Lecture 8: Linear Regression with Multiple Regressors

Zheng Tian

1 Introduction

1.1 Overview

This lecture extends the simple regression model to the multiple regression model. Many aspects of multiple regression parallel those of regression with a single regressor. The coefficients of the multiple regression model can be estimated using the OLS estimation method. The algebraic and statistical properties of the OLS estimators of the multiple regression are also similar to those of the simple regression. However, there are some new concepts, such as the omitted variable bias and multicollinearity, to deepen our understanding of the OLS estimation.

1.2 Learning goals

- Be able to set up a multiple regression model with matrix notation.
- Understand the meaning of holding other things constant.
- Estimate the multiple regression model with the OLS estimation.
- Understand the Frisch-Waugh-Lovell theorem.
- Capable of detecting the omitted variable bias and multicollinearity.

1.3 Readings

- *Introduction to Econometrics* by Stock and Watson. Read thoroughly Chapter 6 and Sections 18.1 and 18.2
- *Introductory Econometrics: a Modern Approach* by Wooldridge. Chapter 3.

2 The Multiple Regression Model

2.1 The problem of a simple linear regression

In the last two lectures, we use a simple linear regression model to examine the effect of class sizes on test scores in the California elementary school districts. The simple linear regression model with only one regressor is

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

What are your intuitive challenge against this model?

- It ignores too many other important factors that are presumably included in *OtherFactors*, which is the error term, u_i, in the regression model.
  - What are possible other important factors?

    School characteristics teachers' quality, school buildings, equipment

    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?
  - The districts with high percentages of English learners tend to have large student-teacher ratios. That is, these two variables are correlated with the correlation coefficient being 0.17.
  - The higher percentage of English learners a district has, the lower test scores students there will make.
  - In the simple regression model, the estimated negative coefficient on student-teacher ratios on test scores could include not only the negative influence from class sizes but also that from the percentage of English learners.
  - In the terminology of statistics, the magnitude of the coefficient on student-teacher ratio is **overestimated**.
  - Generally, we commit an **omitted variable bias** by setting up a simple regression model. We will explore the omitted variable bias in the last section in this lecture.

- Solutions to the omitted variable bias

  Include the percentage of English learners in the regression model.

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

  which is a multiple regression model. In fact, we can include other independent variables in the multiple regression model.

2.2 The form of 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$^{\text{th}}$ observation on the dependent variable; $X_{1i}, X_{2i}, \ldots, X_{ki}$ are the i$^{\text{th}}$ observation on each of the k regressors; and u_i is the error term associated with the i$^{\text{th}}$ observation.
- The **population regression line** (or **population regression function**) is the relationship that holds between Y and X on average in the population

$$E(Y_i | X_{1i} = x_1, \ldots, X_{ki} = x_k) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

- $\beta_1, \ldots, \beta_k$ 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_{0i} that equals 1 for all observations.
  - Including X_{0i}, there are $k + 1$ regressors in the multiple regression model.
  - The linear regression model with a single regressor is in fact a multiple regression model with two regressors, 1 and X.
- u_i is the error term. It can be assumed to be **homoskedastic**. That is, $\text{Var}(u_i | X_{1i}, \ldots, X_{ki}) = \sigma_u^2$, which is a constant independent of the value of $X_{1i}, \ldots, X_{ki}$. Otherwise, it can be assumed to be **heteroskedastic**.

2.3 The interpretation of β_i

Holding other things constant

We can suppress the subscript i in Equation (1) so that we can re-write it as

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

In the multiple regression model in Equation (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** or **controlling for other X**.

The phrase of **holding other things constant** (or **ceteris paribus** in Latin) is important in economics. To disentangle the particular causal effect of one explanatory variable X on Y from all other confounding factors, we must hold these other factors constant so that it is meaningful to compare the values of Y before and after a change in X happens.

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 and holding X_2 constant. Then, the new value of Y is

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

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

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

Partial effect

If Y and X_i for $i = 1, \ldots, k$ are continuous and differentiated variables, from Equation (2), we know that 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 on Y holding other X constant.

