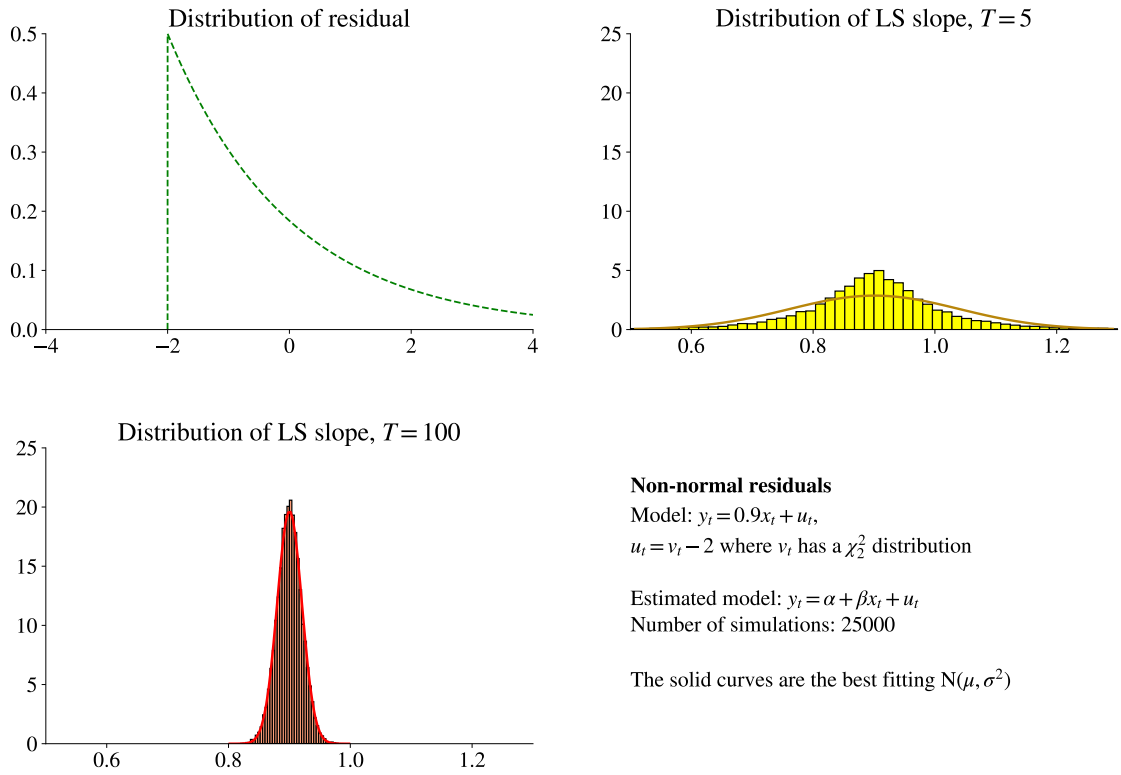


Figure 6.1: Results from a Monte Carlo experiment with fat-tailed errors

define $\varepsilon_t = \mu + Pu_t$. *Note that* $\text{Cov}(\varepsilon_t) = \text{E}\, Pu_t u_t' P' = PIP' = \Sigma$. *(To watch out for: the convention for whether to calculate* P *or* P' *differs across computer languages.)*

It is straightforward to sample the errors from other distributions than the normal, for instance, a student-t distribution. Equipped with uniformly distributed random numbers, you can always (numerically) invert the cumulative distribution function (cdf) of any distribution to generate random variables from any distribution by using the probability transformation method. See *Figure 6.1* for an example.

Remark 6.3 *(The probability transformation method*) A random variable* Y *has the cdf* $u = \Pr(Y \leq y) = F(y)$*, where* y *is a number. Clearly,* u *is a probability and thus between 0 and 1. Draw random numbers* u_i *from a uniform distribution over* $(0, 1)$*. Then, calculate* $y_i = F^{-1}(u_i)$*, where* $F^{-1}()$ *is the inverse of* $F()$*. A sample of* y_i *will have the cdf* F*.*

Example 6.4 *(The probability transformation method*) The exponential cdf is* $u = 1 -$

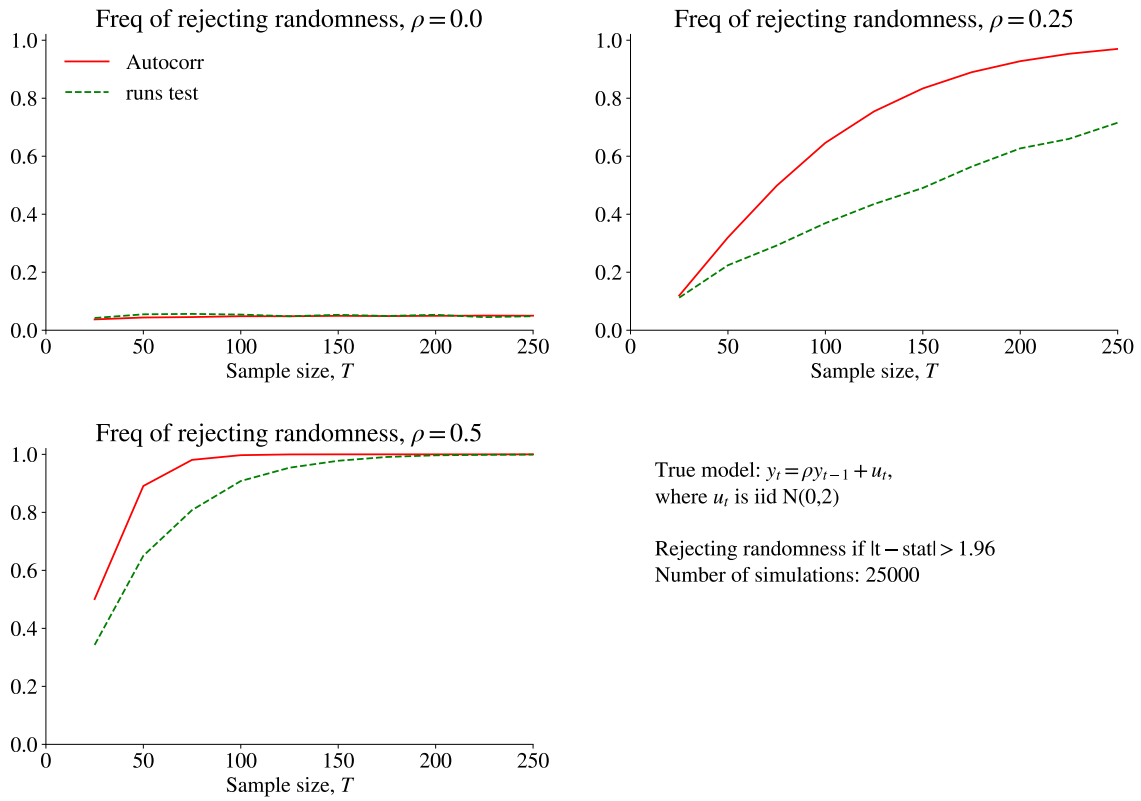


Figure 6.2: Results from a Monte Carlo experiment on two methods of testing for randomness.

$\exp(-\theta y)$ *with inverse* $y = -\ln(1-u)/\theta$. *Draw* u_i *from* $U(0,1)$ *and transform to* y_i *to get an exponentially distributed variable.*

6.2.2 Monte Carlo Simulations when x_t Includes Lags of y_t

When x_t contains lags of y_t, then we must set up the simulations so that temporal link is preserved in every artificial sample which we create. For instance, suppose x_t includes y_{t-1} and another vector z_t of variables which are independent of $u_{t\pm s}$ for all s

$$\begin{aligned} y_t &= x_t'\beta + u_t \\ &= \gamma y_{t-1} + \phi' z_t + u_t. \end{aligned} \tag{6.3}$$

We can then generate an artificial sample as follows. First, create a sample $\tilde{z}_t$ for $t = 1, \ldots, T$ by some time series model (for instance, a VAR) or by taking the observed

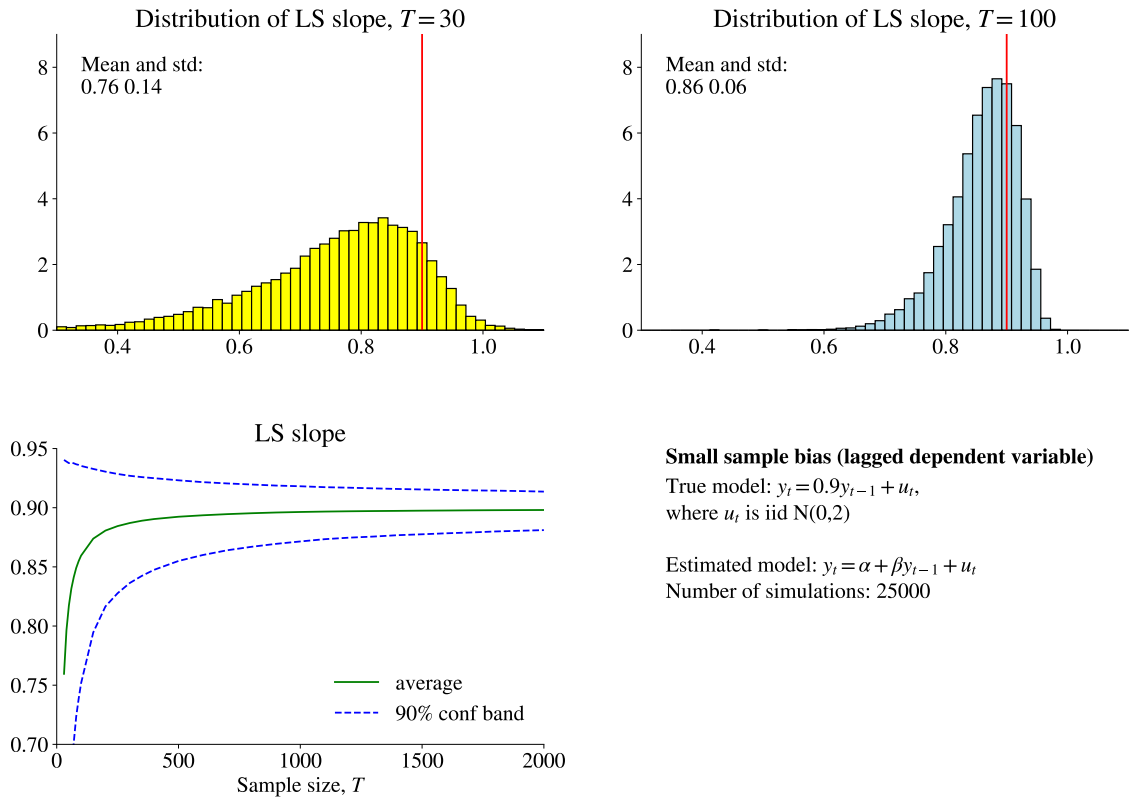


Figure 6.3: Results from a Monte Carlo experiment of LS estimation of the AR coefficient.

sample itself. Second, observation t of $(\tilde{x}_t, \tilde{y}_t)$ is generated recursively as

$$\tilde{y}_t = \gamma \tilde{y}_{t-1} + \phi' \tilde{z}_t + \tilde{u}_t \text{ for } t = 1, \ldots T \tag{6.4}$$

Notice that this makes sure that $\tilde{y}_{t-1}$ is the lagged value of $\tilde{y}_t$ (from the same artificial sample). We clearly need the initial value $\tilde{y}_0$ (for instance, a randomly picked number from the sample of y_t) to start up the artificial sample—and then the rest of the sample ($t = 1, 2, ...$) is calculated recursively. To reduce the importance of the initial value, you may choose to generate $100 + T$ values and then discard the first 100 observations. See Figures 6.2–6.3 for examples.

Remark 6.5 *(Monte Carlo for a VAR system) For a* VAR(2) model *(where there is no z_t)*

$$y_t = A_1 y_{t-1} + A_2 y_{t-2} + u_t,$$

the procedure is straightforward. First, estimate the model on data and record the esti-

mates (A_1, A_2, $\text{Var}(u_t)$). Second, draw a new time series of residuals, $\tilde{u}_t$ for $t = 1, \ldots, T$ and construct an artificial sample recursively (first $t = 1$, then $t = 2$ and so forth) as

$$\tilde{y}_t = A_1\tilde{y}_{t-1} + A_2\tilde{y}_{t-2} + \tilde{u}_t.$$

(This requires some starting values for y_{-1} and y_0.) Third, re-estimate the model on the artificial sample, $\tilde{y}_t$ for $t = 1, \ldots, T$.

6.2.3 Monte Carlo Simulations with non-iid Errors

It is more difficult to handle non-iid errors, like those with autocorrelation and heteroskedasticity. We then need to model the error process and generate the errors from that model.

When the errors are *autocorrelated*, then we could estimate the error process from the fitted errors and then generate artificial samples of errors (here by an AR(2))

$$\tilde{u}_t = a_1\tilde{u}_{t-1} + a_2\tilde{u}_{t-2} + \tilde{\varepsilon}_t, \tag{6.5}$$

where $\tilde{\varepsilon}_t$ are iid.

See Figure 6.4 for an illustration.

Alternatively, *heteroskedastic errors* can be generated by, for instance, a GARCH(1,1) model

$$u_t \sim N(0, \sigma_t^2), \text{ where } \sigma_t^2 = \omega + \alpha u_{t-1}^2 + \beta\sigma_{t-1}^2. \tag{6.6}$$

However, this specification does not account for any link between the volatility and the regressors (squared)—as tested for by White's test. This would invalidate the usual OLS standard errors and therefore deserves to be taken seriously. A simple, but crude, approach is to generate residuals from a $N(0, \sigma_t^2)$ process, but where σ_t^2 is approximated by the fitted values from

$$\varepsilon_t^2 = c'w_t + \eta_t, \tag{6.7}$$

where w_t include the squares and cross product of all the regressors.

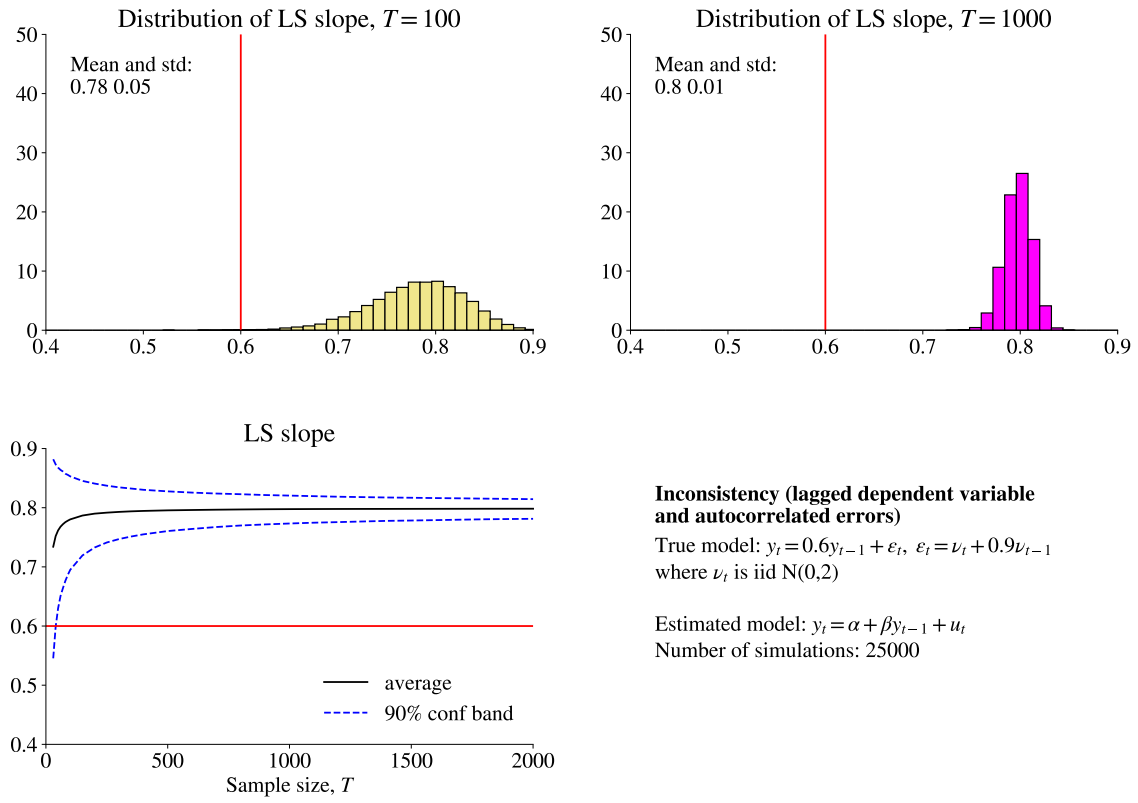


Figure 6.4: Results from a Monte Carlo experiment of LS estimation of the AR coefficient when data is from an ARMA process.

6.3 Bootstrapping

6.3.1 Bootstrapping in the Simplest Case

Bootstrapping is another way to do simulations, where we construct artificial samples by sampling from the actual data (this is sometimes called a non-parametric bootstrap, whereas a parametric bootstrap is basically a Monte Carlo simulation). The advantage of the bootstrap is then that we do not have to try to estimate the process of the errors and regressors (as we do in a Monte Carlo experiment). This means that we do not have to make any strong assumption about the distribution of the errors.

The bootstrap approach works particularly well when the errors are iid and independent of x_{t-s} for all s. (This means, among other things, that x_t cannot include lags of y_t.) We here consider bootstrapping the linear model (6.1), for which we have point estimates (perhaps from LS) and fitted residuals. The procedure is then similar to the Monte Carlo

approach, except that the artificial sample is generated somewhat differently. In particular, Step 1 in the Monte Carlo simulation is replaced by the following:

1. Construct an artificial sample $\tilde{y}_t$ for $t = 1, \ldots, T$ by

$$\tilde{y}_t = x_t'\hat{\beta} + \tilde{u}_t, \tag{6.8}$$

where $\tilde{u}_t$ is drawn with replacement ("residual resampling") from the fitted residuals ($\tilde{u}_t = \hat{u}_s$ where s is the random draw) and where $\hat{\beta}$ is the point estimate from the original sample. Clearly, x_t is just the original data.

Example 6.6 *With $T = 3$, the artificial sample could be*

$$\begin{bmatrix} (\tilde{y}_1, \tilde{x}_1) \\ (\tilde{y}_2, \tilde{x}_2) \\ (\tilde{y}_3, \tilde{x}_3) \end{bmatrix} = \begin{bmatrix} (x_1'\hat{\beta} + \hat{u}_2, x_1) \\ (x_2'\hat{\beta} + \hat{u}_1, x_2) \\ (x_3'\hat{\beta} + \hat{u}_2, x_3) \end{bmatrix}.$$

The approach in (6.8) works also when y_t is a vector of dependent variables. In this case we draw the whole vector $\tilde{u}_t$ together to retain the cross-sectional correlation of the residuals.

The theoretical motivation for why bootstraps work is that the distribution of the fitted residuals converge to the true distribution as the sample size increases. In this sense, the bootstrap relies on asymptotic results, just like most traditional tests rely on a central limit theorem. The key point, however, is that the bootstrap often has smaller distortions (for instance, to the rejection frequency) than traditional tests have.

Remark 6.7 *(Bootstrapped confidence bands) Using the simulated* 0.025*th and* 0.975*th quantiles of the bootstrapped $\tilde{\beta}$ values is a way of creating a 95% confidence band, sometimes called Efron's "bootstrap percentile method". The "bootstrap percentile t-method" (also suggested by Efron) is often considered to be an improvement. To implement it, first define $\tilde{t} = (\tilde{\beta} - \hat{\beta})/\operatorname{Std}(\tilde{\beta})$, where $\operatorname{Std}(\tilde{\beta})$ is the standard deviation across the bootstrap estimates. (Sometimes the centering is done by subtracting the average of $\tilde{\beta}$ values instead of the point estimate $\hat{\beta}$). Let $Q(\tilde{t}; 0.025)$ be the 0.025th quantile of $\tilde{t}$ (that is, the 2.5th percentile) and $Q(\tilde{t}; 0.975)$ be the 0.975th quantile. Then, we could define a 95% confidence band as $[\hat{\beta} + Q(\tilde{t}; 0.025)\operatorname{Std}(\hat{\beta}), \hat{\beta} + Q(\tilde{t}; 0.975)\operatorname{Std}(\hat{\beta})]$, where $\operatorname{Std}(\hat{\beta})$ is a consistent estimate of the standard deviation of $\hat{\beta}$.*

One issue with the bootstrap is that it does not directly create observations that obey the null hypothesis, or even a given alternative hypothesis. For instance, it is not straightforward to create samples where a particular coefficient is zero ($\beta_2 = 0$, say). Actually, (6.8) creates a *distribution around the point estimate*, that is, of $\tilde{\beta} - \hat{\beta}$, where $\hat{\beta}$ is the point estimate from the original sample. This is not important if we just want to understand the standard error of a coefficient (since the standard deviation across the bootstrap simulations is already defined in terms of squared deviations around the average value in the bootstraps.). However, it is crucial when we want to understand percentiles of coefficients, t-stats, or χ^2-stats.

An additional complication is that the average (across bootstrap simulations) estimate $\tilde{\beta}$ may not always equal the point estimate $\hat{\beta}$. If we still want to use the bootstraps to find critical values, then we have to center the test statistic on the the average estimate in the bootstraps, $\tilde{\beta}$. For instance, for a t-test we calculate

$$t = \frac{\tilde{\beta} - \text{average } \tilde{\beta}}{\text{Std}(\tilde{\beta})} \tag{6.9}$$

for each simulation and then take the $(0.025N)$th and $(0.975N)$th observations as the 5% critical values (instead of the ± 1.96 from a standard normal distribution). These critical values can then be used for t-tests based on the original sample, for instance, that $\beta_2 = k$. In most cases, there is little difference between centering on the average $\tilde{\beta}$ and the point estimate $\hat{\beta}$.

A similar reasoning applies to joint tests of coefficients. Consider a linear combination of the coefficients, $R\tilde{\beta}$. If V is the OLS variance-covariance matrix, then for each sample we would calculate the quadratic form

$$\Xi = [R(\tilde{\beta} - \text{average } \tilde{\beta})]'(RVR')^{-1}[R(\tilde{\beta} - \text{average } \tilde{\beta})]. \tag{6.10}$$

Once again, we could take the $(0.95N)$th simulated τ as the 5% critical value (instead of the 95th percentiles for a χ^2_q distribution, for instance, 5.99 for $q = 2$). This critical value can be used to hypotheses like $R\beta - k$ based on the original sample.

In general, the bootstraps of test statistics like the t and Ξ are more precise than the bootstraps of the regression coefficients themselves—provided that we use consistent estimates of the covariance matrix. (In the limit, these statistics do not depend on model parameters—they asymptotically "pivotal"—which often improves the convergence rate.) For instance, in the case of autocorrelated residuals, this suggests that it might be better to

create a bootstrap simulation for t-stats calculated with a Newey-West covariance matrix than a "t-stat" based on a standard OLS covariance matrix since the latter will have an asymptotic distribution which depends on the autocorrelation (that is, model parameters).

6.3.2 Bootstrapping when x_t Includes Lags of y_t

When x_t contains lagged values of y_t, then we have to modify the approach in (6.8) since $\tilde{u}_t$ can become correlated with x_t. For instance, if x_t includes y_{t-1} and we happen to sample $\tilde{u}_t = \hat{u}_{t-1}$, then we get a non-zero correlation between regressor and residual. The easiest way to handle this is as in the Monte Carlo simulations in (6.4), but where $\tilde{u}_t$ are drawn (with replacement) from the sample of fitted residuals. The same carries over to the *VAR model* in Remark 6.5.

6.3.3 Bootstrapping when Errors Are Heteroskedastic

Suppose now that the errors are heteroskedastic, but serially uncorrelated. If the heteroskedasticity is unrelated to the regressors, then we can still use (6.8).

However, if the heteroskedasticity is related to the regressors, then it would be wrong to pair x_t with just any $\tilde{u}_t = \hat{u}_s$ since that destroys the relation between x_t and the variance of the residual. (This is the case that White's test for heteroskedasticity tries to identify.)

An alternative way of bootstrapping can then be used: generate the artificial sample by drawing (with replacement) *pairs* (y_s, x_s), that is, we let the artificial pair for observation t be $(\tilde{y}_t, \tilde{x}_t) = (y_s, x_s)$ for some random draw of s. Since $(y_s, x_s) = (x_s'\hat{\beta} + \hat{u}_s, x_s)$ we are effectively pairing the fitted residual $\hat{u}_s$ with the contemporaneous regressors x_s. This is called a *paired bootstrap* (or "case resampling"). Notice that we are sampling with replacement—otherwise the approach of drawing pairs would be to just re-create the original data set. This approach works also when y_t is a vector of dependent variables.

Example 6.8 *With $T = 3$, the artificial sample could be*

$$\begin{bmatrix} (\tilde{y}_1, \tilde{x}_1) \\ (\tilde{y}_2, \tilde{x}_2) \\ (\tilde{y}_3, \tilde{x}_3) \end{bmatrix} = \begin{bmatrix} (y_2, x_2) \\ (y_3, x_3) \\ (y_3, x_3) \end{bmatrix} = \begin{bmatrix} (x_2'\beta + \hat{u}_2, x_2) \\ (x_3'\beta + \hat{u}_3, x_3) \\ (x_3'\beta + \hat{u}_3, x_3) \end{bmatrix}$$

It could be argued (see, for instance, Davidson and MacKinnon (1993)) that bootstrapping the pairs (y_s, x_s) makes little sense when x_s contains lags of y_s, since the random sampling of the pair (y_s, x_s) destroys the autocorrelation pattern of the regressors.

See Table 6.4 for an application.

	$\gamma = 0$		$\gamma = 1$	
α :	0	1	0	1
Simulated	7.1	18.9	13.4	25.1
OLS formula	7.1	13.3	13.4	19.2
White's	7.0	18.5	13.3	24.3
Bootstrap	7.1	18.5	13.4	24.4
Bootstrap 2	7.0	18.5	13.3	24.3
FGLS	7.5	17.3	14.0	24.1

Table 6.4: Standard error of OLS slope (%) under heteroskedasticity (simulation evidence). Model: $y_t = 1+0.9x_t+\epsilon_t$, where $\epsilon_t \sim N(0, \sigma_t^2)$, with $\sigma_t^2 = (1+\gamma|z_t|+\alpha|x_t|)^2$, where z_t is iid N(0,1) and independent of x_t. Sample length: 200. Number of simulations: 25000. The bootstrap draws pairs (y_s, x_s) with replacement while bootstrap 2 is a wild bootstrap.

Remark 6.9 *(The wild Bootstrap) The wild bootstrap is also aimed at solving the heteroskedasticity problem. In this case, the artificial sample is generated as in (6.8), but we use* $\tilde{u}_t = \hat{u}_t\tilde{\varepsilon}_t$ *where* $\hat{u}_t$ *is the fitted (OLS) residual for observation* t *and* $\tilde{\varepsilon}_t$ *is drawn from an iid random variable with mean 0 and variance 1. For instance,* $\tilde{\varepsilon}_t$ *could have a two-point distribution where it is either* -1 *or* 1 *with equal probabilities.*

6.3.4 Bootstrapping when Errors Are Autocorrelated

It is quite hard to handle the case when the errors are serially dependent, since we must sample in such a way that we do not destroy the autocorrelation structure of the data. A common approach is to fit a model for the residuals, for instance, an AR(1), and then bootstrap the (hopefully iid) innovations to that process.

Another approach amounts to *resampling blocks* of data. For instance, suppose the sample has 10 observations, and we decide to create blocks of 3 observations. The first block is $(\hat{u}_1, \hat{u}_2, \hat{u}_3)$, the second block is $(\hat{u}_2, \hat{u}_3, \hat{u}_4)$, and so forth until the last block, $(\hat{u}_8, \hat{u}_9, \hat{u}_{10})$. If we need a sample of length 3τ, say, then we simply draw τ of those 3-observations blocks randomly (with replacement) and stack them to form a longer series.

Example 6.10 *With* $T = 9$ *and a block size of* 3*, the artificial sample could be*

$$\underbrace{\hat{u}_2, \hat{u}_3, \hat{u}_4}_{\text{block 2}}, \underbrace{\hat{u}_7, \hat{u}_8, \hat{u}_9}_{\text{block 7}}, \underbrace{\hat{u}_4, \hat{u}_5, \hat{u}_6}_{\text{block 4}}.$$

To handle end point effects (so that all data points have the same probability to be drawn), we also create blocks by "wrapping" the data around a circle. In practice, this means that we add the following blocks: $(\hat{u}_{10}, \hat{u}_1, \hat{u}_2)$ and $(\hat{u}_9, \hat{u}_{10}, \hat{u}_1)$.

The length of the blocks should clearly depend on the degree of autocorrelation, but $T^{1/3}$ is sometimes recommended as a rough guide. An alternative approach is to have non-overlapping blocks. See Berkowitz and Kilian (2000) for some other approaches.

See Table 6.5 for an illustration.

ρ :	0.0	0.75
Simulated	5.8	10.2
OLS formula	5.8	7.2
Newey-West	5.7	9.6
VARHAC	5.7	11.1
Bootstrapped	5.5	9.5
FGLS	5.9	10.2

Table 6.5: Standard error of OLS intercept (%) under autocorrelation (simulation evidence). Model: $y_t = 1 + 0.9x_t + \epsilon_t$, where $\epsilon_t = \xi_t + \rho\xi_{t-1}, \xi_t$ is iid N(). NW uses 5 lags. VARHAC uses 5 lags and a VAR(1). The bootstrap uses blocks of size 20. Sample length: 300. Number of simulations: 25000.

6.3.5 Other Approaches

There are many other ways to do bootstrapping. For instance, we could sample the regressors and residuals independently of each other and construct an artificial sample of the dependent variable $\tilde{y}_t = \tilde{x}_t'\hat{\beta} + \tilde{u}_t$. This clearly makes sense if the residuals and regressors are independent of each other and errors are iid. In that case, the advantage of this approach is that we do not keep the regressors fixed.

Chapter 7

A System of OLS Regressions

Reference: Wooldridge (2010) 7.3; Greene (2018) 10
More advanced material is denoted by a star (*). It is not required reading.

7.1 A System of Two OLS Regressions

Consider regressions for two different dependent variables (y_{1t} and y_{2t}, for instance, the returns on two different assets) on the same set of regressors (x_t)

$$y_{1t} = x_t'\beta_1 + u_{1t} \tag{7.1}$$
$$y_{2t} = x_t'\beta_2 + u_{2t}, \tag{7.2}$$

where β_1 is the vector of regression coefficients for y_{1t} and β_2 for y_{2t}.

It is straightforward to show (see below) that if the residuals are iid and independent of all regressors but we allow for $\text{Cov}(u_{1t}, u_{2t}) \neq 0$, then

$$\text{Var}\left(\begin{bmatrix}\hat{\beta}_1\\ \hat{\beta}_2\end{bmatrix}\right) = \begin{bmatrix}\sigma_{11}S_{xx}^{-1} & \sigma_{12}S_{xx}^{-1}\\ \sigma_{21}S_{xx}^{-1} & \sigma_{22}S_{xx}^{-1}\end{bmatrix}, \tag{7.3}$$

where $\sigma_{ij} = \text{Cov}(u_{it}, u_{jt})$ and where $S_{xx} = \Sigma_{t=1}^{T} x_t x_t'$.

More generally, when the residuals are heteroskedastic or autocorrelated,

$$\text{Var}\left(\begin{bmatrix}\hat{\beta}_1\\ \hat{\beta}_2\end{bmatrix}\right) = \begin{bmatrix}S_{xx}^{-1} & \mathbf{0}\\ \mathbf{0} & S_{xx}^{-1}\end{bmatrix} \Omega \begin{bmatrix}S_{xx}^{-1} & \mathbf{0}\\ \mathbf{0} & S_{xx}^{-1}\end{bmatrix}, \text{ where} \tag{7.4}$$

$$\Omega = \text{Var}\left(\sum_{t=1}^{T}\begin{bmatrix} x_t u_{1t} \\ x_t u_{2t} \end{bmatrix}\right)$$
$$= \begin{bmatrix} \text{Var}(\Sigma x_t u_{1t}) & \text{Cov}(\Sigma x_t u_{1t}, \Sigma x_t u_{2t}) \\ \text{Cov}(\Sigma x_t u_{2t}, \Sigma x_t u_{1t}) & \text{Var}(\Sigma x_t u_{2t}) \end{bmatrix}. \tag{7.5}$$

The Ω matrix (which is $2k \times 2k$ if there are k regressors in x_t) could be estimated with the methods of White or Newey-West.

Extensions to more than two regression equations are straightforward.

Proof. (of (7.3)–(7.5)) Similarly to the single-equation OLS, we can write

$$\hat{\beta}_1 = \beta_1 + S_{xx}^{-1}\sum_{t=1}^{T} x_t u_{1t}$$
$$\hat{\beta}_2 = \beta_2 + S_{xx}^{-1}\sum_{t=1}^{T} x_t u_{2t}$$

The variance (matrix) of $\hat{\beta}_1$ or of $\hat{\beta}_2$ follows the same pattern as for single-equation OLS. In contrast, the covariance (matrix) is

$$\text{Cov}(\hat{\beta}_1, \hat{\beta}_1) = S_{xx}^{-1}\,\text{Cov}(\textstyle\sum_{t=1}^{T} x_t u_{1t}, \sum_{t=1}^{T} x_t u_{2t}) S_{xx}^{-1}.$$

Together, this gives (8.10)–(7.5). If the residuals are iid and independent of the regressors, then Ω simplifies to

$$\Omega = \begin{bmatrix} \sigma_{11} S_{xx} & \sigma_{12} S_{xx} \\ \sigma_{21} S_{xx} & \sigma_{22} S_{xx} \end{bmatrix},$$

which gives (7.3). ■

7.2 A System of n OLS Regressions

Remark 7.1 *(Kronecker product) Let $\otimes$ denote the Kronecker product, that is, if A and B are matrices, then*

$$A \otimes B = \begin{bmatrix} a_{11}B & \cdots & a_{1n}B \\ \vdots & & \vdots \\ a_{m1}B & \cdots & a_{mn}B \end{bmatrix}.$$

For instance, with

$$A = \begin{bmatrix} 1 & 3 \\ 2 & 4 \end{bmatrix} \text{ and } B = \begin{bmatrix} 10 & 11 \end{bmatrix}, \text{ we get } A \otimes B = \begin{bmatrix} 10 & 11 & 30 & 33 \\ 20 & 22 & 40 & 44 \end{bmatrix}.$$

Let $\hat{\theta}$ be the vector of where $\hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_n$ are stacked (equation 1 first, then equation

2, etc).

With iid residuals the variance-covariance matrix (instead of (7.3)) is

$$\text{Var}(\hat{\theta}) = \Sigma \otimes S_{xx}^{-1}, \tag{7.6}$$

where $\Sigma = \text{Cov}(u_t)$ is the $n \times n$ variance-covariance matrix of the n residuals.

Similarly, with non-iid residuals we get (instead of (8.10))

$$\text{Var}(\hat{\theta}) = (I_n \otimes S_{xx}^{-1})\Omega(I_n \otimes S_{xx}^{-1}), \tag{7.7}$$

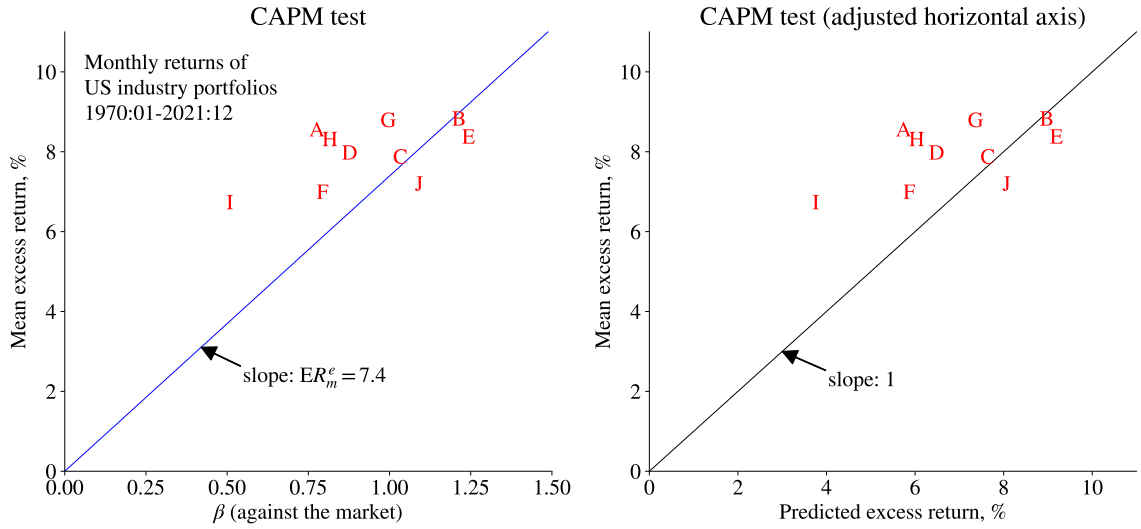
where I_n is the $n \times n$ identity matrix. Let u_t be the vector of the n residuals in period t $(u_{1t}, u_{2t}, ..., u_{nt})$. Then, Ω is (instead of (7.5))

$$\Omega = \text{Var}\left(\sum_{t=1}^{T} u_t \otimes x_t\right), \tag{7.8}$$

which is an $nk \times nk$ matrix which can be estimated by the methods of White or Newey-West.

Remark 7.2 *(Estimating Ω) To estimate Ω in (7.8), create an $T \times nk$ matrix. Let first k columns of row t be $u_{1t}x_t'$, the second k columns be $u_{2t}x_t'$. Then apply, for instance, Newey-West to these nk time series.*

Empirical Example 7.3 *(CAPM on industry portfolios) Figure 7.1 shows results for the intercepts from regressing US industry portfolios on the market. The joint test is for whether all intercepts are zero.*



	α	t-stat	σ
A (NoDur)	2.77	2.31	8.58
B (Durbl)	-0.19	-0.09	15.13
C (Manuf)	0.19	0.22	6.28
D (Enrgy)	1.44	0.64	16.20
E (HiTec)	-0.86	-0.55	11.29
F (Telcm)	1.08	0.72	10.65
G (Shops)	1.38	1.09	9.07
H (Hlth)	2.23	1.46	10.95
I (Utils)	2.93	1.79	11.67
J (Other)	-0.92	-0.96	6.87

CAPM: $R_i^e = \alpha_i + \beta_i R_m^e + e_i$
Predicted excess return: $\beta_i ER_m^e$

α and σ (std of residual) are in annualized %

p-val for testing if all $\alpha_i = 0$: 0.34
10% crit. value (Bonferroni): 2.58

Figure 7.1: CAPM regressions on US industry indices

Chapter 8

A System of Regressions Equations

8.1 A System of OLS Regressions

Reference: Wooldridge (2010) 7.3

Consider the two regressions

$$y_t = x_t'\beta + u_t \tag{8.1}$$

$$z_t = w_t'\gamma + v_t. \tag{8.2}$$

Let $\hat{\Sigma}_{xx} = \sum_{t=1}^{T} x_t x_t'/T$ and similarly for the other second moment matrices. We then know (from basic properties of LS) that

$$\sqrt{T}(\hat{\beta} - \beta) = \hat{\Sigma}_{xx}^{-1}\sqrt{T}\frac{1}{T}\sum_{t=1}^{T} x_t u_t \tag{8.3}$$

$$\sqrt{T}(\hat{\gamma} - \gamma) = \hat{\Sigma}_{ww}^{-1}\sqrt{T}\frac{1}{T}\sum_{t=1}^{T} w_t v_t. \tag{8.4}$$

Let $\Sigma_{ii} = \text{plim}\,\hat{\Sigma}_{ii}$. The remaining terms (typically) obey CLTs, so we can expect the asymptotic distribution to be normal.

The covariance of $\sqrt{T}\hat{\beta}$ and $\sqrt{T}\hat{\gamma}$ is therefore

$$\text{Cov}(\sqrt{T}\hat{\beta}, \sqrt{T}\hat{\gamma}) = \Sigma_{xx}^{-1}\Sigma^{yz}\Sigma_{ww}^{-1}, \text{ with} \tag{8.5}$$

$$\Sigma^{yz} = \text{Cov}\left(\sqrt{T}\frac{1}{T}\sum_{t=1}^{T} x_t u_t, \sqrt{T}\frac{1}{T}\sum_{t=1}^{T} w_t v_t\right) \tag{8.6}$$

where we use the fact that since Σ_{ww}^{-1} is symmetric. Warning: Σ^{yz} is just notation for the covariance matrix of the scaled samples averages of $x_t u_t$ and $w_t v_t$.

The variance-covariance matrices of $\sqrt{T}\hat{\beta}$ and $\sqrt{T}\hat{\gamma}$ are as in the usual OLS setting

$$\text{Cov}(\sqrt{T}\hat{\beta}) = \Sigma_{xx}^{-1}\Sigma^{yy}\Sigma_{xx}^{-1} \text{ and } \text{Cov}(\sqrt{T}\hat{\gamma}) = \Sigma_{ww}^{-1}\Sigma^{zz}\Sigma_{ww}^{-1} \text{ with} \tag{8.7}$$

$$\Sigma^{yy} = \text{Cov}\left(\sqrt{T}\frac{1}{T}\textstyle\sum_{t=1}^{T} x_t u_t\right) \text{ and} \tag{8.8}$$

$$\Sigma^{zz} = \text{Cov}\left(\sqrt{T}\frac{1}{T}\textstyle\sum_{t=1}^{T} w_t v_t\right). \tag{8.9}$$

We can write the full variance-covariance matrix

$$\text{Cov}\left(\sqrt{T}\begin{bmatrix}\hat{\beta}\\ \hat{\gamma}\end{bmatrix}\right) = \begin{bmatrix}\Sigma_{xx}^{-1} & \mathbf{0}\\ \mathbf{0} & \Sigma_{ww}^{-1}\end{bmatrix}\underbrace{\begin{bmatrix}\Sigma^{yy} & \Sigma^{yz}\\ \Sigma^{zy} & \Sigma^{zz}\end{bmatrix}}_{\Sigma \text{ from (8.11)}}\begin{bmatrix}\Sigma_{xx}^{-1} & \mathbf{0}\\ \mathbf{0} & \Sigma_{ww}^{-1}\end{bmatrix}. \tag{8.10}$$

Example 8.1 *(Dimensions) If x_t has 3 elements and w_t has 2 elements, then the dimensions for each of the matrices in (8.10) are*

$$\begin{bmatrix}3\times 3 & 3\times 2\\ 2\times 3 & 2\times 2\end{bmatrix}, \text{ which is } 5\times 5.$$

The middle matrix (on the right hand side) of (8.10) is really the full variance-covariance matrix of the vector where we stack $\sqrt{T}\frac{1}{T}\Sigma x_t u_t$ and $\sqrt{T}\frac{1}{T}\Sigma w_t v_t$, which I denote by Σ

$$\Sigma = \text{Cov}\left(\sqrt{T}\frac{1}{T}\textstyle\sum_{t=1}^{T}\begin{bmatrix}x_t u_t\\ w_t v_t\end{bmatrix}\right). \tag{8.11}$$

This could, for instance, be estimated with the methods of White or Newey-West.

Extensions to more than two regression equations are straightforward: the general patterns of (8.10) and (8.11) are the same.

Remark 8.2 *(iid residuals) If the residuals are iid and independent of all regressors (also across observations), then Σ simplifies to*

$$\begin{bmatrix}\Sigma^{yy} & \Sigma^{yz}\\ \Sigma^{zy} & \Sigma^{zz}\end{bmatrix} = \begin{bmatrix}\sigma_{uu}\Sigma_{xx} & \sigma_{uw}\Sigma_{xw}\\ \sigma_{uw}\Sigma_{wx} & \sigma_{ww}\Sigma_{ww}\end{bmatrix}.$$

Remark 8.3 *(Kronecker product) If A and B are matrices, then*

$$A\otimes B = \begin{bmatrix}a_{11}B & \cdots & a_{1n}B\\ \vdots & & \vdots\\ a_{m1}B & \cdots & a_{mn}B\end{bmatrix}.$$

Remark 8.4 *(SURE) When both regression (8.1) and (8.2) have the same regressors, so* $w_t = x_t$*, then (8.10) and (8.11) simplify to*

$$\text{Cov}\left(\sqrt{T}\begin{bmatrix}\hat{\beta}\\ \hat{\gamma}\end{bmatrix}\right) = \begin{bmatrix}\Sigma_{xx}^{-1} & \mathbf{0}\\ \mathbf{0} & \Sigma_{xx}^{-1}\end{bmatrix}\Sigma\begin{bmatrix}\Sigma_{xx}^{-1} & \mathbf{0}\\ \mathbf{0} & \Sigma_{xx}^{-1}\end{bmatrix} \text{ and } \Sigma = \text{Cov}\left(\sqrt{T}\frac{1}{T}\sum_{t=1}^{T}\begin{bmatrix}u_t\\ v_t\end{bmatrix}\otimes x_t\right),$$

where $\otimes$ *is the Kronecker product. If, in addition, the residuals are iid and independent of* x *as in Remark 8.2, then this simplifies further to*

$$\text{Cov}\left(\sqrt{T}\begin{bmatrix}\hat{\beta}\\ \hat{\gamma}\end{bmatrix}\right) = \begin{bmatrix}\sigma_{uu} & \sigma_{uv}\\ \sigma_{vu} & \sigma_{vv}\end{bmatrix}\otimes\Sigma_{xx}^{-1}.$$

8.2 Applications

8.2.1 CAPM with Several Test Assets

Suppose we have n test assets. Stack the CAPM regressions $i = 1,\ldots,n$ as

$$\begin{bmatrix}R_{1t}^e\\ \vdots\\ R_{nt}^e\end{bmatrix} = \begin{bmatrix}\alpha_1\\ \vdots\\ \alpha_n\end{bmatrix} + \begin{bmatrix}\beta_1\\ \vdots\\ \beta_n\end{bmatrix} f_t + \begin{bmatrix}\varepsilon_{1t}\\ \vdots\\ \varepsilon_{nt}\end{bmatrix}, \text{ where} \tag{8.12}$$
$$\text{E}\,\varepsilon_{it} = 0 \text{ and } \text{Cov}(f_t,\varepsilon_{it}) = 0.$$

This is a system of seemingly unrelated regression equations (SURE)—with the same regressor (see, for instance, Greene (2003) 14). In this case, the efficient estimator (GLS) is LS on each equation separately.

We, we could test, for instance, if $\alpha = \mathbf{0}$ by a Wald test. (In case residuals are iid and independent of the regressors, the covariance matrix of the α vector can be shown to have a particularly simple form.)

8.2.2 A Multifactor Model with Several Test Assets

Reference: Cochrane (2005) 12.1; Campbell, Lo, and MacKinlay (1997) 6.2.1

When the K factors, f_t, are excess returns, the null hypothesis typically says that $\alpha_i = 0$ in

$$R_{it}^e = \alpha_i + \beta_i' f_t + \varepsilon_{it}, \text{ where} \tag{8.13}$$
$$\text{E}\,\varepsilon_{it} = 0 \text{ and } \text{Cov}(f_t,\varepsilon_{it}) = \mathbf{0}_{K\times 1},$$

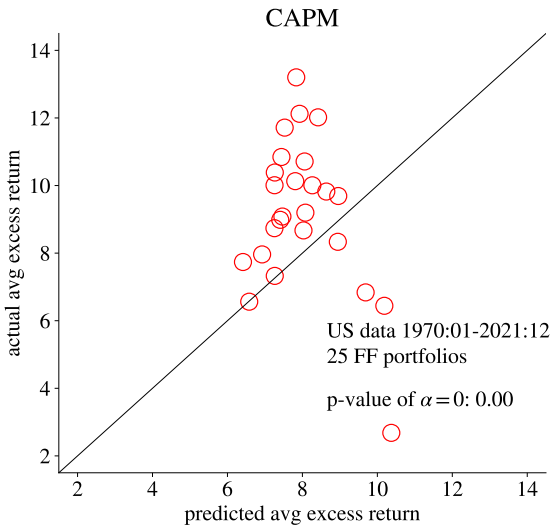


Figure 8.1: CAPM, FF portfolios

and β_i is now an $K \times 1$ vector. We stack the returns for n assets to get

$$\begin{bmatrix} R_{1t}^e \\ \vdots \\ R_{nt}^e \end{bmatrix} = \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{bmatrix} + \begin{bmatrix} \beta_{11} & \dots & \beta_{1K} \\ \vdots & \ddots & \vdots \\ \beta_{n1} & \dots & \beta_{nK} \end{bmatrix} \begin{bmatrix} f_{1t} \\ \vdots \\ f_{Kt} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1t} \\ \vdots \\ \varepsilon_{nt} \end{bmatrix}$$

or in vector form

$$\begin{aligned} R_t^e &= \alpha + \beta f_t + \varepsilon_t, \text{where} \\ \mathrm{E}\,\varepsilon_t &= \mathbf{0}_{n\times 1} \text{ and } \mathrm{Cov}(f_t, \varepsilon_t') = \mathbf{0}_{K\times n}, \end{aligned} \tag{8.14}$$

where α is $n \times 1$ and β is $n \times K$. Notice that β_{ij} shows how the ith asset depends on the jth factor.

Again, we could test, for instance, if $\alpha = \mathbf{0}$.

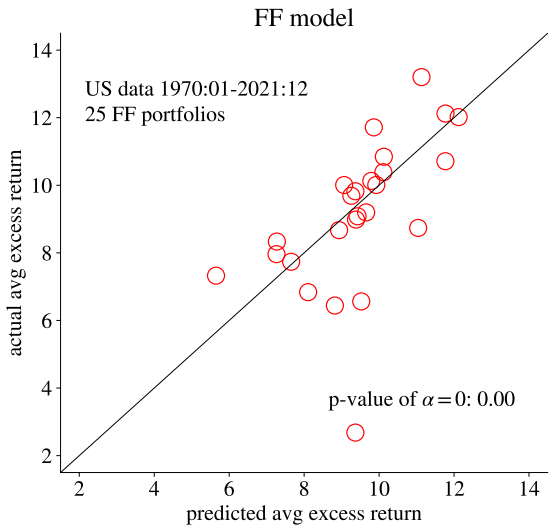


Figure 8.2: FF, FF portfolios

Chapter 9

Portfolio Sorts

9.1 Overview

Reference: Bali, Engle, and Murray (2016)

Portfolio sorts are used to construct portfolios (groups) based on some characteristic, for instance, firm size. Once the portfolios have been defined (to which group/portfolio does i belong?) we often compute the (possible weighted) average with each (group or) portfolio

$$R_{gt} = \sum_{i \in \text{Group}g} w_{it} R_{it}, \tag{9.1}$$

where w_{it} is the relative portfolio weight of asset i in the portfolio ($\Sigma w_{it} = 1$). We often use an unweighted average where $w_{it} = 1/(\text{number of members of the group})$.

A common way (since Jensen, updated in Huberman and Kandel (1987)) is to study the performance of a portfolio by running the following regression

$$\begin{aligned} R^e_{gt} &= \alpha + \beta R^e_{mt} + \varepsilon_t, \text{ with} \\ \mathrm{E}\,\varepsilon_t &= 0 \text{ and } \mathrm{Cov}(R^e_{mt}, \varepsilon_t) = 0, \end{aligned} \tag{9.2}$$

where R^e_{gt} is the excess return on the portfolio being studied and R^e_{mt} the excess returns of a vector of benchmark portfolios (for instance, only the market portfolio if we want to rely on CAPM. Neutral performance requires $\alpha = 0$, which can be tested with a t test.

9.2 Univariate Sorts

A simple and commonly applied method for studying how an asset characteristic (x_i) is related to returns (or some other performance measure) is to do a *univariate sort*. For instance, we could sort the assets $i = 1, ..., n$ according to x_i and then construct three

portfolios: (1) for those i whose x_i belong to the lowest 1/3; (2) those in the mid 1/3 and (3) those in the highest 1/3. Then, we measure the return on equally weighted portfolios—and perhaps analyse the return of portfolio 3 minus the return on portfolio 1. The sorting and portfolio construction is typically repeated at regular intervals. For instance, the Fama-French size portfolios are based on the market capitalization and are rebalanced every June. For a daily momentum strategy, we would rather redo the sort every day based on recent performance.

Empirical Example 9.1 *(Sorting on recent returns) See Table 9.1 for an empirical example where the 25 FF portfolios are sorted into low/low recent 22-day returns, with 5 portfolios in each. The results indicate strong momentum.*

Low 22-day return	4.31
	(0.21)
High 22-day return	14.30
	(0.79)
Difference	9.99
	(0.92)

Table 9.1: Average excess returns and (Sharpe ratios) for 3 portfolios from a univariate sort on recent (22-day) returns (5/5 assets). Daily data on 25 FF portfolios 1979:01-2021:12

Example 9.2 *(Simplified version of "Betting against beta" by Frazzini and Pedersen*) First, find those assets with* $\beta_{i,t-1} < \text{median}(\beta_{i,t-1})$ *where* $\beta_{i,t-1}$ *is the CAPM beta estimated on data up to and including* $t-1$*. (The median is across the assets.) This is the low beta group. Second, calculate the equally weighted portfolio return in* t*. Third, repeat for all periods. Fourth, do points 1–3 also for the high beta assets,* $\beta_{i,t-1} \geq \text{median}(\beta_{i,t-1})$*. Fifth, form the excess return as the difference between the two portfolios.*

9.3 Bivariate Sorts

Bivariate sorts (also called double sorts) are used when there are two important characteristics (here called x and z) and you want to study how z affects returns—controlling for x (that is, holding x "constant"). This may well be important if x and z are correlated.

Bivariate sorts can be done in several ways. An *independent bivariate sort* first does a univariate sort of x_i (say, forming 3 categories: growth, neutral or value), then it makes another univariate sort according to another sorting variable z_i (say, forming two categories: small or big). Then we find the intersections of the two sorts—think of a matrix

$$\begin{array}{lccc} & \underline{\text{Low } x_i} & \underline{\text{Medium } x_i} & \underline{\text{High } x_i} \\ \text{Low } z_i\text{:} & (x_L, z_L) & (x_M, z_L) & (x_H, z_L) \\ \text{High } z_i\text{:} & (x_L, z_H) & (x_M, z_H) & (x_H, z_H) \end{array} \tag{9.3}$$

where, for instance, (x_M, z_H) denote the set of assets that belong to the medium x category and high z category. Notice that this matrix has a different structure than a traditional scatter plot with x on the horizontal axis and z on the vertical axis: in this (and all other tables) we put low z_i on the first line and high z_i on the second.

In an independent bivariate sort we cannot directly control how many assets there will be in each group—and some groups might be empty (see Example 9.3 and Figure 9.1).

Once the portfolio sort is done, we typically calculate the average return (or some other variable of interest) of each portfolio. In the independent sort, you can either compare across rows or across columns.

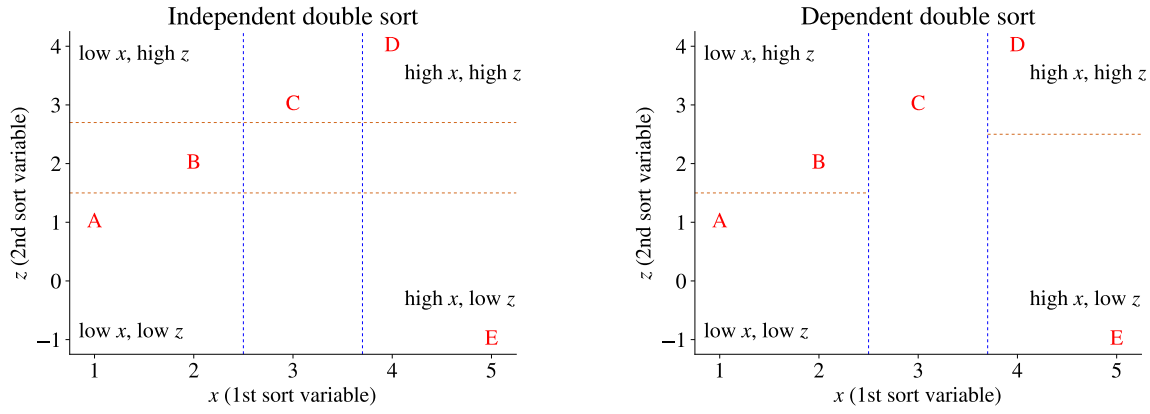


Figure 9.1: Example of bivariate sorts. The data is indicated by letters.

Example 9.3 *(Independent double sort) Suppose there are 5 assets (labelled A, B,...) and*

that the values of x and z are

	x_i	z_i
Asset A	1	1
Asset B	2	2
Asset C	3	3
Asset D	4	4
Asset E	5	−1

We form low/high groups with 2 elements in each

	Assets
Low x:	*A, B*
High x:	*D, E*
Low z:	*E, A*
High z:	*C, D*

The independent double sort then gives

	Low x	*High x*
Low z:	*A*	*E*
High z:		*D*

Notice that there are no assets in the (low x, high z) group. See also Figure 9.1 for an illustration, but notice that the scatter plot has a different structure: low z values are plotted below high z values.

Empirical Example 9.4 *(Independent sorting on recent volatility and returns) We first sort the 25 FF portfolios according to recent volatility, putting 10 into the low group and 10 into the high group. Then we sort on recent returns, also putting 10 into a low group and 10 into a high group. Finally, we form intersections. Figure 9.2 illustrates (for a short subsample) how the number of portfolios in the "low vol, low return" group varies over time. Typically, there are 4–5 portfolios in the group, but it varies considerably over time. The subsequent analysis is therefore focused on the dependent sort (see below).*

In *dependent bivariate sort* we first sort according to x_i as before. Then, *within* an x category we sort according to z_i. This allows us to control the number of assets in each group. Notice that the ordering matters in the dependent sort: letting x represent growth/neutral/value and z small/big will not give the same results as switching the labels. In the dependent sort, we compare across the z categories, that is, *across rows* in

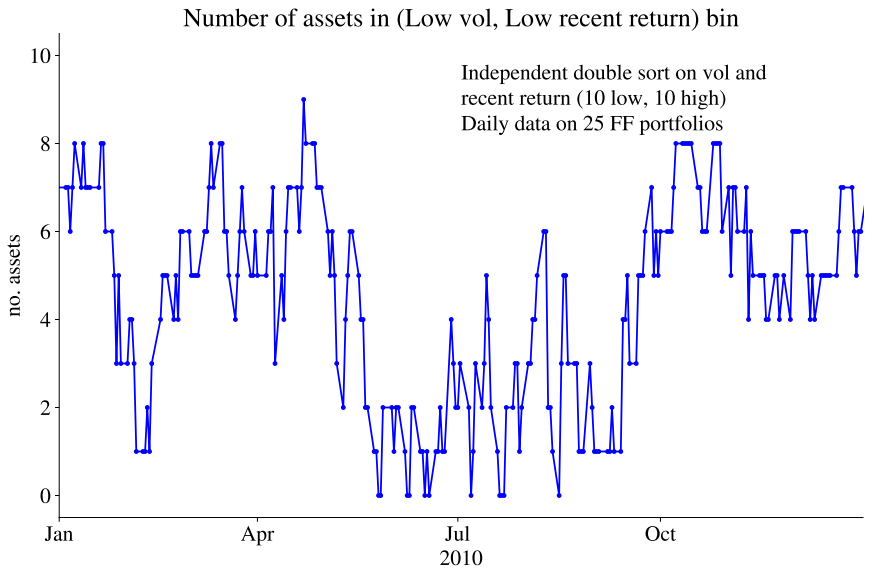


Figure 9.2: Independent bivariate portfolio sort, 25 FF portfolios

(9.3), for instance, the return of (x_L, z_H) minus the return of (x_L, z_L) and so forth. In (9.3) this gives three numbers—which are sometimes averaged: the interpretation is that you are studying the effect of z (here: small/large), but controlling for x (here: one of growth/neutral/value). See Figure 9.1 for an example.

Example 9.5 *(Dependent double sort) Continuing the previous example, the dependent double sort (with one asset in each portfolio) gives*

	Low x	*High x*
Low z:	*A*	*E*
High z:	*B*	*D*

For instance, among the "high x" assets D and E, asset E has a lower z value so it is allocated to the (high x, low z) portfolio, while asset D has a higher z value so it is allocated to the (high x, high z) portfolio. Notice that all portfolios are populated. See also Figure 9.1 for an illustration.

Empirical Example 9.6 *(Dependent sorting on recent volatility and returns) We first sort the 25 FF portfolios according to recent volatility, putting 10 into the low group and 10 into the high group. Within the "low vol" group, we then sort according to recent returns, putting 5 into to the "low vol, low return" group and 5 into the "low vol, high return"*

group. We do the same within the "high volatility" group. This means that there are always 5 FF portfolios in each of the 4 groups. See Figure 9.3 for some (unconditional) background information about the 25 FF portfolios and Figure 9.4 for how often the different FF portfolios are in each of the four groups. Table 9.2 reports the average excess returns and the Sharpe ratios for the four groups.

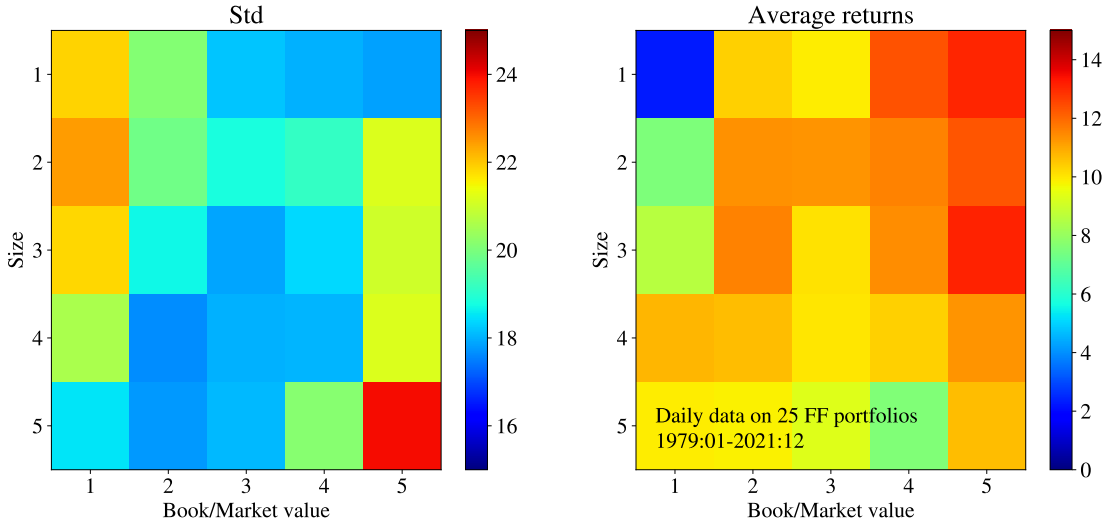


Figure 9.3: Average returns and volatility of the 25 FF portfolios

	Low 22-day vol	High 22-day vol
Low 22-day return	8.81	5.13
	(0.55)	(0.23)
High 22-day return	13.19	12.49
	(0.86)	(0.60)
High - low	4.38	7.36
	(0.87)	(0.94)
High - low, average	5.87	
	(1.12)	

Table 9.2: Average excess returns and (Sharpe ratios) for 4 portfolios from a dependent bivariate sort. The first sort is on volatility (10/10 assets), the second sort (within each volatility bin) is on recent returns (5/5 assets). Daily data on 25 FF portfolios 1979:01-2021:12

The bivariate sort is designed to handle some *correlation* between x and z. If there is no correlation, then a single sort is enough. However, the bivariate sort will break

down if the correlation is too strong. In the independent sort, it can lead to few (or even zero) assets in the off-diagonal portfolios if the correlation is positive (and vice versa if the correlation is negative). In the dependent sort, it may simply leads to results that cannot be trusted, see Figure 9.5 for an example. The figure illustrates how the "high z" portfolios have clearly higher x values than the "low z" portfolios have, so the approach is only moderately successful in controlling for x. This could be solved by having smaller x bins (which may require many assets), so that that the variation in x within each bin is small compared to the variation in z. For instance, the low x bin could contain 20% of the assets, the high x also 20%, leaving out the 60% in the middle. As an alternative, we could consider an orthogonalisation (see below).

Remark 9.7 *(When x and z are perfectly correlated*) If we change Example 9.3 so z equals x, then the independent double sort gives*

	Low x	*High x*
Low z:		*D, E*
High z:	*A, B*	

This has the problem that the off-diagonal portfolios are empty. In contrast, the dependent sort gives

	Low x	*High x*
Low z:	*A*	*D*
High z:	*B*	*E*

The latter has the problem that comparing across rows does not control for x. For instance, asset B has a higher x (and z) value than asset A.

Remark 9.8 *(The Fama-French factors*) The SMB and HML are created by an independent bivariate sort. First, classify firms according to the book/market value: low (growth stocks, using 30th percentile as cutoff), neutral or high (value stocks, using 70th percentile as cutoff). Second, classify firms according to size: small or big, using the median as a cutoff. Create six value weighted portfolios from the intersection of those categories*

	Low book/market	*Medium book/market*	*High book/market*
Small:	*Small Growth (SG)*	*Small Neutral (SN)*	*Small Value (SV)*
Big:	*Big Growth (BG)*	*Big Neutral (BN)*	*Big Value (BV)*

The SMB is the average of the small portfolios minus the average of the big portfolios: $SMB = 1/3(SG + SN + SV) - 1/3(BG + BN + BV)$. Rearranging gives $SMB =$

$1/3(SG-BG)+1/3(SN-BN)+SV)+1/3(SV-BV)$, *which shows that it represents the return on small stocks (for a given book/market) minus the return on big stocks (for same book/market). The HML is the average of the value stocks minus the growth stocks,* $HML = 1/2(SV + BV) - 1/2(SG + BG)$, *which can be rearranged as* $HML = 1/2(SV - SG) + 1/2(BV - BG)$, *which shows that it represents the return on value stocks (for a given size) minus the return on growth stocks (for the same size).*

9.4 Orthogonalisation

Single sort on orthogonalised data is an alternative to a double sort and may be better at handling strong (linear) correlation of x and z. It involves two steps. First, run a regression of (z on x and a constant) to get coefficients (a, b). Second, do a single sort on the residual

$$\varepsilon_i = z_i - (a + bx_i). \tag{9.4}$$

The regression can be done in several different ways: *(1)* a cross-sectional regression as in Figure 9.6; *(2)* time series regressions; *(3)* a panel regression. In cases (1) and (2) there is also a choice between using the full sample or just data up to $t - 1$.

9.5 Trading Strategies

Dynamic trading strategies are similar (and sometimes identical) to portfolio sorts. The basic idea is to create a portfolio based on some kind of sorting of a trading signal.

Empirical Example 9.9 *(Momentum for daily returns on the 25 FF portfolios) Figure 18.9 suggests that there is considerable momentum in the cross-section of the 25 FF portfolios. Investing in past winners earns high returns.*

Empirical Example 9.10 *(Mean reversion of daily S&P 500 returns) Figure 9.8 shows that extreme S&P 500 returns are followed by mean-reverting movements the following day (negative autocorrelation)—which suggests that a trading strategy should sell after a high return and buy after a low return.*

Empirical Example 9.11 *(Mean reversion of daily returns for different size categories) Figure 9.9 compares the results for daily returns on different size categories —and illustrates that there is more predictability (indicating positive autocorrelation) for small stocks.*

Empirical Example 9.12 *(Long run S&P 500 after different p/e values) Figure 9.10 shows average one-year return on S&P 500 for different bins of the p/e ratio (at the beginning of the year). The figure illustrates that buying when the market is undervalued (low p/e) might be a winning strategy.*

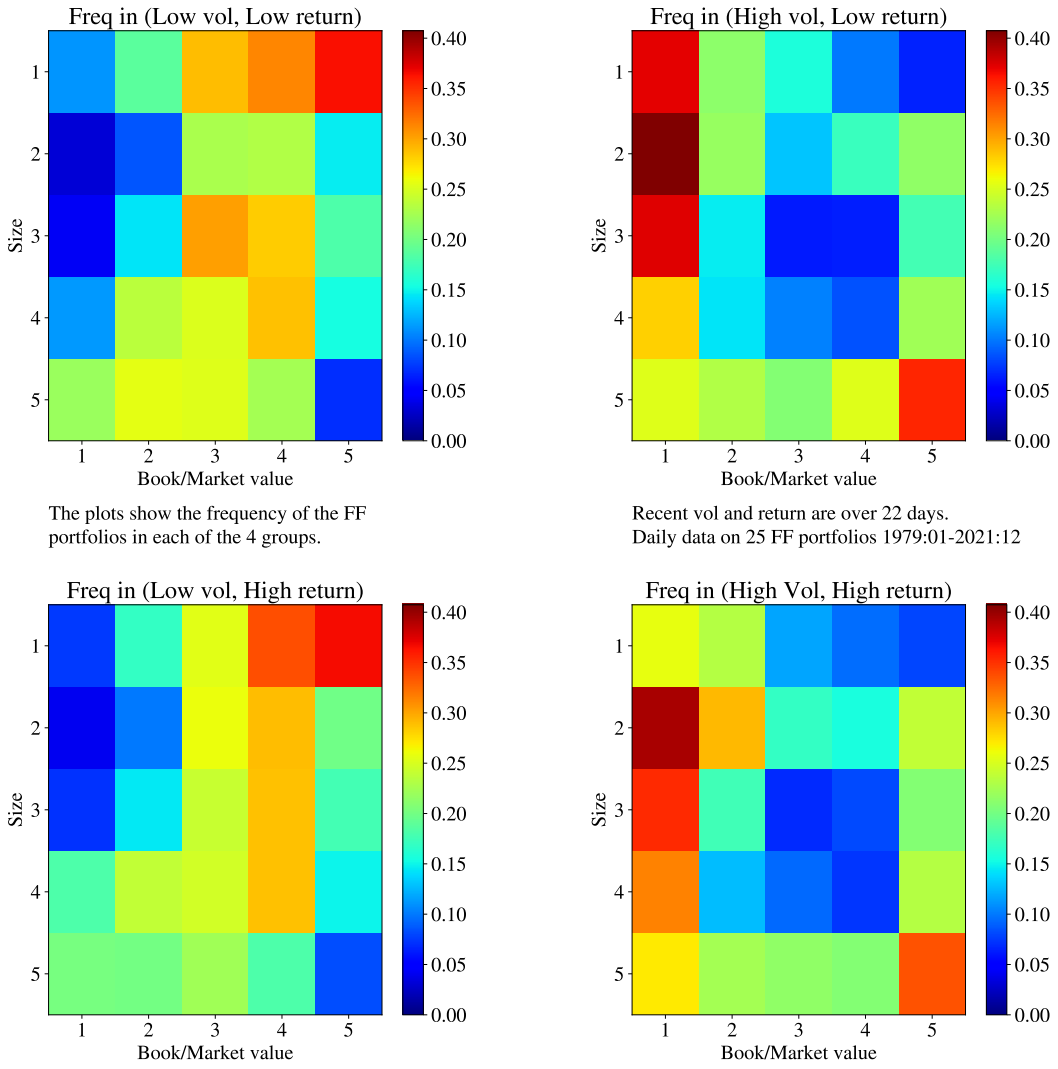


Figure 9.4: Dependent bivariate portfolio sort, 25 FF portfolios

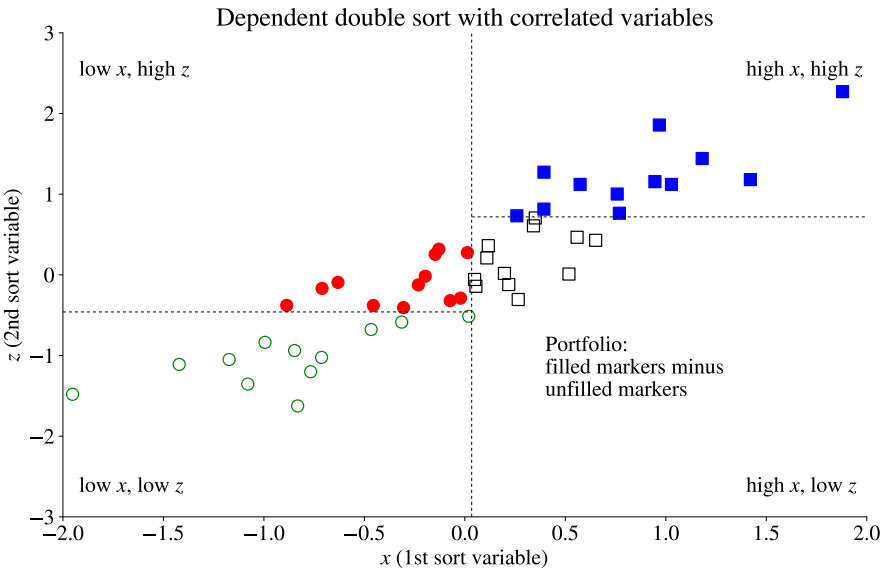


Figure 9.5: Example of dependent bivariate sort with correlated variables

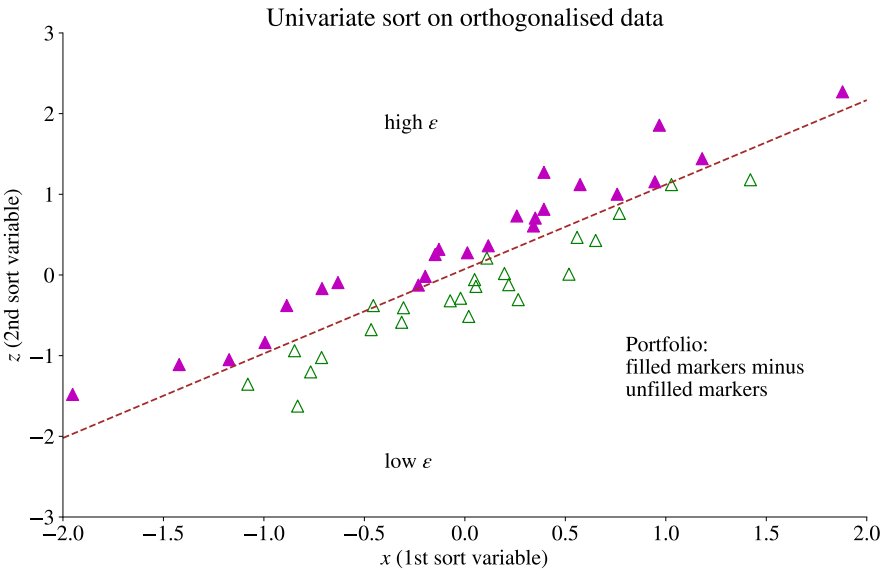


Figure 9.6: Example of univariate sort on orthogonalised data

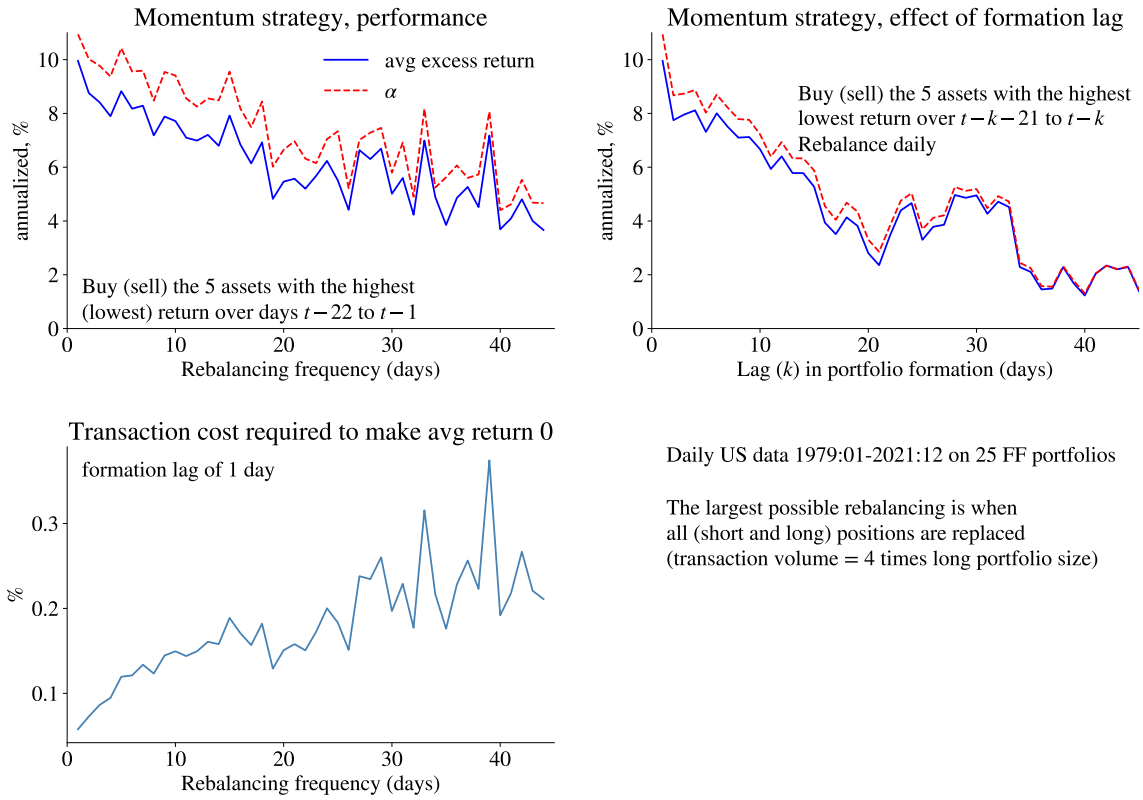


Figure 9.7: Predictability of daily US stock returns, momentum strategy

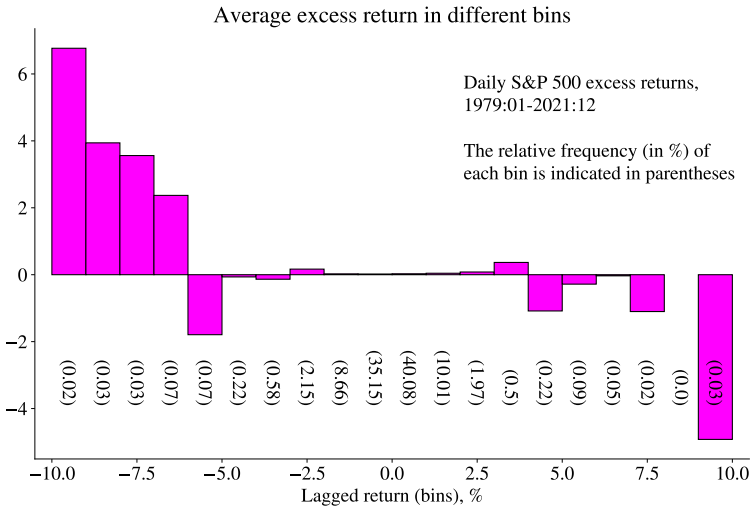


Figure 9.8: Predictability of daily US stock returns, out-of-sample

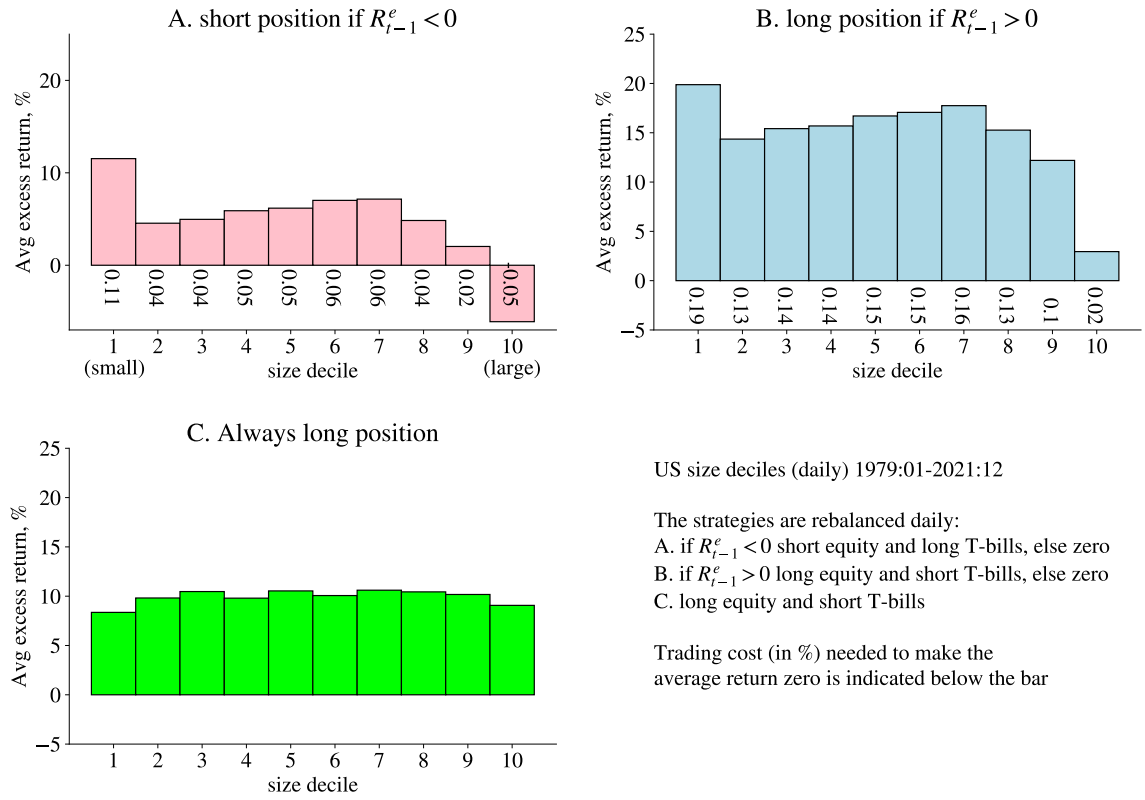


Figure 9.9: Predictability of daily US stock returns, out-of-sample

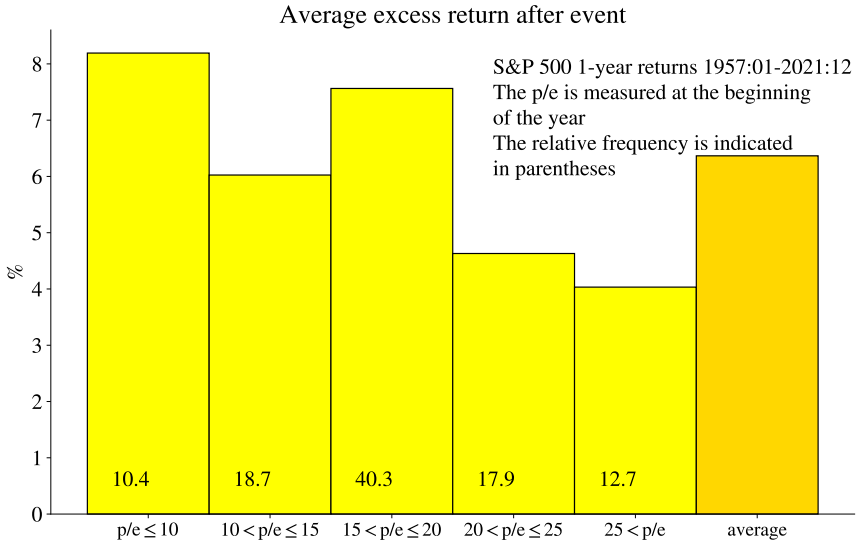


Figure 9.10: Predictability of annual US stock returns, out-of-sample

Chapter 10

GMM*

Sections denoted by a star (*) is not required reading.

Reference: Cochrane (2005) 11 and 14; Campbell (2018) 4; Singleton (2006) 2–4; Greene (2018) 13

10.1 The Basic GMM

In general, the $q \times 1$ vector of sample moment conditions in GMM are written

$$\bar{g}(\beta) = \frac{1}{T}\sum_{t=1}^{T} g_t(\beta) = \mathbf{0}_{q\times 1}, \tag{10.1}$$

where $\bar{g}(\beta)$ is short hand notation for the sample average. The notation $g_t(\beta)$ is meant to show that moments conditions depend on the parameter vector (β) and on data for period t. We let β_0 denote the true value of the $k \times 1$ parameter vector.

The GMM estimator is

$$\hat{\beta}_{k\times 1} = \arg\min \bar{g}(\beta)' W \bar{g}(\beta), \tag{10.2}$$

where W is some symmetric positive definite $q \times q$ weighting matrix. When the model is exactly identified ($q = k$), then we do not have to perform an explicit minimization, since all sample moment conditions can be set equal to zero (there are as many parameters as there are moment conditions). However, we may still have to apply a numerical algorithm to find the $\hat{\beta}$ values that make $\bar{g}(\hat{\beta}) = \mathbf{0}_{q\times 1}$ hold, in particular, if $g_t()$ are non-linear functions.

It can be shown that choosing $W = S_0^{-1}$, where S_0 is the covariance matrix of $\sqrt{T}\bar{g}(\beta_0)$ evaluated at the true parameter values, gives the most efficient estimates (for a given set of moment conditions). To approximate this, an iterative procedure is often

used: start with $W = I_q$ (or some other reasonable weighting matrix), estimate the parameters and use them to create a $T \times q$ matrix of moment conditions, estimate S_0, then (in a second step) use $W = \hat{S}_0^{-1}$ and reestimate. In most cases this iteration is stopped at this stage, but you could also continue iterating until the point estimates converge.

Example 10.1 *(Moment condition for a mean) To estimate the mean of x_t, use*

$$g_t = x_t - \mu.$$

There is one parameter and one moment condition: exactly identified.

Example 10.2 *(Moments conditions for OLS) Consider the linear model $y_t = x_t'\beta_0 + u_t$, where x_t and β are $k \times 1$ vectors. The k moments are*

$$g_t = x_t(y_t - x_t'\beta).$$

There are as many parameters as moment conditions: exactly identified.

Example 10.3 *(Moment conditions for estimating a normal distribution) Suppose you specify four moments for estimating the mean and variance of a normal distribution*

$$g_t = \begin{bmatrix} x_t - \mu \\ (x_t - \mu)^2 - \sigma^2 \\ (x_t - \mu)^3 \\ (x_t - \mu)^4 - 3\sigma^4 \end{bmatrix}$$

This case is overidentified ($q = 4$ and $k = 2$), so a weighting matrix is needed.

Example 10.4 *(Moment conditions for variances and a covariance) For expositional simplicity, assume that both variables have zero means. The variances and the covariance can then be estimated by the moment conditions*

$$\textstyle\sum_{t=1}^{T} g_t(\beta)/T = \mathbf{0}_{3\times 1} \text{ where } g_t = \begin{bmatrix} x_t^2 - \sigma_{xx} \\ y_t^2 - \sigma_{yy} \\ x_t y_t - \sigma_{xy} \end{bmatrix} \text{ and } \beta = \begin{bmatrix} \sigma_{xx} \\ \sigma_{yy} \\ \sigma_{xy} \end{bmatrix}.$$

10.1.1 Distribution of the Basic GMM

GMM estimators are typically asymptotically normally distributed, with a covariance matrix that depends on the covariance matrix of the moment conditions (S_0) and the mapping

from the parameters to the moment conditions (D_0). The details of these matrices are discussed below. For now, notice that the distribution of the GMM estimates is

$$\sqrt{T}(\hat{\beta} - \beta_0) \xrightarrow{d} N(0, V) \text{ if } W = S_0^{-1}, \text{ where}$$
$$V = \left(D_0' S_0^{-1} D_0\right)^{-1}, \tag{10.3}$$

provided we have used S_0^{-1} as the weighting matrix ($W = S_0^{-1}$) in (10.2). The choice of the weighting matrix is irrelevant if the model is exactly identified, so (11.17) can be applied to this case (even if we did not specify any weighting matrix at all). It can also be noticed that when the model is exactly identified, then we can typically rewrite the covariance matrix as $V = D_0^{-1} S_0 (D_0^{-1})'$, which might be easier to calculate. (The case of using another W matrix is discussed below.)

Let S_0 be the $(q \times q)$ covariance matrix of $\sqrt{T}\bar{g}(\beta_0)$, evaluated at the true parameter values

$$S_0 = \text{Cov}[\sqrt{T}\bar{g}(\beta_0)], \tag{10.4}$$

where Cov() is a matrix of covariances. When there is no autocorrelation of the moments, then (10.4) becomes

$$S_0 = \text{Cov}\left[g_t(\beta_0)\right], \text{ if } g_t \text{ is not autocorrelated.} \tag{10.5}$$

When there is autocorrelation, then we may use the Newey-West approach to estimate S_0.

In practice, S_0 is estimated by using the estimated coefficients in the moments to get the data series $g_t(\hat{\beta})$, a $T \times q$ matrix, from which we estimate the covariances needed for (10.4) or (10.5).

Example 10.5 *(Estimating a mean, variance) The moment in Example 10.1 (assuming iid data, so we can use (10.5)) gives*

$$S_0 = \text{Var}(x_t) = \sigma^2.$$

In practice, we replace the variance by a sample estimate. If we suspect that x_t is autocorrelated, then we may use the NW estimator of $\text{Var}(\sqrt{T}\bar{g})$.

Example 10.6 *(OLS, covariance) For the moments in Example 10.2, using $u_t = y_t - x_t'\beta$, we have*

$$S_0 = \text{Cov}\left[\frac{\sqrt{T}}{T}\sum_{t=1}^{T} x_t u_t\right]$$

In practice, replace u_t by the fitted residuals and calculate a sample covariance. It can be shown that under the Gauss-Markov assumptions $S_0 = \sigma^2 \Sigma_{xx}$. If we suspect that the variance of u_t is related to x_t, then we should calculate the covariance matrix of g_t, which gives White's covariance estimator. In addition we suspect that g_t is autocorrelated, then we may use the NW estimator of $\text{Var}(\sqrt{T}\bar{g})$.

Example 10.7 *(Estimating/testing a normal distribution, covariance) Assuming iid normally distributed data (so we can use (10.5)) the moments in Example 10.3 would have the following variance-covariance matrix*

$$S_0 = \begin{bmatrix} \sigma^2 & 0 & 3\sigma^4 & 0 \\ 0 & 2\sigma^4 & 0 & 12\sigma^6 \\ 3\sigma^4 & 0 & 15\sigma^6 & 0 \\ 0 & 12\sigma^6 & 0 & 96\sigma^8 \end{bmatrix}.$$

In practice, we instead use the point estimates in the moments and calculate the sample covariance matrix. If we suspect that g_t is autocorrelated, then we may use the NW estimator of $\text{Var}(\sqrt{T}\bar{g})$.

Let D_0 be the $(q \times k)$ probability limit of the gradient (Jacobian) of the sample moment conditions with respect to the parameters (also evaluated at the true parameters)

$$D_0 = \text{plim} \frac{\partial \bar{g}(\beta_0)}{\partial \beta'}. \tag{10.6}$$

In practice, the gradient D_0 is approximated by using the point estimates and the available sample of data.

Remark 10.8 *(Jacobian) The Jacobian is of the following format*

$$\frac{\partial \bar{g}(\beta_0)}{\partial \beta'} = \begin{bmatrix} \frac{\partial \bar{g}_1(\beta)}{\partial \beta_1} & \cdots & \frac{\partial \bar{g}_1(\beta)}{\partial \beta_k} \\ \vdots & & \vdots \\ \vdots & & \vdots \\ \frac{\partial \bar{g}_q(\beta)}{\partial \beta_1} & \cdots & \frac{\partial \bar{g}_q(\beta)}{\partial \beta_k} \end{bmatrix} \textit{(evaluated at } \beta_0\textit{)}.$$

Example 10.9 *(Estimating a mean, Jacobian) For the moment in Example 10.1*

$$D_0 = \frac{\partial}{\partial \mu} \frac{1}{T} \sum_{t=1}^{T} (x_t - \mu) = -1,$$

which does not involve any parameters or any data.

Example 10.10 *(OLS, Jacobian) For the moments in Example 10.2*

$$D_0 = \text{plim}\left(-\frac{1}{T}\sum_{t=1}^{T} x_t x_t'\right) = -\Sigma_{xx}.$$

This does not contain any parameters either, but includes data. In practice, we replace Σ_{xx} by a sample estimate.

Example 10.11 *(Estimating/testing a normal distribution, covariance) For the moments in Example 10.3 (assuming iid normally distributed data) we have (the rows are for the four different moment conditions, the columns for the parameters: μ and σ^2)*

$$\begin{aligned} D_0 &= \text{plim}\,\frac{1}{T}\sum_{t=1}^{T}\begin{bmatrix} -1 & 0 \\ -2(x_t-\mu) & -1 \\ -3(x_t-\mu)^2 & 0 \\ -4(x_t-\mu)^3 & -6\sigma^2 \end{bmatrix} \\ &= \begin{bmatrix} -1 & 0 \\ 0 & -1 \\ -3\sigma^2 & 0 \\ 0 & -6\sigma^2 \end{bmatrix}. \end{aligned}$$

Element (4,1) of the second equality holds only if the data has a symmetric distribution (for instance, a normal distribution). In practice, we would use the point estimates in the matrix on the first line and calculate the sample average.

