Lecture 8: Linear Regression with Multiple Regressors

Zheng Tian

Outline

The problem of a simple linear regression

The simple linear regression model

$$TestScore = \beta_0 + \beta_1 \times STR + OtherFactors$$

Question: Is this model adequate to characterize the determination of test scores?

- It ignores many important factors, simply lumped into *OtherFactors*, the error term, u_i, in the regression model.
- What are possible other important factors?
  - School district characteristics: average income level, demographic components
  - School characteristics: teachers' quality, school buildings,
  - Student characteristics: family economic conditions, individual ability

Percentage of English learners as an example

The percentage of English learners in a school district could be an relevant and important determinant of test scores, which is omitted in the simple regression model.

How can it affect the estimate of the effect of student-teacher ratios on test score?

- High percentage of English learners $\Rightarrow$ large student-teacher ratios.
- High percentage of English learners $\Rightarrow$ lower test scores.
- The estimated effect of student-teacher ratios may in fact include the influence from the high percentage of English learners.
- In the terminology of statistics, the magnitude of the coefficient on student-teacher ratio is overestimated.
- The problem is called the omitted variable bias

Solutions to the problem of ignoring important factors

We can include these important but ignored variables, like the percentage of English learners ($PctEL$), in the regression model.

$$TestScore_i = \beta_0 + \beta_1 STR_i + \beta_2 PctEL_i + OtherFactors_i$$

A regression model with more than one regressors is a multiple regression model.

A multiple regression model

The general form of a multiple regression model is

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i,\ i = 1, \ldots, n \tag{1}$$

where

- Y_i is the i^{th} observation on the dependent variable;
- $X_{1i}, X_{2i}, \ldots, X_{ki}$ are the i^{th} observation on each of the k regressors; and
- u_i is the error term associated with the i^{th} observation, representing all other factors that are not included in the model.

The components in a multiple regression model

- The population regression line (or population regression function)

$$E(Y_i|X_{1i}, \ldots, X_{ki}) = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki}$$

- $\beta_1, \ldots, \beta_k$ are the coefficients on the corresponding X_i, $i = 1, \ldots, k$.
- β_0 is the intercept, which can also be thought of the coefficient on a regressor X_0 that equals 1 for all observations.
  - Including X_0, there are $k + 1$ regressors in the multiple regression model.

The interpretation of β_i: Holding other things constant

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + u \tag{2}$$

The coefficient β_i on the regressor X_i for $i = 1, \ldots, k$ measures the effect on Y of a unit change in X_i, holding other X constant.

An example

Suppose we have two regressors X_1 and X_2 and we are interested in the effect of X_1 on Y. We can let X_1 change by ΔX_1 and holding X_2 constant. Then, the new value of Y is

$$Y + \Delta Y = \beta_0 + \beta_1(X_1 + \Delta X_1) + \beta_2 X_2$$

Subtracting $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$, we have $\Delta Y = \beta_1 \Delta X_1$. That is

$$\beta_1 = \frac{\Delta Y}{\Delta X}, \text{ holding } X_2 \text{ constant}$$

The partial effect

If Y and X_i for $i = 1, \ldots, k$ are continuous and differentiable variables, β_i is as simply as the partial derivative of Y with respect to X_i. That is

$$\beta_i = \frac{\partial Y}{\partial X_i}$$

By the definition of a partial derivative, β_i is just the effect of a marginal change in X_i on Y holding other X constant.

