Chapter 3

Least Squares

Notation: y_i is a scalar, and $x_i = (x_{i1}, \ldots, x_{iK})'$ is a $K \times 1$ vector. $Y = (y_1, \ldots, y_n)'$ is an $n \times 1$ vector, and

$$X = \begin{bmatrix} x_1' \\ x_2' \\ \vdots \\ x_n' \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1K} \\ x_{21} & x_{22} & \cdots & x_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{22} & \cdots & x_{nK} \end{bmatrix}$$

is an $n \times K$ matrix. I_n is an $n \times n$ identity matrix.

Ordinary least squares (OLS) is the most basic estimation technique in econometrics. It is simple and transparent. Understanding it thoroughly paves the way to study more sophisticated linear estimators. Moreover, many nonlinear estimators resemble the behavior of linear estimators in a neighborhood of the true value. In this lecture, we learn a series of facts from the linear algebra operation.

3.1 Estimator

As we have learned from the linear projection model, the projection coefficient β in the regression

$$y = x'\beta + e$$

can be written as

$$\beta = \left(E\left[xx' \right] \right)^{-1} E\left[xy \right]. \tag{3.1}$$

We draw a pair of (y, x) from the joint distribution, and we mark it as (y_i, x_i) for $i = 1, \ldots, n$ repeated experiments. We possess a *sample* $(y_i, x_i)_{i=1}^n$.

Remark 3.1. Is (y_i, x_i) random or deterministic? Before we make the observation, they are treated as random variables whose realized values are uncertain. (y_i, x_i) is treated as random when we talk about statistical properties — statistical properties of a fixed number is meaningless. After we make the observation, they become deterministic values which cannot vary anymore.

Remark 3.2. In reality, we have at hand fixed numbers (more recently, words, photos, audio clips, video clips, etc., which can all be represented in digital formats with 0 and 1)

to feed into a computational operation, and the operation will return one or some numbers. All statistical interpretation about these numbers are drawn from the probabilistic thought experiments. A *thought experiment* is an academic jargon for a *story* in plain language. Under the axiomatic approach of probability theory, such stories are mathematical consistent and coherent. But mathematics is a tautological system, not science. The scientific value of a probability model depends on how close it is to the *truth* or implications of the truth. In this course, we suppose that the data are generated from some mechanism, which is taken as the truth. In the linear regression model for example, the joint distribution of (y, x) is the truth, while we are interested in the linear projection coefficient β, which is an implication of the truth as in (3.1). Probabilists suppose there is a dragon and try to tell the dragon's behaviors. Statisticians observe many snakes on earth, and try to tell what a dragon looks like.

The sample mean is a natural estimator of the population mean. Replace the population mean $E[\cdot]$ in (3.1) by the sample mean $\frac{1}{n} \sum_{i=1}^{n} \cdot$, and the resulting estimator is

$$\widehat{\beta} = \left(\frac{1}{n} \sum_{i=1}^{n} x_i x_i' \right)^{-1} \frac{1}{n} \sum_{i=1}^{n} x_i y_i = \left(\frac{X'X}{n} \right)^{-1} \frac{X'y}{n} = (X'X)^{-1} X'y$$

if $X'X$ is invertible. This is one way to motivate the OLS estimator.

The operation of OLS bears a natural geometric interpretation. Notice $\mathcal{X} = \left\{ Xb : b \in \mathbb{R}^K \right\}$ is the linear space spanned by the K columns of $X = [X_{\cdot 1}, \ldots, X_{\cdot K}]$, which is of K-dimension if the columns are linearly independent. The OLS estimator is the minimizer of $\min_{b \in \mathbb{R}^K} \|Y - Xb\|$ (Square the Euclidean norm or not does not change the minimizer because a^2 is a monotonic transformation for $a \geq 0$). In other words, $X\widehat{\beta}$ is the point in $\mathcal{X}$ such that it is the closest to the vector Y in terms of the Euclidean norm.

The relationship $Y = X\widehat{\beta} + \widehat{e}$ decomposes Y into two orthogonal vectors $X\widehat{\beta}$ and $\widehat{e}$ as $\left\langle X\widehat{\beta}, \widehat{e} \right\rangle = \widehat{\beta}' X' \widehat{e} = 0_K$, where $\langle \cdot, \cdot \rangle$ is the *inner product* of two vectors. Therefore $X\widehat{\beta}$ is the *projection* of Y onto $\mathcal{X}$, and $\widehat{e}$ is the corresponding *projection residuals.* The Pythagorean theorem implies $\|Y\|^2 = \|X\widehat{\beta}\|^2 + \|\widehat{e}\|^2$.

