

**LECTURE 3**  
**GEOMETRY OF LS, PROPERTIES OF  $\hat{\sigma}^2$ , PARTITIONED REGRESSION, GOODNESS OF FIT**

## Geometry of LS

We can think of  $y$  and the columns of  $X$  as members of the  $n$ -dimensional Euclidean space  $\mathbb{R}^n$ . One can define a subspace of  $\mathbb{R}^n$  called the *column space* of an  $n \times k$  matrix  $X$ , that is, the collection of all vectors in  $\mathbb{R}^n$  that can be written as linear combinations of the columns of  $X$ :

$$\mathcal{S}(X) = \{z \in \mathbb{R}^n : z = Xb, b \in \mathbb{R}^k\}.$$

For two vectors  $a, b$  in  $\mathbb{R}^n$ , the distance between  $a$  and  $b$  is given by the Euclidean norm<sup>1</sup> of their difference  $\|a - b\| = \sqrt{(a - b)^\top (a - b)}$ . Thus, the LS problem, the minimization of the sum of squared residuals  $(y - Xb)^\top (y - Xb)$ , is to find, out of all elements of  $\mathcal{S}(X)$ , the one closest to  $y$ :

$$\min_{\tilde{y} \in \mathcal{S}(X)} \|y - \tilde{y}\|^2.$$

The closest point is found by dropping the perpendicular from  $y$  onto  $\mathcal{S}(X)$ . That is, a solution to the LS problem,  $\hat{y} = X\hat{\beta}$ , must be chosen so that the residual vector  $\hat{u} = y - \hat{y}$  is orthogonal (perpendicular) to each column of  $X$ :

$$\hat{u}^\top X = 0.$$

Consequently,  $\hat{u}$  is orthogonal to every element of  $\mathcal{S}(X)$ . Indeed, if  $z \in \mathcal{S}(X)$ , then there exists  $b \in \mathbb{R}^k$  such that  $z = Xb$ , and

$$\begin{aligned} \hat{u}^\top z &= \hat{u}^\top Xb \\ &= 0. \end{aligned}$$

The collection of the elements of  $\mathbb{R}^n$  orthogonal to  $\mathcal{S}(X)$  is called the *orthogonal complement* of  $\mathcal{S}(X)$ :

$$\mathcal{S}^\perp(X) = \{z \in \mathbb{R}^n : z^\top X = 0\}.$$

Every element of  $\mathcal{S}^\perp(X)$  is orthogonal to every element in  $\mathcal{S}(X)$ .

As shown in Lecture 2, the solution to the LS problem is

$$\begin{aligned} \hat{y} &= X\hat{\beta} \\ &= X(X^\top X)^{-1} X^\top y \\ &= P_X y, \end{aligned}$$

where

$$P_X = X(X^\top X)^{-1} X^\top$$

is called the *orthogonal projection matrix*. For any vector  $y \in \mathbb{R}^n$ ,

$$P_X y \in \mathcal{S}(X).$$

Furthermore, the residual vector lies in  $\mathcal{S}^\perp(X)$ :

$$y - P_X y \in \mathcal{S}^\perp(X). \tag{1}$$

---

<sup>1</sup>For a vector  $x = (x_1, x_2, \dots, x_n)^\top$ , its Euclidean norm is defined as  $\|x\| = \sqrt{x^\top x} = \sqrt{\sum_{i=1}^n x_i^2}$ .

To show (1), observe first that since the columns of  $X$  are in  $\mathcal{S}(X)$ ,

$$\begin{aligned} P_X X &= X (X^\top X)^{-1} X^\top X \\ &= X, \end{aligned}$$

and, since  $(X^\top X)^{-1}$  is symmetric,  $P_X^\top = X (X^\top X)^{-1} X^\top = P_X$ , so

$$X^\top P_X = X^\top.$$

Now,

$$\begin{aligned} X^\top (y - P_X y) &= X^\top y - X^\top P_X y \\ &= X^\top y - X^\top y \\ &= 0. \end{aligned}$$

Thus, by definition, the residuals  $y - P_X y \in \mathcal{S}^\perp(X)$ . The residuals can be written as

$$\begin{aligned} \hat{u} &= y - P_X y \\ &= (I_n - P_X) y \\ &= M_X y, \end{aligned}$$

where

$$\begin{aligned} M_X &= I_n - P_X \\ &= I_n - X (X^\top X)^{-1} X^\top \end{aligned}$$

is a projection matrix onto  $\mathcal{S}^\perp(X)$ .