Example 10.12 *(Estimating a mean, distribution) For the moment condition in Example 10.1 we have (assuming iid data)*

$$\sqrt{T}(\hat{\mu} - \mu_0) \xrightarrow{d} N(0, \sigma^2), \text{ so } \text{“}\hat{\mu} \sim N(\mu_0, \sigma^2/T).\text{”}$$

Example 10.13 *(OLS, distribution) For the moment conditions in Example 10.2*

$$V = \left(\Sigma_{xx} S_0^{-1} \Sigma_{xx}\right)^{-1}.$$

Under the Gauss-Markov assumptions $S_0 = \sigma^2 \Sigma_{xx}$, so

$$V = \left[\Sigma_{xx}\left(\sigma^2 \Sigma_{xx}\right)^{-1} \Sigma_{xx}\right]^{-1} = \sigma^2 \Sigma_{xx}^{-1}.$$

Example 10.14 *(Estimating/testing a normal distribution, distribution) For the moment conditions in Example 10.3 (assuming iid normally distributed data) we have that the*

asymptotic covariance matrix of the estimated mean and variance is then $((D_0'S_0^{-1}D_0)^{-1})$

$$\left(\begin{bmatrix} -1 & 0 \\ 0 & -1 \\ -3\sigma^2 & 0 \\ 0 & -6\sigma^2 \end{bmatrix}' \begin{bmatrix} \sigma^2 & 0 & 3\sigma^4 & 0 \\ 0 & 2\sigma^4 & 0 & 12\sigma^6 \\ 3\sigma^4 & 0 & 15\sigma^6 & 0 \\ 0 & 12\sigma^6 & 0 & 96\sigma^8 \end{bmatrix}^{-1} \begin{bmatrix} -1 & 0 \\ 0 & -1 \\ -3\sigma^2 & 0 \\ 0 & -6\sigma^2 \end{bmatrix}\right)^{-1} = \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{bmatrix}^{-1}$$

$$= \begin{bmatrix} \sigma^2 & 0 \\ 0 & 2\sigma^4 \end{bmatrix}.$$

In an overidentified model ($k < q$), we can test if the k parameters make all q moment conditions hold. Notice that under the null hypothesis (that the model is correctly specified)

$$\sqrt{T}\bar{g}(\beta_0) \xrightarrow{d} N\left(\mathbf{0}_{q\times 1}, S_0\right), \tag{10.7}$$

where q is the number of moment conditions. Since $\hat{\beta}$ is chosen in such a way that k linear combinations of the moment conditions are zero, there are effectively only $q-k$ non-degenerate random variables. We can therefore test the hypothesis that $\bar{g}(\beta_0) = 0$ by the "J test"

$$T\bar{g}(\hat{\beta})'S_0^{-1}\bar{g}(\hat{\beta}) \xrightarrow{d} \chi^2_{q-k}, \text{ if } W = S_0^{-1}. \tag{10.8}$$

The left hand side equals T times of value of the loss function in (10.2) evaluated at the point estimates. With no overidentifying restrictions ($q = k$) there are no restrictions to test. Indeed, the loss function value is then always zero at the point estimates.

Example 10.15 *(Estimating/testing a normal distribution, testing) After having estimated the mean and the variance, we can test if all four moment conditions in Example 10.3 hold. If data is drawn from a normal distribution, they should hold (give and take some randomness).*

Empirical Example 10.16 *(Estimating a mean a variance with GMM) Table 10.1 reports estimates of the mean and variance of the FF equity market return. The first approach (column) uses only the first two moment conditions of 10.3 (an exactly identified case). Other approaches apply different (suboptimal) W matrices in solving (10.2). Finally, the last approaches apply (11.36) by using different A matrices. Table 10.2 reports results from iterating on the W matrix when solving (10.2). The W matrix used in the final iteration is shown in Table 10.3.*

	ex. ident.	$\bar{g}'W_1\bar{g}$	$\bar{g}'W_2\bar{g}$	$A_1\bar{g}$	$A_2\bar{g}$
μ	0.62	0.62	0.39	0.62	0.57
σ^2	20.86	20.86	20.91	20.86	20.86

Table 10.1: Estimates of mean and variance of the FF equity market factor, 1970:01-2021:12. The W_1 and A_1 matrices put equal weights on moment conditions 1–2 and zero weight moment conditions 3–4, while W_2 and A_2 put also a very small weight on moment condition 3.

	iteration				
	0	1	2	3	4
μ	0.62	0.75	0.76	0.76	0.76
σ^2	20.86	18.59	18.55	18.55	18.55

Table 10.2: Estimates of mean and variance of the FF equity market factor, 1970:01-2021:12. The estimates minimize $\bar{g}'W_i\bar{g}$, where $W_i = S_{i-1}^{-1}$

10.2 GMM with a Suboptimal Weighting Matrix

The distribution of the GMM estimates when we use a sub-optimal weighting matrix is similar to (11.17), but the variance-covariance matrix is different (basically, reflecting the fact that the approach does not produce the lowest possible variances anymore).

Example 10.17 *(Estimating/testing a normal distribution) Example 10.3 is overidentified since there are four moment conditions but only two parameters. Instead of using the optimal weighting matrix (the inverse of S_0 from Example 10.7, assuming the data is iid normally distributed), we could use any other (positive definite) 4×4 matrix. For instance, $W = I_4$ or a matrix that puts almost all weight on the first two moment conditions.*

It can be shown that if we use another weighting matrix than $W = S_0^{-1}$, then the

1210.00	58.17	−9.61	−0.40
58.17	18.99	−0.62	−0.07
−9.61	−0.62	0.14	0.01
−0.40	−0.07	0.01	0.00

Table 10.3: $W_i \times 10000$ used in the last iteration when, minimizing $\bar{g}'W_i\bar{g}$ to estimate the mean and variance of the FF equity market factor, 1970:01-2021:12.

variance-covariance matrix in (11.17) should be changed to

$$\begin{aligned} V_2 &= V_{A2} D_0' W S_0 W' D_0 V_{A2}', \text{ where} \\ V_{A2} &= \left(D_0' W D_0\right)^{-1}. \end{aligned} \tag{10.9}$$

Similarly, the test of overidentifying restrictions becomes

$$T \bar{g}(\hat{\beta})' \Psi_2^+ \bar{g}(\hat{\beta}) \xrightarrow{d} \chi^2_{q-k}, \tag{10.10}$$

where Ψ_2^+ is a generalized inverse of

$$\begin{aligned} \Psi_2 &= \Psi_{A2} S_0 \Psi_{A2}', \text{ where} \\ \Psi_{A2} &= I_q - D_0 \left(D_0' W D_0\right)^{-1} D_0' W. \end{aligned} \tag{10.11}$$

The covariance matrix Ψ_2 has a reduced rank, so we must use a generalized inverse in the test.

Remark 10.18 *(Quadratic form with degenerate covariance matrix) If the $n \times 1$ vector $X \sim N(0, \Sigma)$, where Σ has rank $r \leq n$ then $Y = X'\Sigma^+ X \sim \chi^2_r$ where Σ^+ is the pseudo inverse of Σ.*

Example 10.19 *(Pseudo inverse of a square matrix) For the matrix*

$$A = \begin{bmatrix} 1 & 2 \\ 3 & 6 \end{bmatrix}, \text{ we have } A^+ = \begin{bmatrix} 0.02 & 0.06 \\ 0.04 & 0.12 \end{bmatrix}.$$

10.3 GMM without a Loss Function

Suppose we sidestep the whole optimization issue and instead specify k linear combinations of the q moment conditions directly

$$\mathbf{0}_{k\times 1} = \underbrace{A}_{k\times q} \underbrace{\bar{g}(\hat{\beta})}_{q\times 1}, \tag{10.12}$$

where the matrix A is chosen by the researcher. We can solve (possibly with a numerical algorithm) for the $\hat{\beta}$ values that make these equations hold.

Example 10.20 *(Overidentified example: estimating/testing a normal distribution) Example 10.3 is overidentified since there are four moment conditions but only two param-*

eters. One possible A matrix would put all weight on the first two moment conditions

$$A = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}.$$

It is straightforward to show that the variance-covariance matrix in (11.17) should be changed to

$$\begin{aligned} V_3 &= V_{A3} A_0 S_0 A_0' V_{A3}', \text{ where} \\ V_{A3} &= (A_0 D_0)^{-1}, \end{aligned} \tag{10.13}$$

where A_0 is the probability limit of A (if it is random).

Similarly, in the test of overidentifying restrictions (11.35), we should replace Ψ_2 by

$$\begin{aligned} \Psi_3 &= \Psi_{A3} S_0 \Psi_{A3}', \text{ where} \\ \Psi_{A3} &= I_q - D_0 (A_0 D_0)^{-1} A_0. \end{aligned} \tag{10.14}$$

The covariance matrix Ψ_3 has a reduced rank, so we must again use a generalized inverse in the test.

Example 10.21 *(Estimating/testing a normal distribution) Continuing Example 10.20, we have that $A_0 D_0$ in (11.37) is*

$$V_{A3} = \left(\underbrace{\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}}_{A_0} \underbrace{\begin{bmatrix} -1 & 0 \\ 0 & -1 \\ -3\sigma^2 & 0 \\ 0 & -6\sigma^2 \end{bmatrix}}_{D_0} \right)^{-1} = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}.$$

Example 10.22 *(Estimating/testing a normal distribution) Continuing the previous ex-*

ample, Ψ_{A3} in (11.39) is

$$\Psi_{A3} = \underbrace{\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}}_{I_4} - \underbrace{\begin{bmatrix} -1 & 0 \\ 0 & -1 \\ -3\sigma^2 & 0 \\ 0 & -6\sigma^2 \end{bmatrix}}_{D_0} \left(\underbrace{\begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}}_{A_0 D_0} \right)^{-1} \underbrace{\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}}_{A_0}$$

$$= \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ -3\sigma^2 & 0 & 1 & 0 \\ 0 & -6\sigma^2 & 0 & 1 \end{bmatrix}.$$

Ψ_3 in (11.39) is therefore

$$\Psi_3 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ -3\sigma^2 & 0 & 1 & 0 \\ 0 & -6\sigma^2 & 0 & 1 \end{bmatrix} \begin{bmatrix} \sigma^2 & 0 & 3\sigma^4 & 0 \\ 0 & 2\sigma^4 & 0 & 12\sigma^6 \\ 3\sigma^4 & 0 & 15\sigma^6 & 0 \\ 0 & 12\sigma^6 & 0 & 96\sigma^8 \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ -3\sigma^2 & 0 & 1 & 0 \\ 0 & -6\sigma^2 & 0 & 1 \end{bmatrix}'$$

$$= \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 6\sigma^6 & 0 \\ 0 & 0 & 0 & 24\sigma^8 \end{bmatrix}$$

Example 10.23 *(Estimating/testing a normal distribution) Continuing the previous example, we have that the test of the overidentifying restrictions (11.35) (assuming iid normally distributed data to calculate S_0) is (notice the generalized inverse of Ψ_3)*

$$= T \begin{bmatrix} 0 \\ 0 \\ \Sigma_{t=1}^T (x_t - \mu)^3 / T \\ \Sigma_{t=1}^T [(x_t - \mu)^4 - 3\sigma^4] / T \end{bmatrix}' \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1/(6\sigma^6) & 0 \\ 0 & 0 & 0 & 1/(24\sigma^8) \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ \Sigma_{t=1}^T (x_t - \mu)^3 / T \\ \Sigma_{t=1}^T [(x_t - \mu)^4 - 3\sigma^4] / T \end{bmatrix}$$

$$= \frac{T}{6} \frac{\left[\Sigma_{t=1}^T (x_t - \mu)^3 / T\right]^2}{\sigma^6} + \frac{T}{24} \frac{\left\{\Sigma_{t=1}^T [(x_t - \mu)^4 - 3\sigma^4] / T\right\}^2}{\sigma^8}.$$

When we replace μ and σ by their estimates, then this is the same as the Jarque-Bera test of normality.

10.4 GMM Example: The Means and Second Moments of Returns

Let R_t be a vector of net returns of N assets. We want to estimate the mean vector and the covariance matrix. The moment conditions for the mean vector are

$$\mathrm{E}\, R_t - \mu = \mathbf{0}_{N\times 1}, \tag{10.15}$$

and the moment conditions for the unique elements of the second moment matrix are

$$\mathrm{E}\,\mathrm{vech}(R_t R_t') - \mathrm{vech}(\Gamma) = \mathbf{0}_{N(N+1)/2\times 1}. \tag{10.16}$$

Remark 10.24 *(The vech operator) vech(A) where A is $m \times m$ gives an $m(m+1)/2 \times 1$ vector with the elements on and below the principal diagonal A stacked on top of each other (column wise). For instance, vech* $\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = \begin{bmatrix} a_{11} \\ a_{21} \\ a_{22} \end{bmatrix}$.

Stack (10.15) and (10.16) and substitute the sample mean for the population expectation to get the GMM estimator

$$\frac{1}{T}\sum_{t=1}^{T}\begin{bmatrix} R_t \\ \mathrm{vech}(R_t R_t') \end{bmatrix} - \begin{bmatrix} \hat{\mu} \\ \mathrm{vech}(\hat{\Gamma}) \end{bmatrix} = \begin{bmatrix} \mathbf{0}_{N\times 1} \\ \mathbf{0}_{N(N+1)/2\times 1} \end{bmatrix}. \tag{10.17}$$

In this case, $D_0 = -I$, so the covariance matrix of the parameter vector $(\hat{\mu}, \mathrm{vech}(\hat{\Gamma}))$ is just S_0 (defined in (10.4)), which is straightforward to estimate.

Chapter 11

GMM

References: Greene (2000) 4.7 and 11.5-6
Additional references: Hayashi (2000) 3-4; Verbeek (2004) 5; Hamilton (1994) 14; Ogaki (1993), Johnston and DiNardo (1997) 10; Harris and Matyas (1999); Pindyck and Rubinfeld (1998) Appendix 10.1; Cochrane (2001) 10-11; Hansen (1982)

11.1 Method of Moments

Let $g(x_t)$ be a $k \times 1$ vector valued continuous function of a stationary process, and let the probability limit of the mean of $g(.)$ be a function $\gamma(.)$ of a $k \times 1$ vector β of parameters. We want to estimate β. The method of moments (MM, not yet generalized to GMM) estimator is obtained by replacing the probability limit with the sample mean and solving the system of k equations

$$\frac{1}{T}\sum_{t=1}^{T} g(x_t) - \gamma(\beta) = \mathbf{0}_{k\times 1} \tag{11.1}$$

for the parameters β.

It is clear that this is a consistent estimator of β if γ is continuous. (Proof: the sample mean is a consistent estimator of $\gamma(.)$, and by Slutsky's theorem $\operatorname{plim}\gamma(\hat{\beta}) = \gamma(\operatorname{plim}\hat{\beta})$ if γ is a continuous function.)

Example 11.1 *(Moment conditions for variances and covariance) Suppose the series x_t and y_t have zero means. The following moment conditions define the traditional variance*

and covariance estimators

$$\begin{bmatrix} \frac{1}{T}\sum_{t=1}^{T} x_t^2 - \sigma_{xx} \\ \frac{1}{T}\sum_{t=1}^{T} y_t^2 - \sigma_{yy} \\ \frac{1}{T}\sum_{t=1}^{T} x_t y_t - \sigma_{xy} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}.$$

It does not matter if the parameterers are estimated separately or jointly. In contrast, if we want the correlation, ρ_{xy}, instead of the covariance, then we change the last moment condition to

$$\frac{1}{T}\sum_{t=1}^{T} x_t y_t - \rho_{xy}\sqrt{\sigma_{xx}}\sqrt{\sigma_{yy}} = 0,$$

which must be estimated jointly with the first two conditions.

Example 11.2 *(MM for an MA(1).) For an MA(1), $y_t = \epsilon_t + \theta\epsilon_{t-1}$, we have*

$$\begin{array}{rcccl} \mathrm{E}\, y_t^2 & = & \mathrm{E}\left(\epsilon_t + \theta\epsilon_{t-1}\right)^2 & = & \sigma_\epsilon^2\left(1+\theta^2\right) \\ \mathrm{E}\left(y_t y_{t-1}\right) & = & \mathrm{E}\left[\left(\epsilon_t + \theta\epsilon_{t-1}\right)\left(\epsilon_{t-1} + \theta\epsilon_{t-2}\right)\right] & = & \sigma_\epsilon^2\theta. \end{array}$$

The moment conditions could therefore be

$$\begin{bmatrix} \frac{1}{T}\sum_{t=1}^{T} y_t^2 - \sigma_\epsilon^2\left(1+\theta^2\right) \\ \frac{1}{T}\sum_{t=1}^{T} y_t y_{t-1} - \sigma_\epsilon^2\theta \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix},$$

which allows us to estimate θ and σ^2.

11.2 Generalized Method of Moments

GMM extends MM by allowing for more orthogonality conditions than parameters. This could, for instance, increase efficiency and/or provide new aspects which can be tested.

Many (most) traditional estimation methods, like LS, IV, and MLE are special cases of GMM. This means that the properties of GMM are very general, and therefore fairly difficult to prove.

11.3 Moment Conditions in GMM

Suppose we have q (unconditional) moment conditions,

$$\begin{aligned} \mathrm{E}\, g(w_t, \beta_0) &= \begin{bmatrix} \mathrm{E} g_1(w_t, \beta_0) \\ \vdots \\ \mathrm{E} g_q(w_t, \beta_0) \end{bmatrix} \\ &= \mathbf{0}_{q\times 1}, \end{aligned} \tag{11.2}$$

from which we want to estimate the $k \times 1$ ($k \leq q$) vector of parameters, β. The true values are β_0. We assume that w_t is a stationary and ergodic (vector) process (otherwise the sample means does not converge to anything meaningful as the sample size increases). The sample averages, or "sample moment conditions," evaluated at some value of β, are

$$\bar{g}(\beta) = \frac{1}{T}\sum_{t=1}^{T} g(w_t, \beta). \tag{11.3}$$

The sample average $\bar{g}\,(\beta)$ is a vector of functions of random variables, so they are random variables themselves and depend on the sample used. It will later be interesting to calculate the variance of $\bar{g}\,(\beta)$. Note that $\bar{g}(\beta_1)$ and $\bar{g}(\beta_2)$ denote sample means obtained by using two different parameter vectors, but on the same sample of data.

Example 11.3 *(Moments conditions for IV/2SLS.) Consider the linear model $y_t = x_t'\beta_0 + u_t$, where x_t and β are $k \times 1$ vectors. Let z_t be a $q \times 1$ vector, with $q \geq k$. The moment conditions and their sample analogues are*

$$\mathbf{0}_{q\times 1} = \mathrm{E}\, z_t u_t = \mathrm{E}[z_t(y_t - x_t'\beta_0)], \text{ and } \bar{g}\,(\beta) = \frac{1}{T}\sum_{t=1}^{T} z_t(y_t - x_t'\beta),$$

(or $Z'(Y - X\beta)/T$ in matrix form). Let $q = k$ to get IV; let $z_t = x_t$ to get LS.

Example 11.4 *(Moments conditions for MLE.) The maximum likelihood estimator maximizes the log likelihood function, $\frac{1}{T}\Sigma_{t=1}^{T} \ln L\,(w_t; \beta)$, which requires $\frac{1}{T}\Sigma_{t=1}^{T} \partial \ln L\,(w_t; \beta)\,/\partial\beta = 0$, which is just like a GMM moment condition.*

11.3.1 Digression: From Conditional to Unconditional Moment Conditions*

Suppose we are instead given *conditional* moment restrictions

$$\mathrm{E}\left[u(x_t, \beta_0)|z_t\right] = \mathbf{0}_{m\times 1}, \tag{11.4}$$

where z_t is a vector of conditioning (predetermined) variables. We want to transform this to unconditional moment conditions.

Remark 11.5 *(*$\mathrm{E}(u|z) = 0$ *implies* $\mathrm{E}\,uz = 0$*.) The condition* $\mathrm{E}(u|z) = 0$ *implies (a)* $Cov(z,u) = 0$ *(since* $Cov(z,u) = Cov[z, \mathrm{E}(u|z)]$*) and (b)* $\mathrm{E}\,u = 0$ *(since* $\mathrm{E}\,u = \mathrm{E}_z\,\mathrm{E}(u|z)$*).*

Example 11.6 *(Euler equation for optimal consumption.) The standard Euler equation for optimal consumption choice which with isoelastic utility* $U(C_t) = C_t^{1-\gamma}/(1-\gamma)$ *is*

$$\mathrm{E}\left[R_{t+1}\beta\left(\frac{C_{t+1}}{C_t}\right)^{-\gamma} - 1\middle|\,\Omega_t\right] = 0,$$

where R_{t+1} *is a gross return on an investment and* Ω_t *is the information set in* t*. Let* $z_t \in \Omega_t$*, for instance asset returns or consumption* t *or earlier. The Euler equation then implies*

$$\mathrm{E}\left[R_{t+1}\beta\left(\frac{C_{t+1}}{C_t}\right)^{-\gamma} z_t - z_t\right] = 0.$$

Let $z_t = (z_{1t}, ..., z_{nt})'$, and define the new (unconditional) moment conditions as

$$g(w_t, \beta) = u(x_t, \beta) \otimes z_t = \begin{bmatrix} u_1(x_t, \beta)z_{1t} \\ \vdots \\ u_1(x_t, \beta)z_{nt} \\ u_2(x_t, \beta)z_{1t} \\ \vdots \\ u_m(x_t, \beta)z_{nt} \end{bmatrix}_{q\times 1}, \tag{11.5}$$

which by (11.4) must have an expected value of zero, that is

$$\mathrm{E}\,g(w_t, \beta_0) = \mathbf{0}_{q\times 1}. \tag{11.6}$$

This a set of unconditional moment conditions—just as in (11.2). The sample moment conditions (11.3) are therefore valid also in the conditional case, although we have to specify $g(w_t, \beta)$ as in (11.5).

Note that the choice of instruments is often arbitrary: it often amounts to using only a subset of the information variables. GMM is often said to be close to economic theory, but it should be admitted that economic theory sometimes tells us fairly little about which instruments, z_t, to use.

11.4 The Optimization Problem in GMM

11.4.1 The Loss Function

The GMM estimator $\hat{\beta}$ minimizes the weighted quadratic form

$$J = \begin{bmatrix} \bar{g}_1(\beta) \\ \vdots \\ \vdots \\ \bar{g}_q(\beta) \end{bmatrix}' \begin{bmatrix} W_{11} & \cdots & \cdots & W_{1q} \\ \vdots & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ W_{1q} & \cdots & \cdots & W_{qq} \end{bmatrix} \begin{bmatrix} \bar{g}_1(\beta) \\ \vdots \\ \vdots \\ \bar{g}_q(\beta) \end{bmatrix} \tag{11.7}$$

$$= \bar{g}(\beta)' W \bar{g}(\beta), \tag{11.8}$$

where $\bar{g}(\beta)$ is the sample average of $g(w_t, \beta)$ given by (11.3), and where W is some $q \times q$ symmetric positive definite weighting matrix. (We will soon discuss a good choice of weighting matrix.) There are k parameters in β to estimate, and we have q moment conditions in $\bar{g}(\beta)$. We therefore have $q - k$ *overidentifying moment restrictions*.

With $q = k$ the model is exactly identified (as many equations as unknowns), and it should be possible to set all q sample moment conditions to zero by a choosing the $k = q$ parameters. It is clear that the choice of the weighting matrix has no effect in this case since $\bar{g}(\hat{\beta}) = 0$ at the point estimates $\hat{\beta}$. In this case, GMM is just MM.

Example 11.7 *(Simple linear regression.) Consider the model*

$$y_t = x_t \beta_0 + u_t, \tag{11.9}$$

where y_t and x_t are zero mean scalars. The moment condition and loss function are

$$\bar{g}(\beta) = \frac{1}{T} \sum_{t=1}^{T} x_t (y_t - x_t \beta) \text{ and}$$

$$J = W \left[\frac{1}{T} \sum_{t=1}^{T} x_t (y_t - x_t \beta) \right]^2,$$

so the scalar W is clearly irrelevant in this case.

Example 11.8 *(IV/2SLS method continued.) From Example 11.3, we note that the loss function for the IV/2SLS method is*

$$\bar{g}(\beta)'W\bar{g}(\beta) = \left[\frac{1}{T}\sum_{t=1}^{T} z_t(y_t - x_t'\beta)\right]' W \left[\frac{1}{T}\sum_{t=1}^{T} z_t(y_t - x_t'\beta)\right].$$

When $q = k$, then the model is exactly identified, so the estimator could actually be found by setting all moment conditions to zero. We then get the IV estimator

$$\begin{aligned} \mathbf{0} &= \frac{1}{T}\sum_{t=1}^{T} z_t(y_t - x_t'\hat{\beta}_{IV}) \text{ or} \\ \hat{\beta}_{IV} &= \left(\frac{1}{T}\sum_{t=1}^{T} z_t x_t'\right)^{-1} \frac{1}{T}\sum_{t=1}^{T} z_t y_t \\ &= \hat{\Sigma}_{zx}^{-1}\hat{\Sigma}_{zy}, \end{aligned}$$

where $\hat{\Sigma}_{zx} = \Sigma_{t=1}^{T} z_t x_t'/T$ and similarly for the other second moment matrices. Let $z_t = x_t$ to get LS

$$\hat{\beta}_{LS} = \hat{\Sigma}_{xx}^{-1}\hat{\Sigma}_{xy}.$$

11.4.2 First Order Conditions

Remark 11.9 *(Matrix differentiation of quadratic forms.) Let $x_{n\times 1}$, $f(x)_{m\times 1}$, and $A_{m\times m}$ symmetric. Then*

$$\frac{\partial f(x)' A f(x)}{\partial x} = 2\left(\frac{\partial f(x)}{\partial x'}\right)' A f(x).$$

The k first order conditions for minimizing the GMM loss function in (11.8) with respect to the k parameters are that the partial derivatives with respect to β equal zero at

the estimate, $\hat{\beta}$,

$$\begin{aligned}
\mathbf{0}_{k\times 1} &= \frac{\partial \bar{g}(\hat{\beta})' W \bar{m}(\hat{\beta})}{\partial \beta} \\
&= \begin{bmatrix} \frac{\partial \bar{g}_1(\hat{\beta})}{\partial \beta_1} & \cdots & \frac{\partial \bar{g}_1(\hat{\beta})}{\partial \beta_k} \\ \vdots & & \vdots \\ \vdots & & \vdots \\ \frac{\partial \bar{g}_q(\hat{\beta})}{\partial \beta_1} & \cdots & \frac{\partial \bar{g}_q(\hat{\beta})}{\partial \beta_k} \end{bmatrix}' \begin{bmatrix} W_{11} & \cdots & \cdots & W_{1q} \\ \vdots & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ W_{1q} & \cdots & \cdots & W_{qq} \end{bmatrix} \begin{bmatrix} \bar{g}_1(\hat{\beta}) \\ \vdots \\ \vdots \\ \bar{g}_q(\hat{\beta}) \end{bmatrix} \text{ (with } \hat{\beta}_{k\times 1}\text{),} \qquad (11.10) \\
&= \underbrace{\left(\frac{\partial \bar{g}(\hat{\beta})}{\partial \beta'}\right)'}_{k\times q} \underbrace{W}_{q\times q} \underbrace{\bar{g}(\hat{\beta})}_{q\times 1}. \qquad (11.11)
\end{aligned}$$

We can solve for the GMM estimator, $\hat{\beta}$, from (11.11). This set of equations must often be solved by numerical methods, except in linear models (the moment conditions are linear functions of the parameters) where we can find analytical solutions by matrix inversion.

Example 11.10 *(First order conditions of simple linear regression.) The first order conditions of the loss function in Example 11.7 is*

$$\begin{aligned}
0 &= \frac{d}{d\beta} W \left[\frac{1}{T}\sum_{t=1}^{T} x_t(y_t - x_t\hat{\beta})\right]^2 \\
&= \left[-\frac{1}{T}\sum_{t=1}^{T} x_t^2\right] W \left[\frac{1}{T}\sum_{t=1}^{T} x_t(y_t - x_t\hat{\beta})\right], \text{ or} \\
\hat{\beta} &= \left(\frac{1}{T}\sum_{t=1}^{T} x_t^2\right)^{-1} \frac{1}{T}\sum_{t=1}^{T} x_t y_t.
\end{aligned}$$

Example 11.11 *(First order conditions of IV/2SLS.) The first order conditions corre-*

sponding to (11.11) of the loss function in Example 11.8 (when $q \geq k$) are

$$\begin{aligned}
\mathbf{0}_{k\times 1} &= \left[\frac{\partial \bar{g}(\hat{\beta})}{\partial \beta'}\right]' W \bar{g}(\hat{\beta}) \\
&= \left[\frac{\partial}{\partial \beta'}\frac{1}{T}\sum_{t=1}^{T} z_t(y_t - x_t'\hat{\beta})\right]' W \frac{1}{T}\sum_{t=1}^{T} z_t(y_t - x_t'\hat{\beta}) \\
&= \left[-\frac{1}{T}\sum_{t=1}^{T} z_t x_t'\right]' W \frac{1}{T}\sum_{t=1}^{T} z_t(y_t - x_t'\hat{\beta}) \\
&= -\hat{\Sigma}_{xz} W(\hat{\Sigma}_{zy} - \hat{\Sigma}_{zx}\hat{\beta}).
\end{aligned}$$

We can solve for $\hat{\beta}$ from the first order conditions as

$$\hat{\beta}_{2SLS} = \left(\hat{\Sigma}_{xz} W \hat{\Sigma}_{zx}\right)^{-1} \hat{\Sigma}_{xz} W \hat{\Sigma}_{zy}.$$

When $q = k$, then the first order conditions can be premultiplied with $(\hat{\Sigma}_{xz}W)^{-1}$, since $\hat{\Sigma}_{xz}W$ is an invertible $k \times k$ matrix in this case, to give

$$\mathbf{0}_{k\times 1} = \hat{\Sigma}_{zy} - \hat{\Sigma}_{zx}\hat{\beta}, \text{ so } \hat{\beta}_{IV} = \hat{\Sigma}_{zx}^{-1}\hat{\Sigma}_{zy}.$$

11.5 Asymptotic Properties of GMM

We know very little about the general small sample properties, including bias, of GMM. We therefore have to rely either on simulations (Monte Carlo or bootstrap) or on the asymptotic results. This section is about the latter.

GMM estimates are typically consistent and normally distributed, even if the series $g(w_t, \beta)$ in the moment conditions (11.3) are serially correlated and heteroskedastic—provided w_t is a stationary and ergodic process. The reason is essentially that the estimators are (at least as a first order approximation) linear combinations of sample means which typically are consistent (LLN) and normally distributed (CLT). More about that later. The proofs are hard, since the GMM is such a broad class of estimators. This section discusses, in an informal way, how we can arrive at those results.

11.5.1 Consistency

Sample moments are typically consistent, so plim $g(\beta) = \mathrm{E}\, g(w_t, \beta)$. This must hold at any parameter vector in the relevant space (for instance, those inducing stationarity and

variances which are strictly positive). Then, if the moment conditions (11.2) are true only at the true parameter vector, β_0, (otherwise the parameters are "unidentified") and that they are continuous in β, then GMM is consistent. From Slutsky's theorem

$$\operatorname{plim} \bar{g}(\hat{\beta}) = \bar{g}(\operatorname{plim} \hat{\beta}),$$

and we know that this must equal $\mathrm{E}\, g(w_t, \operatorname{plim} \hat{\beta})$ and $\mathrm{E}\, g() = 0$ only at β_0.

Example 11.12 *(Consistency of 2SLS.) By using $y_t = x_t'\beta_0 + u_t$, the first order conditions in Example 11.11 can be rewritten*

$$\begin{aligned}
\mathbf{0}_{k\times 1} &= \hat{\Sigma}_{xz} W \frac{1}{T}\sum_{t=1}^{T} z_t(y_t - x_t'\hat{\beta}) \\
&= \hat{\Sigma}_{xz} W \frac{1}{T}\sum_{t=1}^{T} z_t\left[u_t + x_t'\left(\beta_0 - \hat{\beta}\right)\right] \\
&= \hat{\Sigma}_{xz} W \hat{\Sigma}_{zu} + \hat{\Sigma}_{xz} W \hat{\Sigma}_{zx}\left(\beta_0 - \hat{\beta}\right).
\end{aligned}$$

Take the probability limit

$$\mathbf{0}_{k\times 1} = \operatorname{plim} \hat{\Sigma}_{xz} W \operatorname{plim} \hat{\Sigma}_{zu} + \operatorname{plim} \hat{\Sigma}_{xz} W \operatorname{plim} \hat{\Sigma}_{zx}\left(\beta_0 - \operatorname{plim} \hat{\beta}\right).$$

In most cases, $\operatorname{plim} \hat{\Sigma}_{xz}$ *is some matrix of constants, and* $\operatorname{plim} \hat{\Sigma}_{zu} = \mathrm{E}\, z_t u_t = \mathbf{0}_{q\times 1}$. *It then follows that* $\operatorname{plim} \hat{\beta} = \beta_0$. *Note that the whole argument relies on that the moment condition,* $\mathrm{E}\, z_t u_t = \mathbf{0}_{q\times 1}$, *is true. If it is not, then the estimator is inconsistent. For instance, when the instruments are invalid (correlated with the residuals) or when we use LS ($z_t = x_t$) when there are measurement errors or in a system of simultaneous equations.*

11.5.2 Asymptotic Normality

To derive the asymptotic distribution of $\sqrt{T}(\hat{\beta} - \beta_0)$, we need to define three things. (As usual, we also need to scale with $\sqrt{T}$ to get a non-trivial asymptotic distribution; the asymptotic distribution of $\hat{\beta} - \beta_0$ is a spike at zero.) *First*, let S_0 (a $q \times q$ matrix) denote the asymptotic covariance matrix (as sample size goes to infinity) of $\sqrt{T}$ times the sample

moment conditions evaluated at the true parameters

$$S_0 = \text{Cov}\left[\sqrt{T}\bar{g}\left(\beta_0\right)\right] \tag{11.12}$$

$$= \text{Cov}\left[\frac{1}{\sqrt{T}}\sum_{t=1}^{T} g(w_t, \beta_0)\right], \tag{11.13}$$

where we use the definition of $\bar{g}\left(\beta_0\right)$ in (11.3). In practice, we often estimate this by using the Newey-West estimator (or something similar).

Second, let D_0 (a $q \times k$ matrix) denote the probability limit of the gradient of the sample moment conditions with respect to the parameters, evaluated at the true parameters

$$D_0 = \text{plim}\frac{\partial \bar{g}(\beta_0)}{\partial \beta'}, \text{ where} \tag{11.14}$$

$$\frac{\partial \bar{g}(\beta_0)}{\partial \beta'} = \begin{bmatrix} \frac{\partial \bar{g}_1(\beta)}{\partial \beta_1} & \cdots & \frac{\partial \bar{g}_1(\beta)}{\partial \beta_k} \\ \vdots & & \vdots \\ \vdots & & \vdots \\ \frac{\partial \bar{g}_q(\beta)}{\partial \beta_1} & \cdots & \frac{\partial \bar{g}_q(\beta)}{\partial \beta_k} \end{bmatrix} \text{ at the true } \beta \text{ vector.} \tag{11.15}$$

Notice that a similar gradient, but evaluated at $\hat{\beta}$, also shows up in the first order conditions (11.11).

Third, let the weighting matrix be the inverse of the covariance matrix of the moment conditions (once again evaluated at the true parameters)

$$W = S_0^{-1}. \tag{11.16}$$

It can be shown that this choice of weighting matrix gives the asymptotically most efficient estimator for a *given* set of orthogonality conditions. For instance, in 2SLS, this means a given set of instruments and (11.16) then shows only how to use these instruments in the most efficient way. Of course, another set of instruments might be better (in the sense of giving a smaller $\text{Cov}(\hat{\beta})$).

With the definitions in (11.12) and (11.14) and the choice of weighting matrix in (11.16) and the added assumption that the rank of D_0 equals k (the number of parameters) then we can show (under fairly general conditions) that

$$\sqrt{T}(\hat{\beta} - \beta_0) \xrightarrow{d} N(\mathbf{0}_{k\times 1}, V), \text{ where } V = \left(D_0' S_0^{-1} D_0\right)^{-1}. \tag{11.17}$$

This holds also when the model is exactly identified, so we really do not use any weighting

matrix.

To prove this note the following.

Proof. (The asymptotic distribution (11.17). Sketch of proof.) This proof is essentially an application of the delta rule. By the mean-value theorem the sample moment condition evaluated at the GMM estimate, $\hat{\beta}$, is

$$\bar{g}(\hat{\beta}) = \bar{g}(\beta_0) + \frac{\partial \bar{g}(\beta_1)}{\partial \beta'}(\hat{\beta} - \beta_0) \tag{11.18}$$

for some values β_1 between $\hat{\beta}$ and β_0. (This point is different for different elements in $\bar{g}$.) Premultiply with $[\partial \bar{g}(\hat{\beta})/\partial \beta']' W$. By the first order condition (11.11), the left hand side is then zero, so we have

$$\mathbf{0}_{k\times 1} = \left(\frac{\partial \bar{g}(\hat{\beta})}{\partial \beta'}\right)' W \bar{g}(\beta_0) + \left(\frac{\partial \bar{g}(\hat{\beta})}{\partial \beta'}\right)' W \frac{\partial \bar{g}(\beta_1)}{\partial \beta'}(\hat{\beta} - \beta_0). \tag{11.19}$$

Multiply with $\sqrt{T}$ and solve as

$$\sqrt{T}\left(\hat{\beta} - \beta_0\right) = \underbrace{-\left[\left(\frac{\partial \bar{g}(\hat{\beta})}{\partial \beta'}\right)' W \frac{\partial \bar{g}(\beta_1)}{\partial \beta'}\right]^{-1} \left(\frac{\partial \bar{g}(\hat{\beta})}{\partial \beta'}\right)' W}_{\Gamma} \sqrt{T}\bar{g}(\beta_0). \tag{11.20}$$

If

$$\operatorname{plim} \frac{\partial \bar{g}(\hat{\beta})}{\partial \beta'} = \frac{\partial \bar{g}(\beta_0)}{\partial \beta'} = D_0, \text{ then } \operatorname{plim} \frac{\partial \bar{g}(\beta_1)}{\partial \beta'} = D_0,$$

since β_1 is between β_0 and $\hat{\beta}$. Then

$$\operatorname{plim} \Gamma = -\left(D_0' W D_0\right)^{-1} D_0' W. \tag{11.21}$$

The last term in (11.20), $\sqrt{T}\bar{g}(\beta_0)$, is $\sqrt{T}$ times a vector of sample averages, so by a CLT it converges in distribution to $\mathrm{N}(0, S_0)$, where S_0 is defined as in (11.12). By the continuous mapping theorem we then have that

$$\begin{aligned}
&\sqrt{T}\left(\hat{\beta} - \beta_0\right) \xrightarrow{d} \operatorname{plim} \Gamma \ \times \text{ something that is } N\left(0, S_0\right), \text{ that is,} \\
&\sqrt{T}\left(\hat{\beta} - \beta_0\right) \xrightarrow{d} N\left[\mathbf{0}_{k\times 1}, (\operatorname{plim} \Gamma) S_0 (\operatorname{plim} \Gamma')\right].
\end{aligned}$$

The covariance matrix is then

$$\begin{aligned}\text{Cov}[\sqrt{T}(\hat{\beta}-\beta_0)] &= (\text{plim}\,\Gamma)S_0(\text{plim}\,\Gamma')\\ &= \left(D_0'WD_0\right)^{-1}D_0'WS_0[\left(D_0'WD_0\right)^{-1}D_0'W]' \qquad (11.22)\\ &= \left(D_0'WD_0\right)^{-1}D_0'WS_0W'D_0\left(D_0'WD_0\right)^{-1}. \qquad (11.23)\end{aligned}$$

If $W = W' = S_0^{-1}$, then this expression simplifies to (11.17). (See, for instance, Hamilton (1994) 14 (appendix) for more details.) ■

It is straightforward to show that the difference between the covariance matrix in (11.23) and $\left(D_0'S_0^{-1}D_0\right)^{-1}$ (as in (11.17)) is a positive semi-definite matrix: any linear combination of the parameters has a smaller variance if $W = S_0^{-1}$ is used as the weighting matrix.

All the expressions for the asymptotic distribution are supposed to be evaluated at the true parameter vector β_0, which is unknown. However, D_0 in (11.14) can be estimated by $\partial\bar{g}(\hat{\beta})/\partial\beta'$, where we use the point estimate instead of the true value of the parameter vector. In practice, this means plugging in the point estimates into the sample moment conditions and calculate the derivatives with respect to parameters (for instance, by a numerical method).

Similarly, S_0 in (11.13) can be estimated by, for instance, Newey-West's estimator of $\text{Cov}[\sqrt{T}\bar{g}(\hat{\beta})]$, once again using the point estimates in the moment conditions.

Example 11.13 *(Covariance matrix of 2SLS.) Define*

$$S_0 = \text{Cov}\left[\sqrt{T}\bar{g}\left(\beta_0\right)\right] = \text{Cov}\left(\frac{\sqrt{T}}{T}\sum_{t=1}^{T}z_t u_t\right)$$

$$D_0 = \text{plim}\,\frac{\partial\bar{g}(\beta_0)}{\partial\beta'} = \text{plim}\left(-\frac{1}{T}\sum_{t=1}^{T}z_t x_t'\right) = -\Sigma_{zx}.$$

This gives the asymptotic covariance matrix of $\sqrt{T}(\hat{\beta}-\beta_0)$

$$V = \left(D_0'S_0^{-1}D_0\right)^{-1} = \left(\Sigma_{zx}'S_0^{-1}\Sigma_{zx}\right)^{-1}.$$

11.6 Summary of GMM

$$\begin{aligned}
\text{Economic model}: &\ \mathrm{E}\, g(w_t, \beta_0) = \mathbf{0}_{q\times 1},\ \beta \text{ is } k \times 1 \\
\text{Sample moment conditions}: &\ \bar{g}(\beta) = \frac{1}{T}\sum_{t=1}^{T} g(w_t, \beta) \\
\text{Loss function}: &\ J = \bar{g}(\beta)' W \bar{g}(\beta) \\
\text{First order conditions}: &\ \mathbf{0}_{k\times 1} = \frac{\partial \bar{g}(\hat{\beta})' W \bar{g}(\hat{\beta})}{\partial \beta} = \left(\frac{\partial \bar{g}(\hat{\beta})}{\partial \beta'}\right)' W \bar{g}(\hat{\beta}) \\
\text{Consistency}: &\ \hat{\beta} \text{ is typically consistent if } \mathrm{E}\, g(w_t, \beta_0) = \mathbf{0} \\
\text{Define}: &\ S_0 = \mathrm{Cov}\left[\sqrt{T}\bar{g}\left(\beta_0\right)\right] \text{ and } D_0 = \operatorname{plim} \frac{\partial \bar{g}(\beta_0)}{\partial \beta'} \\
\text{Choose}: &\ W = S_0^{-1} \\
\text{Asymptotic distribution}: &\ \sqrt{T}(\hat{\beta} - \beta_0) \xrightarrow{d} N(\mathbf{0}_{k\times 1}, V), \text{ where } V = \left(D_0' S_0^{-1} D_0\right)^{-1}
\end{aligned}$$

11.7 Efficient GMM and Its Feasible Implementation

The efficient GMM (remember: for a *given* set of moment conditions) requires that we use $W = S_0^{-1}$, which is tricky since S_0 should be calculated by using the true (unknown) parameter vector. However, the following *two-stage procedure* usually works fine:

- First, estimate model with some (symmetric and positive definite) weighting matrix. The identity matrix is typically a good choice for models where the moment conditions are of the same order of magnitude (if not, consider changing the moment conditions). This gives consistent estimates of the parameters β. Then a consistent estimate $\hat{S}$ can be calculated (for instance, with Newey-West).
- Use the consistent $\hat{S}$ from the first step to define a new weighting matrix as $W = \hat{S}^{-1}$. The algorithm is run again to give asymptotically efficient estimates of β.
- Iterate at least once more. (You may want to consider iterating until the point estimates converge.)

Example 11.14 *(Implementation of 2SLS.) Under the classical 2SLS assumptions, there is no need for iterating since the efficient weighting matrix is $\Sigma_{zz}^{-1}/\sigma^2$. Only σ^2 depends*

on the estimated parameters, but this scaling factor of the loss function does not affect $\hat{\beta}_{2SLS}$.

One word of warning: if the number of parameters in the covariance matrix $\hat{S}$ is large compared to the number of data points, then $\hat{S}$ tends to be unstable (fluctuates a lot between the steps in the iterations described above) and sometimes also close to singular. The *saturation ratio* is sometimes used as an indicator of this problem. It is defined as the number of data points of the moment conditions (qT) divided by the number of estimated parameters (the k parameters in $\hat{\beta}$ and the unique $q(q+1)/2$ parameters in $\hat{S}$ if it is estimated with Newey-West). A value less than 10 is often taken to be an indicator of problems. A possible solution is then to impose restrictions on S, for instance, that the autocorrelation is a simple AR(1) and then estimate S using these restrictions (in which case you cannot use Newey-West, or course).

11.8 Testing in GMM

The result in (11.17) can be used to do *Wald tests of the parameter vector*. For instance, suppose we want to test the s linear restrictions that $R\beta_0 = r$ (R is $s \times k$ and r is $s \times 1$) then it must be the case that under null hypothesis

$$\sqrt{T}(R\hat{\beta} - r) \xrightarrow{d} N(\mathbf{0}_{s\times 1}, RVR'). \tag{11.24}$$

Remark 11.15 *(Distribution of quadratic forms.) If the* $n \times 1$ *vector* $x \sim N(0, \Sigma)$*, then* $x'\Sigma^{-1}x \sim \chi^2_n$.

From this remark and the continuous mapping theorem in Remark (5.10) it follows that, under the null hypothesis that $R\beta_0 = r$, the Wald test statistics is distributed as a χ^2_s variable

$$T(R\hat{\beta} - r)'\left(RVR'\right)^{-1}(R\hat{\beta} - r) \xrightarrow{d} \chi^2_s. \tag{11.25}$$

We might also want to *test the overidentifying restrictions*. The first order conditions (11.11) imply that k linear combinations of the q moment conditions are set to zero by solving for $\hat{\beta}$. Therefore, we have $q - k$ remaining overidentifying restrictions which should also be close to zero if the model is correct (fits the data). Under the null hypothesis that the moment conditions hold (so the overidentifying restrictions hold), we know that $\sqrt{T}\bar{g}(\beta_0)$ is a (scaled) sample average and therefore has (by a CLT) an asymptotic normal

distribution. It has a zero mean (the null hypothesis) and the covariance matrix in (11.12). In short,

$$\sqrt{T}\bar{g}\left(\beta_0\right) \xrightarrow{d} N\left(\mathbf{0}_{q\times 1}, S_0\right). \tag{11.26}$$

If would then perhaps be natural to expect that the quadratic form $T\bar{m}(\hat{\beta})'S_0^{-1}\bar{g}(\hat{\beta})$ should be converge in distribution to a χ^2_q variable. That is not correct, however, since $\hat{\beta}$ chosen is such a way that k linear combinations of the first order conditions always (in every sample) are zero. There are, in effect, only $q-k$ nondegenerate random variables in the quadratic form (see Davidson and MacKinnon (1993) 17.6 for a detailed discussion). The correct result is therefore that if we have used optimal weight matrix is used, $W = S_0^{-1}$, then

$$\sqrt{T}\bar{g}(\hat{\beta}) \to^d N\left(\mathbf{0}_{q\times 1}, \Psi_1\right), \text{ with} \tag{11.27}$$

$$\Psi_1 = S_0 - D_0\left(D_0'S_0^{-1}D_0\right)^{-1}D_0'. \tag{11.28}$$

This covariance matrix has reduced rank. It is therefore convenient to use the result that

$$T\bar{g}(\hat{\beta})'S_0^{-1}\bar{g}(\hat{\beta}) \xrightarrow{d} \chi^2_{q-k}, \text{ if } W = S_0^{-1}. \tag{11.29}$$

(A proof is given in the next section.) The left hand side equals T times of value of the loss function (11.8) evaluated at the point estimates, so we could equivalently write what is often called the *J test*

$$TJ(\hat{\beta}) \sim \chi^2_{q-k}, \text{ if } W = S_0^{-1}. \tag{11.30}$$

This also illustrates that with no overidentifying restrictions (as many moment conditions as parameters) there are, of course, no restrictions to test. Indeed, the loss function value is then always zero at the point estimates.

Example 11.16 *(Test of overidentifying assumptions in 2SLS.) In contrast to the IV method, 2SLS allows us to test overidentifying restrictions (we have more moment conditions than parameters, that is, more instruments than regressors). This is a test of whether the residuals are indeed uncorrelated with all the instruments. If not, the model should be rejected. It can be shown that test (11.30) is (asymptotically, at least) the same as the traditional (Sargan (1964), see Davidson (2000) 8.4) test of the overidentifying restrictions in 2SLS. In the latter, the fitted residuals are regressed on the instruments; TR^2 from that regression is χ^2 distributed with as many degrees of freedom as the number of overidentifying restrictions.*

Another test is to compare a restricted and a less restricted model, where we have used the optimal weighting matrix for the less restricted model in estimating both the less restricted and more restricted model (the weighting matrix is treated as a fixed matrix in the latter case). It can be shown that the test of the s restrictions (the "D test", similar in flavour to an LR test), is

$$T[J(\hat{\beta}^{restricted}) - J(\hat{\beta}^{less\ restricted})] \sim \chi^2_s, \text{ if } W = S_0^{-1}. \tag{11.31}$$

The weighting matrix is typically based on the unrestricted model. Note that (11.30) is a special case, since the model with allows q non-zero parameters (as many as the moment conditions) always attains $J = 0$, and that by imposing $s = q - k$ restrictions we get a restricted model.

11.9 GMM with Sub-Optimal Weighting Matrix*

When the optimal weighting matrix is not used, that is, when (11.16) does not hold, then the asymptotic covariance matrix of the parameters is given by (11.23) instead of the result in (11.17). That is,

$$\sqrt{T}(\hat{\beta} - \beta_0) \xrightarrow{d} N(\mathbf{0}_{k\times 1}, V_2), \text{ where } V_2 = \left(D_0'WD_0\right)^{-1} D_0'WS_0W'D_0 \left(D_0'WD_0\right)^{-1}. \tag{11.32}$$

The consistency property is not affected.

The test of the overidentifying restrictions (11.29) and (11.30) are not longer valid. Instead, the result is that

$$\sqrt{T}\bar{g}(\hat{\beta}) \to^d N\left(\mathbf{0}_{q\times 1}, \Psi_2\right), \text{ with} \tag{11.33}$$

$$\Psi_2 = [I - D_0\left(D_0'WD_0\right)^{-1} D_0'W]S_0[I - D_0\left(D_0'WD_0\right)^{-1} D_0'W]'. \tag{11.34}$$

This covariance matrix has rank $q - k$ (the number of overidentifying restriction). This distribution can be used to test hypotheses about the moments, for instance, that a particular moment condition is zero.

Proof. (Sketch of proof of (11.33)–(11.34)) Use (11.20) in (11.18) to get

$$\begin{aligned}\sqrt{T}\bar{g}(\hat{\beta}) &= \sqrt{T}\bar{g}(\beta_0) + \sqrt{T}\frac{\partial \bar{g}(\beta_1)}{\partial \beta'}\Gamma\bar{g}(\beta_0) \\ &= \left[I + \frac{\partial \bar{g}(\beta_1)}{\partial \beta'}\Gamma\right]\sqrt{T}\bar{g}(\beta_0).\end{aligned}$$

The term in brackets has a probability limit, which by (11.21) equals $I - D_0 \left(D_0' W D_0\right)^{-1} D_0' W$. Since $\sqrt{T}\bar{g}(\beta_0) \to^d N\left(\mathbf{0}_{q\times 1}, S_0\right)$ we get (11.33). ■

Remark 11.17 *If the $n \times 1$ vector $X \sim N(0, \Sigma)$, where Σ has rank $r \leq n$ then $Y = X'\Sigma^+ X \sim \chi^2_r$ where Σ^+ is the pseudo inverse of Σ.*

Remark 11.18 *The symmetric Σ can be decomposed as $\Sigma = Z\Lambda Z'$ where Z are the orthogonal eigenvectors ($Z'Z = I$) and Λ is a diagonal matrix with the eigenvalues along the main diagonal. The pseudo inverse can then be calculated as $\Sigma^+ = Z\Lambda^+ Z'$, where*

$$\Lambda^+ = \begin{bmatrix} \Lambda_{11}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix},$$

with the reciprocals of the non-zero eigen values along the principal diagonal of Λ_{11}^{-1}.

This remark and (11.34) implies that the test of overidentifying restrictions (Hansen's J statistics) analogous to (11.29) is

$$T\bar{g}(\hat{\beta})'\Psi_2^+ \bar{g}(\hat{\beta}) \xrightarrow{d} \chi^2_{q-k}. \tag{11.35}$$

It requires calculation of a generalized inverse (denoted by superscript $^+$), but this is fairly straightforward since Ψ_2 is a symmetric matrix. It can be shown that this simplifies to (11.29) when the optimal weighting matrix is used (see below for a proof).

Proof. (Sketch of proof of (11.35)) From the proof of (11.33)–(11.34), notice $W = S_0^{-1}$ gives (in the limit)

$$\sqrt{T}\bar{g}(\hat{\beta}) = \left[I - D_0 \left(D_0' S_0^{-1} D_0\right)^{-1} D_0' S_0^{-1}\right] \sqrt{T}\bar{g}(\beta_0).$$

Premultiply $\bar{g}(\beta_0)$ by $S_0^{1/2} S_0^{-1/2}$ and then multiply both sides by $S_0^{-1/2}$ to get

$$\begin{aligned} \sqrt{T} S_0^{-1/2} \bar{g}(\hat{\beta}) &= \Lambda_0 \sqrt{T} S_0^{-1/2} \bar{g}(\beta_0), \text{ where} \\ \Lambda_0 &= I - S_0^{-1/2} D_0 \left(D_0' S_0^{-1} D_0\right)^{-1} D_0' S_0^{-1/2}. \end{aligned} \tag{(*)}$$

It is clear that Λ_0 is symmetric and that $\Lambda_0 \Lambda_0 = \Lambda_0$ (it is idempotent). It can also be shown that the rank is $q - k$. The square of the left hand side of (*) must equal the square of the right hand side, so (using the fact that $\Lambda_0 \Lambda_0' = \Lambda_0$), so

$$T\bar{g}(\hat{\beta})' S_0^{-1} \bar{g}(\hat{\beta}) = Z'\Lambda_0 Z, \text{ where } Z = \sqrt{T} S_0^{-1/2} \bar{g}(\beta_0). \tag{(**)}$$

Notice that $Z \sim N(0, I)$, so we can use the rule that says that, if $Z \sim N(0, I)$ then $Z'\Lambda_0 Z \sim \chi^2_{q-k}$ (provided Λ_0 is symmetric, idempotent and of rank $q-k$). This clearly means that also the left hand side of (**) has a χ^2_{q-k} distribution. ■

Remark 11.19 *((11.32) and (11.34) when $W = S_0^{-1}$) when $W = S_0^{-1}$, then (11.32) gives $V_2 = \left(D_0' S_0^{-1} D_0\right)^{-1}$, which is the same as in (11.17). However, when $W = S_0^{-1}$, then (11.34) gives $\Psi_2 = S - D\left(D'S^{-1}D\right)^{-1} D'$. Actually, using this in (11.35) gives (at least asymptotically) the same result as using (11.29).*

11.10 GMM without a Loss Function*

Suppose we sidestep the whole optimization issue and instead specify k linear combinations (as many as there are parameters) of the q moment conditions directly. That is, instead of the first order conditions (11.11) we postulate that the estimator should solve

$$\mathbf{0}_{k\times 1} = \underbrace{A}_{k\times q}\underbrace{\bar{g}(\hat{\beta})}_{q\times 1} \ (\hat{\beta} \text{ is } k \times 1). \tag{11.36}$$

The matrix A is chosen by the researcher and it must have rank k (lower rank means that we effectively have too few moment conditions to estimate the k parameters in β). If A is random, then it should have a finite probability limit A_0 (also with rank k).

One case when this approach makes particular sense is when we want to use a subset of the moment conditions to estimate the parameters (some columns in A are then filled with zeros), but we want to study the distribution of all the moment conditions.

By comparing (11.11) and (11.36) we see that A plays the same role as $[\partial \bar{g}(\hat{\beta})/\partial \beta']' W$, but with the difference that A is chosen and not allowed to depend on the parameters. In the asymptotic distribution, it is the probability limit of these matrices that matter, so we can actually substitute A_0 for $D_0' W$ in the proof of the asymptotic distribution. The covariance matrix in (11.32) then becomes

$$\begin{aligned} V_3 &= (A_0 D_0)^{-1} A_0 S_0 [(A_0 D_0)^{-1} A_0]' \\ &= (A_0 D_0)^{-1} A_0 S_0 A_0' [(A_0 D_0)^{-1}]', \end{aligned} \tag{11.37}$$

which can be used to test hypotheses about the parameters.

Similarly, the asymptotic distribution of the moment conditions is

$$\sqrt{T}\bar{g}(\hat{\beta}) \to^d N\left(\mathbf{0}_{q\times 1}, \Psi_3\right), \text{ with} \tag{11.38}$$

$$\Psi_3 = [I - D_0 (A_0 D_0)^{-1} A_0] S_0 [I - D_0 (A_0 D_0)^{-1} A_0]', \tag{11.39}$$

where Ψ_3 has reduced rank. As before, this covariance matrix can be used to construct both t type and χ^2 tests of the moment conditions. For instance, the test of overidentifying restrictions (Hansen's J statistics)

$$T\bar{g}(\hat{\beta})'\Psi_3^+\bar{g}(\hat{\beta}) \xrightarrow{d} \chi^2_{q-k}, \tag{11.40}$$

where Ψ_3^+ is a generalized inverse of Ψ_3.

11.11 Simulated Moments Estimator*

Reference: Ingram and Lee (1991)

It sometimes happens that it is not possible to calculate the theoretical moments in GMM explicitly. For instance, suppose we want to match the variance of the model with the variance of data

$$\mathrm{E}\, g(w_t, \beta_0) = 0, \text{ where} \tag{11.41}$$

$$g(w_t, \beta) = (w_t - \mu)^2 - \text{Var\_in\_model}(\beta), \tag{11.42}$$

but the model is so non-linear that we cannot find a closed form expression for Var_of_model(β_0). Similarly, we could match a covariance of

The SME involves *(i)* drawing a set of random numbers for the stochastic shocks in the model; *(ii)* for a given set of parameter values generate a model simulation with T_{sim} observations, calculating the moments and using those instead of Var_of_model(β_0) (or similarly for other moments), which is then used to evaluate the loss function J_T. This is repeated for various sets of parameter values until we find the one which minimizes J_T.

Basically all GMM results go through, but the covariance matrix should be scaled up with $1 + T/T_{sim}$, where T is the sample length. Note that the same sequence of random numbers should be reused over and over again (as the parameter values are changed).

Example 11.20 *Suppose w_t has two elements, x_t and y_t, and that we want to match both variances and also the covariance. For simplicity, suppose both series have zero means.*

Then we can formulate the moment conditions

$$g(x_t, y_t, \beta) = \begin{bmatrix} x_t^2 - Var(x)\_in\_model(\beta) \\ y_t^2 - Var(y)\_in\_model(\beta) \\ x_t y_t - Cov(x,y)\_in\_model(\beta) \end{bmatrix}. \tag{11.43}$$

Chapter 12

Examples and Applications of GMM

12.1 GMM and Classical Econometrics: Examples

12.1.1 The LS Estimator (General)

The model is

$$y_t = x_t'\beta_0 + u_t, \tag{12.1}$$

where β is a $k \times 1$ vector.

The k moment conditions are

$$\bar{g}(\beta) = \frac{1}{T}\sum_{t=1}^{T} x_t(y_t - x_t'\beta) = \frac{1}{T}\sum_{t=1}^{T} x_t y_t - \frac{1}{T}\sum_{t=1}^{T} x_t x_t'\beta. \tag{12.2}$$

The point estimates are found by setting all moment conditions to zero (the model is exactly identified), $\bar{g}(\beta) = \mathbf{0}_{k\times 1}$, which gives

$$\hat{\beta} = \left(\frac{1}{T}\sum_{t=1}^{T} x_t x_t'\right)^{-1} \frac{1}{T}\sum_{t=1}^{T} x_t y_t \beta. \tag{12.3}$$

If we define

$$S_0 = \text{ACov}\left[\sqrt{T}\bar{g}(\beta_0)\right] = \text{ACov}\left(\frac{\sqrt{T}}{T}\sum_{t=1}^{T} x_t u_t\right) \tag{12.4}$$

$$D_0 = \text{plim}\,\frac{\partial \bar{g}(\beta_0)}{\partial \beta'} = \text{plim}\left(-\frac{1}{T}\sum_{t=1}^{T} x_t x_t'\right) = -\Sigma_{xx}. \tag{12.5}$$

then the asymptotic covariance matrix of $\sqrt{T}(\hat{\beta}-\beta_0)$

$$V_{LS} = \left(D_0' S_0^{-1} D_0\right)^{-1} = \left(\Sigma_{xx}' S_0^{-1} \Sigma_{xx}\right)^{-1} = \Sigma_{xx}^{-1} S_0 \Sigma_{xx}^{-1}. \tag{12.6}$$

We can then either try to estimate S_0 by Newey-West, or make further assumptions to simplify S_0 (see below).

12.1.2 The IV/2SLS Estimator (General)

The model is (12.1), but we use an IV/2SLS method. The q moment conditions (with $q \geq k$) are

$$\bar{g}(\beta) = \frac{1}{T}\sum_{t=1}^{T} z_t(y_t - x_t'\beta) = \frac{1}{T}\sum_{t=1}^{T} z_t y_t - \frac{1}{T}\sum_{t=1}^{T} z_t x_t'\beta. \tag{12.7}$$

The loss function is (for some positive definite weighting matrix W, not necessarily the optimal)

$$\bar{g}(\beta)' W \bar{g}(\beta) = \left[\frac{1}{T}\sum_{t=1}^{T} z_t(y_t - x_t'\beta)\right]' W \left[\frac{1}{T}\sum_{t=1}^{T} z_t(y_t - x_t'\beta)\right], \tag{12.8}$$

and the k first order conditions, $(\partial \bar{g}(\hat{\beta})/\partial\beta')' W \bar{g}(\hat{\beta}) = 0$, are

$$\begin{aligned}
\mathbf{0}_{k\times 1} &= \left[\frac{\partial}{\partial\beta'}\frac{1}{T}\sum_{t=1}^{T} z_t(y_t - x_t'\hat{\beta})\right]' W \frac{1}{T}\sum_{t=1}^{T} z_t(y_t - x_t'\hat{\beta}) \\
&= \left[-\frac{1}{T}\sum_{t=1}^{T} z_t x_t'\right]' W \frac{1}{T}\sum_{t=1}^{T} z_t(y_t - x_t'\hat{\beta}) \\
&= -\hat{\Sigma}_{xz} W(\hat{\Sigma}_{zy} - \hat{\Sigma}_{zx}\hat{\beta}).
\end{aligned} \tag{12.9}$$

We solve for $\hat{\beta}$ as

$$\hat{\beta} = \left(\hat{\Sigma}_{xz} W \hat{\Sigma}_{zx}\right)^{-1} \hat{\Sigma}_{xz} W \hat{\Sigma}_{zy}. \tag{12.10}$$

Define

$$S_0 = \text{ACov}\left[\sqrt{T}\bar{g}(\beta_0)\right] = \text{ACov}\left(\frac{\sqrt{T}}{T}\sum_{t=1}^{T} z_t u_t\right) \tag{12.11}$$

$$D_0 = \text{plim}\,\frac{\partial \bar{g}(\beta_0)}{\partial\beta'} = \text{plim}\left(-\frac{1}{T}\sum_{t=1}^{T} z_t x_t'\right) = -\Sigma_{zx}. \tag{12.12}$$

This gives the asymptotic covariance matrix of $\sqrt{T}(\hat{\beta}-\beta_0)$

$$V = \left(D_0' S_0^{-1} D_0\right)^{-1} = \left(\Sigma_{zx}' S_0^{-1} \Sigma_{zx}\right)^{-1}, \tag{12.13}$$

assuming that we have used $W = S_0^{-1}$.

When the model is exactly identified ($q = k$), then we can make some simplifications since $\hat{\Sigma}_{xz}$ is then invertible. This is the case of the classical IV estimator. We get

$$\hat{\beta} = \hat{\Sigma}_{zx}^{-1} \hat{\Sigma}_{zy} \text{ and } V = \Sigma_{zx}^{-1} S_0 \left(\Sigma_{zx}'\right)^{-1} \text{ if } q = k. \tag{12.14}$$

(Use the rule $(ABC)^{-1} = C^{-1}B^{-1}A^{-1}$ to show this.)

12.1.3 Classical LS Assumptions

Reference: Greene (2000) 9.4 and Hamilton (1994) 8.2.

This section returns to the LS estimator in Section (12.1.1) in order to highlight the classical LS assumptions that give the variance matrix $\sigma^2 \Sigma_{xx}^{-1}$.

We allow the regressors to be stochastic, but require that x_t is independent of all u_{t+s} and that u_t is iid. It rules out, for instance, that u_t and x_{t-2} are correlated and also that the variance of u_t depends on x_t. Expand the expression for S_0 as

$$\begin{aligned} S_0 &= \mathrm{E}\left(\frac{\sqrt{T}}{T}\sum_{t=1}^{T} x_t u_t\right)\left(\frac{\sqrt{T}}{T}\sum_{t=1}^{T} u_t x_t'\right) \\ &= \frac{1}{T}\mathrm{E}\left(... + x_{s-1}u_{s-1} + x_s u_s + ...\right)\left(... + u_{s-1}x_{s-1}' + u_s x_s' + ...\right). \end{aligned} \tag{12.15}$$

Note that

$$\begin{aligned} \mathrm{E}\, x_{t-s}u_{t-s}u_t x_t' &= \mathrm{E}\, x_{t-s}x_t' \,\mathrm{E}\, u_{t-s}u_t \text{ (since } u_t \text{ and } x_{t-s} \text{ independent)} \\ &= \begin{cases} 0 \text{ if } s \neq 0 \text{ (since } \mathrm{E}u_{s-1}u_s = 0 \text{ by iid } u_t) \\ \mathrm{E}\, x_t x_t' \,\mathrm{E}\, u_t u_t \text{ else.} \end{cases} \end{aligned} \tag{12.16}$$

This means that all cross terms (involving different observations) drop out and that we

can write

$$S_0 = \frac{1}{T}\sum_{t=1}^{T} \mathrm{E}\, x_t x_t' \,\mathrm{E}\, u_t^2 \tag{12.17}$$

$$= \sigma^2 \frac{1}{T}\mathrm{E}\sum_{t=1}^{T} x_t x_t' \text{ (since } u_t \text{ is iid and } \sigma^2 = \mathrm{E}\, u_t^2) \tag{12.18}$$

$$= \sigma^2 \Sigma_{xx}. \tag{12.19}$$

Using this in (12.6) gives

$$V = \sigma^2 \Sigma_{xx}^{-1}. \tag{12.20}$$

12.1.4 Almost Classical LS Assumptions: White's Heteroskedasticity.

Reference: Greene (2000) 12.2 and Davidson and MacKinnon (1993) 16.2.

The only difference compared with the classical LS assumptions is that u_t is now allowed to be heteroskedastic, but this heteroskedasticity is not allowed to depend on the moments of x_t. This means that (12.17) holds, but (12.18) does not since $\mathrm{E}u_t^2$ is not the same for all t.

However, we can still simplify (12.17) a bit more. We assumed that $\mathrm{E}x_t x_t'$ and $\mathrm{E}u_t^2$ (which can both be time varying) are not related to each other, so we could perhaps multiply $\mathrm{E}\, x_t x_t'$ by $\Sigma_{t=1}^{T} \mathrm{E}\, u_t^2/T$ instead of by $\mathrm{E}\, u_t^2$. This is indeed true asymptotically—where any possible "small sample" relation between $\mathrm{E}\, x_t x_t'$ and $\mathrm{E}\, u_t^2$ must wash out due to the assumptions of independence (which are about population moments).

In large samples we therefore have

$$\begin{aligned} S_0 &= \left(\frac{1}{T}\sum_{t=1}^{T} \mathrm{E}\, u_t^2\right)\left(\frac{1}{T}\sum_{t=1}^{T} \mathrm{E}\, x_t x_t'\right) \\ &= \left(\frac{1}{T}\sum_{t=1}^{T} \mathrm{E}\, u_t^2\right)\left(\mathrm{E}\,\frac{1}{T}\sum_{t=1}^{T} x_t x_t'\right) \\ &= \omega^2 \Sigma_{xx}, \end{aligned} \tag{12.21}$$

where ω^2 is a scalar. This is very similar to the classical LS case, except that ω^2 is the average variance of the residual rather than the constant variance. In practice, the estimator of ω^2 is the same as the estimator of σ^2, so we can actually apply the standard LS formulas in this case.

This is the motivation for why White's test for heteroskedasticity makes sense: if the

heteroskedasticity is not correlated with the regressors, then the standard LS formula is correct (provided there is no autocorrelation).

12.1.5 Estimating the Mean of a Process

Suppose u_t is heteroskedastic, but not autocorrelated. In the regression $y_t = \alpha + u_t$, $x_t = z_t = 1$. This is a special case of the previous example, since $\mathrm{E}\, u_t^2$ is certainly unrelated to $\mathrm{E}\, x_t x_t' = 1$ (since it is a constant). Therefore, the LS covariance matrix is the correct variance of the sample mean as an estimator of the mean, even if u_t are heteroskedastic (provided there is no autocorrelation).

12.1.6 The Classical 2SLS Assumptions*

Reference: Hamilton (1994) 9.2.

The classical 2SLS case assumes that z_t is independent of all u_{t+s} and that u_t is iid. The covariance matrix of the moment conditions are

$$S_0 = \mathrm{E}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} z_t u_t\right)\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} u_t z_t'\right), \tag{12.22}$$

so by following the same steps in (12.16)-(12.19) we get $S_0 = \sigma^2 \Sigma_{zz}$.The optimal weighting matrix is therefore $W = \Sigma_{zz}^{-1}/\sigma^2$ (or $(Z'Z/T)^{-1}/\sigma^2$ in matrix form). We use this result in (12.10) to get

$$\hat{\beta}_{2SLS} = \left(\hat{\Sigma}_{xz}\hat{\Sigma}_{zz}^{-1}\hat{\Sigma}_{zx}\right)^{-1}\hat{\Sigma}_{xz}\hat{\Sigma}_{zz}^{-1}\hat{\Sigma}_{zy}, \tag{12.23}$$

which is the classical 2SLS estimator.

Since this GMM is efficient (for a given set of moment conditions), we have established that 2SLS uses its given set of instruments in the efficient way—provided the classical 2SLS assumptions are correct. Also, using the weighting matrix in (12.13) gives

$$V = \left(\Sigma_{xz}\frac{1}{\sigma^2}\Sigma_{zz}^{-1}\Sigma_{zx}\right)^{-1}. \tag{12.24}$$

12.1.7 Non-Linear Least Squares

Consider the non-linear regression

$$y_t = F(x_t;\beta_0) + \varepsilon_t, \tag{12.25}$$

where $F(x_t;\beta_0)$ is a potentially non-linear equation of the regressors x_t, with a $k \times 1$ vector of parameters β_0. The non-linear least squares (NLS) approach is minimize the sum of squared residuals, that is, to solve

$$\hat{\beta} = \arg\min \sum_{t=1}^{T} [y_t - F(x_t;\beta)]^2. \tag{12.26}$$

To express this as a GMM problem, use the first order conditions for (12.26) as moment conditions

$$\bar{g}(\beta) = -\frac{1}{T}\sum_{t=1}^{T} \frac{\partial F(x_t;\beta)}{\partial \beta} [y_t - F(x_t;\beta)]. \tag{12.27}$$

The model is then exactly identified so the point estimates are found by setting all moment conditions to zero , $\bar{g}(\beta) = \mathbf{0}_{k\times 1}$.

As usual, $S_0 = \text{Cov}[\sqrt{T}\bar{g}(\beta_0)]$, while the Jacobian is

$$\begin{aligned} D_0 &= \text{plim}\, \frac{\partial \bar{g}(\beta_0)}{\partial \beta'} \\ &= \text{plim}\, \frac{1}{T}\sum_{t=1}^{T} \frac{\partial F(x_t;\beta)}{\partial \beta} \frac{\partial F(x_t;\beta)}{\partial \beta'} - \text{plim}\, \frac{1}{T}\sum_{t=1}^{T} [y_t - F(x_t;\beta)] \frac{\partial^2 F(x_t;\beta)}{\partial \beta \partial \beta'}. \end{aligned} \tag{12.28}$$

Example 12.1 *(With two parameters) With $\beta = [\beta_1, \beta_2]'$ we have*

$$\frac{\partial F(x_t;\beta)}{\partial \beta} = \begin{bmatrix} \partial F(x_t;\beta)/\partial \beta_1 \\ \partial F(x_t;\beta)/\partial \beta_2 \end{bmatrix}, \frac{\partial F(x_t;\beta)}{\partial \beta'} = \begin{bmatrix} \partial F(x_t;\beta)/\partial \beta_1 & \partial F(x_t;\beta)/\partial \beta_2 \end{bmatrix}.$$

The moment conditions are

$$\bar{g}(\beta) = -\frac{1}{T}\sum_{t=1}^{T} \begin{bmatrix} \partial F(x_t;\beta)/\partial \beta_1 \\ \partial F(x_t;\beta)/\partial \beta_2 \end{bmatrix} [y_t - F(x_t;\beta)],$$

which is a 2×1 vector. Notice that the outer product of the gradient (first term) in (12.28) is a 2×2 matrix. Similarly, the matrix with the second derivatives (the Hessian) is also a 2×2 matrix

$$\frac{\partial^2 F(x_t;\beta)}{\partial \beta \partial \beta'} = \begin{bmatrix} \frac{\partial^2 F(x_t;\beta)}{\partial \beta_1 \partial \beta_1} & \frac{\partial^2 F(x_t;\beta)}{\partial \beta_1 \partial \beta_2} \\ \frac{\partial^2 F(x_t;\beta)}{\partial \beta_2 \partial \beta_1} & \frac{\partial^2 F(x_t;\beta)}{\partial \beta_2 \partial \beta_2} \end{bmatrix}.$$

Example 12.2 *(Linear regression function as a special case) When $F(x_t;\beta) = x_t'\beta$, then $\partial F(x_t;\beta)/\partial \beta = x_t$, so the moment conditions are $\bar{g}(\beta) = -\sum_{t=1}^{T} x_t \left(y_t - x_t'\beta\right)/T$. Since the second derivatives are zero, (12.28) becomes $D_0 = \text{plim}\sum_{t=1}^{T} x_t x_t'/T$, which is the same in the LS case (except possibly for the sign of D_0, but that is of no consequence since it is only the square of D_0 that matters.)*

Example 12.3 *(Logistic smooth transition regression) Let $G(z)$ be a logistic (increasing but "S-shaped") function*

$$G(z) = \frac{1}{1 + \exp[-\gamma(z - c)]},$$

where the parameter c is the central location (where $G(z) = 1/2$) and $\gamma > 0$ determines the steepness of the function (a high γ implies that the function goes quickly from 0 to 1 around $z = c$.) See Figure 13.14 for an illustration. A logistic smooth transition regression is

$$\begin{aligned} y_t &= \underbrace{\left\{[1 - G(z_t)]\,\beta_1' + G(z_t)\beta_2'\right\} x_t}_{F(x_t;\beta) \text{ in (12.25)}} + \varepsilon_t \\ &= [1 - G(z_t)]\,\beta_1' x_t + G(z_t)\beta_2' x_t + \varepsilon_t. \end{aligned}$$

The regression coefficients vary smoothly with z_t: from β_1 at low values of z_t to β_2 at high values of z_t. See Figure 13.14 for an illustration. The parameter vector ($\gamma, c, \beta_1, \beta_2$—called just β in (12.25)) is easily estimated by NLS by concentrating the loss function: optimize (numerically) over (γ, c) and let (for each value of (γ, c)) the parameters (β_1, β_2) be the OLS coefficients on the vector of "regressors" $([1 - G(z_t)]\,x_t, G(z_t)x_t)$. The most common application of this model is obtained by letting $x_t = y_{t-s}$ (this is the LSTAR model—logistic smooth transition auto regression model), see Franses and van Dijk (2000).

12.1.8 Moment Conditions with Spuriously Extended Sample 1

One way to handle unbalanced panels (when there is more data on sone variables than on others), is to artificially expand the sample and then interact the moment conditions with a dummy variable to pick out the correct subsample. This example illustrates how and why that works. To keep it simple, the example discusses the case of estimating a sample mean of x_t—for which we have data over the sample $t = 1$ to τ and the sample is artificially extended with $T - \tau$ data points.

To estimate the mean we specify the moment condition

$$g_t = d_t\,(x_t - \mu)\,, \text{ with } d_t = \begin{cases} 1 & t = 1, .., \tau \\ 0 & t = \tau + 1, ..., T \end{cases} \tag{12.29}$$

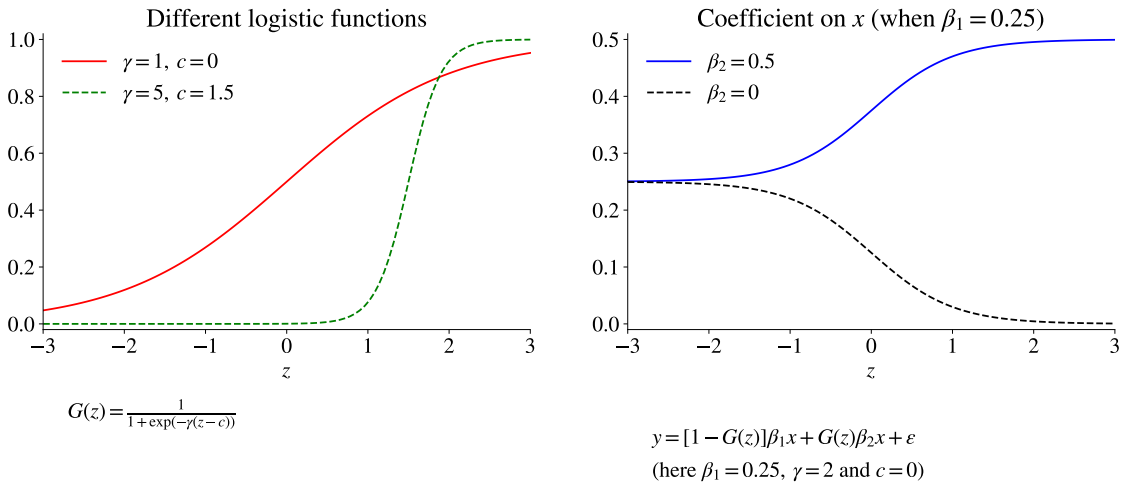


Figure 12.1: Logistic function and the effective slope coefficient in a Logistic smooth transition regression

so the moment conditions look like

$$\begin{bmatrix} x_1 - \mu \\ \vdots \\ x_\tau - \mu \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{12.30}$$

With iid data, the variance $S_0 = \text{Cov}[\sqrt{T}\bar{g}(\beta_0)] = \text{Cov}(g_t)$, that is,

$$S_0 = \frac{\tau\sigma^2 + (T-\tau)0}{T} = \frac{\tau}{T}\sigma^2, \tag{12.31}$$

where $\sigma^2 = \text{Var}(x_t)$. Also, the Jacobian (plim $\partial\bar{g}(\beta_0)/\partial\beta'$) is

$$D_0 = \frac{-\tau}{T}. \tag{12.32}$$

Combining gives

$$\begin{aligned}\sqrt{T}(\hat{\mu}-\mu) &\xrightarrow{d} N(0,V), \text{ where}\\ V &= \left(\frac{-\tau}{T}\times\left(\frac{\tau}{T}\sigma^2\right)^{-1}\times\frac{-\tau}{T}\right)^{-1}\\ &= \sigma^2\times\frac{T}{\tau}.\end{aligned} \tag{12.33}$$

Therefore,

$$\text{Var}(\hat{\mu}) = V/T = \frac{\sigma^2}{\tau}, \tag{12.34}$$

which is the correct result—the artificial extension of the sample does not lead to a spuriously low uncertainty. This demonstrates that the aritificial extension of the sample actually does no harm: the inference based on standard GMM formulas is still correct.

12.1.9 Moment Conditions with Spuriously Extended Sample 2 (Dummies for Missing Values)

Consider the simple regression equation

$$y_t = bx_t + \varepsilon_t \tag{12.35}$$

and suppose the sample length is T, but only the first τ observations have full data, while the last $T-\tau$ observations include some missing values. (Putting these observations last is just a matter of convenience.)

Suppose we prune ("excise") the sample by simply skipping the observations with missing values. Under the standard iid assumptions, we then have that the LS estimate ($\hat{b}$) is distributed as

$$\sqrt{T}(\hat{b}-b_0) \xrightarrow{d} N\left(\mathbf{0}_{k\times 1}, V\right), \tag{12.36}$$

where the covariance matrix is

$$V = \sigma^2\left(\text{plim}\,\frac{1}{\tau}\sum_{t=1}^{\tau}x_t x_t\right)^{-1} \text{ and } \sigma^2 = \text{plim}\,\frac{1}{\tau}\sum_{t=1}^{\tau}\varepsilon_t^2. \tag{12.37}$$

Instead, suppose we use all T observations, but let $d_t = 1$ if there is data for period t and zero otherwise. This gives the sample moment condition

$$\bar{g} = \frac{1}{T}\sum_{t=1}^{T} d_t x_t\left(y_t - bx_t\right) \tag{12.38}$$

The Jacobian is

$$D_0 = -\operatorname{plim}\frac{1}{T}\sum_{t=1}^{T} d_t x_t x_t \tag{12.39}$$

and the covariance of the moment conditions (under the standard iid assumptions)

$$S_0 = \operatorname{plim}\frac{1}{T}\sum_{t=1}^{T} d_t x_t x_t d_t \varepsilon_t^2 = s^2 \operatorname{plim}\frac{1}{T}\sum_{t=1}^{T} d_t x_t x_t, \text{ where } s^2 = \operatorname{plim}\frac{1}{T}\sum_{t=1}^{T} d_t \varepsilon_t^2. \tag{12.40}$$

Combining as in (12.6) gives the covariance matrix

$$V^b = s^2\left(\frac{1}{T}\sum_{t=1}^{T} d_t x_t x_t\right)^{-1}. \tag{12.41}$$

To see that this is the same as in (12.37), notice that

$$\sum_{t=1}^{T} d_t x_t x_t = \sum_{t=1}^{\tau} x_t x_t, \text{ and} \tag{12.42}$$

$$s^2 = \frac{1}{T}\sum_{t=1}^{T} d_t \varepsilon_t^2 = \frac{1}{T}\sum_{t=1}^{\tau} \varepsilon_t^2 = \frac{\tau}{T}\sigma^2.$$

Using this in (12.40)–(12.41) gives

$$V^b = \frac{\tau}{T}\sigma^2\left(\frac{1}{T}\sum_{t=1}^{\tau} x_t x_t\right)^{-1} = \sigma^2\left(\frac{1}{\tau}\sum_{t=1}^{\tau} x_t x_t\right)^{-1}.$$

which is the same as in (12.37). This makes a lot of sense since the dummy approach is just about nullifying the effect of the periods with missing values. In a sense this makes the Jacobian too small, but that is compensated for by making S_0 too large. This demonstrates that the estimation could be done in either way.

12.2 Identification of Systems of Simultaneous Equations

Reference: Greene (2000) 16.1–3

This section shows how the GMM moment conditions can be used to understand if the parameters in a system of simultaneous equations are identified or not.

The structural model (form) is

$$Fy_t + Gz_t = u_t, \tag{12.43}$$

where y_t is a vector of endogenous variables, z_t a vector of predetermined (exogenous) variables, F is a square matrix, and G is another matrix.[1] We can write the jth equation of the structural form (12.43) as

$$y_{jt} = x_t'\beta + u_{jt}, \tag{12.44}$$

where x_t contains the endogenous and exogenous variables that enter the jth equation with non-zero coefficients, that is, subsets of y_t and z_t.

We want to estimate β in (12.44). Least squares is inconsistent if some of the regressors are endogenous variables (in terms of (12.43), this means that the jth row in F contains at least one additional non-zero element apart from coefficient on y_{jt}). Instead, we use IV/2SLS. By assumption, the structural model summarizes all relevant information for the endogenous variables y_t. This implies that the only useful instruments are the variables in z_t. (A valid instrument is uncorrelated with the residuals, but correlated with the regressors.) The moment conditions for the jth equation are then

$$\mathrm{E}\, z_t \left(y_{jt} - x_t'\beta\right) = \mathbf{0} \text{ with sample moment conditions } \frac{1}{T}\sum_{t=1}^{T} z_t \left(y_{jt} - x_t'\beta\right) = \mathbf{0}. \tag{12.45}$$

If there are as many moment conditions as there are elements in β, then this equation is *exactly identified*, so the sample moment conditions can be inverted to give the Instrumental variables (IV) estimator of β. If there are more moment conditions than elements in β, then this equation is *overidentified* and we must devise some method for weighting the different moment conditions. This is the 2SLS method. Finally, when there are fewer moment conditions than elements in β, then this equation is *unidentified*, and we cannot hope to estimate the structural parameters of it.

We can partition the vector of regressors in (12.44) as $x_t' = [\tilde{z}_t', \tilde{y}_t']$, where y_{1t} and z_{1t} are the subsets of z_t and y_t respectively, that enter the right hand side of (12.44). Partition z_t conformably $z_t' = [\tilde{z}_t', z_t^{*\prime}]$, where z_t^* are the exogenous variables that do not enter (12.44). We can then rewrite the moment conditions in (12.45) as

$$\mathrm{E}\begin{bmatrix} \tilde{z}_t \\ z_t^* \end{bmatrix}\left(y_{jt} - \begin{bmatrix} \tilde{z}_t \\ \tilde{y}_t \end{bmatrix}'\beta\right) = \mathbf{0}. \tag{12.46}$$

[1] By premultiplying with F^{-1} and rearranging we get the reduced form $y_t = \Pi z_t + \varepsilon_t$, with $\Pi = -F^{-1}$ and $\mathrm{Cov}(\varepsilon_t) = F^{-1}\mathrm{Cov}(u_t)(F^{-1})'$.

$$\begin{aligned} y_{jt} &= -G_j \tilde{z}_t - F_j \tilde{y}_t + u_{jt} \\ &= x_t'\beta + u_{jt}, \text{ where } x_t' = \left[\tilde{z}_t', \tilde{y}_t'\right], \end{aligned} \tag{12.47}$$

This shows that we need at least as many elements in z_t^* as in $\tilde{y}_t$ to have this equations identified, which confirms the old-fashioned rule of thumb: *there must be at least as many excluded exogenous variables (z_t^*) as included endogenous variables ($\tilde{y}_t$) to have the equation identified.*

This section has discussed identification of structural parameters when 2SLS/IV, one equation at a time, is used. There are other ways to obtain identification, for instance, by imposing restrictions on the covariance matrix. See, for instance, Greene (2000) 16.1-3 for details.

Example 12.4 *(Supply and Demand. Reference: GR 16, Hamilton 9.1.) Consider the simplest simultaneous equations model for supply and demand on a market. Supply is*

$$q_t = \gamma p_t + u_t^s, \ \gamma > 0,$$

and demand is

$$q_t = \beta p_t + \alpha A_t + u_t^d, \ \beta < 0,$$

where A_t is an observable exogenous demand shock (perhaps income). The only meaningful instrument is A_t. From the supply equation we then get the moment condition

$$\mathrm{E}\, A_t \left(q_t - \gamma p_t\right) = 0,$$

which gives one equation in one unknown, γ. The supply equation is therefore exactly identified. In contrast, the demand equation is unidentified, since there is only one (meaningful) moment condition

$$\mathrm{E}\, A_t \left(q_t - \beta p_t - \alpha A_t\right) = 0,$$

but two unknowns (β and α).

Example 12.5 *(Supply and Demand: overidentification.) If we change the demand equation in Example 12.4 to*

$$q_t = \beta p_t + \alpha A_t + bB_t + u_t^d, \ \beta < 0.$$

There are now two moment conditions for the supply curve (since there are two useful instruments)

$$\mathrm{E}\begin{bmatrix} A_t\left(q_t-\gamma p_t\right) \\ B_t\left(q_t-\gamma p_t\right) \end{bmatrix}=\begin{bmatrix} 0 \\ 0 \end{bmatrix},$$

but still only one parameter: the supply curve is now overidentified. The demand curve is still underidentified (two instruments and three parameters).

12.3 Testing for Autocorrelation

This section discusses how GMM can be used to test if a series is autocorrelated. The analysis focuses on first-order autocorrelation, but it is straightforward to extend it to higher-order autocorrelation.

Consider a scalar random variable x_t with a zero mean (it is easy to extend the analysis to allow for a non-zero mean). Consider the moment conditions

$$g_t(\beta)=\begin{bmatrix} x_t^2-\sigma^2 \\ x_t x_{t-1}-\rho\sigma^2 \end{bmatrix},\ \text{so}\ \bar{g}(\beta)=\frac{1}{T}\sum_{t=1}^{T}\begin{bmatrix} x_t^2-\sigma^2 \\ x_t x_{t-1}-\rho\sigma^2 \end{bmatrix},\ \text{with}\ \beta=\begin{bmatrix} \sigma^2 \\ \rho \end{bmatrix}. \tag{12.48}$$

σ^2 is the variance and ρ the first-order autocorrelation so $\rho\sigma^2$ is the first-order autocovariance. We want to test if $\rho=0$. We could proceed along two different routes: estimate ρ and test if it is different from zero or set ρ to zero and then test overidentifying restrictions. We analyze how these two approaches work when the null hypothesis of $\rho=0$ is true.

12.3.1 Estimating the Autocorrelation Coefficient

We estimate both σ^2 and ρ by using the moment conditions (18.5) and then test if $\rho=0$. To do that we need to calculate the asymptotic variance of $\hat{\rho}$ (there is little hope of being able to calculate the small sample variance, so we have to settle for the asymptotic variance as an approximation).

We have an exactly identified system so the weight matrix does not matter—we can then proceed as if we had used the optimal weighting matrix (all those results apply).

To find the asymptotic covariance matrix of the parameters estimators, we need the probability limit of the Jacobian of the moments and the covariance matrix of the moments—evaluated at the true parameter values. Let $\bar{g}_i(\beta_0)$ denote the ith element of the $\bar{g}(\beta)$

vector—evaluated at the true parameter values. The probability of the Jacobian is

$$D_0 = \operatorname{plim} \begin{bmatrix} \partial \bar{g}_1(\beta_0)/\partial\sigma^2 & \partial \bar{g}_1(\beta_0)/\partial\rho \\ \partial \bar{g}_2(\beta_0)/\partial\sigma^2 & \partial \bar{g}_2(\beta_0)/\partial\rho \end{bmatrix} = \begin{bmatrix} -1 & 0 \\ -\rho & -\sigma^2 \end{bmatrix} = \begin{bmatrix} -1 & 0 \\ 0 & -\sigma^2 \end{bmatrix}, \tag{12.49}$$

since $\rho = 0$ (the true value). Note that we differentiate with respect to σ^2, not σ, since we treat σ^2 as a parameter.