2.4 The matrix notation of a multiple regression model

Consider the matrix notation as a way to organize data

When we save the data set of California school districts in Excel, it is saved in a spreadsheet like the following,

Each row represents an observation of all variables pertaining to a school district, and each column represents a variable with all observations. This format of data display can be concisely denoted using vectors and a 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 1: The California data set in Excel

Let us first define the following vectors and matrices:

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \mathbf{X} = \begin{pmatrix} 1 & X_{11} & \cdots & X_{k1} \\ 1 & X_{12} & \cdots & X_{k2} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{1n} & \cdots & X_{kn} \end{pmatrix} = \begin{pmatrix} \mathbf{X}_1' \\ \mathbf{X}_2' \\ \vdots \\ \mathbf{X}_n' \end{pmatrix}, \mathbf{U} = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}$$

- $\mathbf{Y}$ is an $n \times 1$ vector of n observations on the dependent variable.
- $\mathbf{X}$ is an $n \times (k+1)$ matrix of n observations on $k+1$ regressors which include the intercept term as a regressor of 1's.
- $\mathbf{X}_i$ is a $(k+1) \times 1$ vector of the i$^{\text{th}}$ observation on all $(k+1)$ regressors. Thus, $\mathbf{X}_i'$ denotes the i$^{\text{th}}$ row in $\mathbf{X}$.
- $\boldsymbol{\beta}$ is a $(k+1) \times 1$ vector of the $(k+1)$ regression coefficients.
- $\mathbf{U}$ is an $n \times 1$ vector of the n error terms.

Write a multiple regression model with matrix notation

- The multiple regression model for one observation

  The multiple regression model in Equation (1) for the i$^{\text{th}}$ observation can be written as

  $$Y_i = \mathbf{X}_i'\boldsymbol{\beta} + u_i, \ i = 1, \ldots, n \tag{3}$$

- The multiple regression model for all observations

  Stacking all n observations in Equation (3) yields the multiple regression model in matrix form:

  $$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{U} \tag{4}$$

  $\mathbf{X}$ can also be written in terms of column vectors as

  $$\mathbf{X} = [\boldsymbol{X}_0, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_k]$$

  where $\boldsymbol{X}_i = [X_{i1}, X_{i2}, \ldots, X_{in}]'$ is a $n \times 1$ vector of n observations of the k$^{\text{th}}$ regressor. $\boldsymbol{X}_0$ is a vector of 1s. That is, $\boldsymbol{X}_0 = [1, 1, \ldots, 1]'$. More often, we use $\boldsymbol{\iota}$ to denote such a vector of 1s. [1]

[1] $\boldsymbol{\iota}$ has the following properties: (1) $\boldsymbol{\iota}'\mathbf{x} = \sum_{i=1}^{n} x_i$ for an $n \times 1$ vector $\mathbf{x}$, (2) $\boldsymbol{\iota}'\boldsymbol{\iota} = n$ and $(\boldsymbol{\iota}'\boldsymbol{\iota})^{-1} = 1/n$, (3) $\boldsymbol{\iota}' (\boldsymbol{\iota}'\boldsymbol{\iota})^{-1} \mathbf{x} = \bar{x}$, and (4) $\boldsymbol{\iota}'\mathbf{X}\boldsymbol{\iota} = \sum_{i=1}^{n} \sum_{j=1}^{n} x_{ij}$ for an $n \times n$ matrix $\mathbf{X}$.

Thus, Equation (4) can be re-written as

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

3 The OLS Estimator in Multiple Regression

3.1 The OLS estimator

The minimization problem

The idea of the ordinary least squares estimation for a multiple regression model is exactly the same as for a simple regression model. The OLS estimators of the multiple regression model are obtained by minimizing the sum of the squared prediction mistakes.

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 can be obtained by

$$\hat{Y}_i = b_0 + b_1 X_{1i} + \cdots + b_k X_{ki} = \mathbf{X}_i' \mathbf{b},\ i = 1, \ldots, n$$

or

$$\hat{\mathbf{Y}} = \mathbf{X}\mathbf{b}$$

The prediction mistakes with $\mathbf{b}$, or called the residuals, are

$$\hat{u}_i = Y_i - b_0 - b_1 X_{1i} - \cdots - b_k X_{ki} = Y_i - \mathbf{X}_i' \mathbf{b}$$

or in vector notation, the residual vector is

$$\hat{\mathbf{u}} = \mathbf{Y} - \mathbf{X}\mathbf{b}$$

Then the sum of the squared prediction mistakes (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

The formula for the OLS estimator is obtained by taking the derivative of the sum of squared prediction mistakes, $S(b_0, b_1, \ldots, b_k)$, with respect to each coefficient, setting these derivatives to zero, and solving for the estimator $\hat{\boldsymbol{\beta}}$.

