2. OLS Part II

The OLS residuals are orthogonal to the regressors. If the model includes an intercept, the orthogonality of the residuals and regressors gives rise to three results, which have limited practical usefulness but are good exercises for understanding the algebra of least squares.

Partitioned and partial regression is not treated as seriously as it might be in a graduate course, and is only given a cursory presentation. However, the residual maker matrix M_1 is presented, and is used in to define R^2, and in several other parts of the course.

2.1 Some basic properties of OLS

First, note that the LS residuals are "orthogonal" to the regressors –

$$X'X\boldsymbol{b} - X'\boldsymbol{y} = \mathbf{0} \qquad \text{("normal equations"; } (k\times 1)\text{)}$$

So,

$$-X'(\boldsymbol{y} - X\boldsymbol{b}) = -X'\boldsymbol{e} = 0 \ ;$$

or,

$$X'\boldsymbol{e} = 0$$

Orthogonality implies linear independence (but not *vice versa*). See "Linearly Independent, Orthogonal, and Uncorrelated Variables" (Rodgers et al., 1984), for the definition and relationship between the three terms.

If the model includes an intercept term, then one regressor (say, the first column of X) is a unit vector.

In this case we get some further results:

1. The LS residuals sum to zero

$$X'\boldsymbol{e} = \begin{pmatrix} 1 & \cdots & x_{1k} \\ \vdots & \ddots & \vdots \\ 1 & \cdots & x_{nk} \end{pmatrix}' \begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix} = \begin{pmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ x_{1k} & \cdots & x_{nk} \end{pmatrix} \begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix}$$

$$= \begin{pmatrix} \sum_i e_i \\ ? \\ ? \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}$$

From the first element:

$$\sum_{i=1}^{n} e_i = 0$$

2. Fitted regression passes through sample mean

$$X'\boldsymbol{y} = X'X\boldsymbol{b} \ ,$$

or, $$\begin{pmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ x_{1k} & \cdots & x_{nk} \end{pmatrix}\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ x_{1k} & \cdots & x_{nk} \end{pmatrix}\begin{pmatrix} 1 & \cdots & x_{1k} \\ \vdots & \ddots & \vdots \\ 1 & \cdots & x_{nk} \end{pmatrix}\begin{pmatrix} b_1 \\ \vdots \\ b_k \end{pmatrix} .$$

So, $$\begin{pmatrix} \sum_i y_i \\ ? \\ ? \end{pmatrix} = \begin{pmatrix} n & \sum_i x_{i2} & \ldots \\ ? & \ldots & ? \\ ? & \ldots & ? \end{pmatrix}\begin{pmatrix} b_1 \\ \vdots \\ b_k \end{pmatrix} .$$

From the first row of this vector equation –

$$\sum_i y_i = nb_1 + b_2\sum_i x_{i2} + \cdots + b_k\sum_i x_{ik}$$

or,

$$\bar{y} = b_1 + b_2\overline{x_2} + \cdots + b_k\overline{x_k}$$

3. Sample mean of the fitted *y*-values equals sample mean of actual *y*-values

$$y_i = \boldsymbol{x_i'\beta} + \varepsilon_i = \boldsymbol{x_i'b} + e_i = \hat{y}_i + e_i \ .$$

So,

$$\tfrac{1}{n}\textstyle\sum_{i=1}^{n} y_i = \tfrac{1}{n}\sum_{i=1}^{n}\hat{y}_i + \tfrac{1}{n}\sum_{i=1}^{n} e_i \ ,$$

or,

$$\bar{y} = \bar{\hat{y}} + 0 = \bar{\hat{y}}$$

Note: These last 3 results use the fact that the model *includes an intercept*.

2.2 Partitioned and Partial Regression

We can solve for "blocks" of $\boldsymbol{b}$. Suppose we partition our model into 2 blocks:

$$\underset{(n\times 1)}{\boldsymbol{y}} = \underset{(n\times k_1)(k_1\times 1)}{X_1\boldsymbol{\beta}_1} + \underset{(n\times k_2)(k_2\times 1)}{X_2\boldsymbol{\beta}_2} + \underset{(n\times 1)}{\boldsymbol{\varepsilon}} \ .$$

We could solve for the OLS estimator of $\boldsymbol{\beta}_1$ only:

$$\boldsymbol{b_1} = (X_1'X_1)^{-1}X_1'\boldsymbol{y} - (X_1'X_1)^{-1}X_1'X_2\boldsymbol{b_2}$$

When does $\boldsymbol{b_1} = (X_1'X_1)^{-1}X_1'\boldsymbol{y}$?

Above we have a solution for $\boldsymbol{b_1}$ in terms of $\boldsymbol{b_2}$. We can substitute in a similar solution for $\boldsymbol{b_2}$ to obtain a solution for $\boldsymbol{b_1}$ that is a function of only the *X* and *y* data.

