LECTURE 3
GEOMETRY OF LS, PROPERTIES OF $\widehat{\sigma}^2$, 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 R^n. One can define a subspace of R^n called the *column space* of a $n \times k$ matrix X, that is a collection of all vectors in R^n that can be written as linear combinations of the columns of X:

$$\mathcal{S}(X) = \left\{ z \in R^n : z = Xb, b = (b_1, b_2, \dots, b_k)' \in R^k \right\}.$$

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

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

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

$$\widehat{u}'X = 0.$$

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

$$\begin{aligned} \widehat{u}'z &= \widehat{u}'Xb \\ &= 0. \end{aligned}$$

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

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

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

As we have seen in Lecture 2, the solution to the LS problem is given by

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

where

$$P_X = X(X'X)^{-1}X'$$

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

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

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

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

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

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

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

and, since P_X is a symmetric matrix,

$$X'P_X = X'.$$

Now,

$$\begin{aligned} X'\left(y - P_X y\right) &= X'y - X'P_X y \\ &= X'y - X'y \\ &= 0. \end{aligned}$$

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

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

where

$$\begin{aligned} M_X &= I_n - P_X \\ &= I_n - X\left(X'X\right)^{-1} X', \end{aligned}$$

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

The projection matrices P_X and M_X have the following properties:

- $P_X + M_X = I$. This implies, that for any $y \in R^n$,

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

- Symmetric:

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

- Idempotent: $P_X P_X = P_X$, and $M_X M_X = M_X$.

$$\begin{aligned} P_X P_X &= X\left(X'X\right)^{-1} X'X\left(X'X\right)^{-1} X' \\ &= X\left(X'X\right)^{-1} X' \\ &= P_X \\ M_X M_X &= \left(I_n - P_X\right)\left(I_n - P_X\right) \\ &= I_n - 2P_X + P_X P_X \\ &= I_n - P_X \\ &= M_X. \end{aligned}$$

- Orthogonal:

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

This property implies that $M_X X = 0$. Indeed,

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

Note that, in the above discussion, none of the regression assumptions have been used. Given data, y and X, one can always perform least squares, regardless of what data generating process stands behind the data. However, one needs a model to discuss the statistical properties of an estimator (such as unbiasedness and etc).

Properties of $\widehat{\sigma}^2$

The following estimator for σ^2 was suggested in Lecture 2:

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

It turns out that, under the usual regression assumptions (A1)-(A4), $\widehat{\sigma}^2$ is a biased estimator. First, write

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

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

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

Now, since $U' M_X U$ is a scalar,

$$U' M_X U = tr\left(U' M_X U \right),$$

where $tr(A)$ denotes the trace of a matrix A.

$$\begin{aligned} E\left(U' M_X U | X \right) &= E\left(tr\left(U' M_X U \right) | X \right) \\ &= E\left(tr\left(M_X U U' \right) | X \right) \ \text{(because } tr(ABC) = tr(BCA)\text{)} \\ &= tr\left(M_X E\left(UU' | X \right) \right) \ \text{(because } tr \text{ and expectation are linear operators)} \\ &= \sigma^2 tr\left(M_X \right). \end{aligned}$$

The last equality follows, because by Assumption (A3), $E\left(UU' | X \right) = \sigma^2 I_n$. Next,

$$\begin{aligned} tr\left(M_X \right) &= tr\left(I_n - X \left(X'X \right)^{-1} X' \right) \\ &= tr\left(I_n \right) - tr\left(X \left(X'X \right)^{-1} X' \right) \\ &= tr\left(I_n \right) - tr\left(\left(X'X \right)^{-1} X'X \right) \\ &= tr\left(I_n \right) - tr\left(I_k \right) \\ &= n - k. \end{aligned}$$

Thus,

$$E\widehat{\sigma}^2 = \frac{n-k}{n}\sigma^2. \tag{2}$$

The estimator $\widehat{\sigma}^2$ is biased, but it is easy to modify $\widehat{\sigma}^2$ to obtain unbiasedness. Define

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

It follows from (2) that

$$Es^2 = \sigma^2.$$

Partitioned regression

We can partition the matrix of regressors X as follows:

$$X = (X_1 \quad X_2),$$

and write the model as

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

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

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

where β_1 and β_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 β_1. If

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

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

$$(X'X)\,\widehat{\beta} = X'Y$$

as

$$\begin{pmatrix} X_1'X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{pmatrix} \begin{pmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_2 \end{pmatrix} = \begin{pmatrix} X_1'Y \\ X_2'Y \end{pmatrix}.$$

One can obtain the expressions for $\widehat{\beta}_1$ and $\widehat{\beta}_2$ by inverting the partitioned matrix on the left-hand side of the equation above.