The projection matrices  $P_X$  and  $M_X$  have the following properties:

- $P_X + M_X = I_n$ . This implies that for any  $y \in \mathbb{R}^n$ ,

$$y = P_X y + M_X y.$$

- Pythagoras: by orthogonality of  $P_X y$  and  $M_X y$  (see the Orthogonal property below),

$$\|y\|^2 = \|P_X y\|^2 + \|M_X y\|^2 = \|\hat{y}\|^2 + \|\hat{u}\|^2.$$

- Symmetric:

$$\begin{aligned} P_X^\top &= P_X, \\ M_X^\top &= M_X. \end{aligned}$$

- Idempotent:  $P_X^2 = P_X$  and  $M_X^2 = M_X$ :

$$\begin{aligned} P_X P_X &= X (X^\top X)^{-1} X^\top X (X^\top X)^{-1} X^\top \\ &= X (X^\top X)^{-1} X^\top \\ &= P_X \end{aligned}$$

$$\begin{aligned} M_X M_X &= (I_n - P_X) (I_n - P_X) \\ &= I_n - 2P_X + P_X P_X \\ &= I_n - 2P_X + P_X \text{ (since } P_X \text{ is idempotent)} \\ &= I_n - P_X \\ &= M_X. \end{aligned}$$

- Orthogonal:

$$\begin{aligned}
 P_X M_X &= P_X (I_n - P_X) \\
 &= P_X - P_X P_X \\
 &= P_X - P_X \\
 &= 0.
 \end{aligned}$$

This property implies  $M_X X = 0$ :

$$\begin{aligned}
 M_X X &= (I_n - P_X) X \\
 &= X - P_X X \\
 &= X - X \\
 &= 0.
 \end{aligned}$$

The preceding discussion does not use any regression assumptions. Given data  $y$  and  $X$ , one can always perform least squares estimation, regardless of the data-generating process. However, a model is needed to study the statistical properties of an estimator, such as unbiasedness.

## Properties of $\hat{\sigma}^2$

The following estimator of  $\sigma^2$  was introduced in Lecture 2:

$$\begin{aligned}
 \hat{\sigma}^2 &= n^{-1} \sum_{i=1}^n (Y_i - X_i^\top \hat{\beta})^2 \\
 &= n^{-1} \hat{U}^\top \hat{U}.
 \end{aligned}$$

Under Assumptions (A1)–(A4),  $\hat{\sigma}^2$  is biased. To show this, first write

$$\begin{aligned}
 \hat{U} &= M_X Y \\
 &= M_X (X\beta + U) \\
 &= M_X U.
 \end{aligned}$$

The last equality holds because  $M_X X = 0$ . Next,

$$\begin{aligned}
 n\hat{\sigma}^2 &= \hat{U}^\top \hat{U} \\
 &= U^\top M_X^\top M_X U \\
 &= U^\top M_X U.
 \end{aligned}$$

Now, since  $U^\top M_X U$  is a scalar,

$$U^\top M_X U = \text{tr}(U^\top M_X U),$$

where  $\text{tr}(A)$  denotes the trace of the square matrix  $A$ .