The covariance matrix is more complicated. The definition is

$$S_0 = \mathrm{E} \left[\frac{\sqrt{T}}{T} \sum_{t=1}^{T} g_t(\beta_0) \right] \left[\frac{\sqrt{T}}{T} \sum_{t=1}^{T} g_t(\beta_0) \right]'.$$

Assume that there is no autocorrelation in $g_t(\beta_0)$. We can then simplify as

$$S_0 = \mathrm{E}\, g_t(\beta_0) g_t(\beta_0)'.$$

This assumption is stronger than assuming that $\rho = 0$, but we make it here in order to illustrate the asymptotic distribution. To get anywhere, we assume that x_t is iid $N(0, \sigma^2)$. In this case (and with $\rho = 0$ imposed) we get

$$\begin{aligned} S_0 &= \mathrm{E} \begin{bmatrix} x_t^2 - \sigma^2 \\ x_t x_{t-1} \end{bmatrix} \begin{bmatrix} x_t^2 - \sigma^2 \\ x_t x_{t-1} \end{bmatrix}' = \mathrm{E} \begin{bmatrix} (x_t^2 - \sigma^2)^2 & (x_t^2 - \sigma^2) x_t x_{t-1} \\ (x_t^2 - \sigma^2) x_t x_{t-1} & (x_t x_{t-1})^2 \end{bmatrix} \\ &= \begin{bmatrix} \mathrm{E}\, x_t^4 - 2\sigma^2 \mathrm{E}\, x_t^2 + \sigma^4 & 0 \\ 0 & \mathrm{E}\, x_t^2 x_{t-1}^2 \end{bmatrix} = \begin{bmatrix} 2\sigma^4 & 0 \\ 0 & \sigma^4 \end{bmatrix}. \end{aligned} \tag{12.50}$$

To make the simplification in the second line we use the facts that $\mathrm{E}\, x_t^4 = 3\sigma^4$ if $x_t \sim N(0, \sigma^2)$, and that the normality and the iid properties of x_t together imply $\mathrm{E}\, x_t^2 x_{t-1}^2 = \mathrm{E}\, x_t^2 \mathrm{E}\, x_{t-1}^2$ and $\mathrm{E}\, x_t^3 x_{t-1} = \mathrm{E}\, \sigma^2 x_t x_{t-1} = 0$.

By combining (12.49) and (12.50) we get that

$$\begin{aligned} \operatorname{ACov} \left(\sqrt{T} \begin{bmatrix} \hat{\sigma}^2 \\ \hat{\rho} \end{bmatrix} \right) &= \left(D_0' S_0^{-1} D_0 \right)^{-1} \\ &= \left(\begin{bmatrix} -1 & 0 \\ 0 & -\sigma^2 \end{bmatrix}' \begin{bmatrix} 2\sigma^4 & 0 \\ 0 & \sigma^4 \end{bmatrix}^{-1} \begin{bmatrix} -1 & 0 \\ 0 & -\sigma^2 \end{bmatrix} \right)^{-1} \\ &= \begin{bmatrix} 2\sigma^4 & 0 \\ 0 & 1 \end{bmatrix}. \end{aligned} \tag{12.51}$$

This shows the standard expression for the uncertainty of the variance and that the $\sqrt{T}\hat{\rho}$. Since GMM estimators typically have an asymptotic distribution we have $\sqrt{T}\hat{\rho} \rightarrow^d N(0,1)$, so we can test the null hypothesis of no first-order autocorrelation by the test statistics

$$T\hat{\rho}^2 \sim \chi_1^2. \tag{12.52}$$

This is the same as the *Box-Ljung test for first-order autocorrelation*.

This analysis shows that we are able to arrive at simple expressions for the sampling uncertainty of the variance and the autocorrelation—provided we are willing to make strong assumptions about the data generating process. In particular, ewe assumed that data was iid $N(0,\sigma^2)$. One of the strong points of GMM is that we could perform similar tests without making strong assumptions—provided we use a correct estimator of the asymptotic covariance matrix S_0 (for instance, Newey-West).

12.3.2 Testing the Overidentifying Restriction of No Autocorrelation*

We can estimate σ^2 alone and then test if both moment condition are satisfied at $\rho = 0$. There are several ways of doing that, but the perhaps most straightforward is skip the loss function approach to GMM and instead specify the "first order conditions" directly as

$$\begin{aligned} 0 &= A\bar{g} \\ &= \begin{bmatrix} 1 & 0 \end{bmatrix} \frac{1}{T}\sum_{t=1}^{T} \begin{bmatrix} x_t^2 - \sigma^2 \\ x_t x_{t-1} \end{bmatrix}, \end{aligned} \tag{12.53}$$

which sets $\hat{\sigma}^2$ equal to the sample variance.

The only parameter in this estimation problem is σ^2, so the matrix of derivatives becomes

$$D_0 = \operatorname{plim} \begin{bmatrix} \partial \bar{g}_1(\beta_0)/\partial\sigma^2 \\ \partial \bar{g}_2(\beta_0)/\partial\sigma^2 \end{bmatrix} = \begin{bmatrix} -1 \\ 0 \end{bmatrix}. \tag{12.54}$$

By using this result, the A matrix in (12.54) and the S_0 matrix in (12.50,) it is straighforward to calculate the asymptotic covariance matrix the moment conditions. In general, we have

$$\operatorname{ACov}[\sqrt{T}\bar{g}(\hat{\beta})] = [I - D_0 (A_0 D_0)^{-1} A_0] S_0 [I - D_0 (A_0 D_0)^{-1} A_0]'. \tag{12.55}$$

The term in brackets is here (since $A_0 = A$ since it is a matrix with constants)

$$\underbrace{\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}}_{I_2} - \underbrace{\begin{bmatrix} -1 \\ 0 \end{bmatrix}}_{D_0} \left(\underbrace{\begin{bmatrix} 1 & 0 \end{bmatrix}}_{A_0} \underbrace{\begin{bmatrix} -1 \\ 0 \end{bmatrix}}_{D_0} \right)^{-1} \underbrace{\begin{bmatrix} 1 & 0 \end{bmatrix}}_{A_0} = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}. \tag{12.56}$$

We therefore get

$$\text{ACov}[\sqrt{T}\bar{g}(\hat{\beta})] = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 2\sigma^4 & 0 \\ 0 & \sigma^4 \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}' = \begin{bmatrix} 0 & 0 \\ 0 & \sigma^4 \end{bmatrix}. \tag{12.57}$$

Note that the first moment condition has no sampling variance at the estimated parameters, since the choice of $\hat{\sigma}^2$ always sets the first moment condition equal to zero.

The test of the overidentifying restriction that the second moment restriction is also zero is

$$T\bar{g}' \left(\text{ACov}[\sqrt{T}\bar{g}(\hat{\beta})] \right)^{+} \bar{g} \sim \chi^2_1, \tag{12.58}$$

where we have to use a generalized inverse if the covariance matrix is singular (which it is in (12.57)).

In this case, we get the test statistics (note the generalized inverse)

$$T \begin{bmatrix} 0 \\ \Sigma_{t=1}^T x_t x_{t-1}/T \end{bmatrix}' \begin{bmatrix} 0 & 0 \\ 0 & 1/\sigma^4 \end{bmatrix} \begin{bmatrix} 0 \\ \Sigma_{t=1}^T x_t x_{t-1}/T \end{bmatrix} = T \frac{\left[\Sigma_{t=1}^T x_t x_{t-1}/T\right]^2}{\sigma^4}, \tag{12.59}$$

which is the T times the square of the sample covariance divided by σ^4. A sample correlation, $\hat{\rho}$, would satisfy $\Sigma_{t=1}^T x_t x_{t-1}/T = \hat{\rho}\hat{\sigma}^2$, which we can use to rewrite (12.59) as $T\hat{\rho}^2\hat{\sigma}^4/\sigma^4$. By approximating σ^4 by $\hat{\sigma}^4$ we get the same test statistics as in (12.52).

12.4 Estimating and Testing a Normal Distribution

12.4.1 Estimating the Mean and Variance

This section discusses how the GMM framework can be used to test if a variable is normally distributed. The analysis cold easily be changed in order to test other distributions as well.

Suppose we have a sample of the scalar random variable x_t and that we want to test if the series is normally distributed. We analyze the asymptotic distribution under the null

hypothesis that x_t is $N(\mu, \sigma^2)$.

We specify four moment conditions

$$g_t = \begin{bmatrix} x_t - \mu \\ (x_t - \mu)^2 - \sigma^2 \\ (x_t - \mu)^3 \\ (x_t - \mu)^4 - 3\sigma^4 \end{bmatrix} \text{ so } \bar{g} = \frac{1}{T}\sum_{t=1}^{T} \begin{bmatrix} x_t - \mu \\ (x_t - \mu)^2 - \sigma^2 \\ (x_t - \mu)^3 \\ (x_t - \mu)^4 - 3\sigma^4 \end{bmatrix} \tag{12.60}$$

Note that $\mathrm{E}\, g_t = \mathbf{0}_{4\times 1}$ if x_t is normally distributed.

Let $\bar{g}_i(\beta_0)$ denote the ith element of the $\bar{g}(\beta)$ vector—evaluated at the true parameter values. The probability of the Jacobian is

$$\begin{aligned} D_0 &= \operatorname{plim} \begin{bmatrix} \partial \bar{g}_1(\beta_0)/\partial\mu & \partial \bar{g}_1(\beta_0)/\partial\sigma^2 \\ \partial \bar{g}_2(\beta_0)/\partial\mu & \partial \bar{g}_2(\beta_0)/\partial\sigma^2 \\ \partial \bar{g}_3(\beta_0)/\partial\mu & \partial \bar{g}_3(\beta_0)/\partial\sigma^2 \\ \partial \bar{g}_4(\beta_0)/\partial\mu & \partial \bar{g}_4(\beta_0)/\partial\sigma^2 \end{bmatrix} \\ &= \operatorname{plim} \frac{1}{T}\sum_{t=1}^{T} \begin{bmatrix} -1 & 0 \\ -2(x_t - \mu) & -1 \\ -3(x_t - \mu)^2 & 0 \\ -4(x_t - \mu)^3 & -6\sigma^2 \end{bmatrix} = \begin{bmatrix} -1 & 0 \\ 0 & -1 \\ -3\sigma^2 & 0 \\ 0 & -6\sigma^2 \end{bmatrix}. \end{aligned} \tag{12.61}$$

(Recall that we treat σ^2, not σ, as a parameter.)

The covariance matrix of the scaled moment conditions (at the true parameter values) is

$$S_0 = \mathrm{E}\left[\frac{\sqrt{T}}{T}\sum_{t=1}^{T} g_t(\beta_0)\right]\left[\frac{\sqrt{T}}{T}\sum_{t=1}^{T} g_t(\beta_0)\right]', \tag{12.62}$$

which can be a very messy expression. Assume that there is no autocorrelation in $g_t(\beta_0)$, which would certainly be true if x_t is iid. We can then simplify as

$$S_0 = \mathrm{E}\, g_t(\beta_0) g_t(\beta_0)', \tag{12.63}$$

which is the form we use here for illustration. We therefore have (provided $g_t(\beta_0)$ is not

autocorrelated)

$$S_0 = \mathrm{E}\left(\begin{bmatrix} x_t - \mu \\ (x_t-\mu)^2 - \sigma^2 \\ (x_t-\mu)^3 \\ (x_t-\mu)^4 - 3\sigma^4 \end{bmatrix}\right)\left(\begin{bmatrix} x_t - \mu \\ (x_t-\mu)^2 - \sigma^2 \\ (x_t-\mu)^3 \\ (x_t-\mu)^4 - 3\sigma^4 \end{bmatrix}\right)' = \begin{bmatrix} \sigma^2 & 0 & 3\sigma^4 & 0 \\ 0 & 2\sigma^4 & 0 & 12\sigma^6 \\ 3\sigma^4 & 0 & 15\sigma^6 & 0 \\ 0 & 12\sigma^6 & 0 & 96\sigma^8 \end{bmatrix}. \tag{12.64}$$

It is straightforward to derive this result once we have the information in the following remark.

Remark 12.6 *If* $X \sim N(\mu, \sigma^2)$*, then the first few moments around the mean of a are* $\mathrm{E}(X-\mu) = 0$*,* $\mathrm{E}(X-\mu)^2 = \sigma^2$*,* $\mathrm{E}(X-\mu)^3 = 0$ *(all odd moments are zero),* $\mathrm{E}(X-\mu)^4 = 3\sigma^4$*,* $\mathrm{E}(X-\mu)^6 = 15\sigma^6$*, and* $\mathrm{E}(X-\mu)^8 = 105\sigma^8$*.*

Suppose we use the efficient weighting matrix. The asymptotic covariance matrix of the estimated mean and variance is then $((D_0' S_0^{-1} D_0)^{-1})$

$$\begin{aligned}\left(\begin{bmatrix} -1 & 0 \\ 0 & -1 \\ -3\sigma^2 & 0 \\ 0 & -6\sigma^2 \end{bmatrix}' \begin{bmatrix} \sigma^2 & 0 & 3\sigma^4 & 0 \\ 0 & 2\sigma^4 & 0 & 12\sigma^6 \\ 3\sigma^4 & 0 & 15\sigma^6 & 0 \\ 0 & 12\sigma^6 & 0 & 96\sigma^8 \end{bmatrix}^{-1} \begin{bmatrix} -1 & 0 \\ 0 & -1 \\ -3\sigma^2 & 0 \\ 0 & -6\sigma^2 \end{bmatrix}\right)^{-1} &= \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{bmatrix}^{-1} \\ &= \begin{bmatrix} \sigma^2 & 0 \\ 0 & 2\sigma^4 \end{bmatrix}.\end{aligned} \tag{12.65}$$

This is the same as the result from maximum likelihood estimation which use the sample mean and sample variance as the estimators. The extra moment conditions (overidentifying restrictions) does not produce any more efficient estimators—for the simple reason that the first two moments completely characterizes the normal distribution.

12.4.2 Testing Normality*

The payoff from the overidentifying restrictions is that we can test if the series is actually normally distributed. There are several ways of doing that, but the perhaps most straightforward is skip the loss function approach to GMM and instead specify the "first order

conditions" directly as

$$
\begin{aligned}
0 &= A\bar{g} \\
&= \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \frac{1}{T} \sum_{t=1}^{T} \begin{bmatrix} x_t - \mu \\ (x_t - \mu)^2 - \sigma^2 \\ (x_t - \mu)^3 \\ (x_t - \mu)^4 - 3\sigma^4 \end{bmatrix}.
\end{aligned} \tag{12.66}
$$

The asymptotic covariance matrix the moment conditions is as in (12.55). In this case, the matrix with brackets is

$$
\begin{aligned}
&\underbrace{\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}}_{I_4} - \underbrace{\begin{bmatrix} -1 & 0 \\ 0 & -1 \\ -3\sigma^2 & 0 \\ 0 & -6\sigma^2 \end{bmatrix}}_{D_0} \left(\underbrace{\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}}_{A_0} \underbrace{\begin{bmatrix} -1 & 0 \\ 0 & -1 \\ -3\sigma^2 & 0 \\ 0 & -6\sigma^2 \end{bmatrix}}_{D_0} \right)^{-1} \underbrace{\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}}_{A_0} \\
&= \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ -3\sigma^2 & 0 & 1 & 0 \\ 0 & -6\sigma^2 & 0 & 1 \end{bmatrix}
\end{aligned} \tag{12.67}
$$

We therefore get

$$
\begin{aligned}
\text{ACov}[\sqrt{T}\bar{g}(\hat{\beta})] &= \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ -3\sigma^2 & 0 & 1 & 0 \\ 0 & -6\sigma^2 & 0 & 1 \end{bmatrix} \begin{bmatrix} \sigma^2 & 0 & 3\sigma^4 & 0 \\ 0 & 2\sigma^4 & 0 & 12\sigma^6 \\ 3\sigma^4 & 0 & 15\sigma^6 & 0 \\ 0 & 12\sigma^6 & 0 & 96\sigma^8 \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ -3\sigma^2 & 0 & 1 & 0 \\ 0 & -6\sigma^2 & 0 & 1 \end{bmatrix}' \\
&= \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 6\sigma^6 & 0 \\ 0 & 0 & 0 & 24\sigma^8 \end{bmatrix}
\end{aligned} \tag{12.68}
$$

We now form the test statistics for the overidentifying restrictions as in (12.58). In

this case, it is (note the generalized inverse)

$$T\begin{bmatrix} 0 \\ 0 \\ \Sigma_{t=1}^{T}(x_t-\mu)^3/T \\ \Sigma_{t=1}^{T}[(x_t-\mu)^4-3\sigma^4]/T \end{bmatrix}'\begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1/(6\sigma^6) & 0 \\ 0 & 0 & 0 & 1/(24\sigma^8) \end{bmatrix}\begin{bmatrix} 0 \\ 0 \\ \Sigma_{t=1}^{T}(x_t-\mu)^3/T \\ \Sigma_{t=1}^{T}[(x_t-\mu)^4-3\sigma^4]/T \end{bmatrix}$$
$$= \frac{T}{6}\frac{\left[\Sigma_{t=1}^{T}(x_t-\mu)^3/T\right]^2}{\sigma^6} + \frac{T}{24}\frac{\left\{\Sigma_{t=1}^{T}[(x_t-\mu)^4-3\sigma^4]/T\right\}^2}{\sigma^8}. \tag{12.69}$$

When we approximate σ by $\hat{\sigma}$ then this is the same as the *Jarque and Bera test of normality*.

The analysis shows (once again) that we can arrive at simple closed form results by making strong assumptions about the data generating process. In particular, we assumed that the moment conditions were serially uncorrelated. The GMM test, with a modified estimator of the covariance matrix S_0, can typically be much more general.

12.5 IV on a System of Equations*

Suppose we have two equations

$$y_{1t} = x_{1t}'\beta_1 + u_{1t}$$
$$y_{2t} = x_{2t}'\beta_2 + u_{2t},$$

and two sets of instruments, z_{1t} and z_{2t} with the same dimensions as x_{1t} and x_{2t}, respectively. The sample moment conditions are

$$\bar{g}(\beta_1,\beta_2) = \frac{1}{T}\sum_{t=1}^{T}\begin{bmatrix} z_{1t}\left(y_{1t}-x_{1t}'\beta_1\right) \\ z_{2t}\left(y_{2t}-x_{2t}'\beta_2\right) \end{bmatrix},$$

Let $\beta = (\beta_1', \beta_2')'$. Then

$$\frac{\partial \bar{g}(\beta_1,\beta_2)}{\partial \beta'} = \begin{bmatrix} \frac{\partial}{\partial \beta_1'}\frac{1}{T}\sum_{t=1}^{T} z_{1t}\left(y_{1t}-x_{1t}'\beta_1\right) & \frac{\partial}{\partial \beta_2'}\frac{1}{T}\sum_{t=1}^{T} z_{1t}\left(y_{1t}-x_{1t}'\beta_1\right) \\ \frac{\partial}{\partial \beta_1'}\frac{1}{T}\sum_{t=1}^{T} z_{2t}\left(y_{2t}-x_{2t}'\beta_2\right) & \frac{\partial}{\partial \beta_2'}\frac{1}{T}\sum_{t=1}^{T} z_{2t}\left(y_{2t}-x_{2t}'\beta_2\right) \end{bmatrix}$$
$$= \begin{bmatrix} \frac{1}{T}\sum_{t=1}^{T} z_{1t}x_{1t}' & \mathbf{0} \\ \mathbf{0} & \frac{1}{T}\sum_{t=1}^{T} z_{2t}x_{2t}' \end{bmatrix}.$$

This is invertible so we can premultiply the first order condition with the inverse of $[\partial \bar{g}(\beta)/\partial \beta']' A$ and get $\bar{g}(\beta) = \mathbf{0}_{k\times 1}$. We can solve this system for β_1 and β_2 as

$$\begin{aligned}
\begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} &= \begin{bmatrix} \frac{1}{T}\sum_{t=1}^{T} z_{1t}x_{1t}' & \mathbf{0} \\ \mathbf{0} & \frac{1}{T}\sum_{t=1}^{T} z_{2t}x_{2t}' \end{bmatrix}^{-1} \begin{bmatrix} \frac{1}{T}\sum_{t=1}^{T} z_{1t}y_{1t} \\ \frac{1}{T}\sum_{t=1}^{T} z_{2t}y_{2t} \end{bmatrix} \\
&= \begin{bmatrix} \left(\frac{1}{T}\sum_{t=1}^{T} z_{1t}x_{1t}'\right)^{-1} & \mathbf{0} \\ \mathbf{0} & \left(\frac{1}{T}\sum_{t=1}^{T} z_{2t}x_{2t}'\right)^{-1} \end{bmatrix} \begin{bmatrix} \frac{1}{T}\sum_{t=1}^{T} z_{1t}y_{1t} \\ \frac{1}{T}\sum_{t=1}^{T} z_{2t}y_{2t} \end{bmatrix}.
\end{aligned}$$

This is IV on each equation separately, which follows from having an exactly identified system.

Chapter 13

Factor Models

Sections denoted by a star (*) is not required reading.

13.1 CAPM Tests: Overview

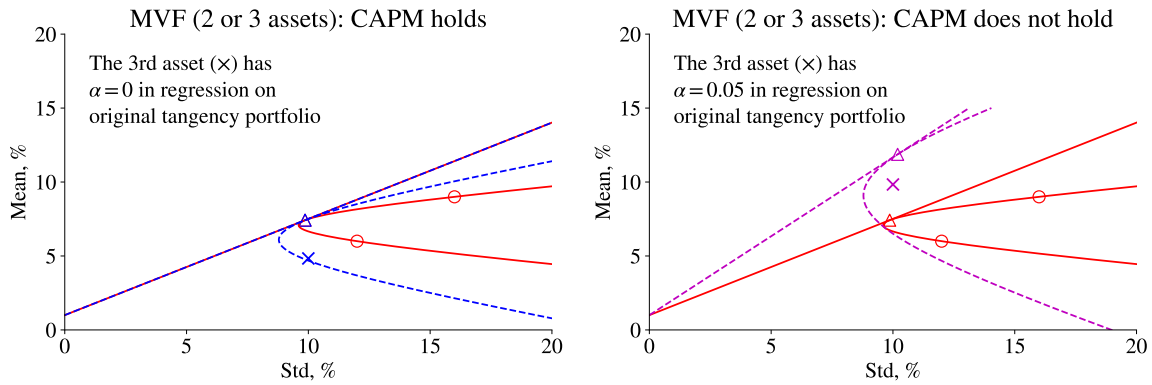
Reference: Cochrane (2005) 12.1; Campbell, Lo, and MacKinlay (1997) 5; Campbell (2018) 3

Let R_{it}^e be the excess return on asset i in excess over the riskfree asset, and let f_t be the excess return on the market portfolio (f for factor). CAPM with a riskfree return says that $\alpha_i = 0$ in

$$\begin{aligned} R_{it}^e &= \alpha_i + \beta_i f_t + \varepsilon_{it}, \text{ where} \\ \mathrm{E}\,\varepsilon_{it} &= 0 \text{ and } \mathrm{Cov}(f_t, \varepsilon_{it}) = 0. \end{aligned} \tag{13.1}$$

The basic test of CAPM is to estimate (13.1) on a single asset and then test if the intercept is zero. This can easily be extended to several assets, where we test if all the intercepts are zero.

Notice that the test of CAPM can be given two interpretations. If we assume that the factor (f_t) is the correct benchmark, then it is a test of whether asset i is "correctly" priced (this is the approach in mutual fund evaluations). Alternatively, if we assume that asset i is correctly priced, then it is a test of the mean-variance efficiency of the factor (compare the Roll critique).



Tangency portfolios:

	original	$\alpha=0$	$\alpha=0.05$
Asset 1	0.47	0.47	0.21
Asset 2	0.53	0.53	-0.58
Asset 3		-0.00	1.36

Figure 13.1: MV frontiers with 2 and 3 assets

13.2 Testing CAPM: Traditional LS Approach

13.2.1 CAPM with One Asset: Traditional LS Approach

If the residuals in the CAPM regression are iid (and independent of the regressor), then the traditional LS approach is just fine: estimate (13.1) and form a t-test of the null hypothesis that the intercept is zero.

The variance of the estimated intercept in the CAPM regression (13.1) is

$$\text{Var}(\hat{\alpha}_i) = (1 + SR^2)\sigma_i^2/T, \tag{13.2}$$

where σ_i^2 is the variance of the residual in (13.1) and SR^2 is the squared Sharpe ratio of the market portfolio (recall: f_t is the excess return on market portfolio). The result is well known, but a simple proof is found in Appendix 13.10. Equation (13.2) shows that the uncertainty about the intercept is high when the disturbance is volatile and when the sample is short, but also when the Sharpe ratio of the market is high. Note that a large market Sharpe ratio means that the market asks for a high compensation for taking on risk. A bit uncertainty about how risky asset i is then gives a large uncertainty about what the risk-adjusted return should be. Clearly, (13.2) can be used to construct a t-test.

Instead of a t-test, we can use the equivalent chi-square test

$$\frac{\hat{\alpha}_i^2}{\text{Var}(\hat{\alpha}_i)} \xrightarrow{d} \chi_1^2 \text{ under } H_0\text{: } \alpha_0 = 0. \tag{13.3}$$

It is quite straightforward to use the properties of mean-variance frontiers (see Gibbons, Ross, and Shanken (1989), MacKinlay (1995) and the simple proof in Appendix 13.10) to show that the test statistic in (13.3) can be written

$$\frac{\hat{\alpha}_i^2}{\text{Var}(\hat{\alpha}_i)} = \frac{(\widehat{SR}_c)^2 - SR^2}{(1 + SR^2)/T}, \tag{13.4}$$

where SR is the Sharpe ratio of the market portfolio and SR_c is the Sharpe ratio of the tangency portfolio when investment in both the market return and asset i is possible. (Recall that the tangency portfolio is the portfolio with the highest possible Sharpe ratio.) If the market portfolio has the same (squared) Sharpe ratio as the tangency portfolio of the mean-variance frontier of asset i and the market portfolio (so the market portfolio is mean-variance efficient also when we take the test asset into account) then the test statistic, $\hat{\alpha}_i^2/\text{Var}(\hat{\alpha}_i)$, is zero—and CAPM is not rejected. The economic importance of a non-zero intercept (α) is thus that the tangency portfolio changes if the test asset is added to the investment opportunity set. See Figure 13.1 for an illustration.

13.2.2 CAPM with Several Assets: Traditional LS Approach

Suppose we have n test assets. Stack the expressions (13.1) for $i = 1, \ldots, n$ as

$$\begin{bmatrix} R_{1t}^e \\ \vdots \\ R_{nt}^e \end{bmatrix} = \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{bmatrix} + \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_n \end{bmatrix} f_t + \begin{bmatrix} \varepsilon_{1t} \\ \vdots \\ \varepsilon_{nt} \end{bmatrix}, \text{ where} \tag{13.5}$$

$$\text{E}\,\varepsilon_{it} = 0 \text{ and } \text{Cov}(f_t, \varepsilon_{it}) = 0.$$

This is a system of seemingly unrelated regressions (SUR)—with the same regressor (see, for instance, Greene (2003) 14). In this case, the efficient estimator (GLS) is LS on each equation separately. Moreover, the covariance matrix of the coefficients is particularly simple.

Under the null hypothesis of zero intercepts and iid residuals (although possibly correlated across regressions), the LS estimate of the intercept has the following asymptotic

distribution

$$\sqrt{T}\hat{\alpha} \to^d N\left[\mathbf{0}_{n\times 1}, \Sigma(1+SR^2)\right], \text{ where} \tag{13.6}$$
$$\Sigma = \begin{bmatrix} \sigma_{11} & \dots & \sigma_{1n} \\ \vdots & & \vdots \\ \sigma_{n1} & \dots & \hat{\sigma}_{nn} \end{bmatrix} \text{ with } \sigma_{ij} = \text{Cov}(\varepsilon_{it}, \varepsilon_{jt})$$

and where $SR^2 = (\text{E}\, f)^2/\text{Var}(f)$.

In practice, we use the sample moments for the covariance matrix, $\sigma_{ij} = \sum_{t=1}^{T} \hat{\varepsilon}_{it}\hat{\varepsilon}_{jt}/T$. This result is well known, but a simple proof is found in Appendix 13.11. To test the null hypothesis that all intercepts are zero, we then use the test statistic

$$T\hat{\alpha}'(1+SR^2)^{-1}\Sigma^{-1}\hat{\alpha} \sim \chi^2_n. \tag{13.7}$$

13.2.3 CAPM with Several Assets: Bonferroni Test

Remark 13.1 *(The Bonferroni inequality) Suppose we perform $i = 1...n$ different tests, each at the significance level p_i. The Bonferroni inequality then says that if the null hypotheses are all true, then*

$$\Pr(\textit{not rejecting in any of the } n \textit{ tests}) \geq 1 - \Sigma_{i=1}^{n} p_i.$$

It follows that rejecting in at least one of the n tests has a probability of less than or equal to $\Sigma_{i=1}^{n} p_i$. For instance, with $p_i = 0.05/n$, there is 5% chance of rejecting in at least one test: $\Pr(\textit{rejecting in at least one of the } n \textit{ tests}) \leq 0.05$.

As an alternative to the joint test, we could instead study each of the n assets separately. Clearly, if we can safely reject the null hypothesis for at least one asset, then the joint hypothesis is also rejected. However, this cannot be implemented with traditional critical values since the chance of at least one false rejection increases with the number of test assets.

To control this "family-wise error rate," a Bonferroni correction is applied. To do this, let t_i be the t-stat for asset i ($t_i = \hat{\alpha}_i/\text{Std}(\hat{\alpha}_i)$). As usual, we would reject the hypothesis that $\alpha_i = 0$ on the 5% level if is $|t_i| > 1.96$.

Redo this for each asset—and reject the joint hypothesis on the family-wise significance level of 5% if at least one of the individual test statistics exceeds the $0.05/n$ critical value. For instance, with 10 test assets, we compare $|t_i|$ with 2.81 instead of 1.96 (since

2.81 is the 99.75th percentile of a $N(0,1)$ distribution, whereas the 97.5 percentile is 1.96). To use another significance level ρ, use ρ/n instead of $0.05/n$.

It can be noticed that since we focus on the highest individual test statistic, the Bonferroni and the Holm-Bonferroni (Holm, 1979) methods give the same result. This would be different if we wanted to see how many of the alphas that are different from 0. In that case the Holm-Bonferroni method is more powerful.

	1	2	3	4	5
1	−3.16	0.35	0.66	2.39	2.63
2	−2.09	0.81	1.71	2.41	1.99
3	−1.95	1.54	1.38	2.40	2.36
4	−0.53	0.65	1.39	2.10	1.54
5	0.08	1.34	1.29	−0.01	0.79

Table 13.1: t-stats for α in CAPM, 25 FF portfolios 1970:01-2021:12. NW uses 1 lag. The Bonferroni adjusted 10% and 5% critical values are 2.88 and 3.09.

13.3 Testing CAPM: GMM

13.3.1 CAPM with Several Assets: GMM and a Wald Test

To test n assets at the same time when the errors are non-iid we can use the GMM framework.

Write the n regressions in (13.5) on vector form as

$$\begin{aligned} R_t^e &= \alpha + \beta f_t + \varepsilon_t, \text{ where} \\ \mathrm{E}\,\varepsilon_t &= \mathbf{0}_{n\times 1} \text{ and } \mathrm{Cov}(f_t, \varepsilon_t') = \mathbf{0}_{1\times n}, \end{aligned} \tag{13.8}$$

where α and β are $n \times 1$ vectors. Clearly, setting $n = 1$ gives the case of a single test asset.

The $2n$ GMM moment conditions are that, at the true values of α and β,

$$\mathrm{E}\, g_t(\alpha, \beta) = \mathbf{0}_{2n\times 1}, \text{ where} \tag{13.9}$$

$$g_t(\alpha, \beta) = \begin{bmatrix} \varepsilon_t \\ f_t \varepsilon_t \end{bmatrix} = \begin{bmatrix} R_t^e - \alpha - \beta f_t \\ f_t \left(R_t^e - \alpha - \beta f_t\right) \end{bmatrix}. \tag{13.10}$$

There are as many parameters as moment conditions, so the GMM estimator picks values of α and β such that the sample analogues of (13.9) are satisfied exactly, which gives the

LS estimator. For the inference, we allow for the possibility of non-iid errors, but if the errors are actually iid, then we (asymptotically) get the same results as in Section 13.2.

With point estimates and their sampling distribution it is straightforward to set up a Wald test for the hypothesis that all elements in α are zero

$$\hat{\alpha}' \operatorname{Var}(\hat{\alpha})^{-1} \hat{\alpha} \xrightarrow{d} \chi^2_n. \tag{13.11}$$

Remark 13.2 *(Easy coding of the GMM Problem (13.9)) Estimate (13.8) by LS (equation by equation). Then, plug in the fitted residuals in (13.10) to generate time series of the moments (will be important for the tests).*

Remark 13.3 *(Distribution of GMM) Let the parameter vector in the moment condition have the true value b_0. Define*

$$S_0 = \operatorname{Cov}\left[\sqrt{T}\bar{g}(b_0)\right] \text{ and } D_0 = \operatorname{plim} \frac{\partial \bar{g}(b_0)}{\partial b'}.$$

When the estimator solves $\min \bar{g}(b)' S_0^{-1} \bar{g}(b)$ *or when the model is exactly identified, the distribution of the GMM estimator is*

$$\sqrt{T}(\hat{b} - b_0) \xrightarrow{d} N(\mathbf{0}_{k\times 1}, V), \text{ where } V = (D_0 S_0^{-1} D_0')^{-1}.$$

When D_0 is invertible (as it would be in an exactly identified model), then we can also write $V = D_0^{-1} S_0 (D_0^{-1})'$.

Details on the Wald Test*

Note that, with a linear model, the Jacobian of the moment conditions does not involve the parameters that we want to estimate. This means that we do not have to worry about evaluating the Jacobian at the true parameter values. The probability limit of the Jacobian is simply the expected value, which can be written as

$$\begin{aligned} \operatorname{plim} \frac{\partial \bar{g}_t(\alpha, \beta)}{\partial [\alpha, \beta]} = D_0 &= -\mathrm{E}\begin{bmatrix} 1 & f_t \\ f_t & f_t^2 \end{bmatrix} \otimes I_n \\ &= -\mathrm{E}\left(\begin{bmatrix} 1 \\ f_t \end{bmatrix}\begin{bmatrix} 1 \\ f_t \end{bmatrix}'\right) \otimes I_n, \end{aligned} \tag{13.12}$$

where $\otimes$ is the Kronecker product. (The last expression applies also to the case of several factors.) Notice that we order the parameters as a column vector with the alphas first and

the betas second. It might be useful to notice that in this case

$$D_0^{-1} = -\left[\mathrm{E}\left(\begin{bmatrix} 1 \\ f_t \end{bmatrix}\begin{bmatrix} 1 \\ f_t \end{bmatrix}'\right)\right]^{-1} \otimes I_n, \tag{13.13}$$

since $(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}$ (if conformable).

Remark 13.4 *(Kronecker product) If A and B are matrices, then*

$$A \otimes B = \begin{bmatrix} a_{11}B & \cdots & a_{1n}B \\ \vdots & & \vdots \\ a_{m1}B & \cdots & a_{mn}B \end{bmatrix}.$$

Example 13.5 *(Two test assets) With assets 1 and 2, the parameter vector is $b = [\alpha_1, \alpha_2, \beta_1, \beta_2]'$. Write out the sample analogues of (13.9) as*

$$\begin{bmatrix} \bar{g}_1(\alpha,\beta) \\ \bar{g}_2(\alpha,\beta) \\ \bar{g}_3(\alpha,\beta) \\ \bar{g}_4(\alpha,\beta) \end{bmatrix} = \frac{1}{T}\sum_{t=1}^{T}\begin{bmatrix} R^e_{1t} - \alpha_1 - \beta_1 f_t \\ R^e_{2t} - \alpha_2 - \beta_2 f_t \\ f_t(R^e_{1t} - \alpha_1 - \beta_1 f_t) \\ f_t(R^e_{2t} - \alpha_2 - \beta_2 f_t) \end{bmatrix} = \frac{1}{T}\sum_{t=1}^{T}\begin{bmatrix} 1 \\ f_t \end{bmatrix} \otimes \begin{bmatrix} R^e_{1t} - \alpha_1 - \beta_1 f_t \\ R^e_{2t} - \alpha_2 - \beta_2 f_t \end{bmatrix},$$

where $\bar{g}_1(\alpha,\beta)$ denotes the sample average of the first moment condition. The Jacobian is

$$\begin{aligned} \frac{\partial \bar{g}(\alpha,\beta)}{\partial[\alpha_1,\alpha_2,\beta_1,\beta_2]'} &= \begin{bmatrix} \partial\bar{g}_1/\partial\alpha_1 & \partial\bar{g}_1/\partial\alpha_2 & \partial\bar{g}_1/\partial\beta_1 & \partial\bar{g}_1/\partial\beta_2 \\ \partial\bar{g}_2/\partial\alpha_1 & \partial\bar{g}_2/\partial\alpha_2 & \partial\bar{g}_2/\partial\beta_1 & \partial\bar{g}_2/\partial\beta_2 \\ \partial\bar{g}_3/\partial\alpha_1 & \partial\bar{g}_3/\partial\alpha_2 & \partial\bar{g}_3/\partial\beta_1 & \partial\bar{g}_3/\partial\beta_2 \\ \partial\bar{g}_4/\partial\alpha_1 & \partial\bar{g}_4/\partial\alpha_2 & \partial\bar{g}_4/\partial\beta_1 & \partial\bar{g}_4/\partial\beta_2 \end{bmatrix} \\ &= -\frac{1}{T}\sum_{t=1}^{T}\begin{bmatrix} 1 & 0 & f_t & 0 \\ 0 & 1 & 0 & f_t \\ f_t & 0 & f_t^2 & 0 \\ 0 & f_t & 0 & f_t^2 \end{bmatrix} = -\frac{1}{T}\sum_{t=1}^{T}\left(\begin{bmatrix} 1 \\ f_t \end{bmatrix}\begin{bmatrix} 1 \\ f_t \end{bmatrix}'\right) \otimes I_2. \end{aligned}$$

The asymptotic covariance matrix of $\sqrt{T}$ times the sample moment conditions, evaluated at the true parameter values, that is at the true disturbances, is defined as

$$S_0 = \mathrm{Cov}\left(\frac{\sqrt{T}}{T}\sum_{t=1}^{T} g_t(\alpha,\beta)\right). \tag{13.14}$$

The Newey-West estimator is often a good estimator of S_0.

From Remark 13.3, we can write the covariance matrix of the $2n \times 1$ vector of parameters (n parameters in α and another n in β) as

$$\text{Cov}\left(\sqrt{T}\begin{bmatrix}\hat{\alpha}\\ \hat{\beta}\end{bmatrix}\right) = D_0^{-1} S_0 (D_0^{-1})'. \tag{13.15}$$

13.3.2 CAPM and Several Assets: GMM and an LM Test*

We could also construct an "LM test" instead by imposing $\alpha = \mathbf{0}$ in the moment conditions (13.9). The moment conditions are then

$$\text{E}\, g(\beta) = \text{E}\begin{bmatrix} R_t^e - \beta f_t \\ f_t(R_t^e - \beta f_t)\end{bmatrix} = \mathbf{0}_{2n\times 1}. \tag{13.16}$$

Since there are $q = 2n$ moment conditions, but only n parameters (the β vector), this model is overidentified.

We could either use a weighting matrix in the GMM loss function or combine the moment conditions so the model becomes exactly identified.

With a weighting matrix, the estimator solves

$$\min_b \bar{g}(b)' W \bar{g}(b), \tag{13.17}$$

where $\bar{g}(b)$ is the sample average of the moments (evaluated at some parameter vector b), and W is a positive definite (and symmetric) weighting matrix. Once we have estimated the model, we can test the n overidentifying restrictions that all $q = 2n$ moment conditions are satisfied at the estimated n parameters $\hat{\beta}$. If not, the restriction (null hypothesis) that $\alpha = \mathbf{0}_{n\times 1}$ is rejected. The test is based on a quadratic form of the moment conditions, $T\bar{g}(b)'\Psi^{-1}\bar{g}(b)$ which has a chi-square distribution if the correct Ψ matrix is used.

Alternatively, to combine the moment conditions so the model becomes exactly identified, premultiply by a matrix A to get

$$A_{n\times 2n}\, \text{E}\, g(\beta) = \mathbf{0}_{n\times 1}. \tag{13.18}$$

The model is then tested by testing if all $2n$ moment conditions in (13.16) are satisfied at this vector of estimates of the betas. This is the GMM analogue to a classical LM test. Once again, the test is based on a quadratic form of the moment conditions, $T\bar{g}(b)'\Psi^{-1}\bar{g}(b)$ which has a chi-square distribution if the correct Ψ matrix is used.

For instance, to effectively use only the last n moment conditions in the estimation, we specify

$$A\,\mathrm{E}\,g(\beta) = \begin{bmatrix} 0_{n\times n} & I_n \end{bmatrix} \mathrm{E} \begin{bmatrix} R_t^e - \beta f_t \\ f_t(R_t^e - \beta f_t) \end{bmatrix} = \mathbf{0}_{n\times 1}. \tag{13.19}$$

This clearly gives the classical LS estimator without an intercept

$$\hat{\beta} = \frac{\sum_{t=1}^{T} f_t R_t^e / T}{\sum_{t=1}^{T} f_t^2 / T}. \tag{13.20}$$

Example 13.6 *(Combining moment conditions, CAPM on two assets) With two assets we can combine the four moment conditions into only two by*

$$A\,\mathrm{E}\,g_t(\beta_1, \beta_2) = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \mathrm{E} \begin{bmatrix} R_{1t}^e - \beta_1 f_t \\ R_{2t}^e - \beta_2 f_t \\ f_t(R_{1t}^e - \beta_1 f_t) \\ f_t(R_{2t}^e - \beta_2 f_t) \end{bmatrix} = \mathbf{0}_{2\times 1}.$$

Remark 13.7 *(Test of overidentifying assumption in GMM) When the GMM estimator solves the quadratic loss function $\bar{g}(\beta)' S_0^{-1} \bar{g}(\beta)$ (or is exactly identified), then the J test statistic is*

$$T\,\bar{g}(\hat{\beta})' S_0^{-1} \bar{g}(\hat{\beta}) \xrightarrow{d} \chi^2_{q-k},$$

where q is the number of moment conditions and k is the number of parameters.

Remark 13.8 *(Distribution of GMM, more general results) When GMM solves $\min_b \bar{g}(b)' W \bar{g}(b)$ or $A\bar{g}(\hat{\beta}) = \mathbf{0}_{k\times 1}$, the distribution of the GMM estimator and the test of overidentifying assumptions are different from those in Remarks 13.3 and 13.7.*

13.3.3 Size and Power of the CAPM Tests

The size (using asymptotic critical values) and power in small samples is often found to be disappointing. Typically, these tests tend to reject a true null hypothesis too often (see Campbell, Lo, and MacKinlay (1997) Table 5.1) and the power to reject a false null hypothesis is often fairly low. These features are especially pronounced when the sample is small and the number of assets, n, is high. One useful rule of thumb is that a *saturation ratio* (the number of observations per parameter) below 10 (or so) is likely to worsen the performance of the test. In the test here we have nT observations, $2n$ parameters in α and β, and $n(n+1)/2$ unique parameters in S_0, so the saturation ratio is $T/(2 + (n+1)/2)$.

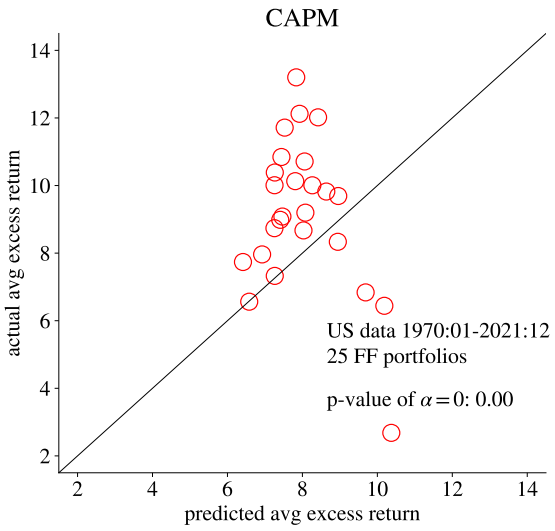


Figure 13.2: CAPM, FF portfolios

For instance, with $T = 60$ and $n = 10$ or at $T = 100$ and $n = 20$, we have a saturation ratio of 8, which is very low (compare Table 5.1 in CLM).

One possible way of dealing with the wrong size of the test is to use critical values from simulations of the small sample distributions (Monte Carlo simulations or bootstrap simulations).

13.3.4 Choice of Portfolios

This type of test is typically done on portfolios of assets, rather than on the individual assets themselves. There are several econometric and economic reasons for this. The econometric techniques we apply need the returns to be (reasonably) stationary in the sense that they have approximately the same means and covariance (with other returns) throughout the sample (individual assets, especially stocks, can change character as the company moves into another business). It might be more plausible that size or industry portfolios are stationary in this sense. Also, individual assets are typically very volatile, which makes it hard to obtain precise estimate and to be able to reject anything.

It sometimes makes economic sense to sort the assets according to a characteristic (size or perhaps book/market)—and then test if the model is true for these portfolios. Rejection of the CAPM for such portfolios may be particularly informative.

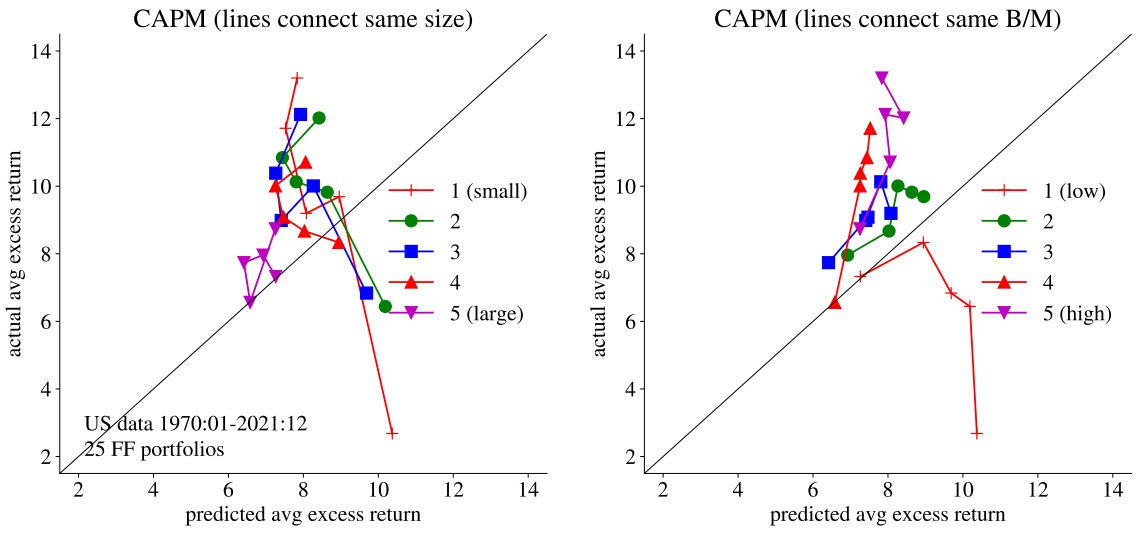


Figure 13.3: CAPM, FF portfolios

13.3.5 Empirical Evidence

See Campbell, Lo, and MacKinlay (1997) 6.5 (Table 6.1 in particular) and Cochrane (2005) 20.2.

One of the more interesting studies is Fama and French (1993) (see also Fama and French (1996)). They construct 25 stock portfolios according to two characteristics of the firm: the size and the book value to market value ratio (BE/ME). In June each year, they sort the stocks according to size and BE/ME. They then form a 5×5 matrix of portfolios, where portfolio ij belongs to the ith size quantile *and* the jth BE/ME quantile (so this is a *double-sort*). This is illustrated in Table 13.2.

	Book value/Market value				
	1	2	3	4	5
Size 1	1	2	3	4	5
2	6	7	8	9	10
3	11	12	13	14	15
4	16	17	18	19	20
5	21	22	23	24	25

Table 13.2: Numbering of the FF portfolios.

Fama and French run a traditional CAPM regression on each of the 25 portfolios (monthly data 1963–1991)—and then study if the expected excess returns are related

to the betas as they should according to CAPM (recall that CAPM implies $\mathrm{E}\, R_{it}^e = \beta_i \mathrm{E}\, R_{mt}^e$).

Empirical Example 13.9 *(CAPM on 25 FF portfolios) In fact, there is little relation between* $\mathrm{E}\, R_{it}^e$ *and* β_i *(see Figure 13.2). This lack of relation is due to the combination of two features of the data, see Figure 13.3. First,* within a size quantile *there is a negative relation (across BE/ME quantiles) between* $\mathrm{E}\, R_{it}^e$ *and* β_i*—in stark contrast to CAPM. Second,* within a BE/ME quantile*, there is a positive relation (across size quantiles) between* $\mathrm{E}\, R_{it}^e$ *and* β_i*—as predicted by CAPM.*

13.4 Testing Multi-Factor Models (Factors are Excess Returns)

Reference: Cochrane (2005) 12.1; Campbell, Lo, and MacKinlay (1997) 6.2.1

13.4.1 A Multi-Factor Model

When the K factors, f_t, are excess returns, the null hypothesis typically says that $\alpha_i = 0$ in

$$\begin{aligned} R_{it}^e &= \alpha_i + \beta_i' f_t + \varepsilon_{it}, \text{ where} \\ \mathrm{E}\, \varepsilon_{it} &= 0 \text{ and } \mathrm{Cov}(f_t, \varepsilon_{it}) = \mathbf{0}_{K\times 1}, \end{aligned} \tag{13.21}$$

and β_i is now an $K \times 1$ vector. The CAPM regression is a special case when the market excess return is the only factor. In other models like ICAPM (see Cochrane (2005) 9.2), we typically have several factors. We stack the returns for n assets to get

$$\begin{bmatrix} R_{1t}^e \\ \vdots \\ R_{nt}^e \end{bmatrix} = \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{bmatrix} + \begin{bmatrix} \beta_{11} & \dots & \beta_{1K} \\ \vdots & \ddots & \vdots \\ \beta_{n1} & \dots & \beta_{nK} \end{bmatrix} \begin{bmatrix} f_{1t} \\ \vdots \\ f_{Kt} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1t} \\ \vdots \\ \varepsilon_{nt} \end{bmatrix}$$

or in vector form

$$\begin{aligned} R_t^e &= \alpha + \beta f_t + \varepsilon_t, \text{ where} \\ \mathrm{E}\, \varepsilon_t &= \mathbf{0}_{n\times 1} \text{ and } \mathrm{Cov}(f_t, \varepsilon_t') = \mathbf{0}_{K\times n}, \end{aligned} \tag{13.22}$$

where α is $n \times 1$ and β is $n \times K$. Notice that β_{ij} shows how the ith asset depends on the jth factor.

This is, of course, very similar to the CAPM (one-factor) model—and both the LS and GMM approaches discussed before are straightforward to extend.

13.4.2 Multi-Factor Model: Traditional LS (SURE)

The results from the LS approach of testing CAPM generalizes directly (see Appendix 13.11 for details). In particular, (13.7) still holds—but where the residuals are from the multi-factor regressions (13.21) and where the Sharpe ratio of the tangency portfolio (based on the factors) depends on the means and covariance matrix of all factors

$$T\hat{\alpha}'(1+SR^2)^{-1}\Sigma^{-1}\hat{\alpha} \sim \chi^2_n, \text{ where} \tag{13.23}$$
$$SR^2 = \mathrm{E}\, f' \,\mathrm{Cov}(f)^{-1}\,\mathrm{E}\, f.$$

13.4.3 Multi-Factor Model: Orthogonalized Factors

Remark 13.10 *(The effect of orthogonalized regressors) Let $\tilde{x}_{2t}$ be the residual from regressing x_{2t} on x_{1t}. Suppose you estimate the following regressions*

$$y_t = x_{1t}'b_1 + e_{1t} \tag{13.24}$$
$$y_t = \tilde{x}_{2t}'b_2 + e_{2t} \tag{13.25}$$
$$y_t = x_{1t}'\beta_1 + x_{2t}'\beta_2 + u_t \tag{13.26}$$
$$y_t = x_{1t}'\gamma_1 + \tilde{x}_{2t}'\gamma_2 + \varepsilon_t, \tag{13.27}$$

Then, (a) $\hat{\gamma}_1 = \hat{b}_1$; (b) $\hat{\gamma}_2 = \hat{b}_2$; and also (c) $\hat{\gamma}_2 = \hat{\beta}_2$.

Proof. (of Remark 13.10)*) Since x_{1t} and $\tilde{x}_{2t}$ in (13.27) are orthogonal, we can estimate (13.24) and (13.25) separately and get the same coefficients as in the joint estimation—which demonstrates results *(a)* and *(b)*. Instead of regression (13.25), we could estimate $(*)$ $e_{1t} = \tilde{x}_{2t}'\delta_2 + \varepsilon_{2t}$ where e_{1t} is the residual from (13.24). It is clear that $\hat{b}_2 = \hat{\delta}_2$ since the movements in y_t that are driven by x_{1t} are orthogonal to $\tilde{x}_{2t}$ and will thus not affect regression (13.25). Result *(c)* then follows from the fact that regression $(*)$ is known to give the same estimate as $\hat{\beta}_2$ from (13.26); this is the Frisch-Waugh-Lovell theorem (see, for instance, Greene (2018) 3). ■

The point of Remark 13.10 is the following. Suppose x_{1t} contains a set of basic factors and a constant, and that we are interested in investigating a set of additional (new?) factors x_{2t}. It might then be tempting to orthogonalize the new x_{2t} factors against the old x_{1t}

factors—perhaps in the hope of getting a cleaner measure of their importance. However, the remark shows that this does *not* change the coefficients of the new factors.

(As a practical aspect, notice that instead of estimating (13.27) we can estimate (13.24) and (13.25) separately and get the same results.)

It is important to understand the difference between the regression of y_t on $\tilde{x}_{2t}$ in (13.25) and a regression of y_t on (the original, not orthogonalized) x_{2t}

$$y_t = x_{2t}'c_2 + v_{2t}. \tag{13.28}$$

If *(a)* x_{1t} and x_{2t} are correlated and *(b)* x_{1t} and y_t also are correlated, then the estimate $\hat{c}_2$ will differ from $\hat{b}_2$. This is similar to an omitted variables bias.

Remark 13.11 *(Omitted variables) It known (see, for instance, Greene (2018) 4) that if the correct regression model is* $y_t = g_t'\beta_g + h_t'\beta_h + u_t$*, but we estimate* $y_t = g_t'c + v_t$*, then the probability limit of* $\hat{c} = \beta_g + [\ \theta_1 \quad \ldots \quad \theta_L\ ]\beta_h$ *where* θ_i *is probability limit of the coefficients from regressing* h_{it} *on* g_t

In terms of Remark 13.11, consider (13.26) to be the correct model and (13.28) as the model with omitted variables (x_{1t} is omitted). We then notice that $\hat{c}_2$ is a mix of two things: *(1)* how x_{2t} influences y_t which is β_2 (or equivalently, γ_2 or b_2); *(2)* how x_{1t} influences y_t (which is β_1) times how x_{1t} influences x_{2t}.

For instance, we could have the case where $\beta_2 = 0$ (x_{2t} has not effect on y_t) but $c_2 > 0$ since x_{1t} affects both y_t and x_{2t} positively. In contrast, the regression on the ortogonalized variable (13.25) would give zero coefficients.

13.4.4 Multi-Factor Model: GMM

The moment conditions are

$$\mathrm{E}\, g_t(\alpha, \beta) = \mathrm{E}\left(\begin{bmatrix} 1 \\ f_t \end{bmatrix} \otimes \varepsilon_t\right) = \mathrm{E}\left(\begin{bmatrix} 1 \\ f_t \end{bmatrix} \otimes (R_t^e - \alpha - \beta f_t)\right) = \mathbf{0}_{n(1+K)\times 1}. \tag{13.29}$$

Note that this expression looks similar to (13.9)—the only difference is that f_t may now be a vector (and we therefore need to use the Kronecker product). It is then intuitively clear that the expressions for the asymptotic covariance matrix of $\hat{\alpha}$ and $\hat{\beta}$ will look very similar too.

When the system is exactly identified, the GMM estimator solves the sample analogues of (13.29), which is the same as LS (equation by equation). The model can be tested by testing if all alphas are zero, as in (13.11).

Instead, when we restrict $\alpha = \mathbf{0}_{n\times 1}$ (overidentified system), then we either specify a weighting matrix W and solve

$$\min_\beta \bar{g}(\beta)' W \bar{g}(\beta), \tag{13.30}$$

or we specify a matrix A to combine the moment conditions and solve

$$A_{nK\times n(1+K)}\bar{g}(\beta) = \mathbf{0}_{nK\times 1}. \tag{13.31}$$

Example 13.12 *(Moment condition with two assets and two factors) The moment conditions for $n = 2$ and $K = 2$ are*

$$\mathrm{E}\, g_t(\alpha,\beta) = \mathrm{E}\begin{bmatrix} R^e_{1t} - \alpha_1 - \beta_{11} f_{1t} - \beta_{12} f_{2t} \\ R^e_{2t} - \alpha_2 - \beta_{21} f_{1t} - \beta_{22} f_{2t} \\ f_{1t}(R^e_{1t} - \alpha_1 - \beta_{11} f_{1t} - \beta_{12} f_{2t}) \\ f_{1t}(R^e_{2t} - \alpha_2 - \beta_{21} f_{1t} - \beta_{22} f_{2t}) \\ f_{2t}(R^e_{1t} - \alpha_1 - \beta_{11} f_{1t} - \beta_{12} f_{2t}) \\ f_{2t}(R^e_{2t} - \alpha_2 - \beta_{21} f_{1t} - \beta_{22} f_{2t}) \end{bmatrix} = \mathbf{0}_{6\times 1}.$$

Restricting $\alpha_1 = \alpha_2 = 0$ gives the moment conditions for the overidentified case.

Details on the Wald Test*

For the exactly identified case, we have the following results. The expressions for the Jacobian D_0 and its inverse are the same as in (13.12)–(13.13). Notice that in this Jacobian we differentiate the moment conditions (13.29) with respect to $\mathrm{vec}(\alpha,\beta)$, that is, where the parameters are stacked in a column vector with the alphas first, then the betas for the first factor, followed by the betas for the second factor etc. The test is based on a quadratic form of the moment conditions, $T\bar{g}(b)'\Psi^{-1}\bar{g}(b)$ which has a chi-square distribution if the correct Ψ matrix is used. The covariance matrix of the average moment conditions are as in (13.14).

13.4.5 Empirical Evidence

Fama and French (1993) also try a multi-factor model. They find that a three-factor model fits the 25 stock portfolios fairly well (two more factors are needed to also fit the seven

bond portfolios that they use). The three factors are: the market excess return, the return on a portfolio of small stocks minus the return on a portfolio of big stocks (SMB), and the return on a portfolio with high BE/ME minus the return on portfolio with low BE/ME (HML).

Remark 13.13 *(The Fama-French factors) The SMB and HML are created by a double sort. First, classify firms according the book/market value: low (growth stocks, using 30th percentile as cutoff), neutral or high (value stocks, using 70th percentile as cutoff). Second, classify firms according to size: small or big, using the median as a cutoff. Create six value weighted portfolios from the intersection of those groups*

	Low book/market	*Medium book/market*	*High book/market*
Small:	*Small Growth (SG)*	*Small Neutral (SN)*	*Small Value (SV)*
Big:	*Big Growth (BG)*	*Big Neutral (BN)*	*Big Value (BV)*

The SMB is the average of the small portfolios minus the average of the big portfolios: $SMB = 1/3(SG + SN + SV) - 1/3(BG + BN + BV)$. *Rearranging gives* $SMB = 1/3(SG-BG)+1/3(SN-BN)+SV)+1/3(SV-BV)$, *which shows that it represents the return on small stocks (for a given book/market) minus the return on big stocks (for same book/market). The HML is the average of the value stocks minus the growth stocks,* $HML = 1/2(SV + BV) - 1/2(SG + BG)$, *which can be rearranged as* $HML = 1/2(SV - SG) + 1/2(BV - BG)$, *which shows that it represents the return on value stocks (for a given size) minus the return on growth stocks (for the same size).*

Empirical Example 13.14 *(A 3-factor model for the 25 FF portfolios) The Fama-French three-factor model is rejected at traditional significance levels (see Campbell, Lo, and MacKinlay (1997) Table 6.1 or Fama and French (1993) Table 9c), but it can still capture a fair amount of the variation of expected returns—see Figures 13.4–13.5.*

Empirical Example 13.15 *(A 3-factor model for 10 industry portfolios) Figure 13.6 suggests that the 3-factor FF models works poorly for industry portfolios.*

Is it a trivial finding that the 25 FF portfolios are better explained once we use the HML and SMB factors? No, as argued by Fama and French (1996) it just shows that there is a common (possibly unknown) pricing factor. To see that in a simplified setting, suppose excess returns are generated by some one-factor model $R^e_{it} = \beta_i f_t + \varepsilon_{it}$, although we

may not know what the factor is. In addition, assume that the portfolio (HML or SMB, say) we create is just an equally weighted average across all n assets like

$$x_t = \beta_x f_t + \sum_{i=1}^{n} \varepsilon_{it}/n, \text{ where } \beta_x = \sum_{i=1}^{n} \beta_i/n. \tag{13.32}$$

(With an appropriate interpretation of the signs of the betas, this could actually be a long-short portfolio.) Regressing R_{it}^e on this portfolio gives a slope coefficient $\gamma_i = \text{Cov}(R_{it}^e, x_t)/\text{Var}(x_t)$. If we assume that all residuals are uncorrelated with the factor and with each other, then the numerator of γ_i can be simplified as

$$\text{Cov}(R_{it}^e, x_t) = \beta_i \beta_x \text{Var}(f_t) + \text{Var}(\varepsilon_{it})/n. \tag{13.33}$$

The last term is due to the fact that ε_{it} shows up both in R_{it}^e and x_t, but its importance decreases as the number of assets in the portfolio (n) increases. This shows that if the cross-section (n) is large, then γ_i depends mostly on the first term. Clearly, the first term is non-zero if all three ingredients are non-zero. This means that both asset i and the portfolio are exposed to a (time-varying) factor, although we may not know what that factor represents. However, the pattern of γ_i across assets may give us a clue. (There are clearly other methods to investigate if there are common factors, for instance, principal component analysis.)

Remark 13.16 *(Factor structure after having controlled for the market movements*) If the purpose is to investigate if there is a remaining factor structure after having controlled for the market movements, we can do the following. First, create "abnormal returns" as $R_{it}^e - \hat{b}_i R_{mt}^e$, where $\hat{b}_i$ is the coefficient obtained from regressing R_{it}^e on R_{mt}^e (and a constant). Then, replace R_{it}^e in (13.32)–(13.33) with this abnormal return. By the properties of OLS, this gives the same as running multiple regressions using R_{mt}^e and x_t as regressors (this is the Frisch-Waugh theorem).*

13.4.6 Calendar Time and Cross Sectional Regression

To investigate how the performance (alpha) or exposure (betas) of different investors/funds are related to investor/fund characteristics, we often use the *calendar time* approach. First define M discrete investor groups (for instance, age 18–30, 31–40, etc) and calculate their respective average excess returns ($\bar{R}_{jt}^e$ for group j)

$$\bar{R}_{jt}^e = \frac{1}{N_j} \sum_{i \in \text{Group } j} R_{it}^e, \tag{13.34}$$

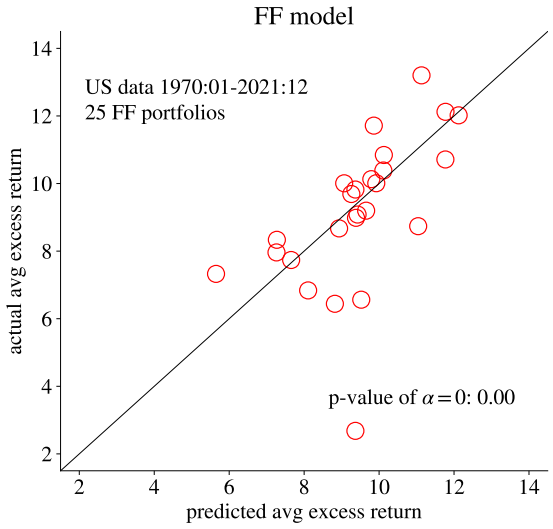


Figure 13.4: FF, FF portfolios

where N_j is the number of individuals in group j.

Then, we run a factor model

$$\bar{R}^e_{jt} = \alpha_j + \beta_j' f_t + v_{jt}, \text{ for } j = 1, 2, \ldots, M \tag{13.35}$$

where f_t typically includes various return factors (for instance, excess returns on equity and bonds). By estimating these M equations as a SURE system with White's (or Newey-West's) covariance estimator, it is straightforward to test various hypotheses, for instance, that the intercept (the "alpha") is higher for the Mth group than for the for first group.

Example 13.17 *(Calendar time approach with two investor groups) With two investor groups, estimate the following SURE system*

$$\bar{R}^e_{1t} = \alpha_1 + \beta_1' f_t + v_{1t},$$
$$\bar{R}^e_{2t} = \alpha_2 + \beta_2' f_t + v_{2t}.$$

The calendar time approach is straightforward and the cross-sectional correlations are fairly easy to handle (in the SURE approach). However, it forces us to define discrete investor groups—which makes it hard to handle several different types of investor characteristics (for instance, age, trading activity and income) at the same time.

The *cross sectional regression* approach is to first estimate the factor model for each

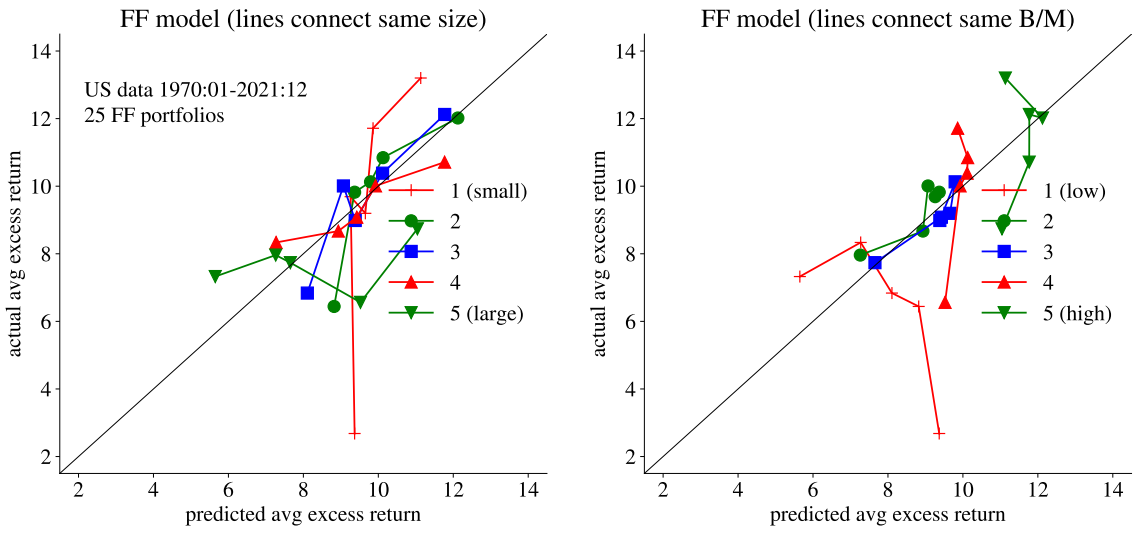


Figure 13.5: FF, FF portfolios

investor

$$R_{it}^e = \alpha_i + \beta_i' f_t + \varepsilon_{it}, \text{ for } i = 1, 2, \ldots, N \tag{13.36}$$

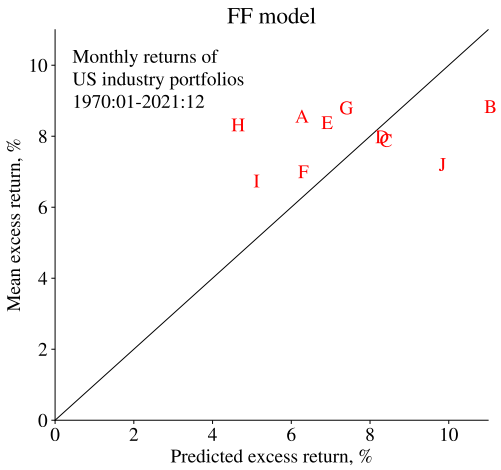
and to then regress the (estimated) betas for the pth factor (for instance, the intercept) on the investor characteristics

$$\hat{\beta}_{pi} = z_i' c_p + w_{pi}. \tag{13.37}$$

In this second-stage regression, the investor characteristics z_i could be a dummy variable (for an age group, say) or a continuous variable (age, say). Notice that using a continuous investor characteristics assumes that the relation between the characteristics and the beta is linear—something that is not assumed in the calendar time approach. (This saves degrees of freedom, but may sometimes be a very strong assumption.) However, a potential problem with the cross sectional regression approach is that it is often important to account for the cross-sectional correlation of the residuals.

13.5 Testing Multi-Factor Models (General Factors)

Reference: Cochrane (2005) 12.2; Campbell, Lo, and MacKinlay (1997) 6.2.3 and 6.3



	α	t-stat	σ
A (NoDur)	2.24	1.90	8.37
B (Durbl)	-2.28	-1.13	14.35
C (Manuf)	-0.58	-0.68	6.03
D (Enrgy)	-0.38	-0.17	15.59
E (HiTec)	1.42	1.05	9.62
F (Telcm)	0.64	0.43	10.44
G (Shops)	1.35	1.06	9.03
H (Hlth)	3.62	2.48	10.33
I (Utils)	1.57	1.01	10.96
J (Other)	-2.69	-3.43	5.56

Fama-French model
Factors: US market, SMB (size), and HML (book-to-market)
Predicted excess return: $\beta_m R_m^e + \beta_{SMB} R_{SMB} + \beta_{HML} R_{HML}$

α and σ (std of residual) are in annualized %

p-val for testing if all $\alpha_i = 0$: 0.01

Figure 13.6: Three-factor model, US industry portfolios

13.5.1 GMM Estimation with General Factors

Linear factor models imply that all expected excess returns are linear functions of the same vector of factor risk premia (λ)

$$\mathrm{E}\, R_{it}^e = \beta_i' \lambda, \text{ where } \lambda \text{ is } K \times 1, \text{ for } i = 1, \ldots n, \tag{13.38}$$

where the β_i are the loading of asset i on the factors, as estimated from (13.21).

Stacking the test assets gives

$$\mathrm{E} \begin{bmatrix} R_{1t}^e \\ \vdots \\ R_{nt}^e \end{bmatrix} = \begin{bmatrix} \beta_{11} & \ldots & \beta_{1K} \\ \vdots & \ddots & \vdots \\ \beta_{n1} & \ldots & \beta_{nK} \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \vdots \\ \lambda_K \end{bmatrix}, \text{ or}$$

$$\mathrm{E}\, R_t^e = \beta \lambda, \tag{13.39}$$

where β is $n \times K$.

When the factors are excess returns, then the factor risk premia must equal the expected excess returns of those factors, $\lambda = \mathrm{E}\, f_t$. (To see this, let the factor also be one of the test assets. It will then get a beta equal to unity on itself.) In any case, if a factor

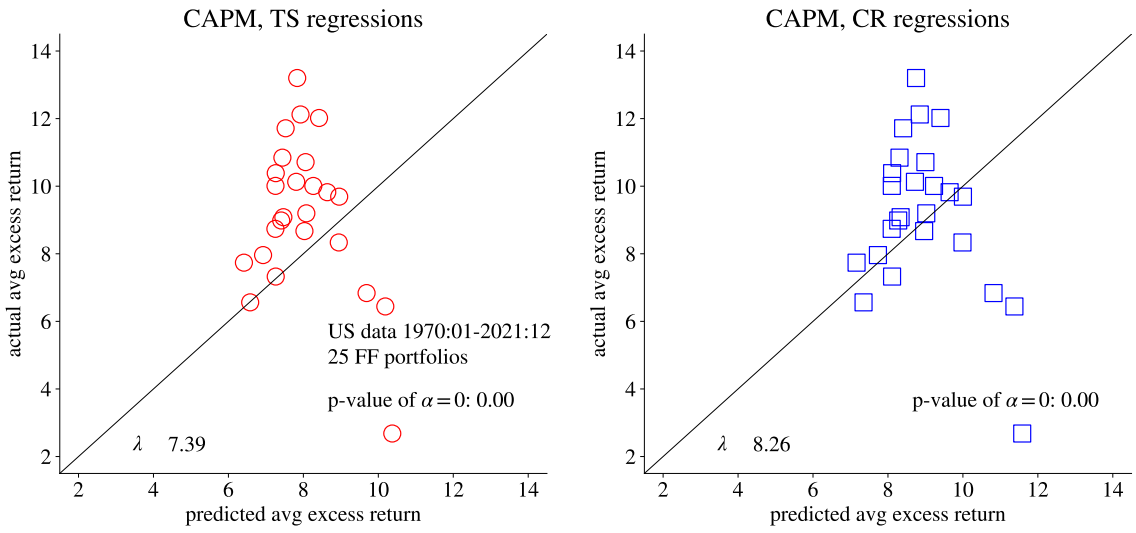


Figure 13.7: CAPM on the 25 FF portfolios, TS and CR regressions

risk premium is negative, then assets that are positively exposed to it (positive betas) will have a negative risk premium, and vice versa.

Remark 13.18 *(Factor mimicking portfolios) It is more difficult to estimate and test a model with general factors than a model with excess return factors. A common approach to get around the difficulties is to replace any general factor with the linear combination of excess returns that best mimics the general factor. This linear combination can be constructed be either forming a regression of the general factor on a vector of excess returns, or by creating an arbitrage portfolio that is long assets that are highly correlated with the general factor and short assets that are less or even negatively correlated with the factor.*

The old way of testing this is to do a two-step estimation: first, estimate the β_i vectors in a time series model like (13.21) (equation by equation); second, use $\hat{\beta}_i$ as regressors in a regression equation of the type (13.38) with a residual added

$$\bar{R}_i^e = \hat{\beta}_i' \lambda + u_i, \tag{13.40}$$

where $\bar{R}_i^e = \Sigma_{t=1}^T R_{it}^e / T$ is the (time-series) average of R_{it}^e.

It is then tested if $u_i = 0$ for all assets $i = 1, \ldots, n$. This approach is often called a *cross-sectional* regression while the previous tests are time series regressions. Clearly,

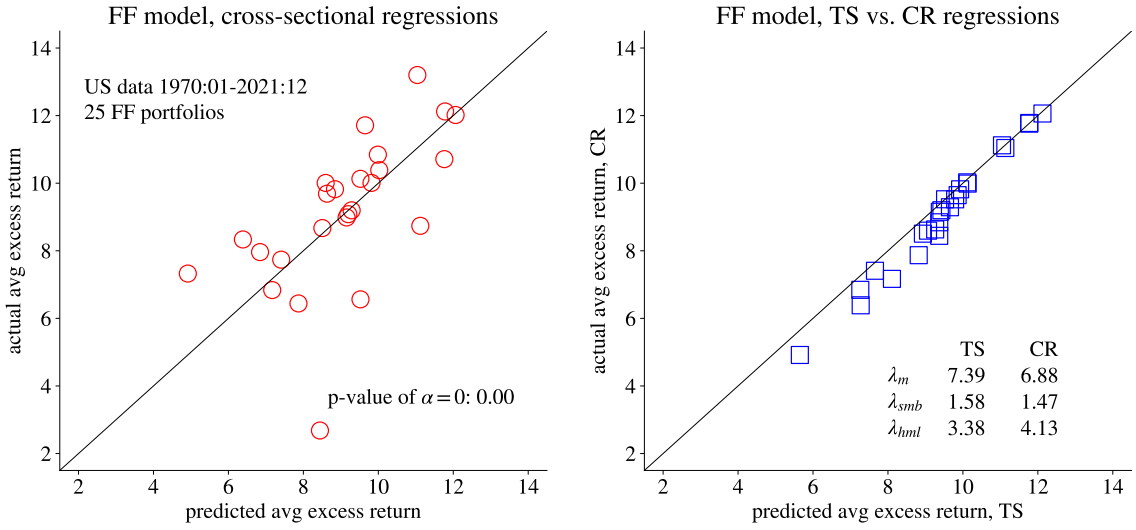


Figure 13.8: FF model

this approach relies on the assumption that the betas are indeed non-zero (and preferably not too similar across the test assets).

An issue with the cross-sectional approach is that we have to account for the fact that the regressors in the second step, $\hat{\beta}_i$, are just estimates and therefore contain estimation errors. This "errors-in-variables" problem is likely to have two effects *(i)* it gives a downwards bias of the estimates of λ and an upward bias of the mean of the fitted residuals; and *(ii)* invalidate the standard expression of the test of λ.

A way to handle these problems is to combine the moment conditions for the time series regressions (13.29) (to estimate β) with (13.39) (to estimate λ) to get a joint system

$$\mathrm{E}\, g_t(\alpha, \beta, \lambda) = \mathrm{E}\begin{bmatrix} \begin{bmatrix} 1 \\ f_t \end{bmatrix} \otimes (R_t^e - \alpha - \beta f_t) \\ R_t^e - \beta\lambda \end{bmatrix} = \mathbf{0}_{n(1+K+1)\times 1}. \tag{13.41}$$

We can then test the overidentifying restrictions of the model. There are $n(1+K+1)$ moment condition (for each asset we have one moment condition for the constant, K moment conditions for the K factors, and one moment condition corresponding to the restriction on the linear factor model). There are only $n(1+K)+K$ parameters (n in α, nK in β and K in λ). We therefore have $n-K$ overidentifying restrictions which can be tested with a chi-square test. From the GMM estimation using (13.41) we get estimates of the factor risk premia and also the variance-covariance of them. This allows us to not

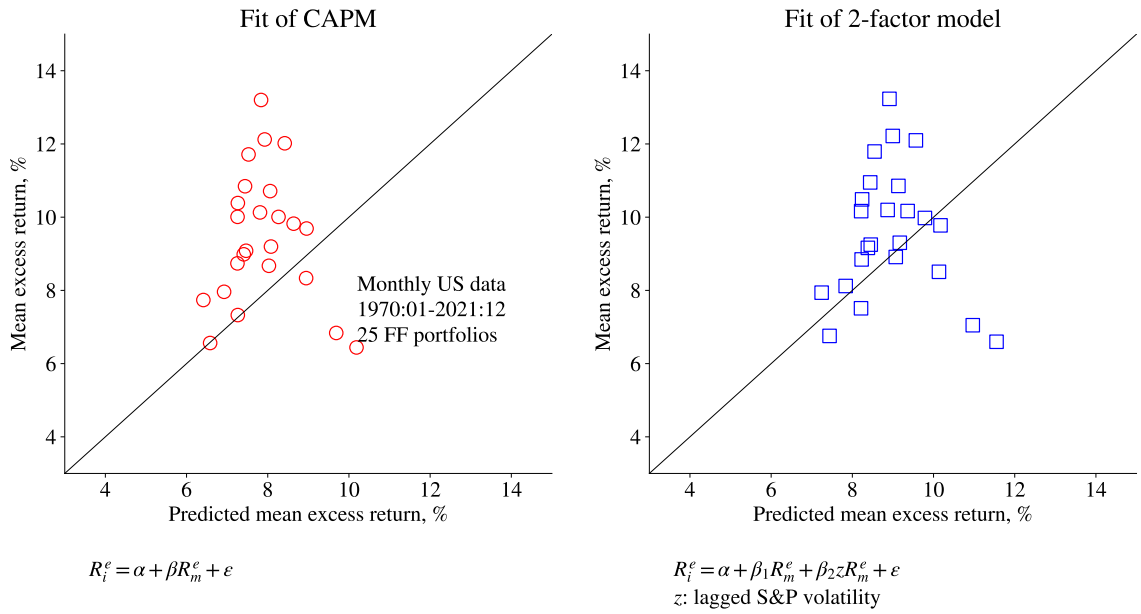


Figure 13.9: CAPM and 2-factor model

only test the moment conditions, but also to characterize the risk factors and to test if they are priced (each of them, or perhaps all jointly) by using a Wald test. Notice that this is, in general, a non-linear estimation problem, since the parameters in β multiply the parameters in λ.

Empirical Example 13.19 *(Time series vs. cross-sectional regression, CAPM and 3-factor model for the 25 FF portfolios) See Figures 13.7–13.8 for an empirical example based on the CAPM and the FF model, and for a comparison with the results from the time series approach. For CAPM, the fit differs across the two methods (since the implied factor risk premia do). For the 3-factor FF model, the fit is more similar across the methods.*

Empirical Example 13.20 *(CAPM vs. 2-factor model) See Figure 13.9 for an empirical comparison of CAPM with a 2-factor model (where one of the factors is not an excess return).*

One approach to estimate the model is to specify a weighting matrix W and then solve a minimization problem like (13.30). The test is based on a quadratic form of the moment conditions, $T\bar{g}(b)'\Psi^{-1}\bar{g}(b)$ which has a chi-square distribution if the correct Ψ matrix is used. In the special case of $W = S_0^{-1}$, the distribution is given by Remark 13.3. For

other choices of the weighting matrix, the expression for the covariance matrix is more complicated.

It is straightforward to show that the Jacobian of these moment conditions (with respect to $\text{vec}(\alpha, \beta, \lambda)$) is

$$D_0 = -\begin{bmatrix} \frac{1}{T}\sum_{t=1}^{T}\left(\begin{bmatrix} 1 \\ f_t \end{bmatrix}\begin{bmatrix} 1 \\ f_t \end{bmatrix}'\right) \otimes I_n & \mathbf{0}_{n(1+K)\times K} \\ \begin{bmatrix} 0 & \lambda' \end{bmatrix} \otimes I_n & \beta_{n\times K} \end{bmatrix} \tag{13.42}$$

where the upper left block is similar to the expression for the case with excess return factors (13.12), while the other blocks are new.

Example 13.21 *(Two assets and one factor) we have the moment conditions*

$$\text{E}\, g_t(\alpha_1, \alpha_2, \beta_1, \beta_2, \lambda) = \text{E}\begin{bmatrix} R_{1t}^e - \alpha_1 - \beta_1 f_t \\ R_{2t}^e - \alpha_2 - \beta_2 f_t \\ f_t(R_{1t}^e - \alpha_1 - \beta_1 f_t) \\ f_t(R_{2t}^e - \alpha_2 - \beta_2 f_t) \\ R_{1t}^e - \beta_1 \lambda \\ R_{2t}^e - \beta_2 \lambda \end{bmatrix} = \mathbf{0}_{6\times 1}.$$

There are then 6 moment conditions and 5 parameters, so there is one overidentifying restriction to test. Note that with one factor, then we need at least two assets for this testing approach to work ($n - K = 2 - 1$). In general, we need at least one more asset than factors. In this case, the Jacobian is

$$\begin{aligned} \frac{\partial \bar{g}}{\partial[\alpha_1, \alpha_2, \beta_1, \beta_2, \lambda]'} &= -\frac{1}{T}\sum_{t=1}^{T}\begin{bmatrix} 1 & 0 & f_t & 0 & 0 \\ 0 & 1 & 0 & f_t & 0 \\ f_t & 0 & f_t^2 & 0 & 0 \\ 0 & f_t & 0 & f_t^2 & 0 \\ 0 & 0 & \lambda & 0 & \beta_1 \\ 0 & 0 & 0 & \lambda & \beta_2 \end{bmatrix} \\ &= -\begin{bmatrix} \frac{1}{T}\sum_{t=1}^{T}\left(\begin{bmatrix} 1 \\ f_t \end{bmatrix}\begin{bmatrix} 1 \\ f_t \end{bmatrix}'\right) \otimes I_2 & \mathbf{0}_{4\times 1} \\ [0, \lambda] \otimes I_2 & \beta \end{bmatrix}. \end{aligned}$$

13.5.2 Traditional Cross-Sectional Regressions as a Special Case

Instead of estimating the overidentified model (13.41) (by specifying a weighting matrix), we could combine the moment equations so they become equal to the number of parameters. This can be done, by specifying a matrix A and combine as $A\,\mathrm{E}\,g_t = \mathbf{0}$. This does not generate any overidentifying restrictions, but it still allows us to test hypotheses about the moment conditions and about λ. One possibility is to let the upper left block of A be an identity matrix and just combine the last n moment conditions, $R_t^e - \beta\lambda$, to just K moment conditions

$$A\,\mathrm{E}\,g_t = \mathbf{0}_{[n(1+K)+K]\times 1} \tag{13.43}$$

$$\begin{bmatrix} I_{n(1+K)} & \mathbf{0}_{n(1+K)\times n} \\ \mathbf{0}_{K\times n(1+K)} & \theta_{K\times n} \end{bmatrix} \mathrm{E}\begin{bmatrix} \begin{bmatrix} 1 \\ f_t \end{bmatrix} \otimes (R_t^e - \alpha - \beta f_t) \\ R_t^e - \beta\lambda \end{bmatrix} = \mathbf{0} \tag{13.44}$$

$$\mathrm{E}\begin{bmatrix} \begin{bmatrix} 1 \\ f_t \end{bmatrix} \otimes (R_t^e - \alpha - \beta f_t) \\ \theta(R_t^e - \beta\lambda) \end{bmatrix} = \mathbf{0} \tag{13.45}$$

Here A has $n(1+K)+K$ rows (which equals the number of parameters (α, β, λ)) and $n(1+K+1)$ columns (which equals the number of moment conditions). (Notice also that θ is $K \times n$, β is $n \times K$ and λ is $K \times 1$.)

Remark 13.22 *(Calculation of the estimates based on (13.44)) In this case, we can estimate α and β with LS equation by equation—as a standard time-series regression of a factor model. To estimate the $K \times 1$ vector λ, notice that we can solve the second set of K moment conditions as*

$$\theta\,\mathrm{E}(R_t^e - \beta\lambda) = \mathbf{0}_{K\times 1} \text{ or } \lambda = (\theta\beta)^{-1}\,\theta\,\mathrm{E}\,R_t^e,$$

which is just like a cross-sectional instrumental variables regression of $\mathrm{E}\,R_t^e = \beta\lambda$ (with β being the regressors, θ the instruments, and $\mathrm{E}\,R_t^e$ the dependent variable).

With $\theta = \beta'$, we get the traditional cross-sectional approach (13.38). The only difference is we here take the uncertainty about the generated betas into account (in the testing). Alternatively, let Σ be the covariance matrix of the residuals from the time-series estimation of the factor model. Then, using $\theta = \beta'\Sigma^{-1}$ gives a traditional GLS cross-sectional approach.

Empirical Example 13.23 *(Factor risk premia for the 3-factor model, different methods) Table 14.2 shows estimates of the factor risk premia from several methods based on the 25 FF portfolios.*

	Data	CR	FMB1	FMB2
Market	7.39	6.88	6.88	−7.36
	(2.20)	(2.29)	(2.23)	(3.94)
SMB	1.58	1.47	1.47	1.08
	(1.48)	(1.55)	(1.52)	(1.52)
HML	3.38	4.13	4.13	3.73
	(1.44)	(1.60)	(1.48)	(1.48)

Table 13.3: Different estimates of factor risk premia, annualized %. Numbers in (parentheses) are standard deviations. The 25 FF portfolios 1970:01-2021:12. Data are the mean excess returns of the factors; CR are estimates of the factor risk premia from a cross-sectional regression; FMB1 are from Fama-MacBeth without intercept in the cross-sectional regression; FMB2 are from Fama-MacBeth with intercept in the cross-sectional regression.In both FMB regressions, the betas are estimated from the full sample.

To test the asset pricing implications, we test if the moment conditions $\mathrm{E}\, g_t = \mathbf{0}$ in (13.43) are satisfied at the estimated parameters. The test is based on a quadratic form of the moment conditions, $T\bar{g}(b)'\Psi^{-1}\bar{g}(b)$ which has a chi-square distribution if the correct Ψ matrix is used (typically more complicated than in Remark 13.3).

Example 13.24 *(LS cross-sectional regression, two assets and one factor) With the moment conditions in Example (13.21) and the weighting vector $\theta = [\beta_1, \beta_2]$ (13.45) is*

$$A\,\mathrm{E}\, g_t(\alpha_1, \alpha_2, \beta_1, \beta_2, \lambda) = \mathrm{E}\begin{bmatrix} R^e_{1t} - \alpha_1 - \beta_1 f_t \\ R^e_{2t} - \alpha_2 - \beta_2 f_t \\ f_t(R^e_{1t} - \alpha_1 - \beta_1 f_t) \\ f_t(R^e_{2t} - \alpha_2 - \beta_2 f_t) \\ \beta_1(R^e_{1t} - \beta_1\lambda) + \beta_2(R^e_{2t} - \beta_2\lambda) \end{bmatrix} = \mathbf{0}_{5\times 1},$$

which has as many parameters as moment conditions. The test of the asset pricing model

is then to test if

$$\mathrm{E}\, g_t(\alpha_1, \alpha_2, \beta_1, \beta_2, \lambda) = \mathrm{E} \begin{bmatrix} R^e_{1t} - \alpha_1 - \beta_1 f_t \\ R^e_{2t} - \alpha_2 - \beta_2 f_t \\ f_t(R^e_{1t} - \alpha_1 - \beta_1 f_t) \\ f_t(R^e_{2t} - \alpha_2 - \beta_2 f_t) \\ R^e_{1t} - \beta_1 \lambda \\ R^e_{2t} - \beta_2 \lambda \end{bmatrix} = \mathbf{0}_{6\times 1},$$

are satisfied at the estimated parameters.

Example 13.25 *(Structure of $\theta\,\mathrm{E}(R^e_t - \beta\lambda)$) If there are 2 factors and three test assets, then $0_{2\times 1} = \theta\,\mathrm{E}(R^e_t - \beta\lambda)$ is*

$$\begin{bmatrix} 0 \\ 0 \end{bmatrix} = \begin{bmatrix} \theta_{11} & \theta_{12} & \theta_{13} \\ \theta_{21} & \theta_{22} & \theta_{23} \end{bmatrix} \left(\begin{bmatrix} \mathrm{E}\, R^e_{1t} \\ \mathrm{E}\, R^e_{2t} \\ \mathrm{E}\, R^e_{3t} \end{bmatrix} - \begin{bmatrix} \beta_{11} & \beta_{12} \\ \beta_{21} & \beta_{22} \\ \beta_{31} & \beta_{32} \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \end{bmatrix} \right).$$

13.5.3 What If the Factors Are Excess Returns?*

It would (perhaps) be natural if the tests discussed in this section coincided with those in Section 13.4 when the factors are in fact excess returns. That is *almost* so. The difference is that we here estimate the $K \times 1$ vector λ (factor risk premia) as a vector of free parameters, while the tests in Section 13.4 *impose* $\lambda = \mathrm{E}\, f_t$. This can be done in (13.44)–(13.45) by doing two things. First, define a new set of test assets by stacking the original test assets and the excess return factors

$$\tilde{R}^e_t = \begin{bmatrix} R^e_t \\ f_t \end{bmatrix}, \tag{13.46}$$

which is an $(n + K) \times 1$ vector. Second, define the $K \times (n + K)$ matrix θ as

$$\tilde{\theta} = \begin{bmatrix} \mathbf{0}_{K\times n} & I_K \end{bmatrix}. \tag{13.47}$$

(Clearly, the betas of f_t (as test assets) must equal I_K and their residuals must be zero. This means that the GLS approach to (13.45), $\theta = \beta'\Sigma^{-1}$, is conceptually the same as (13.47), since all weight is on the betas of f_t. However, (13.47) is numerically more robust.) Together, this gives

$$\lambda = \mathrm{E}\, f_t. \tag{13.48}$$

It is also straightforward to show that this gives precisely the same test statistics as the Wald test on the multifactor model (13.21).

Proof. (of (13.48)) The betas of the $\tilde{R}_t^e$ vector are

$$\tilde{\beta} = \begin{bmatrix} \beta_{n\times K} \\ I_K \end{bmatrix}.$$

The expression corresponding to $\theta \,\mathrm{E}(R_t^e - \beta\lambda) = \mathbf{0}$ is then

$$\begin{bmatrix} \mathbf{0}_{K\times n} & I_K \end{bmatrix} \mathrm{E} \begin{bmatrix} R_t^e \\ f_t \end{bmatrix} - \begin{bmatrix} \mathbf{0}_{K\times n} & I_K \end{bmatrix} \begin{bmatrix} \beta_{n\times K} \\ I_K \end{bmatrix} \lambda = \mathbf{0}, \text{ or}$$
$$\mathrm{E}\, f_t = \lambda.$$

■

Remark 13.26 *(Two assets, one excess return factor) By including the factors among the test assets and using the weighting vector $\theta = [0, 0, 1]$ gives*

$$A \,\mathrm{E}\, g_t(\alpha_1, \alpha_2, \alpha_3, \beta_1, \beta_2, \beta_3, \lambda) = \mathrm{E} \begin{bmatrix} R_{1t}^e - \alpha_1 - \beta_1 f_t \\ R_{2t}^e - \alpha_2 - \beta_2 f_t \\ f_t - \alpha_3 - \beta_3 f_t \\ f_t(R_{1t}^e - \alpha_1 - \beta_1 f_t) \\ f_t(R_{2t}^e - \alpha_2 - \beta_2 f_t) \\ f_t(f_t - \alpha_3 - \beta_3 f_t) \\ 0(R_{1t}^e - \beta_1\lambda) + 0(R_{2t}^e - \beta_2\lambda) + 1(f_t - \beta_3\lambda) \end{bmatrix} = \mathbf{0}_{7\times 1}.$$

Since $\alpha_3 = 0$ and $\beta_3 = 1$, this gives the estimate $\lambda = \mathrm{E}\, f_t$. There are 7 moment conditions and as many parameters. To test the asset pricing model, test if the following

moment conditions are satisfied at the estimated parameters

$$\mathrm{E}\, g_t(\alpha_1,\alpha_2,\alpha_3,\beta_1,\beta_2,\beta_3,\lambda) = \mathrm{E}\begin{bmatrix} R^e_{1t}-\alpha_1-\beta_1 f_t \\ R^e_{2t}-\alpha_2-\beta_2 f_t \\ f_t-\alpha_3-\beta_3 f_t \\ f_t(R^e_{1t}-\alpha_1-\beta_1 f_t) \\ f_t(R^e_{2t}-\alpha_2-\beta_2 f_t) \\ f_t(f_t-\alpha_3-\beta_3 f_t) \\ R^e_{1t}-\beta_1\lambda \\ R^e_{2t}-\beta_2\lambda \\ f_t-\beta_3\lambda \end{bmatrix} = \mathbf{0}_{9\times 1}.$$

In fact, this gives the same test statistic as when testing if α_1 and α_2 are zero in (13.11).