Look at the data in terms of vectors and matrix

	A	B	C	D	E
1	obs_num	dist_cod	testscr	str	el_pct
2	1	75119	690.8000	17.8899	0.0000
3	2	61499	661.2000	21.5247	4.5833
4	3	61549	643.6000	18.6972	30.0000
5	4	61457	647.7000	17.3571	0.0000
6	5	61523	640.8500	18.6713	13.8577
7	6	62042	605.5500	21.4063	12.4088
8	7	68536	606.7500	19.5000	68.7179
9	8	63834	609.0000	20.8941	46.9595
10	9	62331	612.5000	19.9474	30.0792
11	10	67306	612.6500	20.8056	40.2759
12	11	65722	615.7500	21.2381	52.9148
13	12	62174	616.3000	21.0000	54.6099
14	13	71795	616.3000	20.6000	42.7184
15	14	72181	616.3000	20.0082	20.5339
16	15	72298	616.4500	18.0278	80.1233
17	16	72041	617.3500	20.2520	49.4131
18	17	63594	618.0500	16.9779	85.5397
19	18	63370	618.3000	16.5098	58.9074
20	19	64709	619.8000	22.7040	77.0058
21	20	63560	620.3000	19.9111	49.8140
22	21	63230	620.5000	18.3333	40.6818
23	22	72058	621.4000	22.6190	16.2105
24	23	63842	621.7500	19.4483	45.0749
25	24	71811	622.0500	25.0526	39.0756
26	25	65748	622.6000	20.6754	76.6653
27	26	72272	623.1000	18.6824	40.4912
28	27	65961	623.2000	22.8455	73.7202
29	28	63313	623.4500	19.2667	70.0115
30	29	72199	623.6000	19.2500	55.9622

Figure: The California data set in Excel

- Each row represents an observation of all variables pertaining to a school district.
- Each column represents a variable with all observations.
- The whole dataset can be seen as a matrix.

Define variables in matrix notation

Write all the variables in vector and matrix notation

$$\underbrace{\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}}_{\text{Dependent variable}}, \underbrace{\mathbf{X} = \begin{pmatrix} 1 & X_{11} & \cdots & X_{k1} \\ 1 & X_{12} & \cdots & X_{k2} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{1n} & \cdots & X_{kn} \end{pmatrix}}_{\text{Independent variables}}, \underbrace{\mathbf{u} = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}}_{\text{Errors}}, \underbrace{\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}}_{\text{Coefficients}}$$

Write the multiple regression model in matrix notation

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} \tag{3}$$

Why do we use matrix notation

Concise, easy to derive properties; big-picture perspective.

Two other ways to write the regression model

Write X in row vectors

- The i$^{\text{th}}$ row in X is a $(k+1) \times 1$ vector

$$\boldsymbol{x}_i = \begin{pmatrix} 1 \\ X_{1i} \\ \vdots \\ X_{ki} \end{pmatrix}. \text{ Thus, its transpose is } \boldsymbol{x}_i' = (1, X_{1i}, \cdots, X_{ki})$$

- We can write the regression model (Equation 3) as

$$Y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + u_i, \; i = 1, \dots, n \tag{4}$$

Two other ways to write the regression model (cont'd)

Write X in vector vectors

- The i$^{\text{th}}$ column in $\mathbf{X}$ is a $n \times 1$ vector

$$\boldsymbol{X}_i = \begin{pmatrix} X_{i1} \\ \vdots \\ X_{in} \end{pmatrix}. \text{ The first column is } \boldsymbol{\iota} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}. \text{ Thus } \mathbf{X} = (\boldsymbol{\iota}, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_k)$$

- The regression model (Equation 3) can be re-written as

$$\mathbf{Y} = \beta_0 \boldsymbol{\iota} + \beta_1 \boldsymbol{X}_1 + \cdots + \beta_k \boldsymbol{X}_k + \mathbf{u} \tag{5}$$

The minimization problem and the OLS estimator

- The core idea of the OLS estimator for a multiple regression model remains the same as in a simple regression model: minimizing the sum of the squared residuals.
- Let $\mathbf{b} = [b_0, b_1, \ldots, b_k]'$ be some estimators of $\boldsymbol{\beta} = [\beta_0, \beta_1, \ldots, \beta_k]'$.
- The predicted Y_i is

$$\hat{Y}_i = b_0 + b_1 X_{1i} + \cdots + b_k X_{ki} = \mathbf{x}_i'\mathbf{b},\ i = 1, \ldots,$$
$$\text{or in matrix notation } \hat{\mathbf{Y}} = \mathbf{X}\mathbf{b}$$

- The residuals, i.e., the prediction mistakes, with $\mathbf{b}$ is

$$\hat{u}_i = Y_i - b_0 - b_1 X_{1i} - \cdots - b_k X_{ki} = Y_i - \mathbf{x}_i'\mathbf{b}$$
$$\text{or in matrix notation } \hat{\mathbf{u}} = \mathbf{Y} - \mathbf{X}\mathbf{b}$$

The minimization problem and the OLS estimator (cont'd)