$$\begin{aligned}
 \text{E}[U^\top M_X U \mid X] &= \text{E}[\text{tr}(U^\top M_X U) \mid X] \\
 &= \text{E}[\text{tr}(M_X U U^\top) \mid X] \quad (\text{since } \text{tr}(ABC) = \text{tr}(BCA)) \\
 &= \text{tr}(M_X \text{E}[U U^\top \mid X]) \quad (\text{since the trace and expectation are linear}) \\
 &= \sigma^2 \text{tr}(M_X),
 \end{aligned}$$

where the last equality uses Assumption (A3):  $E[UU^\top | X] = \sigma^2 I_n$ , so that  $\text{tr}(M_X \cdot \sigma^2 I_n) = \sigma^2 \text{tr}(M_X)$ . It remains to compute  $\text{tr}(M_X)$ . We have

$$\begin{aligned} \text{tr}(M_X) &= \text{tr}(I_n - X(X^\top X)^{-1}X^\top) \\ &= \text{tr}(I_n) - \text{tr}(X(X^\top X)^{-1}X^\top) \\ &= \text{tr}(I_n) - \text{tr}((X^\top X)^{-1}X^\top X) \quad (\text{by trace cyclicity}) \\ &= \text{tr}(I_n) - \text{tr}(I_k) \\ &= n - k. \end{aligned}$$

Thus,

$$E[\hat{\sigma}^2 | X] = \frac{n-k}{n}\sigma^2, \tag{2}$$

and by iterated expectations,

$$E[\hat{\sigma}^2] = \frac{n-k}{n}\sigma^2.$$

The estimator  $\hat{\sigma}^2$  is biased, but dividing the sum of squared residuals by  $n-k$  instead of  $n$  yields an unbiased estimator of  $\sigma^2$ . Define

$$\begin{aligned} s^2 &= \frac{n}{n-k}\hat{\sigma}^2 \\ &= (n-k)^{-1} \sum_{i=1}^n (Y_i - X_i^\top \hat{\beta})^2. \end{aligned}$$

It follows from (2) and iterated expectations that

$$E[s^2] = \sigma^2.$$

## Partitioned regression

We can partition the matrix of regressors  $X$  as follows:

$$X = \begin{pmatrix} X_1 & X_2 \end{pmatrix},$$

and write the model as

$$Y = X_1\beta_1 + X_2\beta_2 + U,$$

where  $X_1$  is an  $n \times k_1$  matrix,  $X_2$  is an  $n \times k_2$  matrix,  $k_1 + k_2 = k$ , and

$$\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix},$$

where  $\beta_1$  and  $\beta_2$  are  $k_1$ - and  $k_2$ -vectors, respectively. Such a decomposition allows one to focus on a group of variables and their corresponding parameters, say  $X_1$  and  $\beta_1$ . If

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix},$$

then one can write the following version of the normal equations:

$$(X^\top X)\hat{\beta} = X^\top Y$$

as

$$\begin{pmatrix} X_1^\top X_1 & X_1^\top X_2 \\ X_2^\top X_1 & X_2^\top X_2 \end{pmatrix} \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \begin{pmatrix} X_1^\top Y \\ X_2^\top Y \end{pmatrix}.$$

The expressions for  $\hat{\beta}_1$  and  $\hat{\beta}_2$  can be obtained by inverting the partitioned matrix on the left-hand side.

Alternatively, since  $X$  has full column rank by Assumption (A4), so does  $X_2$ ; hence  $X_2^\top X_2$  is invertible. Define  $M_2$  to be the projection matrix onto the orthogonal complement of  $\mathcal{S}(X_2)$  (the residual maker for  $X_2$ ):

$$M_2 = I_n - X_2 (X_2^\top X_2)^{-1} X_2^\top.$$

Like  $M_X$ , the matrix  $M_2$  is symmetric and idempotent. Since  $X = (X_1, X_2)$  has full column rank, the columns of  $M_2 X_1$  are linearly independent, so  $X_1^\top M_2 X_1 = (M_2 X_1)^\top (M_2 X_1)$  is positive definite (hence invertible). Then

$$\hat{\beta}_1 = (X_1^\top M_2 X_1)^{-1} X_1^\top M_2 Y. \quad (3)$$

This identity is known as the *Frisch–Waugh–Lovell (FWL) theorem*. To show this, first write

$$Y = X_1 \hat{\beta}_1 + X_2 \hat{\beta}_2 + \hat{U}. \quad (4)$$

By construction,

$$\begin{aligned} M_2 \hat{U} &= \hat{U} \quad (\hat{U} \text{ is orthogonal to } X_2), \\ M_2 X_2 &= 0, \\ X_1^\top \hat{U} &= 0, \\ X_2^\top \hat{U} &= 0. \end{aligned}$$