Remark 13.27 *(What is an excess return?*) Short answer: the return of a zero cost portfolio. More detailed answer: consider a portfolio with the (net) return*

$$R_p = v_1R_1 + v_2R_2 + v_3R_3 + (1-v_1-v_2-v_3)R_4,$$

where v_i is the portfolio weight on asset i which has the net return R_i. The balance $(1-v_1-v_2-v_3)$ is made up of asset 4 with the net return R_4 (which may be a riskfree asset). Rearrange as

$$R_p - R_4 = v_1(R_1-R_4) + v_2(R_2-R_4) + v_3(R_3-R_4).$$

Clearly, $R_p - R_4$ is an excess return—and it is a linear combination of other excess returns (even if v_1, v_2 and/or v_3 happen to be negative and they do not sum to unity). If $v_3 = -v_2$, then we can rearrange further to get

$$R_p - R_4 = v_1(R_1-R_4) + v_2(R_2-R_3).$$

This is still an excess return, although the "excess" on the right hand side is over different returns.

When Some (but Not All) of the Factors Are Excess Returns*

Partition the vector of factors as

$$f_t = \begin{bmatrix} Z_t \\ F_t \end{bmatrix}, \tag{13.49}$$

where Z_t is an $v \times 1$ vector of excess return factors and F_t is a $w \times 1$ vector of general factors ($K = v + w$).

It makes sense (and is econometrically efficient) to use the fact that the factor risk premia of the excess return factors are just their average excess returns (as in CAPM). This can be done in (13.44)–(13.45) by doing two things. First, define a new set of test assets by stacking the original test assets and the excess return factors

$$\tilde{R}_t^e = \begin{bmatrix} R_t^e \\ Z_t \end{bmatrix}, \tag{13.50}$$

which is an $(n + v) \times 1$ vector. Second, define the $K \times (n + K)$ matrix θ

$$\tilde{\theta} = \begin{bmatrix} \mathbf{0}_{v\times n} & I_v \\ \vartheta_{w\times n} & \mathbf{0}_{w\times v} \end{bmatrix}, \tag{13.51}$$

where ϑ is some $w \times n$ matrix. Together, this ensures that

$$\tilde{\lambda} = \begin{bmatrix} \lambda_Z \\ \lambda_F \end{bmatrix} = \begin{bmatrix} \mathrm{E}\, Z_t \\ (\vartheta\beta^F)^{-1}\vartheta(\mathrm{E}\, R_t^e - \beta^Z\lambda_Z) \end{bmatrix}, \tag{13.52}$$

where the β^Z and β^F are just betas of the original test assets on Z_t and F_t respectively—according to the partitioning

$$\beta_{n\times K} = \begin{bmatrix} \beta^Z_{n\times v} & \beta^F_{n\times w} \end{bmatrix}. \tag{13.53}$$

One possible choice of ϑ is $\vartheta = \beta^{F\prime}$, since then λ_F are the same as when running a cross-sectional regression of the expected "abnormal return" ($\mathrm{E}\, R_t^e - \beta^Z\lambda_Z$) on the betas ($\beta^F$).

Empirical Example 13.28 *(2-factor model on the 25 FF portfolios) Figure 13.10 illustrates that the fit of a 2-factor model, estimated with the cross sectional approach, depends on whether the risk premia of excess return factors are constrained to coincide with the average excess return on them or not.*

Proof. (of (13.52)) The betas of the $\tilde{R}_t^e$ vector are

$$\tilde{\beta} = \begin{bmatrix} \beta^Z_{n\times v} & \beta^F_{n\times w} \\ I_v & 0_{v\times w} \end{bmatrix}.$$

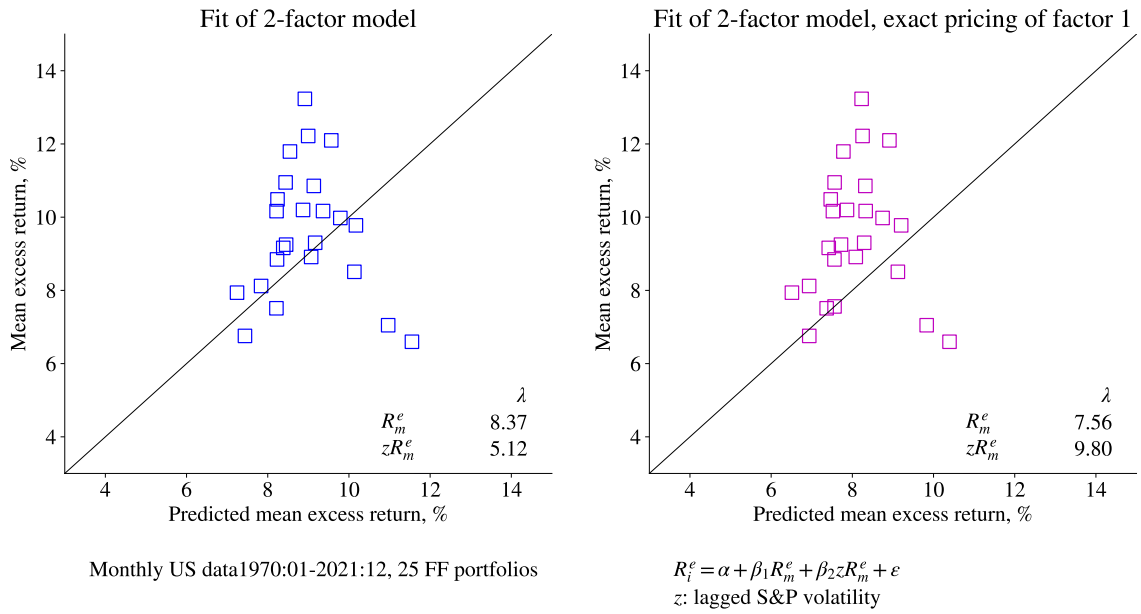


Figure 13.10: Fit of 2-factor model, CR regressions w/wo restrictions

The expression corresponding to $\theta \, \mathrm{E}(R^e_t - \beta\lambda) = \mathbf{0}$ is then

$$
\begin{aligned}
\tilde{\theta} \, \mathrm{E} \, \tilde{R}^e_t &= \tilde{\theta}\tilde{\beta}\tilde{\lambda} \\
\begin{bmatrix} \mathbf{0}_{v\times n} & I_v \\ \vartheta_{w\times n} & \mathbf{0}_{w\times v} \end{bmatrix} \begin{bmatrix} \mathrm{E} \, R^e_t \\ \mathrm{E} \, Z_t \end{bmatrix} &= \begin{bmatrix} \mathbf{0}_{v\times n} & I_v \\ \vartheta_{w\times n} & \mathbf{0}_{w\times v} \end{bmatrix} \begin{bmatrix} \beta^Z_{n\times v} & \beta^F_{n\times w} \\ I_v & 0_{v\times w} \end{bmatrix} \begin{bmatrix} \lambda_Z \\ \lambda_F \end{bmatrix} \\
\begin{bmatrix} \mathrm{E} \, Z_t \\ \vartheta_{w\times n} \, \mathrm{E} \, R^e_t \end{bmatrix} &= \begin{bmatrix} I_v & \mathbf{0}_{v\times w} \\ \vartheta_{w\times n}\beta^Z_{n\times v} & \vartheta_{w\times n}\beta^F_{n\times w} \end{bmatrix} \begin{bmatrix} \lambda_Z \\ \lambda_F \end{bmatrix}.
\end{aligned}
$$

The first v equations give

$$
\lambda_Z = \mathrm{E} \, Z_t.
$$

The remaining w equations give

$$
\begin{aligned}
\vartheta \, \mathrm{E} \, R^e_t &= \vartheta\beta^Z\lambda_Z + \vartheta\beta^F\lambda_F, \text{ so} \\
\lambda_F &= (\vartheta\beta^F)^{-1}\vartheta(\mathrm{E} \, R^e_t - \beta^Z\lambda_Z).
\end{aligned}
$$

■

Example 13.29 *(Structure of θ to identify λ for excess return factors) Continue Example 13.25 (where there are 2 factors and three test assets) and assume that $Z_t = R^e_{3t}$—so the first factor is really an excess return—which we have appended last to set of test assets.*

Then $\beta_{31} = 1$ *and* $\beta_{32} = 0$ *(regressing* Z_t *on* Z_t *and* F_t *gives the slope coefficients* 1 *and* $\dot{0}$*.) If we set* $(\theta_{11}, \theta_{12}, \theta_{13}) = (0, 0, 1)$*, then the moment conditions in Example 13.25 can be written*

$$\begin{bmatrix} 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 \\ \theta_{21} & \theta_{22} & \theta_{23} \end{bmatrix} \left(\begin{bmatrix} \mathrm{E}\, R_{1t}^e \\ \mathrm{E}\, R_{2t}^e \\ \mathrm{E}\, Z_t \end{bmatrix} - \begin{bmatrix} \beta_{11} & \beta_{12} \\ \beta_{21} & \beta_{22} \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \lambda_Z \\ \lambda_F \end{bmatrix} \right).$$

The first line reads

$$0 = \mathrm{E}\, Z_t - \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} \lambda_Z \\ \lambda_F \end{bmatrix}, \text{ so } \lambda_Z = \mathrm{E}\, Z_t.$$

13.5.4 Empirical Evidence

Chen, Roll, and Ross (1986) use a number of macro variables as factors—along with traditional market indices. They find that industrial production and inflation surprises are priced factors, while the market index might not be. Breeden, Gibbons, and Litzenberger (1989) and Lettau and Ludvigson (2001) estimate models where consumption growth is the factor—with mixed results.

13.6 Linear SDF Models

This section discusses how we can estimate and test the asset pricing equation

$$\mathrm{E}\, m_t R_t^e = 0. \tag{13.54}$$

Assume that the SDF is linear in the factors

$$m_t = \bar{m} + b'(f_t - \mathrm{E}\, f_t), \tag{13.55}$$

where the $K \times 1$ vector f_t contains the factors and where $\bar{m} \neq 0$. Combining with (13.54) gives the moment conditions

$$g_t(b) = m_t R_t^e = \bar{m} R_t^e + b'(f_t - \bar{f}_t) R_t^e, \tag{13.56}$$

where m_t is a scalar. There are K parameters (in b) and n moment conditions (the number of assets). The mean of the SDF cannot be estimated from excess returns (it could if we used returns), but it is straightforward to show that the choice of $\bar{m}$ (as long as not zero)

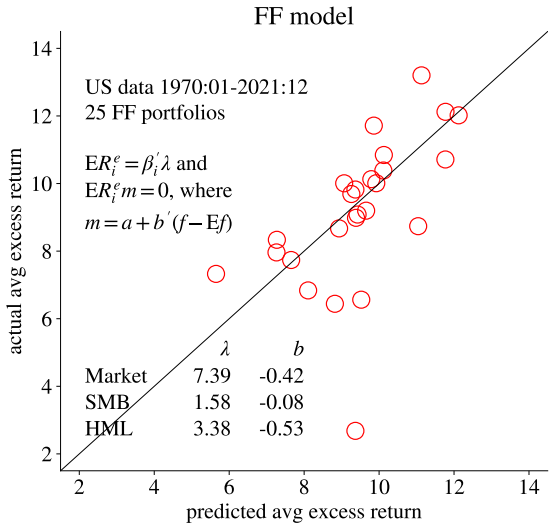


Figure 13.11: CAPM and quadratic model (co-skewness)

does not matter for the test based on excess returns.

Empirical Example 13.30 *(The implied SDF parameters from the 25 FF portfolios) See Figure 13.11 for estimates of the FF model.*

Remark 13.31 *(The SDF model and the mean SDF) Take expectations of the moment conditions (13.56) and set equal to zero to get*

$$b' \operatorname{Cov}(f_t, R_t^e) = -\bar{m} \operatorname{E} R_t^e.$$

This would be satisfied by $(\bar{m}, b) = (0, \mathbf{0})$*, which makes no sense. Instead, for any* $\bar{m} \neq 0$*, we could have*

$$\operatorname{E} R_t^e = \frac{-1}{\bar{m}} b' \operatorname{Cov}(f_t, R_t^e),$$

which allows us to test if there is a $K \times 1$ *vector* b *that prices all* n *assets, given how the covariance matrix of the returns and factors looks like.*

To estimate this model with a weighting matrix W, we minimize the loss function

$$J = \bar{g}(b)' W \bar{g}(b). \tag{13.57}$$

Alternatively, the moment conditions are combined into K effective conditions as

$$A_{K \times n} \bar{g}(b) = \mathbf{0}_{K \times 1}. \tag{13.58}$$

To test the asset pricing implications, we test if the moment conditions $\mathrm{E}\, g_t = \mathbf{0}$ are satisfied at the estimated parameters. The test is based on a quadratic form of the moment conditions, $T\bar{g}(b)'\Psi^{-1}\bar{g}(b)$ which has a chi-square distribution if the correct Ψ matrix is used.

13.6.1 SDF Models versus Linear Factor Models: The Tests*

Reference: Ferson (1995); Jagannathan and Wang (2002) (theoretical results); Cochrane (2005) 15 (empirical comparison); Bekaert and Urias (1996); and Söderlind (1999)

The test of the linear factor model and the test of the linear SDF model are (generally) not the same: they test the same implications of the models, but in slightly different ways. The moment conditions look a bit different—and combined with non-parametric methods for estimating the covariance matrix of the sample moment conditions, the two methods can give different results (in small samples, at least). Asymptotically, they are always the same, as showed by Jagannathan and Wang (2002).

There is one case where we know that the tests of the linear factor model and the SDF model are identical: when the factors are excess returns and the SDF is constructed to price these factors as well. To demonstrate this, let R^e_{1t} be a vector of excess returns on some benchmarks assets. Construct a stochastic discount factor as in Hansen and Jagannathan (1991):

$$m_t = \bar{m} + (R^e_{1t} - \bar{R}^e_{1t})'b, \tag{13.59}$$

where $\bar{m}$ is a constant and b is chosen to make m_t "price" R^e_{1t} in the sample, that is, so

$$\Sigma_{t=1}^{T}\, \mathrm{E}\, R^e_{1t} m_t / T = \mathbf{0}. \tag{13.60}$$

Consider the test assets with excess returns R^e_{2t}, and "SDF-based performance"

$$\bar{g}_{2t} = \frac{1}{T}\sum_{t=1}^{T} R^e_{2t} m_t. \tag{13.61}$$

Compare with the linear factor portfolio model

$$R^e_{2t} = \alpha + \beta R^e_{1t} + \varepsilon_t, \tag{13.62}$$

(where $\mathrm{E}\,\varepsilon_t = \mathbf{0}$ and $\mathrm{Cov}(R^e_{1t}, \varepsilon_t) = \mathbf{0}$) to see that the SDF-performance ("pricing error") is proportional to a traditional alpha

$$\bar{g}_{2t}/\bar{m} = \hat{\alpha}. \tag{13.63}$$

In both cases we are thus testing if α is zero or not.

Proof. (of (13.63)) (Here written in terms of population moments, to simplify the notation.) It follows directly that $b = -\text{Var}(R_{1t}^e)^{-1}\left(\text{E}\, R_{1t}^e \bar{m}\right)$. Using this and the expression for m_t in (13.61) gives

$$\text{E}\, g_{2t} = \text{E}\, R_{2t}^e \bar{m} - \text{Cov}\left(R_{2t}^e, R_{1t}^e\right) \text{Var}(R_{1t}^e)^{-1}\, \text{E}\, R_{1t}^e \bar{m}.$$

We now rewrite this equation in terms of the parameters in the factor portfolio model (13.62). The latter implies $\text{E}\, R_{2t}^e = \alpha + \beta\, \text{E}\, R_{1t}^e$, and the least squares estimator of the slope coefficients is $\beta = \text{Cov}\left(R_{2t}^e, R_{1t}^e\right) \text{Var}\left(R_{1t}^e\right)^{-1}$. Using these two facts in the equation above—and replacing population moments with sample moments, gives (13.63). ■

13.7 Conditional Factor Models

Reference: Cochrane (2005) 8; Ferson and Schadt (1996)

The simplest way of introducing conditional information is to simply state that the factors are not just the usual market indices or macro economic series: the factors are functions of them (this is sometimes called "scaled factors" to indicate that we scale the original factors with instruments). For instance, if R_{mt}^e is the return on the market portfolio and z_{t-1} is something else which is thought to be important for asset pricing (use theory), then the factors could be

$$f_{1t} = R_{mt}^e \text{ and } f_{2t} = z_{t-1} R_{mt}^e. \tag{13.64}$$

Since the second factor is not an excess return, the test is done as in (13.41).

An alternative interpretation of this is that we have only one factor, but that the coefficient of the factor is time varying. This is easiest seen by plugging in the factors in the time-series regression part of the moment conditions (13.41), $R_{it}^e = \alpha + \beta f_t + \varepsilon_{it}$,

$$\begin{aligned} R_{it}^e &= \alpha + \beta_1 R_{mt}^e + \beta_2 z_{t-1} R_{mt}^e + \varepsilon_{it} \\ &= \alpha + (\beta_1 + \beta_2 z_{t-1}) R_{mt}^e + \varepsilon_{it}. \end{aligned} \tag{13.65}$$

The first line looks like a two factor model with constant coefficients, while the second line looks like a one-factor model with a time-varying coefficient $(\beta_1 + \beta_2 z_{t-1})$. This is clearly just a matter of interpretation, since it is the same model (and is tested in the same way). This model can be estimated and tested as in the case of "general factors"—as

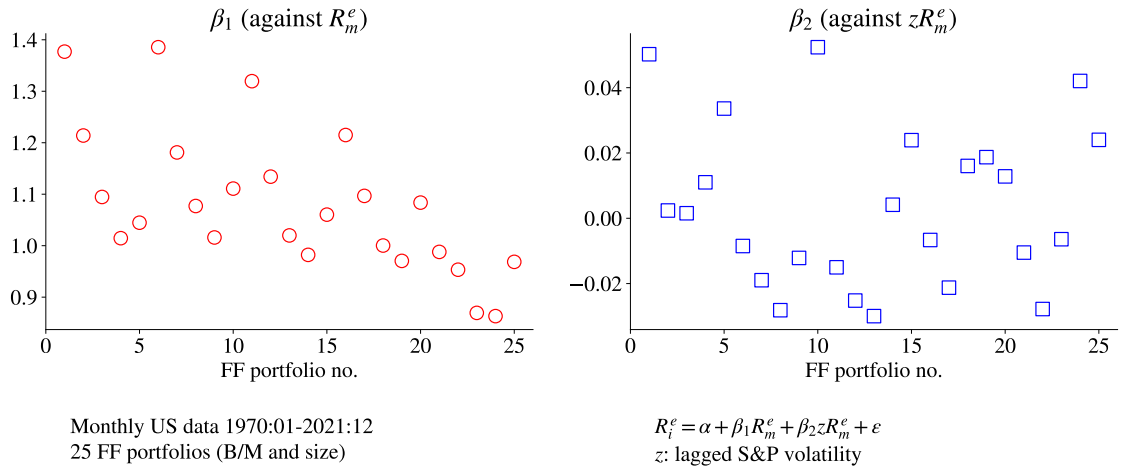


Figure 13.12: Betas of CAPM and 2-factor model

$z_{t-1} R_{mt}^e$ is not a traditional excess return.

Empirical Example 13.32 *(Conditional factor model estimated on the 25 FF portfolios) Figures 13.12–13.13 shows the betas of the conditional model. It seems as if the value firms (portfolios 5, 10, 15, 20, 25) have a somewhat higher exposure to the market when volatility is high. However, the time-variation is not marked. Therefore, the conditional (two-factor model) fits the cross-section of average returns only slightly better than CAPM—see Figure 13.9.*

Conditional models typically have more parameters than unconditional models, which is likely to give small samples issues (in particular with respect to the inference). It is important to remember some of the new factors (original factors times instruments) are probably not an excess returns, so the test is done with an LM test as in (13.41).

Remark 13.33 *(Dynamic Portfolios*) The returns on our factors, f_t, could be the excess return on* dynamic portfolios, *$R_{1t}^e = s_{t-1} \otimes R_{0t}^e$, where s_{t-1} are some information variables (not payoffs as before), for instance, lagged returns or market volatility, and R_{0t}^e are some basic benchmarks (S&P500 and bond, perhaps). The reason is that if R_{0t}^e are excess returns, so are $R_{1t}^e = s_{t-1} \otimes R_{0t}^e$. Therefore, the typical cross-sectional test (of $\mathrm{E}\, R^e = \beta'\lambda$) coincides with the test of the alpha—and also of zero SDF pricing errors. Notice also that the returns of our test assets, R_{it}^e, could be the excess return on dynamic strategies in terms of some basic test assets (mutual funds, say), $R_{2t}^e = z_{t-1} \otimes R_{pt}^e$, where z_{t-1} are information variables and R_{pt}^e are basic test assets. In this case, we are testing the performance of these dynamic strategies.*

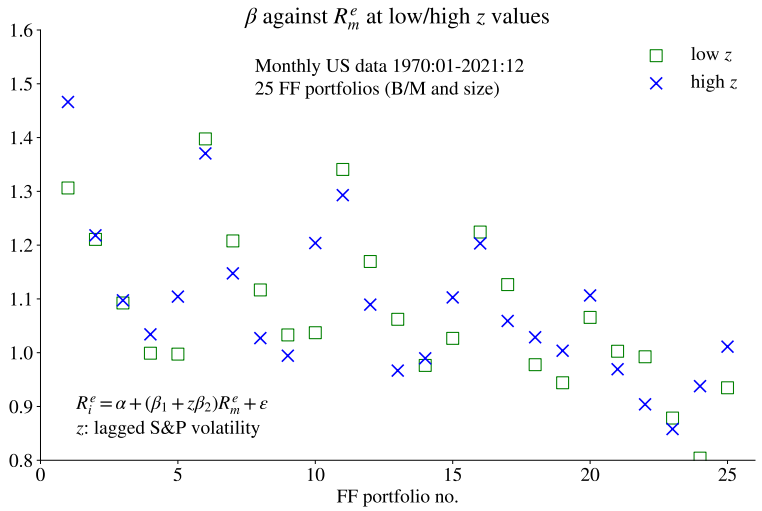


Figure 13.13: Betas of conditional CAPM

13.8 Conditional Models with "Regimes"

Reference: Christiansen, Ranaldo, and Söderlind (2011)

It is also possible to estimate non-linear factor models. The model could be piecewise linear or include higher order times. For instance, Treynor and Mazuy (1966) extend the CAPM regression by including a squared term (of the market excess return) to capture market timing.

Alternatively, the conditional model (13.65) could be changed so that the time-varying coefficients are non-linear in the information variable. In the simplest case, this could be dummy variable regression where the definition of the regimes is exogenous.

More ambitiously, we could use a smooth transition regression, which estimates both the "abruptness" of the transition between regimes as well as the cutoff point. Let $G(z)$ be a logistic (increasing but "S-shaped", sigmoidal) function

$$G(z) = \frac{1}{1 + \exp[-\gamma(z - c)]}, \tag{13.66}$$

where the parameter c is the central location (where $G(z) = 1/2$) and $\gamma > 0$ determines the steepness of the function (a high γ implies that the function goes quickly from 0 to 1 around $z = c$.) See Figure 13.14 for an illustration. A logistic smooth transition

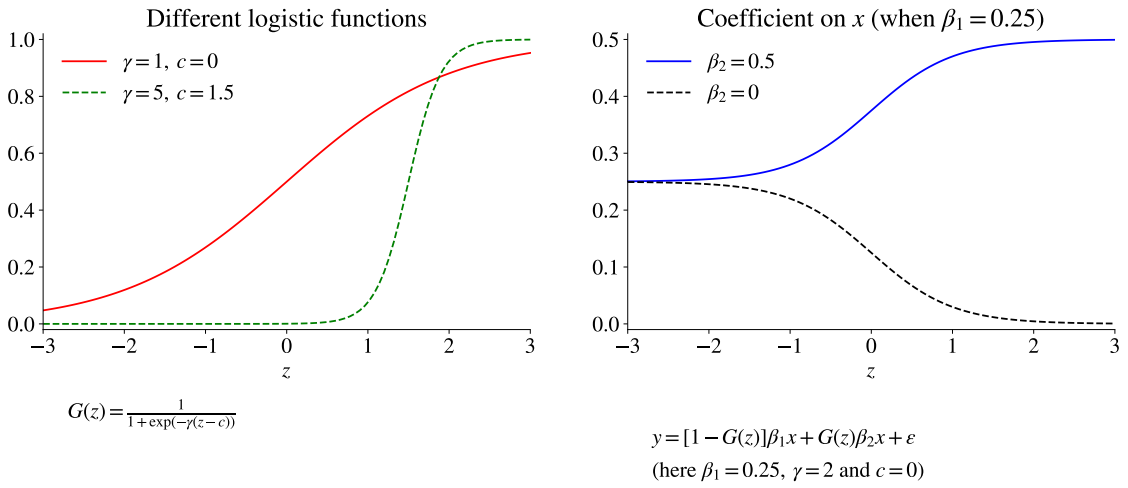


Figure 13.14: Logistic function and the effective slope coefficient in a Logistic smooth transition regression

regression is

$$\begin{aligned} R_{it}^{e} &= \beta(z_{t-1})' x_t + \varepsilon_t \\ &= \left\{ [1 - G(z_{t-1})]\, \beta_1' + G(z_{t-1}) \beta_2' \right\} x_t + \varepsilon_t \\ &= [1 - G(z_{t-1})]\, \beta_1' x_t + G(z_{t-1}) \beta_2' x_t + \varepsilon_t. \end{aligned} \tag{13.67}$$

At low z_t values, the regression coefficients are (almost) β_1 and at high z_t values they are (almost) β_2. See Figure 13.14 for an illustration.

Remark 13.34 *(NLS estimation) The parameter vector $(\gamma, c, \beta_1, \beta_2)$ is easily estimated by Non-Linear least squares (NLS) by concentrating the loss function: optimize (numerically) over (γ, c) and let (for each value of (γ, c)) the parameters (β_1, β_2) be the OLS coefficients on the vector of "regressors" $([1 - G(z_{t-1})]\, x_t, G(z_{t-1}) x_t)$.*

The most common application of this model is by letting $x_t = R_{i,t-s}^{e}$. This is the LSTAR model—logistic smooth transition auto regression model, see Franses and van Dijk (2000).

Empirical Example 13.35 *(L^* model estimated on the 25 FF portfolios) See Figures 13.15–13.16.*

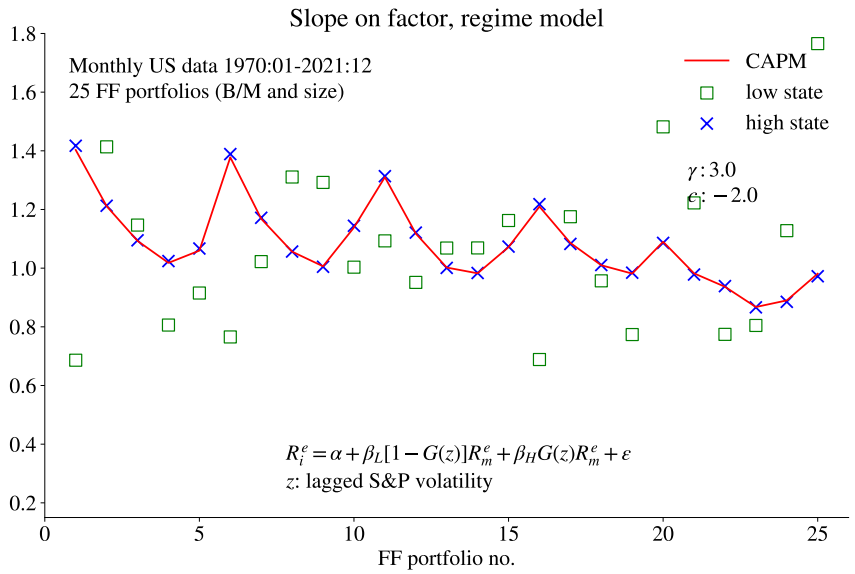


Figure 13.15: Betas on the market in the low and high regimes, 25 FF portfolios

13.9 Fama-MacBeth

Reference: Cochrane (2005) 12.3; Campbell, Lo, and MacKinlay (1997) 5.8; Fama and MacBeth (1973)

The Fama and MacBeth (1973) approach (called FMB below) is a bit different from the regression approaches discussed so far—although it seems most related to what we discussed in Section 13.5. The method has three steps, described below.

- First, estimate the betas β_i ($i = 1, \ldots, n$) from (13.1) (this is a time-series regression). This is often done on the whole sample—assuming the betas are constant. Sometimes, the betas are estimated separately for different sub samples (so we could let $\hat{\beta}_i$ carry a time subscript in the equations below).

- Second, run a cross sectional regression for every t. That is, for period t, estimate λ_t from the cross section (across the assets $i = 1, \ldots, n$) regression

$$R_{it}^e = \gamma_t + \lambda_t' \hat{\beta}_i + \varepsilon_{it}, \tag{13.68}$$

  where $\hat{\beta}_i$ are the regressors. Note the difference to the traditional cross-sectional approach discussed in (13.8), where the second stage regression regressed $\bar{R}_{it}^e$ (the time-series average of R_{it}^e) on $\hat{\beta}_i$, while the Fama-MacBeth approach runs one re-

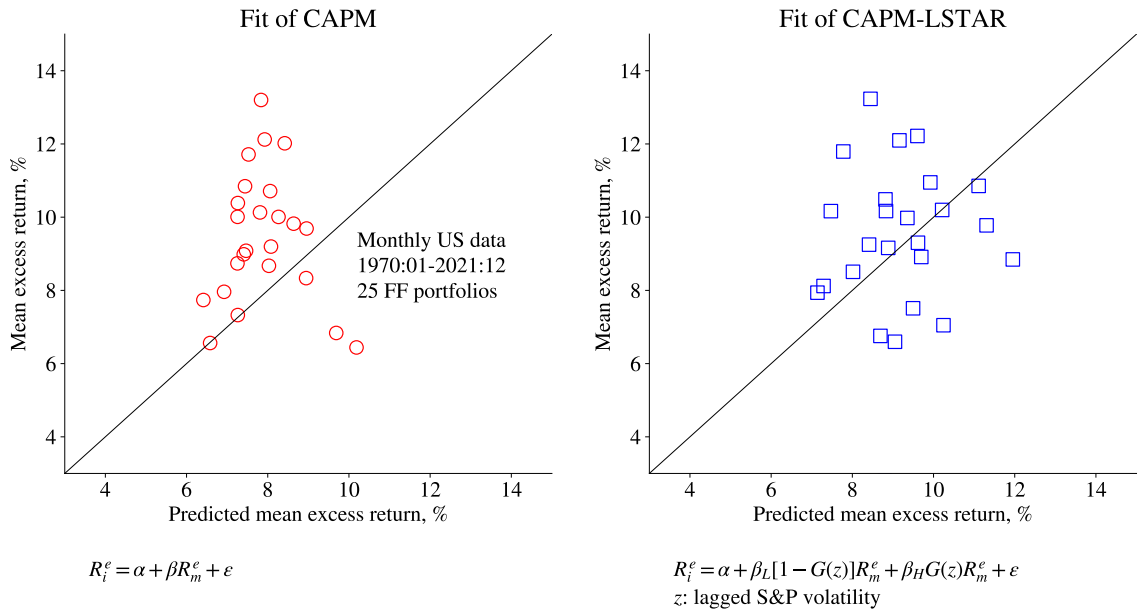


Figure 13.16: Test of 1 and 2-factor models, 25 FF portfolios

gression for every time period. The intercept γ_t (which capture time-fixed effects) is often dropped from the regression.

- Third, estimate the time averages

$$\hat{\varepsilon}_i = \frac{1}{T}\sum_{t=1}^{T}\hat{\varepsilon}_{it} \text{ for } i = 1,\ldots,n, \text{ (for every asset)} \tag{13.69}$$

$$\hat{\lambda} = \frac{1}{T}\sum_{t=1}^{T}\hat{\lambda}_t. \tag{13.70}$$

  Since $\hat{\lambda}_t$ measures the cross-sectional effect, $\hat{\lambda}$ is just the average of the cross-sectional effect.

The second step, using $\hat{\beta}_i$ as regressors, creates an errors-in-variables problem since $\hat{\beta}_i$ are estimated, that is, measured with an error. The effect of this is typically to bias the estimator of λ_t towards zero (and any intercept, or mean of the residual, is biased upward). One way to minimize this problem, used by Fama and MacBeth (1973), is to let the assets be portfolios of assets, for which we can expect that some of the individual noise in the first-step regressions to average out—and thereby make the measurement error in $\hat{\beta}$ smaller.

Empirical Example 13.36 *(FMB vs. other methods, 25 FF portfolios) See Table 14.2.*

Remark 13.37 *(Fama-MacBeth with constant betas) If the betas are (restricted to be) constant across time, then the estimate $\hat{\lambda}$ from (13.70) without intercept (exclude the γ_t term) is the same as from the traditional cross-sectional regression (13.40). To see that, consider the simplifying case of only one factor (so $\hat{\beta}_i$ is a scalar). Then, the FMB from the second step regression (13.68) without an intercept gives $\hat{\lambda}_t = (\Sigma_{i=1}^n \hat{\beta}_i R_{it}^e)/(\Sigma_{i=1}^n \hat{\beta}_i^2)$. Notice that the denominator is the same across time, so we can calculate the time-average (13.70) as*

$$\hat{\lambda} = \frac{1}{T}\sum_{t=1}^{T} \frac{\sum_{i=1}^n \hat{\beta}_i R_{it}^e}{\Sigma_{i=1}^n \hat{\beta}_i^2} = \frac{\sum_{i=1}^n \hat{\beta}_i \bar{R}_{it}^e}{\Sigma_{i=1}^n \hat{\beta}_i^2},$$

which is the same as from the CR approach.

We clearly want portfolios which have different betas, or else the second step regression (13.68) does not work. Fama and MacBeth (1973) choose to construct portfolios according to some initial estimate of asset specific betas. Another way to deal with the errors-in-variables problem is to adjust the tests. Jagannathan and Wang (1996) and Jagannathan and Wang (1998) discuss the asymptotic distribution of this estimator.

We can test the model by studying if $\varepsilon_i = 0$ (recall from (13.69) that ε_i is the time average of the residual for asset i, ε_{it}), by forming a t-test $\hat{\varepsilon}_i / \text{Std}(\hat{\varepsilon}_i)$. Fama and MacBeth (1973) suggest that the standard deviation should be found by studying the time-variation in $\hat{\varepsilon}_{it}$. In particular, they suggest that the variance of $\hat{\varepsilon}_{it}$ (not $\hat{\varepsilon}_i$) can be estimated by the (average) squared variation around its mean

$$\text{Var}(\hat{\varepsilon}_{it}) = \frac{1}{T}\sum_{t=1}^{T} (\hat{\varepsilon}_{it} - \hat{\varepsilon}_i)^2 . \tag{13.71}$$

Since $\hat{\varepsilon}_i$ is the sample average of $\hat{\varepsilon}_{it}$, the variance of the former is the variance of the latter divided by T (the sample size)—provided $\hat{\varepsilon}_{it}$ is iid. That is,

$$\text{Var}(\hat{\varepsilon}_i) = \frac{1}{T}\text{Var}(\hat{\varepsilon}_{it}) = \frac{1}{T^2}\sum_{t=1}^{T} (\hat{\varepsilon}_{it} - \hat{\varepsilon}_i)^2 . \tag{13.72}$$

A similar argument leads to the variance of $\hat{\lambda}$

$$\text{Var}(\hat{\lambda}) = \frac{1}{T^2}\sum_{t=1}^{T} (\hat{\lambda}_t - \hat{\lambda})^2 . \tag{13.73}$$

Fama and MacBeth (1973) found, among other things, that the squared beta is not significant in the second step regression, nor is a measure of non-systematic risk.

The approach can also be extended to include other variables in the cross-sectional regressions, so (13.68) would become

$$R_{it}^e = \gamma_t + \lambda_t' \begin{bmatrix} \hat{\beta}_i \\ z_{it} \end{bmatrix} + \varepsilon_{it}, \tag{13.74}$$

where z_{it} could be a vector of asset (i) specific characteristics in period t (for instance, the leverage). Testing the λ coefficients of z_{it} is done in the same way as before. It can be noticed that when z_{it} is time-varying, then the FMB approach is not the same as OLS on pooled data. In fact, FMB is focused on the average cross-sectional affect, not on the time-series effect. (Actually, regressions where all fixed effects have been taken out by demeaning are the same in FMB and pooled OLS.)

13.10 Appendix: Details of CAPM Regression

Proof. (of (13.2)) Consider the regression equation $y_t = x_t' b_0 + u_t$. With iid errors that are independent of all regressors (also across observations), the LS estimator, $\hat{b}_{LS}$, is asymptotically distributed as

$$\sqrt{T}(\hat{b}_{LS} - b_0) \xrightarrow{d} N(\mathbf{0}, \sigma^2 \Sigma_{xx}^{-1}), \text{ where } \sigma^2 = \mathrm{E}\, u_t^2 \text{ and } \Sigma_{xx} = \mathrm{E}\, \Sigma_{t=1}^T x_t x_t'/T.$$

When the regressors are just a constant (equal to one) and one variable regressor, f_t, so $x_t = [1, f_t]'$, then we have

$$\Sigma_{xx} = \mathrm{E} \sum_{t=1}^T x_t x_t'/T = \mathrm{E} \frac{1}{T} \sum_{t=1}^T \begin{bmatrix} 1 & f_t \\ f_t & f_t^2 \end{bmatrix} = \begin{bmatrix} 1 & \mathrm{E}\, f_t \\ \mathrm{E}\, f_t & \mathrm{E}\, f_t^2 \end{bmatrix}, \text{ so}$$

$$\sigma^2 \Sigma_{xx}^{-1} = \frac{\sigma^2}{\mathrm{E}\, f_t^2 - (\mathrm{E}\, f_t)^2} \begin{bmatrix} \mathrm{E}\, f_t^2 & -\mathrm{E}\, f_t \\ -\mathrm{E}\, f_t & 1 \end{bmatrix} = \frac{\sigma^2}{\mathrm{Var}(f_t)} \begin{bmatrix} \mathrm{Var}(f_t) + (\mathrm{E}\, f_t)^2 & -\mathrm{E}\, f_t \\ -\mathrm{E}\, f_t & 1 \end{bmatrix}.$$

(In the last line we use $\mathrm{Var}(f_t) = \mathrm{E}\, f_t^2 - (\mathrm{E}\, f_t)^2$.) The upper left cell is (13.2). ■

Proof. (of (13.4)) From the CAPM regression (13.1) we have

$$\mathrm{Cov} \begin{bmatrix} R_{it}^e \\ R_{mt}^e \end{bmatrix} = \begin{bmatrix} \beta_i^2 \sigma_m^2 + \mathrm{Var}(\varepsilon_{it}) & \beta_i \sigma_m^2 \\ \beta_i \sigma_m^2 & \sigma_m^2 \end{bmatrix}, \text{ and } \begin{bmatrix} \mu_i^e \\ \mu_m^e \end{bmatrix} = \begin{bmatrix} \alpha_i + \beta_i \mu_m^e \\ \mu_m^e \end{bmatrix}.$$

Suppose we use this information to construct a mean-variance frontier for both R_{it} and

R_{mt}, and we find the tangency portfolio, with excess return R^e_{ct}. It is straightforward to show that the square of the Sharpe ratio of the tangency portfolio is $\mu^{e\prime}\Sigma^{-1}\mu^e$, where μ^e is the vector of expected excess returns and Σ is the covariance matrix. By using the covariance matrix and mean vector above, we get that the squared Sharpe ratio for the tangency portfolio, $\mu^{e\prime}\Sigma^{-1}\mu^e$, (using both R_{it} and R_{mt}) is

$$\left(\frac{\mu^e_c}{\sigma_c}\right)^2 = \frac{\alpha_i^2}{\text{Var}(\varepsilon_{it})} + \left(\frac{\mu^e_m}{\sigma_m}\right)^2,$$

which we can write as

$$(SR_c)^2 = \frac{\alpha_i^2}{\text{Var}(\varepsilon_{it})} + (SR_m)^2.$$

Use the notation $f_t = R_{mt} - R_{ft}$ and combine this with (13.2) and to get (13.4). ■

13.11 Appendix: Details of SURE Systems

Proof. (of (13.6)) Write each of the regression equations in (13.5) on a traditional form

$$R^e_{it} = x_t'\theta_i + \varepsilon_{it}, \text{ where } x_t = \begin{bmatrix} 1 \\ f_t \end{bmatrix}.$$

Define

$$\Sigma_{xx} = \text{plim} \sum_{t=1}^{T} x_t x_t'/T, \text{ and } \sigma_{ij} = \text{plim} \sum_{t=1}^{T} \varepsilon_{it}\varepsilon_{jt}/T.$$

With iid errors that are independent of all regressors (also across observations), the asymptotic covariance matrix of the vectors $\hat{\theta}_i$ and $\hat{\theta}_j$ (assets i and j) is $\sigma_{ij}\Sigma_{xx}^{-1}/T$ (see below for a separate proof). In matrix form,

$$\text{Cov}(\sqrt{T}\hat{\theta}) = \begin{bmatrix} \sigma_{11} & \dots & \sigma_{1n} \\ \vdots & & \vdots \\ \sigma_{n1} & \dots & \hat{\sigma}_{nn} \end{bmatrix} \otimes \Sigma_{xx}^{-1},$$

where $\hat{\theta}$ stacks $\hat{\theta}_1, \dots, \hat{\theta}_n$. As in (13.2), the upper left element of Σ_{xx}^{-1} equals $1 + SR^2$, where SR is the Sharpe ratio of the market. For a link to the GMM based formulas in Remark 13.3, notice that the above expression for $\text{Cov}(\sqrt{T}\hat{\theta})$ can also be written

$$D_0^{-1}(\Sigma \otimes \Sigma_{xx})(D_0^{-1})', \text{ with } D_0^{-1} = (I_n \otimes \Sigma_{xx}^{-1})$$

and where Σ is the matrix with σ_{ij} as the elements. ■

Proof. (of distribution of SURE coefficients, used in proof of (13.6)*) To simplify, consider the SUR system

$$y_t = x_t'\beta + u_t$$
$$z_t = x_t'\gamma + v_t.$$

Let $\hat{\Sigma}_{xx} = \sum_{t=1}^{T} x_t x_t'/T$. We then know (from basic properties of LS) that

$$\sqrt{T}(\hat{\beta} - \beta) = \hat{\Sigma}_{xx}^{-1}\sqrt{T}\frac{1}{T}\sum_{t=1}^{T} x_t u_t$$
$$\sqrt{T}(\hat{\gamma} - \gamma) = \hat{\Sigma}_{xx}^{-1}\sqrt{T}\frac{1}{T}\sum_{t=1}^{T} x_t v_t.$$

Notice that $\Sigma_{xx} = \operatorname{plim}\hat{\Sigma}_{xx}$, while the remaining terms (typically) obey CLTs. The covariance of $\sqrt{T}\hat{\beta}$ and $\sqrt{T}\hat{\gamma}$ is therefore (since Σ_{xx}^{-1} is symmetric)

$$\operatorname{Cov}(\sqrt{T}\hat{\beta}, \sqrt{T}\hat{\gamma}) = \Sigma_{xx}^{-1}\operatorname{Cov}\left(\sqrt{T}\frac{1}{T}\sum_{t=1}^{T} x_t u_t, \sqrt{T}\frac{1}{T}\sum_{t=1}^{T} x_t v_t\right)\Sigma_{xx}^{-1}.$$

With iid errors (although u_t and v_t may be correlated) that are independent of all regressors (also across observations), this simplifies to

$$\Sigma_{xx}^{-1}\operatorname{plim}(\frac{1}{T}\sum_{t=1}^{T} x_t x_t')\Sigma_{xx}^{-1}\sigma_{uv} = \Sigma_{xx}^{-1}\sigma_{uv}.$$

■

Remark 13.38 *(General results on SURE distribution, same regressors) Let the regression equations be*

$$y_{it} = x_t'\theta_i + \varepsilon_{it},\ i = 1, \ldots, n,$$

where x_t is a $K \times 1$ vector (the same in all n regressions). When the moment conditions are arranged so that the first n are $x_{1t}\varepsilon_t$, then next are $x_{2t}\varepsilon_t$

$$\mathrm{E}\, g_t = \mathrm{E}(x_t \otimes \varepsilon_t),$$

then Jacobian (with respect to the coefficients of x_{1t}, then the coefficients of x_{2t}, etc) and its inverse are

$$D_0 = -\Sigma_{xx} \otimes I_n \text{ and } D_0^{-1} = -\Sigma_{xx}^{-1} \otimes I_n.$$

The covariance matrix of the moment conditions is as usual $S_0 = \sum_{s=-\infty}^{\infty} \mathrm{E}\, g_t g_{t-s}'$. As

an example, let $n = 2$, $K = 2$ *with* $x_t' = (1, f_t)$ *and let* $\theta_i = (\alpha_i, \beta_i)$, *then we have*

$$\begin{bmatrix} \bar{g}_1 \\ \bar{g}_2 \\ \bar{g}_3 \\ \bar{g}_4 \end{bmatrix} = \frac{1}{T}\sum_{t=1}^{T} \begin{bmatrix} y_{1t} - \alpha_1 - \beta_1 f_t \\ y_{2t} - \alpha_2 - \beta_2 f_t \\ f_t(y_{1t} - \alpha_1 - \beta_1 f_t) \\ f_t(y_{2t} - \alpha_2 - \beta_2 f_t) \end{bmatrix},$$

and

$$\begin{aligned} \frac{\partial \bar{g}}{\partial[\alpha_1, \alpha_2, \beta_1, \beta_2]'} &= \begin{bmatrix} \partial\bar{g}_1/\partial\alpha_1 & \partial\bar{g}_1/\partial\alpha_2 & \partial\bar{g}_1/\partial\beta_1 & \partial\bar{g}_1/\partial\beta_2 \\ \partial\bar{g}_2/\partial\alpha_1 & \partial\bar{g}_2/\partial\alpha_2 & \partial\bar{g}_2/\partial\beta_1 & \partial\bar{g}_2/\partial\beta_2 \\ \partial\bar{g}_3/\partial\alpha_1 & \partial\bar{g}_3/\partial\alpha_2 & \partial\bar{g}_3/\partial\beta_1 & \partial\bar{g}_3/\partial\beta_2 \\ \partial\bar{g}_4/\partial\alpha_1 & \partial\bar{g}_4/\partial\alpha_2 & \partial\bar{g}_4/\partial\beta_1 & \partial\bar{g}_4/\partial\beta_2 \end{bmatrix} \\ &= -\frac{1}{T}\sum_{t=1}^{T} \begin{bmatrix} 1 & 0 & f_t & 0 \\ 0 & 1 & 0 & f_t \\ f_t & 0 & f_t^2 & 0 \\ 0 & f_t & 0 & f_t^2 \end{bmatrix} = \left(-\frac{1}{T}\sum_{t=1}^{T} x_t x_t'\right) \otimes I_2. \end{aligned}$$

Remark 13.39 *(General results on SURE distribution, same regressors, alternative ordering of moment conditions and parameters*) If instead, the moment conditions are arranged so that the first* K *are* $x_t\varepsilon_{1t}$, *the next are* $x_t\varepsilon_{2t}$ *as in*

$$\mathrm{E}\, g_t = \mathrm{E}(\varepsilon_t \otimes x_t),$$

then the Jacobian (wrt the coefficients in regression 1, then the coefficients in regression 2 etc.) and its inverse are

$$D_0 = I_n \otimes (-\Sigma_{xx}) \text{ and } D_0^{-1} = I_n \otimes (-\Sigma_{xx}^{-1}).$$

Reordering the moment conditions and parameters in Example 13.38 gives

$$\begin{bmatrix} \bar{g}_1 \\ \bar{g}_2 \\ \bar{g}_3 \\ \bar{g}_4 \end{bmatrix} = \frac{1}{T}\sum_{t=1}^{T} \begin{bmatrix} y_{1t} - \alpha_1 - \beta_1 f_t \\ f_t(y_{1t} - \alpha_1 - \beta_1 f_t) \\ y_{2t} - \alpha_2 - \beta_2 f_t \\ f_t(y_{2t} - \alpha_2 - \beta_2 f_t) \end{bmatrix},$$

and

$$\frac{\partial \bar{g}}{\partial[\alpha_1, \beta_1, \alpha_2, \beta_2]'} = \begin{bmatrix} \partial\bar{g}_1/\partial\alpha_1 & \partial\bar{g}_1/\partial\beta_1 & \partial\bar{g}_1/\partial\alpha_2 & \partial\bar{g}_1/\partial\beta_2 \\ \partial\bar{g}_2/\partial\alpha_1 & \partial\bar{g}_2/\partial\beta_1 & \partial\bar{g}_2/\partial\alpha_2 & \partial\bar{g}_2/\partial\beta_2 \\ \partial\bar{g}_3/\partial\alpha_1 & \partial\bar{g}_3/\partial\beta_1 & \partial\bar{g}_3/\partial\alpha_2 & \partial\bar{g}_3/\partial\beta_2 \\ \partial\bar{g}_4/\partial\alpha_1 & \partial\bar{g}_4/\partial\beta_1 & \partial\bar{g}_4/\partial\alpha_2 & \partial\bar{g}_4/\partial\beta_2 \end{bmatrix}$$

$$= -\frac{1}{T}\sum\nolimits_{t=1}^{T} \begin{bmatrix} 1 & f_t & 0 & 0 \\ f_t & f_t^2 & 0 & 0 \\ 0 & 0 & 1 & f_t \\ 0 & 0 & f_t & f_t^2 \end{bmatrix} = I_2 \otimes \left(-\frac{1}{T}\sum\nolimits_{t=1}^{T} x_t x_t' \right).$$

Chapter 14

Financial Panel Data

References: Verbeek (2012) 10, Baltagi (2008), Hoechle (2007), Driscoll and Kraay (1998), Wooldridge (2010) and Petersen (2009).

14.1 Introduction to Panel Data

A panel data set (also called a longitudinal data set) has data on a cross-section ($i = 1, 2, \ldots, N$, individuals or firms) for many time periods ($t = 1, 2, \ldots, T$). Our aim is to estimate a linear relation between the dependent variable and the regressors

$$y_{it} = \alpha_i + x_{it}'\beta_i + u_{it}, \tag{14.1}$$

where the coefficients (α_i,β_i) may or may not be different for different individuals (this is discussed in detail below). As examples of such applications, we may want to evaluate if alphas or betas of different mutual funds are related to fund characteristics, for instance, costs or trading activity. Alternatively, we want to investigate whether firms with different types of board compositions perform differently. Sometimes it will be convenient to put the constant in the x_{it} vector to write the model as $y_{it} = x_{it}'\beta_i + u_{it}$. (This should be clear from the context.)

Data on the dependent variable has this structure:

$$\begin{array}{rccccc} & \underline{i=1} & \underline{i=2} & \cdots & \underline{i=N} \\ t=1: & y_{11} & y_{21} & & y_{N1} \\ t=2: & y_{12} & y_{22} & & y_{N2} \\ \vdots & & & & \\ t=T: & y_{1T} & y_{2T} & & y_{NT} \end{array} \tag{14.2}$$

The structure for each of the regressors is similar, although it can also be the case that (some of) the regressors are the same for all N investors (for instance, when the regressors are pricing factors like the market excess return). When needed for clarity we will use the $y_{i,t}$ notation instead of y_{it}.

The structure in (14.2) implicitly assumes that we have a *balanced panel*, that is, have data for all the cells. However, it is often the case that the panel is *unbalanced* in the sense that some data is missing For instance, we may not have data on regressor 3 for $i = 7$ and $t = 3$. If data is *missing in a random way*, then we can simply exclude (y_{it}, x_{it}) for the missing (i, t). In our example that means just excluding $(y_{7,3}, x_{7,3})$ but keeping all other data. In contrast, if data is missing in a non-random way (for instance, depending on the value of y_{it}), then we have to apply more sophisticated sample-selection models (not discussed in this chapter).

14.2 An Overview of Different Panel Data Models

A *pooled model* assumes that all individuals have the same coefficients (no subscript on β), so (14.1) becomes

$$y_{it} = \alpha + x_{it}'\beta + u_{it}. \tag{14.3}$$

This model can be estimated by pooled OLS (see below).

A *fixed effects model* assumes that all individuals have the same slope coefficients, but that their intercepts might differ

$$y_{it} = \alpha_i + x_{it}'\beta + u_{it}. \tag{14.4}$$

An extension of the fixed effects model is to also allow for *time fixed effects*

$$y_{it} = \lambda_t + \alpha_i + x_{it}'\beta + u_{it}. \tag{14.5}$$

Estimation of these models is discussed below.

A *random effects model* is similar to a fixed effects model, except that the individual "mean" α_i now contains a common component (α) and a random individual component (μ_i). We can then write the model as

$$y_{it} = \alpha + x_{it}'\beta + u_{it} \text{ where } u_{it} = \mu_i + \varepsilon_{it}. \tag{14.6}$$

The ε_{it} is typically assumed to be uncorrelated across time and individuals, but the μ_i

terms make the u_{it} residuals correlated over time (for the same individual). The estimation of this model is discussed later.

The *unrestricted model* (14.1) allows all individuals to have different coefficients (hence a subscript i on β_i). These regressions could be estimated by OLS for each individual separately. Alternatively, a GLS approach can be applied to enhance the efficiency by exploiting the correlation (of the residuals) across individuals. This approach is not discussed in these notes, since it is basically very similar to the SURE model used for testing CAPM and other linear factor models. (See the CAPM notes.)

14.3 Pooled OLS

Consider the regression model

$$y_{it} = x_{it}'\beta + u_{it}, \tag{14.7}$$

where x_{it} is an $k \times 1$ vector. For notational convenience, this section assumes that any constant is included in the x_{it} vector along with the other regressors. Notice that the coefficients are the same across individuals (and time), but that the regressors may vary along both the time series and cross-sectional dimensions. We assume that u_{jt} is uncorrelated with x_{it} (across all i and j).

Define the matrices

$$\Sigma_{xx} = \frac{1}{TN}\sum_{t=1}^{T}\sum_{i=1}^{N} x_{it}x_{it}' \text{ (a } k \times k \text{ matrix)} \tag{14.8}$$

$$\Sigma_{xy} = \frac{1}{TN}\sum_{t=1}^{T}\sum_{i=1}^{N} x_{it}y_{it} \text{ (a } k \times 1 \text{ vector).} \tag{14.9}$$

The LS estimator (stacking all TN observations) is then

$$\hat{\beta} = \Sigma_{xx}^{-1}\Sigma_{xy}. \tag{14.10}$$

In case u_{it} is uncorrelated across time and also across individuals, then the usual expressions for $\text{Std}(\hat{\beta})$ apply. However, it is often the case that there are *clusters* of individuals (all small firms, say) that have correlated residuals. This would require handling those correlations.

Recall that we can (conceptually) decompose the point estimate $\hat{\beta}$ by using (14.7) to substitute for y_{it} in Σ_{xy} (14.9) and then in (14.10). The result is

$$\hat{\beta} = \beta + \Sigma_{xx}^{-1}\left(\frac{1}{TN}\sum_{t=1}^{T}\sum_{i=1}^{N} h_{it}\right), \text{ where } h_{it} = x_{it}u_{it}. \tag{14.11}$$

The variance-covariance matrix ($k \times k$) can then be written

$$\text{Var}(\sqrt{TN}\hat{\beta}) = \Sigma_{xx}^{-1} \text{Var}\left(\frac{1}{\sqrt{TN}} \sum_{t=1}^{T} \sum_{i=1}^{N} h_{it}\right) \Sigma_{xx}^{-1}. \tag{14.12}$$

(Clearly, if $\text{Var}(\sqrt{TN}\hat{\beta}) = A$, then $\text{Var}(\hat{\beta}) = A/(TN)$).

Notice that the middle matrix in (14.12) is the variance-covariance matrix of a sum of the $k \times 1$ vector $x_{it}u_{it}$ divided by $\sqrt{TN}$, that is, the average $x_{it}u_{it}$ multiplied by $\sqrt{TN}$. The sum in this expression looks like

$$\sum_{t=1}^{T} \sum_{i=1}^{N} h_{it} = \underbrace{h_{1,1}}_{i=1,t=1} + \underbrace{h_{2,1}}_{i=2,t=1} + \ldots \underbrace{h_{N-1,T}}_{i=N-1,t=T} + \underbrace{h_{N,T}}_{i=N,t=T}. \tag{14.13}$$

The variance of this sum depends on how the elements are correlated. Different *cluster methods* would account for a non-zero covariance across individuals within the same period (for instance, between h_{it} and h_{jt}), or across time for the same individual (for instance, between $h_{i,t}$ and $h_{i,t-1}$) and sometimes for both.

Remark 14.1 *(Unbalanced panels*) With missing values in (y_{it}, x_{it}) we want to exclude that observation. This can be done in several ways. For instance, by changing the summations in (14.8), (14.9) and (14.12) to skip over such data points. Alternatively, we can set $(y_{it}, z_{it}) = (0, \mathbf{0}_k)$ so all variables related to (t, i) are set to zero—but then keep the standard summation. In either case the TN terms are not entirely correct, but they cancel both in the calculation of $\hat{\beta}$ in (14.10) and of* $\text{Var}(\hat{\beta})$ *in (14.12).*

Remark 14.2 *(*Panel regression vs average coefficient) Consider the regression for investor i*

$$y_{it} = x_t'\beta_i + \varepsilon_{it},\ i = 1...N,$$

where the regressors are the same in all regressions—but where the coefficients might be different across investors. Clearly, we have for each i

$$\hat{\beta}_i = \tilde{S}_{xx}^{-1}\tilde{S}_{xy_i},$$
$$\text{where } \tilde{S}_{xx} = \sum_{t=1}^{T} x_t x_t'/T \text{ and } \tilde{S}_{xy_i} = \sum_{t=1}^{T} x_t y_{it}/T.$$

The cross-sectional average of the regression coefficients is therefore

$$\frac{1}{N}\sum_{i=1}^{N}\hat{\beta}_i = \tilde{S}_{xx}^{-1}\frac{1}{N}\sum_{i=1}^{N}\tilde{S}_{xy_i}.$$

Compare that to (14.8) and notice that since x_t is repeated N times, we have $\tilde{S}_{xx} = \Sigma_{xx}$. Similarly, comparing with (14.9) gives

$$\frac{1}{N}\sum_{i=1}^{N}\tilde{S}_{xy_i} = \Sigma_{xy}.$$

This shows that $\frac{1}{N}\sum_{i=1}^{N}\hat{\beta}_i = \hat{\beta}$, where the latter is from the panel regression (14.10).

14.4 The Within Estimator ("Fixed Effects Estimator")

In the fixed effects model, we allow for different individual intercepts

$$y_{it} = \alpha_i + x_{it}'\beta + u_{it}. \tag{14.14}$$

There are several ways to estimate this model. The conceptually most straightforward is to include individual dummies (N) where dummy i takes the value of one if the data refers to individual i and zero otherwise and estimate the model with pooled OLS. (Clearly, the regression can then not include any intercept. Alternatively, include an intercept but only $N-1$ dummies, for $i = 2-N$.) However, this approach can be difficult to implement since it may involve a very large number of regressors.

As an alternative (which gives the same point estimates as pooled OLS with dummies) consider the following approach. First, take average across time (for a given i) of y_{it} and x_{it} in (14.14). That is, think (but do not run any estimation yet...) of forming the cross-sectional regression

$$\bar{y}_i = \alpha_i + \bar{x}_i'\beta + \bar{u}_i, \text{ where} \tag{14.15}$$

$$\bar{y}_i = \frac{1}{T}\sum_{t=1}^{T}y_{it} \text{ and } \bar{x}_i = \frac{1}{T}\sum_{t=1}^{T}x_{it}. \tag{14.16}$$

Second, transform the data as

$$y_{it}^* = y_{it} - \bar{y}_i \tag{14.17}$$

$$x_{it}^* = x_{it} - \bar{x}_i. \tag{14.18}$$

These variables have zero means.

Use the transformed variables to express the difference between (14.14) and (14.15) as

$$y_{it}^* = x_{it}^{*\prime}\beta + u_{it}^*. \tag{14.19}$$

At this stage, estimate β by running pooled OLS on all observations of (14.19). There is no intercept in this regression, but adding one should not affect the slope coefficients. We denote this estimate $\hat{\beta}_{FE}$ (FE stands for fixed effects) and it is also often called the *within estimator*. The interpretation of this approach is that we estimate the slope coefficients by using only the movements around individual means (not how the individual means differ). Notice that it gives the same results as OLS with dummies. Third and finally, get estimates of individual intercepts as

$$\alpha_i = \bar{y}_i - \bar{x}_i' \hat{\beta}_{FE}. \tag{14.20}$$

Clearly, the within estimator wipes out all regressors that are constant across time for a given individual (say, gender and schooling): they are effectively merged with the individual means (μ_i). In practice, such variables must be excluded from the x_{it} vector since otherwise there will be some transformed variables, $x_{it} - \bar{x}_i$, that are always zero—causing numerical problems.

Remark 14.3 *(The within estimator and the Frisch-Waugh-Lovell theorem*) Regressing y_t and x_t on a set of dummies gives $\bar{y}_i$ and $\bar{x}_i$. The FWL theorem says that regressing y_{it}^* on x_{it}^* gives the same slope coefficients as regressing y_{it} on x_{it} and the dummies.*

We can apply the usual tests on the pooled OLS results from (14.19)—provided the residuals are uncorrelated across time and individuals. Otherwise, we need to apply a cluster method.

Remark 14.4 *(Lagged dependent variable as regressor*) If $y_{i,t-1}$ is among the regressors x_{it}, then the within estimator (14.19) is biased in short samples (that is, when T is small)—and increasing the cross-section (that is, N) does not help. To see the problem, suppose that the lagged dependent variable is the only regressor ($x_{it} = y_{i,t-1}$). The within estimator (14.19) is then*

$$y_{it} - \textstyle\sum_{t=1}^{T} y_{it}/T = [y_{i,t-1} - \sum_{t=2}^{T} y_{i,t-1}/(T-1)]\beta + [u_{it} - \sum_{t=1}^{T} u_{it}/T].$$

The problem is that the regressor $(y_{i,t-1} - ...)$ is correlated with Σu_{it} since the latter contains $u_{i,t-1}$ which affects $y_{i,t-1}$ directly. In addition, $\Sigma y_{i,t-1}$ contains $y_{i,t}$ which is correlated with u_{it}. It can be shown that this bias can be substantial for panels with small T.

Remark 14.5 *(Fixed effects in unbalanced panels*) When (y_{it}, x_{it}) inlcude one or more missing values, then we typically exclude that observation from the estimation. For that reason, they should also be excluded from calculating $\bar{y}_i$ and $\bar{x}_i$. In this way, (y_{it}^*, x_{it}^*) will have zero means in the sample that is used in the estimation.*

	LS	Fixed eff	Between	GLS	1st diff
exper/100	7.84	4.11	10.64	4.57	3.55
	(8.25)	(6.21)	(4.05)	(7.12)	(2.33)
exper^2/100	−0.20	−0.04	−0.32	−0.06	−0.05
	(−5.04)	(−1.50)	(−2.83)	(−2.37)	(−0.93)
tenure/100	1.21	1.39	1.25	1.38	1.29
	(2.47)	(4.25)	(0.90)	(4.32)	(2.98)
tenure^2/100	−0.02	−0.09	−0.02	−0.07	−0.08
	(−0.85)	(−4.36)	(−0.20)	(−3.77)	(−2.45)
south	−0.20	−0.02	−0.20	−0.13	−0.02
	(−13.51)	(−0.45)	(−6.67)	(−5.70)	(−0.56)
union	0.11	0.06	0.12	0.07	0.04
	(6.72)	(4.47)	(3.09)	(5.57)	(3.31)

Table 14.1: Panel estimation of log wages for women, $T = 5$ and $N = 716$ from NLS (1982,1983,1985,1987,1988). Example of fixed and random effects models, Hill et al (2008), Table 15.9. Numbers in parentheses are t-stats.

14.4.1 The Within Estimator with Time Fixed Effects

When we allow for both time fixed effects and individual fixed effects

$$y_{it} = \lambda_t + \alpha_i + x_{it}'\beta + u_{it}, \tag{14.21}$$

then we could once again introduce dummies (now for both time periods and individuals) and apply pooled OLS.

As before, it is often easier to transform the data before estimating with pooled OLS. In this case, we run the regression on transformed variables

$$y_{it}^* = x_{it}^{*\prime}\beta + u_{it}^*. \tag{14.22}$$

The transformations are

$$y_{it}^* = y_{it} - \bar{y}_i - \bar{y}_t + \bar{y} \tag{14.23}$$

$$x_{it}^* = x_{it} - \bar{x}_i - \bar{x}_t + \bar{x}, \tag{14.24}$$

where $\bar{x}_i$ is defined in (14.16) and

$$\bar{x}_t = \textstyle\sum_{i=1}^{N} x_{it}/N \text{ for each } t \text{ and}$$
$$\bar{x} = \textstyle\sum_{t=1}^{T}\sum_{i=1}^{N} x_{it}/(TN).$$

(Similarly for the transformation of y_{it}.) The last terms $(\bar{y}, \bar{x})$ makes sure that the grand mean of the transformed variable is zero. (If we instead add an intercept to (14.22), then this is not important for the slope coefficients.)

The estimation and testing of (14.22) is the same as for the standard within estimator (see above).

Remark 14.6 *(*Fixed effects in unbalanced panels) As with individual fixed effects, the averages should only be calculated from those data points that will be used in the estimation.*

Remark 14.7 *(Only time fixed effects*) When we allow for time fixed effects but no individual fixed effect, then the transformations are*

$$y_{it}^* = y_{it} - \bar{y}_t$$
$$x_{it}^* = x_{it} - \bar{x}_t.$$

Proof. (*of (14.22)) First, take averages over time (for each individual, $i = 1$ to N) of (14.21) to get

$$\bar{y}_i = \bar{\lambda} + \alpha_i + \bar{x}_i'\beta + \bar{u}_i.$$

Second, take averages over the cross section (in each time period, $t = 1$ to T)

$$\bar{y}_t = \lambda_t + \bar{\alpha} + \bar{x}_t'\beta + \bar{u}_t.$$

Third, take averages across both time and the cross-section

$$\bar{y} = \bar{\lambda} + \bar{\alpha} + \bar{x}'\beta.$$

Finally, subtract all three from (14.21) to get (14.22). ■

14.5 The First-Difference Estimator

An another way of estimating the fixed effects model is to difference away the μ_i by taking *first-differences* (in time)

$$\Delta y_{it} = \Delta\lambda_t + \Delta x_{it}'\beta + u_{it}^*, \tag{14.25}$$

where $\Delta y_{it} = y_{it} - y_{i,t-1}$ and similarly for the regressors. Quite often, we interpret $\Delta\lambda_t$ as just a constant. Notice that

$$u_{it}^* = u_{it} - u_{i,t-1}, \tag{14.26}$$

so there are reasons to suspect that u_{it}^* is (negatively) autocorrelated.

Notice that the first-difference approach focuses on how changes in the regressors (over time, for the same individual) affect changes in the dependent variable. Also this method wipes out all regressors that are constant across time (for a given individual).

Regression (14.26) can be estimated by pooled OLS. However, unadjusted standard errors are likely to overstate the uncertainty. This suggests that using the unadjusted standard errors is a conservative approach (harder to reject the null hypothesis). The reason is that if u_{it} is iid, then $\text{Cov}(u_{it}^*, u_{i,t-1}^*) = -\text{Var}(u_{it})$.

Remark 14.8 *(Lagged dependent variable as regressor*) If $y_{i,t-1}$ is among the regressors x_{it}, then the first-difference method (14.25) does not work (OLS is inconsistent and a larger sample does not help). The reason is that the (autocorrelated) residual is then correlated with the lagged dependent variable. This model cannot be estimated by OLS (the instrumental variable method might work).*

14.6 Differences-in-Differences Estimator

Consider the first-difference model (14.25) when one of the regressors is a dummy variable indicating whether individual i was "treated" (for instance, received investment advise) in period t. We can estimate this as before—and interpret the coefficient as the effect of the "treatment" (conditional on all other variables)

In the classical difference-in-difference estimator there are only two periods ($T = 2$): before and after the treatment. *If there are no other regressors*, then (14.25) can be written

$$\Delta y_{it} = \Delta\lambda_t + \delta Q_{it} + u_{it}^*, \tag{14.27}$$

where Q_{it} is the dummy variable. (The restriction that all individuals have the same $\Delta\lambda_t$ term is the so called "parallel trend assumption.") In this case δ can be estimated by the difference between the average Δy_{it} among the treated ($\Delta\bar{y}_{B2}$) and the average Δy_{it} among the non-treated ($\Delta\bar{y}_{A2}$)

$$\hat{\delta} = \Delta\bar{y}_{B2} - \Delta\bar{y}_{A2}. \tag{14.28}$$

(Notice that the change of the average is the same as the average of the change.)

More generally, consider the regression specification

$$y_{it} = \alpha_i + \kappa Q_i + \lambda_t + \delta T_t Q_i + x_{it}'\beta + \varepsilon_{it}, \tag{14.29}$$

where Q_i is a cross-sectional dummy variable (0 if i is non-treated and 1 otherwhise) and T_t is a time-series dummy variable (0 before the treatment, 1 after) and x_{it} contains other regressors. In this specification, δ is the key coefficient. In this specfication, α_i, Q_i and the cross-sectional vatriation in x_{it} capture the differences across individuals that are not related to the treatment. In contrast, λ_t and the time-series variation in x_{it} capture changes over time that are also unrelated to the treatment. In contrast, $T_t Q_i$ captures the treatment effect.

Suppose we only have two time periods (before and after the treatment), then the first difference of (14.29) gives

$$\Delta y_{it} = \underbrace{\Delta\alpha_i + \kappa\Delta Q_i}_{0} + \Delta\lambda_t + \delta\underbrace{\Delta(T_t Q_i)}_{Q_{it}} + \Delta x_{it}'\beta + u_{it}^*, \tag{14.30}$$

where Q_{it} is 0 for non-treated and 1 for treated (like in (14.27)). Notice that the "parallel trend assumption" now amounts to assuming that $\Delta\lambda_t + \Delta x_{it}'\beta$ is the same across the treated and non-treated. If this is questionable, then (14.27) should not be used. Rather, we should estimate (14.30).

14.7 Random Effects Model*

The random effects model allows for *random* individual "intercepts" (μ_i)

$$y_{it} = \beta_0 + x_{it}'\beta + \mu_i + \varepsilon_{it}, \text{ where} \tag{14.31}$$

$$\varepsilon_{it} \text{ is iid } N(0, \sigma_\varepsilon^2) \text{ and } \mu_i \text{ is iid } N(0, \sigma_\mu^2). \tag{14.32}$$

Notice that μ_i is random (across agents) but constant across time, while ε_{it} is just random noise. Hence, μ_i can be interpreted as the permanent "luck" of individual i.

It is sometimes argued that the random effect only makes sense if the data is a sample from a larger population—and then captures the peculiar (relative to the population) features of the individuals that end up in the sample. It is then convenient to merge μ_i with ε_{it}, because it gives fewer parameters to estimate (and thus, saves degrees of freedom). In contrast, if the cross-section effectively contains the population (all mutual funds on a market, say), then a fixed effect is perhaps more reasonable.

Clearly, if the we regard μ_i as non-random, then we are back in the fixed-effects model. (The choice between the two models is not always easy, so it may be wise to try both—and compare the results.)

We could write the regression as

$$y_{it} = \beta_0 + x_{it}'\beta + u_{it}, \text{ where } u_{it} = \mu_i + \varepsilon_{it}. \tag{14.33}$$

and we typically assume that ε_{jt} and μ_i are not correlated with each other or with x_{it}. Notice that u_{it} is autocorrelated even if ε_{it} is not: $\text{Cov}(u_{it}, u_{i,t-1}) = \text{Var}(\mu_i)$.

There are several ways to estimate the random effects model. First, the methods for fixed effects (the within and first-difference estimators) all work—so the "fixed effect" can actually be a random effect. Second, the *between estimator* using only individual time averages (from (14.16))

$$\bar{y}_i = \beta_0 + \bar{x}_i'\beta + \underbrace{\mu_i + \bar{\varepsilon}_i}_{\text{residual}_i}, \tag{14.34}$$

is also consistent, but discards all time-series information. Third, LS on

$$y_{it} = \beta_0 + x_{it}'\beta + \underbrace{\mu_i + \varepsilon_{it}}_{\text{residual}_{it}} \tag{14.35}$$

is consistent (but not really efficient). However, in this case we may need to adjust $\text{Cov}(\hat{\beta})$ since the covariance matrix of the residuals is not diagonal.

In the random effects model, the μ_i variable can be thought of as an *excluded variable*. Excluded variables typically give a bias in the coefficients of all included variables—unless the excluded variable is uncorrelated with all of them. This is the assumption in the random effects model (recall: we assumed that μ_i is uncorrelated with x_{jt}). If this assumption is wrong, then we cannot estimate the RE model by either OLS or GLS, but

the within-estimator (compare with the FE model) works, since it effectively eliminates the excluded variable from the system.

Remark 14.9 *(Generalized least squares*) GLS is an alternative estimation method that exploits correlation structure of residuals to increase the efficiency. In this case, it can be implemented by running OLS on*

$$y_{it} - \vartheta \bar{y}_i = \beta_0(1-\vartheta) + (x_{it} - \vartheta \bar{x}_i)'\beta + \upsilon_{it}, \text{ where}$$
$$\vartheta = 1 - \sqrt{\sigma_u^2/(\sigma_u^2 + T\sigma_\mu^2)}.$$

In this equation, σ_u^2 is the variance of the residuals in the "within regression" as estimated in (14.19) and $\sigma_\mu^2 = \sigma_B^2 - \sigma_u^2/T$, where σ_B^2 is the variance of the residuals in the "between regression" (14.34). Here, σ_μ^2 can be interpreted as the variance of the random effect μ_i. However, watch out for negative values of σ_μ^2 and notice that when $\vartheta \approx 1$, then GLS is similar to the "within estimator" from (14.19). This happens when $\sigma_\mu^2 >> \sigma_u^2$ or when T is large. The intuition is that when σ_μ^2 is large, then it is smart to get rid of that source of noise by using the within estimator, which disregards the information in the differences between individual means.

14.8 Fama-MacBeth

The Fama and MacBeth (1973) approach (called FMB below) is a different method for handling panel data. The method has two main steps, described below.

First, estimate λ_t and β_t

$$y_{it} = \lambda_t + x_{it}'\beta_t + u_{it} \tag{14.36}$$

period by period (using the cross section $i = 1 - N$). The FMB has the nice properties of easily handling unbalanced data sets (the cross-sectional regressions (14.36) are run on the available cross section for each time period).

Second, estimate the time averages

$$\hat{\beta} = \frac{1}{T}\sum_{t=1}^{T}\hat{\beta}_t. \tag{14.37}$$

Remark 14.10 *(Step 0*) The FMB can also be used to test CAPM (or other linear factor models). In this case, y_{it} in (14.36) are the excess returns on asset i in period t (R_{it}^e)*

and x_{it} are the loadings (γ_{it}) of the excess return on the market excess return (or other factors) according to the regression $R^e_{it} = \alpha + f_t'\gamma_{it} + \varepsilon_{it}$. In many cases, the γ_{it} values used as x_{it} are estimated during a previous sample, for instance, during the five years up to and including $t-1$. In other cases, the γ_{it} values are estimated from the full sample, and are thus constant across periods. The latter has the advantage of being more precise estimates, provided the assumption of constant loadings is correct.

Fama and MacBeth (1973) suggest that the standard deviation should be found by studying the time-variation in $\hat{\beta}_t$. In particular, they suggest that the variance of $\hat{\beta}_t$ (notice, not $\hat{\beta}$) can be estimated by the (average) squared variation around its mean

$$\text{Var}(\hat{\beta}_t) = \frac{1}{T}\sum_{t=1}^{T}(\hat{\beta}_t - \hat{\beta})^2. \tag{14.38}$$

Since $\hat{\beta}$ is the sample average of $\hat{\beta}_t$, the variance of the former is the variance of the latter divided by T (the sample size)—provided $\hat{\beta}_t$ is iid. That is,

$$\text{Var}(\hat{\beta}) = \frac{1}{T}\text{Var}(\hat{\beta}_t) = \frac{1}{T^2}\sum_{t=1}^{T}(\hat{\beta}_t - \hat{\beta})^2. \tag{14.39}$$

When x_{it} are common risk factors ($x_{it} = x_t$), then FMB and pooled OLS give the same point estimates (provided (14.36) is estimated without an intercept, effectively setting $\lambda_t = 0$). However, FMB's $\text{Var}(\hat{\beta})$ automatically handles the cross sectional correlations between residuals, while the pooled OLS would require applying a cluster method.

Empirical Example 14.11 *(Estimated factor risk premia from different methods) Table 14.2 shows estimates of the factor risk premia from several methods based on the 25 FF portfolios.*

14.9 Calendar Time and Cross Sectional Regression

14.9.1 Calendar Time Approach

The *calendar time* (CalTime) approach is to first define M discrete investor groups (for instance, age 18–30, 31–40, etc) and calculate their respective average excess returns ($\bar{y}_{jt}$ for group j)

$$\bar{y}_{jt} = \frac{1}{N_j}\sum_{i\in\text{Group}\,j} y_{it}, \tag{14.40}$$

	Data	CR	FMB1	FMB2
Market	7.39	6.88	6.88	−7.36
	(2.20)	(2.29)	(2.23)	(3.94)
SMB	1.58	1.47	1.47	1.08
	(1.48)	(1.55)	(1.52)	(1.52)
HML	3.38	4.13	4.13	3.73
	(1.44)	(1.60)	(1.48)	(1.48)

Table 14.2: Different estimates of factor risk premia, annualized %. Numbers in (parentheses) are standard deviations. The 25 FF portfolios 1970:01-2021:12. Data are the mean excess returns of the factors; CR are estimates of the factor risk premia from a cross-sectional regression; FMB1 are from Fama-MacBeth without intercept in the cross-sectional regression; FMB2 are from Fama-MacBeth with intercept in the cross-sectional regression.In both FMB regressions, the betas are estimated from the full sample.

where N_j is the number of individuals in group j. Notice that $\bar{y}_{jt}$ is just one time series of the equally-weighted portfolio return for investors in group j.

Then, we run a factor model

$$\bar{y}_{jt} = x_t'\beta_j + v_{jt}, \text{ for } j = 1, 2, \ldots, M \tag{14.41}$$

where x_t typically includes a constant and various return factors, for instance, excess returns on equity and bonds. Notice that x_t is the same for all groups. By estimating these M equations as a SURE system with White's (or Newey-West's) covariance estimator, it is straightforward to test various hypotheses, for instance, that the intercept (the "alpha") is higher for the Mth group than for the for first group.

Example 14.12 *(CalTime with two investor groups) With two investor groups, estimate the following SURE system*

$$\bar{y}_{1t} = x_t'\beta_1 + v_{1t},$$
$$\bar{y}_{2t} = x_t'\beta_2 + v_{2t}.$$

The CalTime approach is straightforward and the cross-sectional correlations are fairly easy to handle (in the SURE approach). However, it forces us to define discrete investor groups—which makes it hard to handle several different types of investor characteristics (for instance, age, trading activity and income) at the same time.

Empirical Example 14.13 *(Investor activivity vs performance, calendar time regressions) See Table 14.3 for results on a ten-year panel of some 60,000 Swedish pension savers from*

	Inactive	Active	Higly Active
coef	−0.76	3.08	8.65
t-tstat NW	−0.69	1.77	2.73

Table 14.3: Calendar time regressions. Annualised coefficients and t-stats from Table 10, in Dahlquist et al (RFS 2017). For Inactive the coefficient is the annualised alpha, but for the other two cateogories it is the difference in alpha to the Inactive. Three EW portfolios based on 62640 individuals, 2116 days. The dependent variables are the returns of the EW portfolio based on the activity indicators. The regressions also control for 7 risk factors over 5 days (2 lags, 2 leads).

Dahlquist, Martinez, and Söderlind (2016). In this case, the dependent variable is the return of a pension investment portfolio (on day t, individual i). Each individual is sorted into one EW portfolio based of her/his trading activity over the last year (inactive active, very active). The regressors include a constant, 7 risk factors (global and Swedish market, SMB, HML as a well as a bond factor) on ± 2 days ($1 + 7 \times 5$ regressors).

14.9.2 Cross Sectional Regression

The *cross sectional regression* (CrossReg) approach is to first estimate the factor model for each investor on time series data

$$y_{it} = x_t' \beta_i + \varepsilon_{it}, \text{ for } i = 1, 2, \ldots, N \tag{14.42}$$

and to run cross-sectional regressions of the (estimated) betas (for instance, for the pth factor) on the investor characteristics

$$\hat{\beta}_{p,i} = z_i' \gamma + u_i. \tag{14.43}$$

In this second-stage regression, the investor characteristics z_i could be a dummy variable (for age group, say) or a continuous variable (age, say). Notice that using a continuous investor characteristics assumes that the relation between the characteristics and the beta is linear—something that is not assumed in the CalTime approach. (This saves degrees of freedom, but may sometimes be a very strong assumption.) However, the cross-sectional approach is best suited for the case when the investor characteristics (z_i) are constant across time. This is a clear limitation.

A potential problem with the CrossReg approach is the cross-sectional correlation of the residuals (u_i). For instance, we may have a very large cross-sectional (N is large), but

it so happens that many of the investors follow very similar investment strategies. Notice also that this approach can only handle the (time) average characteristics.

Empirical Example 14.14 *(Investor activivity vs performance, cross-sectional regressions) For an empirical illustration, see Table 14.4 where the t-stats look massively inflated. Also, the investor characteristics used in these regressions are kept constant (across time) by simply using the time-series averages.*

	coef	t-tstat W	t-tstat C1	t-tstat C2	t-tstat C3
Inactive	−1.14	−17.27	−38.04	−13.39	−6.06
Active	5.95	42.13	6.70	18.77	11.62
Higly Active	10.25	25.80	9.91	11.64	25.71
Age	0.00	2.70	2.09	2.54	1.79
Male	0.43	20.66	7.83	23.61	27.88
Pension rights	−0.06	−8.55	−15.26	−6.76	−1.77

Table 14.4: Annualised regression coefficients and t-stats from a cross-sectional regression on the same data as in Dahlquist et al (RFS 2017), except that the investor characteristics are fixed across time (using the time averages). The dependent variable is the alpha of the 62640 individual portfolios. For Inactive the coefficient is the annualised alpha, but for the other two cateogories it is the difference in alpha to the Inactive. The alphas are from time series regressions which control for 7 risk factors over 5 days (2 lags, 2 leads). The t-stats are from White (W), clustering on activity (C1) age (C2), and pension rights (C3).

14.9.3 Cross Sectional Regression with Clustering of Residuals

In a cross-sectional regression $T = 1$. Use this in (14.8)–(14.9) to write (14.12) as

$$\text{Var}(\sqrt{N}\hat{\beta}) = \Sigma_{xx}^{-1} \text{Var}\left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N} h_i\right)\Sigma_{xx}^{-1}. \tag{14.44}$$

When we assume that the residuals (or here, $h_i = x_i u_i$) are iid, then $\text{Var}(\Sigma_{i=1}^{N} h_i)$ is just the sum of the variances of each term, $\Sigma_{i=1}^{N}\text{Var}(h_i)$. With clustering, this is different. The following example illustrates this.

Example 14.15 *(Cluster method on $N = 4$) Assume that individuals 1 and 2 form cluster 1 and that individuals 3 and 4 form cluster 2—and disregard correlations across clusters.*

This means setting the covariances across clusters to zero,

$$\begin{aligned}\text{Var}(\Sigma_{i=1}^{N} h_i) = \text{E}(&h_1^2 + h_2^2 + h_3^2 + h_4^2, \\ &2h_1h_2 + \underbrace{2h_1h_3}_{0} + \underbrace{2h_1h_4}_{0} + \underbrace{2h_2h_3}_{0} + \underbrace{2h_2h_4}_{0} + 2h_3h_4).\end{aligned}$$

(Recall that $\text{E}\, h_i = 0$*, so* $\text{E}\, h_i h_j = \text{Cov}(h_i, h_j)$*.) Notice that this can be written*

$$\text{Var}(\Sigma_{i=1}^{N} h_i) = \text{E}(h_1 + h_2)^2 + \text{E}(h_3 + h_4)^2.$$

Suppose there are C different clusters—and that we know which cluster individual i belongs to. Then, the previous example suggests that we can estimate the middle term of (14.44) as

$$S_C = \frac{1}{N}\sum_{c=1}^{C} h^c (h^c)', \text{ where } h^c = \sum_{i \in \text{ cluster } c} h_i. \tag{14.45}$$

The iid case is when each i is her/his own cluster. In contrast, we cannot allow everyone to be in the same cluster, since this would give $h^c = 0$. This will change when we have $T > 1$. It is often argued that replacing h_i by $h_i C/(C-1)$ improves the small properties of S_C.

14.10 Panel Regressions, Driscoll-Kraay and Cluster Methods

Equation (14.12) says that

$$\text{Var}(\sqrt{TN}\hat{\beta}) = \Sigma_{xx}^{-1} S_0 \Sigma_{xx}^{-1}, \text{ where } S_0 = \text{Cov}(\sqrt{TN}\bar{h}). \tag{14.46}$$

Clearly, the value of $\text{Cov}(\sqrt{TN}\bar{h})$ depends on how the elements in the average $\bar{h}$ are correlated (across time and across individuals).

14.10.1 The Effect of Cross-Sectional Correlations

To simplify the exposition we first focus on the cross-sectional correlations by *assuming that there are no autocorrelations*. In this case, we can simplify as

$$S_0 = \text{Cov}\left(\frac{1}{\sqrt{TN}}\sum_{t=1}^{T}\sum_{i=1}^{N} h_{it}\right) \tag{14.47}$$

$$= \frac{1}{TN}\sum_{t=1}^{T} \text{Cov}(h_t), \text{ where } h_t = \sum_{i=1}^{N} h_{it}. \tag{14.48}$$

In this expression, h_t is the $k \times 1$ vector of cross-sectional ($i = 1, 2, ..., N$) sums in period t. Since we use the covariance matrix of the moment conditions, heteroskedasticity is accounted for (as in White's method).

In general, $\text{Cov}(h_t)$ involves all the cross-sectional covariances. For instance, with $N = 2$ we have

$$\text{Cov}\left(h_{1t} + h_{2t}\right) = \text{Cov}\left(h_{1t}, h_{1t}\right) + \text{Cov}\left(h_{2t}, h_{2t}\right) + \text{Cov}\left(h_{1t}, h_{2t}\right) + \text{Cov}\left(h_{2t}, h_{1t}\right), \tag{14.49}$$

where each term is a $k \times k$ matrix. An iid assumption would assume that covariances across individuals (firms) are zero ($\text{Cov}\left(h_{1t}, h_{2t}\right) = 0$). In contrast, a cluster method may assume that such covariances are zero unless the two individuals belong to the same cluster (town, football club,...). The Driscoll-Kraay method makes no such assumptions.

14.10.2 From Driscoll-Kraay to Standard OLS (no autocorrelations)

We initially rule out autocorrelations. The methods summarised below all aim at estimating S_0 in (14.48) in a consistent way.

The Driscoll and Kraay (1998) (DK) estimates S_0 by

$$S_{DK} = \frac{1}{TN} \sum_{t=1}^{T} h_t h_t', \tag{14.50}$$

where h_t is the $k \times 1$ vector of the cross-sectional sum of h_{it} in period t, as defined in (14.48).

Remark 14.16 *(Relation to the notation in Hoechle (2007)) Hoechle writes* $\text{Cov}(\hat{\beta}) = (X'X)^{-1} \hat{S}_T (X'X)^{-1}$, *where* $\hat{S}_T = \sum_{t=1}^{T} h_t h_t'$. *Clearly,* $X'X/(TN) = \Sigma_{xx}$ *and* $\hat{S}_T / TN = S$. *Combining gives (14.46).*

Example 14.17 *(DK on $N = 4$) As an example, suppose there is one regressor ($k = 1$) and $N = 4$. Then, (14.48) gives the cross-sectional sum in period t*

$$h_t = h_{1t} + h_{2t} + h_{3t} + h_{4t},$$

and the covariance matrix (14.50)

$$\begin{aligned} TN \times S_{DK} &= \sum_{t=1}^{T} h_t h_t' \\ &= \sum_{t=1}^{T} (h_{1t} + h_{2t} + h_{3t} + h_{4t})^2 \\ &= \sum_{t=1}^{T} (h_{1t}^2 + h_{2t}^2 + h_{3t}^2 + h_{4t}^2 \\ &+ 2h_{1t}h_{2t} + 2h_{1t}h_{3t} + 2h_{1t}h_{4t} + 2h_{2t}h_{3t} + 2h_{2t}h_{4t} + 2h_{3t}h_{4t}) \end{aligned}$$

The term in parentheses is the sum of all the elements in this matrix of cross products $(h_{it}h_{jt})$*:*

i	1	2	3	4
1	h_{1t}^2	$h_{1t}h_{2t}$	$h_{1t}h_{3t}$	$h_{1t}h_{4t}$
2	$h_{2t}h_{1t}$	h_{2t}^2	$h_{2t}h_{3t}$	$h_{2t}h_{4t}$
3	$h_{3t}h_{1t}$	$h_{3t}h_{2t}$	h_{3t}^2	$h_{3t}h_{4t}$
4	$h_{4t}h_{1t}$	$h_{4t}h_{2t}$	$h_{4t}h_{3t}$	h_{4t}^2

This means that all cross-sectional covariances are allowed to be non-zero. In case h_{it} *is a* $k \times 1$ *vector, replace (the scalar)* $h_{it}h_{jt}$ *by the (*$k \times k$ *matrix)* $h_{it}h_{jt}'$*.*

A *cluster method* puts restrictions on the covariance terms (of h_{it}) that are allowed to enter the estimate S. In practice, all terms across clusters are left out. This can be implemented by changing the S matrix. In particular, instead of interacting all i with each other, we only allow for interaction within each of the C clusters ($c = 1, ..., C$)

$$S_C = \frac{1}{TN} \sum_{t=1}^{T} \sum_{c=1}^{C} h_t^c (h_t^c)', \text{ where } h_t^c = \sum_{i \in \text{cluster } c} h_{it}. \tag{14.51}$$

(Clearly, with only one cluster, then we are back in the DK method (14.50).)

Example 14.18 *(Cluster method on* $N = 4$*, changing Example 14.17 directly) Reconsider Example 14.17, but assume that individuals* 1 *and* 2 *form cluster* 1 *and that individuals* 3 *and* 4 *form cluster* 2*—and disregard correlations across clusters. This means*

setting the covariances across clusters to zero,

$$\begin{aligned} TN \times S_C &= \sum_{t=1}^{T}[(h_{1t}+h_{2t})^2+(h_{3t}+h_{4t})^2] \\ &= \sum_{t=1}^{T}(h_{1t}^2+h_{2t}^2+h_{3t}^2+h_{4t}^2, \\ &\quad 2h_{1t}h_{2t}+\underbrace{2h_{1t}h_{3t}}_{0}+\underbrace{2h_{1t}h_{4t}}_{0}+\underbrace{2h_{2t}h_{3t}}_{0}+\underbrace{2h_{2t}h_{4t}}_{0}+2h_{3t}h_{4t}). \end{aligned}$$

In this case, the term in parentheses sums all the elements in this matrix:

i	$\underline{1}$	$\underline{2}$	$\underline{3}$	$\underline{4}$
$\underline{1}$	h_{1t}^2	$h_{1t}h_{2t}$	0	0
$\underline{2}$	$h_{1t}h_{2t}$	h_{2t}^2	0	0
$\underline{3}$	0	0	h_{3t}^2	$h_{3t}h_{4t}$
$\underline{4}$	0	0	$h_{3t}h_{4t}$	h_{4t}^2

This disregards any cross-sectional correlations across clusters.