- The sum of the squared residuals is

$$
\begin{aligned}
S(\mathbf{b}) = S(b_0, b_1, \ldots, b_k) &= \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_{1i} - \cdots - b_k X_{ki})^2 \\
&= \sum_{i=1}^{n} (Y_i - \mathbf{x}_i' \mathbf{b})^2 = (\mathbf{Y} - \mathbf{X}\mathbf{b})'(\mathbf{Y} - \mathbf{X}\mathbf{b}) \\
&= \hat{\mathbf{u}}'\hat{\mathbf{u}} = \sum_{i=1}^{n} \hat{u}_i^2
\end{aligned}
$$

- The OLS estimator is the solution to the following minimization problem:

$$
\min_{\mathbf{b}} \; S(\mathbf{b}) = \hat{\mathbf{u}}'\hat{\mathbf{u}} \tag{6}
$$

The OLS estimator of $\boldsymbol{\beta}$ as a solution to the minimization problem

- Solve the minimization problem:

$$\text{F.O.C.: } \frac{\partial S(\mathbf{b})}{\partial b_j} = 0, \text{ for } j = 0, 1, \ldots, k$$

- The derivative of $S(b_0, \ldots, b_k)$ with respect to b_j is

$$\frac{\partial}{\partial b_j} \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_{1i} - \cdots - b_k X_{ki})^2 =$$

$$-2 \sum_{i=1}^{n} X_{ji} (Y_i - b_0 - b_1 X_{1i} - \cdots - b_k X_{ki}) = 0$$

- There are $k+1$ such equations. Solving the system of equations, we obtain the OLS estimator $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \ldots, \hat{\beta}_k)'$.

The OLS estimator in matrix notation

Let $\hat{\beta}$ denote the OLS estimator. Then the expression of $\hat{\beta}$ is given by

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \tag{7}$$

Some useful results of matrix calculus

To prove Equation (7), we need to use some results of matrix calculus.

$$\frac{\partial \mathbf{a}'\mathbf{x}}{\partial \mathbf{x}} = \mathbf{a},\ \frac{\partial \mathbf{x}'\mathbf{a}}{\partial \mathbf{x}} = \mathbf{a},\ \text{ and } \frac{\partial \mathbf{x}'\mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = (\mathbf{A}+\mathbf{A}')\mathbf{x} \tag{8}$$

when $\mathbf{A}$ is symmetric, then $(\partial \mathbf{x}'\mathbf{A}\mathbf{x})/(\partial \mathbf{x}) = 2\mathbf{A}\mathbf{x}$

The proof

Proof of Equation (7).

$$S(\mathbf{b}) = \hat{\mathbf{u}}'\hat{\mathbf{u}} = \mathbf{Y}'\mathbf{Y} - \mathbf{b}'\mathbf{X}'\mathbf{Y} - \mathbf{Y}'\mathbf{X}\mathbf{b} - \mathbf{b}'\mathbf{X}'\mathbf{X}\mathbf{b}$$

The first order conditions for minimizing $S(\mathbf{b})$ with respect to $\mathbf{b}$ is

$$\begin{aligned} -2\mathbf{X}'\mathbf{Y} - 2\mathbf{X}'\mathbf{X}\mathbf{b} &= \mathbf{0} \\ \mathbf{X}'\mathbf{X}\mathbf{b} &= \mathbf{X}'\mathbf{Y} \qquad (9) \end{aligned}$$

Then

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

given that $\mathbf{X}'\mathbf{X}$ is invertible. □

Note that Equation (9) represents a system of equations with $k+1$ equations.

The OLS estimator of $\hat{\beta}_1$ in a simple regression model

The simple linear regression model written in matrix notation is

$$\mathbf{Y} = \beta_0 \boldsymbol{\iota} + \beta_1 \mathbf{X}_1 + \mathbf{u} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$$

where

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \mathbf{X} = \begin{pmatrix} \boldsymbol{\iota} & \mathbf{X}_1 \end{pmatrix} = \begin{pmatrix} 1 & X_{11} \\ \vdots & \vdots \\ 1 & X_{1n} \end{pmatrix}, \mathbf{u} = \begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$$

The OLS estimator of $\hat{\beta}_1$ in a simple regression model (cont'd)

Let's get the components in Equation (7) step by step.

