Econometrics II (ECON 372)
Lecture 1
Matrix Representation of OLS

Teng Wah Leo

1 Short Revision of Matrix Algebra

From our discussion of multiple variable ordinary least squares regression, we saw how complicated the calculations can get. If we were to write a program line by line, you can imagine how many lines of solution we would have to describe for the computer to calculate. The question then is whether we can express the solution more succinctly. The answer is in linear algebra, or matrix algebra. Before we describe the solution, let us revise some matrix algebra. From here on, when we describe a vector of elements, we mean a column vector. For example, let a be a n element vector, then

$$a = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix}$$

Note that the usual rules of simple algebra still holds. What we will be discussing are some operations we do not see in simple algebra.

1. $(A + B) + C = A + (B + C)$
2. $(AB)\,C = A\,(BC)$
3. $A\,(B + C) = AB + AC$
4. $IA = AI = A$, where I is the identity matrix (matrix with ones on the diagonal, and zero off the diagonal).

1.1 Transpose

Let A, B and C be matrices. A transpose of a matrix is where we change the first row of a matrix into the first column, the second row into the second column, ..., and is denote with $'$. Some of the properties of the operation are

1. $(A')' = A$
2. $(A + B)' = A' + B'$
3. $(AB)' = B'A'$
4. $A' = A$, if A is a symmetric matrix where a matrix is a symmetric matrix when the off diagonal elements above the diagonal are the same as those below the diagonal.

1.2 Matrix Multiplication

Let a be as described above, and b be a similar n element vector. Then

$$\mathbf{a'b} = \begin{bmatrix} a_1 & a_2 & \dots & a_n \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix} = \sum_{i=1}^{n} a_i b_i$$

What we are doing above is multiplying the corresponding elements of each the vectors, that is i^{th} element of a with the i^{th} element of b. For example,

$$\mathbf{a'} = \begin{bmatrix} a_1 & a_2 & \dots & a_n \end{bmatrix}$$

And

$$(\mathbf{a'})' = a$$

1.3 Scalar, Dot, or Inner Product

What do we do if we multiply matrices together? This leads to inner product, and is described as follows; Let A and B be $m \times n$ matrix (m rows and n columns) and $n \times p$

matrix (n rows with p columns) respectively. That is let

$$A = \begin{bmatrix} \widetilde{a}_1 \\ \widetilde{a}_2 \\ \vdots \\ \widetilde{a}_m \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix}$$

$$B = \begin{bmatrix} \widetilde{b}_1 & \widetilde{b}_2 & \dots & \widetilde{b}_p \end{bmatrix} = \begin{bmatrix} b_{11} & b_{12} & \dots & b_{1p} \\ b_{21} & b_{22} & \dots & b_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ b_{m1} & b_{m2} & \dots & b_{mp} \end{bmatrix}$$

Then the product of AB is

$$A_{(m\times n)}B_{(n\times p)} = C_{(n\times p)} = \begin{bmatrix} \widetilde{a}_1\widetilde{b}_1 & \widetilde{a}_1\widetilde{b}_2 & \dots & \widetilde{a}_1\widetilde{b}_p \\ \widetilde{a}_2\widetilde{b}_1 & \widetilde{a}_2\widetilde{b}_2 & \dots & \widetilde{a}_2\widetilde{b}_p \\ \vdots & \vdots & \ddots & \vdots \\ \widetilde{a}_m\widetilde{b}_1 & \widetilde{a}_m\widetilde{b}_2 & \dots & \widetilde{a}_m\widetilde{b}_p \end{bmatrix}$$

$$= \begin{bmatrix} \sum_{i=1}^n a_{1i}b_{i1} & \sum_{i=1}^n a_{1i}b_{i2} & \dots & \sum_{i=1}^n a_{1i}b_{ip} \\ \sum_{i=1}^n a_{2i}b_{i1} & \sum_{i=1}^n a_{2i}b_{i2} & \dots & \sum_{i=1}^n a_{2i}b_{ip} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^n a_{mi}b_{i1} & \sum_{i=1}^n a_{mi}b_{i2} & \dots & \sum_{i=1}^n a_{mi}b_{ip} \end{bmatrix}$$

1.4 Matrix Addition

$$a + b = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} + \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix} = \begin{bmatrix} a_1 + b_1 \\ a_2 + b_2 \\ \vdots \\ a_n + b_n \end{bmatrix}$$

1.5 Geometric Interpretation of Matrices and Vectors

1. Vectors can be depicted as directed line segments from the origin, and terminates at the point described by the coordinates.

2. Parallelogram law for the addition of vectors.

3. A vector in $\mathbb{R}^n$ can always be represented by a linear combination of $(n \times 1)$ vectors. Such a vector is called *basis*, and they are not unique. All that is required is for each of the vectors used in the linear combination to point in different directions.

4. A vector space is a collection of vectors with the following properties: 1. If v_1 and v_2 are any two vectors in the space, then $v_1 + v_2$ is in the space. 2. If v is in the space, and λ is a scalar constraint, then λv is in the space.

5. Two vectors are said to be **linearly independent** if the only solution to

$$\lambda_1 a + \lambda_2 b = 0$$

where a and b are two vectors, and λ_1 and λ_2 are scalars such the $\lambda_1 = \lambda_1 = 0$.