Instead, we get *White's covariance matrix* by excluding all cross-sectional cross terms. This can be accomplished by defining

$$S_W = \frac{1}{TN}\sum_{t=1}^{T}\sum_{i=1}^{N} h_{it}h_{it}'. \tag{14.52}$$

(This can be interpreted as a cluster method (14.51) where each i is its own cluster.) Notice that this disregards any cross-sectional correlations.

Example 14.19 *(White's method on $N = 4$) With only one regressor (14.52) gives*

$$TN \times S = \sum_{t=1}^{T}\left(h_{1t}^2+h_{2t}^2+h_{3t}^2+h_{4t}^2\right),$$

so the term in parentheses sums all the elements in this matrix:

i	$\underline{1}$	$\underline{2}$	$\underline{3}$	$\underline{4}$
$\underline{1}$	h_{1t}^2	0	0	0
$\underline{2}$	0	h_{2t}^2	0	0
$\underline{3}$	0	0	h_{3t}^2	0
$\underline{4}$	0	0	0	h_{4t}^2

Finally, the *traditional LS covariance matrix* assumes that White's estimate (14.52) can be simplified by exploiting the fact that $x_{it}x_{it}'$ and u_{it}^2 are not correlated. This changes S_W to $\Sigma_{xx}s^2$ where $s^2 = \Sigma_{t=1}^{T}\Sigma_{i=1}^{N}u_{it}^2/TN$. Using in (14.46) gives

$$\text{Cov}_{LS}(\sqrt{TN}\hat{\beta}) = \Sigma_{xx}^{-1}s^2. \tag{14.53}$$

14.10.3 Reintroducing Autocorrelations

The previous analysis disregarded autocorrelations. We now reintroduce this possibility.

For the DK estimator, first define the estimate of the pth autocovariance matrix by

$$S_{DK,p} = \frac{1}{TN}\sum_{t=1}^{T} h_t h_{t-p}'. \tag{14.54}$$

When $p = 0$, then this is clearly the same as S_{DK} in (14.50). If we allow for P lags, then the estimate of S_0 is

$$S_{DK} = S_{DK,0} + \textstyle\sum_{p=1}^{P} w_p(S_{DK,p} + S_{DK,p}'), \tag{14.55}$$

where $w_p = 1 - p/(P+1)$ in case we use the Bartlett weights (as in Newey-West), but also $w_p = 1$ can be motivated (see Petersen (2009) for a discussion).

The cluster method is to first define

$$S_{C,p} = \frac{1}{TN}\sum_{t=1}^{T}\sum_{c=1}^{C} h_t^c (h_{t-p}^c)' \tag{14.56}$$

and then use

$$S_C = S_{C,0} + \textstyle\sum_{p=1}^{P} w_p(S_{C,p} + S_{C,p}'). \tag{14.57}$$

Finally, if we rule out all correlations across individuals, then we set

$$S_{W,p} = \frac{1}{TN}\sum_{t=1}^{T}\sum_{i=1}^{N} h_{it} h_{i,t-p}' \tag{14.58}$$

and use

$$S_W = S_{W,0} + \textstyle\sum_{p=1}^{P} w_p(S_{W,p} + S_{W,p}'). \tag{14.59}$$

14.10.4 A Monte Carlo Experiment

Reference: Dahlquist, Martinez, and Söderlind (2016)

Basic Setup

This section reports results from a simple Monte Carlo experiment. We use the model

$$y_{it} = \alpha + \beta f_t + \gamma g_i + u_{it}, \tag{14.60}$$

where y_{it} is the return of individual i in period t, f_t a benchmark return and g_i is the (demeaned) number of the cluster $(-2, -1, 0, 1, 2)$ that the individual belongs to. This is a simplified version of the regressions we run in the paper. In particular, δ measures how the performance depends on the number of fund switches.

The experiment uses 3000 artificial samples with $t = 1, \ldots, 2000$ and $i = 1, \ldots, 1665$. Each individual is a member of one of five equally sized groups (333 individuals in each group). The benchmark return f_t is iid normally distributed with a zero mean and a standard deviation equal to $15/\sqrt{250}$, while u_{it} is also normally distributed with a zero mean and a standard deviation of one (different cross-sectional correlations are shown in the table). In generating the data, the true values of α and δ are zero, while β is one—and these are also the hypotheses tested below. To keep the simulations easy to interpret, there is no autocorrelation or heteroskedasticity.

Results for three different GMM-based methods are reported: Driscoll and Kraay (1998), a cluster method and White's method.

MC Covariance Structure

To generate data with correlated (in the cross-section) residuals, let the residual of individual i (belonging to group j) in period t be

$$u_{it} = \varepsilon_{it} + v_{jt} + w_t, \tag{14.61}$$

where $\varepsilon_{it} \sim N(0, \sigma_\varepsilon^2)$, $v_{jt} \sim N(0, \sigma_v^2)$ and $w_t \sim N(0, \sigma_w^2)$—and the three components are uncorrelated. This implies that

$$\begin{aligned} \text{Var}(u_{it}) &= \sigma_\varepsilon^2 + \sigma_v^2 + \sigma_w^2, \\ \text{Cov}(u_{it}, u_{kt}) &= \begin{bmatrix} \sigma_v^2 + \sigma_w^2 & \text{if individuals } i \text{ and } k \text{ belong to the same group} \\ \sigma_w^2 & \text{otherwise.} \end{bmatrix} \end{aligned} \tag{14.62}$$

Clearly, when $\sigma_w^2 = 0$ then the correlation across groups is zero, but there may be correlation within a group. If both $\sigma_v^2 = 0$ and $\sigma_w^2 = 0$, then there is no correlation at all

across individuals. For CalTime portfolios (one per activity group), we expect the individual shocks ε_{it} to average out, so a group portfolio has the variance $\sigma_v^2 + \sigma_w^2$ and the covariance of two different group portfolios is σ_w^2.

The Monte Carlo simulations consider different values of the variances—to illustrate the effect of the correlation structure.

Results from the Monte Carlo Simulations

Table 14.5 reports the fraction of times the absolute value of a t-statistic for a true null hypothesis is higher than 1.96. The table has three panels for different correlation patterns of the residuals (u_{it}): no correlation between individuals, correlations only within the pre-specified clusters and correlation across all individuals.

In the *upper panel*, where the residuals are iid, all three methods have rejection rates around 5% (the nominal size).

In the *middle panel*, the residuals are correlated within each of the five clusters, but there is no correlation between individuals that belong to the different clusters. In this case, but the DK and the cluster method have the right rejection rates, while White's method gives much too high rejection rates (around 85%). The reason is that White's method disregards correlation between individuals—and in this way underestimates the uncertainty about the point estimates. It is also worth noticing that the good performance of the cluster method depends on pre-specifying the correct clustering. Further simulations (not tabulated) show that with a completely random cluster specification (unknown to the econometricians), gives almost the same results as White's method.

The *lower panel* has no cluster correlations, but all individuals are now equally correlated (similar to a fixed time effect). For the intercept (α) and the slope coefficient on the common factor (β), the DK method still performs well, while the cluster and White's methods give too many rejects: the latter two methods underestimate the uncertainty since some correlations across individuals are disregarded. Things are more complicated for the slope coefficient of the cluster number (δ). Once again, DK performs well, but both the cluster and White's methods lead to too few rejections. The reason is the interaction of the common component in the residual with the cross-sectional dispersion of the group number (g_i).

Remark 14.20 *(Interpretation of the simulations results*) To understand this last result, consider a stylised case where* $y_{it} = \delta g_i + u_{it}$ *where* $\delta = 0$ *and* $u_{it} = w_t$ *so all residuals*

are due to an (excluded) time fixed effect. In this case, the matrix above becomes

$$\begin{bmatrix} i & \underline{1} & \underline{2} & \underline{3} & \underline{4} \\ \underline{1} & w_t^2 & \underline{w_t^2} & -w_t^2 & -w_t^2 \\ \underline{2} & \underline{w_t^2} & w_t^2 & -w_t^2 & -w_t^2 \\ \underline{3} & -w_t^2 & -w_t^2 & w_t^2 & \underline{w_t^2} \\ \underline{4} & -w_t^2 & -w_t^2 & \underline{w_t^2} & w_t^2 \end{bmatrix}$$

(This follows from $g_i = (-1,-1,1,1)$ *and since* $h_{it} = g_i \times w_t$ *we get* $(h_{1t}, h_{2t}, h_{3t}, h_{4t}) = (-w_t, -w_t, w_t, w_t)$*.) Both White's and the cluster method sum up only positive cells, so* S *is a strictly positive number. (For this the cluster method, this result relies on the assumption that the clusters used in estimating* S *correspond to the values of the regressor,* g_i*.) However, that is wrong since it is straightforward to demonstrate that the estimated coefficient in any sample must be zero. This is seen by noticing that* $\sum_{i=1}^{N} h_{it} = 0$ *at a zero slope coefficient holds for all* t*, so there is in fact no uncertainty about the slope coefficient. In contrast, the DK method adds the off-diagonal elements which are all equal to* $-w_t^2$*, giving the correct result* $S = 0$*.*

Empirical Example 14.21 *(Panel regressions, different types of t-stats) Based on Table 4, regression [2] in Karnaukh, Ranaldo, and Söderlind (2015), Table 14.6 shows point estimates and Table 14.7 four different sets of t-stats.*

14.11 From CalTime to a Panel Regression

The CalTime estimates can be replicated by using the individual data in the panel. For instance, with two investor groups we could estimate the following two regressions

$$y_{it} = x_t'\beta_1 + u_{it}^{(1)} \text{ for } i \in \text{group 1} \tag{14.63}$$

$$y_{it} = x_t'\beta_2 + u_{it}^{(2)} \text{ for } i \in \text{group 2.} \tag{14.64}$$

More interestingly, these regression equations can be combined into one *single* panel regression (and still give the same estimates) by the help of dummy variables. Let $z_{ji} = 1$ if individual i is a member of group j and zero otherwise. Stacking all the data, we have

(still with two investor groups)

$$\begin{aligned} y_{it} &= (z_{1i}x_t)'\beta_1 + (z_{2i}x_t)'\beta_2 + u_{it} \\ &= \left(\begin{bmatrix} z_{1i}x_t \\ z_{2i}x_t \end{bmatrix}\right)' \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + u_{it} \\ &= (z_i \otimes x_t)'\beta + u_{it}, \text{ where } z_i = \begin{bmatrix} z_{1i} \\ z_{2i} \end{bmatrix}. \end{aligned} \tag{14.65}$$

This is estimated with LS by stacking all NT observations.

To see why the CalTime approach implicitly handles correlations within the groups (clusters), notice that the CalTime approach (14.41) and the panel approach (14.65) give the same coefficients. This makes it clear that the errors in CalTime are just group averages of the errors in the panel regressions

$$v_{jt} = \frac{1}{N_j}\sum_{i\in\text{Group } j} u_{it}^{(j)}. \tag{14.66}$$

We know that

$$\text{Var}(v_{jt}) = \frac{1}{N_j}(\overline{\sigma}_{ii} - \overline{\sigma}_{ih}) + \overline{\sigma}_{ih}, \tag{14.67}$$

where $\overline{\sigma}_{ii}$ is the average $\text{Var}(u_{it}^{(j)})$ and $\overline{\sigma}_{ih}$ is the average $\text{Cov}(u_{it}^{(j)}, u_{ht}^{(j)})$. With a large cross-section, only the covariance matters. A good covariance estimator for the panel approach will therefore have to handle the covariance with a group (and perhaps also the covariance across groups). This suggests that the panel regression needs to handle the cross-correlations (for instance, by using the cluster or DK covariance estimators).

Proof. (*of (14.67)) Write (14.66) as $v = \mathbf{1}'u/N$, where $\mathbf{1}$ is an $N \times 1$ vector of ones. It follows that $\text{Var}(v) = \mathbf{1}'\Sigma\mathbf{1}/N^2$, where Σ is the covariance matrix of u. Clearly, $\mathbf{1}'\Sigma\mathbf{1}$ is just the sum of all elements of Σ. First, the sum of all elements along the diagonal divided by N is the average variance, $\overline{\sigma}_{ii}$. Second, the sum of all off-diagonal elements divided by $N(N-1)$ is the average covariance, $\overline{\sigma}_{ih}$. Therefore, $\mathbf{1}'\Sigma\mathbf{1}/N^2 = \overline{\sigma}_{ii}/N + \overline{\sigma}_{ih}N(N-1)/N^2$, which can be rearranged as (14.67). ■

We could also consider the case *when the characteristics are not dummies* (like young or old), but rather continuous variable (for instance, age measured in years). For this case,

write the model as

$$y_{it} = (z_{it} \otimes x_t)'d + v_{it} \tag{14.68}$$

$$= ([1, z_{1it}, \ldots, z_{mit}] \otimes [1, x_{1t}, \ldots, x_{kt}])'d + u_{it}, \tag{14.69}$$

where z_{jit} measures characteristics j of investor i in period t and where x_{pt} is the pth regressor. In many cases z_{jit} is time-invariant and could even be just a dummy: $z_{jit} = 1$ if investor i belongs to investor group j (for instance, being young). In other cases, z_{jit} is time invariant and contains static information about investor i.

This model is estimated with LS (stacking all NT observations), but the standard errors could be calculated according to Driscoll and Kraay (1998) (DK)—which accounts for cross-sectional correlations, for instance, correlations between the residuals of different investors (say, v_{1t} and v_{7t}).

Example 14.22 *(One investor characteristic and one pricing factor). In this case (14.68) is*

$$y_{it} = \begin{bmatrix} 1 \\ x_t \\ z_{it} \\ z_{it}x_{1t} \end{bmatrix}' d + u_{it},$$

$$= d_0 + d_1 x_t + d_2 z_{it} + d_3 z_{it} x_t + u_{it}.$$

In case we are interested in how the investor characteristics (z_{it}) affect the intercept (alpha), then d_2 is the key coefficient. To see that, rearrange as

$$y_{it} = \underbrace{d_0 + d_2 z_{it}}_{\text{intercept}} + \underbrace{(d_1 + d_3 z_{it})}_{\text{slope}} x_t + u_{it}.$$

Clearly, d_2 shows how the characteristics z_{it} affects the intercept and d_3 how it affects the slope.

14.11.1 An Empirical Illustration

Empirical Example 14.23 *(Panel regressions) See Table 14.8 for results on a ten-year panel of some 60,000 Swedish pension savers from Dahlquist, Martinez, and Söderlind (2016). In this case, the dependent variable is the return of a pension investment portfolio (on day t, individual i). The regressors include a constant, 7 risk factors (global and*

Swedish market, SMB, HML as a well as a bond factor) on ±2 days ($1+7\times5$ regressors), an indicator of trading activity of the individual over the last year (inactive active, very active).

The table illustrates the distinct difference in t-stats obtained by using different ways of handling the cross-sectional correlations (of the residuals). Notice, in particular, that the DK t-stats are the same as in calendar time approach (using Newey-West) in Table 14.3, although the estimation method is very different (here: a panel regression).

Empirical Example 14.24 *(Panel regressions, different t-stats) Table 14.9 replicates the point estimates from the cross-sectional regressions in Table 14.4, but using a panel estimation. Notice that this approach treats the investor characteristics as constant across time. For instance, the activity level is the average activity level across time. Notice that the point estimates are the same as in the cross-sectional regression, but the standard errors differ. In particular, the DK approach gives much lower t-stats.*

Table 14.10 extends the panel estimation in Table 14.8, but it includes more regressors (age, gender and pension rights). This would be difficult to handle in a calendar time approach, and thus illustrates that a panel regression can handle more general cases. Notice that the investor characteristics are here allowed to change across time. For instance, an investor can be active during the early years and then become inactive.

14.12 The Results in Hoechle, Schmid and Zimmermann

Hoechle, Schmid, and Zimmermann (2015) (HSZ) prove the following two propositions about (14.68)–(14.69).

Proposition 14.25 *If the z_{it} vector in (14.68) consists of dummy variables indicating exclusive and constant group membership ($z_{1it} = 1$ means that investor i belongs to group 1, so $z_{jit} = 0$ for $j = 2,...,m$), then the LS estimates and DK standard errors of (14.68) are the same as LS estimates and Newey-West standard errors of the CalTime approach (14.41). (See HSZ for a proof.)*

This proposition basically says that panel regression is as good as the CT approach. So why use a panel regression, then? A. Because it allows for *(a)* many characteristics (poor, old, men) without having to define a very large set of dummies (poor&old&men, poor&old&female, poor&young&men,...); *(b)* a finer (continuous) characteristics grid (age in years, months, days and...).

Proposition 14.26 *(When z_{it} is a measure of constant investor characteristics) The LS estimates and DK standard errors of (14.68) are the same as the LS estimates of CrossReg approach (14.43), but where the standard errors account for the cross-sectional correlations, while those in the CrossReg approach do not. (See HSZ for a proof.)*

	White	Cluster	Driscoll-Kraay
A. No cross-sectional correlation			
α	0.049	0.049	0.050
β	0.044	0.045	0.045
γ	0.050	0.051	0.050
B. Within-cluster correlations			
α	0.853	0.053	0.054
β	0.850	0.047	0.048
γ	0.859	0.049	0.050
C. Within- and between-cluster correlations			
α	0.935	0.377	0.052
β	0.934	0.364	0.046
γ	0.015	0.000	0.050

Table 14.5: **Simulated size of different covariance estimators** This table presents the fraction of rejections of true null hypotheses for three different estimators of the covariance matrix: White's (1980) method, a cluster method, and Driscoll and Kraay's (1998) method. The model of individual i in period t and who belongs to cluster j is $r_{it} = \alpha + \beta f_t + \gamma g_i + u_{it}$, where f_t is a common regressor (iid normally distributed) and g_i is the demeaned number of the cluster that the individual belongs to. The simulations use 3000 repetitions of samples with $t = 1, \ldots, 2000$ and $i = 1, \ldots, 1665$. Each individual belongs to one of five different clusters. The error term is constructed as $u_{it} = \varepsilon_{it} + v_{jt} + w_t$, where ε_{it} is an individual (iid) shock, v_{jt} is a shock common to all individuals who belong to cluster j, and w_t is a shock common to all individuals. All shocks are normally distributed. In Panel A the variances of $(\varepsilon_{it}, v_{jt}, w_t)$ are (1,0,0), so the shocks are iid; in Panel B the variances are (0.67,0.33,0), so there is a 33% correlation within a cluster but no correlation between different clusters; in Panel C the variances are (0.67,0,0.33), so there is no cluster-specific shock and all shocks are equally correlated, effectively having a 33% correlation within a cluster and between clusters.

	Poor	Rich
cap flow	−4.2	−4.4
VIX	−8.6	−6.8
TED	−4.6	−6.8
MSCIw	−4.9	−2.0
FXvol	−24.1	−37.7
Stockvol	−4.2	−1.9
StockLiq	−1.7	−8.6
BondLiq	9.7	7.3
lag	−6.5	−1.3

Table 14.6: Regression coefficients (in %) from Table 4, regression [2] in Karnaukh et al (RFS 2015), 1995:01–2009:12. Panel regressions of 30 FX liquidity time-series on (common) drivers.

	OLS		White's		Cluster		DK	
	Poor	Rich	Poor	Rich	Poor	Rich	Poor	Rich
cap flow	−2.14	−2.12	−2.12	−2.47	−1.50	−1.51	−1.26	−1.35
VIX	−2.63	−1.97	−2.26	−1.81	−1.79	−1.22	−1.64	−1.12
TED	−2.46	−3.43	−2.16	−3.42	−1.79	−2.37	−1.67	−2.19
MSCIw	−2.20	−0.86	−1.95	−0.74	−1.32	−0.46	−1.13	−0.46
FXvol	−10.46	−15.60	−6.61	−9.95	−4.86	−5.26	−4.43	−5.11
Stockvol	−1.61	−0.71	−1.25	−0.48	−0.83	−0.33	−1.03	−0.31
StockLiq	−0.75	−3.71	−0.60	−2.63	−0.43	−2.00	−0.40	−1.90
BondLiq	4.68	3.33	3.94	2.83	2.49	1.93	2.06	1.89
lag	−3.38	−0.62	−2.93	−0.53	−1.98	−0.34	−1.72	−0.30

Table 14.7: Different t-stats for Table 4, regression [2] in Karnaukh et al (RFS 2015), 1995:01–2009:12. Clustering is done according rich/poor (on average). All methods, except OLS, allow for first-order autocorrelation.

	coef	t-tstat W	tstat DK
Inactive	−0.76	−56.89	−0.69
Active	3.08	37.48	1.77
Higly Active	8.65	28.73	2.73

Table 14.8: Annualised regression coefficients and different t-stats from Table 10, regressions I and II in Dahlquist et al (RFS 2017). For Inactive the coefficient is the annualised alpha, but for the other two cateogories it is the difference in alpha to the Inactive. Panel regressions, 62640 individuals, 2116 days. The dependent variable is the return of the individual portfolio (day t, individual i). The regressions also control for 7 risk factors over 5 days (2 lags, 2 leads).

	coef	t-tstat W	tstat DK
Inactive	−1.14	−13.01	−0.88
Active	5.95	40.99	1.69
Higly Active	10.25	23.50	2.54
Age	0.00	2.05	0.49
Male	0.43	15.50	3.71
Pension rights	−0.06	−6.57	−1.46

Table 14.9: Annualised regression coefficients and different t-stats similar to Table 10, regressions I and II in Dahlquist et al (RFS 2017), but where the investor characteristics are fixed across time (using the time averages). Panel regressions, 62640 individuals, 2116 days. The dependent variable is the return of the individual portfolio (day t, individual i). The regressions also control for 7 risk factors over 5 days (2 lags, 2 leads).

	coef	t-tstat W	tstat DK
Inactive	−1.10	−1.63	−0.69
Active	3.10	34.61	1.79
Higly Active	8.69	28.44	2.74
Age	0.00	0.19	0.11
Male	0.62	2.94	2.22
Pension rights	−0.03	−0.39	−0.33

Table 14.10: Annualised regression coefficients and different t-stats from Table 10, regressions I and II in Dahlquist et al (RFS 2017). Panel regressions, 62640 individuals, 2116 days. The dependent variable is the return of the individual portfolio (day t, individual i). The regressions also control for 7 risk factors over 5 days (2 lags, 2 leads).

Chapter 15

Predicting Asset Returns: Nonparametric Estimation

15.1 Basics of Kernel Regressions

Reference: Campbell, Lo, and MacKinlay (1997) 12.3; Härdle (1990); Pagan and Ullah (1999); Mittelhammer, Judge, and Miller (2000) 21; Hansen (forthcoming (2021) 19

15.1.1 Introduction

Nonparametric regressions are used when we are unwilling to impose a parametric form on the regression equation—and we have a lot of data.

Let the scalars y_t and x_t be related as

$$y_t = b(x_t) + \varepsilon_t, \tag{15.1}$$

where ε_t is uncorrelated over time and where $\mathrm{E}\,\varepsilon_t = 0$ and $\mathrm{E}(\varepsilon_t|x_t) = 0$. The function $b()$ is unknown and possibly non-linear. In comparison, in a linear regression we have $b(x_t) = \beta x_t$.

One possibility of estimating such a function is to approximate $b(x)$ by a polynomial (or some other basis). This will give quick estimates, but the results are "global" in the sense that the value of $b(x)$ at a particular x value ($x = 1.9$, say) will depend on all the data points—and potentially very strongly so. The approach in this section is more "local" by down weighting information from data points where x_t is far from x.

As a starting point, suppose we want to estimate $b(x)$ at $x = 1.9$. If our sample has 3 observations (say, $t = 3$, 27, and 99) with $x_t = 1.9$, then it would be straightforward to average over these three observations to estimate $b(1.9)$ as $(y_3 + y_{27} + y_{99})/3$. This makes sense, since the average of the error terms $(\varepsilon_3, \varepsilon_{27}, \varepsilon_{99})$ is likely to be close to zero.

Unfortunately, we seldom have repeated observations of this type. Moreover, it seems

to be a waste to disregard data points where x_t is close, but not equal, to x. Instead, we may try to estimate the value of $b(x)$ by averaging over (y) observations where x_t is close to x (here 1.9). The general form of this type of estimator is

$$\hat{b}(x) = \frac{\sum_{t=1}^{T} w(x_t - x) y_t}{\sum_{t=1}^{T} w(x_t - x)}, \tag{15.2}$$

where $w(x_t - x)/\Sigma_{t=1}^{T} w(x_t - x)$ is the weight on data in t (in practice, y_t). This weight is non-negative and (weakly) decreasing in the the distance of x_t from x. Note that the denominator makes the weights sum to unity. The basic assumption behind (15.2) is that the $b(x)$ function is smooth so local averaging (around x) makes sense.

As an example of a $w(.)$ function, it could be to give equal weight all values of x_t that are in a certain bin ("mean-bin") and zero weight to all other observations. See Figure 15.1 for an example. Alternatively, we can give equal weights to the k values of x_t which are closest to x and zero to all other observations (this is the "k-nearest neighbor" estimator, see Härdle (1990) 3.2). As another example, the weight function could be defined so that it trades off the expected squared errors, $\mathrm{E}[y_t - \hat{b}(x)]^2$, and the expected squared acceleration, $\mathrm{E}[d^2\hat{b}(x)/dx^2]^2$. This defines a cubic spline (often used in macroeconomics when $x_t = t$, and is then called the Hodrick-Prescott filter).

15.1.2 Kernel Regression

A *kernel regression* uses a pdf as the weight function, $w(x_t - x) = K(u_t)$, where $u_t = (x_t - x)/h$. The choice of h (also called bandwidth) allows us to easily vary the relative weights of different observations, in particular the importance of nearby vs. distant observations.

The perhaps simplest choice is a flat function over an interval/bin (and zero outside). For this case, write the weighting function as

$$w(u_t) = \begin{cases} \frac{1}{2\sqrt{3}} \text{ for } |u| \leq \sqrt{3} \\ 0 \text{ otherwise,} \end{cases} \tag{15.3}$$

$$\text{with } u_t = (x_t - x)/h. \tag{15.4}$$

The reason for the $\sqrt{3}$ terms is that it makes area under the function equal to 1 ($\int w(u)du = 1$) and the variance also equal to one ($\int w(u)u^2du = 1$). This standardisation makes it easy to compare with a $N(0,1)$ distribution. In any case, we can adjust h to get the intervals we want. Since (15.3) implies a flat function over $\pm h\sqrt{3}$, we can set $h = \gamma/\sqrt{3}$ to

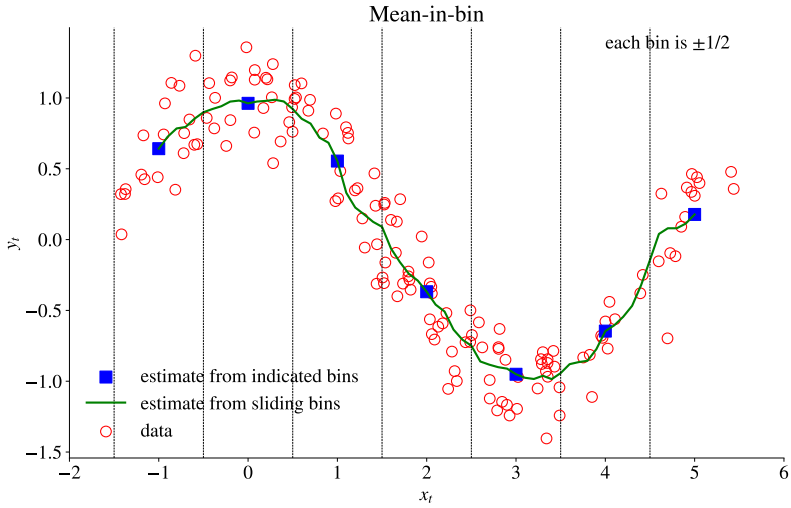


Figure 15.1: Example of a mean-in-bin estimation

get flat weights for all data points (x_t) that satisfy $\gamma \leq x_t - x \leq \gamma$. (Notice: some authors use the convention of a uniform distribution over $(x-h, x+h)$ or $(x-h/2, x+h/2)$ instead.) The mean-in-bin approach in Figure 15.1 is implemented by using (15.3).

Remark 15.1 *(Interpretation of the pdf in (15.3)) If $w(u_t)$ is a pdf of the u_t variable, then $w(u_t)/h$ is the pdf of x_t. Notice that both give the same result in (15.2).*

However, we can gain efficiency and get a smoother (across x values) estimate by using a density function that tapers off more smoothly. With an $N(0,1)$ kernel applied to $u_t = (x_t - x)/h$, we get the following weights

$$w(u_t) = \frac{1}{\sqrt{2\pi}} \exp\left(-u_t^2/2\right), \text{ with } u_t = (x_t - x)/h. \tag{15.5}$$

When $h \to 0$, then no averaging is done ($\hat{b}(x)$ evaluated at $x = x_t$ is just y_t). In contrast, as $h \to \infty$, $\hat{b}(x)$ becomes the sample average of y_t so we have global averaging. Clearly, some value of h in between is needed.

Remark 15.2 *(The Epanechnikov kernel) Let $u_t = (x_t - x)/h$. The Epanechnikov kernel is*

$$w(u_t) = \begin{cases} \frac{3}{4\sqrt{5}}\left(1 - \frac{u_t^2}{5}\right) \text{ for } |u| \leq \sqrt{5} \\ 0 \text{ otherwise.} \end{cases}$$

It can be noticed that $\int w(u)du = 1$ and $\int w(u)u^2du = 1$ (the variance is 1).

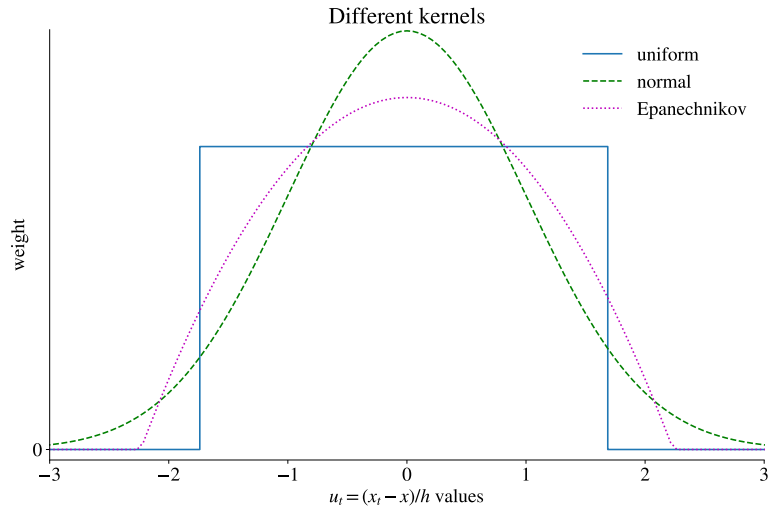


Figure 15.2: Different weighting functions for non-parametric regression

See Figure 15.2 for a comparison of weighting functions (also called kernels). The choice between them is typically less important than the choice of the bandwitdh h.

In practice we have to estimate $\hat{b}(x)$ at a finite number of points x. This could, for instance, be 100 evenly spread points in the interval between the minimum and the maximum values observed in the sample. See Figures 15.3–15.6 for illustrations of the method. Special corrections might be needed if there are a lot of observations stacked close to the boundary of the support of x (see Härdle (1990) 4.4).

Example 15.3 *(Kernel regression) Suppose the sample has three data points* $[x_1, x_2, x_3] = [1.5, 2, 2.5]$ *and* $[y_1, y_2, y_3] = [5, 4, 3.5]$. *Consider the estimation of* $b(x)$ *at* $x = 1.9$. *With* $h = 1$, *the numerator in (15.5) is*

$$\begin{aligned}\sum_{t=1}^{T} w(x_t - x)y_t &= \left(e^{-(1.5-1.9)^2/2} \times 5 + e^{-(2-1.9)^2/2} \times 4 + e^{-(2.5-1.9)^2/2} \times 3.5\right)/\sqrt{2\pi} \\ &\approx (0.92 \times 5 + 1.0 \times 4 + 0.84 \times 3.5)/\sqrt{2\pi} \\ &= 11.52/\sqrt{2\pi}.\end{aligned}$$

The denominator is

$$\begin{aligned}\sum_{t=1}^{T} w(x_t - x) &= \left(e^{-(1.5-1.9)^2/2} + e^{-(2-1.9)^2/2} + e^{-(2.5-1.9)^2/2}\right)/\sqrt{2\pi} \\ &\approx 2.75/\sqrt{2\pi}.\end{aligned}$$

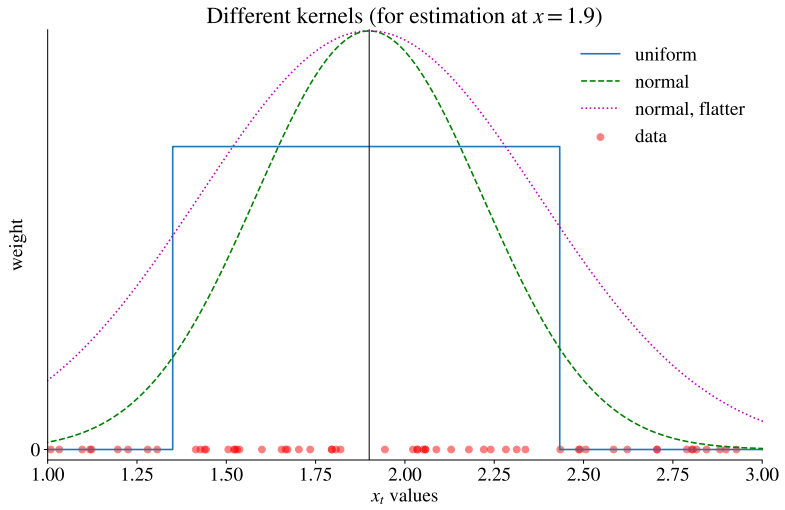


Figure 15.3: Different weighting functions for non-parametric regression

The estimate at $x = 1.9$ *is therefore*

$$\hat{b}(1.9) \approx 11.52/2.75 \approx 4.19.$$

15.1.3 Multivariate Kernel Regression

Suppose that y_t depends on two variables (x_t and z_t)

$$y_t = b(x_t, z_t) + \varepsilon_t, \tag{15.6}$$

where ε_t is uncorrelated over time and where $\mathrm{E}\,\varepsilon_t = 0$ and $\mathrm{E}(\varepsilon_t | x_t, z_t) = 0$.

This makes the estimation problem more data demanding. To see why, suppose we use a uniform density function as weighting function (see in (15.3)). However, with two regressors, the interval becomes a rectangle. With as little as a 20 intervals of each of x and z, we get 400 bins, so we need a large sample to have a reasonable number of observations in every bin.

In any case, the most common way to implement the kernel regressor is to let

$$\hat{b}(x, z) = \frac{\sum_{t=1}^{T} w(x_t - x) v(z_t - z) y_t}{\sum_{t=1}^{T} w(x_t - x) v(z_t - z)}, \tag{15.7}$$

where $w()$ and $v()$ are two kernels like in (15.5) and where we may allow the bandwidth (h) to be different for x_t and z_t (and depend on the variance of x_t and y_t). In this case.

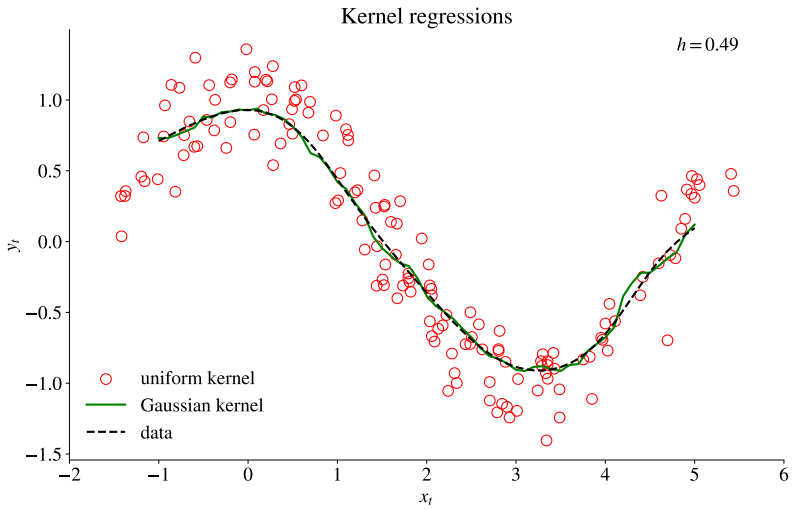


Figure 15.4: Example of kernel estimations

the weight of the observation (x_t, z_t) is proportional to $w(x_t - x)v(z_t - z)$, which is high if both x_t and z_t are close to x and z respectively.

Empirical Example 15.4 *(Kernel regression of an AR(2) for equity returns) See Figure 15.8.*

15.2 Distribution of the Kernel Regression and Choice of Bandwidth

Kernel regressions are typically consistent, provided longer samples are accompanied by smaller values of h, so the weighting function becomes more and more local as the sample size increases. It can be shown (see Härdle (1990) 3.1 and Pagan and Ullah (1999) 3.3–4) that under the assumption that x_t is iid, the mean squared error, variance and bias of the estimator at the value x are approximately (for general kernel functions)

$$\text{MSE}(x) = \text{Var}[\hat{b}(x)] + \text{Bias}[\hat{b}(x)]^2, \text{ with} \tag{15.8}$$

$$\text{Var}[\hat{b}(x)] = \frac{1}{Th}\frac{\sigma^2(x)}{f(x)} \times \int_{-\infty}^{\infty} K(u)^2 du \tag{15.9}$$

$$\text{Bias}[\hat{b}(x)] = h^2 \times \left[\frac{1}{2}\frac{d^2b(x)}{dx^2} + \frac{df(x)}{dx}\frac{1}{f(x)}\frac{db(x)}{dx}\right] \times \int_{-\infty}^{\infty} K(u)u^2 du. \tag{15.10}$$

In these expressions, $\sigma^2(x)$ is the variance of the residuals in (15.1) which may depend on the x value, $f(x)$ the marginal density of x and $K(u)$ the kernel (pdf) used as a weighting

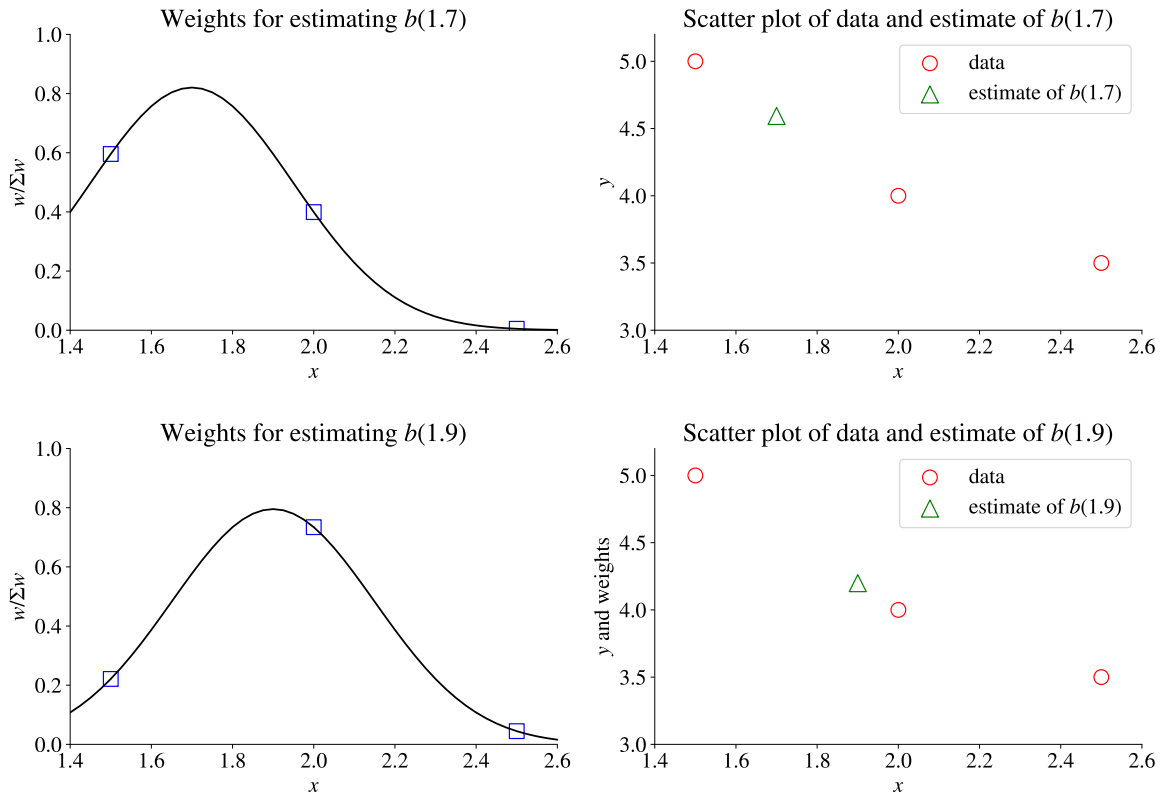


Figure 15.5: Example of kernel regression with three data points

function for $u_t = (x_t - x)/h$. The remaining terms are functions of the true regression function $b(x)$.

Remark 15.5 *(*Value of $\int_{-\infty}^{\infty} K(u)^2 du$). When $K(u)$ is a standard normal pdf, then the integral is $1/(2\sqrt{\pi})$, for the uniform distribution in (15.3) it is $1/(2\sqrt{3})$ and for an Epanechnikov kernel in Remark 15.2 it is $3\sqrt{5}/25$.*

As a comparison, a linear regression has $\hat{b}(x) = z'\hat{\gamma}$ where $z = [1, x]$, so the variance of the fitted value is

$$\text{Var}(z'\hat{\gamma}) = z'V(\hat{\gamma})z, \tag{15.11}$$

where $V(\hat{\gamma})$ is the variance-covariance matrix of $\hat{\gamma}$. (Notice that this is different from the variance of a forecast error, since the latter also includes the variance of the residual.)

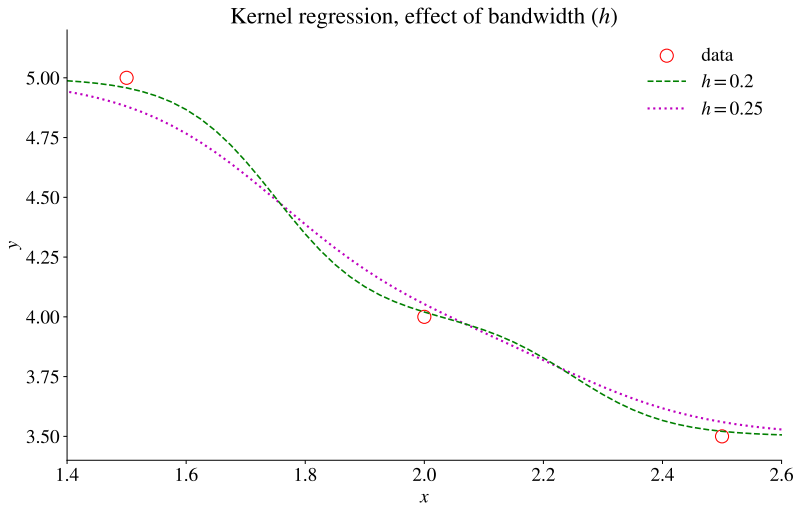


Figure 15.6: Example of kernel regression with three data points

With a Gaussian kernel these expressions can be simplified to

$$\text{Var}[\hat{b}(x)] = \frac{1}{Th}\frac{\sigma^2(x)}{f(x)} \times \frac{1}{2\sqrt{\pi}} \tag{15.12}$$

$$\text{Bias}[\hat{b}(x)] = h^2 \times \left[\frac{1}{2}\frac{d^2b(x)}{dx^2} + \frac{df(x)}{dx}\frac{1}{f(x)}\frac{db(x)}{dx}\right]. \tag{15.13}$$

Proof. (of (15.12)–(15.13)) We know that

$$\int_{-\infty}^{\infty} K(u)^2 du = \frac{1}{2\sqrt{\pi}} \text{ and } \int_{-\infty}^{\infty} K(u)u^2 du = 1,$$

if $K(u)$ is the density function of a standard normal distribution. (We are using the $N(0,1)$ pdf for the variable $u_t = (x_t - x)/h$.) Use in (15.9)–(15.10). ■

Remark 15.6 *(*$\text{Var}[\hat{b}(x)]$ *with other kernels). Use Remark 15.5 to replace the* $1/(2\sqrt{\pi})$ *term in (15.12).*

Equations (15.9) and (15.12) show that smaller h increases the variance (we effectively use fewer data points to estimate $b(x)$) but decreases the bias of the estimator (it becomes more local to x). If h decreases less than proportionally with the sample size (so hT in the denominator of the first term increases with T), then the variance goes to zero and the estimator is consistent (since the bias in the second term decreases as h does). It is clear that the choice of h has a major importance on the estimation results.

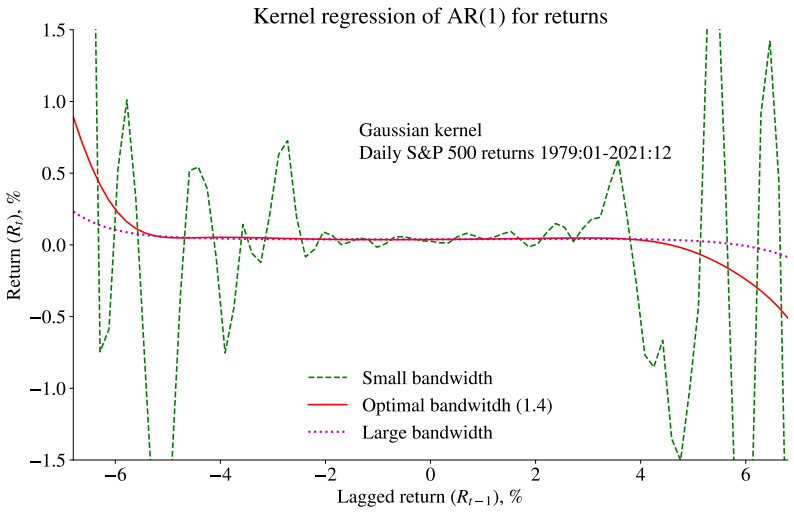


Figure 15.7: Non-parametric regression, importance of bandwidth

Empirical Example 15.7 *(Kernel regression of an AR(1) for equity returns) Figure 15.7 clearly illustrates the importance of the bandwidth.*

The variance is also a function of the variance of the residuals and the "peakedness" of the kernel, but not of the $b(x)$ function. The more concentrated the kernel is ($\int K(u)^2 du$ large) around x (for a given h), the less information is used in forming the average around x, and the uncertainty is therefore larger—which is similar to using a small h. A low density of the regressors ($f(x)$ low) means that we have little data at x which drives up the uncertainty of the estimator.

Equations (15.10) and (15.13) show that the bias increases (in magnitude) with the curvature of the $b(x)$ function (that is, $(d^2b(x)/dx^2)^2$). This makes sense, since rapid changes of the slope of $b(x)$ make it hard to get $b(x)$ right by averaging at nearby x values. It also increases with the variance of the kernel since a large kernel variance is similar to a large h.

Remark 15.8 *(Rule of thumb value of h) In a simplified case, we can find the h value that minimizes the MSE by an analytical approach. Use (15.13) to construct the MSE* $=$ $\text{Var}(b)+bias(b)^2$. *To simplify, assume the distribution of x is uniform, so* $f(x) = 1/(x_{\max}-x_{\min})$ *and* $df(x)/dx = 0$. *In addition, run the regression* $y = \alpha+\beta x+\gamma x^2+\varepsilon$ *as an approximation of* $b(x)$. *With this we have* $d^2b(x)/dx^2 \approx 2\gamma$ *and we approximate*

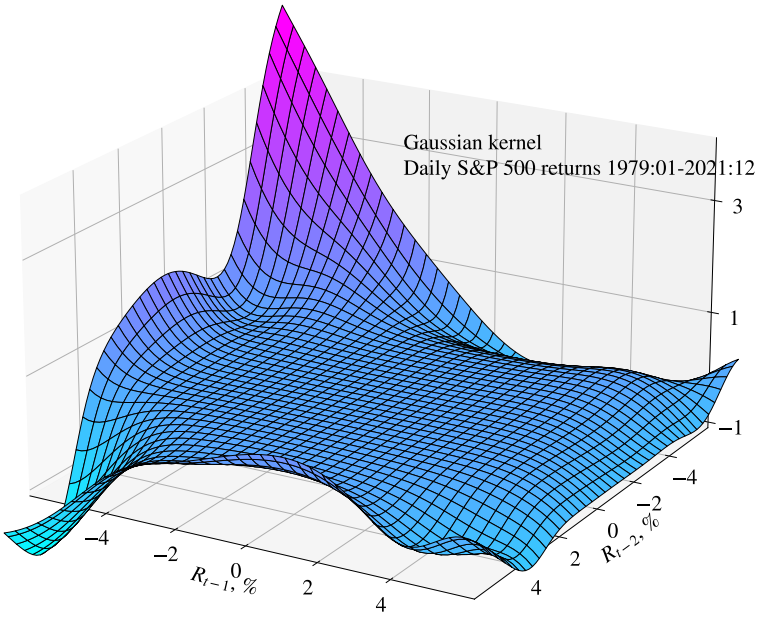


Figure 15.8: Non-parametric regression with two regressors

σ^2 *by the variance of the fitted residuals,* σ_ε^2*. Combining, we have*

$$MSE = \frac{1}{Th}\sigma_\varepsilon^2(x_{\max} - x_{\min})\frac{1}{2\sqrt{\pi}} + h^4\gamma^2.$$

Minimizing with respect to h *gives the first order condition*

$$0 = -\frac{1}{Th^2}\sigma_\varepsilon^2(x_{\max} - x_{\min})\frac{1}{2\sqrt{\pi}} + 4h^3\gamma^2, \text{ so}$$
$$h = 0.6\left(\frac{\sigma_\varepsilon^2(x_{\max} - x_{\min})}{T\gamma^2}\right)^{1/5}$$

In practice, replace $x_{\max} - x_{\min}$ *by the difference between the 90th and 10th percentiles of* x*.*

A good (but computationally intensive) approach to choose h is by the leave-one-out *cross-validation* technique. This approach would, for instance, choose h to minimize the

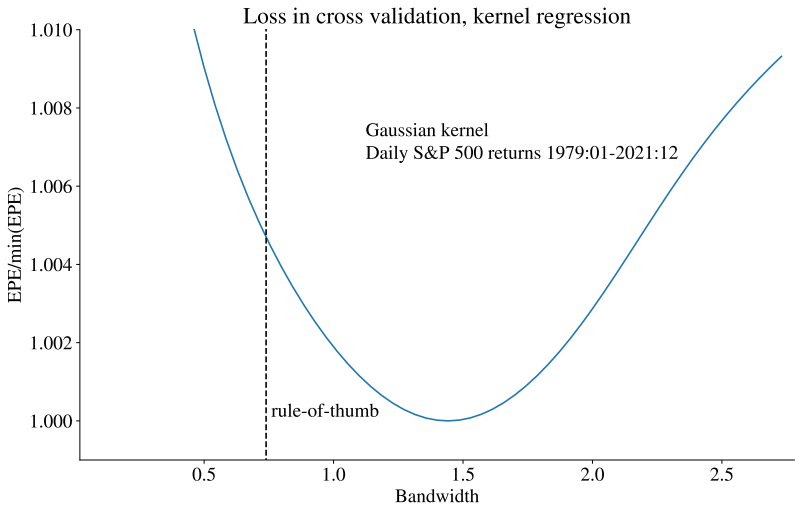


Figure 15.9: Cross-validation

expected (or average) prediction error

$$\text{EPE}(h) = \sum_{t=1}^{T} [y_t - \hat{b}_{-t}(x_t, h)]^2 / T, \tag{15.14}$$

where $\hat{b}_{-t}(x_t, h)$ is the fitted value of the regression function evaluated at $x = x_t$. Notice that the regression function $\hat{b}_{-t}(x_t, h)$ is estimated (using the bandwidth h) on a sample that excludes observation (y_t, x_t). This means that each prediction is out-of-sample. To calculate (15.14) we clearly need to make T estimations, that is, we have to estimate $\hat{b}_{-t}(x_t, h)$ for each t. Then we repeat this for different values of h to find the minimum.

Empirical Example 15.9 *(Kernel regression of an AR(1) for equity returns) Figure 15.9 shows the EPE for different values of the bandwidth for the kernel regressions previously illustrated in Figure 15.7.*

Remark 15.10 *(EPE calculations) Step 1: pick a value for h*
Step 2: estimate the $b(x)$ function on all data, but exclude $t = 1$, then calculate $\hat{b}_{-1}(x_1)$ and the error $y_1 - \hat{b}_{-1}(x_1)$
Step 3: redo Step 2, but now exclude $t = 2$ and. calculate the error $y_2 - \hat{b}_{-2}(x_2)$. Repeat this for $t = 3, 4, ..., T$. Calculate the EPE as in (15.14).
Step 4: redo Steps 2–3, but for another value of h. Keep doing this until you find the best h (the one that gives the lowest EPE)

Remark 15.11 *(Speed and fast Fourier transforms) A a fast Fourier transform can help speeding up the calculation of the kernel estimator.*

If the observations are independent, then it can be shown (see Härdle (1990) 4.2, Pagan and Ullah (1999) 3.3–6, and also (15.12)) that, with a Gaussian kernel, the estimator at point x is asymptotically normally distributed

$$\sqrt{Th}[\hat{b}(x) - b(x)] \to^d N\left(0, \frac{1}{2\sqrt{\pi}} \frac{\sigma^2(x)}{f(x)}\right), \tag{15.15}$$

where $f(x)$ is the density of x and $\sigma^2(x)$ is the variance of the residuals in (15.1) which could also depend on the x value. (A similar expression for the distribution holds for other choices of the kernel.) This expression assumes that the asymptotic bias is zero, which is guaranteed if h is decreased (as T increases) slightly faster than $T^{-1/5}$ (for instance, suppose $h = T^{-1.1/5}h_0$, where h_0 is a constant). To estimate the density of x, we can apply a standard method, for instance using a Gaussian kernel and the bandwidth (for the density estimate only) of $1.06\,\text{Std}(x_t)T^{-1/5}$.

Remark 15.12 *(Asymptotic bias) The condition that h decreases faster than $T^{-1/5}$ ensures that the bias of $\sqrt{Th}\hat{b}(x)$ vanishes as $T \to \infty$. This is seen by noticing that the bias of $\hat{b}(x)$ is proportional to h^2 (see (15.13)). Multiplying by $\sqrt{Th}$ gives the bias of $\sqrt{Th}\hat{b}(x)$ as being proportional to $T^{1/2}h^{5/2}$. With $h = T^{-1.1/5}h_0$, this bias is proportional to $T^{1/2}(T^{-1.1/5}h_0)^{5/2}$, that is, to $T^{-0.05}h_0^{5/2}$ which is decreases to zero as T increases.*

To estimate $\sigma^2(x)$ in (15.15), we may assume that it does not depend on x, so we just estimate the variance of the fitted residuals. (Clearly, this requires estimating $\hat{b}(x)$ at every point $x = x_t$ in the sample, not just a small grid of x values.) Alternatively, we use a non-parametric regression of the squared fitted residuals on x_t

$$\hat{\varepsilon}_t^2 = \sigma^2(x_t), \text{ where } \hat{\varepsilon}_t = y_t - \hat{b}(x_t), \tag{15.16}$$

where $\hat{b}(x_t)$ are the fitted values from the non-parametric regression (15.1). To draw confidence bands, it is typically assumed that the asymptotic bias is zero ($\text{E}\,\hat{b}(x) = b(x)$).

Empirical Example 15.13 *(Kernel regression of an AR(1) for equity returns) See Figure 15.10 for an example where the width of the confidence band varies across x values—mostly because the sample contains few observations of extreme x values as shown in*

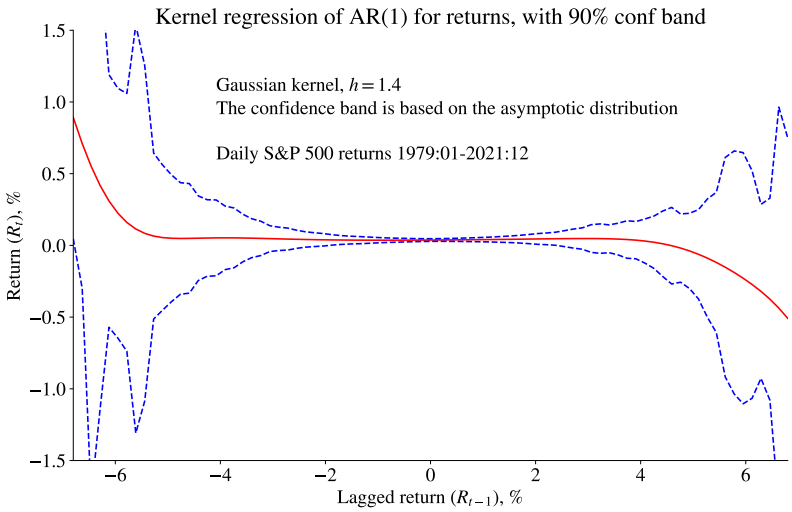


Figure 15.10: Kernel regression with confidence bands

Figure 15.11. In particular, compare with the confidence bands of a linear regression in Figure 15.12, which do account for the lack of data points with extreme x values.

Empirical Example 15.14 *(Bootstrapping the confidence bands for the kernel regression I) Figure 15.13 shows a bootstrapped confidence band. The bootstrap simulations account for the (non-linear) autocorrelation (the returns are generated recursively using the estimated regression function) and the residuals have regressor-dependent heteroskedasticity. The latter is achieved by first estimating (by a non-parametric approach) how the squared fitted residuals depend on the regressor (lagged return). Then standardized fitted residuals are calculated. In the simulations, the fitted residuals are drawn (with replacement) and then scaled up by the regressor dependent volatility.*

Empirical Example 15.15 *(Bootstrapping the confidence bands for the kernel regression II) Figures 15.13 and 15.14 show bootstrapped confidence bands for the kernel regression in Figure 15.10. Figure 15.13 plots the regression function (from Figure 15.10),* ± 1.64 *times the bootstrapped standard deviation (for each regressor value). In contrast, Figure 15.14 shows the 5th and 95th percentiles across the bootstrap simulations (also for each regressor value).*

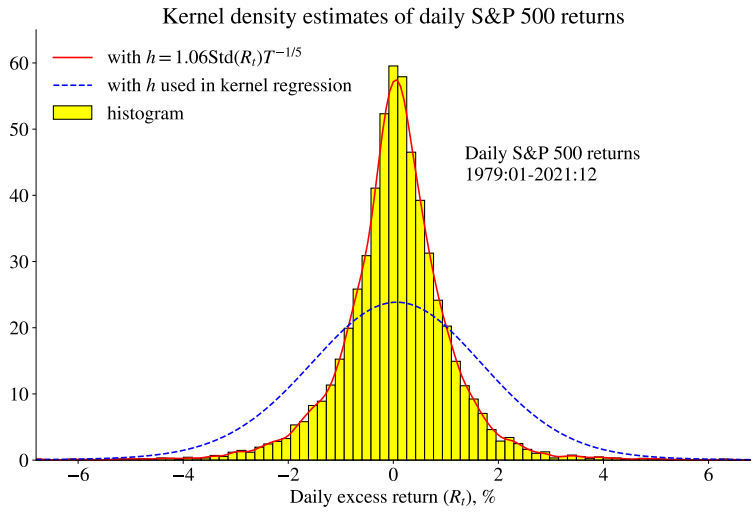


Figure 15.11: Distribution of daily stock returns

15.3 Local Linear Regressions

Notice that (15.2) solves the problem $\min_{a_x} \sum_{t=1}^{T} w(x_t - x)(y_t - \alpha_x)^2$ for each value of x. For a given value of x, α_x is a constant—but it can vary across x values. The first order condition (at a given x value) is $\sum_{t=1}^{T} w(x_t - x)(y_t - \alpha_x) = 0$, so the solution is as in (15.2), that is, $\hat{\alpha}_x = \hat{b}(x)$. This can be interpreted as a "local constant" regression model: for each x it is just a constant.

This can be extended to solving a problem like

$$\min_{\alpha_x, \gamma_x} \sum_{t=1}^{T} w(x_t - x)[y_t - \alpha_x - \beta_x(x_t - x)]^2, \tag{15.17}$$

which defines the local linear estimator. (Yes, the convention is to use $x_t - x$ as the regressor, but this could easily be changed.) The first order conditions are similar to the usual normal equations for LS (except that data point t has the weight $w(x_t - x)$ and that we use $x_t - x$ as the regressor). In fact, if we let $z_t = [1, x_t - x]'$ and collect the coefficients in $\theta_x = [\alpha_x, \beta_x]'$, then the first order conditions can be written

$$\sum_{t=1}^{T} w(x_t - x) z_t y_t = \sum_{t=1}^{T} w(x_t - x) z_t z_t' \hat{\theta}_x. \tag{15.18}$$

It is straightforward to solve these, but perhaps even easier if we create $\tilde{z}_t = \sqrt{w(x_t - x)} z_t$ and $\tilde{y}_t = \sqrt{w(x_t - x)} y_t$, because (15.18) is then the same as the first order conditions for a regression of $\tilde{y}_t$ on $\tilde{z}_t$ (without a constant). (An extension to a quadratic or higher

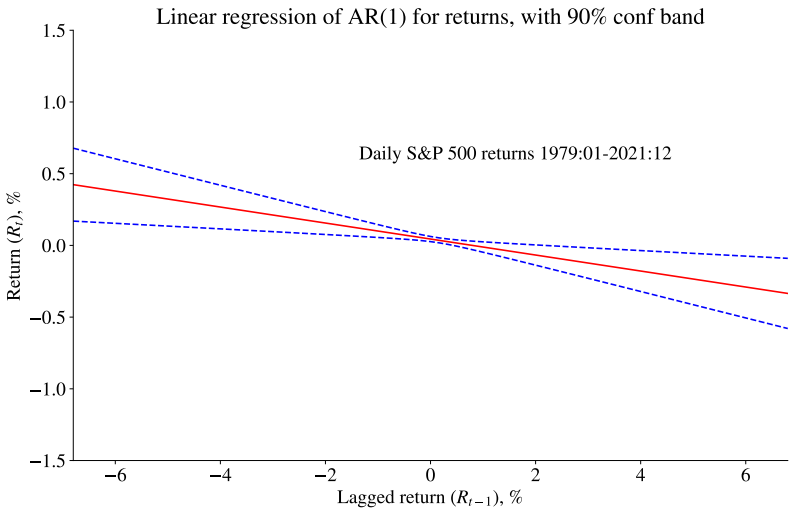


Figure 15.12: Linear regression with confidence bands

function seems straightforward.)

Clearly, solving (15.18)) gives one $\hat{\theta}_x$ vector for each x value that we consider. Once we have the estimates, the fitted value at the value x is just $\hat{\alpha}_x$ (since the regression function is $y_t = \alpha_x + \beta_x(x_t - x) + \varepsilon_t$ and we evaluate it at $x_t = x$.)

It can be shown that the local-linear estimator has the same asymptotic variance as the kernel regression, and that the bias only includes the $d^2b(x)/dx^2$ term (not the linear term). The latter means that the bias does not depend on the pdf of the regressor ($f(x)$), which is an advantage.

The bandwidth parameter (which only shows up in the calculations of the weights, $w(x_t - x)$) can be chosen by a leave-one-out cross validation approach or use the same rule of thumb choice as in Remark 15.8.

Remark 15.16 *(Rule of thumb value of h) Since Remark 15.8 effectively disregards the linear term in the bias (by assuming $df(x)/dx = 0$), it actually solves the same problem as for the local linear regression. The optimal h values is thus the same.*

Empirical Example 15.17 *(Local linear regression of AR(1) for daily S&P 500 returns) See Figures 15.15 – 15.18.*

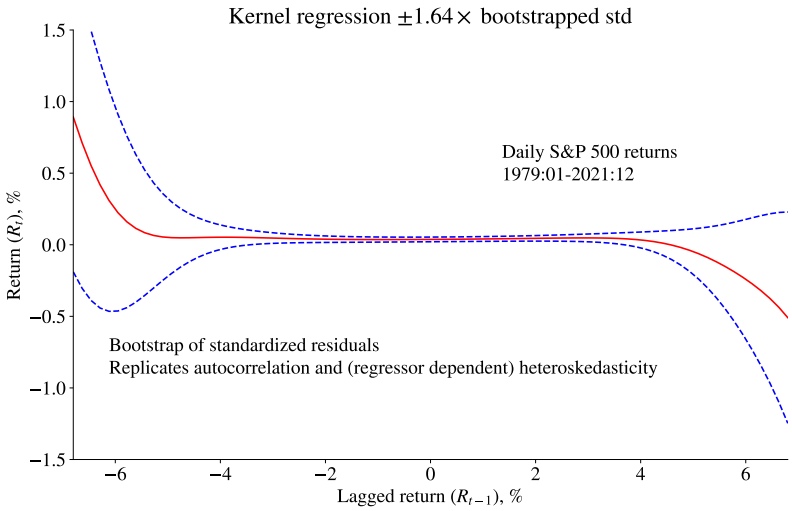


Figure 15.13: Kernel regression with bootstrapped confidence bands

15.4 Applications of Kernel Regressions

15.4.1 "Nonparametric Estimation of State-Price Densities Implicit in Financial Asset Prices," by Ait-Sahalia and Lo (1998)

Reference: Ait-Sahalia and Lo (1998)

There seem to be systematic deviations from the Black-Scholes model. For instance, implied volatilities are often higher for options far from the current spot (or forward) price—the volatility smile. This is sometimes interpreted as if the beliefs about the future log asset price put larger probabilities on very large movements than what is compatible with the normal distribution ("fat tails").

This has spurred many efforts to both describe the distribution of the underlying asset price and to amend the Black-Scholes formula by adding various adjustment terms. One strand of this literature uses nonparametric regressions to fit observed option prices to the variables that also show up in the Black-Scholes formula (spot price of underlying asset, strike price, time to expiry, interest rate, and dividends). For instance, Ait-Sahalia and Lo (1998) applies this to daily data for Jan 1993 to Dec 1993 on S&P 500 index options (14,000 observations).

This paper estimates nonparametric option price functions and calculates the implicit risk-neutral distribution as the second partial derivative of this function with respect to the strike price.

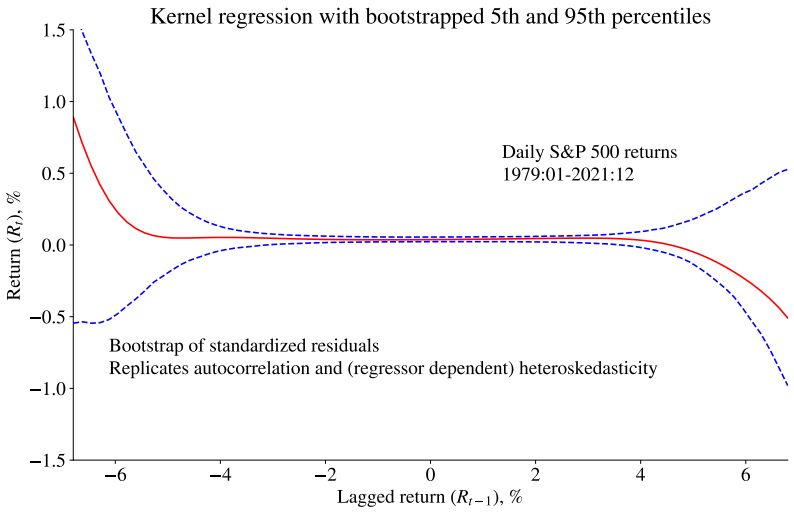


Figure 15.14: Kernel regression with bootstrapped confidence bands

1. First, the call option price, H_{it}, is estimated as a multivariate kernel regression

$$H_{it} = b(S_t, X, \tau, r_{\tau t}, \delta_{\tau t}) + \varepsilon_{it}, \tag{15.19}$$

   where S_t is the price of the underlying asset, X is the strike price, τ is time to expiry, $r_{\tau t}$ is the interest rate between t and $t + \tau$, and $\delta_{\tau t}$ is the dividend yield (if any) between t and $t + \tau$. It is very hard to estimate a five-dimensional kernel regression, so various ways of reducing the dimensionality are tried. For instance, by making $b()$ a function of the forward price, $S_t[\tau \exp(r_{\tau t} - \delta_{\tau t})]$, instead of S_t, $r_{\tau t}$, and $\delta_{\tau t}$ separably.

2. Second, the implicit risk-neutral pdf of the future asset price is calculated as $\partial^2 b(S_t, X, \tau, r_{\tau t}, \delta_{\tau t})/\partial X^2$, properly scaled so it integrates to unity.

3. This approach is used on daily data for Jan 1993 to Dec 1993 on S&P 500 index options (14,000 observations). They find interesting patterns of the implied moments (mean, volatility, skewness, and kurtosis) as the time to expiry changes. In particular, the nonparametric estimates suggest that distributions for longer horizons have increasingly larger skewness and kurtosis: whereas the distributions for short horizons are not too different from normal distributions, this is not true for longer horizons. (See their Fig 7.)

4. They also argue that there is little evidence of instability in the implicit pdf over

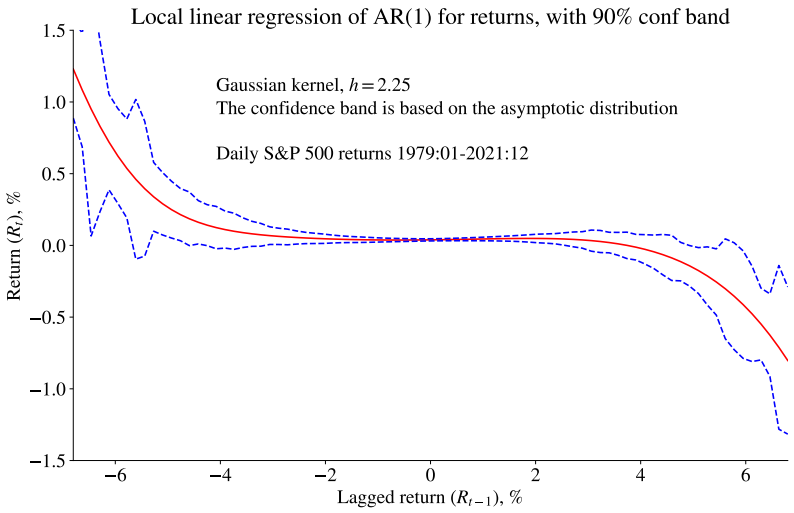


Figure 15.15: Non-parametric local linear regression with confidence bands

their sample.

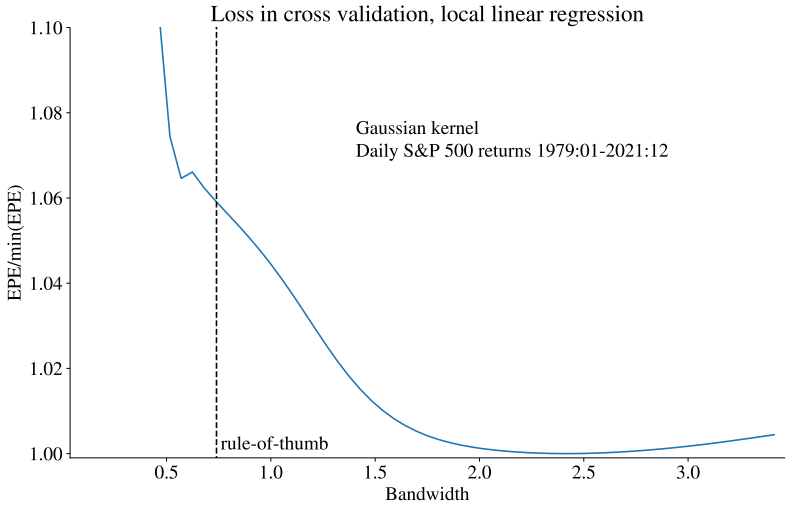


Figure 15.16: Cross-validation

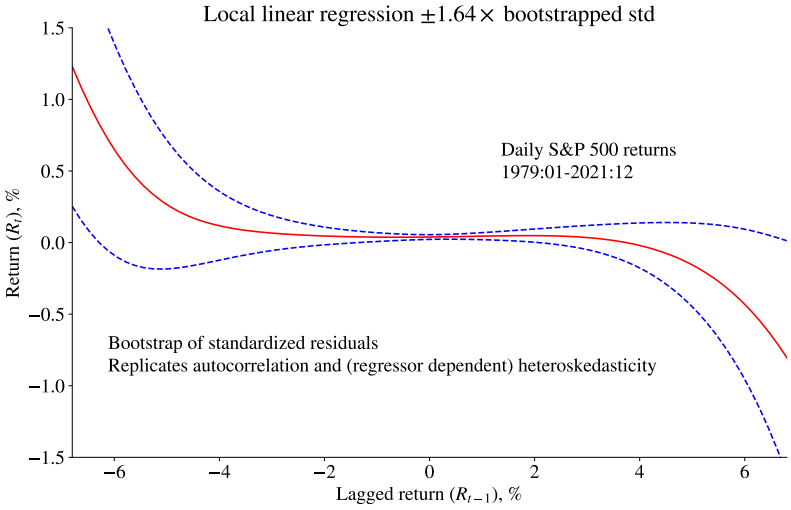


Figure 15.17: Non-parametric local linear regression with bootstrapped confidence bands

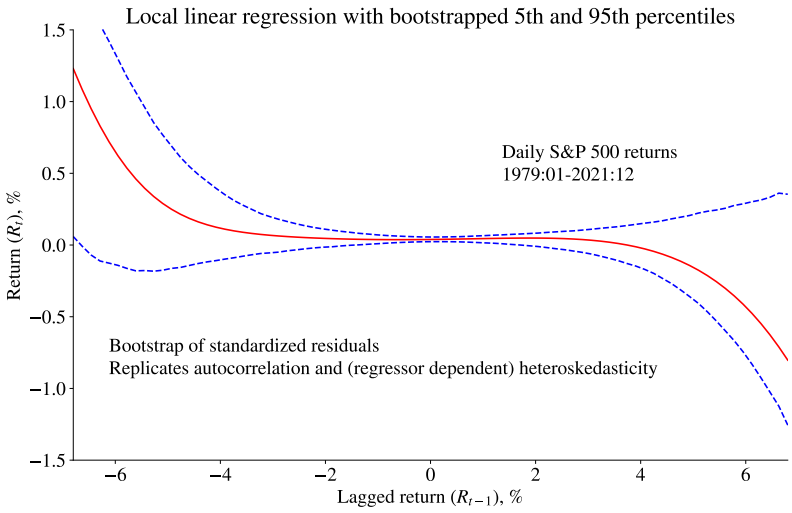


Figure 15.18: Non-parametric local linear regression with bootstrapped confidence bands

Chapter 16

Regression Discontinuity

(This file is a quick conversion from a file of slides. The formatting is thus not the best.)

Reference: Wooldridge ch. 21, Lee and Lemieux (JEL 2010), Lee (Journal of Econometrics 2008), Hansen 21

$$y_i = \alpha + \tau D_i + f(X_i;\beta) + \varepsilon_i, \text{ where} \tag{16.1}$$
$$D_i = \begin{cases} 0 \text{ if } X_i < c \\ 1 \text{ if } X_i \geq c \end{cases}$$

τ is the "treatment effect" (which depends on X, and X has also a direct effect)

Classical *example* (Thistlethwaite and Campbell, 1960): X is test scores, get merit award if $X \geq c$, Y future academic outcomes

Basic idea: X has 2 effects, one of which is the τ (due to the treatment kicking in at c)

Figure 16.1: linear case where $f(X_i;\beta) = \beta' X_i$

Why not IV? X_i affects treatment, but also has a direct effect

Why not "OLS"/other estimation: we want to *use only information around c* to measure the effect

You could think of the local X_i value to be partly "randomized" (could potentially handle a range of potential issues like excluded variables).

Gives *local* identification of the effect (valid more generally?)

Key assumption: the effect of treatment is discrete, all other effects (related to X) are continuous

Figure 16.2:

1. $\mathrm{E}\, Y(1)|X$ is expected Y at X if treated: $\alpha + \tau + f(X_i;\beta)$ in (16.1)

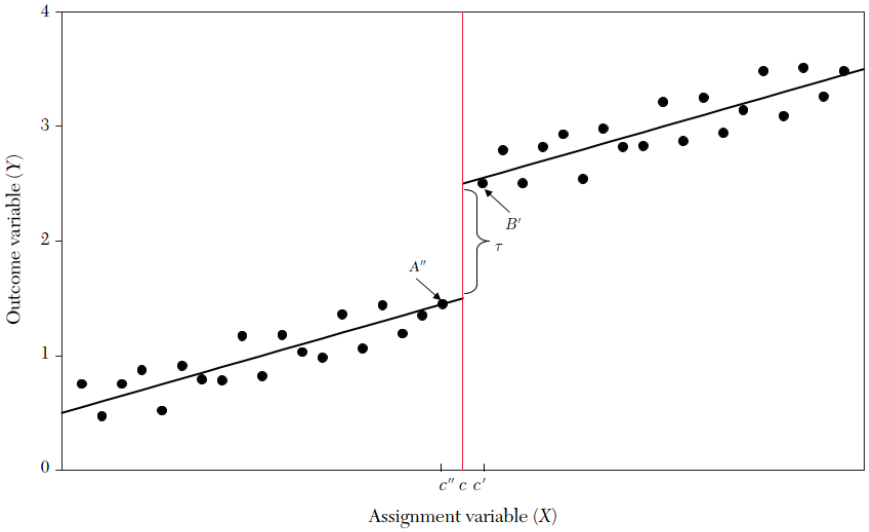


Figure 1. Simple Linear RD Setup

Figure 16.1: Lee and Lemieux Fig 1

2. $\mathrm{E}\,Y(0)|X$ is expected Y at X if not treated: $\alpha + f(X_i;\beta)$

3. We can only isolate τ if $f(X_i;\beta)$ is continuous at c (else we may measure $\tau +$ jump in $f()$)

Sharp or fuzzy? ($\Pr(\text{treatment}|X) = 0/1$ or just jumps up at c?) Here: focus is on sharp designs.

(Fuzzy designs require stronger assumptions and involve an IV like estimation)

use only information around c? In practice, we need to use a bit more...

1. Assuming linearity or some polynomial? Perhaps, but leads to global estimation

2. Average below/above? Overestimates the effect if $f'() > 0$: cf. *Figure 16.2*: $B' - A'$

3. Linear regression below/above? cf $B - A$

Notice:

2. is a kernel regression with a uniform kernel
3. is a local-linear regression with a *uniform kernel* (no contamination)

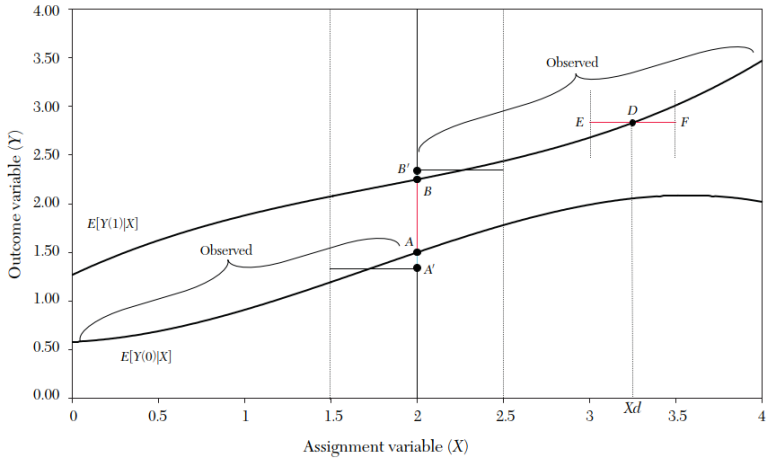


Figure 2. Nonlinear RD

Figure 16.2: Lee and Lemieux Fig 2

16.1 The Data

from Lee 2008: individual_final.dta: $N = 27,176$

notice the close overlap of y values (there are only 141 unique y values), Figure 16.3

...in my figures, I only use the unique observations, Figure 16.4

16.2 Parametric Estimates below/above c

$$y_i = \alpha + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^2 + e_i \text{ for } X_i < c$$
$$y_i = \delta + \gamma_1 X_i + \gamma_2 X_i^2 + \gamma_3 X_i^2 + e_i \text{ for } X_i \geq c$$

global estimates: "precise" but are they correct? Figure 16.5

16.3 Kernel Regression with a Uniform Kernel

...is very easy, see Figures 16.6 (as kernel regression) and Figure 16.7 (as mean-in-bin)

1. create bins for X_i values (make sure the bin boundary is at c)
2. for i st. X_i is in bin k, find the average y_i
3. repeat for every bin

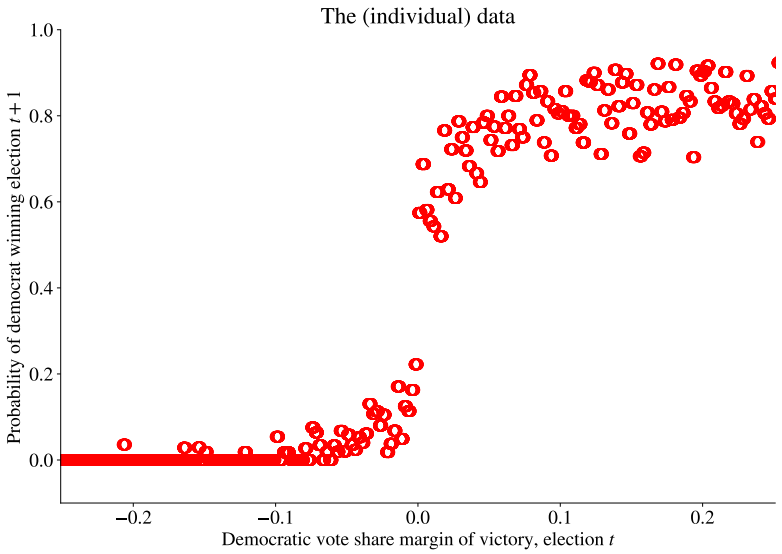


Figure 16.3: Lee (2008) data

4. Trickier:

   (a) choice of bins ("bandwidth"): rule of thumb or cross-validation

   (b) the confidence bands

Drawback: may fail to capture the fact that $f'(X_i;\beta) \neq 0$ inside the bin.

16.4 Variance of Mean in Bin

Figure 16.7 (mean-in-bin)

$\bar{y}_b$: mean in bin b, T_b data points, variance (within bin) σ_b^2

Classical expression

$$\text{Var}(\bar{y}_b) = \sigma_b^2 / T_b$$

From kernel regression with uniform kernel over bin $\pm h\sqrt{3}$

$$\text{Var}[\hat{b}(x)] = \frac{1}{Th}\frac{\sigma^2}{f(x)} \times \frac{1}{2\sqrt{3}}$$

Suppose $f(x)$ is estimated as a (normalised) histogram:

$$f_b(x) = \frac{T_b}{T \times (\text{bin width})} = \frac{T_b}{T 2h\sqrt{3}}$$

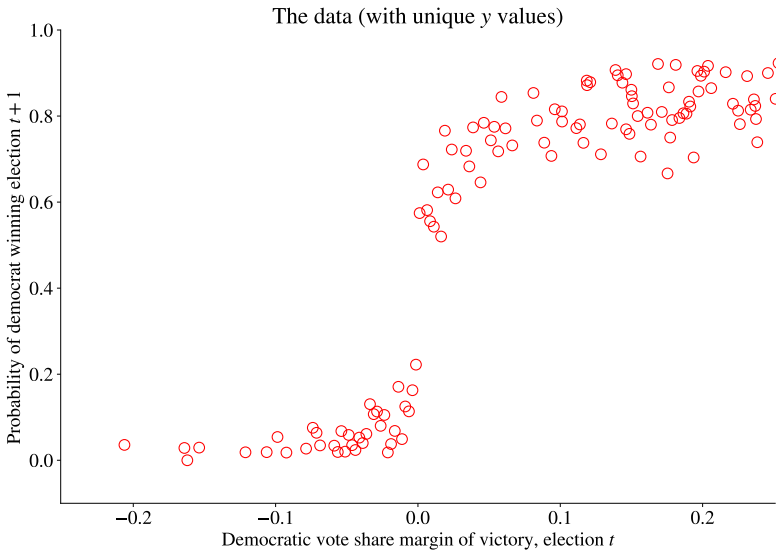


Figure 16.4: Lee (2008) data, unique cases

Combine to see that $\text{Var}[\hat{b}(x)] = \text{Var}(\bar{y}_b)$

16.5 Local Linear Regression with a Uniform Kernel

...is also very easy, see Figure 16.8 (kernel) and Figure 16.9 (LS in bin)

1. as before
2. for i st. X_i is in bin k, regress y_i on X_i
3. as before
4. Trickier: as before

16.6 More Regressors

$$y_i = \alpha + \tau D_i + f(X_i; \beta) + Z_i'\gamma + \varepsilon_i \tag{16.2}$$

Simplest approach: regress y_i on Z_i and use the residual to replace y_i in the "RDD estimation"

Alternative: estimate (16.2)

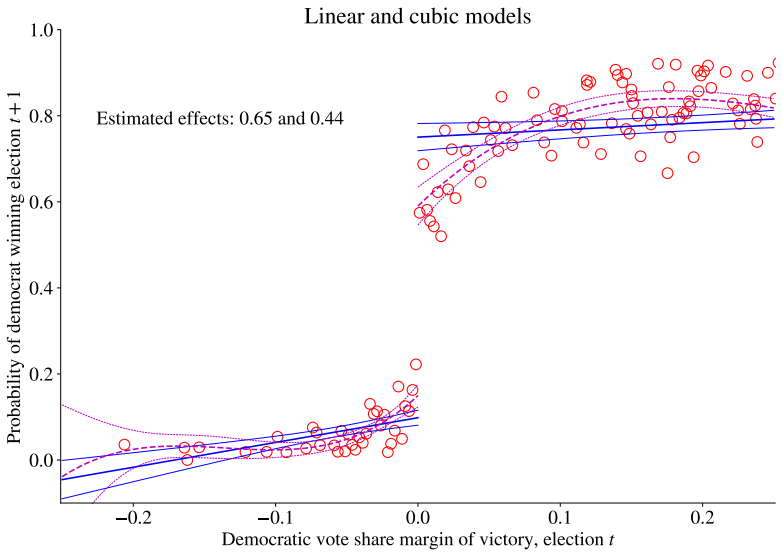


Figure 16.5: Parametric estimates (above/below threshold)

16.7 Distribution of Assignment Variable (X_i): Local Randomization or Not?

...no jump in cdf of X_i at c (informal test), see Figure 16.10

16.8 Regression Kink Designs

Reference: Card et al (Econometrica, 2015)

No jump, but slope changes

$$y_i = \alpha + D_i f_1(X_i;\beta_1) + (1 - D_i) f_2(X_i;\beta_2) + \varepsilon_i, \text{ where} \tag{16.3}$$

$$D_i = \begin{cases} 0 \text{ if } X_i < c \\ 1 \text{ if } X_i \geq c \end{cases}$$

where $f_1()$ and $f_2()$ have different derivatives ("slopes"), but where we require $f_1(c;\beta_2) = f_2(c;\beta_1)$ (continuous), see Figure 16.11

an example of a kinked regression function (but from a very different field): Figure 16.12

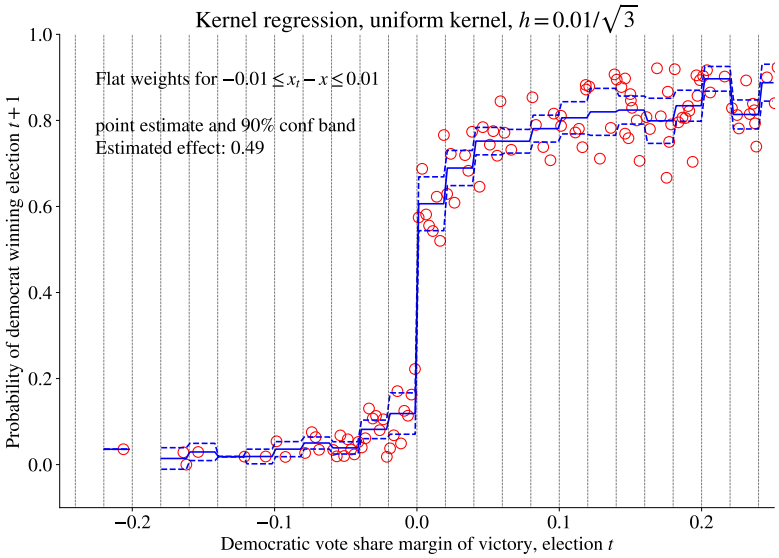


Figure 16.6: Non-parametric estimates, kernel regression

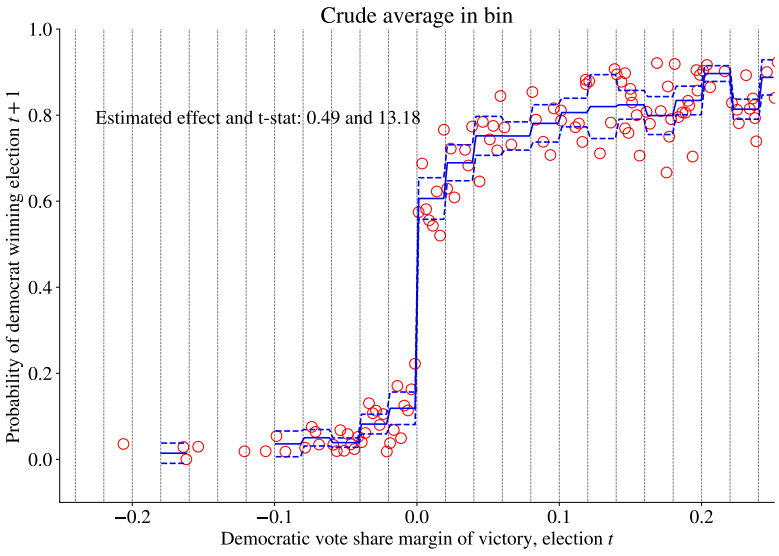


Figure 16.7: Mean in bin

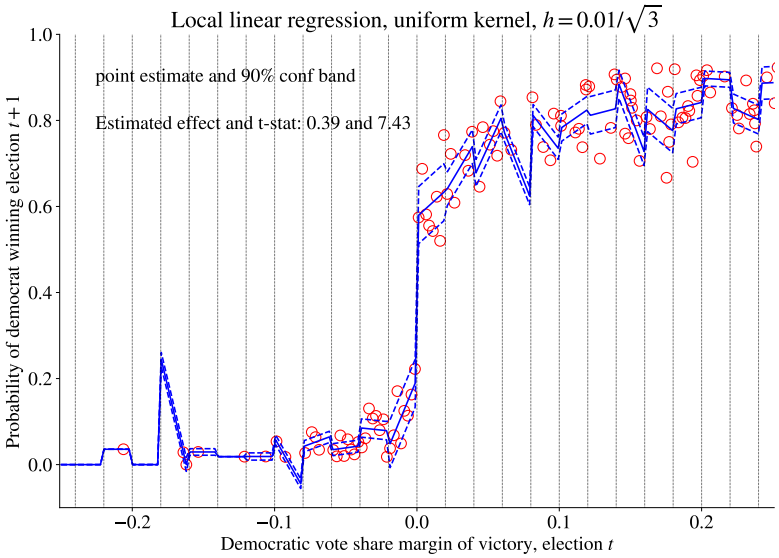


Figure 16.8: Non-parametric estimates, local linear regression

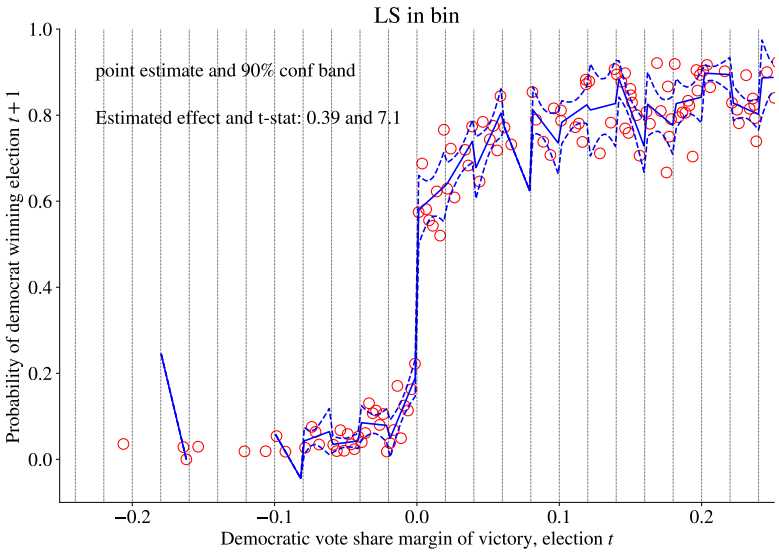


Figure 16.9: local linear regression

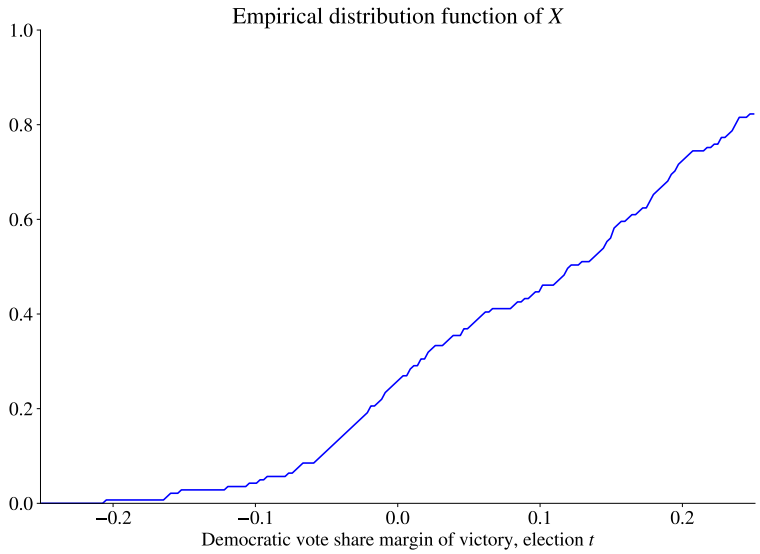


Figure 16.10: edf

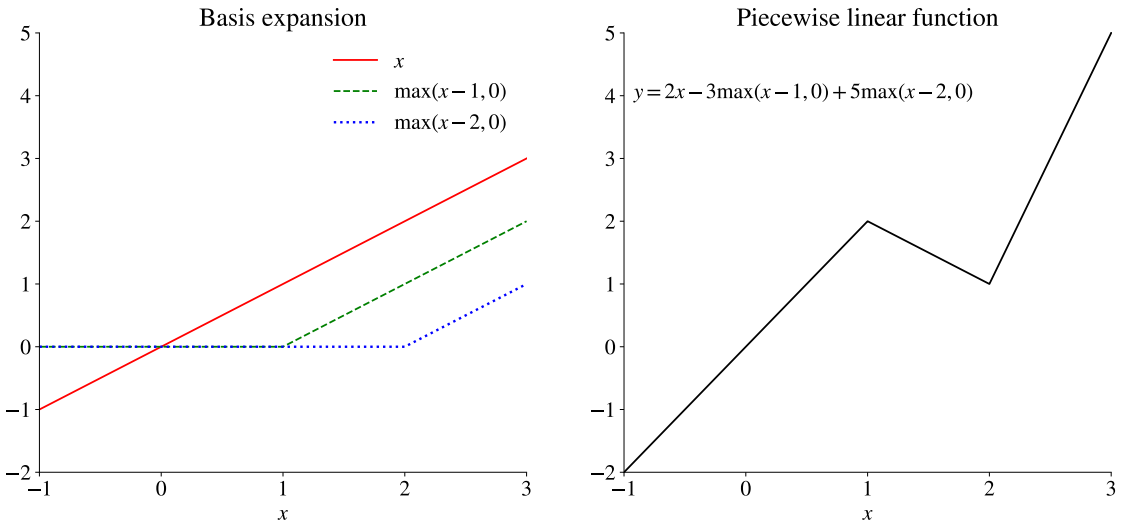
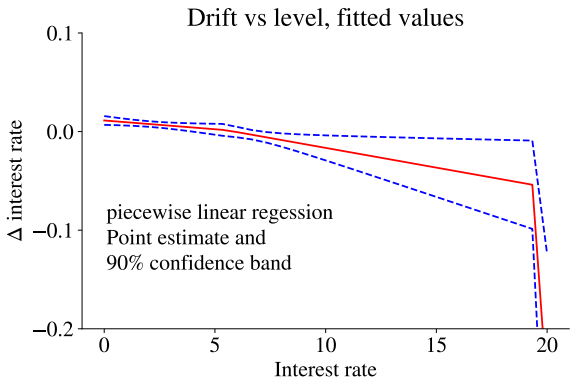


Figure 16.11: Example of piecewise linear function, created by basis expansion



Daily federal funds rates 1954:07-2021:12

	coeff	Std
knot 1	5.500	0.000
knot 2	19.516	0.301
slope seg 1	-0.002	0.001
extra slope seg 2	-0.002	0.003
extra slope seg 3	-0.427	0.176
const	0.011	0.003

Figure 16.12: Federal funds rate, piecewise linear model

Chapter 17

Instrumental Variables Method (IV)*

Reference: Verbeek (2012) 5, Greene (2018) 8.3; Hamilton (1994) 9.2; and Pindyck and Rubinfeld (1998) 7.

