**POLSCI 630 Introduction to Regression Methods** Spring 2026

# Lecture 2 — Algebra of Regression Estimation

*Jiawei Fu, Duke University* *Scribe: Jiawei Fu*

## 1 Overview

In the last lecture, we discussed the target parameter and identification assumptions in different linear regression models. In this lecture, we will work with sample data and use it to estimate the parameter.

## 2 Algebra of Least Squares

In econometrics, we assume that our data is a random sample from population.

**Assumption 1.** *The random variables* $\{(Y_1, X_1), ...(Y_n, X_n)\}$ *are independent and identically distributed; they are draws from a common distribution* $P$*.*

Throughout the course, we use the matrix $n \times k$

$$X = \begin{bmatrix} X_{11} & X_{12} & ... & X_{1k} \\ \vdots & \vdots & ... & \vdots \\ X_{i1} & X_{i2} & ... & X_{ik} \\ \vdots & \vdots & ... & \vdots \\ X_{n1} & X_{n2} & ... & X_{nk} \end{bmatrix}$$

to denote the explanatory variables of the sample. Each row is for an individual $i$, each column is for each variable $X_k$. Similarly, the outcome of the sample is the $n \times 1$ vector $Y = [Y_1, ..., Y_n]^T$.

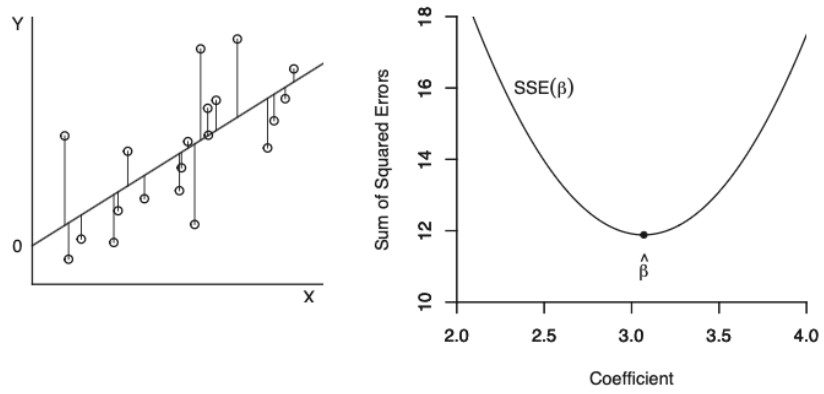
To avoid confusion, I will add subscript $i$ to denote the population model. For example, for the structural linear model, under identification assumptions, the parameter is identified as $\beta = (\mathbb{E}[X_i X_i'])^{-1}\mathbb{E}[X_i Y_i]$. $X_i$ is a vector $n \times 1$ containing $k$ independent variables $X_{i1}, ..., X_{ik}$.

How should we use data to estimate it? When the parameter $\theta = \mathbb{E}[Y]$ is an expectation, we know that the best estimator is the mean of the sample $\frac{1}{n}\sum_{i=1}^n Y_i$. We use the sample moment to replace the population moment.

**Remark 1.** *Formally, this is the idea of the plug-in estimator. The parameter* $\theta = \psi(P)$ *is a function of the distribution. Then, a plug-in estimator is* $\hat{\theta} = \psi(\hat{P})$*, where* $\hat{P}$ *is the empirical distribution* $\mathbb{P}_n(A) = \frac{1}{n}\sum_{i=1}^n I(X_i \in A)$*. For example,* $\theta = \mathbb{E}[Y_i] = \int Y_i dP$*. The plug-in estimator is* $\hat{\theta} = \int Y d\mathbb{P}_n = \frac{1}{n}\sum_{i=1}^n Y_i$*.*

Therefore, for $\beta = (\mathbb{E}[XX'])^{-1}\mathbb{E}[XY]$, by the plug-in principle, we replace the population estimator with the sample averages. We obtain the estimator (verify the last equation by yourself)

$$\hat{\beta} = (\frac{1}{n}\sum_{i=1}^n X_i X_i')^{-1}(\frac{1}{n}\sum_{i=1}^n X_i Y_i) = (\sum_{i=1}^n X_i X_i')^{-1}(\sum_{i=1}^n X_i Y_i) = (X'X)^{-1}X'Y.$$




(a) Deviation from Fitted Line

(b) Sum of Squared Error Function

Figure 1: From Hansen (2022)

Traditionally, in structural econometrics, people often specify a population criterion function $Q : \Theta \to \mathcal{R}$, which is uniquely maximized in the true parameter. The choice of $Q$ and the existence of $\theta_0$ is suggested by the identification of the model. One general way to estimate $\theta_0$ is by maximizing $\hat{Q}_n$, the empirical criterion. Then we get the *extremum* estimate.