Substitute equation (4) into the right-hand side of equation (3):

$$\begin{aligned} & (X_1^\top M_2 X_1)^{-1} X_1^\top M_2 Y \\ &= (X_1^\top M_2 X_1)^{-1} X_1^\top M_2 (X_1 \hat{\beta}_1 + X_2 \hat{\beta}_2 + \hat{U}) \\ &= (X_1^\top M_2 X_1)^{-1} X_1^\top M_2 X_1 \hat{\beta}_1 \\ &+ (X_1^\top M_2 X_1)^{-1} X_1^\top \hat{U} \quad (\text{since } M_2 X_2 = 0 \text{ and } M_2 \hat{U} = \hat{U}) \\ &= \hat{\beta}_1 \quad (\text{since } X_1^\top \hat{U} = 0). \end{aligned}$$

Since  $M_2$  is symmetric and idempotent, one can write

$$\begin{aligned} \hat{\beta}_1 &= ((M_2 X_1)^\top (M_2 X_1))^{-1} (M_2 X_1)^\top M_2 Y \\ &= (\tilde{X}_1^\top \tilde{X}_1)^{-1} \tilde{X}_1^\top \tilde{Y}, \end{aligned}$$

where

$$\begin{aligned} \tilde{X}_1 &= M_2 X_1 \\ &= X_1 - X_2 (X_2^\top X_2)^{-1} X_2^\top X_1 \quad (\text{residuals from the regression of } X_1 \text{ on } X_2), \\ \tilde{Y} &= M_2 Y \\ &= Y - X_2 (X_2^\top X_2)^{-1} X_2^\top Y \quad (\text{residuals from the regression of } Y \text{ on } X_2). \end{aligned}$$

Thus, to obtain the coefficients on the first  $k_1$  regressors, instead of running the full regression with  $k_1 + k_2$  regressors, one can regress  $Y$  on  $X_2$  to obtain the residuals  $\tilde{Y}$ , regress  $X_1$  on  $X_2$  to obtain the residuals  $\tilde{X}_1$ , and then regress  $\tilde{Y}$  on  $\tilde{X}_1$  to obtain  $\hat{\beta}_1$ . In other words,  $\hat{\beta}_1$  shows the effect of  $X_1$  *after controlling* for  $X_2$ .

Analogously, for  $\hat{\beta}_2$ :

$$\begin{aligned} \hat{\beta}_2 &= (X_2^\top M_1 X_2)^{-1} X_2^\top M_1 Y, \quad \text{where} \\ M_1 &= I_n - X_1 (X_1^\top X_1)^{-1} X_1^\top. \end{aligned}$$

For example, consider a simple regression

$$Y_i = \beta_1 + \beta_2 X_i + U_i,$$

for  $i = 1, \dots, n$ .

Define an  $n$ -vector of ones:

$$\ell = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}.$$

In this case, the matrix of regressors is given by

$$\begin{pmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{pmatrix} = (\ell \quad X).$$

Consider

$$M_1 = I_n - \ell (\ell^\top \ell)^{-1} \ell^\top,$$

and

$$\widehat{\beta}_2 = \frac{X^\top M_1 Y}{X^\top M_1 X}.$$

Now,  $\ell^\top \ell = n$ . Therefore,

$$\begin{aligned} M_1 &= I_n - \frac{1}{n} \ell \ell^\top, \text{ and} \\ M_1 X &= X - \ell \frac{\ell^\top X}{n} \\ &= X - \bar{X} \ell \\ &= \begin{pmatrix} X_1 - \bar{X} \\ X_2 - \bar{X} \\ \vdots \\ X_n - \bar{X} \end{pmatrix}, \end{aligned}$$

where

$$\begin{aligned} \bar{X} &= \frac{\ell^\top X}{n} \\ &= n^{-1} \sum_{i=1}^n X_i. \end{aligned}$$