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 for $j = 0, \ldots, k$. Solving this system of equations, we obtain the OLS estimator $\hat{\boldsymbol{\beta}} = (b_0, \ldots, b_k)'$.

Using matrix notation, the formula for the OLS estimator $\hat{\boldsymbol{\beta}}$ is

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

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 $(\partial \mathbf{x}'\mathbf{A}\mathbf{x})/(\partial \mathbf{x}) = 2\mathbf{A}\mathbf{x}$

Proof of Equation (7). The sum of squared prediction mistakes is

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

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

$$-2\mathbf{X}'\mathbf{Y} - 2\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{0} \tag{9}$$

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.

3.2 Show that the OLS estimator of $\hat{\beta}_1$ in a simple regression model

Let take a simple linear regression model as an example. 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

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

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

First, the most important part is $(\mathbf{X}'\mathbf{X})^{-1}$.

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

Recall that the inverse of a 2×2 matrix can be calculated as follows

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

Thus, the inverse of $\mathbf{X}'\mathbf{X}$ is

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

Next, we compute $\mathbf{X}'\mathbf{Y}$.

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

Finally, we compute $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$, which is

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

Therefore, $\hat{\beta}_1$ is the second element of the vector pre-multiplied by the fraction, that is,

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

It follows that

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

3.3 Application to Test Scores and the Student-Teacher Ratio

Now we can apply the OLS estimation method of multiple regression to the application of California school districts. Recall that the estimated simple linear regression model is

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

Since we concern that the estimated coefficient on STR may be overestimated without considering the percentage of English learners in the districts, we include this new variable in the multiple regression model to control for the effect of English learners, yielding a new estimated regression model as

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

- 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, which verifies our concern that we may omit important variables in the simple linear regression model.

4 Measures of Fit in Multiple Regression

4.1 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

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

In the multiple regression model, SSR needs to be divided by $(n-k-1)$ because there are $(k+1)$ coefficients to be estimated using n samples.

4.2 The R^2

Like in the regression model with single regressor, we can define TSS, ESS, and SSR in the multiple regression model.

- **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)**: $SSR = \sum_{i=1}^{n}\hat{u}_i^2$

In matrix notation, we can write $Y_i - \bar{Y},\ i = 1, \ldots, n$ using the following vectors.

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix},\ \boldsymbol{\iota} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix},\ \mathbf{y} = \mathbf{Y} - \bar{Y}\boldsymbol{\iota} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} - \begin{pmatrix} \bar{Y} \\ \bar{Y} \\ \vdots \\ \bar{Y} \end{pmatrix}$$

Therefore, $\mathbf{y}$ represents the deviation from the mean of $Y_i,\ i = 1, \ldots, n$. Similarly, we can define the deviation from the mean of $\hat{Y}_i$, $i = 1, \ldots, n$ as $\hat{\mathbf{y}} = \hat{\mathbf{Y}} - \bar{Y}\boldsymbol{\iota}$. Then we can rewrite $TSS, ESS,$ and SSR as

$$TSS = \mathbf{y}'\mathbf{y},\ ESS = \hat{\mathbf{y}}'\hat{\mathbf{y}},\ \text{ and } SSR = \hat{\mathbf{u}}'\hat{\mathbf{u}}$$

In multiple regression, the relationship that

$$TSS = ESS + SSR,\ \text{or, } \mathbf{y}'\mathbf{y} = \hat{\mathbf{y}}'\hat{\mathbf{y}} + \hat{\mathbf{u}}'\hat{\mathbf{u}}$$

still holds so that we can define R^2 as

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

Limitations of R^2

1. R^2 is valid only if a regression model is estimated using the OLS since otherwise it would not be true that $TSS = ESS + SSR$.

2. R^2 that is defined using the deviation from the mean is only valid when a constant term is included in regression. Otherwise, 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 $TSS = \sum_{i=1}^n Y_i^2 = \mathbf{Y}'\mathbf{Y}$, $ESS = \sum_{i=1}^2 \hat{Y}_i^2 = \hat{\mathbf{Y}}'\hat{\mathbf{Y}}$, and $SSR = \sum_{i=1}^n \hat{u}_i^2 = \hat{\mathbf{u}}'\hat{\mathbf{u}}$, using the uncentered variables. Note that in a regression without a constant term, the equality $TSS = ESS + SSR$ is still true.