Now, let us apply this for BLP, and we will see why the estimate is called least squares. Recall $\beta = argmin\mathbb{E}\left[(Y - X'\beta)^2\right]$. By the sample analog, we can define $\hat{\beta} = argmin\frac{1}{n}\sum_{i=1}^n (Y_i - X_i'\beta)^2$. We call $\sum_{i=1}^n (Y_i - X_i'\beta)^2$ the sum of squared errors, SSE. In other words, the ordinary least squares (OLS) estimator $\hat{\beta}$ is the minimizer of SSE.

We will show summation form and matrix form.

**Theorem 2.** *If $\sum_{i=1}^n X_i X_i' > 0$, the least squares estimator is unique and equal*

$$\hat{\beta} = (\sum_{i=1}^n X_i X_i')^{-1}(\sum_{i=1}^n X_i Y_i) = (X'X)^{-1}XY$$

*Proof.* Recall, $SSE(\beta) = \sum_{i=1}^n Y_i^2 - 2\beta' \sum_{i=1}^n X_i Y_i + \beta' \sum_{i=1}^n X_i X_i'\beta$

Matrix form: $SSE = (Y - X\beta)'(Y - X\beta) = Y'Y - 2Y'X\beta + \beta' X'X\beta$

Take FOC: $0 = -2\sum_{i=1}^n X_i Y_i + 2\sum_{i=1}^n X_i X_i'\hat{\beta}$. Note that $2\sum_{i=1}^n X_i X_i'\hat{\beta} = \sum_{i=1}^n X_i Y_i$ is a system of $k$ equations with $k$ unknowns.

Matrix form: FOC: $0 = -2X'Y + 2X'X\hat{\beta}$. Then we obtain the least squares normal equation $X'X\hat{\beta} = X'y$.

Again, as we prove the coefficient of BLP, we obtain $\hat{\beta} = (\sum_{i=1}^n X_i X_i')^{-1}(\sum_{i=1}^n X_i Y_i)$ if $\sum_{i=1}^n X_i X_i'$ is invertible.

Matrix form: we need the inverse of $X'X$ to exist. Then $\hat{\beta} = (X'X)^{-1}XY$.

To be complete, we also need SOC: $2\sum_{i=1}^n X_i X_i'$ is positive definite. □

We emphasize the importance of the existence of the inverse of $X'X$. We need it for identification

and estimation. The column of $X$ should be linearly independent and there should be at least $K$ observations. However, in practice, it is quite easy to attempt to calculate a regression with linearly dependent regressors. This can occur for many reasons, including the following:

- Including the same regressor twice.
- Including regressors that are a linear combination of one another.
- Including a dummy variable and its square.
- Estimating a regression on a sub-sample for which a dummy variable is either all zeros or all ones.
- Including more regressors than observations.

## 3 Least Squares Residuals

The residual is defined as the difference between the true value $Y_i$ and the fitted value $\hat{Y}_i$:

$$\hat{e}_i = Y_i - \hat{Y}_i = Y_i - X_i'\hat{\beta}.$$

In matrix form, we can write $Y = X\hat{\beta} + \hat{e}$. Do not confuse with error $e_i$, which is unobservable while the residual $\hat{e}_i$ is an estimator.

Residual has two important algebraic properties. First, $\sum_{i=1}^n X_i\hat{e}_i = 0$ (or in the matrix form, $X'\hat{e} = 0$). To see this,

$$\begin{aligned}
\sum_{i=1}^n X_i\hat{e}_i &= \sum_{i=1}^n X_i(Y_i - X_i'\hat{\beta}) \\
&= \sum_{i=1}^n X_iY_i - \sum_{i=1}^n X_iX_i'\hat{\beta} \\
&= \sum_{i=1}^n X_iY_i - \sum_{i=1}^n X_iX_i'(\sum_{i=1}^n X_iX_i')^{-1}(\sum_{i=1}^n X_iY_i) \\
&= \sum_{i=1}^n X_iY_i - \sum_{i=1}^n X_iY_i \\
&= 0
\end{aligned}$$

Second, if $X_i$ contains a constant, the first property implies that $\frac{1}{n}\sum_{i=1}^n \hat{e}_i = 0$ because the first column in $X_i$ is all 1's.