6. The set of vectors in $\mathbb{R}^n$ that can be used as a linear combination to represent another vector in the vector space is call the **spanning set**. The set of vectors unlike basis need not be linearly independent, therefore, the **spanning set** can be very large.

7. We can always take a subset of a **basis** from the set of **basis** describing a space in $\mathbb{R}^n$ to describe another space. For example, we can always take $k < n$ *basis* to span a new space, which we call a **hyperplane**. Another example is when we take 2 **basis** from $\mathbb{R}^3$, we get a plane.

8. Each vector in $\mathbb{R}^n$ may be expressed as a *unique* linear combination of some appropriate set of n linearly independent vectors.

9. Let $\mathbf{a}$ and $\mathbf{b}$ be two vectors, then they are orthogonal (at right angles to each other) if and only if $\mathbf{a}'\mathbf{b} = 0$

1.6 Rank of a Matrix

1. The maximum number of linearly independent rows is equal to the maximum number of linearly independent columns. The number is the rank of the matrix, denoted by $\rho(A)$.

2. For $A_{m \times n}$, $\rho(A) \leq \min(m, n)$

3. $\rho(A) = \rho(A')$

4. If $\rho(A) = m = n$, which implies that A is nonsingular, then a unique inverse A^{-1} exists.

5. $\rho(A'A) = \rho(AA') = \rho(A)$

6. $\rho(AB) \leq \min[\rho(A), \rho(B)]$

1.7 Matrix Inverse

We will now cover the calculation of an inverse, but note the rules (assuming all matrices below are nonsingular);

1. $(AB)^{-1} = B^{-1}A^{-1}$

2. $(A^{-1})^{-1} = A$

3. $(A')^{-1} = (A^{-1})'$

4. $|A^{-1}| = \frac{1}{|A|}$

5. The inverse of an upper (lower) triangular matrix is also an upper (lower) triangular matrix.

6. Inverse of a block diagonal matrix; Let A be

$$\begin{aligned} A &= \begin{bmatrix} A_{11} & 0 \\ 0 & A_{22} \end{bmatrix} \\ A^{-1} &= \begin{bmatrix} A_{11}^{-1} & 0 \\ 0 & A_{22}^{-1} \end{bmatrix} \end{aligned}$$

7. A is a m by n matrix, and let B be another matrix. A kronecker product is

$$A \otimes B = \begin{bmatrix} a_{11}B & a_{12}B & \dots & a_{1n}B \\ a_{21}B & a_{22}B & \dots & a_{2n}B \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1}B & a_{m2}B & \dots & a_{mn}B \end{bmatrix}$$

then

$$(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}$$

1.8 Quadratic Forms & max and min

A quadratic formulation in the $\mathbb{R}^n$ space is a real value function (you can say it is a matrix counterpart to a simple linear equation) of the form,

$$F(x_1, x_2, \ldots, x_n) = \sum_{i \leq j} \alpha_{ij} x_i x_j$$

where i, $j = \{1, 2, \ldots, n\}$. This function can be written in matrix form as

$$F(\mathbf{x}) = \mathbf{x}' A \mathbf{x}$$

where A is symmetric matrix. Just as in your experience with linear equation, when optimizing, you need to know whether the function you are working on is concave or convex, so too do you need a similar concept in $\mathbb{R}^n$.

Consider a single variable equation, $y = \alpha x^2$. The second order derivative is just 2α, and as you should recall, you would be maximizing if $\alpha < 0$, and minimizing if $\alpha > 0$. If the equation conforms to the former, we say that it is **positive definite**, and in the latter case it is **negative definite**. Of course the equation need not be that "perfect", and an intermediate concept exists. This occurs when the second order derivatives are ≥ 0 and ≤ 0, and they correspond with **positive semidefinite** and **negative semidefinite**. But this discussion relates to simple functional equations. How about in matrices? What are the rules, or truths that will help you figure out whether you are maximizing or minimizing?

We will be dealing with principally symmetric matrices, and consequently will be defining the concepts with reference to them. For a $k \times k$ symmetric matrix A, and $\mathbf{x}$ being a $k \times 1$ column vector, it is,

1. **Positive Definite** if $\mathbf{x}' A \mathbf{x} > 0, \forall \mathbf{x} \neq 0$ in $\mathbb{R}^k$,
2. **Positive Semi-Definite** if $\mathbf{x}' A \mathbf{x} \geq 0, \forall \mathbf{x} \neq 0$ in $\mathbb{R}^k$,
3. **Negative Definite** if $\mathbf{x}' A \mathbf{x} < 0, \forall \mathbf{x} \neq 0$ in $\mathbb{R}^k$,
4. **Negative Semi-Definite** if $\mathbf{x}' A \mathbf{x} \leq 0, \forall \mathbf{x} \neq 0$ in $\mathbb{R}^k$, and
5. **Indefinite** if $\mathbf{x}' A \mathbf{x} > 0$ for some $\mathbf{x} \in \mathbb{R}^k$, and $\mathbf{x}' A \mathbf{x} < 0$ for other $\mathbf{x} \in \mathbb{R}^k$.