3. 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
$$\mathbf{Y} = \beta_0 + \beta_1 \mathbf{X}_1 + \mathbf{u} \tag{13}$$
$$\mathbf{Y} = \beta_0 + \beta_1 \mathbf{X}_1 + \beta_2 \mathbf{X}_2 + \mathbf{u} \tag{14}$$
Since both models use the same $\mathbf{Y}$, TSS must be the same. If the OLS estimator $\hat{\beta}_2$ does not equal 0, then SSR in Equation (13) is always larger than that of Equation (14) since the former SSR is minimized with respect to β_0, β_1 and with the constraint of $\beta_2 = 0$ and the latter is minimized without the constraint over β_2.

4.3 The adjusted R^2

The adjusted R^2 is, or $\bar{R}^2$, is a modified version of R^2 in Equation (11). 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, which is given in Equation (10); s_Y^2 is the sample variance of Y.
- 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 $\bar{R}^2 < R^2$.

- On one hand $k \uparrow \Rightarrow \frac{SSR}{TSS} \downarrow$. On the other hand, $k \uparrow \Rightarrow \frac{n-1}{n-k-1} \uparrow$. 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 $\mathbf{R^2}$ 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.

5 The Frisch-Waugh-Lovell Theorem

5.1 The grouped regressors

Consider a multiple regression model

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

which has k regressors. We can group these k regressors into two subset, $\mathbf{X}_1$ with k_1 regressors and $\mathbf{X}_2$ with k_2 regressors, with which we rewrite the multiple regression model above as

$$\mathbf{Y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{u} \tag{16}$$

5.2 An estimation strategy

Suppose that we are interested in $\boldsymbol{\beta}_1$ in Equation (16). We can perform the following steps to estimate $\boldsymbol{\beta}_1$:

1. Regress each regressor in $\mathbf{X}_1$ on all regressors in $\mathbf{X}_2$, denoting the residuals from this regression as $\widetilde{\mathbf{X}}_1$.
2. Regress $\mathbf{Y}$ on all regressors in $\mathbf{X}_2$, denoting the residuals from this regression as $\widetilde{\mathbf{Y}}$.
3. Regress $\widetilde{\mathbf{Y}}$ on $\widetilde{\mathbf{X}}_1$, and obtain the estimates of $\boldsymbol{\beta}_1$ as $(\widetilde{\mathbf{X}}_1'\widetilde{\mathbf{X}}_1)^{-1}\widetilde{\mathbf{X}}_1'\widetilde{\mathbf{Y}}$.

5.3 The Frisch-Waugh-Lovell Theorem

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

1. the OLS estimates of $\boldsymbol{\beta}_1$ using the steps above and the OLS estimates of $\boldsymbol{\beta}_1$ computed directly from Equation (16) are numerically identical.
2. the residuals from the regression of $\widetilde{\mathbf{Y}}$ on $\widetilde{\mathbf{X}}_1$ and the residuals from Equation (16) are numerically identical.

The proof of the FWL theorem is beyond the scope of this proof. Interested students may refer to Exercise 18.7. Understanding the meaning of this theorem is much more important than understanding the proof.

The FWL theorem provides a mathematical statement of how the multiple regression coefficient $\hat{\boldsymbol{\beta}}_1$ estimates the effect on $\mathbf{Y}$ of $\mathbf{X}_1$, controlling for other $\mathbf{X}$.

- Step 1 purges the effects of other X's on X_1
- Step 2 purges the effects of other X's on Y
- Step 3 estimates the effect of X_1 on Y using what is left over after removing the effect of other X's.

5.4 An example of the FWL theorem

Consider a regression model with single regressor

$$Y_i = \beta_0 + \beta_1 X_i + u_i,\ i = 1, \ldots, n$$

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_{1i} on 1. Then the residuals from these two regressions are $x_{1i} = X_{1i} - \bar{X}_1$.
3. Regress y_i on x_{1i} without intercept. That is, estimate the model

   $$y_i = \beta_1 x_{1i} + v_i$$

Then the OLS estimate of β_1 in the reduced model is the same as that in the original model.

We can obtain $\hat{\beta}_1$ directly by applyin the formula in Equation (7). That is

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

6 The Least Squares Assumptions in Multiple Regression

6.1 The least squares assumptions

We introduce four least squares assumptions for a multiple regression model. The first three assumptions directly follows those in the simple regression model with minor modification to allow for multiple regressors. The fourth assumption is new.