Alternatively, let's define M_2 to be the projection matrix on the space orthogonal to the space $\mathcal{S}(X_2)$:

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

Then,

$$\widehat{\beta}_1 = (X_1'M_2X_1)^{-1} X_1'M_2Y. \tag{3}$$

In order to show that, first write

$$Y = X_1\widehat{\beta}_1 + X_2\widehat{\beta}_2 + \widehat{U}. \tag{4}$$

Note that by the construction,

$$\begin{aligned} M_2\widehat{U} &= \widehat{U} \ (\widehat{U} \text{ is orthogonal to } X_2), \\ M_2X_2 &= 0, \\ X_1'\widehat{U} &= 0, \\ X_2'\widehat{U} &= 0. \end{aligned}$$

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

$$\begin{aligned} & \left(X_1'M_2X_1\right)^{-1} X_1'M_2Y \\ &= \left(X_1'M_2X_1\right)^{-1} X_1'M_2\left(X_1\widehat{\beta}_1 + X_2\widehat{\beta}_2 + \widehat{U}\right) \\ &= \left(X_1'M_2X_1\right)^{-1} X_1'M_2X_1\widehat{\beta}_1 \\ &+ \left(X_1'M_2X_1\right)^{-1} X_1'\widehat{U} \qquad \left(M_2X_2 = 0 \text{ and } M_2\widehat{U} = \widehat{U}\right) \\ &= \widehat{\beta}_1. \end{aligned}$$

Since M_2 is symmetric and idempotent, one can write

$$\begin{aligned} \widehat{\beta}_1 &= \left(\left(M_2X_1\right)'\left(M_2X_1\right)\right)^{-1}\left(M_2X_1\right)'\left(M_2Y\right) \\ &= \left(\widetilde{X}_1'\widetilde{X}_1\right)^{-1}\widetilde{X}_1'\widetilde{Y}, \end{aligned}$$

where

$$\begin{aligned} \widetilde{X}_1 &= M_2X_1 \\ &= X_1 - X_2\left(X_2'X_2\right)^{-1}X_2'X_1 \text{ residuals from the regression of } X_1 \text{ on } X_2, \\ \widetilde{Y} &= M_2Y \\ &= Y - X_2\left(X_2'X_2\right)^{-1}X_2'Y \text{ residuals from the regression of } Y \text{ on } X_2. \end{aligned}$$

Thus, to obtain coefficients for 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 $\widetilde{Y}$, regress X_1 on X_2 to obtain the residuals $\widetilde{X}_1$, and then regress $\widetilde{Y}$ on $\widetilde{X}_1$ to obtain $\widehat{\beta}_1$. In other words, $\widehat{\beta}_1$ shows the effect of X_1 *after controlling* for X_2.

Similarly to $\widehat{\beta}_1$, one can write:

$$\begin{aligned} \widehat{\beta}_2 &= \left(X_2'M_1X_2\right)^{-1}X_2'M_1Y, \text{ where} \\ M_1 &= I_n - X_1\left(X_1'X_1\right)^{-1}X_1'. \end{aligned}$$

For example, consider a simple regression

$$Y_i = \beta_1 + \beta_2X_i + U_i,$$

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

Let's define a n-vector of ones:

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

In this case, the matrix of regressors is given by

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

Consider

$$M_1 = I_n - \ell\left(\ell'\ell\right)^{-1}\ell',$$

and

$$\widehat{\beta}_2 = \frac{X'M_1Y}{X'M_1X}.$$

Now, $\ell'\ell = n$. Therefore,

$$\begin{aligned}
M_1 &= I_n - \frac{1}{n}\ell\ell', \text{ and} \\
M_1X &= X - \ell\frac{\ell'X}{n} \\
&= X - \overline{X}\ell \\
&= \begin{pmatrix} X_1 - \overline{X} \\ X_2 - \overline{X} \\ \vdots \\ X_n - \overline{X} \end{pmatrix},
\end{aligned}$$

where

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

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

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

Goodness of fit

Write

$$\begin{aligned}
Y &= P_XY + M_XY \\
&= \widehat{Y} + \widehat{U},
\end{aligned}$$

where, by the contruction,

$$\begin{aligned}
\widehat{Y}'\widehat{U} &= \left(P_XY\right)'\left(M_XY\right) \\
&= Y'P_XM_XY \\
&= 0.
\end{aligned}$$