Further, note that a matrix that is positive (negative) definite is also by definition a positive (negative) semi-definite matrix.

There is however a simple test for the definiteness of a quadratic form or a symmetric matrix by focusing on the matrix A. However, before we describe this test, we need to know additional notation about matrices. For a $n \times n$ matrix A, a submatrix or the subset of the matrix A that is obtained from the elimination of $n-k$ rows and the same $n-k$ columns is known as a k^{th} **order principal submatrix of** A. The determinant of a $k \times k$ k^{th} order principal submatrix is known as a k^{th} **order principal minor of** A. As an example, consider a 3×3 matrix such as that below,

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}$$

In the case above there are 3 first order submatrix, 3 second order submatrix, and 1 third order submatrix with the same number of principal minors for each order of submatrix. The three first order principal minors are, $|a_{11}|$, $|a_{22}|$ and $|a_{33}|$. The three second order principal minors are,

1. $$\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}$$

2. $$\begin{vmatrix} a_{11} & a_{13} \\ a_{31} & a_{33} \end{vmatrix}$$

3. $$\begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix}$$

Finally, with reference to the same $n \times n$ matrix. The k^{th} order principal submatrix of A obtained by deleting the *last* $n-k$ rows and columns from the matrix A is called the k^{th} **order leading principal submatrix**, and the determinant of this submatrix is known as the k^{th} **order leading principal minor**. Typically, the leading principal submatrix is denoted by the matrix name and including its order as a subscript. Refering to the 3×3 example, the 3 leading principal minors are,

1. $|A_1| = |a_{11}|$

2. $$|A_2| = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}$$

3. $$|A_3| = \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix}$$

Theorem 1 *For a $n \times n$ symmetric matrix A, we have the following,*

1. *A is positive definite if and only if all the n leading principal minors are strictly positive.*

2. *A is negative definite if and only if its n leading principal minors alternate in sign as follows, $|A_1| < 0$, $|A_2| > 0$, $|A_3| < 0$, ... All more succinctly $(-1)^k |A_k|$ where $k = \{1, 2, \ldots, n\}$.*

3. *When the matrix A does not abide by the above rules or signs, it is said to be a indefinite matrix.*

A possibility that the matrix may fail, in some sense marginally, is when one or some of the leading principal minors are equal to zero. In those cases, we have to check all the principal minors as opposed to just the leading principal minors.

Theorem 2 *For a $n \times n$ symmetric matrix A,*

1. *A is positive semi-definite if and only if every principal minor is greater than or equal to zero, ≥ 0.*

2. *A is negative semi-definite if and only if every principal minor of odd order is less than or equal to 0, and every principal minor of even order is greater than or equal to 0.*

3. *If neither of the above is fulfilled, A is indefinite.*

Consider the 3×3 matrix again,

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}$$

So A is positive semi-definite if and only if $|a_{11}| \geq 0$, $|a_{22}| \geq 0$ and $|a_{33}| \geq 0$, and for the three second order principal minors,

1.
$$\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} \geq 0$$

2.
$$\begin{vmatrix} a_{11} & a_{13} \\ a_{31} & a_{33} \end{vmatrix} \geq 0$$

3.
$$\begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} \geq 0$$

And finally that the determinant of the entire matrix is ≥ 0.

The same matrix is negative semi-definite if and only if, $|a_{11}| \leq 0$, $|a_{22}| \leq 0$ and $|a_{33}| \leq 0$, and for the three second order principal minors,

1.
$$\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} \geq 0$$

2.
$$\begin{vmatrix} a_{11} & a_{13} \\ a_{31} & a_{33} \end{vmatrix} \geq 0$$

3.
$$\begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} \geq 0$$

And that the determinant of the entire matrix is ≤ 0.

1.9 Matrix Differentiation

Let

$$f(\mathbf{b}) = \mathbf{a}'_{(1\times k)}\mathbf{b}_{(k\times 1)}$$

The

$$\frac{\partial \mathbf{a}'\mathbf{b}}{\partial \mathbf{b}} = \frac{\partial \mathbf{b}'\mathbf{a}}{\partial \mathbf{b}} = \mathbf{a}_{(k\times 1)}$$

If

$$f(\mathbf{b}) = \mathbf{b}'A\mathbf{b}$$

if A is symmetric

$$\frac{\partial \mathbf{b}'A\mathbf{b}}{\partial \mathbf{b}} = 2A\mathbf{b}$$

2 Ordinary Least Squares Revisited

Our typical k variable regression is written as follows,

$$y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + ... + b_k x_{ik} + e_i$$

Or in the population form

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i$$

You should realize that this could easily be represented in matrix form for all the n observations. That is

$$\begin{aligned} y_1 &= b_0 + b_1 x_{11} + b_2 x_{12} + \cdots + b_k x_{1k} + e_1 \\ y_2 &= b_0 + b_1 x_{21} + b_2 x_{22} + \cdots + b_k x_{2k} + e_2 \\ &\vdots \\ y_n &= b_0 + b_1 x_{n1} + b_2 x_{n2} + \cdots + b_k x_{nk} + e_n \end{aligned}$$

can be written as

$$\mathbf{y} = X\mathbf{b} + \mathbf{e} \tag{1}$$

where $\mathbf{y}$ is a vector of y_i, X is the matrix of variables where the rows are for each observation, and each column is for a particular variable in question, $\mathbf{b}$ is the vector of slope coefficients, and $\mathbf{e}$ is a vector of residuals or idiosyncratic error terms.