Define:

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

- M_2 is an "*idempotent*" matrix
- $M_2M_2 = M_2M_2' = M_2 = M_2'M_2$.

Then, we can write:

$$\boldsymbol{b_1} = (X_1'M_2X_1)^{-1}X_1'M_2\boldsymbol{y}$$

By the symmetry of the problem, we can interchange the "1" and "2" subscripts, and get:

$$\boldsymbol{b_2} = (X_2'M_1X_2)^{-1}X_2'M_1\boldsymbol{y}$$

So, finally, we can write:

$$\boldsymbol{b_1} = (X_1^{*\prime}X_1^*)^{-1}X_1^{*\prime}\boldsymbol{y_1^*} \qquad \boldsymbol{b_2} = (X_2^{*\prime}X_2^*)^{-1}X_2^{*\prime}\boldsymbol{y_2^*}$$

where:

$$X_1^* = M_2X_1 \;;\; X_2^* = M_1X_2 \;;\; \boldsymbol{y_1^*} = M_2\boldsymbol{y} \;;\; \boldsymbol{y_2^*} = M_1\boldsymbol{y}$$

These results are important for several reasons:

- Historically, for computational reasons. See the "Frisch-Waugh-Lovell Theorem".
- In certain situations $\boldsymbol{b_1}$ and $\boldsymbol{b_2}$ may have different properties. This is difficult to show without have separate formulas.
- Having established the M_i "residual maker" matrix, it is used frequently to derive other results.

Why is M_i called a "residual maker" matrix?

2.3 Goodness-of-Fit

- One way of measuring the "quality" of fitted regression model is by the extent to which the model "explains" the *sample variation* for $\boldsymbol{y}$.
- Sample variance of $\boldsymbol{y}$ is $\frac{1}{(n-1)}\sum_{i=1}^{n}(y_i-\bar{y})^2$.
- Or, we could just use $\sum_{i=1}^{n}(y_i-\bar{y})^2$ to measure *variability*.
- Our "fitted" regression model, using LS, gives us

$$\boldsymbol{y} = X\boldsymbol{b} + \boldsymbol{e} = \hat{\boldsymbol{y}} + \boldsymbol{e}$$

where $$\hat{\boldsymbol{y}} = X\boldsymbol{b} = X(X'X)^{-1}X'\boldsymbol{y}$$

- Recall that *if the model includes an intercept*, then the residuals sum to zero, and $\bar{y} = \bar{\hat{y}}$.

To simplify things, introduce the following matrix:

$$M^0 = [I_n - \tfrac{1}{n}\boldsymbol{i}\boldsymbol{i}']$$

where: $$\boldsymbol{i} = \begin{pmatrix}1\\ \vdots \\ 1\end{pmatrix} \quad ; \quad (n\times 1)$$

Note that:

- $M^0$is an idempotent matrix.
- $M^0\boldsymbol{i} = \boldsymbol{0}$.
- M^0 transforms elements of a vector into deviations from sample mean.
- $\boldsymbol{y}'M^0\boldsymbol{y} = \boldsymbol{y}'M^0M^0\boldsymbol{y} = \sum_{i=1}^{n}(y_i-\bar{y})^2$.

Let's check the third of these results:

$$M^0\boldsymbol{y} = \left\{\begin{bmatrix}1 & \cdots & 0\\ \vdots & \ddots & \vdots\\ 0 & \cdots & 1\end{bmatrix} - \left[\begin{bmatrix}1/n & \cdots & 1/n\\ \vdots & \ddots & \vdots\\ 1/n & \cdots & 1/n\end{bmatrix}\right]\right\}\begin{pmatrix}y_1\\ \vdots\\ y_n\end{pmatrix}$$

$$= \begin{bmatrix} y_1 - \frac{1}{n}y_1 - \frac{1}{n}y_2 \cdots - \frac{1}{n}y_n \\ \vdots \\ y_n - \frac{1}{n}y_1 - \frac{1}{n}y_2 - \cdots - \frac{1}{n}y_n \end{bmatrix} = \begin{pmatrix} y_1 - \bar{y} \\ \vdots \\ y_n - \bar{y} \end{pmatrix}.$$

Returning to our "fitted" model:

$$\boldsymbol{y} = X\boldsymbol{b} + \boldsymbol{e} = \hat{\boldsymbol{y}} + \boldsymbol{e}$$

So, we have:

$$M^0\boldsymbol{y} = M^0\hat{\boldsymbol{y}} + M^0\boldsymbol{e} = M^0\hat{\boldsymbol{y}} + \boldsymbol{e}\ .$$

[$M^0\boldsymbol{e} = \boldsymbol{e}$; because the residuals sum to zero.]