17.1 Instrumental Variables Method

When OLS is inconsistent (see Figure 17.1 for an example), then we typically apply MLE or the instrumental variables (IV) or 2SLS methods. This section describes the latter.

We want to estimate β in

$$y_t = x_t'\beta + u_t, \tag{17.1}$$

where x_t and β are vectors with k elements. Recall that OLS is defined by making the fitted residuals orthogonal (uncorrelated) with the regressors

$$\mathbf{0}_{kx1} = \sum_{t=1}^{T} x_t(y_t - x_t'\hat{\beta}). \tag{17.2}$$

Example 17.1 *(ARMA(1,1)) Consider the time series process $y_t = 0.9y_{t-1} + \varepsilon_t$ where $\varepsilon_t = v_t + 0.5v_{t-1}$. Notice that the regressor (y_{t-1}) is correlated with the residual (especially the v_{t-1} part), so OLS is inconsistent.*

The IV method replaces (17.2) by

$$\mathbf{0}_{kx1} = \sum_{t=1}^{T} z_t(y_t - x_t'\hat{\beta}_{iv}), \tag{17.3}$$

where z_t is a vector of k elements that have two key properties: (1) z_t is uncorrelated with the true residual (u_t) so z_t are *valid instruments*; but (2) correlated with the regressors (x_t) so z_t are *relevant instruments*. The first property cannot be directly checked since we never observe the true residuals. Instead, theoretical arguments must be used, but the

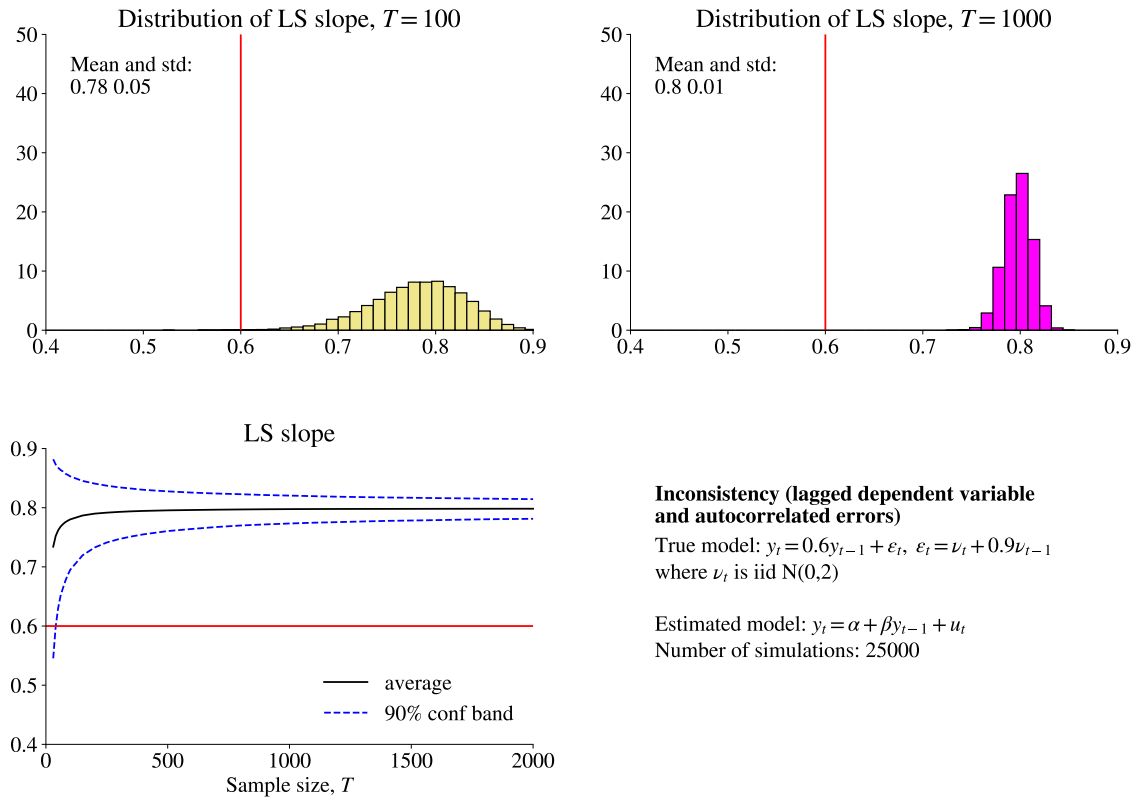


Figure 17.1: Results from a Monte Carlo experiment of LS estimation of the AR coefficient when data is from an ARMA process.

Hausman test can be of some help (see below). In particular, a valid instrument *cannot be endogenous* with respect to (that is, caused by) y_t and it *cannot be an erroneously excluded regressor*, because both cases would lead to $\text{Cov}(z_t, u_t) \neq 0$. A good application of the IV method must argue why that is not the case.

In contrast, the second property is easily checked by, for instance, regressing x_t on z_t and studying the t-statistics. (A perhaps more intuitive discussion of what the IV estimator does is found in the section on 2SLS.)

Example 17.2 *(ARMA(1,1) continued) Continuing Example 17.1, notice that y_{t-2} (or earlier lags) are not correlated with the residual so they could be used as instruments.*

Notice that some regressors (elements of x_t) may also be used as instruments (z_t). For instance, if just one of the regressors is an endogenous variable then we need (at least one) new instrument for that regressor, while the other regressors can be instruments for themselves.

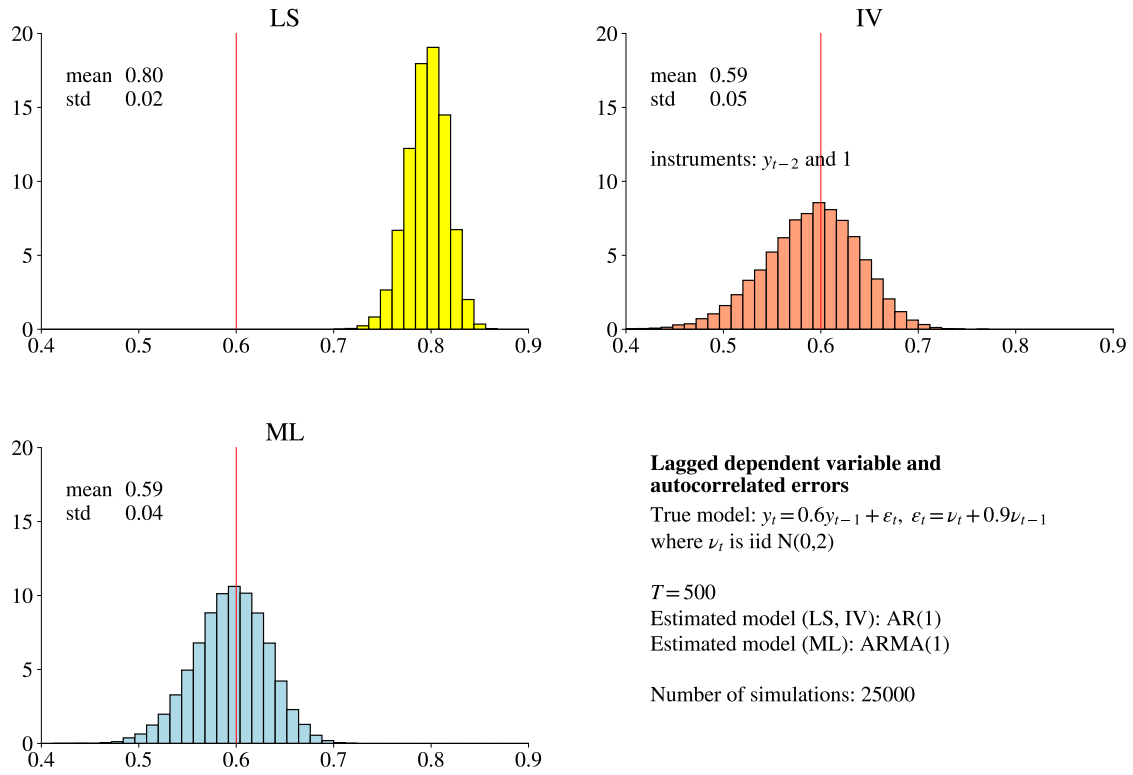


Figure 17.2: Results from a Monte Carlo experiment when data is from an ARMA process.

Example 17.3 *(Supply and Demand) Consider the simplest simultaneous equations model for supply and demand on a market are*

$$q_t = \gamma p_t + u_t^s,\ \gamma > 0 \text{ (supply)}$$
$$q_t = \beta p_t + \alpha A_t + u_t^d,\ \beta < 0 \text{ (demand),}$$

where A_t is an observable demand shock (perhaps income). To estimate the supply curve, the observable demand shocks A_t can be used as an instrument. See Figure 17.3 for an illustration.

Solving (17.3) gives the IV estimator

$$\hat{\beta}_{iv} = \left(\sum_{t=1}^{T} z_t x_t'\right)^{-1} \sum_{t=1}^{T} z_t y_t. \tag{17.4}$$

Clearly, this is the same as OLS when $z_t = x_t$. Notice that $\Sigma z_t x_t'$ must be invertible (have full rank) for this to work, that is, z_t and x_t must be correlated (or else z_t are not valid

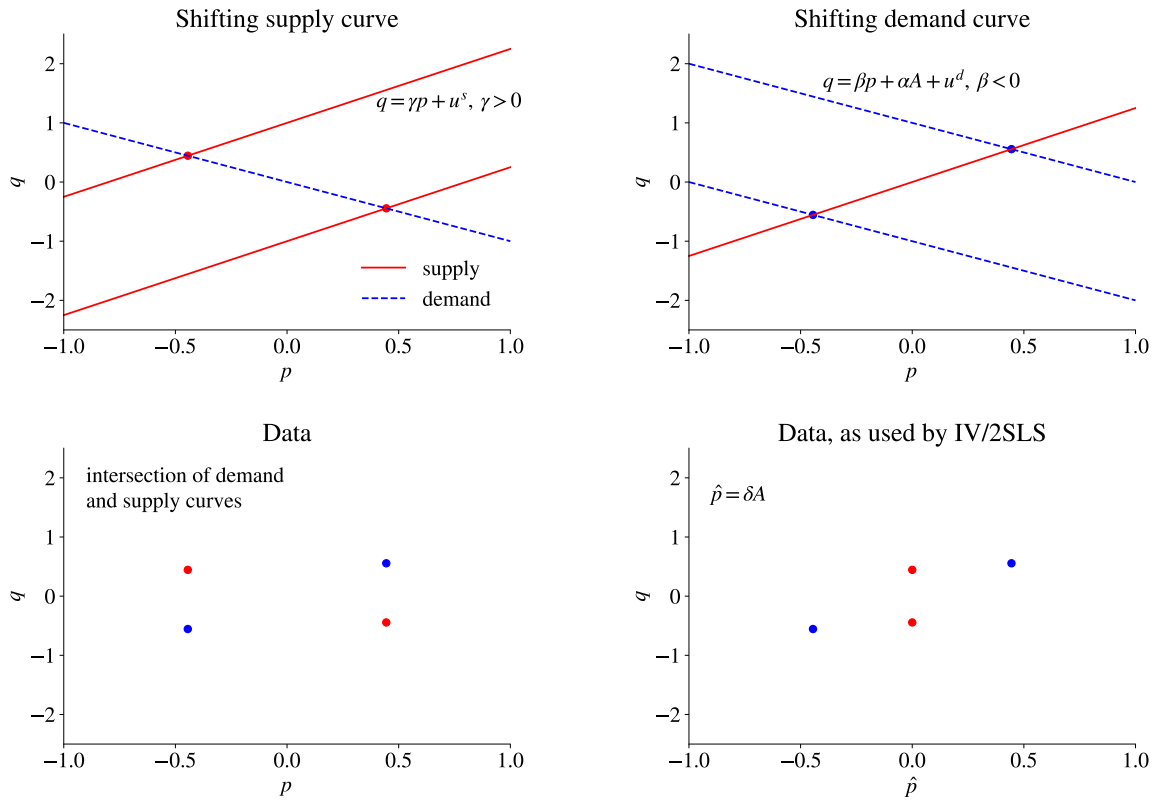


Figure 17.3: Illustration of demand and supply curves

instruments). We need as many instruments as regressors, but some can be instruments for themselves (if it can be argued that this regressor is not correlated with the true residual). There are few results on the small sample properties, although it is often found that there is a small sample bias.

Remark 17.4 *(Matrix notation) Let z_t' be the t^{th} row of Z and similarly for X. We then have $\hat{\beta}_{iv} = (Z'X)^{-1}(Z'Y)$.*

Figure 17.2 shows an example with an *ARMA(1,1)* process. The regressor (y_{t-1}) is correlated with the residual (v_{t-1}), so OLS is inconsistent. The IV method uses (1, y_{t-2}) as instruments for (1, y_{t-1}). Notice that (1, y_{t-2}) are indeed uncorrelated with the residual (which include shocks in t and $t-1$ but not earlier), but correlated with the regressors (because of the persistence of the y_t series).

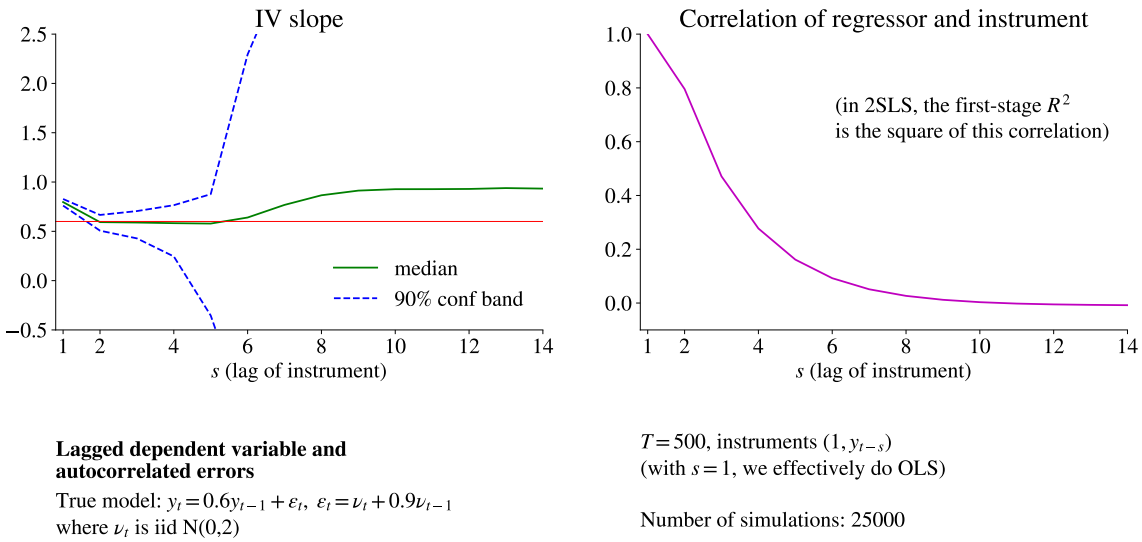


Figure 17.4: Results from a Monte Carlo experiment when data is from an ARMA process.

$\hat{\beta}_{iv}$ is (asymptotically) normally distributed so

$$\text{“}\hat{\beta}_{iv} \rightarrow^d \text{”}N(\beta, V), \text{ with} \tag{17.5}$$
$$V = S_{zx}^{-1} S S_{xz}^{-1} \text{ where } S = \text{Var}(\textstyle\sum_{t=1}^{T} z_t u_t)$$

and $S_{zx} = \Sigma_{t=1}^{T} z_t x_t'$. (See Section 17.3 for details.) This general expression is valid for both autocorrelated and heteroskedastic residuals—all such features are loaded into the S matrix. We can estimate S by replacing u_t by fitted residuals

$$\hat{u}_t = y_t - x_t' \beta_{iv}. \tag{17.6}$$

If the residuals are iid and independent of z_t (so $S = \sigma^2 S_{zz}$), then

$$V = \sigma^2 S_{zx}^{-1} S_{zz} S_{xz}^{-1}, \text{ if } u_t \text{ are iid.} \tag{17.7}$$

The IV estimator has often large standard deviations, especially with “weak instruments” (weak correlation with regressors). This is illustrated in Figure 17.4.

Example 17.5 *(*$\text{Var}(\hat{\beta}_{iv})$ *in the simplest case) Assume* y_t, x_t *and* z_t *are zero mean variables and that* z_t *and* u_t *are independent. Equation (17.7) for a simple regression can*

then be written

$$\begin{aligned}\text{Var}(\hat{\beta}_{iv}) &= \frac{\sigma^2 \text{Var}(z_t)/T}{\text{Cov}(x,z)^2} \\ &= \frac{\sigma^2/T}{\text{Var}(x_t)} \frac{1}{\text{Corr}(x_t,z_t)^2}.\end{aligned}$$

If $\text{Corr}(x_t, z_t) = 1$ *or* -1*, then this is the same as with OLS (but the consistency can be questioned in this case). Instead, with a low* $\text{Corr}(x_t, z_t)^2$ *value (weak instruments), then the uncertainty increases.*

17.2 Two-stages-least squares (2SLS)

2SLS is the same as IV when there are as many instruments (L) as there are regressors (k). When there are more instruments than regressors ($L > k$), then 2SLS can produce more precise (efficient) estimates than IV. It proceeds in two steps.

First, regress each of the elements in x_t on

$$x_{it} = \delta_i' z_t + \varepsilon_t, \text{ for } i = 1 \text{ to } k. \tag{17.8}$$

where δ_i is a vector with L elements and let $\hat{x}_{it}$ be the fitted values

$$\hat{x}_{it} = \delta_i' z_t. \tag{17.9}$$

We can stack the equations as

$$\hat{x}_t = \delta' z_t, \tag{17.10}$$

where δ is an $L \times k$ matrix with δ_i in column i. The fit (or t-stats) of these regressions are often used to assess if the instruments are relevant, but we typically require very high |t-stats|, see the discussion of weak instruments.

Second, regress y_t on the fitted values $\hat{x}_t$

$$y_t = \beta' \hat{x}_t + v_t. \tag{17.11}$$

Remark 17.6 *(Alternative to (17.11)*) We could equally well use* $\hat{x}_t$ *instead of* z_t *in the IV estimator (17.4). This gives the same result as (17.11), provided that the instruments in the first stage estimation (17.8) include all "non-problematic" regressors.*

Similarly to IV, the small sample properties are poor if the first-stage regression has a

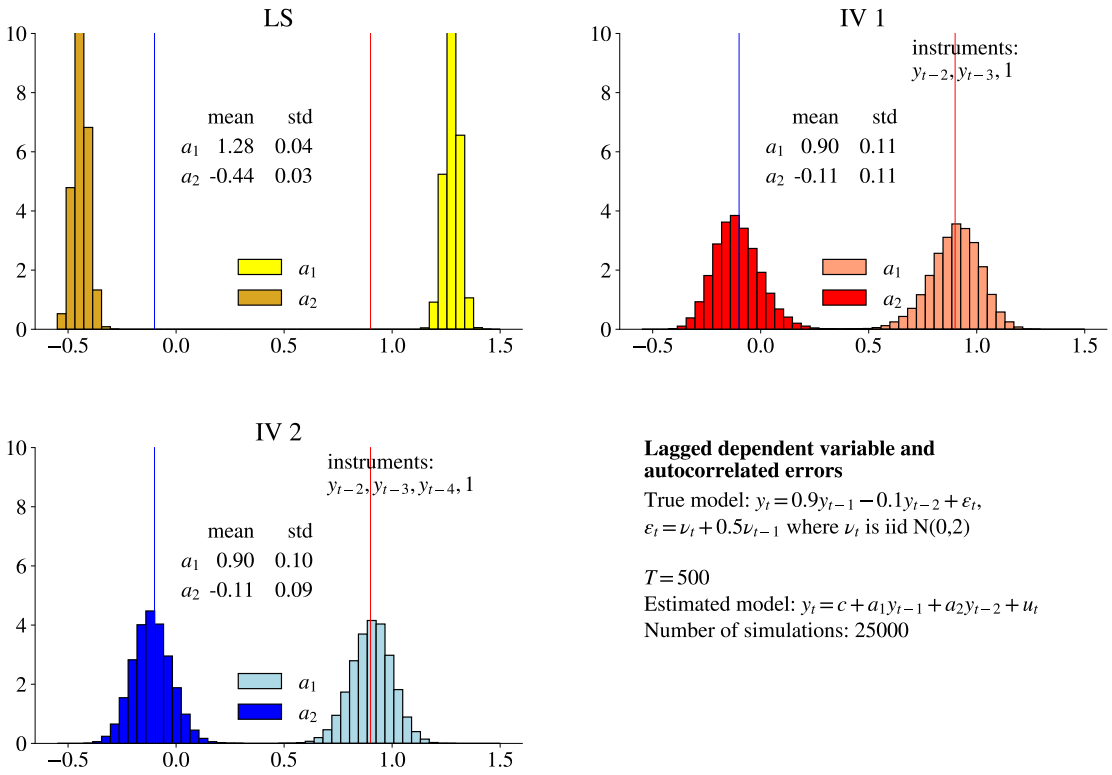


Figure 17.5: Results from a Monte Carlo experiment when data is from an ARMA process.

low R^2 ("weak instruments"), but also when R^2 is extremely high, since then the orthogonality conditions are likely to be violated — see Figure 17.4.

The 2SLS approach highlights the key idea of IV (and 2SLS): in the regression (17.11) we only consider those movements in the regressors that are correlated with z_t (as captured by $\hat{x}_t$). Since z_t is chosen to be uncorrelated with the residuals, but correlated with x_t, we are only using the "clean" co-movements of x_t and y_t to estimate the coefficients. See Figure 17.3 for an illustration.

See Figure 17.5 for a case where using more instruments gives more precise estimates.

It can be shown (see Section 17.3 for details) that the (asymptotically valid) variance-covariance matrix for $\hat{\beta}_{SLS}$ is

$$\begin{aligned} V &= BSB', \text{ where} \\ B &= (S_{xz}S_{zz}^{-1}S_{zx})^{-1}S_{xz}S_{zz}^{-1} \text{and } S = \text{Var}(\textstyle\sum_{t=1}^{T} z_t u_t). \end{aligned} \tag{17.12}$$

This general expression is valid for both autocorrelated and heteroskedastic residuals

since thoses features are loaded into the S matrix. We can estimate S as in the IV case: by replacing u_t by the fitted residuals

$$\hat{u}_t = y_t - x_t'\beta_{iv}. \tag{17.13}$$

Notice that these are *not* the same as the fitted residuals from the 2nd stage regression. If the residuals are iid and independent of z_t (so $S = \sigma^2 S_{zz}$), then V simplifies to

$$V = \sigma^2(S_{xz}S_{zz}^{-1}S_{zx})^{-1} \text{ if } u_t \text{ are iid.} \tag{17.14}$$

Example 17.7 *(ARMA(2,1)) $y_t = 0.6y_{t-1} + 0.3y_{t-2} + \varepsilon_t$ where $\varepsilon_t = v_t + 0.5v_{t-1}$. Notice that y_{t-1} is correlated with v_{t-1} but y_{t-2} is not. We could therefore use $y_{t-2}, y_{t-3}, ...$ as instruments for the two regressors.*

Example 17.8 *(Supply and Demand) Continuing Example 17.3,it can be shown that regressing q_t on $\hat{p}_t$ (where $\hat{p}_t = \delta A_t$) will give a consistent estimate of γ: see Figure 17.3 for an illustration. (Example 17.11) calculates the probability limit of $\hat{\delta}$).*

Empirical Example 17.9 *(Wage equation) Tables 17.1–17.2 shows results from an example in Hill, Griffiths, and Lim (2008) 10.3.3. The purpose is to estimate how log wages depend on education, experience and experience^2, while treating education as an endogenous variable. The instruments are experience, experience^2 and the the education of the mother: see the first stage regression in 17.1. The result is fairly different from the OLS regression: see Table 17.2.*

Remark 17.10 *(Overidentifying restrictions in 2SLS*) When we use 2SLS, then we can test if instruments affect the dependent variable only via their correlation with the regressors. If not, something is wrong with the model since some relevant variables are excluded from the regression. A simple test is to first estimate with 2SLS to get the fitted residuals $\hat{u}_t$, then regress those on z_t. The TR^2 from this second regression is (under the null hypothesis) χ^2_{df} with df being the number of overidentifying restrictions.*

17.3 Consistency and Asymptotic Distributions of the IV and 2SLS Estimators*

This section gives some details of the asymptotic properties of IV and 2SLS.

	1st stage
c	9.775
	(23.753)
exper	0.049
	(1.152)
exper^2	−0.001
	(−0.959)
mothereduc	0.268
	(8.481)
T	428

Table 17.1: First stage estimation of the 'educ' variable. Example of IV estimation, Hill et al (2008), section 10.3.3. Instruments: c, exper, exper^2, and mothereduc. Numbers in parentheses are t-stats (from White's method).

17.3.1 Asymptotic Results on the IV Estimator*

There are few results on small sample properties, but it is often noticed that IV is often imprecise and even biased. In large samples, we typically get consistency and a normal distribution.

Use (17.1) to substitute for y_t in (17.4), multiply both sides by $\sqrt{T}$ and rearrange as

$$\sqrt{T}(\hat{\beta}_{iv} - \beta) = \hat{\Sigma}_{zx}^{-1}\sqrt{T}\textstyle\sum_{t=1}^{T} z_t u_t / T, \text{ where} \qquad (17.15)$$
$$\hat{\Sigma}_{zx} = \textstyle\sum_{t=1}^{T} z_t x_t' / T.$$

Since we have strong beliefs that $\text{Cov}(z_t, u_t) = 0$, this expression shows that $\hat{\beta}_{iv}$ should be consistent.

Example 17.11 *(Supply and Demand) Continuing Example 17.3, we can solve for the two endogenous variables (the "reduced form") as*

$$\begin{bmatrix} q_t \\ p_t \end{bmatrix} = \begin{bmatrix} \frac{\gamma\alpha}{\gamma-\beta} \\ \frac{\alpha}{\gamma-\beta} \end{bmatrix} A_t + \begin{bmatrix} \frac{\beta}{\beta-\gamma} & -\frac{\gamma}{\beta-\gamma} \\ \frac{1}{\beta-\gamma} & -\frac{1}{\beta-\gamma} \end{bmatrix} \begin{bmatrix} u_t^s \\ u_t^d \end{bmatrix}.$$

Suppose we estimate the supply curve by using A_t as the instrument. For simplicity, assume all variables have zero means. Then the probability limit of (17.4) is

$$\text{plim}\, \hat{\gamma}_{iv} = \text{Cov}(p_t, A_t)^{-1}\, \text{Cov}(q_t, A_t).$$

From the reduced form we have (assuming A_t, u_t^d and u_t^s are uncorrelated) $\text{Cov}(p_t, A_t) =$

	OLS	IV/2SLS
c	−0.522	0.198
	(−2.641)	(0.407)
educ	0.107	0.049
	(7.634)	(1.301)
exper	0.042	0.045
	(3.170)	(2.888)
exper^2	−0.001	−0.001
	(−2.073)	(−2.145)
T	428	428

Table 17.2: IV estimation of wage equation. Example of IV estimation, Hill et al (2008), section 10.3.3. Instruments: c, exper, exper^2, and mothereduc. Numbers in parentheses are t-stats (from White's method).

$\frac{\alpha}{\gamma-\beta}\text{Var}(A_t)$ *and* $\text{Cov}(q_t, A_t) = \frac{\gamma\alpha}{\gamma-\beta}\text{Var}(A_t)$. *Combining gives* $\text{plim}\,\hat{\gamma}_{iv} = \gamma$. *Also notice that* $\text{plim}\,\hat{\delta} = \text{Cov}(p_t, A_t)/\text{Var}(A_t) = \alpha/(\gamma-\beta)$.

Example 17.12 *(Supply and Demand) Continuing Examples 17.3 and 17.11, we have* $\text{plim}\,\hat{\delta} = \alpha/(\gamma-\beta)$. *Thus, regressing* q_t *on* $\hat{p}_t = \delta A_t$ *has a probability limit of* $\text{Cov}(q_t, \delta A_t)/\text{Var}(\delta A_t)$, *which (using Example 17.11) can be simplified to* $\frac{\gamma\alpha}{\gamma-\beta}/\delta = \gamma$.

Since $\sqrt{T}\Sigma_{t=}^{T} z_t u_t/T$ in (17.15) is $\sqrt{T}\times$a sample average, it is plausible that a CLT applies so the asymptotic distribution $\sqrt{T}(\hat{\beta}_{iv}-\beta)$ might be normal with a zero mean and a variance-covariance matrix

$$\text{Var}(\sqrt{T}\hat{\beta}_{iv}) = \Sigma_{zx}^{-1}\Sigma\Sigma_{xz}^{-1}, \text{ where } \Sigma = \text{Var}\left(\textstyle\sum_{t=1}^{T} z_t u_t/\sqrt{T}\right). \tag{17.16}$$

and where Σ_{zx} is the probability limit of $\hat{\Sigma}_{zx}$. The last matrix in the covariance matrix follows from $(\Sigma_{zx}^{-1})' = (\Sigma_{zx}')^{-1} = \Sigma_{xz}^{-1}$. Dividing both sides by T and rewriting gives (17.5). (Details: Σ_{zx} is the probability limit of S_{zx}/T and $\Sigma = S/T$. Use this in (17.16) and simplify to get the probability limit of $TS_{zx}^{-1}SS_{xz}^{-1}$. Divide both sides by T to get $\text{Var}(\hat{\beta}_{iv})$ which then equals (17.5).)

17.3.2 Asymptotic Results on 2SLS*

From (17.8)–(17.9) we have

$$\begin{aligned}\hat{\delta} &= \hat{\Sigma}_{zz}^{-1}\hat{\Sigma}_{zx},\\ \hat{\beta} &= \hat{\Sigma}_{\hat{x}\hat{x}}^{-1}\hat{\Sigma}_{\hat{x}y}\end{aligned} \tag{17.17}$$

where $\hat{\Sigma}_{zz} = \sum_{t=1}^{T} z_t z_t'/T$, and so forth. Notice that $\hat{\Sigma}_{zz}^{-1}$ is an $L \times L$ matrix and $\hat{\Sigma}_{zx}$ is an $L \times k$ matrix, so $\hat{\delta}$ is $L \times k$ as mentioned around (17.10). The fitted values in (17.10) can then be written

$$\begin{aligned}\hat{x}_t &= \hat{\delta}' z_t\\ &= \hat{\Sigma}_{xz}\hat{\Sigma}_{zz}^{-1} z_t,\end{aligned} \tag{17.18}$$

so

$$\begin{aligned}\hat{\Sigma}_{\hat{x}\hat{x}} &= \textstyle\sum_{t=1}^{T}\hat{x}_t\hat{x}_t'/T = \hat{\Sigma}_{xz}\hat{\Sigma}_{zz}^{-1}\hat{\Sigma}_{zx} \text{ and}\\ \hat{\Sigma}_{\hat{x}y} &= \textstyle\sum_{t=1}^{T}\hat{x}_t y_t/T = \hat{\Sigma}_{xz}\hat{\Sigma}_{zz}^{-1}\hat{\Sigma}_{zy}.\end{aligned} \tag{17.19}$$

(Substitute for $\hat{x}$ from (17.18) and simplify to derive this.)

Using these results in the equations of $\hat{\delta}$ and $\hat{\beta}$ (17.17) gives

$$\hat{\beta} = (\hat{\Sigma}_{xz}\hat{\Sigma}_{zz}^{-1}\hat{\Sigma}_{zx})^{-1}\hat{\Sigma}_{xz}\hat{\Sigma}_{zz}^{-1}\hat{\Sigma}_{zy} \tag{17.20}$$

Substituting for y_t by using (17.1) and expanding gives

$$\begin{aligned}\hat{\beta} &= (\hat{\Sigma}_{xz}\hat{\Sigma}_{zz}^{-1}\hat{\Sigma}_{zx})^{-1}\hat{\Sigma}_{xz}\hat{\Sigma}_{zz}^{-1}\textstyle\sum_{t=1}^{T} z_t(x_t'\beta + u_t)/T\\ &= \beta + \underbrace{(\hat{\Sigma}_{xz}\hat{\Sigma}_{zz}^{-1}\hat{\Sigma}_{zx})^{-1}\hat{\Sigma}_{xz}\hat{\Sigma}_{zz}^{-1}}_{A}\textstyle\sum_{t=1}^{T} z_t u_t.\end{aligned} \tag{17.21}$$

Consistency follows from plim $\sum_{t=1}^{T} z_t u_t/T = 0$ and asymptotic normality from a CLT applied to $\sqrt{T}\sum_{t=1}^{T} z_t u_t/T$ and the asymptotic variance-covariance matrix is

$$\text{Var}(\sqrt{T}\hat{\beta}_{iv}) = A\Sigma A', \text{ where } \Sigma = \text{Var}(\Sigma_{t=}^{T} z_t u_t/\sqrt{T}) \tag{17.22}$$

This can be rewritten as (17.12). (Details: Σ_{zx} is the probability limit of S_{zx}/T etc and $\Sigma = S/T$. Divide both sides by T.)

Remark 17.13 *(*Alternative expression for A) By using (17.17), A in (17.21) can also*

be written $A = (\hat{\delta}' \Sigma_{zz} \hat{\delta})^{-1} \hat{\delta}'$.

17.4 Hausman's Specification Test

Reference: Greene (2018) 8.6

This test is constructed to test if an efficient estimator (like LS) gives (approximately) the same estimate as a consistent estimator (like IV). If not, the efficient estimator is most likely inconsistent. It is therefore a way to test for the presence of endogeneity and/or measurement errors.

Let $\hat{\beta}_e$ be an estimator that is consistent and asymptotically efficient when the null hypothesis, H_0, is true, but inconsistent when H_0 is false. Let $\hat{\beta}_c$ be an estimator that is consistent under both H_0 and the alternative hypothesis. When H_0 is true, the asymptotic distribution is such that

$$\text{Cov}(\hat{\beta}_e, \hat{\beta}_c) = \text{Var}(\hat{\beta}_e). \tag{17.23}$$

Proof. (of 17.23, univariate version*) Consider the estimator $\lambda \hat{\beta}_c + (1-\lambda)\hat{\beta}_e$, which is clearly consistent under H_0 since both $\hat{\beta}_c$ and $\hat{\beta}_e$ are. The asymptotic variance of this estimator is

$$\lambda^2 \text{Var}(\hat{\beta}_c) + (1-\lambda)^2 \text{Var}(\hat{\beta}_e) + 2\lambda(1-\lambda)\text{Cov}(\hat{\beta}_c, \hat{\beta}_e),$$

which is minimized at $\lambda = 0$ (since $\hat{\beta}_e$ is asymptotically efficient). The first order condition with respect to λ

$$2\lambda \text{Var}(\hat{\beta}_c) - 2(1-\lambda)\text{Var}(\hat{\beta}_e) + 2(1-2\lambda)\text{Cov}(\hat{\beta}_c, \hat{\beta}_e) = 0$$

should therefore be zero at $\lambda = 0$ so

$$\text{Var}(\hat{\beta}_e) = \text{Cov}(\hat{\beta}_c, \hat{\beta}_e).$$

(See Davidson (2000) 8.1) ■

This means that we can write

$$\begin{aligned} \text{Var}(\hat{\beta}_e - \hat{\beta}_c) &= \text{Var}(\hat{\beta}_e) + \text{Var}(\hat{\beta}_c) - 2\,\text{Cov}(\hat{\beta}_e, \hat{\beta}_c) \\ &= \text{Var}(\hat{\beta}_c) - \text{Var}(\hat{\beta}_e). \end{aligned} \tag{17.24}$$

We can use this to test, for instance, if the estimates from least squares ($\hat{\beta}_e$, since LS is efficient if errors are iid normally distributed) and instrumental variable method ($\hat{\beta}_c$,

since consistent even if the true residuals are correlated with the regressors) are the same. In this case, H_0 is that the true residuals are uncorrelated with the regressors.

All we need for this test are the point estimates and consistent estimates of the variance-covariance matrices. Testing one of the coefficient can be done by a t test, and testing all the parameters by a χ^2 test

$$(\hat{\beta}_e - \hat{\beta}_c)' \operatorname{Var}(\hat{\beta}_e - \hat{\beta}_c)^{-1} (\hat{\beta}_e - \hat{\beta}_c) \sim \chi^2_j, \tag{17.25}$$

where the covariance matrix is from (17.24) and where j equals the number of regressors that are potentially endogenous or measured with error. Note that the covariance matrix is likely to have a reduced rank, so the inverse needs to be calculated as a generalized (pseudo) inverse.

Chapter 18

Predicting Asset Returns

Sections denoted by a star (*) is not required reading.

Reference: Cochrane (2005) 20.1; Campbell, Lo, and MacKinlay (1997) 2 and 7; Campbell (2018) 5; Taylor (2005) 5–7; Elliot and Timmermann (2016)

18.1 A Little Financial Theory and Predictability

The traditional interpretation of autocorrelation in asset returns is that there are some "irrational traders." For instance, feedback trading would create positive short term autocorrelation in returns. If there are non-trivial market imperfections, then predictability can be used to generate economic profits.

In contrast, if there are no important market imperfections, then predictability of excess returns should be thought of as predictable movements in risk premia. To see the latter, let R^e_{t+1} be the excess return on an asset. The canonical asset pricing equation says

$$\mathrm{E}_t\, m_{t+1} R^e_{t+1} = 0, \tag{18.1}$$

where m_{t+1} is the stochastic discount factor.

Remark 18.1 *(A consumption-based model) Suppose we want to maximize the expected discounted sum of utility* $\mathrm{E}_t\, \Sigma_{s=0}^{\infty} \beta^s u(c_{t+s})$. *Let* Q_t *be the consumer price index in* t. *Then, we have*

$$m_{t+1} = \begin{cases} \beta \frac{u'(c_{t+1})}{u'(c_t)} \frac{Q_t}{Q_{t+1}} & \textit{if returns are nominal} \\ \beta \frac{u'(c_{t+1})}{u'(c_t)} & \textit{if returns are real.} \end{cases}$$

We can rewrite (18.1) (using $\mathrm{Cov}(x, y) = \mathrm{E}\, xy - \mathrm{E}\, x\, \mathrm{E}\, y$) as

$$\mathrm{E}_t\, R^e_{t+1} = -\mathrm{Cov}_t(m_{t+1}, R^e_{t+1}) / \mathrm{E}_t\, m_{t+1}. \tag{18.2}$$

This says that the expected excess return will vary if risk (the covariance) does. If we can model how these expected returns change over time, then we have a forecasting model for returns. (If the expectations are not too crazy, then the forecasting model may actually forecast future returns...)

Example 18.2 *(Epstein-Zin utility function) Epstein and Zin (1991) define a certainty equivalent of future utility as* $Z_t = [\mathrm{E}_t(U_{t+1}^{1-\gamma})]^{1/(1-\gamma)}$ *where* γ *is the risk aversion—and then use a CES aggregator function to govern the intertemporal trade-off between current consumption and the certainty equivalent:* $U_t = [(1-\delta)C_t^{1-1/\psi} + \delta Z_t^{1-1/\psi}]^{1/(1-1/\psi)}$ *where* ψ *is the elasticity of intertemporal substitution. If returns are iid (so the consumption-wealth ratio is constant), then it can be shown that this utility function has the same pricing implications as the CRRA utility, that is,*

$$\mathrm{E}[(C_t/C_{t-1})^{-\gamma} R_t] = \textit{constant}.$$

(See Söderlind (2006) for a simple proof.) The point is that without predictability, the Epstein-Zin utility function has the same implications as the CRRA utility function. Establishing whether there is predictability is therefore a way to assess the importance of the theory.

Example 18.3 *(Portfolio choice with predictable returns) Campbell and Viceira (1999) specify a model where the log return of the only risky asset follows the time series process*

$$r_{t+1} = r_f + x_t + u_{t+1},$$

where r_f *is a constant riskfree rate,* u_{t+1} *is unpredictable, and the state variable follows (constant suppressed)*

$$x_{t+1} = \phi x_t + \eta_{t+1},$$

where η_{t+1} *is also unpredictable. Clearly,* $\mathrm{E}_t(r_{t+1} - r_f) = x_t$. $\mathrm{Cov}_t(u_{t+1}, \eta_{t+1})$ *can be non-zero. For instance, with* $\mathrm{Cov}_t(u_{t+1}, \eta_{t+1}) < 0$*, a high return (*$u_{t+1} > 0$*) is typically associated with an expected low future return (*x_{t+1} *is low since* $\eta_{t+1} < 0$*). With Epstein-Zin preferences, the portfolio weight on the risky asset is (approximately) of the form*

$$v_t = a_0 + a_1 x_t,$$

where a_0 *and* a_1 *are complicated expression (in terms of the model parameters—can be calculated numerically). There are several interesting results. First, if returns are not*

predictable (x_t is constant since η_{t+1} is), then the portfolio choice is constant. Second, when returns are predictable, but the relative risk aversion is unity (no intertemporal hedging), then $v_t = 1/(2\gamma) + x_t/[\gamma \operatorname{Var}_t(u_{t+1})]$, so predictability does still not matter. Third, with a higher risk aversion and $\operatorname{Cov}_t(u_{t+1}, \eta_{t+1}) < 0$, there is a positive hedging demand for the risky asset: it pays off (today) when the future investment opportunities are poor.

Example 18.4 *(Habit persistence) The habit persistence model of Campbell and Cochrane (1999) has a CRRA utility function, but the argument is the difference between consumption and a habit level, $C_t - X_t$, instead of just consumption. The habit is parameterised in terms of the "surplus ratio" $S_t = (C_t - X_t)/C_t$. The log surplus ratio.(s_t)is assumed to be a non-linear AR(1)*

$$s_t = \phi s_{t-1} + \lambda(s_{t-1})\Delta c_t.$$

It can be shown (see Söderlind (2006)) that if $\lambda(s_{t-1})$ is a constant λ and the excess return is unpredictable (by s_t) then the habit persistence model is virtually the same as the CRRA model, but with $\gamma(1+\lambda)$ as the "effective" risk aversion.

Example 18.5 *(Reaction to news and the autocorrelation of returns) Let the log asset price, p_t, be the sum of a random walk and a temporary component (with perfectly correlated innovations, to make things simple)*

$$\begin{aligned} p_t &= u_t + \theta\varepsilon_t, \text{ where } u_t = u_{t-1} + \varepsilon_t \\ &= u_{t-1} + (1+\theta)\varepsilon_t. \end{aligned}$$

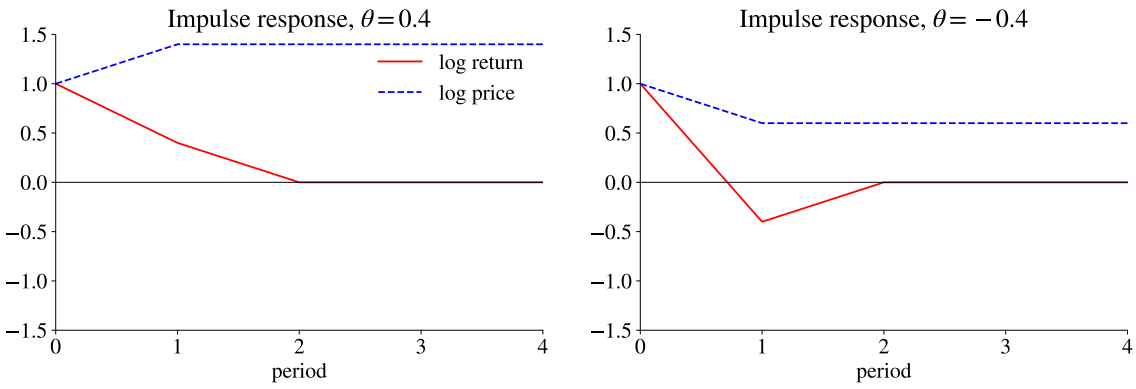
Let $r_t = p_t - p_{t-1}$ be the log return. It is straightforward to calculate that

$$\operatorname{Cov}(r_{t+1}, r_t) = -\theta(1+\theta)\operatorname{Var}(\varepsilon_t),$$

so $0 < \theta < 1$ (initial overreaction of the price) gives a negative autocorrelation. In short, mean reversion in the price level means negative autocorrelation of the returns—and vice versa. See Figure 18.1 for the impulse responses with respect to a piece of news, ε_t.

18.2 Autocorrelations

Reference: Campbell, Lo, and MacKinlay (1997) 2



Log return is an MA(1): $r_t = \varepsilon_t + \theta\varepsilon_{t-1}$
The log price level is: $p_t = r_0 + r_1 + \ldots + r_t$
The figure traces out the response to $\varepsilon_0 = 1$

Positive (negative) autocorrelation of the returns
⇔
Momentum (mean reversion) of the price

Figure 18.1: Impulse responses when price is random walk plus temporary component

18.2.1 Autocorrelation Coefficients and the Box-Pierce Test

The sampling properties of autocorrelations ($\hat{\rho}_s$) are complicated, but there are several useful large sample results for Gaussian processes (these results typically carry over to processes which are similar to the Gaussian—a homoskedastic process with finite 6th moment is typically enough, see Priestley (1981) 5.3 or Brockwell and Davis (1991) 7.2–7.3). When the true autocorrelations are all zero (not ρ_0, of course), then we have

$$\sqrt{T}\begin{bmatrix} \hat{\rho}_i \\ \hat{\rho}_j \end{bmatrix} \to^d N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right), \tag{18.3}$$

provided $(i, j) \neq 0$ and $i \neq j$. This result can be used to construct tests for both single autocorrelations (t-test or χ^2 test) and several autocorrelations at once (χ^2 test). To apply this on returns, the return horizon can be whatever (seconds, years,...), but it is important that the returns are non-overlapping (time aggregation can easily introduce spurious serial correlation).

Example 18.6 *(t-test) We want to test the hypothesis that* $\rho_1 = 0$. *Since the* $N(0, 1)$ *distribution has 5% of the probability mass below -1.64 and another 5% above 1.64, we can reject the null hypothesis at the 10% level if* $\sqrt{T}|\hat{\rho}_1| > 1.64$. *With* $T = 100$, *we therefore need* $|\hat{\rho}_1| > 1.64/\sqrt{100} = 0.165$ *for rejection, and with* $T = 1000$ *we need* $|\hat{\rho}_1| > 1.64/\sqrt{1000} \approx 0.053$.

The *Box-Pierce test* follows directly from the result in (18.3), since it shows that $\sqrt{T}\hat{\rho}_i$

and $\sqrt{T}\hat{\rho}_j$ are iid N(0,1) variables. Therefore, the sum of the square of them is distributed as an χ^2 variable. The test statistic typically used is

$$Q_L = T\sum_{s=1}^{L}\hat{\rho}_s^2 \to^d \chi_L^2. \tag{18.4}$$

Example 18.7 *(Box-Pierce) Let $\hat{\rho}_1 = 0.165$, and $T = 100$, so $Q_1 = 100 \times 0.165^2 = 2.72$. The 10% critical value of the χ_1^2 distribution is 2.71, so the null hypothesis of no autocorrelation is rejected.*

The choice of lag order in (18.4), L, should be guided by theoretical considerations, but it may also be wise to try different values. There is clearly a trade off: too few lags may miss a significant high-order autocorrelation, but too many lags can destroy the power of the test (as the test statistic is not affected much by increasing L, but the critical values increase).

Empirical Example 18.8 *(Autocorrelations for different lags, daily equity returns) See Figure 18.2 for autocorrelations (different lags) of S&P 500 returns. The figure suggests little autocorrelation in returns (R_t^e) , but considerable autocorrelation for the absolute value ($|R_t^e|$). Since, $R_t^e = \text{sign}(R_t^e)|R_t^e|$, this suggests that it is very difficult to predict the sign of the returns. Also, see Figure 18.3 for ten size-sorted equity portfolios which suggests that most size categories have more autocorrelations than large cap (which are fairly closer to S&P 500).*

The main problem with these tests is that the assumptions behind the results in (18.3) may not be reasonable. For instance, data may be heteroskedastic. One way of handling these issues is to make use of the GMM framework. Alternatively, a non-parametric test like the "runs test" can be used.

Remark 18.9 *(Runs test*) A "runs test" is a non-parametric test of randomness. Let d_t be an indicator variable*

$$d_t = \begin{cases} 0 & \text{if } y_t \leq q \\ 1 & \text{if } y_t > q \end{cases}$$

where q typically (but not necessarily) is the mean of y_t. Let $T_1 = \sum_{t=1}^{T} d_t$, that is the number of occasions when $y_t > q$, and $T_2 = T - T_1$ (the number of occasions when $y_t \leq q$). Also define the numbers of runs (r), that is, the number of changes in the d_t

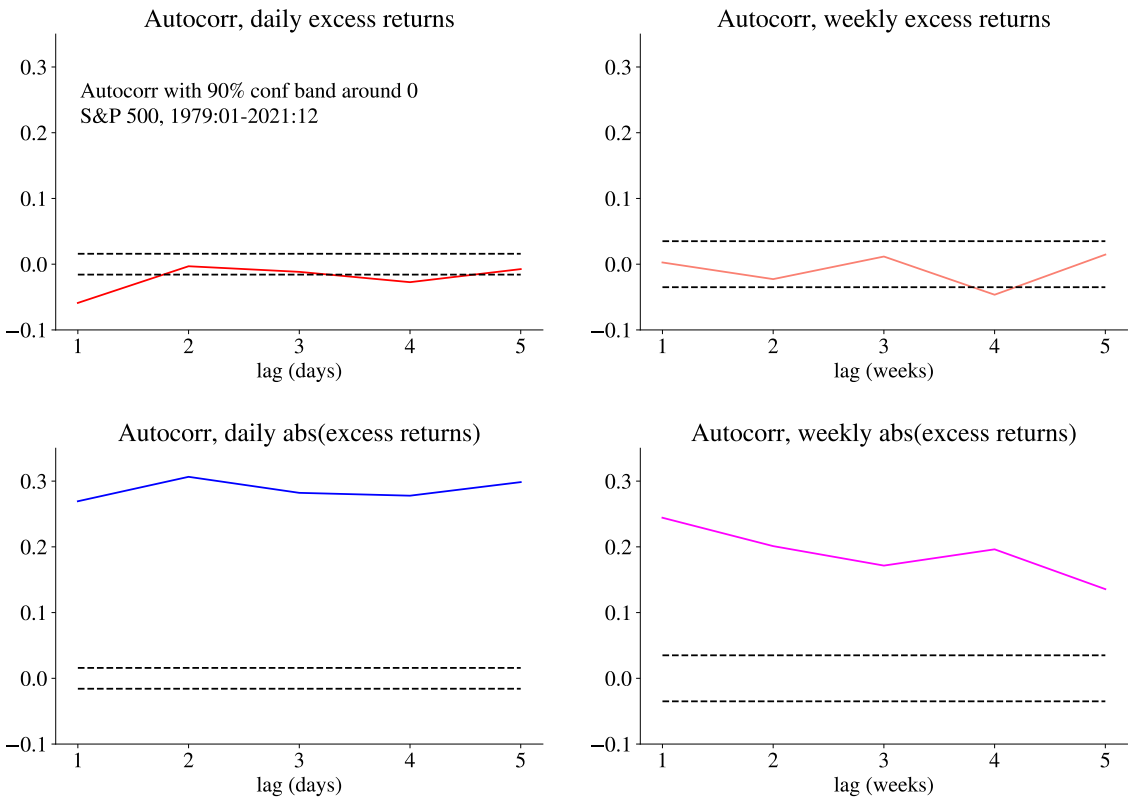


Figure 18.2: Predictability of US stock returns

series (where the first observation is counted as a change)

$$r = 1 + \sum_{t=2}^{T} |d_t - d_{t-1}|.$$

(Warning: r indicates "runs," not returns.) It is straightforward (but tedious) to show that, under the null hypothesis of randomness,

$$\mathrm{E}\, r = 2\frac{T_1 T_2}{T} + 1 \text{ and}$$
$$\mathrm{Var}(r) = \frac{(\mathrm{E}\, r - 1)(\mathrm{E}\, r - 2)}{T - 1}.$$

We can therefore test the null hypothesis of randomness by a t-stat

$$\frac{r - \mathrm{E}\, r}{\sqrt{\mathrm{Var}(r)}} \to^{d} N(0, 1).$$

The basic intuition of the test is that a positive autocorrelation would lead to too few runs

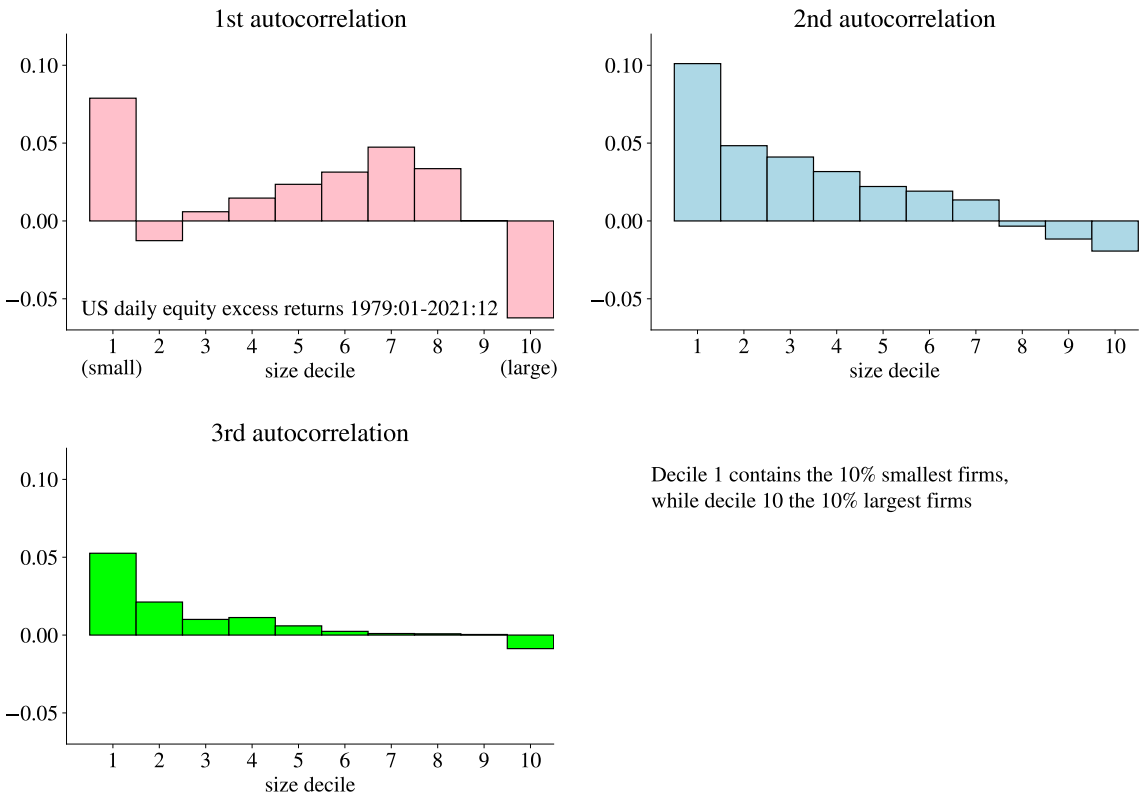


Figure 18.3: Predictability of US stock returns, size deciles

($r < \mathrm{E}\, r$): the y_t variable would stay on one side of the threshold q for long spells of time—and hence there would be few changes in x_t. Negative autocorrelation is just the opposite, since it tends to give a zigzag pattern around the mean. See Figure 18.4 for an example.

18.2.2 GMM Test of Autocorrelation*

This section discusses how GMM can be used to test if a series is autocorrelated. The analysis focuses on first-order autocorrelation, but it is straightforward to extend it to higher-order autocorrelation.

Consider a scalar random variable x_t with a zero mean (it is easy to extend the analysis

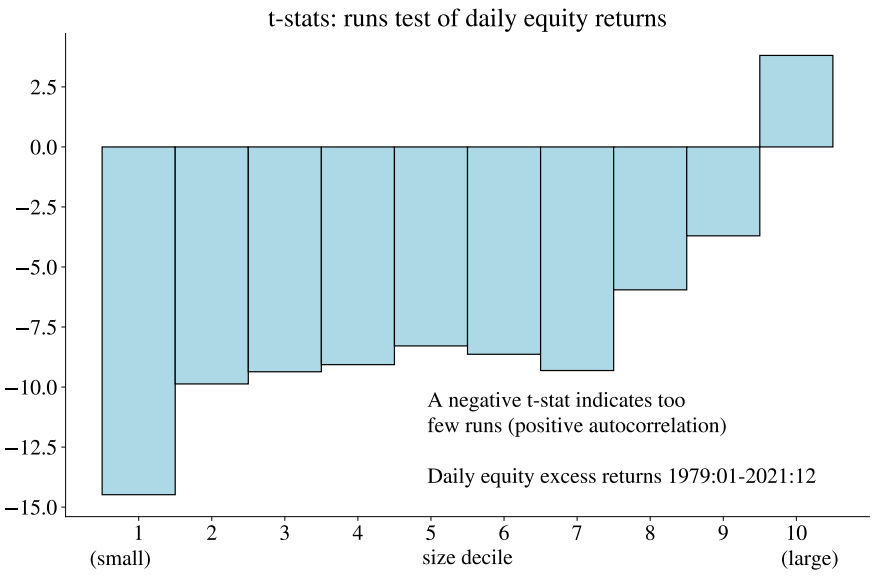


Figure 18.4: Runs test

to allow for a non-zero mean). Consider the moment conditions

$$g_t(\beta) = \begin{bmatrix} x_t^2 - \sigma^2 \\ x_t x_{t-1} - \rho\sigma^2 \end{bmatrix}, \text{ so } \bar{g}(\beta) = \frac{1}{T}\sum_{t=1}^{T} \begin{bmatrix} x_t^2 - \sigma^2 \\ x_t x_{t-1} - \rho\sigma^2 \end{bmatrix}, \text{ with } \beta = \begin{bmatrix} \sigma^2 \\ \rho \end{bmatrix}. \tag{18.5}$$

σ^2 is the variance and ρ the first-order autocorrelation so $\rho\sigma^2$ is the first-order autocovariance. We want to test if $\rho = 0$. We could proceed along two different routes: estimate ρ and test if it is different from zero or set ρ to zero and then test overidentifying restrictions.

We are able to arrive at simple expressions for these tests—provided we are willing to make strong assumptions about the data generating process. (These tests then typically coincide with classical tests like the Box-Pierce test.) One of the strong points of GMM is that we could perform similar tests without making strong assumptions—provided we use a correct estimator of the asymptotic covariance matrix of the moment conditions.

Remark 18.10 *(Box-Pierce as an Application of GMM) (18.5) is an exactly identified system so the weight matrix does not matter, so the asymptotic distribution is*

$$\sqrt{T}(\hat{\beta} - \beta_0) \xrightarrow{d} N(0, V), \text{ where } V = \left(D_0' S_0^{-1} D_0\right)^{-1}, V = D_0^{-1} S_0 (D_0^{-1})',$$

where D_0 is the Jacobian of the moment conditions and S_0 the covariance matrix of the

moment conditions (at the true parameter values). We have

$$D_0 = \operatorname{plim} \begin{bmatrix} \partial \bar{g}_1(\beta_0)/\partial\sigma^2 & \partial \bar{g}_1(\beta_0)/\partial\rho \\ \partial \bar{g}_2(\beta_0)/\partial\sigma^2 & \partial \bar{g}_2(\beta_0)/\partial\rho \end{bmatrix} = \begin{bmatrix} -1 & 0 \\ -\rho & -\sigma^2 \end{bmatrix} = \begin{bmatrix} -1 & 0 \\ 0 & -\sigma^2 \end{bmatrix},$$

since $\rho = 0$ (the true value). The definition of the covariance matrix is

$$S_0 = \mathrm{E} \left[\frac{\sqrt{T}}{T} \sum_{t=1}^{T} g_t(\beta_0) \right] \left[\frac{\sqrt{T}}{T} \sum_{t=1}^{T} g_t(\beta_0) \right]'.$$

Assume that there is no autocorrelation in $g_t(\beta_0)$ (which means, among other things, that volatility, x_t^2, is not autocorrelated). We can then simplify as

$$S_0 = \mathrm{E}\, g_t(\beta_0) g_t(\beta_0)'.$$

This assumption is stronger than assuming that $\rho = 0$, but we make it here in order to illustrate the asymptotic distribution. Moreover, assume that x_t is iid $N(0, \sigma^2)$. In this case (and with $\rho = 0$ imposed) we get

$$\begin{aligned} S_0 &= \mathrm{E} \begin{bmatrix} x_t^2 - \sigma^2 \\ x_t x_{t-1} \end{bmatrix} \begin{bmatrix} x_t^2 - \sigma^2 \\ x_t x_{t-1} \end{bmatrix}' = \mathrm{E} \begin{bmatrix} (x_t^2 - \sigma^2)^2 & (x_t^2 - \sigma^2) x_t x_{t-1} \\ (x_t^2 - \sigma^2) x_t x_{t-1} & (x_t x_{t-1})^2 \end{bmatrix} \\ &= \begin{bmatrix} \mathrm{E}\, x_t^4 - 2\sigma^2 \mathrm{E}\, x_t^2 + \sigma^4 & 0 \\ 0 & \mathrm{E}\, x_t^2 x_{t-1}^2 \end{bmatrix} = \begin{bmatrix} 2\sigma^4 & 0 \\ 0 & \sigma^4 \end{bmatrix}. \end{aligned}$$

To make the simplification in the second line we use the facts that $\mathrm{E}\, x_t^4 = 3\sigma^4$ if $x_t \sim N(0, \sigma^2)$, and that the normality and the iid properties of x_t together imply $\mathrm{E}\, x_t^2 x_{t-1}^2 = \mathrm{E}\, x_t^2 \mathrm{E}\, x_{t-1}^2$ and $\mathrm{E}\, x_t^3 x_{t-1} = \mathrm{E}\, \sigma^2 x_t x_{t-1} = 0$. Combining gives

$$\begin{aligned} \operatorname{Cov} \left(\sqrt{T} \begin{bmatrix} \hat{\sigma}^2 \\ \hat{\rho} \end{bmatrix} \right) &= D_0^{-1} S_0 (D_0^{-1})' \\ &= \begin{bmatrix} -1 & 0 \\ 0 & -1/\sigma^2 \end{bmatrix} \begin{bmatrix} 2\sigma^4 & 0 \\ 0 & \sigma^4 \end{bmatrix} \begin{bmatrix} -1 & 0 \\ 0 & -1/\sigma^2 \end{bmatrix} \\ &= \begin{bmatrix} 2\sigma^4 & 0 \\ 0 & 1 \end{bmatrix}. \end{aligned}$$

This shows that $\sqrt{T}\hat{\rho} \to^d N(0, 1)$.

18.2.3 Autoregressions

An alternative way of testing autocorrelations is to estimate an AR model

$$R_t = c + a_1 R_{t-1} + a_2 R_{t-2} + ... + a_p R_{t-p} + \varepsilon_t, \tag{18.6}$$

and then test if all the slope coefficients are zero with a χ^2 test. The return horizon can be whatever (seconds, years,...), but it is important that the returns are non-overlapping.

This approach is somewhat less general than the Box-Pierce test, but most stationary time series processes can be well approximated by an AR of relatively low order. To account for heteroskedasticity and other problems, we can estimate the covariance matrix of the parameters by an estimator like Newey-West. It can be noticed that when $R_t = c + aR_{t-1} + \varepsilon_t$, then a equals the first autocorrelation coefficient.

The autoregression can easily allow for the coefficients to depend on the market situation. For instance, consider an AR(1), but where the autoregression coefficient may be different depending on the sign of last period's return

$$R_t = \alpha + \beta Q_{t-1} R_{t-1} + \gamma(1 - Q_{t-1}) R_{t-1} + \varepsilon_t, \text{ where} \tag{18.7}$$

$$Q_{t-1} = \begin{cases} 1 \text{ if } R_{t-1} < 0 \\ 0 \text{ else.} \end{cases}$$

Empirical Example 18.11 *(AR(1) and asymmetric AR(1) for daily S&P 500 returns) See Figure 18.5.*

Autoregressions have also been used to study the predictability of long-run returns.

Empirical Example 18.12 *(AR(1) for long run equity returns) See Figure 18.6 for AR(1) results for different (long) investment horizons.*

Remark 18.13 *(Pitfall I in testing long-run returns) Let the return in (18.6) be a two period return, $r_t = \tilde{r}_t + \tilde{r}_{t-1}$, where $\tilde{r}_t$ is a one-period (log) return. An AR(1) on overlapping data would then be*

$$\tilde{r}_t + \tilde{r}_{t-1} = c + a(\tilde{r}_{t-1} + \tilde{r}_{t-2}) + \varepsilon_t.$$

Even if the one-period returns are uncorrelated, a would tend to be positive and significant—since $\tilde{r}_{t-1}$ shows up on both the left and right hand sides: the returns are overlapping. Instead, the correct specification is

$$\tilde{r}_t + \tilde{r}_{t-1} = c + a(\tilde{r}_{t-2} + \tilde{r}_{t-3}) + \varepsilon_t.$$

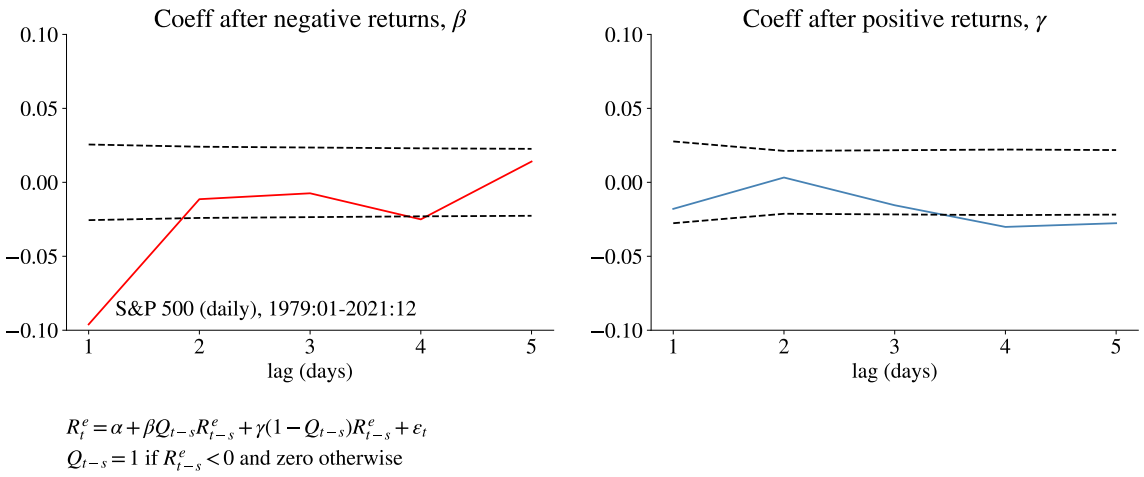


Figure 18.5: Predictability of US stock returns, results from a regression with interactive dummies

Remark 18.14 *(Pitfall 2 in testing long-run returns) A less serious pitfall is to use all available returns on the left hand side, for instance, all daily two-day returns. Two successive observations are then*

$$\tilde{r}_t + \tilde{r}_{t-1} = c + a(\tilde{r}_{t-2} + \tilde{r}_{t-3}) + \varepsilon_t$$
$$\tilde{r}_{t+1} + \tilde{r}_t = c + a(\tilde{r}_{t-1} + \tilde{r}_{t-2}) + \varepsilon_{t+1}$$

There is no problem with the point estimate of a, since the left and right hand side returns do not overlap. However, the residuals (ε_t and ε_{t+1}) are likely to be correlated which has to be handled in order to make correct inference. To see this, suppose $\tilde{r}_t = c/2 + u_t$ where u_t is iid. Clearly, the left and right hand sides are uncorrelated, so $a = 0$. With this we have

$$\tilde{r}_t + \tilde{r}_{t-1} = c + \varepsilon_t, \text{ where } \varepsilon_t = u_t + u_{t-1}$$
$$\tilde{r}_{t+1} + \tilde{r}_t = c + \varepsilon_{t+1}, \text{ where } \varepsilon_{t+1} = u_{t+1} + u_t.$$

Since u_t shows up in both ε_t and ε_{t+1}, the latter are correlated. See Figure 18.7. This can be solved by using a Newey-West approach (or something similar), or by skipping every second observation (there is then no overlap of the residuals).

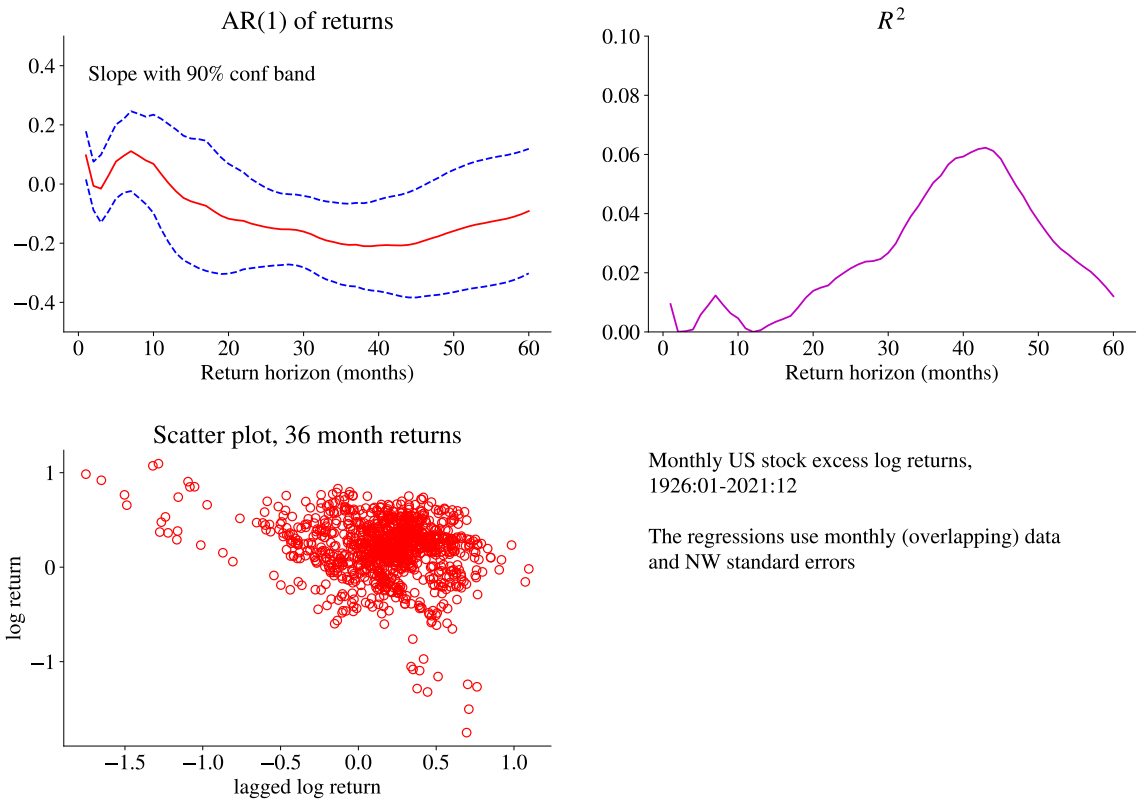


Figure 18.6: Predictability of US stock returns

18.3 Multivariate (Auto-)correlations

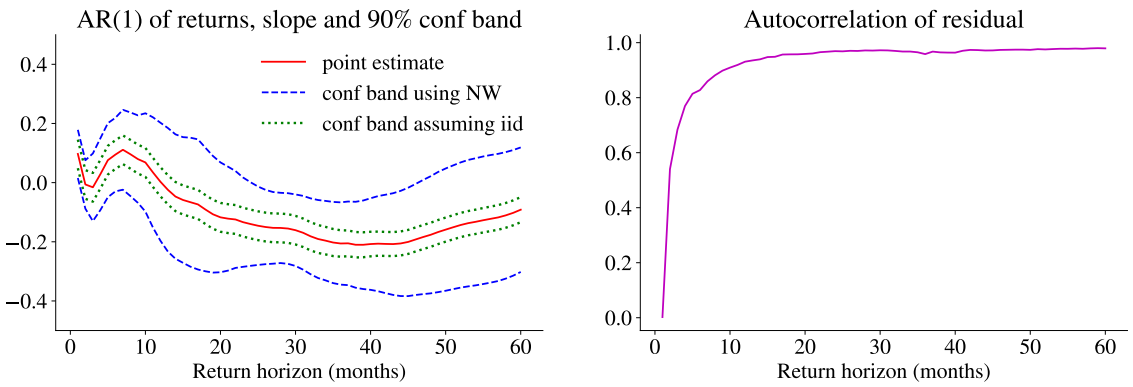
There is no reason to restrict the prediction model to only use the lagged returns of the same asset.

Empirical Example 18.15 *(Augmented AR(1) regressions) Figure 18.8 shows results from augmented AR(1) estimations for each of the ten size-sorted equity portfolios: the lagged return of the largest firms (decile 10) is added as a regressor.*

18.3.1 Momentum or Contrarian Strategy?

Reference: Lo and MacKinlay (1990)

A momentum strategy invests in assets that have performed well recently—and often goes short in those that have underperformed. The performance is driven by both autocorrelation and spill-over effects from other assets.



Slope with two different 90% conf bands (assuming iid or using NW)

Monthly US stock excess returns 1926:01-2021:12, overlapping data

Figure 18.7: Slope coefficient, LS vs Newey-West standard errors

Empirical Example 18.16 *(Momentum for daily returns on the 25 FF portfolios) Figure 18.9 suggests that there is considerable momentum in the cross-section of the 25 FF portfolios. Investing in past winners earns high returns.*

To disentangle the drivers of the return on a dynamic strategy, let there be N assets with returns R, with means and a cross autocovariance matrix

$$\begin{aligned} \mathrm{E}\, R &= \mu \text{ and} \\ \Gamma(k) &= \mathrm{E}[(R_t - \mu)(\tilde{R}_{t-k} - \mu)'], \end{aligned} \tag{18.8}$$

where $\tilde{R}_{t-k}$ can be the returns in $t-k$ (so $\Gamma(k)$ is an autocovariance matrix) or instead the (time series) average returns over a period ending in $t-k$ (for instance, a moving average over 22 trading days).

Empirical Example 18.17 *(Correlations of $R_{i,t}$ and $R_{j,t-s}$) See Figure 18.10 for cross-autocovariances of the daily 25 FF portfolios. For instance, cell (i,j) shows the correlation of $R_{i,t}$ and $R_{j,t-1}$.*

Example 18.18 *($\Gamma(k)$ with two assets) We have*

$$\Gamma(k) = \begin{bmatrix} \mathrm{Cov}(R_{1,t}, \tilde{R}_{1,t-k}) & \mathrm{Cov}(R_{1,t}, \tilde{R}_{2,t-k}) \\ \mathrm{Cov}(R_{2,t}, \tilde{R}_{1,t-k}) & \mathrm{Cov}(R_{2,t}, \tilde{R}_{2,t-k}) \end{bmatrix}.$$

When $\tilde{R}_{t-k} = \tilde{R}_{t-k}$, then this is the autocovariance matrix for lag 1.

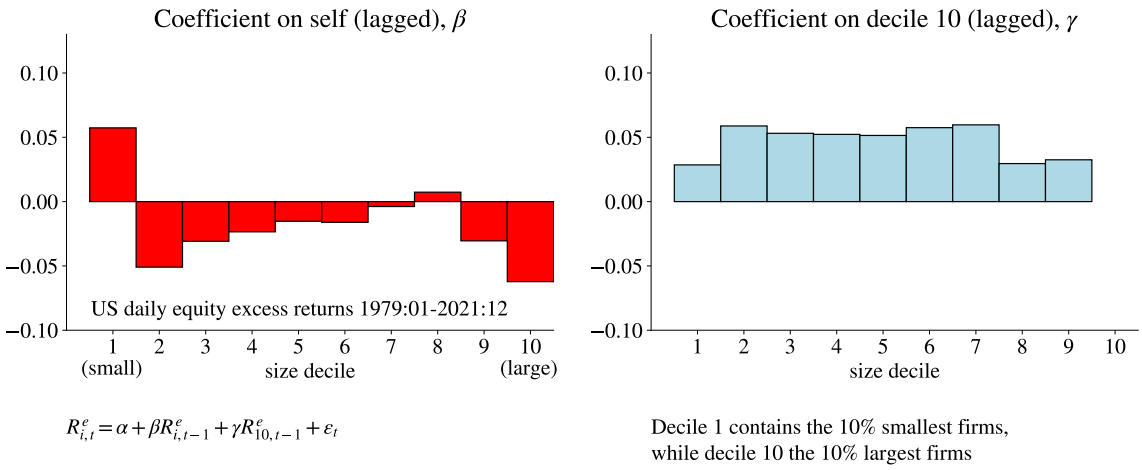


Figure 18.8: Coefficients from multiple prediction regressions

Define the equal weighted market portfolio return as

$$R_{mt} = \frac{1}{N}\sum_{i=1}^{N} R_{it}, \tag{18.9}$$

with the corresponding mean return $\mu_m = \mathrm{E}\, R_{mt}$.

A *momentum strategy* could (for instance) use the $N \times 1$ vector or portfolio weights

$$w_t(k) = \frac{\tilde{R}_{t-k} - \tilde{R}_{mt-k}}{N}, \tag{18.10}$$

which says that $w_{it}(k)$ is positive for assets with a return above (the cross-sectional) average return k periods back. (To analyse a contrarian strategy, reverse the sign of (18.10).) Notice that the portfolio weights depend on $\tilde{R}_{t-k} - \tilde{R}_{mt-k}$, which can be just the returns in $t-k$ or perhaps a moving average of returns for a period ending in $t-k$. For instance, with daily data the weights for day t may depend on the returns over the last month.

Notice that the weights sum to zero, so this is a zero cost portfolio. However, the weights differ from fixed weights (which would, for instance, be to put $1/5$ into the best 5 assets, and $-1/5$ into the 5 worst assets) since the overall size of the exposure ($1'|w_t|$) changes over time. A large dispersion of the past returns means large positions and vice versa.

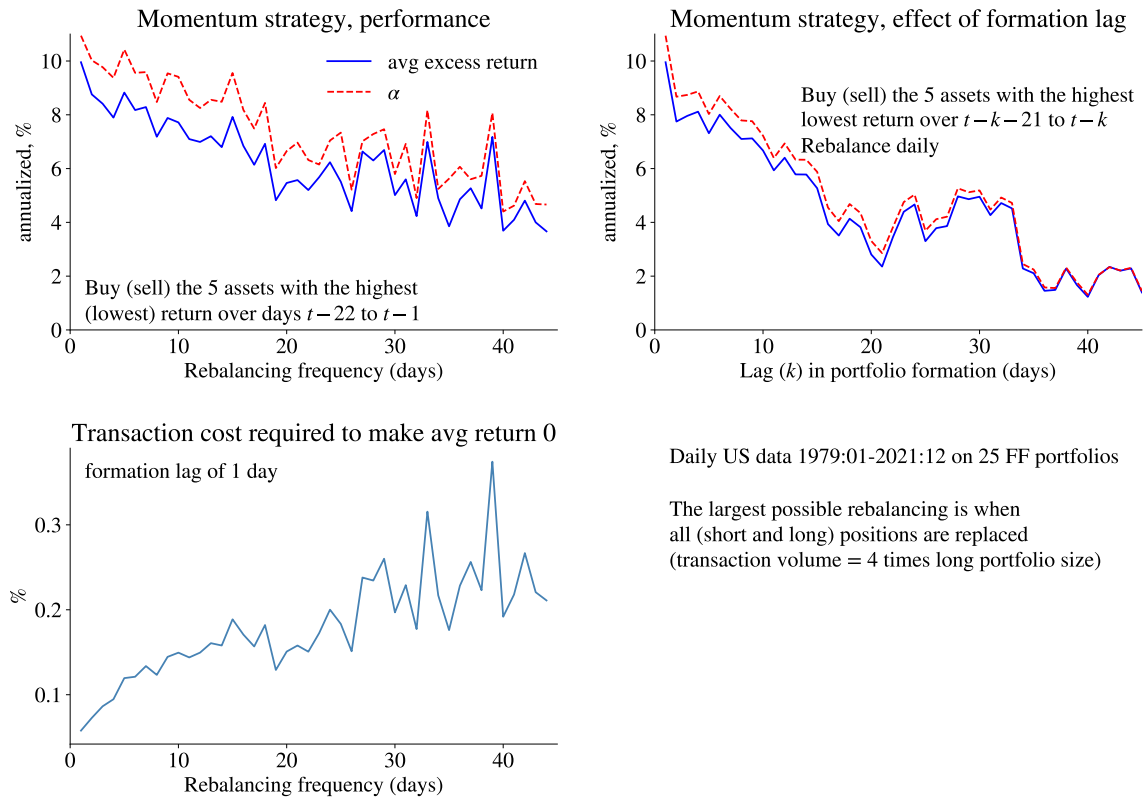


Figure 18.9: Performance of momentum investing

The profit from this strategy is

$$\pi_t(k) = \sum_{i=1}^{N} \underbrace{\frac{\tilde{R}_{it-k} - \tilde{R}_{mt-k}}{N}}_{w_{it}} R_{it} = \sum_{i=1}^{N} \frac{\tilde{R}_{it-k} R_{it}}{N} - \tilde{R}_{mt-k} R_{mt}, \tag{18.11}$$

where the last term uses the fact that $\Sigma_{i=1}^{N} \tilde{R}_{mt-k} R_{it}/N = \tilde{R}_{mt-k} R_{mt}$.

The expected profit is

$$\mathrm{E}\,\pi_t(k) = \frac{N-1}{N^2} \operatorname{tr} \Gamma(k) - \frac{1}{N^2} \left[\mathbf{1}'\Gamma(k)\mathbf{1} - \operatorname{tr} \Gamma(k)\right] + \frac{1}{N} \sum_{i=1}^{N} (\mu_i - \mu_m)^2, \tag{18.12}$$

where the $\mathbf{1}'\Gamma(k)\mathbf{1}$ sums all the elements of $\Gamma(k)$ and $\operatorname{tr} \Gamma(k)$ sums the elements along the main diagonal. (See below for a proof.)

With a random walk, $\Gamma(k) = 0$, (18.12) shows that the momentum strategy wins money: the first two terms are zero, while the third term contributes to a positive performance. The reason is that the momentum strategy (on average) invests in assets with high

average returns ($\mu_i > \mu_m$).

The *first term* of (18.12) depends only on own autocovariances, that is, how a return is correlated with the lagged return of the same asset. If these own autocovariances are (on average) positive, then a strongly performing asset in $t-k$ tends to perform well in t, which helps a momentum strategy (as the strongly performing asset is overweighted).

Notice that the *second term* of (18.12) sums all elements in the autocovariance matrix and then subtracts the sum of the diagonal elements—so it only depends on the sum of the cross-covariances, that is, how a return is correlated with the lagged return of other assets. In general, negative cross-covariances benefit a momentum strategy. To see why, consider the case with only two assets and suppose we observe a higher lagged return on asset 1 than on asset 2. If this predicts a low return on asset 2 (since $\text{Cov}(R_{2,t}, R_{1,t-k}) < 0$), but asset 2 does not predict asset 1 (since $\text{Cov}(R_{1,t}, R_{2,t-k}) = 0$), then the momentum strategy will profit. The reason is that we have a negative portfolio weight of asset 2 (since it performed relatively worse than asset 1 in the previous period).

Empirical Example 18.19 *(Decomposing momentum profits, daily returns on the 25 FF portfolios) See Tables 18.1 and 18.2.*

Example 18.20 *((18.12) with 2 assets) With*

$$\Gamma(k) = \begin{bmatrix} 0.1 & 0 \\ 0 & 0.1 \end{bmatrix},$$

then

$$\frac{N-1}{N^2}\operatorname{tr}\Gamma(k) = \frac{2-1}{2^2} \times (0.1+0.1) = 0.05, \text{ and}$$
$$-\frac{1}{N^2}\left[\mathbf{1}'\Gamma(k)\mathbf{1} - \operatorname{tr}\Gamma(k)\right] = -\frac{1}{2^2}(0.2-0.2) = 0$$

so the sum of the first two terms of (18.12) is positive (good for a momentum strategy).

Example 18.21 *((18.12) with 2 assets) Suppose we have*

$$\Gamma(k) = \begin{bmatrix} \text{Cov}(R_{1,t}, R_{1,t-k}) & \text{Cov}(R_{1,t}, R_{2,t-k}) \\ \text{Cov}(R_{2,t}, R_{1,t-k}) & \text{Cov}(R_{2,t}, R_{2,t-k}) \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ -0.1 & 0 \end{bmatrix}.$$

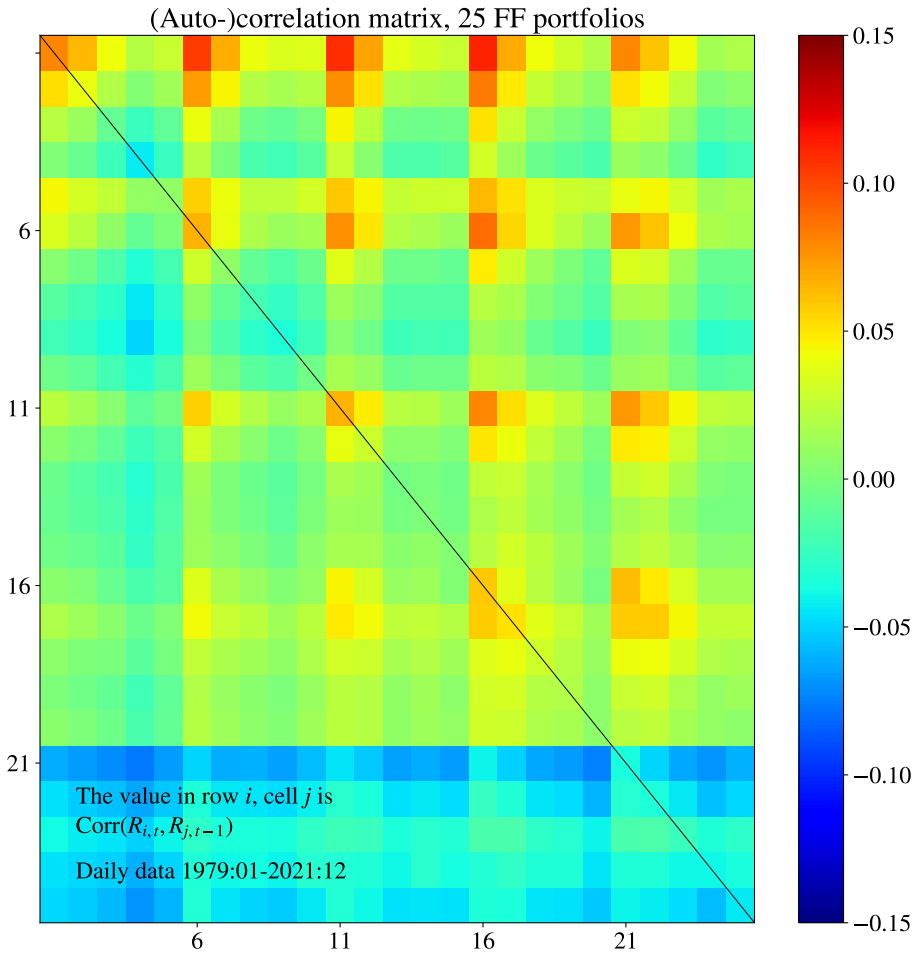


Figure 18.10: Illustration of the cross-autocorrelations, $\text{Corr}(R_t, R_{t-k})$, daily FF data. Dark colours indicate high correlations, light colours indicate low correlations.