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1k} \\ x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

For the population form, we write

$$\mathbf{y} = X\beta + \epsilon \tag{2}$$

Where β and ϵ are column vectors of the slope coefficient and the population error terms.

Recall that we would like to minimize the sum of square residuals. While that would be equivalent to

$$\begin{aligned} \min_{\mathbf{b}} \mathbf{e}'\mathbf{e} &= \min_{\mathbf{b}} (\mathbf{y} - X\mathbf{b})' (\mathbf{y} - X\mathbf{b}) \\ &= \min_{\mathbf{b}} (\mathbf{y}'\mathbf{y} - 2\mathbf{b}'X'\mathbf{y} + \mathbf{b}'X'X\mathbf{b}) \end{aligned}$$

Therefore the first order condition is

$$\begin{aligned} \tfrac{\partial \mathbf{e}'\mathbf{e}}{\partial \mathbf{b}} &= -2X'\mathbf{y} + 2X'X\mathbf{b} = 0 \\ \Rightarrow \quad X'X\mathbf{b} &= X'\mathbf{y} \\ \Rightarrow \quad \mathbf{b} &= (X'X)^{-1} X'\mathbf{y} \end{aligned} \tag{3}$$

And we have a very succinct way of describing the solution! And that is what you tell the computer to find. Further note that

$$\frac{\partial^2 (\mathbf{e}'\mathbf{e})}{\partial \mathbf{b}^2} = 2X'X > 0$$

The matrix is positive definite (actually all we need is for it to be positive semi-definite), which means the objective function is convex, and consequently we know that we are minimizing the objective function.

Next substituting the population regression expression into our solution for the slope coefficient

$$\begin{aligned} \mathbf{b} &= (X'X)^{-1} X'\mathbf{y} = (X'X)^{-1} X' (X\beta + \epsilon) \\ &= (X'X)^{-1} (X'X)\beta + (X'X)^{-1} X'\epsilon \\ &= \beta + (X'X)^{-1} X'\epsilon \\ &\Rightarrow \mathbf{E}(\mathbf{b}) = \beta \end{aligned}$$

Since

$$\mathbf{E}\left((X'X)^{-1} X'\epsilon\right) = (X'X)^{-1} X'\,\mathbf{E}(\epsilon) = 0_{(k\times 1)}$$

Further the variance-covariance matrix of our OLS estimators is;

$$\begin{aligned} \mathbf{b} - \mathbf{E}(\mathbf{b}) &= (X'X)^{-1} X'\epsilon \\ \Rightarrow \quad \mathbf{Var}(\beta) &= \mathbf{E}\left((X'X)^{-1} X'\epsilon\right)\left((X'X)^{-1} X'\epsilon\right)' \\ &= \mathbf{E}\left((X'X)^{-1} X'\epsilon\epsilon' X (X'X)^{-1}\right) \\ &= (X'X)^{-1} X'\,\mathbf{E}(\epsilon\epsilon') X (X'X)^{-1} \\ &= (X'X)^{-1} X'\sigma_\epsilon^2 \mathbf{I} X (X'X)^{-1} \\ &= \sigma_\epsilon^2 (X'X)^{-1} (X'X) (X'X)^{-1} \\ &= \sigma_\epsilon^2 (X'X)^{-1} \end{aligned}$$

Where $\mathbf{I}$ is an identity matrix, and σ_ϵ^2 is a scalar since if you recall, OLS relies on homogeneity, i.e. the error terms are all normally distributed with mean zero, and variance of σ_ϵ^2. Further note that $\epsilon\epsilon'$ is an $n \times n$ matrix, and the off diagonals are all zero since

we have also assumed that errors between observations are uncorrelated with each other. That is

$$\begin{bmatrix} \sigma_\epsilon^2 & 0 & \dots & 0 \\ 0 & \sigma_\epsilon^2 & \dots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \dots & \sigma_\epsilon^2 \end{bmatrix} = \sigma_\epsilon^2 \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \dots & 1 \end{bmatrix}$$

What this means then is that there are essentially two lines to write in a program;

$$\begin{aligned} \mathbf{b} &= (X'X)^{-1} X'\mathbf{y} \\ \mathbf{Var}(\mathbf{b}) &= \sigma_\epsilon^2 (X'X)^{-1} \end{aligned}$$

Note that the off diagonal elements of the $(X'X)$ matrix are nonzero. Essentially, the explanatory variables or covariates are very likely correlated. Further as you recall, since we typically do not know σ_ϵ^2 we would estimate is as follows

$$\frac{(\mathbf{e}'\mathbf{e})}{n-(k+1)}$$

which adds a third line to the program. To find the vector of t statistics, you will need

$$t = \left(diag\left(s.d.\left(\mathbf{b} \right) \right)_{(k\times k)} \right)^{-1} \mathbf{b}_{(k\times 1)}$$

where $(s.d.(\mathbf{b}))$ would be the vector of standard deviations, and $diag(.)$ changes the off diagonal elements to 0, and this is the final line.