Suppose that the model contains an intercept, i.e. the first column of X is the vector of ones ℓ. The *total variation* in Y is

$$\begin{aligned}
\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2 &= Y'M_1Y \\
&= \left(\widehat{Y} + \widehat{U}\right)' M_1 \left(\widehat{Y} + \widehat{U}\right) \\
&= \widehat{Y}'M_1\widehat{Y} + \widehat{U}'M_1\widehat{U} + 2\widehat{Y}'M_1\widehat{U}.
\end{aligned}$$

Since the model contains an intercept,

$$\begin{aligned} \ell' \widehat{U} &= 0, \text{ and} \\ M_1 \widehat{U} &= \widehat{U}. \end{aligned}$$

However, $\widehat{Y}'\widehat{U} = 0$, and, therefore,

$$\begin{aligned} Y' M_1 Y &= \widehat{Y}' M_1 \widehat{Y} + \widehat{U}'\widehat{U}, \text{ or} \\ \sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2 &= \sum_{i=1}^{n} \left(\widehat{Y}_i - \overline{\widehat{Y}}\right)^2 + \sum_{i=1}^{n} \widehat{U}_i^2. \end{aligned}$$

Note that

$$\begin{aligned} \overline{Y} &= \frac{\ell' Y}{n} \\ &= \frac{\ell' \widehat{Y}}{n} + \frac{\ell' \widehat{U}}{n} \\ &= \frac{\ell' \widehat{Y}}{n} \\ &= \overline{\widehat{Y}}. \end{aligned}$$

Hence, the averages of Y and its predicted values $\widehat{Y}$ are equal, and we can write:

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

where

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

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

$$\begin{aligned} R^2 &= \frac{\sum_{i=1}^{n} \left(\widehat{Y}_i - \overline{Y}\right)^2}{\sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2} \\ &= 1 - \frac{\sum_{i=1}^{n} \widehat{U}_i^2}{\sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2} \\ &= 1 - \frac{\widehat{U}'\widehat{U}}{Y' M_1 Y}. \end{aligned}$$

Properties of R^2:

- Bounded between 0 and 1 as implied by decomposition (5). This property does not hold if the model does not have an intercept, and one should not use the above definition of R^2 in this case. If $R^2 = 1$ then $\widehat{U}'\widehat{U} = 0$, which can happen only if $Y \in \mathcal{S}(X)$, i.e. Y is *exactly* a linear combination of the columns of X.

- Increases by adding more regressors.

  Proof. Consider a partitioned matrix $X = (Z \;\; W)$. Let's study the effect of adding W on R^2. Let

$$P_X = X\left(X'X\right)^{-1}X' \text{ projection matrix corresponding to the full regression,}$$
$$P_Z = Z\left(Z'Z\right)^{-1}Z' \text{ projection matrix corresponding to the regression without } W.$$

  Define also

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

  Note that since Z is a part of X,

$$P_X Z = Z,$$

  and

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

  Consequently,

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

  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:

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

  Next,

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

  Hence,

$$\widehat{U}_Z'\widehat{U}_Z \geq \widehat{U}_X'\widehat{U}_X.$$

- R^2 shows how much of the *sample* variation in y was explained by X. However, our objective is to estimate *population* relationships and not to explain the *sample* variation. High R^2 is not necessary an indicator of the good regression model, and a low R^2 is not an evidence against it.

- One can always *find* an X that makes $R^2 = 1$, just take any n linearly independent vectors. Because such a set spans the whole R^n space, any $y \in R^n$ can be written as an exact linear combination of the columns of that X.

Since R^2 increases with inclusion of additional regressors, instead researchers often report the *adjusted coefficient of determination* $\overline{R}^2$:

$$\begin{aligned}\overline{R}^2 &= 1 - \frac{n-1}{n-k}\left(1 - R^2\right) \\ &= 1 - \frac{\widehat{U}'\widehat{U}/\left(n-k\right)}{Y'M_1Y/\left(n-1\right)}.\end{aligned}$$

The adjusted coefficient of determination discounts the fit when the number of the regressors k is large relative to the number of observations n. $\overline{R}^2$ may decrease with k. However, there is no strong argument for using such an adjustment.