Then

$$\frac{N-1}{N^2}\operatorname{tr}\Gamma(k) = \frac{2-1}{2^2} \times 0 = 0, \textit{ and}$$
$$-\frac{1}{N^2}\left[\mathbf{1}'\Gamma(k)\mathbf{1} - \operatorname{tr}\Gamma(k)\right] = -\frac{1}{2^2}\left[-0.1 - 0\right] = 0.025,$$

so the sum of the first two terms of (18.12) is positive (good for a momentum strategy). For instance, suppose $R_{1,t-k} > 0$, then $R_{2,t}$ tends to be low which is good (we have a negative portfolio weight on asset 2).

Proof. (of (18.12)) Take expectations of (18.11) and use the fact that $\text{E}\, xy = \text{Cov}(x, y) +$

	Portfolio 1	Portfolio 2	Portfolio 3
1	0.49	12.18	8.84
2	0.37	9.09	6.87
3	0.38	9.41	7.31
4	0.36	8.77	6.61
5	0.34	8.44	6.36

Table 18.1: Returns on different momentum portfolios, annualized %. The rows are for different formation lags (days). Portfolio 1 follows Lo and MacKinlay (1990), except that the portfolio weights depend on the average return over the previous month. Portfolio 2 applies a static scaling of the portfolio weights to get an average long (short) exposure of 1. Portfolio 3 instead scales the weights in each period. Daily US data 1979:01-2021:12 on 25 FF portfolios.

	auto cov	Cross cov	means	sum
1	1.14	−0.67	0.02	0.49
2	1.07	−0.72	0.02	0.37
3	0.83	−0.46	0.02	0.38
4	0.60	−0.26	0.02	0.36
5	0.62	−0.30	0.02	0.34

Table 18.2: Contributions to the average returns on a momentum portfolio, annualized %. The rows are for different formation lags (days). The strategy follows Lo and MacKinlay (1990), except that the portfolio weights depend on the average return over the previous month. Daily US data 1979:01-2021:12 on 25 FF portfolios.

$\mathrm{E}\, x\, \mathrm{E}\, y$ to get

$$\mathrm{E}\,\pi_t(k) = \frac{1}{N}\sum_{i=1}^{N}\left[\mathrm{Cov}(R_{it}, \tilde{R}_{it-k}) + \mu_i^2\right] - \left[\mathrm{Cov}(R_{mt}, \tilde{R}_{mt-k}) + \mu_m^2\right].$$

(using the fact that $\mathrm{E}\,\tilde{R}_{it-k} = \mathrm{E}\,R_{it}$ and $\mathrm{E}\,\tilde{R}_{mt-k} = \mathrm{E}\,R_{mt}$). Define the $N \times N$ cross-covariance matrix $\Gamma(k) = \mathrm{Cov}(R_t, \tilde{R}_{t-k})$ and recall that $R_{mt} = \mathbf{1}'R_t/N$ (and $\tilde{R}_{mt} = \mathbf{1}'\tilde{R}_t/N$). We can then rewrite the terms as

$$\begin{aligned}
\frac{1}{N}\sum_{i=1}^{N}\mathrm{Cov}(R_{it}, \tilde{R}_{it-k}) &= \frac{1}{N}\,\mathrm{tr}\,\Gamma(k) \\
\mathrm{Cov}(R_{mt}, \tilde{R}_{mt-k}) &= \mathbf{1}'\Gamma(k)\mathbf{1}/N^2 \\
\frac{1}{N}\sum_{i=1}^{N}\mu_i^2 - \mu_m^2 &= \frac{1}{N}\sum_{i=1}^{N}(\mu_i - \mu_m)^2.
\end{aligned}$$

Combine to rewrite $\mathrm{E}\,\pi_t$ as

$$\mathrm{E}\,\pi_t(k) = \frac{1}{N}\operatorname{tr}\Gamma(k) - \frac{1}{N^2}\mathbf{1}'\Gamma(k)\mathbf{1} + \frac{1}{N}\sum_{i=1}^{N}(\mu_i - \mu_m)^2,$$

which can be rearranged as (18.12). ■

18.4 Other Predictors

There are many other, perhaps more economically plausible, possible predictors of future stock returns. For instance, both the dividend-price ratio and nominal interest rates have been used to predict long-run returns, and short-run returns on other assets have been used to predict short-run returns.

18.4.1 Prices and Dividends

Reference: Campbell and Shiller (1988), Campbell, Lo, and MacKinlay (1997) 7 and Cochrane (2005) 20.1.

Recall that the asset price P_t, gross return R_{t+1} and dividends are related according to

$$P_t = \frac{D_{t+1} + P_{t+1}}{R_{t+1}}. \tag{18.13}$$

(This is an identity, since it defines the gross return.) Recursively solving this equation forward gives an expression of the price (or price/dividend ratio) is terms of the present value of future dividends, where the discounting is made by the actual returns. (This is also an identity.) See Appendix for details. We now log-linearise this present value expression in order to tie it more closely to the (typically linear) econometrics methods for detecting predictability The result is

$$p_t - d_t \approx \sum_{s=0}^{\infty}\rho^s[(d_{t+1+s} - d_{t+s}) - \tilde{r}_{t+1+s}], \tag{18.14}$$

where p_t is the log price, d_t the log dividend and $\tilde{r}_{t+1+s}$ is a one-period log return. Also, $\rho = 1/(1 + \overline{D/P})$ where $\overline{D/P}$ is a steady state dividend-price ratio ($\rho = 1/1.04 \approx 0.96$ if $\overline{D/P}$ is 4%) and where See Appendix for details. Clearly, a high price-dividend ratio must imply future dividend growth and/or low future returns.

One of the most successful attempts to forecast long-run return is by using the dividend-

price ratio

$$r_{t+q} = \alpha + \beta_q(d_t - p_t) + \varepsilon_{t+q}, \tag{18.15}$$

where r_{t+q} is the log return between t and $t+q$. For instance, CLM Table 7.1, report R^2 values from this regression which are close to zero for monthly returns, but they increase to 0.4 for 4-year returns (US, value weighted index, mid 1920s to mid 1990s).

Empirical Example 18.22 *(Predicting long run equity returns with E/P) See Figure 18.11.*

By comparing with (18.14), we see that the dividend-ratio in (18.15) is only asked to predict a finite (unweighted) sum of future returns and not dividend growth. We should therefore expect (18.15) to work particularly well if the horizon is long (high q) and if dividends are stable over time, which seems to be the case.

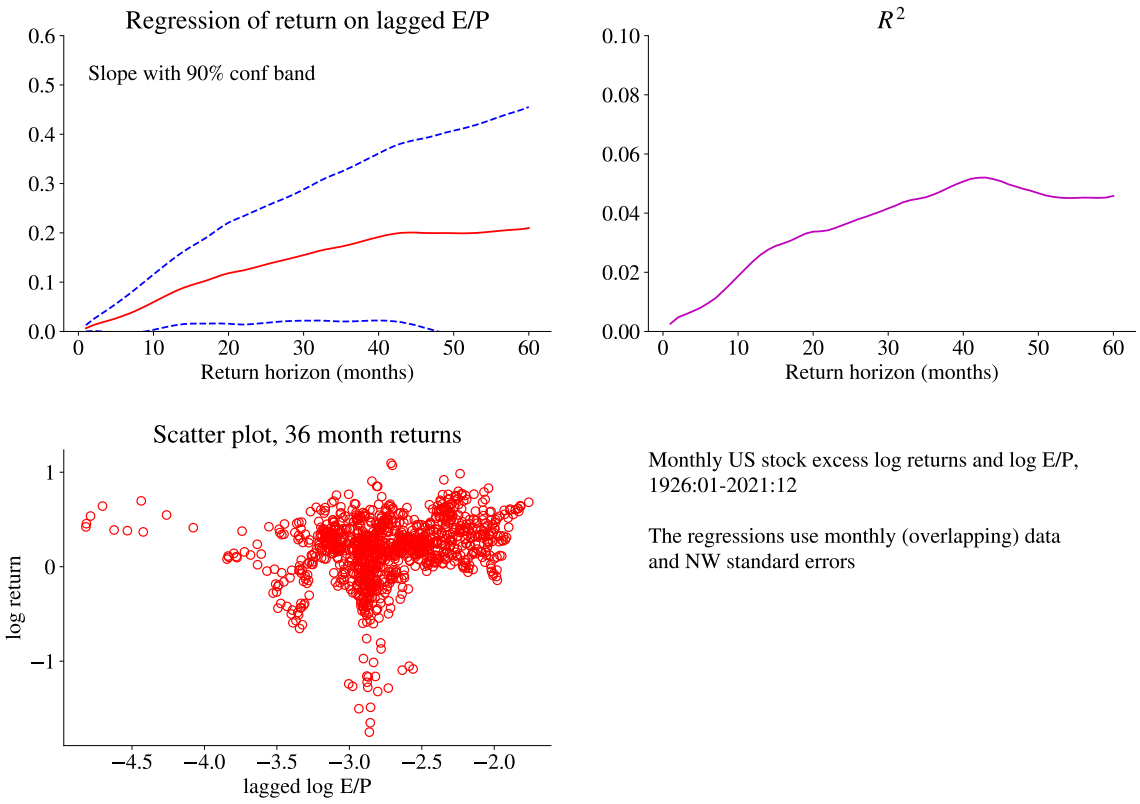


Figure 18.11: Predictability of US stock returns

18.4.2 Predictability but No Autocorrelation

The evidence for US stock returns is that long-run returns may perhaps be predicted by using dividend-price ratio or interest rates, but that the long-run autocorrelations are weak (long run US stock returns appear to be "weak-form efficient" but not "semi-strong efficient"). Both CLM 7.1.4 and Cochrane 20.1 use small models for discussing this case. The key in these discussions is to make changes in dividends unforecastable, but let the return be forecastable by some state variable ($E_t\, d_{t+1+s} - E_t\, d_{t+s} = 0$ and $E_t\, r_{t+1} = r + x_t$), but in such a way that there is little autocorrelation in returns. By taking expectations of (18.14) we see that price-dividend will then reflect expected future returns and therefore be useful for forecasting.

18.5 Spurious Regressions and In-Sample Overfitting

References: Ferson, Sarkissian, and Simin (2003)

18.5.1 Spurious Regressions

Ferson, Sarkissian, and Simin (2003) argue that many prediction equations suffer from "spurious regression" features—and that data mining tends to make things even worse.

Their simulation experiment is based on a simple model where the return predictions are

$$R_{t+1} = \alpha + \delta Z_t + v_{t+1}, \tag{18.16}$$

where Z_t is a regressor (predictor). The true model is that returns follow the process

$$R_{t+1} = \mu + Z_t^* + u_{t+1}, \tag{18.17}$$

where the residual is white noise. In this equation, Z_t^* represents movements in expected returns. The predictors follow a diagonal VAR(1)

$$\begin{bmatrix} Z_t \\ Z_t^* \end{bmatrix} = \begin{bmatrix} \rho & 0 \\ 0 & \rho^* \end{bmatrix} \begin{bmatrix} Z_{t-1} \\ Z_{t-1}^* \end{bmatrix} + \begin{bmatrix} \varepsilon_t \\ \varepsilon_t^* \end{bmatrix}, \text{ with } \operatorname{Cov}\left(\begin{bmatrix} \varepsilon_t \\ \varepsilon_t^* \end{bmatrix} \right) = \Sigma. \tag{18.18}$$

In the case of a "pure spurious regression," the innovations to the predictors are uncorrelated (Σ is diagonal). In this case, δ ought to be zero—and their simulations show that the estimates are almost unbiased. Instead, there is a problem with the standard deviation of $\hat{\delta}$. If ρ^* is high, then the returns will be autocorrelated. See Table 18.3 for an

illustration.

	$\kappa = 0.0$		$\kappa = 0.75$	
ρ :	0.0	0.75	0.0	0.75
Simulated	5.8	8.7	3.9	10.9
OLS formula	5.8	8.6	3.9	5.8
Newey-West	5.7	8.4	3.8	8.9
VARHAC	5.7	8.5	3.8	10.5
Bootstrapped	5.8	8.5	3.8	10.1
FGLS	5.8	4.7	3.9	5.9

Table 18.3: Standard error of OLS slope (%) under autocorrelation (simulation evidence). Model: $y_t = 1 + 0.9x_t + \epsilon_t$, where $\epsilon_t = \rho\epsilon_{t-1} + \xi_t$, ξ_t is iid N(). $x_t = \kappa x_{t-1} + \eta_t$, η_t is iid N(). NW uses 5 lags. VARHAC uses 5 lags and a VAR(1). The bootstrap uses blocks of size 20. Sample length: 300. Number of simulations: 25000.

Under the null hypothesis of $\delta = 0$, this autocorrelation is loaded onto the residuals. For that reason, the simulations use a Newey-West estimator of the covariance matrix (with an automatic choice of lag order). This should, ideally, solve the problem with the inference—but the simulations show that it doesn't: when Z_t^* is very autocorrelated (0.95 or higher) and reasonably important (so an R^2 from running (18.17), if we could, would be 0.05 or higher), then the 5% critical value (for a t-test of the hypothesis $\delta = 0$) would be 2.7 (to be compared with the nominal value of 1.96). Since the point estimates are almost unbiased, the interpretation is that the standard deviations are underestimated. In contrast, with low autocorrelation and/or low importance of Z_t^*, the standard deviations are much more in line with nominal values.

See *Table 18.3* for an illustration. The table shows that we need a combination of an autocorrelated residuals and an autocorrelated regressor to create a problem for the usual LS formula for the standard deviation of a slope coefficient. When the autocorrelation is very high, even the Newey-West estimator is likely to underestimate the true uncertainty.

To study the interaction between spurious regressions and data mining, Ferson, Sarkissian, and Simin (2003) let Z_t be chosen from a vector of L possible predictors—which all are generated by a diagonal VAR(1) system as in (18.18) with uncorrelated errors. It is assumed that the researchers choose Z_t by running L regressions, and then picks the one with the highest R^2. When $\rho^* = 0.15$ and the researcher chooses between $L = 10$ predictors, the simulated 5% critical value is 3.5. Since this does not depend on the importance of Z_t^*, it is interpreted as a typical feature of "data mining," which is bad enough.

When the autocorrelation is 0.95, then the importance of Z_t^* start to become important—"spurious regressions" interact with the data mining to create extremely high simulated critical values. A possible explanation is that the data mining exercise is likely to pick out the most autocorrelated predictor, and that a highly autocorrelated predictor exacerbates the spurious regression problem.

18.6 Model Selection

Selecting a good prediction model is often very different from constructing a model to test a theoretical hypothesis or to establish economic causality. In particular, theory plays a somewhat smaller role (just to help identifying a set of reasonable predictors) and there is a greater emphasis on having a small model. The focus on small models is driven by a considerable amount of evidence suggesting that large prediction models often perform poorly out-of-sample.

This section summarises some standard approaches to keeping the model small, while still providing a good in-sample fit. They can be applied to the full sample or data, or on recursive/moving data windows.

18.6.1 Traditional Model Selection

Remember that R^2 can never decrease by adding more regressors, so it is not really a good guide in selecting a model (unless you have already decided on the number of predictors, for instance, only one). To avoid overfitting, we "punish" models with too many parameters by using the adjusted R^2, defined as

$$\bar{R}^2 = 1 - (1 - R^2)\frac{T-1}{T-k}, \tag{18.19}$$

where T is the sample size and k is the number of regressors (including the constant). This measure includes trade-off between fit and the number of regressors (per data point). Notice that $\bar{R}^2$ can be negative (while $0 \leq R^2 \leq 1$). Clearly, the model must include a constant for R^2 (and therefore $\bar{R}^2$) to make sense. Alternatively, apply Akaike's Information Criterion (AIC) and the Bayesian information criterion (BIC). They are

$$AIC = \ln \sigma^2 + 2\frac{k}{T} \tag{18.20}$$

$$BIC = \ln \sigma^2 + \frac{k}{T}\ln T, \tag{18.21}$$

where σ^2 is the variance of the fitted residuals.

These measures also involve trade-offs between fit (low σ^2) and number of parameters (k, including the intercept). Choose the model with the *highest* $\bar{R}^2$ or *lowest* AIC or BIC. It can be shown (by using $R^2 = 1 - \sigma^2/\operatorname{Var}(y_t)$) that AIC and BIC can be rewritten as

$$AIC = \ln \operatorname{Var}(y_t) + \ln(1 - R^2) + 2\frac{k}{T} \tag{18.22}$$

$$BIC = \ln \operatorname{Var}(y_t) + \ln(1 - R^2) + \frac{k}{T} \ln T. \tag{18.23}$$

This shows that both are decreasing in R^2 (which is good), but increasing in the number of regressors per data point (k/T). It therefore leads to a similar trade-off as in $\bar{R}^2$. Recall that the model should always include a constant.

Empirical Example 18.23 *(Empirical application of model selection) See Table 18.4 for an empirical example showing a number of possible model specifications. The dependent variable is the monthly realized variance of S&P 500 returns (calculated from daily returns). The possible regressors are lags of the dependent variable, the VIX index and the S&P 500 returns. Similarly, Table 18.5 for the the best specification according to AIC. Notice that AIC tend to favour fairly large models with many regressors.*

18.6.2 Sequential Model Selection

Reference: Hastie, Tibshirani, and Friedman (2001) 3

If there are k potential regressors, then there are $2^k - 1$ different models. If the list of models is not too long, then we can try them all and use the AIC and BIC in (18.20)–(18.21), see Table 18.5. Otherwise, we need some type of sequential approach.

Example 18.24 *(3 potential regressors) If the three potential regressors are* 1, x_1 *and* x_2, *then the list of models has* $2^3 - 1 = 7$ *possibilities:* (1); (x_1); (x_2); $(1, x_1)$; $(1, x_2)$; (x_1, x_2); $(1, x_1, x_2)$.

A *forward stepwise selection* is as follows

(1) start with an intercept (18.24)

(2) add the variable that improves the fit the most

(3) repeat (2) until the fit does not improve much.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
RV_{t-1}	0.66						0.16
	(8.79)						(1.94)
RV_{t-2}		0.45					−0.01
		(5.60)					(−0.14)
VIX_{t-1}			0.95				1.08
			(10.33)				(3.84)
VIX_{t-2}				0.67			−0.38
				(8.74)			(−1.70)
R_{t-1}					−0.86		0.03
					(−3.41)		(0.16)
R_{t-2}						−0.49	−0.13
						(−2.29)	(−1.39)
constant	5.13	8.35	−2.97	2.44	15.98	15.67	−0.38
	(4.81)	(6.45)	(−1.98)	(1.85)	(17.47)	(16.85)	(−0.32)
R^2	0.44	0.21	0.53	0.26	0.15	0.05	0.56
$\bar{R}^2$	0.44	0.20	0.53	0.26	0.14	0.05	0.56
obs	362	362	362	362	362	362	362

Table 18.4: Regression of monthly realized S&P 500 return volatility 1990:02-2021:12. Numbers in parentheses are t-stats, based on Newey-West with 4 lags.

To specify a stopping rule, first define the residual sum of squares (for a given vector of coefficients, β) as

$$RSS(\beta) = \sum_{t=1}^{T}(y_t - x_t'\beta)^2. \tag{18.25}$$

In step (2) we would then add the variable that gives the lowest RSS (when added to the previous selection). In step (3), it is often recommended that we stop adding regressors when

$$\frac{RSS(\hat{\beta}_{\text{old}}) - RSS(\hat{\beta}_{\text{new}})}{RSS(\hat{\beta}_{\text{new}})/(T-k-1)} < c_{1,T-k-1}, \tag{18.26}$$

where k is the number of coefficients in $\hat{\beta}_{\text{old}}$ (including the intercept) so there are $k+1$ coefficients in $\hat{\beta}_{\text{new}}$ and $c_{1,T-k-1}$ is the 90% or 95% critical value of an $F_{1,T-k-1}$ distribution. For instance, the 90% critical value of $F_{1,100}$ equals 2.76.

As an alternative to the RSS based rule in (18.25)–(18.26), we could instead use t-stats: in step (2) add the variable with the highest |t-stat| and in step (3) stop adding variables when that |t-stat| is lower than 1.64 (or 1.96).

Empirical Example 18.25 *(Forward stepwise selection) Applying the forward step se-*

	(1)	(2)	(3)	(4)
RV_{t-1}	0.18			0.15
	(2.04)			(1.74)
RV_{t-2}				
VIX_{t-1}	1.02	1.20	1.22	1.06
	(5.81)	(6.60)	(6.76)	(5.71)
VIX_{t-2}	−0.34	−0.31	−0.37	−0.37
	(−2.50)	(−2.26)	(−2.50)	(−2.62)
R_{t-1}				
R_{t-2}			−0.18	−0.13
			(−1.95)	(−1.40)
constant	−0.67	−1.85	−0.98	−0.26
	(−0.72)	(−1.58)	(−1.05)	(−0.31)
R^2	0.56	0.55	0.56	0.56
BIC	3.80	3.80	3.80	3.81
obs	362	362	362	362

Table 18.5: Regression of monthly realized S&P 500 return volatility 1990:02-2021:12. Ordered from best (1) according to BIC to fourth best (4). Numbers in parentheses are t-stats, based on Newey-West with 4 lags.

lection approach (based on t-stats) to the regression discussed in Example 18.23 gives a sequence of larger and larger models shown in Table 18.6.

18.6.3 The Lasso Method*

An alternative approach to model selection is the *Lasso method*, which minimizes the sum of squared residuals (just like OLS), but with a penalty on $\Sigma_{i=1}^{K}|b_i|$,

$$\min_b \sum_{t=1}^{T}(y_t - \alpha - x_t'b)^2 + \gamma\sum_{i=1}^{K}|b_i|, \tag{18.27}$$

where the value of γ is chosen a priori, but where we typically consider different values of γ. Having the same penalty on all $|b_i|$ makes perhaps most sense when the regressors have the same scale (for instance, zero mean and unit standard deviation). The *adaptive Lasso* instead uses weighted penalties, $\gamma\Sigma_{i=}^{K}w_i|b_i|$, often with $w_i = 1/|b_i^{OLS}|$. This will clearly give a larger effective penalty on variables whose OLS coefficients are small (in absolute terms). It is sometimes an advantage to divide the first term in (18.27) by T,

	(1)	(2)	(3)	(4)
RV_{t-1}			0.18	0.15
			(2.04)	(1.74)
RV_{t-2}				
VIX_{t-1}	0.95	1.20	1.02	1.06
	(10.33)	(6.60)	(5.81)	(5.71)
VIX_{t-2}		−0.31	−0.34	−0.37
		(−2.26)	(−2.50)	(−2.62)
R_{t-1}				
R_{t-2}				−0.13
				(−1.40)
constant	−2.97	−1.85	−0.67	−0.26
	(−1.98)	(−1.58)	(−0.72)	(−0.31)
R^2	0.53	0.55	0.56	0.56
obs	362	362	362	362

Table 18.6: Best four regressions of monthly realized S&P 500 return volatility according to a forward step selection (based on t-stats), 1990:02-2021:12. Ordered from smallest model (1) to fourth smallest model (4). Numbers in parentheses are t-stats, based on Newey-West with 4 lags.

to make the interpretation of γ independent of the sample size (and sometimes also for numerical reasons).

Remark 18.26 *(Alternative formulation) The same problem can also be written as a constrained optimisation problem*

$$\min_b \sum_{t=1}^{T}(y_t - \alpha - x_t'b)^2 \text{ subject to } \sum_{i=1}^{K}|b_i| \leq t,$$

where a small t corresponds to a high γ.

Clearly, when $\gamma = 0$, then the lasso approach reproduces the OLS estimates. For larger values of γ, the lasso will give smaller coefficients: some b_i will be zero and others tend to be closer to zero than OLS would suggest (similar to other "shrinkage" methods like a ridge estimation).

The lasso method can be used as a model selection technique by estimating a sequence of models with with different γ values. With a sufficiently high γ, only one coefficient is

	(1)	(2)	(3)	(4)
RV_{t-1}			0.18	0.15
			(2.04)	(1.74)
RV_{t-2}				
VIX_{t-1}	0.95	1.20	1.02	1.06
	(10.33)	(6.60)	(5.81)	(5.71)
VIX_{t-2}		−0.31	−0.34	−0.37
		(−2.26)	(−2.50)	(−2.62)
R_{t-1}				
R_{t-2}				−0.13
				(−1.40)
constant	−2.97	−1.85	−0.67	−0.26
	(−1.98)	(−1.58)	(−0.72)	(−0.31)
R^2	0.53	0.55	0.56	0.56
$\sum \lvert b_i \rvert$	0.73	1.17	1.23	1.31
obs	362	362	362	362

Table 18.7: Best four regressions of monthly realized S&P 500 return volatility where the model are selected by lasso, but then estimated with OLS, 1990:02-2021:12. Ordered from smallest model (1) to fourth smallest model (4). Numbers in parentheses are t-stats, based on Newey-West with 4 lags. The $\sum |b_i|$ is for regression using standardized variables.

non-zero, for a somewhat lower γ value two coefficients are non-zero and so on. See Figure 18.12. Once the L (five, say) smallest specifications are found, we could re-estimate each of them with OLS. (This is the lars-OLS hybrid discussed in Efron, Hasti, Johnstone, and Tibshirani (2004).)

Empirical Example 18.27 *(Lasso regression) Applying the Lasso approach to the regression discussed in Example 18.23 gives a sequence of smaller and smaller models. Figure 18.12 shows how the coefficients of the normalised variables change as penalty parameter is increased. Re-estimating the four smallest of those models with OLS gives the results in Table 18.7.*

Remark 18.28 *(Ridge regression*) The ridge regression solves* $\min_b \sum_{t=1}^{T}(y_t - \alpha - x_t'b)^2 + \lambda\sum_{i=1}^{K} b_i^2$*,where* $\lambda > 0$*, so it forms a compromise between OLS and zero coefficients. This is easiest to see if* y_t *and* x_t *are demeaned so* $\alpha = 0$*. Then, the first order*

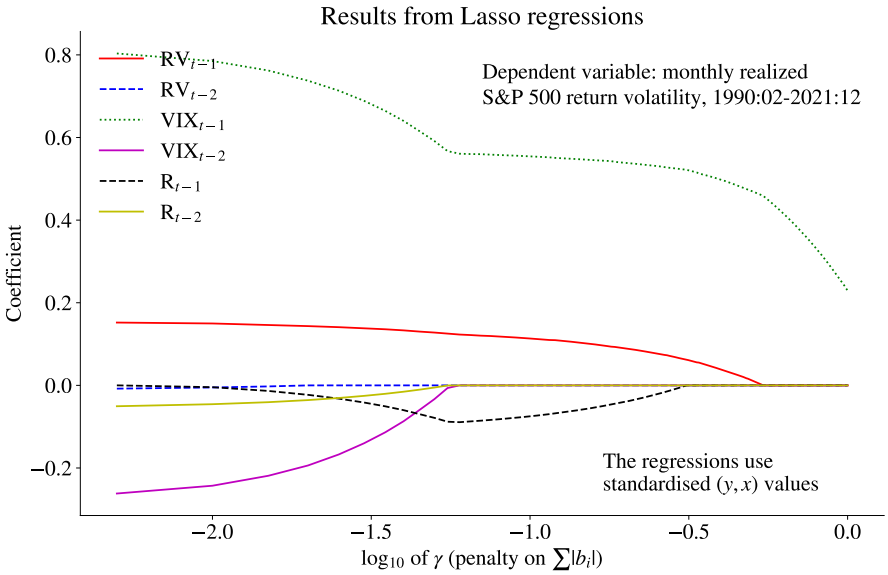


Figure 18.12: Lasso regressions

conditions for minimization are $\sum_{t=1}^{T} x_t(y_t - x_t'\tilde{b}) - \lambda\tilde{b} = 0$, *so* $\tilde{b} = (\sum_{t=1}^{T} x_t x_t' + \lambda I)^{-1}\sum_{t=1}^{T} x_t y_t$. *Notice that* $\lambda = 0$ *gives OLS, while* $\lambda = \infty$ *gives* $\tilde{b} = \mathbf{0}$.

Remark 18.29 *(Application of the lasso/lars algorithms) These algorithms often standardize* x_t *to have zero means and unit standard deviations, and* y_t *to have zero means (and perhaps unit standard deviation).*

Remark 18.30 *(Elastic net regression*) An elastic net regression is a mix of a Lasso regression and a ridge regression. It solves* $\min_b \sum_{t=1}^{T}(y_t - \alpha - x_t'b)^2 + \gamma\sum_{i=1}^{K}|b_i| + \lambda\sum_{i=1}^{K} b_i^2$.

18.7 Forecast Averaging

Reference: Elliot and Timmermann (2016) 14

Averaging across forecasts have often proved to be a good way of producing a superior forecast.

There are two main cases: (1) when we have access to the forecasts and also the data/model that produced them and (2) when we have access to the forecasts only. We discuss them in reverse order.

Suppose we have access to K different forecasts ($\hat{R}_t^i$ for $i = 1$ to K) of the return R_t. All these forecasts are made in period $t-h$ (with $h \geq 1$). We form a weighted average as

$$R_t^* = \sum_{i=1}^K w_i \hat{R}_t^i, \text{ with } \sum_{i=1}^K w_i = 1. \tag{18.28}$$

For instance, w be be chosen as to minimize the forecast error variance or the MSE over the sample up to and including $t-h$. In practice, it seems difficult to beat an unweighted average or an unweighted average after having pruned the most extreme forecasts ("trimmed mean").

Remark 18.31 *(*Minimising the MSE) Let Σ be the variance-covariance matrix of the forecast errors from K different models. If the forecasts are unbiased (so the forecast errors have zero means), then the MSE of a combined forecast is $w'\Sigma w$. Therefore, minimize $w'\Sigma w/2 + \lambda(1-\mathbf{1}'w)$ with respect to w to get the first order conditions $\Sigma w = \mathbf{1}\lambda$ and $1 = \mathbf{1}'w$, which together imply $w = \Sigma^{-1}\mathbf{1}/\mathbf{1}'\Sigma^{-1}\mathbf{1}$.*

Instead, suppose we have access also to the models and data that produces the various forecasts. It can then be argued that the proper way to proceed is to pool all the data and apply the model selection techniques. However, the unweighted average across forecasts often perform reasonably well.

Empirical Example 18.32 *(Forecast combination, out-of-sample evaluations) See Table 18.8.*

18.8 Out-of-Sample Forecasting Performance

References: Goyal and Welch (2008), and Campbell and Thompson (2008)

The idea of out-of-sample forecasting is to replicate real life forecasting. The prediction equation is estimated on data up to and including $t-1$, and then a forecast is made for period t. The forecasting performance of the equation is then compared to some benchmark prediction model like the historical average (also estimated on data up to and including $t-1$). See Figure 18.13 for an illustration. Then, the sample is extended with one period (t) and a forecast is made for $t+1$. This continues until the sample is exhausted.

Goyal and Welch (2008) find that the evidence of predictability of equity returns disappears when out-of-sample forecasts are considered.

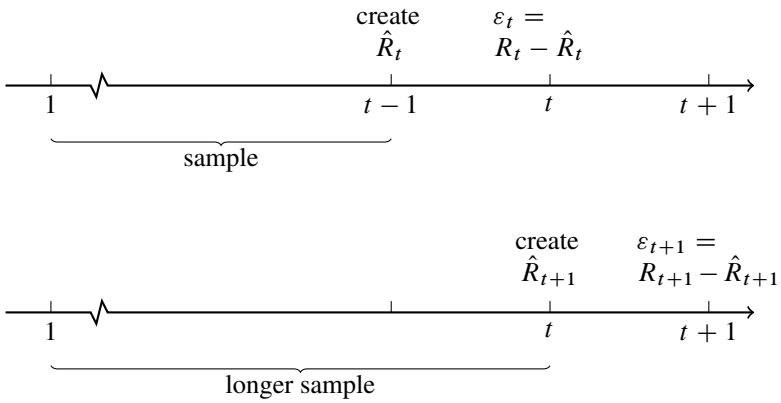


Figure 18.13: Out-of-sample forecasting

In contrast, Campbell and Thompson (2008) claim that there is still some out-of-sample predictability, provided we put restrictions on the estimated models. They first report that only few variables (earnings price ratio, T-bill rate and the inflation rate) have significant predictive power for one-month stock returns in the full sample (1871–2003 or early 1920s–2003, depending on predictor). The comparison is done in terms of the MSE and an "out-of-sample R^2"

$$R^2_{OS} = 1 - \sum_{t=s}^{T} \varepsilon_t^2 / \sum_{t=s}^{T} e_t^2, \tag{18.29}$$

where s is the first period with an out-of-sample forecast, $\varepsilon_t = R_t - \hat{R}_t$ is the forecast based on the prediction model (estimated on data up to and including $t-1$) and $e_t = R_t - \tilde{R}_t$ is the prediction from some benchmark model (also estimated on data up to and including $t-1$). The paper uses the historical average (also estimated on data up to and including $t-1$) as the benchmark prediction. The evidence shows that the out-of-sample forecasting performance is very weak—as claimed by Goyal and Welch (2008).

Campbell and Thompson (2008) argue that forecasting equations can easily give strange results when they are estimated on a small data set (as they are early in the sample). They therefore try different restrictions: setting the slope coefficient to zero whenever the sign is "wrong," setting the prediction (or the historical average) to zero whenever the value is negative. This improves the results a bit—although the predictive performance is still weak.

Empirical Example 18.33 *(Out-of-sample prediction of equity returns) Figure 18.14 shows results for daily size-sorted equity returns. There is some short-run predictability for*

small firm returns also out-of-sample. Figure 18.15 shows results on predicting long run equity returns with E/P. The evidence suggests that the in-sample long-run predictability vanishes out-of-sample.

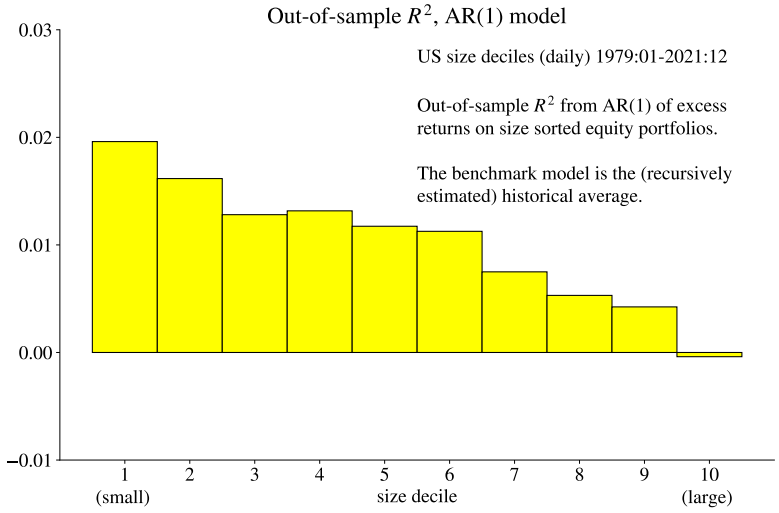


Figure 18.14: Short-run predictability of US stock returns, out-of-sample.

18.9 Evaluating Forecasting Performance

Further reading: Diebold (2001) 11; Stekler (1991); Diebold and Mariano (1995); Clark and West (2007)

To do a solid evaluation of the forecast performance (of some forecaster/forecast method/forecast institute), we need a sample (history) of the forecasts and the resulting forecast errors. The reason is that the forecasting performance for a single period is likely to be dominated by luck, so we can only expect to find systematic patterns by looking at results for several periods.

To set up tests of the forecasting performance, let ε_t be the forecast error in period t

$$\varepsilon_t = R_t - \hat{R}_t, \tag{18.30}$$

where $\hat{R}_t$ is the forecast (made in $t-h$) and R_t the actual outcome. (Warning: some authors prefer to work with $\hat{R}_t - R_t$ as the forecast error instead.)

Quite often, we compare a forecast method (or forecasting institute) with a benchmark forecast like a "no change," a random walk or the historical average. The idea of such a

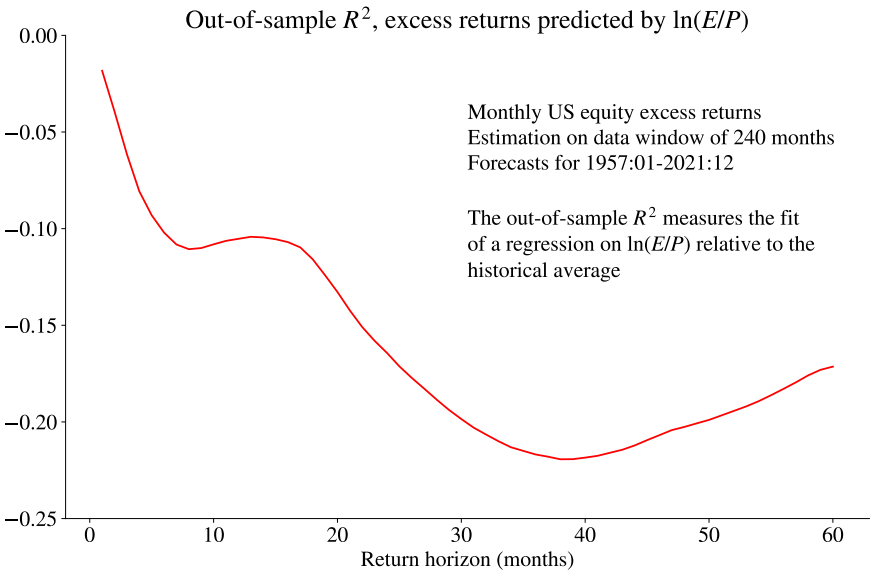


Figure 18.15: Predictability of US stock returns, out-of-sample

comparison is to study if the resources employed in creating the forecast really bring value added compared to a very simple (and inexpensive) forecast.

Ultimately, the ranking of forecasting methods should be done based on the true benefits/costs of forecast errors—which may differ between organizations. For instance, a forecasting agency has a reputation (and eventually customers) to lose, while an investor has more immediate pecuniary concerns. Unless the relation between the forecast error and the losses are immediately understood, the ranking of two forecast methods is typically done based on a number of standard criteria. Several of those criteria are inspired by basic statistics.

Most statistical forecasting methods are based on the idea of minimizing the sum of squared forecast errors, $\Sigma_{t=1}^{T}\varepsilon_t^2$. For instance, the least squares (LS) method picks the regression coefficient in

$$R_t = \beta_0 + \beta_1 x_{t-h} + \varepsilon_t \tag{18.31}$$

to minimize the sum of squared residuals. This will, among other things, give a zero mean of the fitted residuals and also a zero correlation between the fitted residual and the regressor. As usual, rational forecasts should have forecast errors that cannot be predicted (by past regressors or forecast errors).

Evaluation of a forecast often involve extending these ideas to the forecast method, irrespective of whether a LS regression has been used or not. In practice, this means

studying *(i)* whether the forecast error, e_t, has a zero mean; *(ii)* the mean squared (or absolute value) of the forecast error ; *(iii)* the fraction of times the squared (or absolute value) of the forecast error is lower than some threshold; *(iv)* the profit from investing by following a forecasting model; *(v)* if the forecast errors are autocorrelated or correlated with past information.

Remark 18.34 *(Autocorrelation of forecast errors*) An efficient h-step-ahead forecast error has a zero correlation with the forecast error h (and more) periods earlier. For instance, with* $h = 2$*, let* $e_{t+2,t} = y_{t+2} - \mathrm{E}_t\, y_{t+2}$ *be the error of forecasting* y_{t+2} *using the information in period* t*. It should be uncorrelated with* $e_{t,t-2} = y_t - \mathrm{E}_{t-2}\, y_t$*, since the latter is known when the forecast* $\mathrm{E}_t\, y_{t+2}$ *is formed.*

To perform formal tests of forecasting performance a Diebold and Mariano (1995) test is typically performed. It is an application of GMM. To implement it, consider two different forecasts. For instance, the first forecast could come from a naive forecasting model (for instance, no change) that you hope to beat (forecast errors e_t) and the other is your estimated model (forecast errors ε_t). To test the different aspects discussed before, let $\delta(x)$ be an indicator function that is one if x is true and zero otherwise, and let R_t^e and R_t^ε denote the returns from following trading strategies based on the different forecasts. Then, we could consider, for instance, the following moment conditions

$$g_t = e_t - \varepsilon_t, \text{ or} \tag{18.32}$$

$$g_t = e_t^2 - \varepsilon_t^2 \text{ or } g_t = |e_t| - |\varepsilon_t|, \text{ or} \tag{18.33}$$

$$g_t = \delta[\text{sign}(\tilde{R}_t) \neq \text{sign}(R_t)] - \delta[\text{sign}(\hat{R}_t) \neq \text{sign}(R_t)], \text{ or} \tag{18.34}$$

$$g_t = R_t^e - R_t^\varepsilon, \text{ or} \tag{18.35}$$

$$g_t = e_t e_{t-1} - \varepsilon_t \varepsilon_{t-1} \text{ or } g_t = e_t x_{t-h} - \varepsilon_t x_{t-h}. \tag{18.36}$$

The different moment conditions correspond to the different aspects of the forecasts discussed above. For instance, (18.32) is for testing if the two methods have the same average forecast error, while (18.33) tests the MSE, which is an application of the Mariano-Diebold approach. In contrast, (18.34) tests if the e model forecasts the wrong sign of the return more often than the ε model does. Finally, (18.35) compares the returns of a trading strategy (not specified here) that depends on the forecasts and (18.36) tests if the e_t errors are more predictable than the ε_t errors.

From the usual properties of GMM, we have typically have that (if the null hypothesis

is true)

$$\sqrt{T}\bar{g} \to^d N(0, S_0), \tag{18.37}$$

where $\bar{g} = \sum_{t=1}^{T} g_t/T$ is the average moment condition and S_0 is the variance of $\sqrt{T}\bar{g}$. When g_t has no autocorrelation, then we can use $S_0 = \text{Var}(g_t)/T$. Otherwise, S_0 can be estimated by, for instance, a Newey-West approach. It is especially important to handle autocorrelations in the forecast errors when we are forecasting multi-period returns using overlapping data (for instance, monthly data on annual returns). This can be used to construct a t-test.

The null hypothesis is that $\text{E}\, g_t = 0$ (the two models perform equally well). In a two-sided test the alternative hypothesis is $\text{E}\, g_t \neq 0$ (are the forecast errors different?). The null is then rejected whenever the t-stat calculated from (18.37) is large in absolute value (for instance, $|t| > 1.645$ for the 10% significance level). A one-sided test (where the alternative is $\text{E}\, g_t > 0$ or $\text{E}\, g_t < 0$) is also easy to perform.

However, when the models behind e and ε are *nested* (say, e is generated by a special case of the model that generates ε), then the asymptotic distribution is non-normal so other critical values must be applied (see Clark and McCracken (2001)). This is, for instance, the case when model behind e includes just an intercept and model behind ε has an intercept and a slope coefficient of some predictor x. If applied to returns, the model behind e would just pick up the historical average return, while the model behind ε would also capture the predictive changes related to x. The basic reason for the non-normal behaviour is that, even under the null hypothesis of equal performance, the average $e_t^2 - \varepsilon_t^2$ is likely to negative since ε_t^2 is affected by the noise caused by estimating too many parameters. Clark and West (2007) suggest another way of handling this problem. In particular, they suggest replacing the squared forecast errors in (18.33) with

$$g_t = e_t^2 - [\varepsilon_t^2 - (\hat{R}_t^e - \hat{R}_t^\varepsilon)^2], \tag{18.38}$$

and then use (18.37). This approach adjusts for the fact that the model behind ε is affected by noise caused by the estimation of the extra parameters. (This logic assumes that the smaller model is the true one, so the larger model includes parameters that ought to be set to zero.)

Since $R_t^e = \hat{R}_t^e + e_t$ (and similarly for ε_t), (18.38) can be rewritten in terms of the

forecast errors only

$$g_t = e_t^2 - [\varepsilon_t^2 - (e_t - \varepsilon_t)^2] \tag{18.39}$$

$$= 2e_t(e_t - \varepsilon_t). \tag{18.40}$$

(Recall that e_t is the error from the smaller model, while ε_t is from the larger model.) The null hypothesis is $\mathrm{E}\, g_t = 0$ and the alternative that $\mathrm{E}\, g_t \neq 0$ for a two sided test (the permance is different) or $\mathrm{E}\, g_t > 0$ for a one-sided test (the smaller "e" model is worse than the larger "ε" model).

The simulation evidence in Clark and West (2007) suggests that using (18.40) or applying a bootstrap to (18.33) have similar properties. The bootstrap approach can also be readily applied to the other evaluation criteria (18.32)–(18.36).

	AR(1)		E/P		Combination	
	mean	t-stat	mean	t-stat	mean	t-stat
MSE in-sample	288.66		276.29			
R^2_{oos}	−0.04		−0.06		−0.02	
$e - \varepsilon$	0.23	1.96	−1.40	−1.38	−0.59	−1.13
$e^2 - \varepsilon^2$	−12.96	−1.55	−17.70	−0.76	−7.21	−0.64
$\lvert e\rvert - \lvert\varepsilon\rvert$	−0.21	−1.47	−0.73	−1.03	−0.28	−0.78
$2e(e - \varepsilon)$	−12.03	−1.49	13.54	0.57	0.75	0.07

Table 18.8: Mariano-Diebold (and Clark-West) tests of forecasting 1-year S&P returns with different models. The total sample is 1946–2020, but the forecasts are made for 1971–2020. The e forecasts are the historical average returns while the ε forecasts are out-of-sample and based on the different regressions. Estimation is done on an expanding data window. The std use a NW approach with 1 lag (year).

	5th percentile	95th percentile
$e^2 - \varepsilon^2$	−2.20	0.21
$\lvert e\rvert - \lvert\varepsilon\rvert$	−2.53	0.04
$2e(e - \varepsilon)$	−1.71	1.67

Table 18.9: Bootstrapped percentiles of the Mariano-Diebold (and Clark-West) tests of the E/P model in Table 18.8. The simulations are done under the null hypothesis by randomly drawing (with replacement) the returns from the prediction sample.

Empirical Example 18.35 *(Empirical results on predicting annual S&P 500 returns) Tables 18.8–18.9 and Figure 18.16 summarize the results. The combined model seems to do*

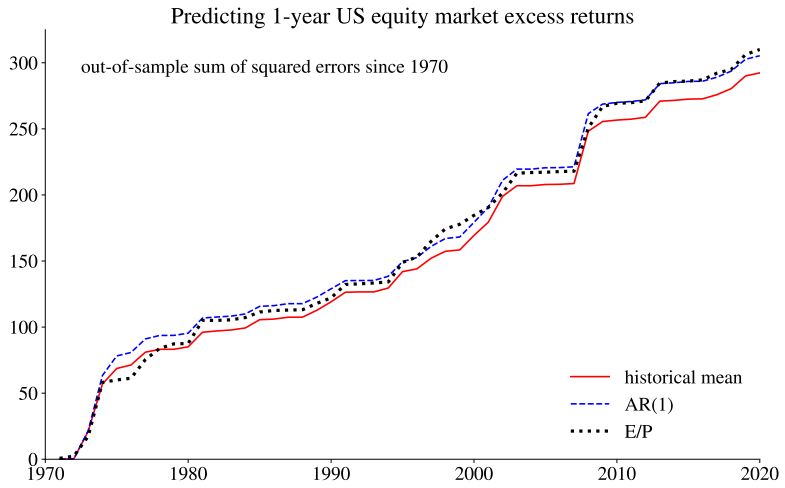


Figure 18.16: Accumumation of the (oos) MSE from three different forecasting models

slightly better than the two individual models. The build-up in the oos MSE shows some jumps, but the ranking of the three methods do not change dramatically over time. Notice also that the bootstrapped confidence bands in 18.9 appear very asymmetric, in spite of being simulated under the null hypothesis. Only the Clark-West has a symmetric confidence band.

For instance, Leitch and Tanner (1991) analyse the profits from selling 3-month T-bill futures when the forecasted interest rate is above futures rate (forecasted bill price is below futures price). The profit from this strategy is (not surprisingly) strongly related to measures of correct direction of change (see above), but (perhaps more surprisingly) not very strongly related to mean squared error, or absolute errors.

18.10 Appendix: Prices and Dividends

The gross return, R_{t+1}, is defined as

$$R_{t+1} = \frac{D_{t+1} + P_{t+1}}{P_t}, \text{ so } P_t = \frac{D_{t+1} + P_{t+1}}{R_{t+1}}. \tag{18.41}$$

Substituting for P_{t+1} (and then $P_{t+2}, ...$) gives

$$P_t = \frac{D_{t+1}}{R_{t+1}} + \frac{D_{t+2}}{R_{t+1}R_{t+2}} + \frac{D_{t+3}}{R_{t+1}R_{t+2}R_{t+3}} + \dots \tag{18.42}$$

$$= \sum_{j=1}^{\infty} \frac{D_{t+j}}{\prod_{k=1}^{j} R_{t+k}}, \tag{18.43}$$

provided the discounted value of P_{t+j} goes to zero as $j \to \infty$. This is simply an accounting identity. It is clear that a high price in t must lead to low future returns and/or high future dividends—which (by rational expectations) also carry over to expectations of future returns and dividends.

It is sometimes more convenient to analyse the price-dividend ratio. Dividing (18.42) and (18.43) by D_t gives

$$\frac{P_t}{D_t} = \frac{1}{R_{t+1}}\frac{D_{t+1}}{D_t} + \frac{1}{R_{t+1}R_{t+2}}\frac{D_{t+2}}{D_{t+1}}\frac{D_{t+1}}{D_t} + \frac{1}{R_{t+1}R_{t+2}R_{t+3}}\frac{D_{t+3}}{D_{t+2}}\frac{D_{t+2}}{D_{t+1}}\frac{D_{t+1}}{D_t} + \dots \tag{18.44}$$

$$= \sum_{j=1}^{\infty}\prod_{k=1}^{j}\frac{D_{t+k}/D_{t+k-1}}{R_{t+k}}. \tag{18.45}$$

As with (18.43) it is just an accounting identity. It must therefore also hold in expectations. Since expectations are good (the best?) predictors of future values, we have the implication that the asset price should predict a discounted sum of future dividends, (18.43), and that the price-dividend ratio should predict a discounted sum of future changes in dividends.

Proof. (of (18.14)—slow version) Rewrite (18.41) as

$$R_{t+1} = \frac{D_{t+1} + P_{t+1}}{P_t} = \frac{P_{t+1}}{P_t}\left(1 + \frac{D_{t+1}}{P_{t+1}}\right) \text{ or in logs}$$

$$\tilde{r}_{t+1} = p_{t+1} - p_t + \ln\left[1 + \exp(d_{t+1} - p_{t+1})\right].$$

Make a first order Taylor approximation of the last term around a steady state value of $d_{t+1} - p_{t+1}$, denoted $\overline{d-p}$,

$$\ln\left[1 + \exp(d_{t+1} - p_{t+1})\right] \approx \ln\left[1 + \exp(\overline{d-p})\right] + \frac{\exp(\overline{d-p})}{1 + \exp(\overline{d-p})}\left[d_{t+1} - p_{t+1} - \left(\overline{d-p}\right)\right]$$

$$\approx \text{ constant} + (1-\rho)\left(d_{t+1} - p_{t+1}\right),$$

where $\rho = 1/[1 + \exp(\overline{d - p})] = 1/(1 + \overline{D/P})$. Combine and forget about the constant. The result is

$$\begin{aligned}\tilde{r}_{t+1} &\approx p_{t+1} - p_t + (1-\rho)(d_{t+1} - p_{t+1}) \\ &= \rho p_{t+1} - p_t + (1-\rho) d_{t+1},\end{aligned}$$

where $0 < \rho < 1$. Add and subtract d_t from the right hand side and rearrange

$$\begin{aligned}\tilde{r}_{t+1} &\approx \rho(p_{t+1} - d_{t+1}) - (p_t - d_t) + (d_{t+1} - d_t), \text{ or} \\ p_t - d_t &\approx \rho(p_{t+1} - d_{t+1}) + (d_{t+1} - d_t) - \tilde{r}_{t+1}\end{aligned}$$

This is a (forward looking, unstable) difference equation, which we can solve recursively forward. Provided $\lim_{s\to\infty} \rho^s (p_{t+s} - d_{t+s}) = 0$, the solution is (18.14). (Trying to solve for the log price level instead of the log price-dividend ratio is problematic since the condition $\lim_{s\to\infty} \rho^s p_{t+s} = 0$ may not be satisfied.) ■

Chapter 30

Appendix: A Primer in Matrix Algebra*

30.0.1 Adding and Multiplying: A Matrix and a Scalar

For this appendix, let c be a scalar and define the matrices

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, z = \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}, A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}, \text{ and } B = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix}.$$

Multiplying a matrix by a scalar means multiplying each element by the scalar

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} c = \begin{bmatrix} A_{11}c & A_{12}c \\ A_{21}c & A_{22}c \end{bmatrix}.$$

Example 30.1 *(Matrix × scalar)*

$$\begin{bmatrix} 1 & 3 \\ 3 & 4 \end{bmatrix} 10 = \begin{bmatrix} 10 & 30 \\ 30 & 40 \end{bmatrix}.$$

Adding/subtracting a scalar to each element of a matrix can be done by

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} + cJ = \begin{bmatrix} A_{11}+c & A_{12}+c \\ A_{21}+c & A_{22}+c \end{bmatrix},$$

where J is a matrix (of the same size as A) filled with ones. This is sometimes written $A + c$, although that notation is not universally liked. In some applications, $\mathbf{1}_n$ (or just $\mathbf{1}$) is used to denote a vector of n ones.

Example 30.2 *(Matrix ± scalar)*

$$\begin{bmatrix} 10 \\ 11 \end{bmatrix} - 10 \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$
$$\begin{bmatrix} 1 & 3 \\ 3 & 4 \end{bmatrix} + 10 \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 11 & 13 \\ 13 & 14 \end{bmatrix}.$$

30.0.2 Adding and Multiplying: Two Matrices

Matrix *addition* (or subtraction) is element by element

$$A + B = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} + \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix} = \begin{bmatrix} A_{11} + B_{11} & A_{12} + B_{12} \\ A_{21} + B_{21} & A_{22} + B_{22} \end{bmatrix}.$$

Example 30.3 *(Matrix addition and subtraction)*

$$\begin{bmatrix} 10 \\ 11 \end{bmatrix} - \begin{bmatrix} 2 \\ 5 \end{bmatrix} = \begin{bmatrix} 8 \\ 6 \end{bmatrix}$$
$$\begin{bmatrix} 1 & 3 \\ 3 & 4 \end{bmatrix} + \begin{bmatrix} 1 & 2 \\ 3 & -2 \end{bmatrix} = \begin{bmatrix} 2 & 5 \\ 6 & 2 \end{bmatrix}$$

To turn a column into a row vector, use the *transpose* operator like in x'

$$x' = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}' = \begin{bmatrix} x_1 & x_2 \end{bmatrix}.$$

Matrix *multiplication* requires the two matrices to be conformable: the first matrix has as many columns as the second matrix has rows. Element ij of the result is the multiplication of the ith row of the first matrix with the jth column of the second matrix

$$AB = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix} = \begin{bmatrix} A_{11}B_{11} + A_{12}B_{21} & A_{11}B_{12} + A_{12}B_{22} \\ A_{21}B_{11} + A_{22}B_{21} & A_{21}B_{12} + A_{22}B_{22} \end{bmatrix}.$$

Multiplying a square matrix A with a column vector z gives a column vector

$$Az = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = \begin{bmatrix} A_{11}z_1 + A_{12}z_2 \\ A_{21}z_1 + A_{22}z_2 \end{bmatrix}.$$

Example 30.4 *(Matrix multiplication)*

$$\begin{bmatrix} 1 & 3 \\ 3 & 4 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 3 & -2 \end{bmatrix} = \begin{bmatrix} 10 & -4 \\ 15 & -2 \end{bmatrix}$$
$$\begin{bmatrix} 1 & 3 \\ 3 & 4 \end{bmatrix} \begin{bmatrix} 2 \\ 5 \end{bmatrix} = \begin{bmatrix} 17 \\ 26 \end{bmatrix}$$

30.0.3 Transpose

Similarly, transposing a matrix is like flipping it around the main diagonal

$$A' = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}' = \begin{bmatrix} A_{11} & A_{21} \\ A_{12} & A_{22} \end{bmatrix}.$$

Example 30.5 *(Matrix transpose)*

$$\begin{bmatrix} 10 \\ 11 \end{bmatrix}' = \begin{bmatrix} 10 & 11 \end{bmatrix}$$
$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}' = \begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix}$$

30.0.4 Inner and Outer Products, Quadratic Forms

For two column vectors x and z, the product $x'z$ is called the *inner product* (a scalar)

$$x'z = \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = x_1 z_1 + x_2 z_2,$$

and xz' the *outer product* (a matrix)

$$xz' = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \begin{bmatrix} z_1 & z_2 \end{bmatrix} = \begin{bmatrix} x_1 z_1 & x_1 z_2 \\ x_2 z_1 & x_2 z_2 \end{bmatrix}.$$

(Notice that xz does not work).

Example 30.6 *(Inner and outer products)*

$$\begin{bmatrix} 10 \\ 11 \end{bmatrix}' \begin{bmatrix} 2 \\ 5 \end{bmatrix} = \begin{bmatrix} 10 & 11 \end{bmatrix} \begin{bmatrix} 2 \\ 5 \end{bmatrix} = 75$$
$$\begin{bmatrix} 10 \\ 11 \end{bmatrix} \begin{bmatrix} 2 \\ 5 \end{bmatrix}' = \begin{bmatrix} 10 \\ 11 \end{bmatrix} \begin{bmatrix} 2 & 5 \end{bmatrix} = \begin{bmatrix} 20 & 50 \\ 22 & 55 \end{bmatrix}$$

If x is a column vector and A a square matrix, then the product $x'Ax$ is a quadratic form *(a scalar).*

Example 30.7 *(Quadratic form)*

$$\begin{bmatrix} 10 \\ 11 \end{bmatrix}' \begin{bmatrix} 1 & 3 \\ 3 & 4 \end{bmatrix} \begin{bmatrix} 10 \\ 11 \end{bmatrix} = 1244$$

30.0.5 Matrix Inverse

A matrix *inverse* is the closest we get to "dividing" by a matrix. The inverse of a matrix A, denoted A^{-1}, is such that

$$AA^{-1} = I \text{ and } A^{-1}A = I,$$

where I is the *identity matrix* (ones along the diagonal, and zeros elsewhere). The matrix inverse is useful for solving systems of linear equations, $y = Ax$ as $x = A^{-1}y$.

For a 2×2 matrix we have

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}^{-1} = \frac{1}{A_{11}A_{22} - A_{12}A_{21}} \begin{bmatrix} A_{22} & -A_{12} \\ -A_{21} & A_{11} \end{bmatrix}.$$

Example 30.8 *(Matrix inverse) We have*

$$\begin{bmatrix} -4/5 & 3/5 \\ 3/5 & -1/5 \end{bmatrix} \begin{bmatrix} 1 & 3 \\ 3 & 4 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \text{ so}$$
$$\begin{bmatrix} 1 & 3 \\ 3 & 4 \end{bmatrix}^{-1} = \begin{bmatrix} -4/5 & 3/5 \\ 3/5 & -1/5 \end{bmatrix}.$$

30.0.6 Solving Systems of Linear Equations

If A is $n \times n$ and invertible and b and y are $n \times 1$ vectors, then we can solve

$$Ab = y \text{ as } b = A^{-1}y.$$

This solution is unique.

$$\begin{bmatrix} 1 & 3 \\ 3 & 4 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} 10 \\ 11 \end{bmatrix}, \text{ gives}$$
$$\begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} 1 & 3 \\ 3 & 4 \end{bmatrix}^{-1} \begin{bmatrix} 10 \\ 11 \end{bmatrix} = \begin{bmatrix} -1.4 \\ 3.8 \end{bmatrix}.$$

30.0.7 OLS Notation: $X'X$ or $\sum_{t=1}^{T} x_t x_t'$?

Let x_t be a $K \times 1$ vector of (of data in period t). We can the calculate the outer product ($K \times K$) as $x_t x_t'$ and summing each element across T observations gives the $K \times K$ matrix $S_{xx} = \sum_{t=1}^{T} x_t x_t'$.

Alternatively, let X be a $T \times K$ matrix with x_t' in row t. Then we can also calculate S_{xx} as $X'X$.

Example 30.9 *(Sum of outer product, $\sum_{t=1}^{T} x_t x_t'$)*

$$x_1 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}, x_2 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \textit{and } x_3 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

We then have

$$\begin{aligned} \sum_{t=1}^{T} x_t x_t' &= \begin{bmatrix} 1 \\ -1 \end{bmatrix} \begin{bmatrix} 1 & -1 \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \end{bmatrix} \begin{bmatrix} 1 & 1 \end{bmatrix} \\ &= \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \\ &= \begin{bmatrix} 3 & 0 \\ 0 & 2 \end{bmatrix}. \end{aligned}$$

In this example, the matrix happens to be diagonal, but that is not a general result. However, it will always be symmetric.

Example 30.10 *(Sum of outer product, $X'X$) Define*

$$X = \begin{bmatrix} 1 & -1 \\ 1 & 0 \\ 1 & 1 \end{bmatrix}.$$

It is straightforward to calculate that $X'X = \begin{bmatrix} 3 & 0 \\ 0 & 2 \end{bmatrix}$.

30.0.8 Derivatives of Matrix Expressions

Let z and x be $n \times 1$ vectors. The *derivative of the inner product* is $\partial(z'x)/\partial z = x$.

Example 30.11 *(Derivative of an inner product) With* $n = 2$

$$z'x = z_1x_1 + z_2x_2, \text{ so } \frac{\partial(z'x)}{\partial z} = \frac{\partial(z_1x_1 + z_2x_2)}{\begin{bmatrix} \partial z_1 \\ \partial z_2 \end{bmatrix}} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}.$$

Let x be $n \times 1$ and A a symmetric $n \times n$ matrix. The *derivative of the quadratic form* is $\partial(x'Ax)/\partial x = 2Ax$.

Example 30.12 *(Derivative of a quadratic form) With* $n = 2$*, the quadratic form is*

$$x'Ax = \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} A_{11} & A_{12} \\ A_{12} & A_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = x_1^2A_{11} + x_2^2A_{22} + 2x_1x_2A_{12}.$$

The derivatives with respect to x_1 *and* x_2 *are*

$$\frac{\partial(x'Ax)}{\partial x_1} = 2x_1A_{11} + 2x_2A_{12} \text{ and } \frac{\partial(x'Ax)}{\partial x_2} = 2x_2A_{22} + 2x_1A_{12}, \text{ or}$$
$$\frac{\partial(x'Ax)}{\begin{bmatrix} \partial x_1 \\ \partial x_2 \end{bmatrix}} = 2 \begin{bmatrix} A_{11} & A_{12} \\ A_{12} & A_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}.$$

Chapter 31

Some Statistics

This section summarizes some useful facts about statistics. Heuristic proofs are given in a few cases.

Some references: Mittelhammer (1996), DeGroot (1986), Greene (2000), Davidson (2000), Johnson, Kotz, and Balakrishnan (1994).

31.1 Distributions and Moment Generating Functions

Most of the stochastic variables we encounter in econometrics are continuous. For a continuous random variable X, the range is uncountably infinite and the probability that $X \leq x$ is $\Pr(X \leq x) = \int_{-\infty}^{x} f(q)dq$ where $f(q)$ is the continuous probability density function of X. Note that X is a random variable, x is a number (1.23 or so), and q is just a dummy argument in the integral.

Fact 31.1 *(cdf and pdf) The cumulative distribution function of the random variable X is $F(x) = \Pr(X \leq x) = \int_{-\infty}^{x} f(q)dq$. Clearly, $f(x) = dF(x)/dx$. Note that x is just a number, not random variable.*

Fact 31.2 *(Moment generating function of X) The moment generating function of the random variable X is $mgf(t) = \mathrm{E}\, e^{tX}$. The rth moment is the rth derivative of $mgf(t)$ evaluated at $t = 0$: $\mathrm{E}\, X^r = d^r mgf(0)/dt^r$. If a moment generating function exists (that is, $\mathrm{E}\, e^{tX} < \infty$ for some small interval $t \in (-h, h)$), then it is unique.*

Fact 31.3 *(Moment generating function of a function of X) If X has the moment generating function $mgf_X(t) = \mathrm{E}\, e^{tX}$, then $g(X)$ has the moment generating function $\mathrm{E}\, e^{tg(X)}$. The affine function $a + bX$ (a and b are constants) has the moment generating function $mgf_{g(X)}(t) = \mathrm{E}\, e^{t(a+bX)} = e^{ta}\,\mathrm{E}\, e^{tbX} = e^{ta} mgf_X(bt)$. By setting $b = 1$ and*

$a = -\mathrm{E}\,X$ we obtain a mgf for central moments (variance, skewness, kurtosis, etc), $mgf_{(X-\mathrm{E}\,X)}(t) = e^{-t\,\mathrm{E}\,X} mgf_X(t)$.

Example 31.4 *When $X \sim N(\mu, \sigma^2)$, then $mgf_X(t) = \exp\left(\mu t + \sigma^2 t^2/2\right)$. Let $Z = (X-\mu)/\sigma$ so $a = -\mu/\sigma$ and $b = 1/\sigma$. This gives $mgf_Z(t) = \exp(-\mu t/\sigma) mgf_X(t/\sigma) = \exp\left(t^2/2\right)$. (Of course, this result can also be obtained by directly setting $\mu = 0$ and $\sigma = 1$ in mgf_X.)*

Fact 31.5 *(Characteristic function and the pdf) The characteristic function of a random variable x is*

$$\begin{aligned} g(\phi) &= \mathrm{E}\exp(i\phi x) \\ &= \textstyle\int_x \exp(i\phi x) f(x) dx, \end{aligned}$$

where $f(x)$ is the pdf. This is a Fourier transform of the pdf (if x is a continuous random variable). The pdf can therefore be recovered by the inverse Fourier transform as

$$f(x) = \frac{1}{2\pi} \textstyle\int_{-\infty}^{\infty} \exp(-i\phi x) g(\phi) d\phi.$$

In practice, we typically use a fast (discrete) Fourier transform to perform this calculation, since there are very quick computer algorithms for doing that.

Fact 31.6 *The charcteristic function of a $N(\mu, \sigma^2)$ distribution is $\exp(i\phi\mu - \phi^2\sigma^2/2)$ and of a lognormal(μ, σ^2) distribuion (where $\ln x \sim N(\mu, \sigma^2)$) $\sum_{j=0}^{\infty} \frac{(i\phi)^j}{j!} \exp(j\mu + j^2\sigma^2/2)$.*

Fact 31.7 *(Change of variable, univariate case, monotonic function) Suppose X has the probability density function $f_X(c)$ and cumulative distribution function $F_X(c)$. Let $Y = g(X)$ be a continuously differentiable function with $dg/dX > 0$ (so $g(X)$ is increasing for all c such that $f_X(c) > 0$. Then the cdf of Y is*

$$F_Y(c) = \Pr[Y \leq c] = \Pr[g(X) \leq c] = \Pr[X \leq g^{-1}(c)] = F_X[g^{-1}(c)],$$

where g^{-1} is the inverse function of g such that $g^{-1}(Y) = X$. We also have that the pdf of Y is

$$f_Y(c) = f_X[g^{-1}(c)] \left| \frac{dg^{-1}(c)}{dc} \right|.$$

If, instead, $dg/dX < 0$ (so $g(X)$ is decreasing), then we instead have the cdf of Y

$$F_Y(c) = \Pr[Y \leq c] = \Pr[g(X) \leq c] = \Pr[X \geq g^{-1}(c)] = 1 - F_X[g^{-1}(c)],$$

but the same expression for the pdf.

Proof. Differentiate $F_Y(c)$, that is, $F_X[g^{-1}(c)]$ with respect to c. ■

Example 31.8 *Let $X \sim U(0,1)$ and $Y = g(X) = F^{-1}(X)$ where $F(c)$ is a strictly increasing cdf. We then get*

$$f_Y(c) = \frac{dF(c)}{dc}.$$

The variable Y then has the pdf $dF(c)/dc$ and the cdf $F(c)$. This shows how to generate random numbers from the $F()$ distribution: draw $X \sim U(0,1)$ and calculate $Y = F^{-1}(X)$.

Example 31.9 *Let $Y = \exp(X)$, so the inverse function is $X = \ln Y$ with derivative $1/Y$. Then, $f_Y(c) = f_X(\ln c)/c$. Conversely, let $Y = \ln X$, so the inverse function is $X = \exp(Y)$ with derivative $\exp(Y)$. Then, $f_Y(c) = f_X[\exp(c)]\exp(c)$.*

Example 31.10 *Let $Y = (X - \mu)/\sigma$ and suppose X has a $N(\mu, \sigma^2)$ distribution. The inverse function is $X = Y\sigma + \mu$ and its derivative is σ. Combine this with $f_X(c)$ in 31.56 to get $f_Y(c) = \exp(-\frac{1}{2}c^2)/\sqrt{2\pi}$, which is the $N(0,1)$ pdf.*

Example 31.11 *Let $X \sim U(0,2)$, so the pdf and cdf of X are then $1/2$ and $c/2$ respectively. Now, let $Y = g(X) = -X$ gives the pdf and cdf as $1/2$ and $1 + y/2$ respectively. The latter is clearly the same as $1 - F_X[g^{-1}(c)] = 1 - (-c/2)$.*

Fact 31.12 *(Distribution of truncated a random variable) Let the probability distribution and density functions of X be $F(x)$ and $f(x)$, respectively. The corresponding functions, conditional on $a < X \leq b$ are $[F(x) - F(a)]/[F(b) - F(a)]$ and $f(x)/[F(b) - F(a)]$. Clearly, outside $a < X \leq b$ the pdf is zero, while the cdf is zero below a and unity above b.*

31.2 Joint and Conditional Distributions and Moments

31.2.1 Joint and Conditional Distributions

Fact 31.13 *(Joint and marginal cdf) Let X and Y be (possibly vectors of) random variables and let x and y be two numbers. The joint cumulative distribution function of X and Y is $H(x,y) = \Pr(X \leq x, Y \leq y) = \int_{-\infty}^{x}\int_{-\infty}^{y} h(q_x, q_y)dq_y dq_x$, where $h(x,y) = \partial^2 F(x,y)/\partial x \partial y$ is the joint probability density function.*

Fact 31.14 *(Joint and marginal pdf) The marginal cdf of X is obtained by integrating out Y: $F(x) = \Pr(X \leq x, Y \text{ anything}) = \int_{-\infty}^{x} \left[\int_{-\infty}^{\infty} h(q_x, q_y) dq_y\right] dq_x$. This shows that the marginal pdf of x is $f(x) = dF(x)/dx = \int_{-\infty}^{\infty} h(q_x, q_y) dq_y$.*

Fact 31.15 *(Conditional distribution) The pdf of Y conditional on $X = x$ (a number) is $g(y|x) = h(x, y)/f(x)$. This is clearly proportional to the joint pdf (at the given value x).*

Fact 31.16 *(Change of variable, multivariate case, monotonic function) The result in Fact 31.7 still holds if X and Y are both $n \times 1$ vectors, but the derivative are now $\partial g^{-1}(c)/\partial dc'$ which is an $n \times n$ matrix. If g_i^{-1} is the ith function in the vector g^{-1} then*

$$\frac{\partial g^{-1}(c)}{\partial dc'} = \begin{bmatrix} \frac{\partial g_1^{-1}(c)}{\partial c_1} & \cdots & \frac{\partial g_1^{-1}(c)}{\partial c_n} \\ \vdots & & \vdots \\ \frac{\partial g_n^{-1}(c)}{\partial c_1} & \cdots & \frac{\partial g_n^{-1}(c)}{\partial c_m} \end{bmatrix}.$$

31.2.2 Moments of Joint Distributions

Fact 31.17 *(Caucy-Schwartz)* $(\mathrm{E}\,XY)^2 \leq \mathrm{E}(X^2)\,\mathrm{E}(Y^2)$.

Proof. $0 \leq \mathrm{E}[(aX+Y)^2] = a^2\,\mathrm{E}(X^2) + 2a\,\mathrm{E}(XY) + \mathrm{E}(Y^2)$. Set $a = -\mathrm{E}(XY)/\mathrm{E}(X^2)$ to get

$$0 \leq -\frac{[\mathrm{E}(XY)]^2}{\mathrm{E}(X^2)} + \mathrm{E}(Y^2), \text{ that is, } \frac{[\mathrm{E}(XY)]^2}{\mathrm{E}(X^2)} \leq \mathrm{E}(Y^2).$$

■

Fact 31.18 *($-1 \leq Corr(X, y) \leq 1$). Let Y and X in Fact 31.17 be zero mean variables (or variables minus their means). We then get $[\mathrm{Cov}(X, Y)]^2 \leq \mathrm{Var}(X)\,\mathrm{Var}(Y)$, that is, $-1 \leq \mathrm{Cov}(X, Y)/[Std(X)Std(Y)] \leq 1$.*

31.2.3 Conditional Moments

Fact 31.19 *(Conditional moments)* $\mathrm{E}(Y|x) = \int y g(y|x) dy$ *and* $\mathrm{Var}(Y|x) = \int [y - \mathrm{E}(Y|x)] g(y|x) dy$.

Fact 31.20 *(Conditional moments as random variables) Before we observe X, the conditional moments are random variables—since X is. We denote these random variables by* $\mathrm{E}(Y|X)$, $\mathrm{Var}(Y|X)$, *etc.*

Fact 31.21 *(Law of iterated expectations)* $\mathrm{E}\,Y = \mathrm{E}[\mathrm{E}\,(Y|X)]$. *Note that* $\mathrm{E}\,(Y|X)$ *is a random variable since it is a function of the random variable* X*. It is not a function of* Y*, however. The outer expectation is therefore an expectation with respect to* X *only.*

Proof. $\mathrm{E}[\mathrm{E}\,(Y|X)] = \int\left[\int yg(y|x)dy\right]f(x)dx = \int\int yg(y|x)f(x)dydx = \int\int yh(y,x)dydx = \mathrm{E}\,Y$. ■

Fact 31.22 *(Conditional vs. unconditional variance)* $\mathrm{Var}\,(Y) = \mathrm{Var}\,[\mathrm{E}\,(Y|X)]+\mathrm{E}\,[\mathrm{Var}\,(Y|X)]$.

Fact 31.23 *(Properties of Conditional Expectations) (a)* $Y = \mathrm{E}\,(Y|X)+U$ *where* U *and* $\mathrm{E}\,(Y|X)$ *are uncorrelated:* $\mathrm{Cov}\,(X,Y) = \mathrm{Cov}\,[X,\mathrm{E}\,(Y|X)+U] = \mathrm{Cov}\,[X,\mathrm{E}\,(Y|X)]$. *It follows that (b)* $\mathrm{Cov}[Y,\mathrm{E}\,(Y|X)] = \mathrm{Var}[\mathrm{E}\,(Y|X)]$*; and (c)* $\mathrm{Var}\,(Y) = \mathrm{Var}\,[\mathrm{E}\,(Y|X)] + \mathrm{Var}\,(U)$*. Property (c) is the same as Fact 31.22, where* $\mathrm{Var}\,(U) = \mathrm{E}\,[\mathrm{Var}\,(Y|X)]$.

Proof. $\mathrm{Cov}\,(X,Y) = \int\int x(y-\mathrm{E}\,y)h(x,y)dydx = \int x\left[\int(y-\mathrm{E}\,y)g(y|x)dy\right]f(x)dx$, but the term in brackets is $\mathrm{E}\,(Y|X) - \mathrm{E}\,Y$. ■

Fact 31.24 *(Conditional expectation and unconditional orthogonality)* $\mathrm{E}\,(Y|Z) = 0 \Rightarrow \mathrm{E}\,YZ = 0$.

Proof. Note from Fact 31.23 that $\mathrm{E}(Y|X) = 0$ implies $\mathrm{Cov}\,(X,Y) = 0$ so $\mathrm{E}\,XY = \mathrm{E}\,X\,\mathrm{E}\,Y$ (recall that $\mathrm{Cov}\,(X,Y) = \mathrm{E}\,XY - \mathrm{E}\,X\,\mathrm{E}\,Y$). Note also that $\mathrm{E}\,(Y|X) = 0$ implies that $\mathrm{E}\,Y = 0$ (by iterated expectations). We therefore get

$$\mathrm{E}\,(Y|X) = 0 \Rightarrow \begin{bmatrix} \mathrm{Cov}\,(X,Y) = 0 \\ \mathrm{E}\,Y = 0 \end{bmatrix} \Rightarrow \mathrm{E}\,YX = 0.$$

■

31.2.4 Regression Function and Linear Projection

Fact 31.25 *(Regression function) Suppose we use information in some variables* X *to predict* Y*. The choice of the forecasting function* $\hat{Y} = k(X) = \mathrm{E}\,(Y|X)$ *minimizes* $\mathrm{E}[Y-k(X)]^2$*. The conditional expectation* $\mathrm{E}\,(Y|X)$ *is also called the regression function of* Y *on* X*. See Facts 31.23 and 31.24 for some properties of conditional expectations.*

Fact 31.26 *(Linear projection) Suppose we want to forecast the scalar* Y *using the* $k\times 1$ *vector* X *and that we restrict the forecasting rule to be linear* $\hat{Y} = X'\beta$*. This rule is a linear projection, denoted* $P(Y|X)$*, if* β *satisfies the orthogonality conditions* $\mathrm{E}[X(Y - X'\beta)] = \mathbf{0}_{k\times 1}$*, that is, if* $\beta = (\mathrm{E}\,XX')^{-1}\,\mathrm{E}\,XY$*. A linear projection minimizes* $\mathrm{E}[Y - k(X)]^2$ *within the class of linear* $k(X)$ *functions.*

Fact 31.27 *(Properties of linear projections) (a) The orthogonality conditions in Fact 31.26 mean that*

$$Y = X'\beta + \varepsilon,$$

where $\mathrm{E}(X\varepsilon) = \mathbf{0}_{k\times 1}$. *This implies that* $\mathrm{E}[P(Y|X)\varepsilon] = 0$, *so the forecast and forecast error are orthogonal. (b) The orthogonality conditions also imply that* $\mathrm{E}[XY] = \mathrm{E}[XP(Y|X)]$. *(c) When* X *contains a constant, so* $\mathrm{E}\,\varepsilon = 0$, *then (a) and (b) carry over to covariances:* $\mathrm{Cov}[P(Y|X), \varepsilon] = 0$ *and* $\mathrm{Cov}[X, Y] = \mathrm{Cov}[XP, (Y|X)]$.

Example 31.28 *($P(1|X)$) When* $Y_t = 1$, *then* $\beta = (\mathrm{E}\,XX')^{-1}\mathrm{E}\,X$. *For instance, suppose* $X = [x_{1t}, x_{t2}]'$. *Then*

$$\beta = \begin{bmatrix} \mathrm{E}\,x_{1t}^2 & \mathrm{E}\,x_{1t}x_{2t} \\ \mathrm{E}\,x_{2t}x_{1t} & \mathrm{E}\,x_{2t}^2 \end{bmatrix}^{-1} \begin{bmatrix} \mathrm{E}\,x_{1t} \\ \mathrm{E}\,x_{2t} \end{bmatrix}.$$

If $x_{1t} = 1$ *in all periods, then this simplifies to* $\beta = [1, 0]'$.

Remark 31.29 *Some authors prefer to take the transpose of the forecasting rule, that is, to use* $\hat{Y} = \beta' X$. *Clearly, since* XX' *is symmetric, we get* $\beta' = \mathrm{E}(YX')(\mathrm{E}\,XX')^{-1}$.

Fact 31.30 *(Linear projection with a constant in* X*) If* X *contains a constant, then* $P(aY + b|X) = aP(Y|X) + b$.

Fact 31.31 *(Linear projection versus regression function) Both the linear regression and the regression function (see Fact 31.25) minimize* $\mathrm{E}[Y - k(X)]^2$, *but the linear projection imposes the restriction that* $k(X)$ *is linear, whereas the regression function does not impose any restrictions. In the special case when* Y *and* X *have a joint normal distribution, then the linear projection is the regression function.*

Fact 31.32 *(Linear projection and OLS) The linear projection is about population moments, but OLS is its sample analogue.*

31.3 Convergence in Probability, Mean Square, and Distribution

Fact 31.33 *(Convergence in probability) The sequence of random variables* $\{X_T\}$ *converges in probability to the random variable* X *if (and only if) for all* $\varepsilon > 0$

$$\lim_{T\to\infty} \Pr(|X_T - X| < \varepsilon) = 1.$$

We denote this $X_T \xrightarrow{p} X$ *or* $\operatorname{plim} X_T = X$ *(*X *is the probability limit of* X_T*). Note: (a)* X *can be a constant instead of a random variable; (b) if* X_T *and* X *are matrices, then* $X_T \xrightarrow{p} X$ *if the previous condition holds for every element in the matrices.*

Example 31.34 *Suppose* $X_T = 0$ *with probability* $(T-1)/T$ *and* $X_T = T$ *with probability* $1/T$*. Note that* $\lim_{T\to\infty} \Pr(|X_T - 0| = 0) = \lim_{T\to\infty}(T-1)/T = 1$*, so* $\lim_{T\to\infty} \Pr(|X_T - 0| = \varepsilon) = 1$ *for any* $\varepsilon > 0$*. Note also that* $\mathrm{E}\, X_T = 0 \times (T-1)/T + T \times 1/T = 1$*, so* X_T *is biased.*

Fact 31.35 *(Convergence in mean square) The sequence of random variables* $\{X_T\}$ *converges in mean square to the random variable* X *if (and only if)*

$$\lim_{T\to\infty} \mathrm{E}(X_T - X)^2 = 0.$$

We denote this $X_T \xrightarrow{m} X$*. Note: (a)* X *can be a constant instead of a random variable; (b) if* X_T *and* X *are matrices, then* $X_T \xrightarrow{m} X$ *if the previous condition holds for every element in the matrices.*

Fact 31.36 *(Convergence in mean square to a constant) If* X *in Fact 31.35 is a constant, then then* $X_T \xrightarrow{m} X$ *if (and only if)*

$$\lim_{T\to\infty} (\mathrm{E}\, X_T - X)^2 = 0 \text{ and } \lim_{T\to\infty} \operatorname{Var}(X_T{}^2) = 0.$$

This means that both the variance and the squared bias go to zero as $T \to \infty$*.*

Proof. $\mathrm{E}(X_T - X)^2 = \mathrm{E}\, X_T^2 - 2X\, \mathrm{E}\, X_T + X^2$. Add and subtract $(\mathrm{E}\, X_T)^2$ and recall that $\operatorname{Var}(X_T) = \mathrm{E}\, X_T^2 - (\mathrm{E}\, X_T)^2$. This gives $\mathrm{E}(X_T - X)^2 = \operatorname{Var}(X_T) - 2X\, \mathrm{E}\, X_T + X^2 + (\mathrm{E}\, X_T)^2 = \operatorname{Var}(X_T) + (\mathrm{E}\, X_T - X)^2$. ■

Fact 31.37 *(Convergence in distribution) Consider the sequence of random variables* $\{X_T\}$ *with the associated sequence of cumulative distribution functions* $\{F_T\}$*. If* $\lim_{T\to\infty} F_T = F$ *(at all points), then* F *is the limiting cdf of* X_T*. If there is a random variable* X *with cdf* F*, then* X_T *converges in distribution to* X*:* $X_T \xrightarrow{d} X$*. Instead of comparing cdfs, the comparison can equally well be made in terms of the probability density functions or the moment generating functions.*

Fact 31.38 *(Relation between the different types of convergence) We have* $X_T \xrightarrow{m} X \Rightarrow X_T \xrightarrow{p} X \Rightarrow X_T \xrightarrow{d} X$*. The reverse implications are not generally true.*

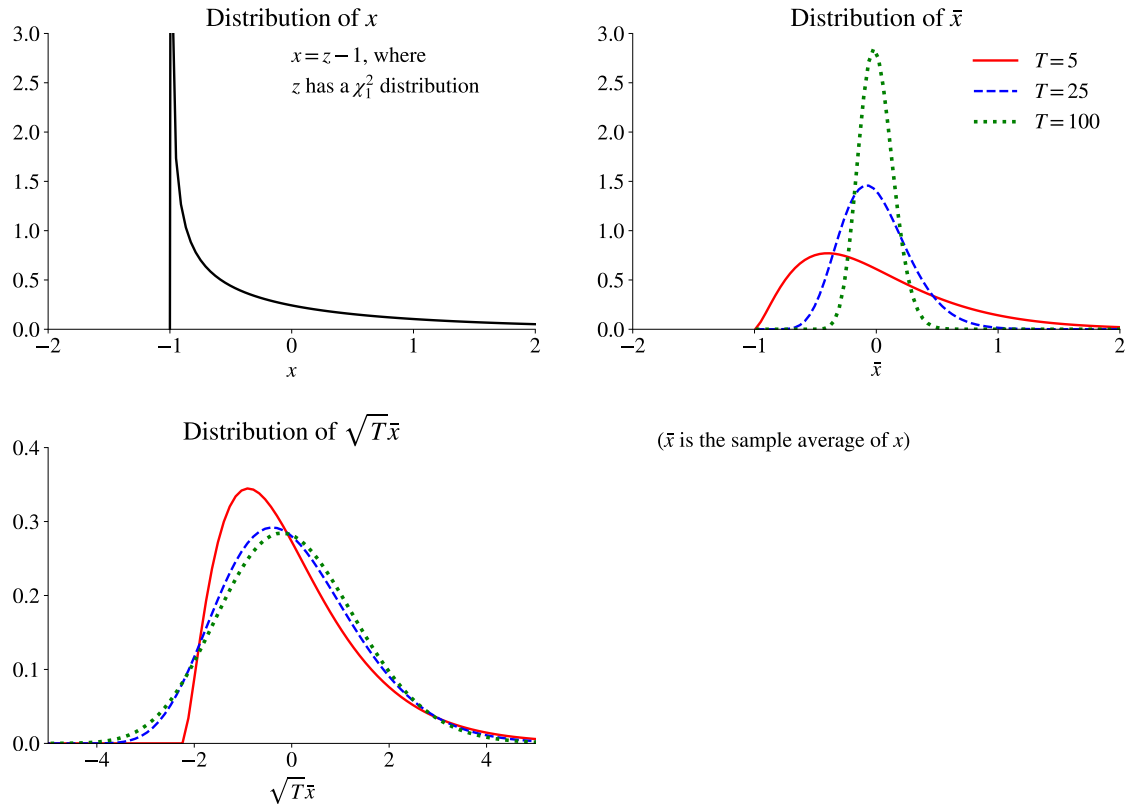


Figure 31.1: Sampling distributions

Example 31.39 *Consider the random variable in Example 31.34. The expected value is* $\mathrm{E}\, X_T = 0(T-1)/T + T/T = 1$. *This means that the squared bias does not go to zero, so* X_T *does not converge in mean square to zero.*

Fact 31.40 *(Slutsky's theorem) If* $\{X_T\}$ *is a sequence of random matrices such that* $\operatorname{plim} X_T = X$ *and* $g(X_T)$ *a continuous function, then* $\operatorname{plim} g(X_T) = g(X)$.

Fact 31.41 *(Continuous mapping theorem) Let the sequences of random matrices* $\{X_T\}$ *and* $\{Y_T\}$, *and the non-random matrix* $\{a_T\}$ *be such that* $X_T \xrightarrow{d} X$, $Y_T \xrightarrow{p} Y$, *and* $a_T \to a$ *(a traditional limit). Let* $g(X_T, Y_T, a_T)$ *be a continuous function. Then* $g(X_T, Y_T, a_T) \xrightarrow{d} g(X, Y, a)$.

31.4 Laws of Large Numbers and Central Limit Theorems

Fact 31.42 *(Khinchine's theorem) Let* X_t *be independently and identically distributed (iid) with* $\mathrm{E}\, X_t = \mu < \infty$. *Then* $\Sigma_{t=1}^{T} X_t / T \xrightarrow{p} \mu$.

Fact 31.43 *(Chebyshev's theorem) If* $\mathrm{E}\, X_t = 0$ *and* $\lim_{T\to\infty} \mathrm{Var}(\Sigma_{t=1}^{T} X_t/T) = 0$*, then* $\Sigma_{t=1}^{T} X_t/T \xrightarrow{p} 0$.

Fact 31.44 *(The Lindeberg-Lévy theorem) Let* X_t *be independently and identically distributed (iid) with* $\mathrm{E}\, X_t = 0$ *and* $\mathrm{Var}(X_t) < \infty$*. Then* $\frac{1}{\sqrt{T}} \Sigma_{t=1}^{T} X_t/\sigma \xrightarrow{d} N(0,1)$.

31.5 Stationarity

Fact 31.45 *(Covariance stationarity)* X_t *is covariance stationary if*

$$\begin{aligned} \mathrm{E}\, X_t &= \mu \text{ is independent of } t, \\ \mathrm{Cov}\,(X_{t-s}, X_t) &= \gamma_s \text{ depends only on } s, \text{ and} \\ &\text{both } \mu \text{ and } \gamma_s \text{ are finite.} \end{aligned}$$

Fact 31.46 *(Strict stationarity)* X_t *is strictly stationary if, for all* s*, the joint distribution of* $X_t, X_{t+1}, ..., X_{t+s}$ *does not depend on* t.

Fact 31.47 *(Strict stationarity versus covariance stationarity) In general, strict stationarity does not imply covariance stationarity or vice versa. However, strict stationary with finite first two moments implies covariance stationarity.*

31.6 Martingales

Fact 31.48 *(Martingale) Let* Ω_t *be a set of information in* t*, for instance* $Y_t, Y_{t-1}, ...$ *If* $\mathrm{E}\,|Y_t| < \infty$ *and* $\mathrm{E}(Y_{t+1}|\Omega_t) = Y_t$*, then* Y_t *is a martingale.*

Fact 31.49 *(Martingale difference) If* Y_t *is a martingale, then* $X_t = Y_t - Y_{t-1}$ *is a martingale difference:* X_t *has* $\mathrm{E}\,|X_t| < \infty$ *and* $\mathrm{E}(X_{t+1}|\Omega_t) = 0$.

Fact 31.50 *(Innovations as a martingale difference sequence) The forecast error* $X_{t+1} = Y_{t+1} - \mathrm{E}(Y_{t+1}|\Omega_t)$ *is a martingale difference.*

Fact 31.51 *(Properties of martingales) (a) If* Y_t *is a martingale, then* $\mathrm{E}(Y_{t+s}|\Omega_t) = Y_t$ *for* $s \geq 1$*. (b) If* X_t *is a martingale difference, then* $\mathrm{E}(X_{t+s}|\Omega_t) = 0$ *for* $s \geq 1$.

Proof. (a) Note that $\mathrm{E}(Y_{t+2}|\Omega_{t+1}) = Y_{t+1}$ and take expectations conditional on Ω_t: $\mathrm{E}[\mathrm{E}(Y_{t+2}|\Omega_{t+1})|\Omega_t] = \mathrm{E}(Y_{t+1}|\Omega_t) = Y_t$. By iterated expectations, the first term equals $\mathrm{E}(Y_{t+2}|\Omega_t)$. Repeat this for $t+3$, $t+4$, etc. (b) Essentially the same proof. ■

Fact 31.52 *(Properties of martingale differences) If X_t is a martingale difference and g_{t-1} is a function of Ω_{t-1}, then $X_t g_{t-1}$ is also a martingale difference.*

Proof. $\mathrm{E}(X_{t+1}g_t|\Omega_t) = \mathrm{E}(X_{t+1}|\Omega_t)g_t$ since g_t is a function of Ω_t. ■

Fact 31.53 *(Martingales, serial independence, and no autocorrelation) (a) X_t is serially uncorrelated if $\mathrm{Cov}(X_t, X_{t+s}) = 0$ for all $s \neq 0$. This means that a linear projection of X_{t+s} on $X_t, X_{t-1,...}$ is a constant, so it cannot help predict X_{t+s}. (b) X_t is a martingale difference with respect to its history if $\mathrm{E}(X_{t+s}|X_t, X_{t-1}, ...) = 0$ for all $s \geq 1$. This means that no function of $X_t, X_{t-1}, ...$ can help predict X_{t+s}. (c) X_t is serially independent if $\mathrm{pdf}(X_{t+s}|X_t, X_{t-1}, ...) = \mathrm{pdf}(X_{t+s})$. This means than no function of $X_t, X_{t-1}, ...$ can help predict any function of X_{t+s}.*

Fact 31.54 *(WLN for martingale difference) If X_t is a martingale difference, then $\operatorname{plim} \Sigma_{t=1}^T X_t/T = 0$ if either (a) X_t is strictly stationary and $\mathrm{E}|x_t| < 0$ or (b) $\mathrm{E}|x_t|^{1+\delta} < \infty$ for $\delta > 0$ and all t. (See Davidson (2000) 6.2)*

Fact 31.55 *(CLT for martingale difference) Let X_t be a martingale difference. If $\operatorname{plim} \Sigma_{t=1}^T (X_t^2 - \mathrm{E}\,X_t^2)/T = 0$ and either*

(a) X_t is strictly stationary or
(b) $\max_{t\in[1,T]} \frac{(\mathrm{E}|X_t|^{2+\delta})^{1/(2+\delta)}}{\Sigma_{t=1}^T \mathrm{E}\,X_t^2/T} < \infty$ for $\delta > 0$ and all $T > 1$,

then $(\Sigma_{t=1}^T X_t/\sqrt{T})/(\Sigma_{t=1}^T \mathrm{E}\,X_t^2/T)^{1/2} \xrightarrow{d} N(0,1)$. (See Davidson (2000) 6.2)

31.7 Special Distributions

31.7.1 The Normal Distribution

Fact 31.56 *(Univariate normal distribution) If $X \sim N(\mu, \sigma^2)$, then the probability density function of X, $f(x)$ is*

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}.$$

The moment generating function is $mgf_X(t) = \exp\left(\mu t + \sigma^2 t^2/2\right)$ and the moment generating function around the mean is $mgf_{(X-\mu)}(t) = \exp\left(\sigma^2 t^2/2\right)$.