In other words, residuals have a sample mean of zero, and the sample correlation between the regressors and the residual is zero. These are algebraic results and hold for all linear regression estimates. They are not assumptions.

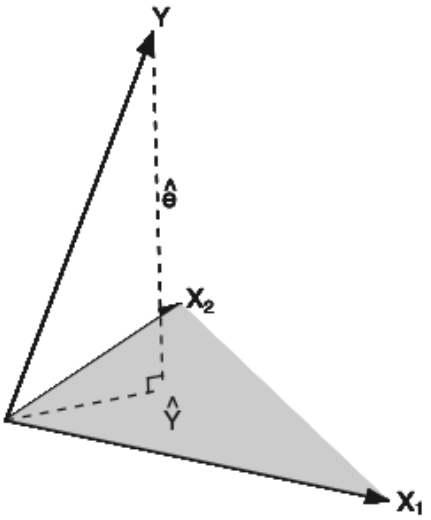



Figure 2: Caption

# 4 Projection

Let us return to $Y = \hat{Y} + \hat{e} = X\hat{\beta} + \hat{e}$. The regressor matrix: $X = [X_1 X_2 ... X_k]$. Linear combinations of columns form a space $R(X)$. If $k = 2$, then it is a plane. $\hat{Y} = X\hat{\beta}$ is in the span of $X$. In other words, we want to find a linear combination of $X$ to get $\hat{Y}$. The distance between $Y$ and $\hat{Y}$ is $\hat{e}$. Recall that we hope that the distance between $Y$ and $\hat{Y}$ is minimized ($\min \sum_{i=1}^{n} (Y_i - X_i'\beta)^2$). It turns out that if $\hat{e}$ is perpendicular to the space of $X$ (The angle between the vectors $\hat{Y}$ and $\hat{e}$ is 90 degrees), the distance is minimized. We call $\hat{Y}$ the projection of $Y$ onto $R(X)$.

Now, we hope to find a linear transformation , or, say, orthogonal projection, $P$ such that $PY = \hat{Y}$. We can find it in $\hat{Y} = X\hat{\beta} = X(X'X)^{-1}(X'Y)$. Therefore, $P = X(X'X)^{-1}X'$. Intuitively, if we project $X$ on its own space, we get it: $PX = X(X'X)^{-1}X'X = X$. In addition, $P' = P$ (symmetric) and $PP = P$ (idempotent).

The diagonal elements if the projection matrix is called leverage values for the regressor matrix. Since

$$P = \begin{bmatrix} X_1' \\ X_2' \\ \vdots \\ X_n' \end{bmatrix} (X'X)^{-1}(X_1\ X_2\ ...\ X_n)$$

they are $h_{ii} = X_i'(X'X)^{-1}X_i$. From the formula, the leverage value $h_{ii}$ is a normalized length of the observed regressor vector $X_i$.

We can also find a matrix $M$ such that $MY = \hat{e}$. We call it annihilator matrix, or residual maker. We can find it by $\hat{e} = Y - X(X'X)^{-1}X'Y = (I - X(X'X)^{-1}X')Y$. Therefore, $M = I - X(X'X)^{-1}X' = I - P$. Intuitively, any matrix in the R(X) will be annihilated; therefore,

$MX = 0$ and $MP = PM = 0$ (show it). Also, similar to $P$, $M' = M$(symmetric) and $MM = M$ (idempotent).

# 5 Frisch-Waugh-Lovell (FWL)

Partition $X = [X_1\ X_2]$ and $\beta = (\beta_1, \beta_2)$. The regression model can be written as

$$Y = X_1\beta_1 + X_2\beta_2 + e \tag{1}$$

We are interested in algebraic expressions for $\hat{\beta}_1$ and $\hat{\beta}_2$.