Step (1): compute $(\mathbf{X}'\mathbf{X})$

$$
\begin{aligned}
\mathbf{X}'\mathbf{X} &= \begin{pmatrix} \boldsymbol{\iota}' \\ \mathbf{X}_1' \end{pmatrix} \begin{pmatrix} \boldsymbol{\iota} & \mathbf{X}_1 \end{pmatrix} = \begin{pmatrix} 1 & \cdots & 1 \\ X_{11} & \cdots & X_{1n} \end{pmatrix} \begin{pmatrix} 1 & X_{11} \\ \vdots & \vdots \\ 1 & X_{1n} \end{pmatrix} \\
&= \begin{pmatrix} \boldsymbol{\iota}'\boldsymbol{\iota} & \boldsymbol{\iota}'\mathbf{X}_1 \\ \mathbf{X}_1'\boldsymbol{\iota} & \mathbf{X}_1'\mathbf{X}_1 \end{pmatrix} = \begin{pmatrix} n & \sum_{i=1}^{n} X_{1i} \\ \sum_{i=1}^{n} X_{1i} & \sum_{i=1}^{n} X_{1i}^2 \end{pmatrix}
\end{aligned}
$$

The OLS estimator of $\hat{\beta}_1$ in a simple regression model (cont'd)

Step (2): compute $(\mathbf{X}'\mathbf{X})^{-1}$

The inverse of a 2×2 matrix

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}^{-1} = \frac{1}{a_{11}a_{22} - a_{12}a_{21}} \begin{pmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{pmatrix}$$

The inverse of $\mathbf{X}'\mathbf{X}$

$$(\mathbf{X}'\mathbf{X})^{-1} = \frac{1}{n\sum_{i=1}^{n} X_{1i}^2 - (\sum_{i=1}^{n} X_{1i})^2} \begin{pmatrix} \sum_{i=1}^{n} X_{1i}^2 & -\sum_{i=1}^{n} X_{1i} \\ -\sum_{i=1}^{n} X_{1i} & n \end{pmatrix}$$

The OLS estimator of $\hat{\beta}_1$ in a simple regression model (cont'd)

Step (3): compute $\mathbf{X}'\mathbf{Y}$

$$\mathbf{X}'\mathbf{Y} = \begin{pmatrix} \iota' \\ \mathbf{X}_1' \end{pmatrix} \mathbf{Y} = \begin{pmatrix} 1 & \cdots & 1 \\ X_{11} & \cdots & X_{1n} \end{pmatrix} \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} \iota'\mathbf{Y} \\ \mathbf{X}_1'\mathbf{Y} \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n Y_i \\ \sum_{i=1}^n X_{1i} Y_i \end{pmatrix}$$

Step (4): compute $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$

$$\begin{aligned} \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} &= \frac{1}{n\sum_{i=1}^n X_{1i}^2 - (\sum_{i=1}^n X_{1i})^2} \begin{pmatrix} \sum_{i=1}^n X_{1i}^2 & -\sum_{i=1}^n X_{1i} \\ -\sum_{i=1}^n X_{1i} & n \end{pmatrix} \begin{pmatrix} \sum_{i=1}^n Y_i \\ \sum_{i=1}^n X_{1i} Y_i \end{pmatrix} \\ &= \frac{1}{n\sum_{i=1}^n X_{1i}^2 - (\sum_{i=1}^n X_{1i})^2} \begin{pmatrix} \sum_{i=1}^n X_{1i}^2 \sum_{i=1}^n Y_i - \sum_{i=1}^n X_{1i} \sum_{i=1}^n X_{1i} Y_i \\ -\sum_{i=1}^n X_{1i} \sum_{i=1}^n Y_i + n\sum_{i=1}^n X_{1i} Y_i \end{pmatrix} \end{aligned}$$

The OLS estimator of $\hat{\beta}_1$ in a simple regression model (cont'd)

The formula of $\hat{\beta}_1$

$$\hat{\beta}_1 = \frac{n\sum_{i=1}^{n} X_{1i}Y_i - \sum_{i=1}^{n} X_{1i}\sum_{i=1}^{n} Y_i}{n\sum_{i=1}^{n} X_{1i}^2 - (\sum_{i=1}^{n} X_{1i})^2} = \frac{\sum_{i=1}^{n}(X_{1i} - \bar{X}_1)(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_{1i} - \bar{X}_1)^2}$$