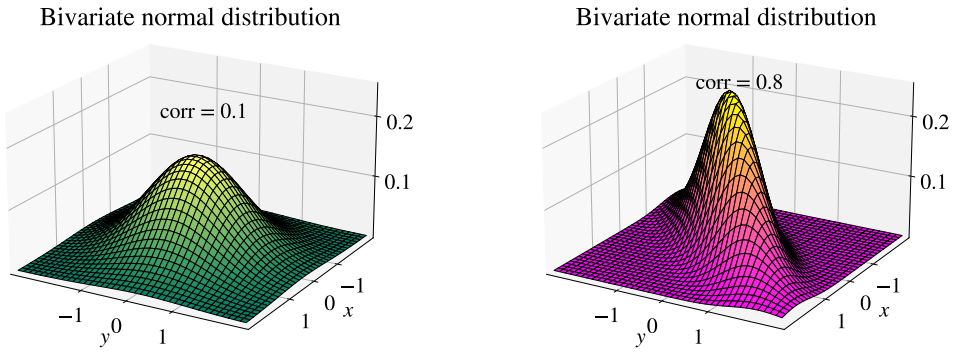


Figure 31.2: Normal distributions

Example 31.57 *The first few moments around the mean are* $\mathrm{E}(X-\mu)=0$, $\mathrm{E}(X-\mu)^2=\sigma^2$, $\mathrm{E}(X-\mu)^3=0$ *(all odd moments are zero),* $\mathrm{E}(X-\mu)^4=3\sigma^4$, $\mathrm{E}(X-\mu)^6=15\sigma^6$, *and* $\mathrm{E}(X-\mu)^8=105\sigma^8$. *More generally, for even n, we have* $\mathrm{E}\,(X-\mu)^n=\sigma^2(n-1)!!$ *where* $(n-1)!!$ *is the product of all odd numbers up to and including* $n-1$, $(n-1)\times(n-3)\times\ldots\times 3\times 1$.

Fact 31.58 *(Standard normal distribution) If* $X\sim N(0,1)$, *then the moment generating function is* $mgf_X(t)=\exp\left(t^2/2\right)$. *Since the mean is zero,* $m(t)$ *gives central moments. The first few are* $\mathrm{E}\,X=0$, $\mathrm{E}\,X^2=1$, $\mathrm{E}\,X^3=0$ *(all odd moments are zero), and* $EX^4=3$. *The distribution function,* $\Pr(X\leq a)=\Phi(a)=1/2+1/2\,\mathrm{erf}(a/\sqrt{2})$, *where* erf() *is the error function,* $\mathrm{erf}(z)=\frac{2}{\sqrt{\pi}}\int_0^z\exp(-t^2)dt$*). The complementary error function is* $\mathrm{erfc}(z)=1-\mathrm{erf}(z)$. *Since the distribution is symmetric around zero, we have* $\Phi(-a)=\Pr(X\leq -a)=\Pr(X\geq a)=1-\Phi(a)$. *Clearly,* $1-\Phi(-a)=\Phi(a)=1/2\,\mathrm{erfc}(-a/\sqrt{2})$. *This latter is often a better method for calculating probabilities in the far left tail.*

Fact 31.59 *(Multivariate normal distribution) If* X *is an* $n\times 1$ *vector of random variables with a multivariate normal distribution, with a mean vector* μ *and variance-covariance matrix* Σ, $N(\mu,\Sigma)$, *then the density function is*

$$f(x)=\frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}}\exp\left[-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right].$$

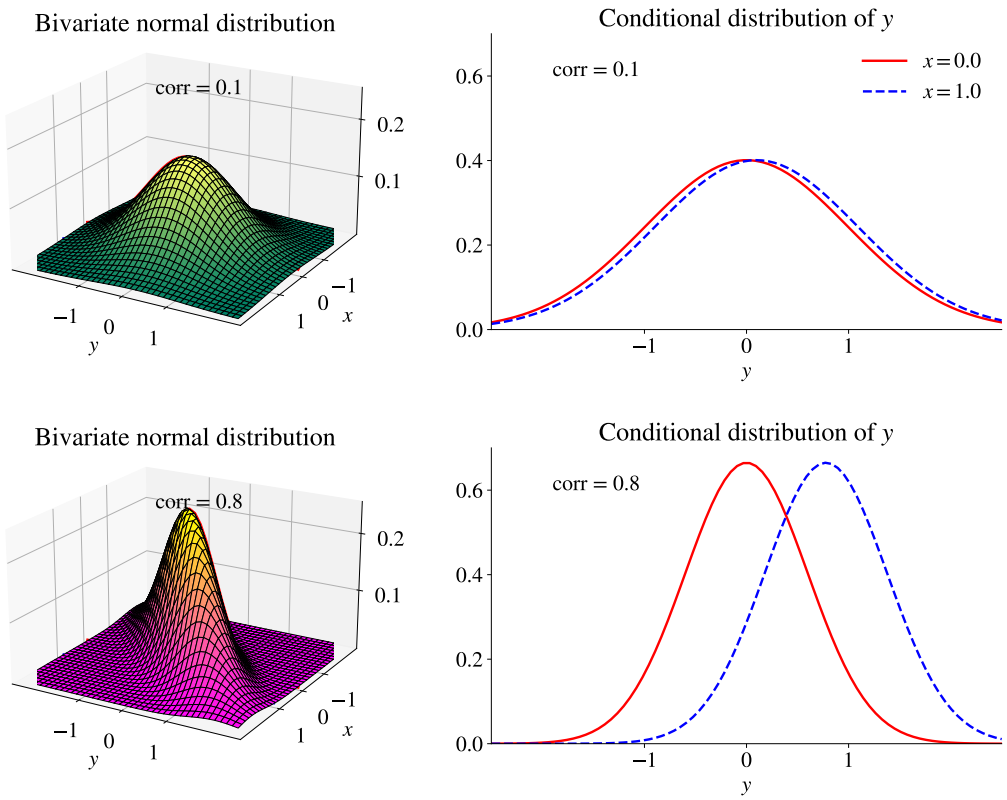


Figure 31.3: Density functions of normal distributions

Fact 31.60 *(Conditional normal distribution) Suppose* $Z_{m\times 1}$ *and* $X_{n\times 1}$ *are jointly normally distributed*

$$\begin{bmatrix} Z \\ X \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_Z \\ \mu_X \end{bmatrix}, \begin{bmatrix} \Sigma_{ZZ} & \Sigma_{ZX} \\ \Sigma_{XZ} & \Sigma_{XX} \end{bmatrix}\right).$$

The distribution of the random variable Z *conditional on that* $X = x$ *(a number) is also normal with mean*

$$\mathrm{E}\left(Z|x\right) = \mu_Z + \Sigma_{ZX}\Sigma_{XX}^{-1}\left(x - \mu_X\right),$$

and variance (variance of Z *conditional on that* $X = x$*, that is, the variance of the prediction error* $Z - \mathrm{E}\left(Z|x\right)$*)*

$$\mathrm{Var}\left(Z|x\right) = \Sigma_{ZZ} - \Sigma_{ZX}\Sigma_{XX}^{-1}\Sigma_{XZ}.$$

Note that the conditional variance is constant in the multivariate normal distribution ($\mathrm{Var}\left(Z|X\right)$ *is not a random variable in this case). Note also that* $\mathrm{Var}\left(Z|x\right)$ *is less than*

$\text{Var}(Z) = \Sigma_{ZZ}$ *(in a matrix sense) if* X *contains any relevant information (so* Σ_{ZX} *is not zero, that is,* $\text{E}(Z|x)$ *is not the same for all* x*).*

Example 31.61 *(Conditional normal distribution) Suppose* Z *and* X *are scalars in Fact 31.60 and that the joint distribution is*

$$\begin{bmatrix} Z \\ X \end{bmatrix} \sim N\left(\begin{bmatrix} 3 \\ 5 \end{bmatrix}, \begin{bmatrix} 1 & 2 \\ 2 & 6 \end{bmatrix}\right).$$

The expectation of Z *conditional on* $X = x$ *is then*

$$\text{E}(Z|x) = 3 + \frac{2}{6}(x-5) = 3 + \frac{1}{3}(x-5).$$

Similarly, the conditional variance is

$$\text{Var}(Z|x) = 1 - \frac{2 \times 2}{6} = \frac{1}{3}.$$

Fact 31.62 *(Stein's lemma) If* Y *has normal distribution and* $h()$ *is a differentiable function such that* $\text{E}|h'(Y)| < \infty$*, then* $\text{Cov}[Y, h(Y)] = \text{Var}(Y)\,\text{E}\,h'(Y)$.

Proof. $\text{E}[(Y-\mu)h(Y)] = \int_{-\infty}^{\infty}(Y-\mu)h(Y)\phi(Y;\mu,\sigma^2)dY$, where $\phi(Y;\mu,\sigma^2)$ is the pdf of $N(\mu,\sigma^2)$. Note that $d\phi(Y;\mu,\sigma^2)/dY = -\phi(Y;\mu,\sigma^2)(Y-\mu)/\sigma^2$, so the integral can be rewritten as $-\sigma^2\int_{-\infty}^{\infty}h(Y)d\phi(Y;\mu,\sigma^2)$. Integration by parts ("$\int udv = uv - \int vdu$") gives $-\sigma^2\left[h(Y)\phi(Y;\mu,\sigma^2)\big|_{-\infty}^{\infty} - \int_{-\infty}^{\infty}\phi(Y;\mu,\sigma^2)h'(Y)dY\right] = \sigma^2\,\text{E}\,h'(Y)$. ■

Fact 31.63 *(Stein's lemma 2) It follows from Fact 31.62 that if* X *and* Y *have a bivariate normal distribution and* $h()$ *is a differentiable function such that* $\text{E}|h'(Y)| < \infty$*, then* $\text{Cov}[X, h(Y)] = \text{Cov}(X,Y)\,\text{E}\,h'(Y)$.

Example 31.64 *(a) With* $h(Y) = \exp(Y)$ *we get* $\text{Cov}[X, \exp(Y)] = \text{Cov}(X,Y)\,\text{E}\exp(Y)$*; (b) with* $h(Y) = Y^2$ *we get* $\text{Cov}[X, Y^2] = \text{Cov}(X,Y)2\,\text{E}\,Y$ *so with* $\text{E}\,Y = 0$ *we get a zero covariance.*

Fact 31.65 *(Stein's lemma 3) Fact 31.63 still holds if the joint distribution of* X *and* Y *is a mixture of* n *bivariate normal distributions, provided the mean and variance of* Y *is the same in each of the* n *components. (See Söderlind (2009) for a proof.)*

Fact 31.66 *(Truncated normal distribution) Let* $X \sim N(\mu,\sigma^2)$*, and consider truncating the distribution so that we want moments conditional on* $a < X \leq b$*. Define* $a_0 =$

$(a-\mu)/\sigma$ *and* $b_0 = (b-\mu)/\sigma$ *and let* $\phi()$ *be the pdf and* $\Phi()$ *the cdf of a* $N(0,1)$ *variable. Then,*

$$\mathrm{E}(X|a < X \leq b) = \mu - \sigma \frac{\phi(b_0) - \phi(a_0)}{\Phi(b_0) - \Phi(a_0)} \text{ and}$$
$$\mathrm{Var}(X|a < X \leq b) = \sigma^2 \left\{ 1 - \frac{b_0\phi(b_0) - a_0\phi(a_0)}{\Phi(b_0) - \Phi(a_0)} - \left[\frac{\phi(b_0) - \phi(a_0)}{\Phi(b_0) - \Phi(a_0)} \right]^2 \right\}.$$

Fact 31.67 *(Lower truncation) In Fact 31.66, let* $b \to \infty$*, so we only have the truncation* $a < X$*. Then, we have*

$$\mathrm{E}(X|a < X) = \mu + \sigma \frac{\phi(a_0)}{1 - \Phi(a_0)} \text{ and}$$
$$\mathrm{Var}(X|a < X) = \sigma^2 \left\{ 1 + \frac{a_0\phi(a_0)}{1 - \Phi(a_0)} - \left[\frac{\phi(a_0)}{1 - \Phi(a_0)} \right]^2 \right\}.$$

(The latter follows from $\lim_{b\to\infty} b_0\phi(b_0) = 0$*.)*

Example 31.68 *Suppose* $X \sim N(0, \sigma^2)$ *and we want to calculate* $\mathrm{E}|x|$*. This is the same as* $\mathrm{E}(X|X > 0) = 2\sigma\phi(0)$*.*

Fact 31.69 *(Upper truncation) In Fact 31.66, let* $a \to -\infty$*, so we only have the truncation* $X \leq b$*. Then, we have*

$$\mathrm{E}(X|X \leq b) = \mu - \sigma \frac{\phi(b_0)}{\Phi(b_0)} \text{ and}$$
$$\mathrm{Var}(X|X \leq b) = \sigma^2 \left\{ 1 - \frac{b_0\phi(b_0)}{\Phi(b_0)} - \left[\frac{\phi(b_0)}{\Phi(b_0)} \right]^2 \right\}.$$

(The latter follows from $\lim_{a\to-\infty} a_0\phi(a_0) = 0$*.)*

Fact 31.70 *(Delta method) Consider an estimator* $\hat{\beta}_{k\times 1}$ *which satisfies*

$$\sqrt{T}\left(\hat{\beta} - \beta_0\right) \xrightarrow{d} N(0, \Omega),$$

and suppose we want the asymptotic distribution of a transformation of β

$$\gamma_{q\times 1} = g(\beta),$$

where $g(.)$ is has continuous first derivatives. The result is

$$\sqrt{T}\left[g\left(\hat{\beta}\right)-g\left(\beta_0\right)\right] \xrightarrow{d} N\left(0,\Psi_{q\times q}\right), \text{ where}$$
$$\Psi=\frac{\partial g\left(\beta_0\right)}{\partial \beta'}\Omega\frac{\partial g\left(\beta_0\right)'}{\partial \beta}, \text{ where } \frac{\partial g\left(\beta_0\right)}{\partial \beta'} \text{ is } q\times k.$$

Proof. By the mean value theorem we have

$$g\left(\hat{\beta}\right)=g\left(\beta_0\right)+\frac{\partial g\left(\beta^*\right)}{\partial \beta'}\left(\hat{\beta}-\beta_0\right),$$

where

$$\frac{\partial g\left(\beta\right)}{\partial \beta'}=\begin{bmatrix}\frac{\partial g_1(\beta)}{\partial \beta_1} & \cdots & \frac{\partial g_1(\beta)}{\partial \beta_k}\\ \vdots & \ddots & \vdots \\ \frac{\partial g_q(\beta)}{\partial \beta_1} & \cdots & \frac{\partial g_q(\beta)}{\partial \beta_k}\end{bmatrix}_{q\times k},$$

and we evaluate it at β^* which is (weakly) between $\hat{\beta}$ and β_0. Premultiply by $\sqrt{T}$ and rearrange as

$$\sqrt{T}\left[g\left(\hat{\beta}\right)-g\left(\beta_0\right)\right]=\frac{\partial g\left(\beta^*\right)}{\partial \beta'}\sqrt{T}\left(\hat{\beta}-\beta_0\right).$$

If $\hat{\beta}$ is consistent ($\operatorname{plim}\hat{\beta}=\beta_0$) and $\partial g\left(\beta^*\right)/\partial\beta'$ is continuous, then by Slutsky's theorem $\operatorname{plim}\partial g\left(\beta^*\right)/\partial\beta'=\partial g\left(\beta_0\right)/\partial\beta'$, which is a constant. The result then follows from the continuous mapping theorem. ■

31.7.2 The Lognormal Distribution

Fact 31.71 *(Univariate lognormal distribution) If $x\sim N(\mu,\sigma^2)$ and $y=\exp(x)$ then the probability density function of y, $f(y)$ is*

$$f(y)=\frac{1}{y\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2}\left(\frac{\ln y-\mu}{\sigma}\right)^2},\ y>0.$$

The rth moment of y is $\mathrm{E}\,y^r=\exp(r\mu+r^2\sigma^2/2)$. See 31.4 for an illustration.

Example 31.72 *The first two moments are $\mathrm{E}\,y=\exp\left(\mu+\sigma^2/2\right)$ and $\mathrm{E}\,y^2=\exp(2\mu+2\sigma^2)$. We therefore get $\operatorname{Var}(y)=\exp\left(2\mu+\sigma^2\right)\left[\exp\left(\sigma^2\right)-1\right]$ and $\operatorname{Std}(y)/\mathrm{E}\,y=\sqrt{\exp(\sigma^2)-1}$.*

Fact 31.73 *(Moments of a truncated lognormal distribution) If $x\sim N(\mu,\sigma^2)$ and $y=\exp(x)$ then $\mathrm{E}(y^r|y>a)=\mathrm{E}(y^r)\Phi(r\sigma-a_0)/\Phi(-a_0)$, where $a_0=(\ln a-\mu)/\sigma$ and*

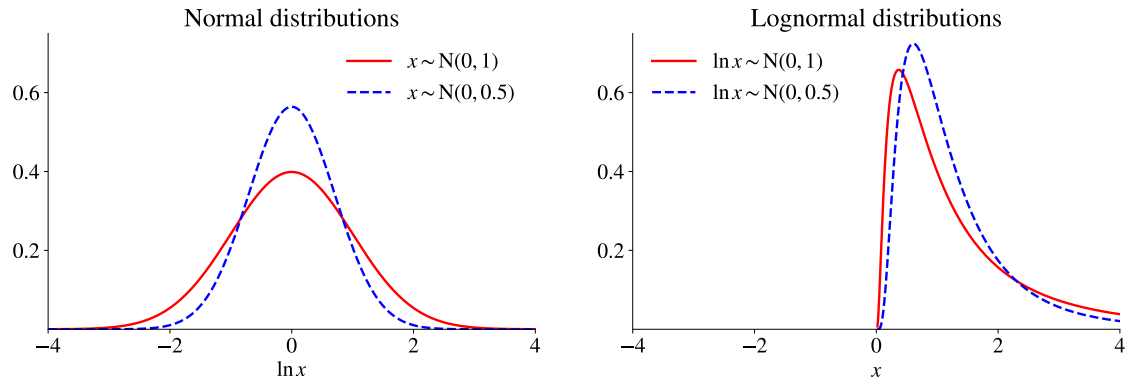


Figure 31.4: Lognormal distribution

where $\phi()$ *is the pdf and* $\Phi()$ *the cdf of a* $N(0,1)$ *variable. Notice that the denominator is* $\Pr(y > a) = \Phi(-a_0)$. *In contrast,* $\mathrm{E}(y^r|y \leq b) = \mathrm{E}(y^r)\Phi(-r\sigma + b_0)/\Phi(b_0)$, *where* $b_0 = (\ln b - \mu)/\sigma$. *The denominator is* $\Pr(y \leq b) = \Phi(b_0)$. *Clearly,* $\mathrm{E}(y^r) = \exp(r\mu + r^2\sigma^2/2)$

Fact 31.74 *(Moments of a truncated lognormal distribution, two-sided truncation) If* $x \sim N(\mu, \sigma^2)$ *and* $y = \exp(x)$ *then*

$$\mathrm{E}(y^r|a > y < b) = \mathrm{E}(y^r)\frac{\Phi(r\sigma - a_0) - \Phi(r\sigma - b_0)}{\Phi(b_0) - \Phi(a_0)},$$

where $a_0 = (\ln a - \mu)/\sigma$ *and* $b_0 = (\ln b - \mu)/\sigma$. *Note that the denominator is* $\Pr(a > y < b) = \Phi(b_0) - \Phi(a_0)$. *Clearly,* $\mathrm{E}(y^r) = \exp(r\mu + r^2\sigma^2/2)$.

Example 31.75 *The first two moments of the truncated (from below) lognormal distribution are* $\mathrm{E}(y|y > a) = \exp\left(\mu + \sigma^2/2\right)\Phi(\sigma - a_0)/\Phi(-a_0)$ *and* $\mathrm{E}(y^2|y > a) = \exp\left(2\mu + 2\sigma^2\right)\Phi(2\sigma - a_0)/\Phi(-a_0)$.

Example 31.76 *The first two moments of the truncated (from above) lognormal distribution are* $\mathrm{E}(y|y \leq b) = \exp\left(\mu + \sigma^2/2\right)\Phi(-\sigma + b_0)/\Phi(b_0)$ *and* $\mathrm{E}(y^2|y \leq b) = \exp\left(2\mu + 2\sigma^2\right)\Phi(-2\sigma + b_0)/\Phi(b_0)$.

Fact 31.77 *(Multivariate lognormal distribution) Let the* $n \times 1$ *vector* x *have a mulivariate normal distribution*

$$x \sim N(\mu, \Sigma), \text{ where } \mu = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_n \end{bmatrix} \text{ and } \Sigma = \begin{bmatrix} \sigma_{11} & \cdots & \sigma_{1n} \\ \vdots & \ddots & \vdots \\ \sigma_{n1} & \cdots & \sigma_{nn} \end{bmatrix}.$$

Then $y = \exp(x)$ *has a lognormal distribution, with the means and covariances*

$$\begin{aligned}
\mathrm{E}\, y_i &= \exp\left(\mu_i + \sigma_{ii}/2\right) \\
\mathrm{Cov}(y_i, y_j) &= \exp\left[\mu_i + \mu_j + (\sigma_{ii} + \sigma_{jj})/2\right]\left[\exp(\sigma_{ij}) - 1\right] \\
\mathrm{Corr}(y_i, y_j) &= \left[\exp(\sigma_{ij}) - 1\right] / \sqrt{\left[\exp(\sigma_{ii}) - 1\right]\left[\exp(\sigma_{jj}) - 1\right]}.
\end{aligned}$$

Cleary, $\mathrm{Var}(y_i) = \exp\left[2\mu_i + \sigma_{ii}\right]\left[\exp(\sigma_{ii}) - 1\right]$. $\mathrm{Cov}(y_1, y_2)$ *and* $\mathrm{Corr}(y_1, y_2)$ *have the same sign as* $\mathrm{Corr}(x_i, x_j)$ *and are increasing in it. However,* $\mathrm{Corr}(y_i, y_j)$ *is closer to zero.*

31.7.3 The Chi-Square Distribution

Fact 31.78 *(The* χ^2_n *distribution) If* $Y \sim \chi^2_n$*, then the pdf of* Y *is* $f(y) = \frac{1}{2^{n/2}\Gamma(n/2)} y^{n/2-1} e^{-y/2}$*, where* $\Gamma()$ *is the gamma function. The moment generating function is* $mgf_Y(t) = (1 - 2t)^{-n/2}$ *for* $t < 1/2$*. The first moments of* Y *are* $\mathrm{E}\, Y = n$ *and* $\mathrm{Var}(Y) = 2n$*.*

Fact 31.79 *(Quadratic forms of normally distribution random variables) If the* $n \times 1$ *vector* $X \sim N(0, \Sigma)$*, then* $Y = X'\Sigma^{-1}X \sim \chi^2_n$*. Therefore, if the* n *scalar random variables* X_i*,* $i = 1, ..., n$*, are uncorrelated and have the distributions* $N(0, \sigma_i^2)$*,* $i = 1, ..., n$*, then* $Y = \Sigma_{i=1}^n X_i^2/\sigma_i^2 \sim \chi^2_n$*.*

Fact 31.80 *(Distribution of* $X'AX$*) If the* $n \times 1$ *vector* $X \sim N(\mathbf{0}, I)$*, and* A *is a symmetric idempotent matrix (*$A = A'$ *and* $A = AA = A'A = AA'$*) of rank* r*, then* $Y = X'AX \sim \chi^2_r$*.*

Fact 31.81 *(Distribution of* $X'\Sigma^+X$*) If the* $n \times 1$ *vector* $X \sim N(0, \Sigma)$*, where* Σ *has rank* $r \leq n$ *then* $Y = X'\Sigma^+X \sim \chi^2_r$ *where* Σ^+ *is the pseudo inverse of* Σ*.*

Proof. Σ is symmetric, so it can be decomposed as $\Sigma = C\Lambda C'$ where C are the orthogonal eigenvectors ($C'C = I$) and Λ is a diagonal matrix with the eigenvalues along the main diagonal. We therefore have $\Sigma = C\Lambda C' = C_1\Lambda_{11}C_1'$ where C_1 is an $n \times r$ matrix associated with the r non-zero eigenvalues (found in the $r \times r$ matrix Λ_{11}). The generalized inverse can be shown to be

$$\Sigma^+ = \begin{bmatrix} C_1 & C_2 \end{bmatrix} \begin{bmatrix} \Lambda_{11}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} C_1 & C_2 \end{bmatrix}' = C_1\Lambda_{11}^{-1}C_1',$$

We can write $\Sigma^+ = C_1 \Lambda_{11}^{-1/2} \Lambda_{11}^{-1/2} C_1'$. Consider the $r \times 1$ vector $Z = \Lambda_{11}^{-1/2} C_1' X$, and note that it has the covariance matrix

$$\mathrm{E}\, ZZ' = \Lambda_{11}^{-1/2} C_1' \,\mathrm{E}\, XX' C_1 \Lambda_{11}^{-1/2} = \Lambda_{11}^{-1/2} C_1' C_1 \Lambda_{11} C_1' C_1 \Lambda_{11}^{-1/2} = I_r,$$

since $C_1' C_1 = I_r$. This shows that $Z \sim N(\mathbf{0}_{r\times 1}, I_r)$, so $Z'Z = X'\Sigma^+ X \sim \chi_r^2$. ■

Fact 31.82 *(Convergence to a normal distribution) Let $Y \sim \chi_n^2$ and $Z = (Y - n)/n^{1/2}$. Then $Z \xrightarrow{d} N(0, 2)$.*

Example 31.83 *If $Y = \Sigma_{i=1}^n X_i^2/\sigma_i^2$, then this transformation means $Z = (\Sigma_{i=1}^n X_i^2/\sigma_i^2 - 1)/n^{1/2}$.*

Proof. We can directly note from the moments of a χ_n^2 variable that $\mathrm{E}\, Z = (\mathrm{E}\, Y - n)/n^{1/2} = 0$, and $\mathrm{Var}(Z) = \mathrm{Var}(Y)/n = 2$. From the general properties of moment generating functions, we note that the moment generating function of Z is

$$mgf_Z(t) = e^{-t\sqrt{n}} \left(1 - 2\frac{t}{n^{1/2}}\right)^{-n/2} \text{ with } \lim_{n\to\infty} mgf_Z(t) = \exp(t^2).$$

This is the moment generating function of a $N(0, 2)$ distribution, which shows that $Z \xrightarrow{d} N(0, 2)$. This result should not come as a surprise as we can think of Y as the sum of n variables; dividing by $n^{1/2}$ is then like creating a scaled sample average for which a central limit theorem applies. ■

Fact 31.84 *(Non-central Chi-square) If the $n \times 1$ vector $X \sim N(\mu, \Sigma)$, then $Y = X'\Sigma^{-1}X \sim \chi_n^2(\lambda)$ where $\lambda = \mu'\Sigma^{-1}\mu$. This is a non-central Chi-square distribution with n degrees of freedom and the non-centrality parameter λ. (Warning: some authors instead define λ as $\mu'\Sigma^{-1}\mu/2$.) If $\lambda = 0$, then it is the same as a χ_n^2 distribution.*

31.7.4 The t and F Distributions

Fact 31.85 *(The $F(n_1, n_2)$ distribution) If $Y_1 \sim \chi_{n_1}^2$ and $Y_2 \sim \chi_{n_2}^2$ and Y_1 and Y_2 are independent, then $Z = (Y_1/n_1)/(Y_2/n_2)$ has an $F(n_1, n_2)$ distribution. This distribution has no moment generating function, but $\mathrm{E}\, Z = n_2/(n_2 - 2)$ for $n > 2$.*

Fact 31.86 *(Convergence of an $F(n_1, n_2)$ distribution) In Fact (31.85), the distribution of $n_1 Z = Y_1/(Y_2/n_2)$ converges to a $\chi_{n_1}^2$ distribution as $n_2 \to \infty$. (The idea is essentially*

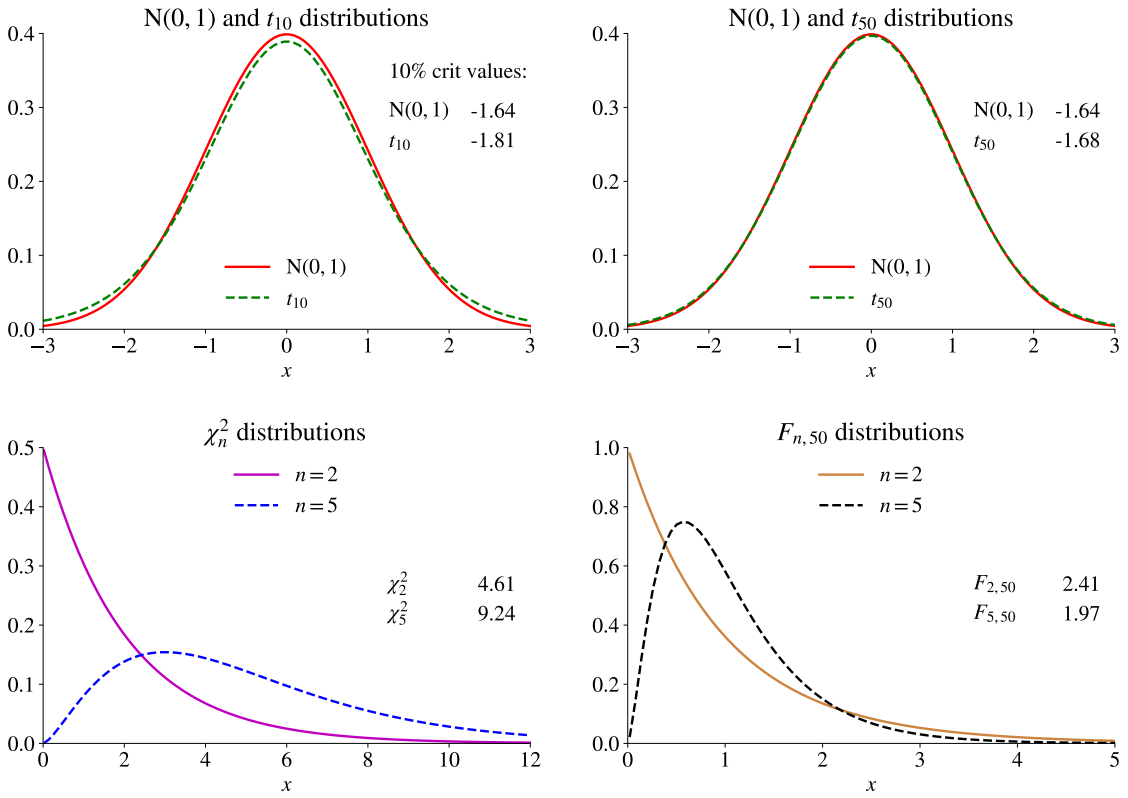


Figure 31.5: t, χ^2, and F distributions

that $n_2 \to \infty$ the denominator converges to the mean, which is $\mathrm{E}\, Y_2/n_2 = 1$. Only the numerator is then left, which is a $\chi^2_{n_1}$ variable.)

Fact 31.87 *(The t_n distribution) If $X \sim N(0,1)$ and $Y \sim \chi^2_n$ and X and Y are independent, then $Z = X/(Y/n)^{1/2}$ has a t_n distribution. The moment generating function does not exist, but $\mathrm{E}\, Z = 0$ for $n > 1$ and $Var(Z) = n/(n-2)$ for $n > 2$.*

Fact 31.88 *(Convergence of a t_n distribution) The t distribution converges to a $N(0,1)$ distribution as $n \to \infty$.*

Fact 31.89 *(t_n versus $F(1,n)$ distribution) If $Z \sim t_n$, then $Z^2 \sim F(1,n)$.*

31.7.5 The Bernouilli and Binomial Distributions

Fact 31.90 *(Bernoulli distribution) The random variable X can only take two values: 1 or 0, with probability p and $1-p$ respectively. The moment generating function is $mgf(t) = pe^t + 1 - p$. This gives $\mathrm{E}(X) = p$ and $\mathrm{Var}(X) = p(1-p)$.*

Example 31.91 *(Shifted Bernoulli distribution) Suppose the Bernoulli variable takes the values a or b (instead of 1 and 0) with probability p and* $1-p$ *respectively. Then* $\mathrm{E}(X) = pa + (1-p)b$ *and* $\mathrm{Var}(X) = p(1-p)(a-b)^2$.

Fact 31.92 *(Binomial distribution). Suppose* $X_1, X_2, ..., X_n$ *all have Bernoulli distributions with the parameter p. Then, the sum* $Y = X_1 + X_2 + ... + X_n$ *has a Binomial distribution with parameters p and n. The pdf is* $pdf(Y) = n!/[y!(n-y)!]p^y(1-p)^{n-y}$ *for* $y = 0, 1, ..., n$. *The moment generating function is* $mgf(t) = [pe^t + 1 - p]^n$. *This gives* $\mathrm{E}(Y) = np$ *and* $\mathrm{Var}(Y) = np(1-p)$.

Example 31.93 *(Shifted Binomial distribution) Suppose the Bernuolli variables* $X_1, X_2, ..., X_n$ *take the values a or b (instead of 1 and 0) with probability p and* $1-p$ *respectively. Then, the sum* $Y = X_1 + X_2 + ... + X_n$ *has* $\mathrm{E}(Y) = n[pa + (1-p)b]$ *and* $\mathrm{Var}(Y) = n[p(1-p)(a-b)^2]$.

31.7.6 The Skew-Normal Distribution

Fact 31.94 *(Skew-normal distribution) Let* ϕ *and* Φ *be the standard normal pdf and cdf respectively. The pdf of a skew-normal distribution with shape parameter* α *is then*

$$f(z) = 2\phi(z)\Phi(\alpha z).$$

If Z has the above pdf and

$$Y = \mu + \omega Z \text{ with } \omega > 0,$$

then Y is said to have a $SN(\mu, \omega^2, \alpha)$ *distribution (see Azzalini (2005)). Clearly, the pdf of Y is*

$$f(y) = 2\phi\left[(y-\mu)/\omega\right]\Phi\left[\alpha\left(y-\mu\right)/\omega\right]/\omega.$$

The moment generating function is $mgf_y(t) = 2\exp\left(\mu t + \omega^2 t^2/2\right)\Phi(\delta\omega t)$ *where* $\delta = \alpha/\sqrt{1+\alpha^2}$. *When* $\alpha > 0$ *then the distribution is positively skewed (and vice versa)—and when* $\alpha = 0$ *the distribution becomes a normal distribution. When* $\alpha \to \infty$, *then the density function is zero for* $Y \leq \mu$, *and* $2\phi\left[(y-\mu)/\omega\right]/\omega$ *otherwise—this is a half-normal distribution.*

Example 31.95 *The first three moments are as follows. First, notice that* $\mathrm{E}\,Z = \sqrt{2/\pi}\delta$,

$\text{Var}(Z) = 1 - 2\delta^2/\pi$ *and* $\text{E}(Z - \text{E}\,Z)^3 = (4/\pi - 1)\sqrt{2/\pi}\delta^3$. *Then we have*

$$\begin{aligned}
\text{E}\,Y &= \mu + \omega\,\text{E}\,Z \\
\text{Var}(Y) &= \omega^2\,\text{Var}(Z) \\
\text{E}\,(Y - \text{E}\,Y)^3 &= \omega^3\,\text{E}(Z - \text{E}\,Z)^3.
\end{aligned}$$

Notice that with $\alpha = 0$ *(so* $\delta = 0$*), then these moments of* Y *become* μ, ω^2 *and* 0 *respecively.*

31.7.7 Generalized Pareto Distribution

Fact 31.96 *(Cdf and pdf of the generalized Pareto distribution) The generalized Pareto distribution is described by a scale parameter (*$\beta > 0$*) and a shape parameter (*ξ*). The cdf (*$\Pr(Z \leq z)$*, where* Z *is the random variable and* z *is a value) is*

$$G(z) = \begin{cases} 1 - (1 + \xi z/\beta)^{-1/\xi} & \text{if } \xi \neq 0 \\ 1 - \exp(-z/\beta) & \xi = 0, \end{cases}$$

for $0 \leq z$ *and* $0 \leq z \leq -\beta/\xi$ *in case* $\xi < 0$. *The pdf is therefore*

$$g(z) = \begin{cases} \frac{1}{\beta}(1 + \xi z/\beta)^{-1/\xi - 1} & \text{if } \xi \neq 0 \\ \frac{1}{\beta}\exp(-z/\beta) & \xi = 0. \end{cases}$$

The mean is defined (finite) if $\xi < 1$ *and is then* $\text{E}(z) = \beta/(1 - \xi)$, *the median is* $(2^\xi - 1)\beta/\xi$ *and the variance is defined if* $\xi < 1/2$ *and is then* $\beta^2/[(1 - \xi)^2(1 - 2\xi)]$. *To include also a "location" parameter (*μ*), substitute* $x - \mu$ *for* z.

31.7.8 Uniform Distribution

Fact 31.97 *(Cdf and pdf of a uniform distribution on the interval* $a \leq X \leq b$*). The cdf (*G*) and pdf (*g*) are (between* a *and* b*)*

$$\begin{aligned}
G(x) &= \frac{x - a}{b - a} \\
g(x) &= \frac{1}{b - a}.
\end{aligned}$$

The cdf is zero below a and unity above b. The pdf is zero outside the interval $[a, b]$. The first three central moments are

$$\mathrm{E}\, X = (a+b)/2, \mathrm{Var}(X) = (b-a)^2/12 \text{ and } \mathrm{E}\,(X - \mathrm{E}\, X)^3 = 0.$$

31.7.9 Mixture Distributions

Fact 31.98 *(Pdf of mixture distribution) Let $f_i(x)$ be a pdf for some continuous distribution and $0 \leq \pi_i \leq 1$ be a constant. A mixture distribution with n components has the pdf $f(x) = \Sigma_{i=1}^n \pi_i f_i(x)$, where $\Sigma_{i=1}^n \pi_i = 1$.*

Fact 31.99 *(Uncentered moments of mixture distribution) The jth uncentered moment of X is $\mathrm{E}\, X^j = \int_{-\infty}^{\infty} f(q) q^j dq$. By the definition of the pdf, we can write this as $\mathrm{E}\, X^j = \Sigma_{i=1}^n \pi_i \mathrm{E}\, X_i^j$, where $\mathrm{E}\, X_i^j$ is the jth uncentered moment of the ith component. (To see this, notice that the integral can be written $\int_{-\infty}^{\infty} \Sigma_{i=1}^n \pi_i f_i(q) q^j dq$ and that we can switch the order of the summation and the integration.)*

Fact 31.100 *(Centered moments of mixture distribution) To calculate the jth centered moment, first calculate the first j uncentered moments (the $\mathrm{E}\, X^i$ for $i = 1, 2, .., j$) and define the grand mean as $\mu = \Sigma_{i=1}^n \pi_i \mathrm{E}\, X_i$. Then use the binomial theorem $\mathrm{E}(X-\mu)^j = \Sigma_{k=0}^j \binom{j}{k} \mathrm{E}\, X^{j-k}(-\mu)^k$, where $\mathrm{E}\, X^{j-k}$ denotes the $(j-k)$th uncentered moment.*

Example 31.101 *(Centered moments) For $j = 1$ we get $\mathrm{E}(X-\mu) = 0$, for $j = 2$ we get $\mathrm{E}(X-\mu)^2 = \mathrm{E}\, X^2 - \mu^2$ and for $j = 3$ we get $\mathrm{E}(X-\mu)^3 = \mathrm{E}\, X^3 - 3\,\mathrm{E}\, X^2 \mu + 2\mu^3$.*

31.8 Inference

Fact 31.102 *(Comparing variance-covariance matrices) Let $\mathrm{Var}(\hat{\beta})$ and $\mathrm{Var}(\beta^*)$ be the variance-covariance matrices of two estimators, $\hat{\beta}$ and β^*, and suppose $\mathrm{Var}(\hat{\beta}) - \mathrm{Var}(\beta^*)$ is a positive semi-definite matrix. This means that for any non-zero vector R that $R' \mathrm{Var}(\hat{\beta}) R \geq R' \mathrm{Var}(\beta^*) R$, so every linear combination of $\hat{\beta}$ has a variance that is as large as the variance of the same linear combination of β^*. In particular, this means that the variance of every element in $\hat{\beta}$ (the diagonal elements of $\mathrm{Var}(\hat{\beta})$) is at least as large as variance of the corresponding element of β^*.*

Fact 31.103 *(The Bonferroni inequality) Suppose we perform $i = 1...n$ different tests, each at the significance level p_i. The Bonferroni inequality then says that if the null*

hypotheses are all true, then

$$\Pr(\textit{not rejecting in any of the } n \textit{ tests}) \geq 1 - \Sigma_{i=1}^{n} p_i.$$

It follows that rejecting in at least one of the n tests has a probability of less than or equal to $\Sigma_{i=1}^{n} p_i$. *For instance, with* $p_i = 0.05/n$, *there is 5% chance of rejecting in at least one test:* $\Pr(\textit{rejecting in at least one of the } n \textit{ tests}) \leq 0.05$.

Chapter 32

Some Facts about Matrices

Some references: Greene (2000), Golub and van Loan (1989), Björk (1996), Anton (1987), Greenberg (1988).

32.1 Rank

Fact 32.1 *(Submatrix) Any matrix obtained from the $m \times n$ matrix A by deleting at most $m-1$ rows and at most $n-1$ columns is a submatrix of A.*

Fact 32.2 *(Rank) The rank of the $m \times n$ matrix A is ρ if the largest submatrix with non-zero determinant is $\rho \times \rho$. The number of linearly independent row vectors (and column vectors) of A is then ρ.*

32.2 Vector Norms

Fact 32.3 *(Vector p-norm) Let x be an $n \times 1$ matrix. The p-norm is defined as/*

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}.$$

The Euclidian norm corresponds to $p = 2$

$$\|x\|_2 = \left(\sum_{i=1}^{n} x_i^2 \right)^{1/2} = \sqrt{x'x}.$$

32.3 Systems of Linear Equations and Matrix Inverses

Fact 32.4 *(Linear systems of equations) Consider the linear system $Ax = c$ where A is $m \times n$, x is $n \times 1$, and c is $m \times 1$. A solution is a vector x such that $Ax = c$. It has a unique solution if and only if* $\text{rank}(A) = \text{rank}([\ A \quad c\ ]) = n$*; an infinite number of solutions if and only if* $\text{rank}(A) = \text{rank}([\ A \quad c\ ]) < n$*; and no solution if and only if* $\text{rank}(A) \neq \text{rank}([\ A \quad c\ ])$.

Example 32.5 *(Linear systems of equations, unique solution when $m = n$) Let x be 2×1, and consider the linear system*

$$Ax = c \text{ with } A = \begin{bmatrix} 1 & 5 \\ 2 & 6 \end{bmatrix} \text{ and } c = \begin{bmatrix} 3 \\ 6 \end{bmatrix}.$$

Here $\text{rank}(A) = 2$ *and* $\text{rank}([\ A \quad c\ ]) = 2$*. The unique solution is* $x = [\ 3 \quad 0\ ]'$.

Example 32.6 *(Linear systems of equations, no solution when $m > n$) Let x be a scalar, and consider the linear system*

$$Ax = c \text{ with } A = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \text{ and } c = \begin{bmatrix} 3 \\ 7 \end{bmatrix}.$$

Here $\text{rank}(A) = 1$ *and* $\text{rank}([\ A \quad c\ ]) = 2$*. There is then no solution.*

Example 32.7 *(Inverse of 2×2 matrices). For a 2×2 matrix we have*

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}.$$

In particular, for a triangular matrix we have

$$\begin{bmatrix} a & 0 \\ c & d \end{bmatrix}^{-1} = \begin{bmatrix} 1/a & 0 \\ -c/(ad) & 1/d \end{bmatrix}.$$

Fact 32.8 *(Least squares) Suppose that no solution exists to $Ax = c$. The best approximate solution, in the sense of minimizing (the square root of) the sum of squared errors, $\left[(c - A\hat{x})'(c - A\hat{x})\right]^{1/2} = \|c - A\hat{x}\|_2$, is $\hat{x} = (A'A)^{-1} A'c$, provided the inverse exist.*

This is obviously the least squares solution. In the example with $c = [\ 3 \quad 7\ ]'$*, it is*

$$\begin{aligned}\hat{x} &= \left(\begin{bmatrix}1\\2\end{bmatrix}'\begin{bmatrix}1\\2\end{bmatrix}\right)^{-1}\begin{bmatrix}1\\2\end{bmatrix}'\begin{bmatrix}3\\7\end{bmatrix} \\ &= \frac{17}{5} \text{ or } 3.4.\end{aligned}$$

(Translation to OLS notation: c *is the vector of dependent variables for* m *observations,* A *is the matrix with explanatory variables with the* t^{th} *observation in row* t*, and* x *is the vector of parameters to estimate).*

Fact 32.9 *(Pseudo inverse or generalized inverse) Suppose that no solution exists to* $Ax = c$*, and that* $A'A$ *is not invertible. There are then several approximations,* $\hat{x}$*, which all minimize* $\|c - A\hat{x}\|_2$*. The one with the smallest* $\|\hat{x}\|_2$ *is given by* $\hat{x} = A^+c$*, where* A^+ *is the Moore-Penrose pseudo (generalized) inverse of* A*. See Fact 32.58.*

Example 32.10 *(Linear systems of equations, unique solution when* $m > n$*) Change* c *in Example 32.6 to* $c = [\ 3 \quad 6\ ]'$*. Then* $\text{rank}(A) = 1$ *and* $\text{rank}([\ A \quad c\ ]) = 1$*, and the unique solution is* $x = 3$*.*

Example 32.11 *(Linear systems of equations, infinite number of solutions,* $m < n$*) Let* x *be* 2×1*, and consider the linear system*

$$Ax = c \text{ with } A = \begin{bmatrix}1 & 2\end{bmatrix} \text{ and } c = 5.$$

Here $\text{rank}(A) = 1$ *and* $\text{rank}([\ A \quad c\ ]) = 1$*. Any value of* x_1 *on the line* $5 - 2x_2$ *is a solution.*

Example 32.12 *(Pseudo inverses again) In the previous example, there is an infinite number of solutions along the line* $x_1 = 5 - 2x_2$*. Which one has the smallest norm* $\|\hat{x}\|_2 = [(5 - 2x_2)^2 + x_2^2]^{1/2}$*? The first order condition gives* $x_2 = 2$*, and therefore* $x_1 = 1$*. This is the same value as given by* $\hat{x} = A^+c$*, since* $A^+ = [0.2, 0.4]$ *in this case.*

Fact 32.13 *(Rank and computers) Numerical calculations of the determinant are poor indicators of whether a matrix is singular or not. For instance,* $\det(0.1 \times I_{20}) = 10^{-20}$*. Use the condition number instead (see Fact 32.55).*

Fact 32.14 *(Some properties of inverses) If* A*,* B*, and* C *are invertible, then* $(ABC)^{-1} = C^{-1}B^{-1}A^{-1}$*;* $(A^{-1})' = (A')^{-1}$*; if* A *is symmetric, then* A^{-1} *is symmetric;* $(A^n)^{-1} = \left(A^{-1}\right)^n$*.*

Fact 32.15 *(Changing sign of column and inverting) Suppose the square matrix A_2 is the same as A_1 except that the i^{th} and j^{th} columns have the reverse signs. Then A_2^{-1} is the same as A_1^{-1} except that the i^{th} and j^{th} rows have the reverse sign.*

32.4 Complex matrices

Fact 32.16 *(Modulus of complex number) If $\lambda = a + bi$, where $i = \sqrt{-1}$, then $|\lambda| = |a + bi| = \sqrt{a^2 + b^2}$.*

Fact 32.17 *(Complex matrices) Let A^H denote the transpose of the complex conjugate of A, so that if*

$$A = \begin{bmatrix} 1 & 2+3i \end{bmatrix} \text{ then } A^H = \begin{bmatrix} 1 \\ 2-3i \end{bmatrix}.$$

A square matrix A is unitary *(similar to orthogonal) if $A^H = A^{-1}$, for instance,*

$$A = \begin{bmatrix} \frac{1+i}{2} & \frac{1+i}{2} \\ \frac{1-i}{2} & \frac{-1+i}{2} \end{bmatrix} \text{ gives } A^H = A^{-1} = \begin{bmatrix} \frac{1-i}{2} & \frac{1+i}{2} \\ \frac{1-i}{2} & \frac{-1-i}{2} \end{bmatrix}.$$

and it Hermitian *(similar to symmetric) if $A = A^H$, for instance*

$$A = \begin{bmatrix} \frac{1}{2} & \frac{1+i}{2} \\ \frac{1-i}{2} & \frac{-1}{2} \end{bmatrix}.$$

A Hermitian matrix has real elements along the principal diagonal and A_{ji} is the complex conjugate of A_{ij}. Moreover, the quadratic form $x^H Ax$ is always a real number.

32.5 Eigenvalues and Eigenvectors

Fact 32.18 *(Homogeneous linear system). Consider the linear system in Fact 32.4 with $c = \mathbf{0}$: $A_{m\times n}x_{n\times 1} = \mathbf{0}_{m\times 1}$. Then* $\text{rank}(A) = \text{rank}([\ A \quad c\ ])$*, so it has a unique solution if and only if* $\text{rank}(A) = n$*; and an infinite number of solutions if and only if* $\text{rank}(A) < n$*. Note that $x = \mathbf{0}$ is always a solution, and it is the unique solution if* $\text{rank}(A) = n$*. We can thus only get a nontrivial solution (not all elements are zero), only if* $\text{rank}\,(A) < n$*.*

Fact 32.19 *(Eigenvalues) The n eigenvalues, λ_i, $i = 1, \ldots, n$, and associated eigenvectors, z_i, of the $n \times n$ matrix A satisfy*

$$(A - \lambda_i I)\, z_i = \mathbf{0}_{n\times 1}.$$

We require the eigenvectors to be non-trivial (not all elements are zero). From Fact 32.18, an eigenvalue must therefore satisfy

$$\det(A - \lambda_i I) = 0.$$

Fact 32.20 *(Right and left eigenvectors) A "right eigenvector" z (the most common) satisfies $Az = \lambda z$, and a "left eigenvector" v (seldom used) satisfies $v'A = \lambda v'$, that is, $A'v = \lambda v$.*

Fact 32.21 *(Rank and eigenvalues) For any $m \times n$ matrix A,* $\text{rank}(A) = \text{rank}(A') = \text{rank}(A'A) = \text{rank}(AA')$ *and equals the number of non-zero eigenvalues of $A'A$ or AA'.*

Example 32.22 *Let x be an $n \times 1$ vector, so* $\text{rank}(x) = 1$. *We then have that the outer product, xx' also has rank 1.*

Fact 32.23 *(Determinant and eigenvalues) For any $n \times n$ matrix A,* $\det(A) = \Pi_{i=1}^{n} \lambda_i$.

32.6 Special Forms of Matrices

32.6.1 Triangular Matrices

Fact 32.24 *(Triangular matrix) A lower (upper) triangular matrix has zero elements above (below) the main diagonal.*

Fact 32.25 *(Eigenvalues of triangular matrix) For a triangular matrix A, the eigenvalues equal the diagonal elements of A. This follows from that*

$$det(A - \lambda I) = (A_{11} - \lambda)(A_{22} - \lambda) \ldots (A_{nn} - \lambda).$$

Fact 32.26 *(Squares of triangular matrices) If T is lower (upper) triangular, then TT is as well.*

32.6.2 Orthogonal Vector and Matrices

Fact 32.27 *(Orthogonal vector) The $n \times 1$ vectors x and y are orthogonal if $x'y = 0$.*

Fact 32.28 *(Orthogonal matrix) The $n \times n$ matrix A is orthogonal if $A'A = I$. Properties: If A is orthogonal, then* $\det(A) = \pm 1$*; if A and B are orthogonal, then AB is orthogonal.*

Example 32.29 *(Rotation of vectors ("Givens rotations").) Consider the matrix* $G = I_n$ *except that* $G_{ik} = c$, $G_{ik} = s$, $G_{ki} = -s$, *and* $G_{kk} = c$. *If we let* $c = \cos\theta$ *and* $s = \sin\theta$ *for some angle* θ, *then* $G'G = I$. *To see this, consider the simple example where* $i = 2$ *and* $k = 3$

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & c & s \\ 0 & -s & c \end{bmatrix}' \begin{bmatrix} 1 & 0 & 0 \\ 0 & c & s \\ 0 & -s & c \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & c^2+s^2 & 0 \\ 0 & 0 & c^2+s^2 \end{bmatrix},$$

which is an identity matrix since $\cos^2\theta + \sin^2\theta = 1$. G *is thus an orthogonal matrix. It is often used to "rotate" an* $n \times 1$ *vector* ε *as in* $u = G'\varepsilon$, *where we get*

$$\begin{aligned} u_t &= \varepsilon_t \text{ for } t \neq i,k \\ u_i &= \varepsilon_i c - \varepsilon_k s \\ u_k &= \varepsilon_i s + \varepsilon_k c. \end{aligned}$$

The effect of this transformation is to rotate the i^{th} *and* k^{th} *vectors counterclockwise through an angle of* θ.

32.6.3 Positive Definite Matrices

Fact 32.30 *(Positive definite matrix) The* $n \times n$ *matrix* A *is positive definite if for any non-zero* $n \times 1$ *vector* x, $x'Ax > 0$. *(It is positive semidefinite if* $x'Ax \geq 0$.*)*

Fact 32.31 *(Some properties of positive definite matrices) If* A *is positive definite, then all eigenvalues are positive and real. (To see why, note that an eigenvalue satisfies* $Ax = \lambda x$. *Premultiply by* x' *to get* $x'Ax = \lambda x'x$. *Since both* $x'Ax$ *and* $x'x$ *are positive real numbers,* λ *must also be.)*

Fact 32.32 *(More properties of positive definite matrices) If* B *is a* $r \times n$ *matrix of rank* r *and* A *is a* $n \times n$ *positive definite matrix, then* BAB' *is also positive definite and has rank* r. *For instance,* B *could be an invertible* $n \times n$ *matrix. If* $A = I_n$, *then we have that* BB' *is positive definite.*

Fact 32.33 *(More properties of positive definite matrices) If* A *is positive definite, then* $\det(A) > 0$ *and all diagional elements are positive; if* A *is positive definite, then* A^{-1} *is too.*

Fact 32.34 *(Cholesky decomposition) See Fact 32.44.*

32.6.4 Symmetric Matrices

Fact 32.35 *(Symmetric matrix) A is symmetric if* $A = A'$.

Fact 32.36 *(Properties of symmetric matrices) If A is symmetric, then all eigenvalues are real, and eigenvectors corresponding to distinct eigenvalues are orthogonal.*

Fact 32.37 *If A is symmetric, then* A^{-1} *is symmetric.*

32.6.5 Idempotent Matrices

Fact 32.38 *(Idempotent matrix) A is idempotent if* $A = AA$. *If A is also symmetric, then* $A = A'A = AA'$.

32.7 Matrix Decompositions

Fact 32.39 *(Diagonal decomposition) An* $n \times n$ *matrix A is diagonalizable if there exists a matrix C such that* $C^{-1}AC = \Lambda$ *is diagonal. We can thus write* $A = C\Lambda C^{-1}$. *The* $n \times n$ *matrix A is diagonalizable if and only if it has n linearly independent eigenvectors. We can then take C to be the matrix of the eigenvectors (in columns), and* Λ *the diagonal matrix with the corresponding eigenvalues along the diagonal.*

Fact 32.40 *(Inverting by using a diagonal decomposition) The inverse of the square matrix A is found by noting that if A is square, then from the diagonal decomposition we have*

$$\begin{aligned} AA^{-1} &= I \text{ or} \\ C\Lambda C^{-1}A^{-1} &= I, \text{ so} \\ A^{-1} &= C^{-1}\Lambda^{-1}C. \end{aligned}$$

Fact 32.41 *(Spectral decomposition.) If the eigenvectors are linearly independent, then we can decompose A as*

$$A = Z\Lambda Z^{-1}, \text{ where } \Lambda = diag(\lambda_1, ..., \lambda_n) \text{ and } Z = \begin{bmatrix} z_1 & z_2 & \cdots & z_n \end{bmatrix},$$

where Λ *is a diagonal matrix with the eigenvalues along the principal diagonal, and Z is a matrix with the corresponding eigenvalues in the columns. In this case,* $A^{-1} = Z^{-1}\Lambda^{-1}Z$.

Fact 32.42 *(Diagonal decomposition of symmetric matrices) If A is symmetric (and possibly singular) then the eigenvectors are orthogonal, $C'C = I$, so $C^{-1} = C'$. In this case, we can diagonalize A as $C'AC = \Lambda$, or $A = C\Lambda C'$. It follows that $A = \sum_{i=1}^{n} \lambda_i z_i z_i'$. If A is $n \times n$ but has rank $r \leq n$, then we can write*

$$A = \begin{bmatrix} C_1 & C_2 \end{bmatrix} \begin{bmatrix} \Lambda_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} C_1 & C_2 \end{bmatrix}' = C_1 \Lambda_1 C_1',$$

where the $n \times r$ matrix C_1 contains the r eigenvectors associated with the r non-zero eigenvalues in the $r \times r$ matrix Λ_1. Also, $A^{-1} = C'\Lambda^{-1}C$.

Fact 32.43 *(Quadratic form) If A is a convariance matrix (symmetric and positive definite, then $x'A^{-1}x$ and be rewritten as $y'y$ where $y = diag(1/\sqrt{\lambda_1}, ..., 1/\sqrt{\lambda_n})Cx$.*

Fact 32.44 *(Cholesky decomposition) Let Ω be an $n \times n$ symmetric positive definite matrix. The Cholesky decomposition gives the unique lower triangular P such that $\Omega = PP'$ (some software returns an upper triangular matrix, that is, Q in $\Omega = Q'Q$ instead). Note that each column of P is only identified up to a sign transformation; they can be reversed at will.*

Example 32.45 *(2×2 matrix) For a 2×2 matrix we have the following Cholesky decomposition*

$$chol\left(\begin{bmatrix} a & b \\ b & d \end{bmatrix}\right) = \begin{bmatrix} \sqrt{a} & 0 \\ b/\sqrt{a} & \sqrt{d - b^2/a} \end{bmatrix}.$$

Fact 32.46 *(Triangular Decomposition) Let Ω be an $n \times n$ symmetric positive definite matrix. There is a unique decomposition $\Omega = ADA'$, where A is lower triangular with ones along the principal diagonal, and D is diagonal with positive diagonal elements. This decomposition is usually not included in econometric software, but it can easily be calculated from the commonly available Cholesky decomposition since P in the Cholesky decomposition is of the form*

$$P = \begin{bmatrix} \sqrt{D_{11}} & 0 & \cdots & 0 \\ \sqrt{D_{11}}A_{21} & \sqrt{D_{22}} & & 0 \\ \vdots & & \ddots & \vdots \\ \sqrt{D_{11}}A_{n1} & \sqrt{D_{22}}A_{n2} & \cdots & \sqrt{D_{nn}} \end{bmatrix}.$$

Fact 32.47 *(Schur decomposition) The decomposition of the $n \times n$ matrix A gives the $n \times n$ matrices T and Z such that*

$$A = ZTZ^H$$

where Z is a unitary $n \times n$ matrix and T is an $n \times n$ upper triangular Schur form with the eigenvalues along the diagonal. Note that premultiplying by $Z^{-1} = Z^H$ and postmultiplying by Z gives

$$T = Z^H AZ,$$

which is upper triangular. The ordering of the eigenvalues in T can be reshuffled, although this requires that Z is reshuffled conformably to keep $A = ZTZ^H$, which involves a bit of tricky "book keeping."

Fact 32.48 *(Generalized Schur Decomposition) The decomposition of the $n \times n$ matrices G and D gives the $n \times n$ matrices S, T, Q and Z such that Q and Z are unitary and S and T upper triangular. They satisfy*

$$G = QSZ^H \text{ and } D = QTZ^H.$$

The generalized Schur decomposition solves the generalized eigenvalue problem $Dx = \lambda Gx$, where λ are the generalized eigenvalues (which will equal the diagonal elements in T divided by the corresponding diagonal element in S). Note that we can write

$$Q^H GZ = S \text{ and } Q^H DZ = T.$$

Example 32.49 *If $G = I$ in the generalized eigenvalue problem $Dx = \lambda Gx$, then we are back to the standard eigenvalue problem. Clearly, we can pick $S = I$ and $Q = Z$ in this case, so $G = I$ and $D = ZTZ^H$, as in the standard Schur decomposition.*

Fact 32.50 *(QR decomposition) Let A be $m \times n$ with $m \geq n$. The QR decomposition is*

$$\begin{aligned} A_{m\times n} &= Q_{m\times m} R_{m\times n} \\ &= \begin{bmatrix} Q_1 & Q_2 \end{bmatrix} \begin{bmatrix} R_1 \\ \mathbf{0} \end{bmatrix} \\ &= Q_1 R_1. \end{aligned}$$

where Q is orthogonal ($Q'Q = I$) and R upper triangular. The last line is the "thin QR decomposition," where Q_1 is an $m \times n$ orthogonal matrix and R_1 an $n \times n$ upper

triangular matrix.

Fact 32.51 *(Inverting by using the QR decomposition) Solving $Ax = c$ by inversion of A can be very numerically inaccurate (no kidding, this is a real problem). Instead, the problem can be solved with QR decomposition. First, calculate Q_1 and R_1 such that $A = Q_1 R_1$. Note that we can write the system of equations as*

$$Q_1 Rx = c.$$

Premultply by Q_1' to get (since $Q_1' Q_1 = I$)

$$Rx = Q_1' c.$$

This is an upper triangular system which can be solved very easily (first solve the first equation, then use the solution is the second, and so forth.)

Fact 32.52 *(Singular value decomposition) Let A be an $m \times n$ matrix of rank ρ. The singular value decomposition is*

$$A = U_{m\times m} S_{m\times n} V'_{n\times n}$$

where U and V are orthogonal and S is diagonal with the first ρ elements being non-zero, that is,

$$S = \begin{bmatrix} S_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}, \text{ where } S_1 = \begin{bmatrix} s_{11} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & s_{\rho\rho} \end{bmatrix}.$$

Fact 32.53 *(Singular values and eigenvalues) The singular values of A are the nonnegative square roots of AA^H if $m \leq n$ and of $A^H A$ if $m \geq n$.*

Remark 32.54 *If the square matrix A is symmetric and idempotent ($A = A'A$), then the singular values are the same as the eigevalues. From Fact (32.42) we know that a symmetric A can be decomposed as $A = C\Lambda C'$. It follows that this is the same as the singular value decomposition.*

Fact 32.55 *(Condition number) The condition number of a matrix is the ratio of the largest (in magnitude) of the singular values to the smallest*

$$c = |s_{ii}|_{\max} / |s_{ii}|_{\min}.$$

For a square matrix, we can calculate the condition value from the eigenvalues of AA^H or $A^H A$ (see Fact 32.53). In particular, for a square matrix we have

$$c = \left|\sqrt{\lambda_i}\right|_{\max} / \left|\sqrt{\lambda_i}\right|_{\min},$$

where λ_i are the eigenvalues of AA^H and A is square.

Fact 32.56 *(Condition number and computers) The determinant is not a good indicator of the realibility of numerical inversion algorithms. Instead, let c be the condition number of a square matrix. If $1/c$ is close to the a computer's floating-point precision (10^{-13} or so), then numerical routines for a matrix inverse become unreliable. For instance, while $\det(0.1 \times I_{20}) = 10^{-20}$, the condition number of $0.1 \times I_{20}$ is unity and the matrix is indeed easy to invert to get $10 \times I_{20}$.*

Fact 32.57 *(Inverting by using the SVD decomposition) The inverse of the square matrix A is found by noting that if A is square, then from Fact 32.52 we have*

$$\begin{aligned} AA^{-1} &= I \text{ or} \\ USV'A^{-1} &= I, \text{ so} \\ A^{-1} &= VS^{-1}U', \end{aligned}$$

provided S is invertible (otherwise A will not be). Since S is diagonal, S^{-1} is also diagonal with the inverses of the diagonal elements in S, so it is very easy to compute.

Fact 32.58 *(Pseudo inverse or generalized inverse) The Moore-Penrose pseudo (generalized) inverse of an $m \times n$ matrix A is defined as*

$$A^+ = VS^+U', \text{ where } S^+_{nxm} = \begin{bmatrix} S_{11}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix},$$

where V and U are from Fact 32.52. The submatrix S_{11}^{-1} contains the reciprocals of the non-zero singular values along the principal diagonal. A^+ satisfies the A^+ satisfies the Moore-Penrose conditions

$$AA^+A = A,\ A^+AA^+ = A^+,\ \left(AA^+\right)' = AA^+, \text{ and } \left(A^+A\right)' = A^+A.$$

See Fact 32.9 for the idea behind the generalized inverse.

Fact 32.59 *(Some properties of generalized inverses) If A has full rank, then $A^+ = A^{-1}$; $(BC)^+ = C^+B^+$; if B, and C are invertible, then $(BAC)^{-1} = C^{-1}A^+B^{-1}$; $(A^+)' = (A')^+$; if A is symmetric, then A^+ is symmetric.*

Example 32.60 *(Pseudo inverse of a square matrix) For the matrix*

$$A = \begin{bmatrix} 1 & 2 \\ 3 & 6 \end{bmatrix}, \text{ we have } A^+ = \begin{bmatrix} 0.02 & 0.06 \\ 0.04 & 0.12 \end{bmatrix}.$$

Fact 32.61 *(Pseudo inverse of symmetric matrix) If A is symmetric, then the SVD is identical to the spectral decomposition $A = Z\Lambda Z'$ where Z is a matrix of the orthogonal eigenvectors ($Z'Z = I$) and Λ is a diagonal matrix with the eigenvalues along the main diagonal. By Fact 32.58) we then have $A^+ = Z\Lambda^+Z'$, where*

$$\Lambda^+ = \begin{bmatrix} \Lambda_{11}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix},$$

with the reciprocals of the non-zero eigen values along the principal diagonal of Λ_{11}^{-1}.

32.8 Matrix Calculus

Fact 32.62 *(Matrix differentiation of non-linear functions, $\partial y/\partial x'$) Let the vector $y_{n\times 1}$ be a function of the vector $x_{m\times 1}$*

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = f(x) = \begin{bmatrix} f_1(x) \\ \vdots \\ f_n(x) \end{bmatrix}.$$

Then, let $\partial y/\partial x'$ be the $n \times m$ matrix

$$\frac{\partial y}{\partial x'} = \begin{bmatrix} \frac{\partial f_1(x)}{\partial x'} \\ \vdots \\ \frac{\partial f_n(x)}{\partial x'} \end{bmatrix} = \begin{bmatrix} \frac{\partial f_1(x)}{\partial x_1} & \cdots & \frac{\partial f_1(x)}{\partial x_m} \\ \vdots & & \vdots \\ \frac{\partial f_n(x)}{\partial x_1} & \cdots & \frac{\partial f_n(x)}{\partial x_m} \end{bmatrix}.$$

This matrix is often called the Jacobian of the f functions. (Note that the notation implies that the derivatives of the first element in y, denoted y_1, with respect to each of the elements in x' are found in the first row of $\partial y/\partial x'$. A rule to help memorizing the format of $\partial y/\partial x'$: y is a column vector and x' is a row vector.)

Fact 32.63 *($\partial y'/\partial x$ instead of $\partial y/\partial x'$) With the notation in the previous Fact, we get*

$$\frac{\partial y'}{\partial x} = \begin{bmatrix} \frac{\partial f_1(x)}{\partial x} & \cdots & \frac{\partial f_n(x)}{\partial x} \end{bmatrix} = \begin{bmatrix} \frac{\partial f_1(x)}{\partial x_1} & \cdots & \frac{\partial f_n(x)}{\partial x_1} \\ \vdots & & \vdots \\ \frac{\partial f_1(x)}{\partial x_m} & \cdots & \frac{\partial f_n(x)}{\partial x_m} \end{bmatrix} = \left(\frac{\partial y}{\partial x'}\right)'.$$

Fact 32.64 *(Matrix differentiation of linear systems) When $y_{n\times 1} = A_{n\times m}x_{m\times 1}$, then $f(x)$ is a linear function*

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} a_{11} & \cdots & a_{1m} \\ \vdots & & \vdots \\ a_{n1} & \cdots & a_{nm} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_m \end{bmatrix}.$$

In this case $\partial y/\partial x' = A$ and $\partial y'/\partial x = A'$.

Fact 32.65 *(Matrix differentiation of inner product) The inner product of two column vectors, $y = z'x$, is a special case of a linear system with $A = z'$. In this case we get $\partial(z'x)/\partial x' = z'$ and $\partial(z'x)/\partial x = z$. Clearly, the derivatives of $x'z$ are the same (a transpose of a scalar).*

Example 32.66 *($\partial(z'x)/\partial x = z$ when x and z are 2×1 vectors)*

$$\frac{\partial}{\partial x}\left(\begin{bmatrix} z_1 & z_2 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right) = \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}.$$

Fact 32.67 *(First order Taylor series) For each element $f_i(x)$ in the $n\times$ vector $f(x)$, we can apply the mean-value theorem*

$$f_i(x) = f_i(c) + \frac{\partial f_i(b_i)}{\partial x'}(x-c),$$

for some vector b_i between c and x. Stacking these expressions gives

$$\begin{bmatrix} f_1(x) \\ \vdots \\ f_n(x) \end{bmatrix} = \begin{bmatrix} f_1(c) \\ \vdots \\ f_n(c) \end{bmatrix} + \begin{bmatrix} \frac{\partial f_1(b_1)}{\partial x_1} & \cdots & \frac{\partial f_1(b_1)}{\partial x_m} \\ \vdots & & \vdots \\ \frac{\partial f_n(b_n)}{\partial x_1} & \cdots & \frac{\partial f_n(b_n)}{\partial x_m} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_m \end{bmatrix} \text{ or}$$

$$f(x) = f(c) + \frac{\partial f(b)}{\partial x'}(x-c),$$

where the notation $f(b)$ is a bit sloppy. It should be interpreted as that we have to evaluate the derivatives at different points for the different elements in $f(x)$.

Fact 32.68 *(Matrix differentiation of quadratic forms) Let $x_{m\times 1}$ be a vector, $A_{m\times m}$ a matrix, and $f(x)_{n\times 1}$ a vector of functions. Then,*

$$\begin{aligned}\frac{\partial f(x)' Af(x)}{\partial x} &= \left(\frac{\partial f(x)}{\partial x'}\right)' \left(A + A'\right) f(x) \\ &= 2\left(\frac{\partial f(x)}{\partial x'}\right)' Af(x) \text{ if } A \text{ is symmetric.}\end{aligned}$$

If $f(x) = x$, then $\partial f(x)/\partial x' = I$, so $\partial (x'Ax)/\partial x = 2Ax$ if A is symmetric.

Example 32.69 *($\partial (x'Ax)/\partial x = 2Ax$ when x is 2×1 and A is 2×2)*

$$\begin{aligned}\frac{\partial}{\partial x}\left(\begin{bmatrix}x_1 & x_2\end{bmatrix}\begin{bmatrix}A_{11} & A_{12}\\ A_{21} & A_{22}\end{bmatrix}\begin{bmatrix}x_1\\ x_2\end{bmatrix}\right) &= \left(\begin{bmatrix}A_{11} & A_{12}\\ A_{21} & A_{22}\end{bmatrix} + \begin{bmatrix}A_{11} & A_{21}\\ A_{12} & A_{22}\end{bmatrix}\right)\begin{bmatrix}x_1\\ x_2\end{bmatrix}, \\ &= 2\begin{bmatrix}A_{11} & A_{12}\\ A_{12} & A_{22}\end{bmatrix}\begin{bmatrix}x_1\\ x_2\end{bmatrix} \text{ if } A_{21} = A_{12}.\end{aligned}$$

Example 32.70 *(Least squares) Consider the linear model $Y_{m\times 1} = X_{m\times n}\beta_{n\times 1} + u_{m\times 1}$. We want to minimize the sum of squared fitted errors by choosing the $n\times 1$ vector β. The fitted errors depend on the chosen β: $u(\beta) = Y - X\beta$, so quadratic loss function is*

$$\begin{aligned}L &= u(\beta)'u(\beta) \\ &= (Y - X\beta)'(Y - X\beta).\end{aligned}$$

In thus case, $f(\beta) = u(\beta) = Y - X\beta$, so $\partial f(\beta)/\partial \beta' = -X$. The first order condition for $u'u$ is thus

$$-2X'\left(Y - X\hat{\beta}\right) = \mathbf{0}_{n\times 1} \text{ or } X'Y = X'X\hat{\beta},$$

which can be solved as

$$\hat{\beta} = \left(X'X\right)^{-1}X'Y.$$

Fact 32.71 *(Matrix of 2nd order derivatives of of a non-linear function, $\partial^2 y/\partial x\partial x'$) Let the scalar y be a function of the vector $x_{m\times 1}$*

$$y = f(x).$$

Then, let $\partial^2 y/\partial x\partial x'$ be the $m\times m$ matrix with $\partial^2 y/\partial x_i\partial x_j$ in cell (i,j)

$$\frac{\partial^2 y}{\partial x\partial x'} = \begin{bmatrix}\frac{\partial^2 f(x)}{\partial x_1\partial x_1} & \cdots & \frac{\partial^2 f(x)}{\partial x_1\partial x_m}\\ \vdots & & \vdots \\ \frac{\partial^2 f(x)}{\partial x_m\partial x_1} & \cdots & \frac{\partial^2 f(x)}{\partial x_m\partial xm}\end{bmatrix}.$$

This matrix is often called the Hessian of the f function. This is clearly a symmetric matrix.

32.9 Miscellaneous

Fact 32.72 *(Some properties of transposes)* $(A+B)' = A'+B'$; $(ABC)' = C'B'A'$ *(if conformable).*

Fact 32.73 *(Kronecker product) If A and B are matrices, then*

$$A \otimes B = \begin{bmatrix} a_{11}B & \cdots & a_{1n}B \\ \vdots & & \vdots \\ a_{m1}B & \cdots & a_{mn}B \end{bmatrix}.$$

Some properties: $(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}$ *(if conformable);* $(A \otimes B)(C \otimes D) = AC \otimes BD$ *(if conformable);* $(A \otimes B)' = A' \otimes B'$; *if a is $m \times 1$ and b is $n \times 1$, then $a \otimes b = (a \otimes I_n)b$; if A is symmetric and positive definite, then* $chol(A \otimes I) = chol(A) \otimes I$ *and* $chol(I \otimes A) = I \otimes chol(A)$.

Fact 32.74 *(Cyclical permutation of trace)* $Trace(ABC) = Trace(BCA) = Trace(CAB)$, *if the dimensions allow the products.*

Fact 32.75 *(The* vec *operator).* vec A *where A is $m \times n$ gives an $mn \times 1$ vector with the columns in A stacked on top of each other. For instance,* $\text{vec}\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = \begin{bmatrix} a_{11} \\ a_{21} \\ a_{12} \\ a_{22} \end{bmatrix}$.

Properties: $\text{vec}(A+B) = \text{vec}\,A + \text{vec}\,B$; $\text{vec}(ABC) = (C' \otimes A)\,\text{vec}\,B$; *if a and b are column vectors, then* $\text{vec}(ab') = b \otimes a$.

Fact 32.76 *(The vech operator) vech$A$ where A is $m \times m$ gives an $m(m+1)/2 \times 1$ vector with the elements on and below the principal diagonal A stacked on top of each other (columnwise). For instance, vech* $\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = \begin{bmatrix} a_{11} \\ a_{21} \\ a_{22} \end{bmatrix}$, *that is, like vec, but uses only the elements on and below the principal diagonal.*

Fact 32.77 *(Duplication matrix) The duplication matrix D_m is defined such that for any symmetric $m \times m$ matrix A we have* $\text{vec}\, A = D_m vech A$. *The duplication matrix is therefore useful for "inverting" the vech operator (the step from* $\text{vec}\, A$ *to A is trivial). For instance, to continue the example of the vech operator*

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} a_{11} \\ a_{21} \\ a_{22} \end{bmatrix} = \begin{bmatrix} a_{11} \\ a_{21} \\ a_{21} \\ a_{22} \end{bmatrix} \text{ or } D_2 vech A = \text{vec}\, A.$$

Fact 32.78 *(OLS notation) Let x_t be $k \times 1$ and y_t be $m \times 1$. Suppose we have T such vectors. The sum of the outer product (a $k \times m$ matrix) is*

$$S = \sum_{t=1}^{T} x_t y_t'.$$

Create matrices $X_{T \times k}$ and $Y_{T \times m}$ by letting x_t' and y_t' be the t^{th} rows

$$X_{T \times k} = \begin{bmatrix} x_1' \\ \vdots \\ x_T' \end{bmatrix} \text{ and } Y_{T \times m} = \begin{bmatrix} y_1' \\ \vdots \\ y_T' \end{bmatrix}.$$

We can then calculate the same sum of outer product, S, as

$$S = X'Y.$$

(To see this, let $X(i,:)$ be the ith row of X, and similarly for Y, so

$$X'Y = \sum_{t=1}^{T} X(t,:)'Y(t,:),$$

which is precisely $\Sigma_{t=1}^{T} x_t y_t'$.) For instance, with

$$x_t = \begin{bmatrix} a_t \\ b_t \end{bmatrix} \text{ and } y_t = \begin{bmatrix} p_t \\ q_t \\ r_t \end{bmatrix},$$

and $T = 2$ *we have*

$$X'Y = \begin{bmatrix} a_1 & a_2 \\ b_1 & b_2 \end{bmatrix} \begin{bmatrix} p_1 & q_1 & r_1 \\ p_2 & q_2 & r_2 \end{bmatrix} = \sum_{t=1}^{T} \begin{bmatrix} a_t \\ b_t \end{bmatrix} \begin{bmatrix} p_t & q_t & r_t \end{bmatrix}.$$

Fact 32.79 *(Matrix geometric series) Suppose the eigenvalues to the square matrix* A *are all less than one in modulus. Then,*

$$I + A + A^2 + \cdots = (1 - A)^{-1}.$$

To see why this makes sense, consider $(1 - A)\, \Sigma_{t=1}^{T} A^t$ *(with the convention that* $A^0 = I$*). It can be written as*

$$(1 - A)\, \Sigma_{t=1}^{T} A^t = \left(I + A + A^2 + \cdots\right) - A\left(I + A + A^2 + \cdots\right) = I - A^{T+1}.$$

If all the eigenvalues are stable, then $\lim_{T\to\infty} A^{T+1} = \mathbf{0}$*, so taking the limit of the previous equation gives*

$$(1 - A) \lim_{T\to\infty} \Sigma_{t=1}^{T} A = I.$$

Fact 32.80 *(Matrix exponential) The matrix exponential of an* $n \times n$ *matrix* A *is defined as*

$$\exp(At) = \sum_{s=0}^{\infty} \frac{(At)^s}{s!}.$$

Bibliography

Ait-Sahalia, Y., and A. W. Lo, 1998, "Nonparametric estimation of state-price densities implicit in financial asset prices," *Journal of Finance*, 53, 499–547.

Andrews, D. W. K., and J. C. Monahan, 1992, "An improved heteroskedasticity and autocorrelation consistent covariance matrix estimator," *Econometrica*, 60, 953–966.

Anton, H., 1987, *Elementary linear algebra*, John Wiley and Sons, New York, 5th edn.

Azzalini, A., 2005, "The skew-normal distribution and related multivariate families," *Scandinavian Journal of Statistics*, 32, 159–188.

Bali, T. G., R. F. Engle, and S. Murray, 2016, *Empirical Asset Pricing*, Wiley, Hoboken, New Jersey.

Baltagi, D. H., 2008, *Econometric Analysis of Panel Data*, Wiley, 4th edn.

Bekaert, G., and M. S. Urias, 1996, "Diversification, integration and emerging market closed-end funds," *Journal of Finance*, 51, 835–869.

Berkowitz, J., and L. Kilian, 2000, "Recent developments in bootstrapping time series," *Econometric-Reviews*, 19, 1–48.

Björk, Å., 1996, *Numerical methods for least squares problems*, SIAM, Philadelphia.

Breeden, D. T., M. R. Gibbons, and R. H. Litzenberger, 1989, "Empirical tests of the consumption-oriented CAPM," *Journal of Finance*, 44, 231–262.

Brockwell, P. J., and R. A. Davis, 1991, *Time series: theory and methods*, Springer Verlag, New York, second edn.

Campbell, J. Y., 2018, *Financial decisions and markets*, Princeton University Press, Princeton, New Jersey.

Campbell, J. Y., and J. H. Cochrane, 1999, "By force of habit: a consumption-based explanation of aggregate stock market behavior," *Journal of Political Economy*, 107, 205–251.

Campbell, J. Y., A. W. Lo, and A. C. MacKinlay, 1997, *The econometrics of financial markets*, Princeton University Press, Princeton, New Jersey.

Campbell, J. Y., and R. J. Shiller, 1988, "The dividend-price ratio and expectations of future dividends and discount factors," *Review of Financial Studies*, 1, 195–227.

Campbell, J. Y., and S. B. Thompson, 2008, "Predicting the equity premium out of sample: can anything beat the historical average," *Review of Financial Studies*, 21, 1509–1531.

Campbell, J. Y., and L. M. Viceira, 1999, "Consumption and portfolio decisions when expected returns are time varying," *Quarterly Journal of Economics*, 114, 433–495.

Chen, N.-F., R. Roll, and S. A. Ross, 1986, "Economic forces and the stock market," *Journal of Business*, 59, 383–403.

Christiansen, C., A. Ranaldo, and P. Söderlind, 2011, "The time-varying systematic risk of carry trade strategies," *Journal of Financial and Quantitative Analysis*, 46, 1107–1125.

Clark, T. E., and M. W. McCracken, 2001, "Tests of equal forecast accuracy and encompassing for nested models," *Journal of Econometrics*, 105, 85–110.

Clark, T. E., and K. D. West, 2007, "Approximately normal tests for equal predictive accuracy in nested models," *Journal of Ec*, 138, 291–311.

Cochrane, J. H., 2001, *Asset pricing*, Princeton University Press, Princeton, New Jersey.

Cochrane, J. H., 2005, *Asset pricing*, Princeton University Press, Princeton, New Jersey, revised edn.

Dahlquist, M., J. V. Martinez, and P. Söderlind, 2016, "Individual Investor Activity and Performance," forthcoming in *The Review of Financial Studies*.

Davidson, J., 2000, *Econometric theory*, Blackwell Publishers, Oxford.

Davidson, R., and J. G. MacKinnon, 1993, *Estimation and inference in econometrics*, Oxford University Press, Oxford.

Davison, A. C., and D. V. Hinkley, 1997, *Bootstrap methods and their applications*, Cambridge University Press.

DeGroot, M. H., 1986, *Probability and statistics*, Addison-Wesley, Reading, Massachusetts.

Diebold, F. X., 2001, *Elements of forecasting*, South-Western, 2nd edn.

Diebold, F. X., and R. S. Mariano, 1995, "Comparing predcitve accuracy," *Journal of Business and Economic Statistics*, 13, 253–265.

Driscoll, J., and A. Kraay, 1998, "Consistent covariance matrix estimation with spatially dependent panel data," *Review of Economics and Statistics*, 80, 549–560.

Efron, B., T. Hasti, I. Johnstone, and R. Tibshirani, 2004, "Least angle regression," *The Annals of Statistics*, 32, 407–499.

Efron, B., and R. J. Tibshirani, 1993, *An introduction to the bootstrap*, Chapman and Hall, New York.

Elliot, G., and A. Timmermann, 2016, *Economic forecasting*, Princeton University Press, Princeton, New Jersey.

Epstein, L. G., and S. E. Zin, 1991, "Substitution, risk aversion, and the temporal behavior of asset returns: an empirical analysis," *Journal of Political Economy*, 99, 263–286.

Fama, E., and J. MacBeth, 1973, "Risk, return, and equilibrium: empirical tests," *Journal of Political Economy*, 71, 607–636.

Fama, E. F., and K. R. French, 1993, "Common risk factors in the returns on stocks and bonds," *Journal of Financial Economics*, 33, 3–56.

Fama, E. F., and K. R. French, 1996, "Multifactor explanations of asset pricing anomalies," *Journal of Finance*, 51, 55–84.

Ferson, W. E., 1995, "Theory and empirical testing of asset pricing models," in Robert A. Jarrow, Vojislav Maksimovic, and William T. Ziemba (ed.), *Handbooks in Operations Research and Management Science* . pp. 145–200, North-Holland, Amsterdam.

Ferson, W. E., S. Sarkissian, and T. T. Simin, 2003, "Spurious regressions in financial economics," *Journal of Finance*, 57, 1393–1413.

Ferson, W. E., and R. Schadt, 1996, "Measuring fund strategy and performance in changing economic conditions," *Journal of Finance*, 51, 425–461.

Franses, P. H., and D. van Dijk, 2000, *Non-linear time series models in empirical finance*, Cambridge University Press.

Gibbons, M., S. Ross, and J. Shanken, 1989, "A test of the efficiency of a given portfolio," *Econometrica*, 57, 1121–1152.

Golub, G. H., and C. F. van Loan, 1989, *Matrix computations*, The John Hopkins University Press, Baltimore, 2nd edn.

Goyal, A., and I. Welch, 2008, "A comprehensive look at the empirical performance of equity premium prediction," *Review of Financial Studies 2008*, 21, 1455–1508.

Greenberg, M. D., 1988, *Advanced engineering mathematics*, Prentice Hall, Englewood Cliffs, New Jersey.

Greene, W. H., 2000, *Econometric analysis*, Prentice-Hall, Upper Saddle River, New Jersey, 4th edn.

Greene, W. H., 2003, *Econometric analysis*, Prentice-Hall, Upper Saddle River, New Jersey, 5th edn.

Greene, W. H., 2018, *Econometric analysis*, Pearson Education Ltd, 8th edn.

Hamilton, J. D., 1994, *Time series analysis*, Princeton University Press, Princeton.

Hansen, B. E., forthcoming (2021), *Econometrics*, Princeton University Press, Princeton.

Hansen, L., 1982, "Large sample properties of generalized instrumental variables estimators," *Econometrica*, 50, 1029–1054.

Hansen, L. P., and R. Jagannathan, 1991, "Implications of security market data for models of dynamic economies," *Journal of Political Economy*, 99, 225–262.

Härdle, W., 1990, *Applied nonparametric regression*, Cambridge University Press, Cambridge.

Harris, D., and L. Matyas, 1999, "Introduction to the generalized method of moments estimation," in Laszlo Matyas (ed.), *Generalized Method of Moments Estimation* . chap. 1, Cambridge University Press.

Hastie, T., R. Tibshirani, and J. Friedman, 2001, *The elements of statistical learning: data mining, inference and prediction*, Springer Verlag.

Hayashi, F., 2000, *Econometrics*, Princeton University Press.

Hill, R. C., W. E. Griffiths, and G. C. Lim, 2008, *Principles of Econometrics*, John Wiley and Sons, 3rd edn.

Hoechle, D., 2007, "Robust Standard Errors for Panel Regressions with Cross-Sectional Dependence," *The Stata Journal*, 7, 281–312.

Hoechle, D., M. M. Schmid, and H. Zimmermann, 2015, "Decomposing Performance," Working paper, University of St. Gallen.

Holm, S., 1979, "A simple sequentially rejective multiple test procedure," *Scandinavian Journal of Statistics*, 6, 65–70.

Horowitz, J. L., 2001, "The Bootstrap," in J.J. Heckman, and E. Leamer (ed.), *Handbook of Econometrics* . , vol. 5, Elsevier.

Huberman, G., and S. Kandel, 1987, "Mean-variance spanning," *Journal of Finance*, 42, 873–888.

Ingram, B.-F., and B.-S. Lee, 1991, "Simulation estimation of time-series models," *Journal of Econometrics*, 47, 197–205.

Jagannathan, R., and Z. Wang, 1996, "The conditional CAPM and the cross-section of expectd returns," *Journal of Finance*, 51, 3–53.

Jagannathan, R., and Z. Wang, 1998, "A note on the asymptotic covariance in Fama-MacBeth regression," *Journal of Finance*, 53, 799–801.

Jagannathan, R., and Z. Wang, 2002, "Empirical evaluation of asset pricing models: a comparison of the SDF and beta methods," *Journal of Finance*, 57, 2337–2367.

Johnson, N. L., S. Kotz, and N. Balakrishnan, 1994, *Continuous univariate distributions*, Wiley, New York, 2nd edn.

Johnston, J., and J. DiNardo, 1997, *Econometric methods*, McGraw-Hill, New York, 4th edn.

Karnaukh, N., A. Ranaldo, and P. Söderlind, 2015, "Understanding FX liquidity," *The Review of Financial Studies*, 28, 3073–3108.

Leitch, G., and J. E. Tanner, 1991, "Economic forecast evaluation: profit versus the conventional error measures," *American Economic Review*, 81, 580–590.

Lettau, M., and S. Ludvigson, 2001, "Resurrecting the (C)CAPM: a cross-sectional test when risk premia are time-varying," *Journal of Political Economy*, 109, 1238–1287.

Lo, A. W., and A. C. MacKinlay, 1990, "When are contrarian profits due to stock market overreaction?," *Review of Financial Studies*, 3, 175–208.

MacKinlay, C., 1995, "Multifactor models do not explain deviations from the CAPM," *Journal of Financial Economics*, 38, 3–28.

Mittelhammer, R. C., 1996, *Mathematical statistics for economics and business*, Springer-Verlag, New York.

Mittelhammer, R. C., G. J. Judge, and D. J. Miller, 2000, *Econometric foundations*, Cambridge University Press, Cambridge.

Newey, W. K., and K. D. West, 1987, "A simple positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix," *Econometrica*, 55, 703–708.

Ogaki, M., 1993, "Generalized method of moments: econometric applications," in G. S. Maddala, C. R. Rao, and H. D. Vinod (ed.), *Handbook of Statistics*, vol. 11, . chap. 17, pp. 455–487, Elsevier.

Pagan, A., and A. Ullah, 1999, *Nonparametric econometrics*, Cambridge University Press.

Pesaran, M. H., 2015, *Time series and panel data econometrics*, Oxford University Press.

Petersen, M. A., 2009, "Estimating standard errors in finance panel data sets: comparing approaches," *The Review of Financial Studies*, 22, 435–480.

Pindyck, R. S., and D. L. Rubinfeld, 1998, *Econometric models and economic forecasts*, Irwin McGraw-Hill, Boston, Massachusetts, 4ed edn.

Priestley, M. B., 1981, *Spectral analysis and time series*, Academic Press.

Singleton, K. J., 2006, *Empirical dynamic asset pricing*, Princeton University Press.

Söderlind, P., 1999, "An interpretation of SDF based performance measures," *European Finance Review*, 3, 233–237.

Söderlind, P., 2006, "C-CAPM Refinements and the cross-section of returns," *Financial Markets and Portfolio Management*, 20, 49–73.

Söderlind, P., 2009, "An extended Stein's lemma for asset pricing," *Applied Economics Letters*, 16, 1005–1008.

Stekler, H. O., 1991, "Macroeconomic forecast evaluation techniques," *International Journal of Forecasting*, 7, 375–384.

Taylor, S. J., 2005, *Asset price dynamics, volatility, and prediction*, Princeton University Press.

Treynor, J. L., and K. Mazuy, 1966, "Can Mutual Funds Outguess the Market?," *Harvard Business Review*, 44, 131–136.

Verbeek, M., 2004, *A guide to modern econometrics*, Wiley, Chichester, 2nd edn.

Verbeek, M., 2012, *A guide to modern econometrics*, Wiley, 4th edn.

Wooldridge, J. M., 2010, *Econometric analysis of cross section and panel data*, MIT Press, 2nd edn.