Thus, the matrix  $M_1$  transforms the vector  $X$  into the vector of deviations from the mean. We can write

$$\begin{aligned} \widehat{\beta}_2 &= \frac{\sum_{i=1}^n (X_i - \bar{X}) Y_i}{\sum_{i=1}^n (X_i - \bar{X})^2} \\ &= \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}. \end{aligned}$$

## Goodness of fit

Write

$$\begin{aligned} Y &= P_X Y + M_X Y \\ &= \hat{Y} + \hat{U}, \end{aligned}$$

where, by construction,

$$\begin{aligned} \hat{Y}^\top \hat{U} &= (P_X Y)^\top M_X Y \\ &= Y^\top P_X M_X Y \\ &= 0. \end{aligned}$$

Suppose the model contains an intercept, that is, the first column of  $X$  is the vector of ones  $\ell$ . The *total variation* in  $Y$  is

$$\begin{aligned} \sum_{i=1}^n (Y_i - \bar{Y})^2 &= Y^\top M_1 Y \\ &= (\hat{Y} + \hat{U})^\top M_1 (\hat{Y} + \hat{U}) \\ &= \hat{Y}^\top M_1 \hat{Y} + \hat{U}^\top M_1 \hat{U} + 2\hat{Y}^\top M_1 \hat{U}. \end{aligned}$$

Since the model contains an intercept,

$$\begin{aligned} \ell^\top \hat{U} &= 0, \text{ and} \\ M_1 \hat{U} &= \hat{U}. \end{aligned}$$

Using  $M_1 \hat{U} = \hat{U}$ , the second term simplifies to  $\hat{U}^\top M_1 \hat{U} = \hat{U}^\top \hat{U}$ , and the cross term vanishes because  $\hat{Y}^\top M_1 \hat{U} = \hat{Y}^\top \hat{U} = 0$ . Therefore,

$$\begin{aligned} Y^\top M_1 Y &= \hat{Y}^\top M_1 \hat{Y} + \hat{U}^\top \hat{U}, \text{ or} \\ \sum_{i=1}^n (Y_i - \bar{Y})^2 &= \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^n \hat{U}_i^2. \end{aligned}$$

Moreover,

$$\begin{aligned} \bar{Y} &= \frac{\ell^\top Y}{n} \\ &= \frac{\ell^\top \hat{Y}}{n} + \frac{\ell^\top \hat{U}}{n} \\ &= \frac{\ell^\top \hat{Y}}{n} \\ &= \overline{\hat{Y}}. \end{aligned}$$

Hence, the sample means of  $Y$  and  $\hat{Y}$  coincide, and

$$\begin{aligned} \sum_{i=1}^n (Y_i - \bar{Y})^2 &= \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^n \hat{U}_i^2, \text{ or} \\ TSS &= ESS + RSS, \end{aligned} \tag{5}$$

where

$$\begin{aligned}
 TSS &= \sum_{i=1}^n (Y_i - \bar{Y})^2 \quad (\text{total sum of squares}), \\
 ESS &= \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 \quad (\text{explained sum of squares}), \\
 RSS &= \sum_{i=1}^n \hat{U}_i^2 \quad (\text{residual sum of squares}).
 \end{aligned}$$

The ratio of  $ESS$  to  $TSS$  is called the *coefficient of determination*, or  $R^2$ :

$$\begin{aligned}
 R^2 &= \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} \\
 &= 1 - \frac{\sum_{i=1}^n \hat{U}_i^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} \\
 &= 1 - \frac{\hat{U}^\top \hat{U}}{Y^\top M_1 Y}.
 \end{aligned}$$

Properties of  $R^2$ :

- $R^2$  lies in  $[0, 1]$ , as implied by decomposition (5). This property need not hold without an intercept, and the above definition of  $R^2$  should not be used in that case.  $R^2 = 1$  if and only if  $Y \in \mathcal{S}(X)$ , that is,  $Y$  is *exactly* a linear combination of the columns of  $X$ . (Indeed,  $R^2 = 1$  implies  $\hat{U}^\top \hat{U} = 0$ , hence  $\hat{U} = 0$  and  $Y = \hat{Y} \in \mathcal{S}(X)$ ; conversely, if  $Y \in \mathcal{S}(X)$ , then  $\hat{U} = 0$  and  $R^2 = 1$ .)
- $R^2$  does not decrease when additional regressors are added.

