Chapter 3

Least Squares

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

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

is an $n \times p$ 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

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

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

3.2 Geometry of OLS

There is natural geometry interpretation of the OLS estimator in a n-dimensional Euclidean space. Notice $\mathcal{X} = \{Xb : b \in \mathbb{R}^p\}$ is the linear space spanned by the p columns of $X = \left[X_{\cdot 1}, \ldots, X_{\cdot p}\right]$, which is of p-dimension if the columns are linearly independent. The OLS estimator is the minimizer of $\min_{b\in\mathbb{R}^p} \|Y - Xb\|$ (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.

Define

$$\widehat{Y} = X\widehat{\beta} = X(X'X)^{-1}X'Y = P_X Y,$$

where $P_X = X(X'X)^{-1}X'$ is the projector to the columns space of X. One the other hand, define

$$\widehat{e} = Y - \widehat{Y} = (I_n - P_X)Y = P_X^{\perp} Y,$$

where $P_X^{\perp}$ is the projector to the null space of X. Since the null space and the column space are orthogonal, it is easy to verify that

$$\left\langle X\widehat{\beta}, \widehat{e} \right\rangle = \widehat{\beta}' X'\widehat{e} = 0_p'.$$

The Pythagorean theorem implies

$$\|Y\|^2 = \|X\widehat{\beta}\|^2 + \|\widehat{e}\|^2.$$

If X has full column rank and we use a SVD

$$X = \underset{(n\times n)}{U}\underset{(p\times p)}{S}\underset{(p\times p)}{V}'$$

where U and V have orthonormal columns and S is diagonal with the p singular values of X, then the fitted values can be written as

$$\widehat{Y} = X\widehat{\beta} = X(X'X)^{-1}X'Y = UU'Y.$$

In other words, in terms of in-sample fitting, OLS using regressors X is equivalent to regressing Y on the orthonormal basis U for the column space of X (the two procedures have the same projection $\widehat{Y}$). The corresponding coefficients satisfy $\widehat{\beta} = VS^{-1}U'Y$.

3.3 FWL Theorem

The Frisch-Waugh-Lovell (FWL) theorem is an algebraic fact about the formula of a subvector of the OLS estimator. Decompose $X = (X_1, X_2)$ where X_1 is $n \times (p-1)$ matrix and X_2 is an $n \times 1$ vector, so that

$$Y = X_1\widehat{\beta}_1 + X_2\widehat{\beta}_2 + \widehat{e}. \tag{3.2}$$

To find $\widehat{\beta}_2$, we first decompose X_2 into its projection onto span(X_1) and its projection residual:

$$X_2 = P_1X_2 + P_1^{\perp}X_2,$$

where $P_1 = X_1(X_1'X_1)^{-1}X_1'$ is the projector onto span(X_1) and $P_1^{\perp} = I_n - P_1$. Pre-multiply $P_1^{\perp}$ on both sides of (3.2), we have

$$P_1^{\perp}Y = P_1^{\perp}X_1\widehat{\beta}_1 + P_1^{\perp}X_2\widehat{\beta}_2 + P_1^{\perp}\widehat{e} = P_1^{\perp}X_2\widehat{\beta}_2 + \widehat{e}$$

since $P_1^{\perp}X_1 = 0$ and $P_1^{\perp}\widehat{e} = \widehat{e}$.

Because $\tilde{X}_2 = P_1^{\perp}X_2$ is orthogonal to span(X_1), the coefficient on X_2 in the full regression is pinned down entirely by $\tilde{X}_2$:

$$\widehat{\beta}_2 = (\tilde{X}_2'\tilde{X}_2)^{-1}\tilde{X}_2'Y = (X_2'P_1^{\perp}X_2)^{-1}X_2'P_1^{\perp}Y.$$

3.4 Omitted Variable Bias

Suppose that researcher A runs a full regression $Y = X_1\widehat{\beta}_1 + X_2\widehat{\beta}_2 + \widehat{e}$. When he sends the data to researcher B, he omits X_2 by accident. Research B can only estimated the coefficient for X_1. What would be its different from researcher A's $\widehat{\beta}_1$?

Notice in researcher A's world:

$$\begin{aligned}
Y &= X_1\widehat{\beta}_1 + X_2\widehat{\beta}_2 + \widehat{e} \\
&= X_1\widehat{\beta}_1 + (P_1X_2 + P_1^{\perp}X_2)\widehat{\beta}_2 + \widehat{e} \\
&= (X_1\widehat{\beta}_1 + P_1X_2\widehat{\beta}_2) + (P_1^{\perp}X_2\widehat{\beta}_2 + \widehat{e}) \\
&= X_1(\widehat{\beta}_1 + \widehat{\pi}\widehat{\beta}_2) + (P_1^{\perp}X_2\widehat{\beta}_2 + \widehat{e}),
\end{aligned}$$

where $\widehat{\pi} = (X_1'X_1)^{-1}X_1'X_2$. In other words, researcher B will obtain the coefficient $(\widehat{\beta}_1 + \widehat{\pi}\widehat{\beta}_2)$ associated with X_1, and his residual is $(P_1^{\perp}X_2\widehat{\beta}_2 + \widehat{e})$.

Intuitively, regressing on X_1 only forces the component $P_1X_2 = X_1\widehat{\pi}$ to be absorbed by the coefficient on X_1, while the orthogonal component $P_1^{\perp}X_2\widehat{\beta}_2$ becomes part of the residual.

3.5 Summary

The derivations above are finite-sample linear algebra in $\mathbb{R}^n$: X_1 and X_2 span subspaces of $\mathbb{R}^n$, and P_1 and $P_1^{\perp}$ are ordinary projection matrices. There are population counterparts of both the FWL theorem and the omitted variable bias formula. In that setting, Y and the components of X are random variables that live in a Hilbert space (typically an L^2 space), and "projection" is defined by the linear projection (orthogonality) operator rather than an $n \times n$ matrix. The same logic goes through with population inner products (expectations) replacing sample inner products, but the geometry is more abstract, so we omit the full display.

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 14, 2026