The formula of $\hat{\beta}_0$

$$\hat{\beta}_0 = \frac{\sum_{i=1}^{n} X_{1i}^2\sum_{i=1}^{n} Y_i - \sum_{i=1}^{n} X_{1i}\sum_{i=1}^{n} X_{1i}Y_i}{n\sum_{i=1}^{n} X_{1i}^2 - (\sum_{i=1}^{n} X_{1i})^2} = \bar{Y} - \hat{\beta}_1\bar{X}_1$$

Application to Test Scores and the Student-Teacher Ratio

The simple regression compared with the multiple regression

The estimated simple linear regression model is

$$\widehat{TestScore} = 698.9 - 2.28 \times STR$$

The estimated multiple linear regression model is

$$\widehat{TestScore} = 686.0 - 1.10 \times STR - 0.65 \times PctEL$$

Explanations

- The interpretation of the new estimated coefficient on STR is, holding the percentage of English learners constant, a unit decrease in STR is estimated to increase test scores by 1.10 points.
- We can also interpret the estimated coefficient on $PctEL$ as, holding STR constant, one unit decrease in $PctEL$ increases test scores by 0.65 point.
- The magnitude of the negative effect of STR on test scores in the multiple regression is approximately half as large as when STR is the only regressor.

The standard errors of the regression (SER)

- The standard error of regression (SER) estimates the standard deviation of the error term $\mathbf{u}$. In multiple regression, the SER is

$$SER = s_{\hat{u}}, \;\; \text{where } s_{\hat{u}}^2 = \frac{\sum_{i=1}^{n} \hat{u}_i^2}{n-k-1} = \frac{SSR}{n-k-1} \tag{10}$$

- SSR is divided by $(n-k-1)$ because there are n observations and $(k+1)$ coefficients to be estimated.

R^2

TSS, ESS, and SSR

- The total sum of squares (TSS): $TSS = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$
- The explained sum of squares (ESS): $ESS = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$
- The sum of squared residuals (SSR): $SSR = \sum_{i=1}^{n} \hat{u}_i^2$

The equality still holds in multiple regression

$$TSS = ESS + SSR$$

Define R^2 as before

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{SSR}{TSS} \tag{11}$$

Limitations of R^2

- R^2 is valid only if a regression model is estimated using the OLS since otherwise it would not be true that $TSS = ESS + SSR$.
- R^2 defined in the form of the deviation from the mean is only valid when a constant term is included in regression.
  In a regression model without an intercept, use the uncentered version of R^2, which is also defined as

$$R_u^2 = \frac{EES}{TSS} = 1 - \frac{SSR}{TSS} \tag{12}$$

  where
  - $TSS = \sum_{i=1}^{n} Y_i^2$, $ESS = \sum_{i=1}^{2} \hat{Y}_i^2$, and $SSR = \sum_{i=1}^{n} \hat{u}_i^2$

  Note that in a regression without a constant term, the equality $TSS = ESS + SSR$ holds.

Limitation of R^2 (cont'd)

- R^2 increases whenever an additional regressor is included in a multiple regression model, unless the estimated coefficient on the added regressor is exactly zero.

Consider two regression models

$$\mathbf{Y} = \beta_0 + \beta_1 \mathbf{X}_1 + \mathbf{u} \tag{13}$$

$$\mathbf{Y} = \beta_0 + \beta_1 \mathbf{X}_1 + \beta_2 \mathbf{X}_2 + \mathbf{u} \tag{14}$$

Which model should have smaller *SSR*?

Limitation of R^2 (cont'd)

- Equation (14) have the smaller *SSR* than equation (13). Why?

  An additional $X_2 \Rightarrow$ More in the total variation of Y is explained $\Rightarrow$ Smaller *SSR* (unless $\hat{\beta}_2 = 0$)

- Since both models use the same **Y**, *TSS* must be the same. Because *SSR* decreases as more regressors are added, R^2 increases.
- In mathematics, this is essentially because the OLS estimation for equation (13) solves a constrained minimization problem, while that for equation (14) solves an unconstrained minimization problem.