Proof. Consider a partitioned matrix  $X = \begin{pmatrix} Z & W \end{pmatrix}$ . We study the effect of adding  $W$  on  $R^2$ . Let

$$\begin{aligned}
 P_X &= X (X^\top X)^{-1} X^\top \quad (\text{projection matrix for the full regression}), \\
 P_Z &= Z (Z^\top Z)^{-1} Z^\top \quad (\text{projection matrix for the regression without } W).
 \end{aligned}$$

Define also

$$\begin{aligned}
 M_X &= I_n - P_X, \\
 M_Z &= I_n - P_Z.
 \end{aligned}$$

Since  $Z$  is part of  $X$ ,

$$P_X Z = Z,$$

and

$$\begin{aligned}
 P_X P_Z &= P_X Z (Z^\top Z)^{-1} Z^\top \\
 &= Z (Z^\top Z)^{-1} Z^\top \\
 &= P_Z.
 \end{aligned}$$

Consequently,

$$\begin{aligned}
 M_X M_Z &= (I_n - P_X) (I_n - P_Z) \\
 &= I_n - P_X - P_Z + P_X P_Z \\
 &= I_n - P_X - P_Z + P_Z \\
 &= M_X.
 \end{aligned}$$

Assume that  $Z$  contains a column of ones, so both short and long regressions have intercepts. Define

$$\begin{aligned}\widehat{U}_X &= M_X Y, \\ \widehat{U}_Z &= M_Z Y.\end{aligned}$$

Write:

$$\begin{aligned}0 &\leq (\widehat{U}_X - \widehat{U}_Z)^\top (\widehat{U}_X - \widehat{U}_Z) \\ &= \widehat{U}_X^\top \widehat{U}_X + \widehat{U}_Z^\top \widehat{U}_Z - 2\widehat{U}_X^\top \widehat{U}_Z.\end{aligned}$$

Next,

$$\begin{aligned}\widehat{U}_X^\top \widehat{U}_Z &= Y^\top M_X M_Z Y \\ &= Y^\top M_X Y \\ &= \widehat{U}_X^\top \widehat{U}_X.\end{aligned}$$

Hence,

$$\widehat{U}_Z^\top \widehat{U}_Z \geq \widehat{U}_X^\top \widehat{U}_X.$$

Since both regressions have the same  $TSS$  (which depends only on  $Y$ ), this implies  $R_X^2 \geq R_Z^2$ , that is,  $R^2$  does not decrease when regressors are added.

- $R^2$  measures the fraction of *sample* variation in  $Y$  explained by  $X$ . However, our objective is to estimate *population* relationships, not to explain *sample* variation. A high  $R^2$  does not necessarily indicate a well-specified model, and a low  $R^2$  is not evidence of misspecification.
- One can construct an  $X$  that makes  $R^2 = 1$ : take any  $n$  linearly independent vectors. Since such a set spans  $\mathbb{R}^n$ , any  $Y \in \mathbb{R}^n$  can be written as an exact linear combination of the columns of  $X$ .

Since  $R^2$  does not decrease with the inclusion of additional regressors, researchers often report instead the *adjusted coefficient of determination*  $\overline{R}^2$ :

$$\begin{aligned}\overline{R}^2 &= 1 - \frac{n-1}{n-k}(1-R^2) \\ &= 1 - \frac{\widehat{U}^\top \widehat{U}/(n-k)}{Y^\top M_1 Y/(n-1)}.\end{aligned}$$

The adjusted coefficient of determination penalizes the fit when the number of regressors  $k$  is large relative to the number of observations  $n$ , and  $\overline{R}^2$  may decrease as  $k$  increases. However, there is no strong theoretical argument for using this adjustment.