This is just the technical element to ordinary least squares. We will discuss the geometry of OLS in the next section. Before we begin, there are two note worthy points you should keep in mind.

Remark 1 *From our understanding of the geometry of vectors, there is something interesting that can be said about the ordinary least squares method of estimation. We know that*

$$\begin{aligned} \mathbf{y} &= X\mathbf{b} + \mathbf{e} \\ \Rightarrow \mathbf{e} &= \mathbf{y} - X\mathbf{b} \end{aligned}$$

We can think of X as the $k+1$ column vectors that span the $\mathbb{R}^{k+1}$ space. Essentially, as noted in the idea of OLS, we are trying to minimize the errors from our estimation. This is equivalent to choosing a $\mathbf{b}$ vector that minimizes the distance between the $X\mathbf{b}$

vector, and that of $\mathbf{y}$*. That distance is* $\mathbf{e}$*. Well, this is achieved only when the vector* $\mathbf{e}$ *is perpendicular to the hyperplane generated by the* $k+1$ *columns of* X*. Thus* $\mathbf{e}$ *must be orthogonal to any linear combination of* X*. Let* $\mathbf{c}$ *be a vector such that* $X\mathbf{c}$ *is an arbitrary linear combination of the* $k+1$ *columns in the* k *variable regression. Then*

$$\begin{aligned} c'X'(\mathbf{y}-X\mathbf{b}) &= c'(X'\mathbf{y}-X'X\mathbf{b})=0 \\ \Rightarrow \quad X'\mathbf{y} &= X'X\mathbf{b} \\ \Rightarrow \quad \mathbf{b} &= (X'X)^{-1}X'\mathbf{y} \end{aligned}$$

which is what OLS does!

Remark 2 *There is another concern, although I have assumed that* $(X'X)$ *is a nonsingular matrix, implying a unique inverse exists, as noted in the rank of a matrix, this need not be true. If so, we may face problems in obtaining our parameter estimates. The trick is to use decompositions that eliminate the need for finding an inverse. As a matter of computation, if a program eliminates the need for finding an inverse, it is in fact computationally more efficient. The most commonly used decomposition is the QR decomposition, the idea of which I will barely graze. The reason I am bringing this up is because this will affect the manner in which we should write our program (specifically in* MATLAB*). QR decomposition decomposes a matrix* X *into the product of an orthonormal matrix* Q *(*$Q'Q=I$*) and a upper triangular matrix* R*.*

$$\begin{aligned} X'\mathbf{y} &= X'X\mathbf{b} \\ \Rightarrow \quad (QR)'\mathbf{y} &= (QR)'QR\mathbf{b} \\ \Rightarrow \quad R'Q'\mathbf{y} &= R'Q'QR\mathbf{b} \\ &= R'R\mathbf{b} \\ \Rightarrow \quad \mathbf{b} &= (R'R)\backslash R'Q'\mathbf{y} \\ \Rightarrow \quad \mathbf{b} &= R'\backslash Q'\mathbf{y} \end{aligned}$$

In this form there is no need to calculate any inverse. We will be doing this in MATLAB*. In truth, what I have done above is uninformative for you. What you need to realize is that what we are doing here is for computational reasons as opposed to algebraic.*

3 Geometry of OLS Estimation

We have solved the Ordinary Least Squares problem twice now, and we have a good idea about the intuition of the method. We will go a little further now by understanding the geometry behind it, more precisely, we will use your knowledge of Matrix & Linear Algebra now. The following notes are culled from both Johnston (1984) and Davidson and MacKinnon (2004).

Before we go on, we will formally introduce some terms, notations, and concepts. A real number is said to be lying on a real line, and is denoted as $\mathbb{R}$. In the case, of a $n-$vector, which is just a column vector with n elements (or a $n \times 1$ matrix), we say it that belongs to a set of $n-$vectors in $\mathbb{R}^n$. We can also say the the $n-$vector is in a **Euclidean space** of n dimensions, and we denote it as E^n. All operations in the $\mathbb{R}^n$ space still applies in E^n, but in addition there is the scalar or inner product. For two vectors $\mathbf{x}$ and $\mathbf{y} \in E^n$, the scalar product as you know is

$$\langle \mathbf{x}, \mathbf{y} \rangle \equiv \mathbf{x}'\mathbf{y}$$

Recall also that $\langle \mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{y}, \mathbf{x} \rangle$, which as you should know, is because the scalar product is commutative.

The importance of the scalar product is that it allows us to link matrix algebra, and the geometry of vectors, or more precisely, it allows us to define length or distance of any vector in E^n. The length of a vector $\mathbf{x}$ is also known is its **norm** and is written as,

$$\begin{aligned} ||\mathbf{x}|| &\equiv (\mathbf{x}'\mathbf{x})^{\frac{1}{2}} \\ &\equiv \left(\sum_{i=1}^{n} x_i^2 \right)^{\frac{1}{2}} \end{aligned}$$

Each $n-$vector essentially defines a point on E^n. This means then that $\mathbf{y}$, and each column of the matrix X in a regression model defines a point in E^n, which then allows us to represent the regression model geometrically. The obvious superficial limitation is that we would have problem representing a $n-$vector diagrammatically. However, if we look close at our population regression equation (2), you should notice that there are only three vectors in the equation. Further, since the left hand side of equation (2) has only $X\beta$ and ϵ only, we can represent the equation in 2 dimensions.