We first define $M_1 = I - X_1(X_1'X_1)^{-1}X_1'$. It is a residual maker where the explanatory variables are $X_1$. Therefore, $M_1X_2$ is a residual matrix in the regression of $X_2$ on $X_1$. Similarly, we define $M_2 = I - X_2(X_2'X_2)^{-1}X_2'$.

**Theorem 3.** *(Frisch–Waugh (1933)–Lovell (1963)) The least squares estimator $\beta = (\beta_1, \beta_2)$ for (1) has the algebraic solution*

$$\hat{\beta}_1 = (X_1'M_2X_1)^{-1}(X_1'M_2Y)$$
$$\hat{\beta}_2 = (X_2'M_1X_2)^{-1}(X_2'M_1Y)$$

Before we prove it, let us see what it means. For example, we can rewrite

$$\begin{aligned}\hat{\beta}_2 &= (X_2'M_1X_2)^{-1}(X_2'M_1Y)\\ &= (X_2'M_1M_1X_2)^{-1}(X_2'M_1M_1Y)\\ &= (\tilde{X}_2'\tilde{X}_2)^{-1}(\tilde{X}_2'\tilde{e}_1)\end{aligned}$$

where $\tilde{X}_2 = M_1X_2$ and $\tilde{e}_1 = M_1Y$. This means that the coefficient estimator $\hat{\beta}_2$ is algebraically equal to the least squares regression of $\tilde{e}_1$ on $\tilde{X}_2$. In other words,

1. Regress $Y$ on $X_1$, obtain residuals $\tilde{e}_1$ (i.e. $M_1Y$)
2. Regress $X_2$ on $X_1$, obtain residuals $\tilde{X}_2$ (i.e. $M_1X_2$)
3. Regress $\tilde{e}_1$ on $\tilde{X}_2$, obtain OLS $\hat{\beta}_2$ (i.e. regress $M_1Y$ on $M_1X_2$)

Steps 1 and 2 are commonly called partialing out or netting out the effect of $X_1$. For this reason, the coefficients in multiple regression are often called partial regression coefficients. This result can be helpful when interpreting regression coefficients.

Consider regress income on age and education. The coefficient before education captures the pure effect of education by partialing out the effect of age. Why? Because we can regress income and education on age and then to compute the residuals from this regression. By construction, age will not have any power in explaining variation in these residuals. Therefore, any correlation between income and education after this "purging" is independent of (or after removing the effect of) age. This is the flavor of *ceteris paribus*.

*Proof.* We first write the normal equation:

$$\begin{bmatrix} X_1'X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{bmatrix} \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} X_1'Y \\ X_2'Y \end{bmatrix}$$

Look at the first row. It is

$$X_1'X_1\hat{\beta}_1 + X_1'X_2\hat{\beta}_2 = X_1'Y$$

Then $X_1'X_1\hat{\beta}_1 = X_1'Y - X_1'X_2\hat{\beta}_2$. Premultiplying by $(X_1'X_1)^{-1}$, we obtain

$$\hat{\beta}_1 = (X_1'X_1)^{-1}X_1'Y - (X_1'X_1)^{-1}X_1'X_2\hat{\beta}_2 = (X_1'X_1)^{-1}X_1'(Y - X_2\hat{\beta}_2) \tag{2}$$

Then, look at the second row, which is

$$X_2'X_1\hat{\beta}_1 + X_2'X_2\hat{\beta}_2 = X_2'Y$$

We can put $\hat{\beta}_1 = (X_1'X_1)^{-1}X_1'(Y - X_2\hat{\beta}_2)$ into the equation:

$$X_2'X_1(X_1'X_1)^{-1}X_1'Y - X_2'X_1(X_1'X_1)^{-1}X_1'X_2\hat{\beta}_2 + X_2'X_2\hat{\beta}_2 = X_2'Y$$

After collecting terms, the solution is

$$\begin{aligned} \hat{\beta}_2 &= [X_2'(I - X_1(X_1'X_1)^{-1}X_1')X_2]^{-1}[X_2'(I - X_1(X_1'X_1)^{-1}X_1')Y] \\ &= (X_2'M_1X_2)^{-1}(X_2'M_1Y) \end{aligned}$$

□

Note that from (2), we observe that if $X_1$ and $X_2$ are orthogonal, $X_1'X_{=}02$, then $\hat{\beta}_1 = (X_1'X_1)^{-1}X_1'Y$. It means that if $X_1$ and $X_2$ are orthogonal, then the coefficient in the multiple linear least squares regression of $Y$ on $X_1$ and $X_2$ can be obtained by separate regressions of $Y$ on $X_1$ alone and $Y$ on $X_2$ alone. I hope you can link this with the OVB we mentioned in the last lecture.