Assumption #1 $E(u_i|\mathbf{X}_i) = 0$. 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 $(Y_i, \mathbf{X}_i')\, i = 1, \ldots, n$ 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 (or to be perfectly multicollinear) if one of the regressor is a perfect linear function of the other regressors.

7 The Statistical Properties of the OLS Estimators in Multiple Regression

7.1 Unbiasedness and consistency

Under the least squares assumptions the OLS estimator $\hat{\boldsymbol{\beta}}$ can be shown to be **unbiased** and **consistent** estimator of $\boldsymbol{\beta}$ in the multiple regression model of Equation (4).

Unbiasedness

The OLS estimators $\hat{\boldsymbol{\beta}}$ is unbiased if $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$.

To show the unbiasedness, we can rewrite $\hat{\boldsymbol{\beta}}$ as follows, LaTeX

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \mathbf{u}) = \boldsymbol{\beta} + \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{u} \tag{17}$$

Thus, the conditional expectation of $\hat{\boldsymbol{\beta}}$ is LaTeX

$$E(\hat{\boldsymbol{\beta}}|\mathbf{X}) = \boldsymbol{\beta} + \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'E(\mathbf{u}|\mathbf{X}) = \boldsymbol{\beta} \tag{18}$$

in which $E(\mathbf{u}|\mathbf{X}) = 0$ from the first least squares assumption.

Using the law of iterated expectation, we have LaTeX

$$E(\hat{\boldsymbol{\beta}}) = E(E(\hat{\boldsymbol{\beta}}|\mathbf{X})) = E(\boldsymbol{\beta}) = \boldsymbol{\beta}$$

Therefore, $\hat{\boldsymbol{\beta}}$ is an unbiased estimator of $\boldsymbol{\beta}$.

Consistency

The OLS estimator $\hat{\boldsymbol{\beta}}$ is consistent if as $n \to \infty$, $\hat{\boldsymbol{\beta}}$ will converge to $\boldsymbol{\beta}$ in probability, that is, $\text{plim}_{n\to\infty}\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}$.

From Equation (17), we can have

$$\operatorname*{plim}_{n\to\infty}\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \operatorname*{plim}_{n\to\infty}\left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1}\operatorname*{plim}_{n\to\infty}\left(\frac{\mathbf{X}'\mathbf{u}}{n}\right)$$

Let us first make an assumption, which is usually true, that

$$\operatorname*{plim}_{n\to\infty}\frac{1}{n}\mathbf{X}'\mathbf{X} = \underset{(k+1)\times(k+1)}{\mathbf{Q}_{\mathbf{X}}} \tag{19}$$

which means as n goes to very large, $\mathbf{X}'\mathbf{X}$ converge to a nonstochastic matrix $\mathbf{Q}_{\mathbf{X}}$ with full rank $(k+1)$. In Chapter 18, we will see that $\mathbf{Q}_{\mathbf{X}} = E(\mathbf{X}_i\mathbf{X}_i')$ where $\mathbf{X}_i = [1, X_{1i}, \ldots, X_{ki}]'$ is the i^{th} row of $\mathbf{X}$.

Now let us look at $\text{plim}_{n\to\infty}\frac{1}{n}\mathbf{X}'\mathbf{u}$ which can be rewritten as

$$\operatorname*{plim}_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i u_i$$

Since $E(u_i|\mathbf{X}_i) = 0$, we know that $E(\mathbf{X}_i u_i) = E(\mathbf{X}_i E(u_i|\mathbf{X}_i)) = 0$. Also, by Assumptions #2 and #3, we know that $\mathbf{X}_i u_i$ are i.i.d. and have positive finite variance. Thus, by the

law of large number

$$\operatorname*{plim}_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i u_i = E(\mathbf{X}_i u_i) = 0$$

Therefore, we can conclude that

$$\operatorname*{plim}_{n\to\infty} \hat{\boldsymbol{\beta}} = \boldsymbol{\beta}$$

That is, $\hat{\boldsymbol{\beta}}$ is consistent.

7.2 The Gauss-Markov theorem and efficiency

The Gauss-Markov conditions

The Gauss-Markov conditions for multiple regression are

1. $E(\mathbf{u}|\mathbf{X}) = 0$,
2. $\text{Var}(\mathbf{u}|\mathbf{X}) = E(\mathbf{u}\mathbf{u}'|\mathbf{X}) = \sigma_u^2 \mathbf{I}_n$ (homoskedasticity),
3. $\mathbf{X}$ has full column rank (no perfect multicollinearity).