Before we can try to understand how it works, we have to know additional definitions and notations. In Euclidean space E^n, there are infinitely numerous points. Any collection

of such vectors would **span** (or make) a space in E^n. We call this space, a **subspace**, and the vectors that span this subspace, **basis vectors**. For k vectors $\mathbf{x}_i$, $i \in \{1, 2, \ldots, k\}$, that span a subspace, we denote it as $\mathfrak{S}(X) \equiv \mathfrak{S}(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k)$. This is a $k-$dimensional subspace. Further, the subspace $\mathfrak{S}(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k)$ includes all vectors that can be formed as linear combinations of $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k)$, or precisely,

$$\mathfrak{S}(x_1, x_2, \ldots, x_k) \equiv \left\{ z \in E^n | \, z = \sum_{i=1}^{k} b_i x_i, b_i \in \mathbb{R} \right\} \tag{4}$$

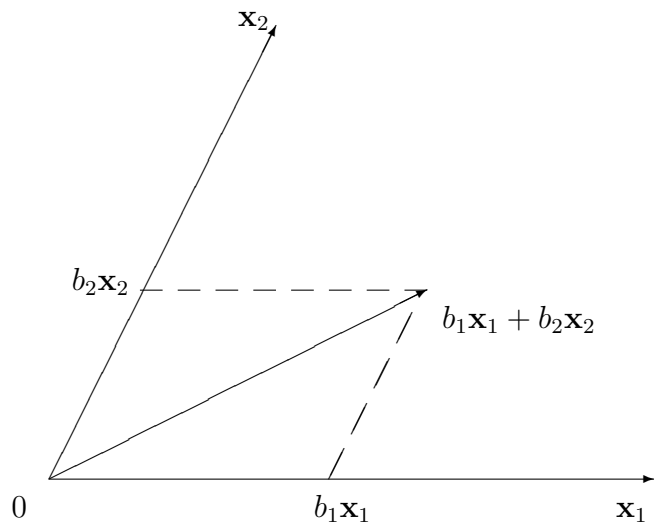
We can also say that the subspace defined above is **span of X** or the **column space of X**. The orthogonal complement of the space $\mathfrak{S}(X)$ is denoted as $\mathfrak{S}^{\perp}(X)$, which is simply the set of all vectors $\mathbf{w}$ in E^n that are orthogonal to everything in $\mathfrak{S}(X)$. In other words, $\langle \mathbf{w}, \mathbf{z} \rangle = \mathbf{w}'\mathbf{z} = 0$. More precisely, we define,

$$\mathfrak{S}^{\perp}(X) \equiv \{ \mathbf{w} \in E^n | \, \mathbf{w}'\mathbf{z} = 0, \forall \mathbf{z} \in \mathfrak{S}(X) \} \tag{5}$$

Further, if the dimension of $\mathfrak{S}(X)$ is k, then the dimension of its complement is $n - k$.

Let us go through an example in two dimensions. Let $\mathbf{x}_1$ and $\mathbf{x}_2$ be two arbitrary vectors where $\mathbf{x}_1 \neq \mathbf{x}_2$ with respective length $||\mathbf{x}_1||$ and $||\mathbf{x}_2||$ extending from a common origin. This is depicted below in figure 1, Then the subspace is just the plane formed by

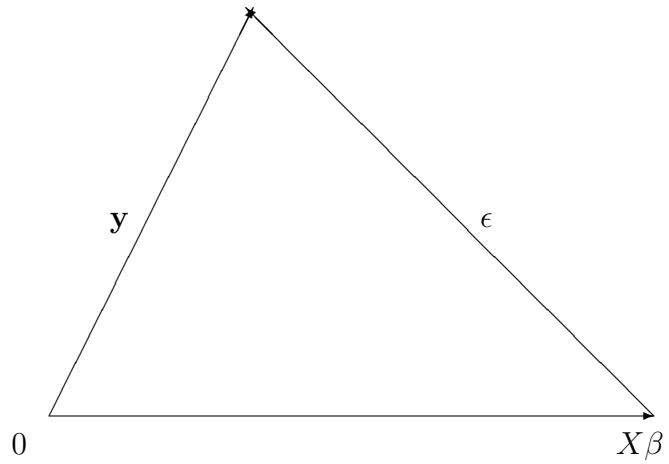
Figure 1: 2−Dimensional Subspace



all the linear combinations of the two vectors $\mathbf{x}_1$ and $\mathbf{x}_2$. This idea extends to technically $n-$dimensions.

Using this idea, we can then represent the regression equation geometrically as in the diagram below in figure 2. Note that we have drawn this diagram with the error

Figure 2: Geometry of OLS



vector ϵ not being orthogonal to the $X\beta$ vector, but that is not to say that the solution is represented this way. We will next define how the solution will look like geometrically.