3.2 Subvector

The Frish-Waugh-Lovell (FWL) theorem is an algebraic fact about the formula of a subvector of the OLS estimator. To derive the FWL theorem we need to use the inverse of partitioned matrix. For a positive definite symmetric matrix $A = \begin{pmatrix} A_{11} & A_{12} \\ A_{12}' & A_{22} \end{pmatrix}$, the inverse can be written as

$$A^{-1} = \begin{pmatrix} \left(A_{11} - A_{12} A_{22}^{-1} A_{12}' \right)^{-1} & -\left(A_{11} - A_{12} A_{22}^{-1} A_{12}' \right)^{-1} A_{12} A_{22}^{-1} \\ -A_{22}^{-1} A_{12}' \left(A_{11} - A_{12} A_{22}^{-1} A_{12}' \right)^{-1} & \left(A_{22} - A_{12}' A_{11}^{-1} A_{12} \right)^{-1} \end{pmatrix}.$$

In our context of OLS estimator, let $X = \begin{pmatrix} X_1 & X_2 \end{pmatrix}$

$$\begin{pmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_2 \end{pmatrix} = \widehat{\beta} = (X'X)^{-1}X'Y$$

$$= \left(\begin{pmatrix} X'_1 \\ X'_2 \end{pmatrix} (X_1 \quad X_2) \right)^{-1} \begin{pmatrix} X'_1 Y \\ X'_2 Y \end{pmatrix}$$

$$= \begin{pmatrix} X'_1 X_1 & X'_1 X_2 \\ X'_2 X_1 & X'_2 X_2 \end{pmatrix}^{-1} \begin{pmatrix} X'_1 Y \\ X'_2 Y \end{pmatrix}$$

$$= \left(\underset{\heartsuit}{\left(X'_1 M'_{X_2} X_1 \right)^{-1}} \quad -\underset{\heartsuit}{\left(X'_1 M'_{X_2} X_1 \right)^{-1}} X'_1 X_2 \left(X'_2 X_2 \right)^{-1} \right) \begin{pmatrix} X'_1 Y \\ X'_2 Y \end{pmatrix}.$$

The subvector

$$\begin{aligned} \widehat{\beta}_1 &= \left(X'_1 M'_{X_2} X_1 \right)^{-1} X'_1 Y - \left(X'_1 M'_{X_2} X_1 \right)^{-1} X'_1 X_2 \left(X'_2 X_2 \right)^{-1} X'_2 Y \\ &= \left(X'_1 M'_{X_2} X_1 \right)^{-1} X'_1 Y - \left(X'_1 M'_{X_2} X_1 \right)^{-1} X'_1 P_{X_2} Y \\ &= \left(X'_1 M'_{X_2} X_1 \right)^{-1} \left(X'_1 Y - X'_1 P_{X_2} Y \right) \\ &= \left(X'_1 M'_{X_2} X_1 \right)^{-1} X'_1 M_{X_2} Y. \end{aligned}$$

Notice that $\widehat{\beta}_1$ can be obtained by the following:

1. Regress Y on X_2, obtain the residual $\tilde{Y}$;

2. Regress X_1 on X_2, obtain the residual $\tilde{X}_1$;

3. Regress $\tilde{Y}$ on $\tilde{X}_1$, obtain OLS estimates $\widehat{\beta}_1$.

Similar derivation can also be carried out in the population linear projection. See Hansen (2020) [E] Chapter 2.22-23.

3.3 Summary

Historical notes: Carl Friedrich Gauss (1777–1855) claimed he had come up with the operation of OLS in 1795. With only three data points at hand, Gauss successfully applied his method to predict the location of the dwarf planet Ceres in 1801. While Gauss did not publish the work on OLS until 1809, Adrien-Marie Legendre (1752–1833) presented this method in 1805. Today people tend to attribute OLS to Gauss, assuming that a giant like Gauss had no need to tell a lie to steal Legendre's discovery.

```
Zhentao Shi.  January 15, 2025
```