Consider an application. Suppose $X_1$ are constant, $1_n$. Then consider the partition $X = [1_n\ X_2]$. In this case, $M_1 = I_n - 1_n(1_n'1_n)^{-1}1_n' = I_n - 1_n(1/n)1_n'$. Then $\tilde{X}_2 = M_1X_2 = X_2 - \overline{X}_2$ and $M_1Y = Y - \overline{Y}$ are the demeaned variables. The FWL tells us that

$$\hat{\beta}_2 = (\sum_{i=1}^{n}(X_{2i} - \overline{X}_2)(X_{2i} - \overline{X}_2))^{-1}(\sum_{i=1}^{n}(X_{2i} - \overline{X}_2)(Y_i - \overline{Y}))$$

You have seen this with one regressor in your homework. The numerator is the covariance of $X_2$ and $Y$, and the denominator is the variance of $X_2$.

## 6 Analysis of Variance

The variation of $Y$ contains information. We hope that the variation from $X$ can remove it. Formally, we ask what the fraction of the sample variance of $Y$ is explained by the least squares fit. This fraction is called the coefficient of determination, or R-squared. This is a crude measure of the regression fit.

Therefore, we need the variance of $Y$, the variance of $X'\hat{\beta}$. Recall that the variance formula is $Var(Y) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$.

We start from $Y = \hat{Y} + \hat{e}$. We first demean.

$$Y - I_n\overline{Y} = \hat{Y} - I_n\overline{Y} + \hat{e}$$

For the squared term, in the matrix form, we just premultiply the self transpose,

$$(Y - I_n\overline{Y})'(Y - I_n\overline{Y}) = (\hat{Y} - I_n\overline{Y})'(\hat{Y} - I_n\overline{Y}) + 2(\hat{Y} - I_n\overline{Y})'\hat{e} + \hat{e}'\hat{e}$$

For the middle part, we show that $(\hat{Y} - I_n\overline{Y})'\hat{e} = 0$ assuming that $X$ contains a constant.

$$\begin{aligned}(\hat{Y} - I_n\overline{Y})'\hat{e} &= \hat{Y}'\hat{e} - \overline{Y}I_n'\hat{e} \\ &= (PY)'(MY) - \overline{Y}(I_n'\hat{e}) \\ &= Y'(PM)Y - 0 \\ &= 0\end{aligned}$$

Therefore, we obtain

$$(Y - I_n\overline{Y})'(Y - I_n\overline{Y}) = (\hat{Y} - I_n\overline{Y})'(\hat{Y} - I_n\overline{Y}) + \hat{e}'\hat{e}.$$

We call the LHS, $(Y - I_n\overline{Y})'(Y - I_n\overline{Y}) = \sum_{i=1}^{n}(Y_i - \overline{Y})^2$ the SST, total sum of squares. The first term of the RHS, $(\hat{Y} - I_n\overline{Y})'(\hat{Y} - I_n\overline{Y}) = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2$ the SSR, regression sum of squares; the last term, $\hat{e}'\hat{e} = \sum_{i=1}^{n}\hat{e}_i^2$ the SSE, error sum of squares. Note that $\frac{1}{n}\sum_{i=1}^{n}\hat{e}_i^2$ is the variance of the residuals.

With those terms, we write

$$SST = SSR + SSE$$

This is commonly called the analysis-of-variance formula for least squares regression. People also use

$$TSS = ESS + RSS$$

TSS: total sum of squares

ESS: explained sum of squares

RSS: residual sum of squares

The $R^2$ is defined as

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}.$$

We observe that is is a number between 0 and 1, and it measures the proportion of the total variation in $Y$ that is accounted for by variation in the regressors. Increases when regressors are added to a regression, so "fit" can always increase by increasing the number of regressors. Therefore, people also propose adjusted $R^2$. Note that it says nothing about causality.

## References

Hansen, B. (2022). *Econometrics*. Princeton University Press.