Let X be the $n \times k$ matrix of **sample** variables, now it becomes clear that $X\mathbf{b}$ is just an $n-$vector in $\mathfrak{S}(X)$ (recall that $\mathbf{b}$ is a $k \times 1$ vector), which in turn is a $k-$dimensional subspace of E^n. From equation (3) recall that the OLS solution is,

$$X'\mathbf{y} - X'X\mathbf{b} = 0 \tag{6}$$

$$\Rightarrow X'(\mathbf{y} - X\mathbf{b}) = 0 \tag{7}$$

Note that the left hand side above is a $k \times 1$ matrix, and each element of the vector is a scalar product,

$$x_i'(\mathbf{y} - X\mathbf{b}) = \langle x_i, \mathbf{y} - X\mathbf{b} \rangle = 0 \tag{8}$$

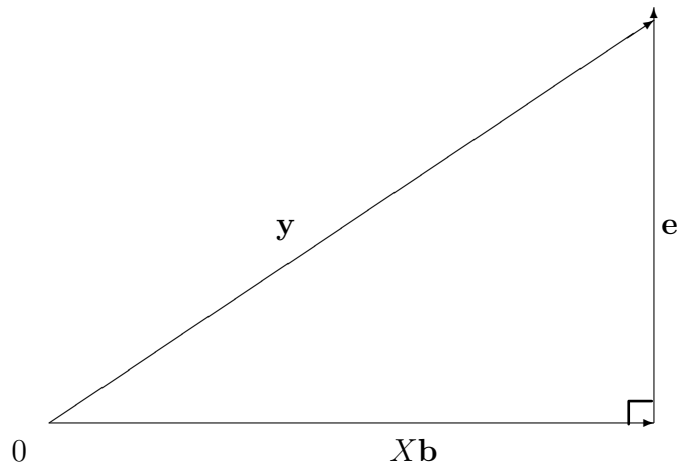
What this means is that each vector of X, which represents each one of the k explanatory variables, is orthogonal to $\mathbf{y} - X\mathbf{b}$. Recall the Gauss-Markov assumption that the errors (since $\mathbf{e} = \widehat{\epsilon} = \mathbf{y} - X\mathbf{b}$) must be independent of the explanatory variables? For this reason, the first order condition of the OLS's objective function is also commonly referred to as the **orthogonality conditions**. Another way to think of this is as follows. Since

$X\mathbf{b}$ is in $\mathfrak{S}(X)$, $\mathbf{e}$ is in fact orthogonal to every vector of $\mathfrak{S}(x)$, the span of X. So that,

$$\langle X\mathbf{b}, \mathbf{e}\rangle = (X\mathbf{b})'\mathbf{e} = b'X'\mathbf{e} = 0 \tag{9}$$

Since $\mathbf{y} = X\mathbf{b} + \mathbf{e}$, and because of the orthogonality condition, the relationship between $X\mathbf{b}$ and $\mathbf{e}$ is as follows in figure 3. This idea can be further consolidated in 3 dimensions.

Figure 3: Geometric Relationship between $X\mathbf{b}$ and $\mathbf{e}$

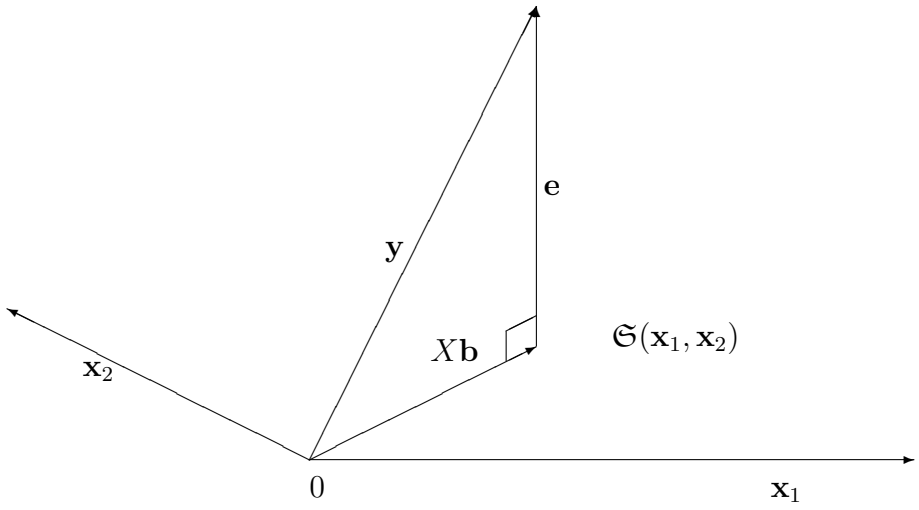


Consider two regressors $\mathbf{x}_1$ and $\mathbf{x}_2$ which spans $\mathfrak{S}(\mathbf{x}_1, \mathbf{x}_2)$ (which is just a two dimensional horizontal plane). Further, as before $\mathbf{y}$ is a vector with length $||\mathbf{y}||$ that extends from the same origin. Since $\mathbf{e}$ is orthogonal to the horizontal plane spanned by $\mathbf{x}_1$ and $\mathbf{x}_2$, it must be a line segment that extends vertically from the horizontal plane $\mathfrak{S}(\mathbf{x}_1, \mathbf{x}_2)$ to $\mathbf{y}$, as depicted in figure 4. You might be asking yourself if this is really what the minimization of the sum of squared errors is doing? Is it really minimizing the distance between the $\mathbf{y}$ vector and the subspace spanned by X? First note that the length/norm of $\mathbf{e}$ is $||\mathbf{e}|| = (\sum_{i=1}^{n} e_i^2)^{1/2}$. But the objective function of the ordinary least squares problem is just $||\mathbf{e}||^2$. Since minimizing the norm is the same as minimizing the square of the norm, this implies the estimator $\mathbf{b}$ of β of OLS does minimize the length $\mathbf{e}$.