The adjusted R^2

- The adjusted R^2 is, or $\bar{R}^2$, is a modified version of R^2.
- The $\bar{R}^2$ improves R^2 in the sense that it does not necessarily increase when a new regressor is added. The $\bar{R}^2$ is

$$\bar{R}^2 = 1 - \frac{SSR/(n-k-1)}{TSS/(n-1)} = 1 - \frac{n-1}{n-k-1}\frac{SSR}{TSS} = 1 - \frac{s_u^2}{s_Y^2} \tag{15}$$

- The adjustment is made by dividing SSR and TSS by their corresponding degrees of freedom, which is $n-k-1$ and $n-1$ respectively.
- s_u^2 is the sample variance of the OLS residuals, and s_Y^2 is the sample variance of Y.

Properties of $\bar{R}^2$

- The definition of the $\bar{R}^2$ in Equation (15) is valid only when a constant term is included in the regression model.
- Since $\frac{n-1}{n-k-1} > 1$, then it is always true that the $\bar{R}^2 < R^2$.
- $k \uparrow \Rightarrow \frac{SSR}{TSS} \downarrow$, but $k \uparrow \Rightarrow \frac{n-1}{n-k-1} \uparrow$.

  Whether $\bar{R}^2$ increases or decreases depends on which of these effects is stronger.
- The $\bar{R}^2$ can be negative. This happens when the regressors, taken together, reduce the sum of squared residuals by such a small amount that his reduction fails to offset the factor $\frac{n-1}{n-k-1}$.

The usefulness of the R^2 and $\bar{R}^2$

- Both R^2 and $\bar{R}^2$ are valid when the regression model is estimated by the OLS estimators. R^2 computed with estimators other than the OLS ones is usually called *pseudo* R^2.

- Their importance as measures of fit cannot be overstated. We cannot heavily reply on R^2 or $\bar{R}^2$ to judge whether some regressors should be included in the model or not.

The grouped regressors

Consider a multiple regression model

$$Y_i = \underbrace{\beta_0 + \beta_1 X_{1i} + \cdots + \beta_{k1} X_{k1,i}}_{\text{k1+1 regressors}} + \underbrace{\beta_{k1+1} X_{k1+1,i} + \cdots \beta_k X_k}_{\text{k2 regressors}} + u_i \tag{16}$$

In matrix notation, we write

$$\mathbf{Y} = \mathbf{X}_1 \boldsymbol{\beta}_1 + \mathbf{X}_2 \boldsymbol{\beta}_2 + \mathbf{u} \tag{17}$$

where

- $\mathbf{X}_1$ is an $n \times (k1 + 1)$ matrix composed of the intercept and the first $k1 + 1$ regressors in Equation (16),
- $\mathbf{X}_2$ is an $n \times k2$ matrix composed of the rest k_2 regressors.
- $\boldsymbol{\beta}_1 = (\beta_0, \beta_1, \ldots, \beta_{k1})'$ and $\boldsymbol{\beta}_2 = (\beta_{k1+1}, \ldots, \beta_k)'$.

Two estimation strategies

Suppose that we are interested in $\boldsymbol{\beta}_2$ but not much in $\boldsymbol{\beta}_1$ in Equation (17). How can we estimate $\boldsymbol{\beta}_2$?

The first strategy: the standard OLS estimation

We can obtain the OLS estimation of $\boldsymbol{\beta}_2$ with Equation (7), i.e., $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$. $\hat{\boldsymbol{\beta}}_2$ is a vector consisting of the last $k2$ elements in $\hat{\boldsymbol{\beta}}$. In matrix notation, we can get $\hat{\boldsymbol{\beta}}_2$ from the following equation

$$\begin{pmatrix} \hat{\boldsymbol{\beta}}_1 \\ \hat{\boldsymbol{\beta}}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{X}_1'\mathbf{X}_1 & \mathbf{X}_1'\mathbf{X}_2 \\ \mathbf{X}_2'\mathbf{X}_1 & \mathbf{X}_2'\mathbf{X}_2 \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{X}_1'\mathbf{Y} \\ \mathbf{X}_2'\mathbf{Y} \end{pmatrix}$$

The second strategy: the step OLS estimation

1. Regress each regressor in $\mathbf{X}_2$ on all regressors in $\mathbf{X}_1$, including the intercept, and get the residuals from this regression, denoted as $\widetilde{\mathbf{X}}_2$. That is, for each regressor $\mathbf{X}_i$ in $\mathbf{X}_2$, $i = k1+1, \ldots, k$, we estimate a multiple regression,

$$\mathbf{X}_i = \gamma_0 + \gamma_1 \mathbf{X}_1 + \cdots + \gamma_{k1} \mathbf{X}_{k1} + v$$