Then –

$$\boldsymbol{y}'M^0\boldsymbol{y} = \boldsymbol{y}'M^{0\prime}M^0\boldsymbol{y} = (M^0\hat{\boldsymbol{y}} + \boldsymbol{e})'(M^0\hat{\boldsymbol{y}} + \boldsymbol{e})$$

$$= \hat{\boldsymbol{y}}'M^0\hat{\boldsymbol{y}} + \boldsymbol{e}'\boldsymbol{e} + 2\boldsymbol{e}'M^0\hat{\boldsymbol{y}}$$

However,

$$\boldsymbol{e}'M^0\hat{\boldsymbol{y}} = \boldsymbol{e}'M^{0\prime}\hat{\boldsymbol{y}} = (M^0\boldsymbol{e})'\hat{\boldsymbol{y}} = \boldsymbol{e}'\hat{\boldsymbol{y}} = \boldsymbol{e}'X(X'X)^{-1}X'\boldsymbol{y} = 0\ .$$

So, we have –

$$\boldsymbol{y}'M^0\boldsymbol{y} = \hat{\boldsymbol{y}}'M^0\hat{\boldsymbol{y}} + \boldsymbol{e}'\boldsymbol{e}$$

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}e_i^2$$

$$\textbf{SST} = \textbf{SSR} + \textbf{SSE}$$

Recall: $\bar{\hat{y}} = \bar{y}$.

This lets us define the "**Coefficient of Determination**" –

$$R^2 = \left(\frac{SSR}{SST}\right) = 1 - \left(\frac{SSE}{SST}\right)$$

Note:

- The second equality in definition of R^2 holds only if model *includes an intercept*.
- $R^2 = \left(\frac{SSR}{SST}\right) \geq 0$
- $R^2 = 1 - \left(\frac{SSE}{SST}\right) \leq 1$
- So, $0 \leq R^2 \leq 1$
- Interpretation of "0" and "1" ?
- R^2 is *unitless* .

What happens if we add ***any*** regressor(s) to the model?

$$\boldsymbol{y} = X_1\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon} \quad ; \quad [1]$$

Then:

$$\boldsymbol{y} = X_1\boldsymbol{\beta}_1 + X_2\boldsymbol{\beta}_2 + \boldsymbol{u} \quad ; \quad [2]$$

(A) Applying LS to [2]:

$min.(\hat{\boldsymbol{u}}'\hat{\boldsymbol{u}})$; $\hat{\boldsymbol{u}} = \boldsymbol{y} - X_1\boldsymbol{b_1} - X_2\boldsymbol{b_2}$

(B) Applying LS to [1]:

$min.(\boldsymbol{e}'\boldsymbol{e})$; $\boldsymbol{e} = \boldsymbol{y} - X_1\hat{\boldsymbol{\beta}}_1$

Problem (B) is just Problem (A), subject to restriction: $\boldsymbol{\beta_2} = 0$. Minimized value in (A) must be $\leq$ minimized value in (B). So, $\hat{\boldsymbol{u}}'\hat{\boldsymbol{u}} \leq \boldsymbol{e}'\boldsymbol{e}$.

What does this imply?

- Adding *any* regressor(s) to the model *cannot increase* (and typically will *decrease*) the sum of squared residuals.
- So, adding *any* regressor(s) to the model *cannot decrease* (and typically will *increase*) the value of R^2.
- Means that R^2 is not really a very interesting measure of the "quality" of the regression model, in terms of explaining sample variability of the dependent variable.
- For these reasons, we usually use the "adjusted" Coefficient of Determination.

  We modify $R^2 = [1 - \frac{\boldsymbol{e}\prime\boldsymbol{e}}{\boldsymbol{y}\prime M^0\boldsymbol{y}}]$ to become:

$$\bar{R}^2 = [1 - \frac{\boldsymbol{e'e}/(n-k)}{\boldsymbol{y'}M^0\boldsymbol{y}/(n-1)}] \, .$$

- What are we doing here?

  We're adjusting for "degrees of freedom" in numerator and denominator.

- "Degrees of freedom" = number of independent pieces of information.

- $\boldsymbol{e} = \boldsymbol{y} - X\boldsymbol{b}$. We estimate *k* parameters from the *n* data-points. We have (*n* – *k*) "degrees of freedom" associated with the fitted model.

- In denominator – have constructed $\bar{y}$ from sample. "Lost" one degree of freedom.

- Possible for $\bar{R}^2 < 0$ (even with intercept in the model).

- $\bar{R}^2$ can *increase or decrease* when we add regressors.

- When will it increase (decrease)?

  In multiple regression, $\bar{R}^2$ will *increase* (decrease) if a variable is deleted, if and only if the associated t-statistic has *absolute value less than* (greater than) unity.

- If model *doesn't* include an intercept, then **SST** ≠ **SSR** + **SSE**, and in this case no longer any guarantee that $0 \leq R^2 \leq 1$.

- Must be careful comparing R^2 and $\bar{R}^2$ values across models.

  Example –

  (1) $\widehat{C}_i = 0.5 + 0.8Y_i$; $R^2 = 0.90$

  (2) $\log(\widehat{C}_i) = 0.2 + 0.75Y_i$; $R^2 = 0.80$

  Sample variation is in *different units*.