Finally, by the Pythagoras' Theorem,

$$\begin{aligned} ||\mathbf{y}||^2 &= ||X\mathbf{b}||^2 + ||\mathbf{e}||^2 && (10) \\ \Rightarrow \mathbf{y}'\mathbf{y} &= (X\mathbf{b})'X\mathbf{b} + \mathbf{e}'\mathbf{e} \\ &= \mathbf{b}'X'X\mathbf{b} + (\mathbf{y} - X\mathbf{b})'(\mathbf{y} - X\mathbf{b}) && (11) \end{aligned}$$

Figure 4: 3−Dimensional Geometric Relationship between $X\mathbf{b}$ & $\mathbf{e}$



where the left hand side is the total sum of squares, and it is equal to the right hand side, which is the sum of the explained sum of squares and the residual sum of squares as we have learned earlier.

4 Extension: Orthogonal Projections

What we have described geometrically is essentially an orthogonal projection when we mapped the vector of $\mathbf{y}$ onto the subspace spanned by X. "A **projection** is a mapping that takes each point of E^n into a point in a subspace of E^n, while leaving all points in that subspace unchanged." (Davidson and MacKinnon 2004) Since the points in this subspace remains unadultered, it is called the **invariant subspace** of the projection. Then a **orthogonal projection** is a projection that maps all points of E^n into a point on the subspace that is *closest* to it. Obviously, there are points in E^n that are already in the invariant subspace, so that an orthogonal projection would leave such points unchanged.

An orthogonal projection formalizes the mapping we performed when we mapped in the previous section the vector $\mathbf{y}$ onto subspace $\mathfrak{S}(X)$ perpendicularly. Technically, we can always perform the procedure by pre−multiplying the vector to be mapped by a **projection matrix**. What is the projection matrix for the OLS procedure we performed?

Recall first your OLS solution for β,

$$\begin{aligned} \mathbf{b} &= (X'X)^{-1}(X'\mathbf{y}) && (12) \\ \Rightarrow X'\mathbf{b} &= (X'(X'X)^{-1}X')\mathbf{y} \\ &= P_X\mathbf{y} && (13) \end{aligned}$$

so that $P_X = (X'(X'X)^{-1}X')$ is the projection matrix that projects $\mathbf{y}$ onto $\mathfrak{S}(X)$. Note that $P_X X = X(X'X)^{-1}X'X = X$ which fulfils the definition of an orthogonal projection since X is already on $\mathfrak{S}(X)$, and it is easy to show that the same is true for $P_X X\mathbf{b}$, since $X\mathbf{b}$ is in $\mathfrak{S}(X)$. In other words, the image of P_X is $\mathfrak{S}(X)$ itself.

We also know that the definition of the residual is,

$$\begin{aligned} \mathbf{e} &= \mathbf{y} - X\mathbf{b} && (14) \\ &= \mathbf{y} - (X'(X'X)^{-1}X')\mathbf{y} \\ &= (\mathbf{I} - (X'(X'X)^{-1}X'))\mathbf{y} && (15) \\ &= (\mathbf{I} - P_X)\mathbf{y} && (16) \\ &= M_X\mathbf{y} && (17) \end{aligned}$$

So that M_X is an orthogonal projection onto the complement of $\mathfrak{S}(X)$, $\mathfrak{S}^{\perp}(X)$. Put another way, the image of M_X is the orthogonal complement of P_X. Note that M_X is a symmetric matrix. To see that M_X is an orthogonal projection onto the complement,

$$M_X X = (I - P_X)X = X - X = \mathbf{0} \quad (18)$$

where $\mathbf{0}$ is a $(n \times k)$ zero matrix. M_X is sometimes referred to as the projection off $\mathfrak{S}(X)$.

Note that a projection matrix must be idempotent. A matrix is idempotent when multiplying the matrix by itself, gives itself again. That is $P_X P_X = P_X$ and $M_X M_X = M_X$. This condition is quite intuitive, since the second projection cannot possibly affect, or do anything more than what the first projection has already done. Nonetheless, you can show this yourself, i.e. that P_X and M_X are idempotent.

Finally, from equation (17),

$$\begin{aligned} \mathbf{I} - P_X &= M_X \\ \Rightarrow M_X + P_X &= \mathbf{I} && (19) \end{aligned}$$

so that $(M_X + P_X)\mathbf{y} = \mathbf{y}$ and we say that the pair of projections P_X and M_X are **complementary projections** that together restores the vector, in this case $\mathbf{y}$, it was meant to project.

References

Davidson, Russell and James G. MacKinnon, *Econometric Theory and Methods*, first ed., Oxford, Oxford University Press, 2004.

Johnston, Jack, *Econometric Methods*, third ed., New York, McGraw-Hill Inc, 1984.