   The residuals from this regression is

$$\widetilde{\mathbf{X}}_i = X_i - \hat{\gamma}_0 - \hat{\gamma}_1 \mathbf{X}_1 - \cdots - \hat{\gamma}_{k1} \mathbf{X}_{k1}$$

   As such, we can get an $n \times k2$ matrix composed of all the residuals $\widetilde{\mathbf{X}}_2 = (\widetilde{\mathbf{X}}_{k1+1} \cdots \widetilde{\mathbf{X}}_k)$.
2. Regress $\mathbf{Y}$ on all regressors in $\mathbf{X}_1$, denoting the residuals from this regression as $\widetilde{\mathbf{Y}}$.
3. Regress $\widetilde{\mathbf{Y}}$ on $\widetilde{\mathbf{X}}_2$, and obtain the estimates of β_2 as $\beta_2 = (\widetilde{\mathbf{X}}_2' \widetilde{\mathbf{X}}_2)^{-1} \widetilde{\mathbf{X}}_2' \widetilde{\mathbf{Y}}$.

The Frisch-Waugh-Lovell Theorem

The Frisch-Waugh-Lovell (FWL) Theorem states that

1. the OLS estimates of β_2 using the second strategy and that from the first strategy are numerically identical.
2. the residuals from the regression of $\widetilde{\mathbf{Y}}$ on $\widetilde{\mathbf{X}}_2$ and the residuals from Equation (17) are numerically identical.

An understanding of the FWL theorem

The FWL theorem provides a mathematical statement of how the multiple regression coefficients in $\hat{\beta}_2$ capture the effects of $\mathbf{X}_2$ on $\mathbf{Y}$, controlling for other $\mathbf{X}$.

- Step 1 purges the effects of the regressors in $\mathbf{X}_1$ on the regressors in $\mathbf{X}_2$
- Step 2 purges the effects of the regressors in $\mathbf{X}_1$ on $\mathbf{Y}$.
- Step 3 estimates the effect of the regressors in $\mathbf{X}_2$ on $\mathbf{Y}$ using the parts in $\mathbf{X}_2$ and $\mathbf{Y}$ that have excluded the effects of $\mathbf{X}_1$.

An example of the FWL theorem

Consider a regression model with single regressor $Y_i = \beta_0 + \beta_1 X_i + u_i$.

Following the estimation strategy in the FWL theorem, we can carry out the following regressions,

1. Regress Y_i on 1. That is, estimate the model $Y_i = \alpha + e_i$. Then, the OLS estimator of α is $\bar{Y}$ and the residuals is $y_i = Y_i - \bar{Y}$
2. Similarly, regress X_i on 1. Then the residuals from this regression is $x_i = X_i - \bar{X}$.
3. Regress y_i on x_i without intercept. That is, estimate the model $y_i = \beta_1 x_i + v_i$
4. We can obtain $\hat{\beta}_1$ directly by applying the formula in Equation (7). That is

$$\hat{\beta}_1 = (\mathbf{x}_1'\mathbf{x}_1)^{-1}\mathbf{x}_1'\mathbf{y} = \frac{\sum_i x_{1i} y_i}{\sum_i x_{1i}^2} = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_i (X_i - \bar{X})^2}$$

The least squares assumptions in Multiple Regression

Assumption #1

$E(u_i|\boldsymbol{x}_i) = 0$. The conditional mean of u_i given $X_{1i}, X_{2i}, \ldots, X_{ki}$ has mean of zero. This is the key assumption to assure that the OLS estimators are unbiased.

Assumption #2

$(Y_i, \boldsymbol{x}_i')$ $i = 1, \ldots, n$ are i.i.d. This assumption holds automatically if the data are collected by simple random sampling.

Assumption #3

Large outliers are unlikely, i.e.,, $0 < E(\mathbf{X}^4) < \infty$ and $0 < E(\mathbf{Y}^4) < \infty$. That is, the dependent variables and regressors have finite kurtosis.

Assumption #4

No perfect multicollinearity. The regressors are said to exhibit perfect multicollinearity if one of the regressor is a perfect linear function of the other regressors.