Understanding the Gauss-Markov conditions

Like in the regression model with single regressor, the least squares assumptions can be summarized by the Gauss-Markov conditions as

- Assumptions #1 and #2 imply that $E(\mathbf{u}|\mathbf{X}) = \mathbf{0}_n$.

  $$E(u_i|\mathbf{X}) = E(u_i|[\mathbf{X_1}, \ldots, \mathbf{X}_i, \ldots, \mathbf{X}_n]') = E(u_i|\mathbf{X}_i) = 0$$

  in which the second equality follows Assumption #2 that $\mathbf{X}_i$, for $i = 1, \ldots, n$ are independent.

- Assumption #1, #2, and the additional assumption of homoskedasticity imply that $\text{Var}(\mathbf{u}|\mathbf{X}) = \sigma_u^2 \mathbf{I}_n$.

  For a random vector $\mathbf{x}$, the variance of $\mathbf{x}$ is a covariance matrix defined as LaTeX

  $$\text{Var}(\mathbf{x}) = E\left((\mathbf{x} - E(\mathbf{x}))(\mathbf{x} - E(\mathbf{x}))'\right)$$

  which also holds for the conditional variance by replacing the expectation operator with the conditional expectation operator.

Since $E(\mathbf{u}|\mathbf{X}) = 0$, its covariance matrix, conditioned on $\mathbf{X}$, is LaTeX

$$\text{Var}(\mathbf{u}|\mathbf{X}) = E(\mathbf{u}\mathbf{u}'|\mathbf{X})$$

where LaTeX

$$\mathbf{u}\mathbf{u}' = \begin{pmatrix} u_1^2 & u_1u_2 & \cdots & u_1u_n \\ u_2u_1 & u_2^2 & \cdots & u_2u_n \\ \vdots & \vdots & \ddots & \vdots \\ u_nu_1 & u_nu_2 & \cdots & u_n^2 \end{pmatrix}$$

Thus, in the matrix $\mathbf{u}\mathbf{u}'$,

- the expectation of the diagonal elements, conditioned on $\mathbf{X}$, are the conditional variance of u_i which is σ_u^2 because of homoskedasticity.
- The conditional expectation of the off-diagonal elements are the covariance of u_i and u_j, conditioned on $\mathbf{X}$. Since u_i and u_j are independent according to Assumption #2, $E(u_iu_j|\mathbf{X}) = 0$.

Therefore, the conditional covariance matrix of $\mathbf{u}$ is LaTeX

$$\text{Var}(\mathbf{u}|\mathbf{X}) = \begin{pmatrix} \sigma_u^2 & 0 & \cdots & 0 \\ 0 & \sigma_u^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_u^2 \end{pmatrix} = \sigma_u^2\mathbf{I}_n$$

The Gauss-Markov Theorem

If the Gauss-Markov conditions hold in the multiple regression model, then the OLS estimator $\hat{\boldsymbol{\beta}}$ is more efficient than any other linear unbiased estimator $\tilde{\boldsymbol{\beta}}$ in the sense that $\text{Var}(\tilde{\boldsymbol{\beta}}) - \text{Var}(\hat{\boldsymbol{\beta}})$ is a positive semidefinite matrix. That is, the OLS estimator is BLUE.

That $\text{Var}(\tilde{\boldsymbol{\beta}}) - \text{Var}(\hat{\boldsymbol{\beta}})$ is a positive semidefinite matrix means that for any nonzero $(k+1) \times 1$ vector $\mathbf{c}$, LaTeX

$$\mathbf{c}' \left(\text{Var}(\tilde{\boldsymbol{\beta}}) - \text{Var}(\hat{\boldsymbol{\beta}}) \right) \mathbf{c} \geq 0$$

or we can simply write as LaTeX

$$\text{Var}(\tilde{\boldsymbol{\beta}}) \geq \text{Var}(\hat{\boldsymbol{\beta}})$$

The equality holds only when $\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}$.[2]

[2]The complete proof of the Gauss-Markov theorem in multiple regression is in Appendix 18.5.

Linear conditionally unbiased estimators

Any linear estimator of $\boldsymbol{\beta}$ can be written as LaTeX

$$\tilde{\boldsymbol{\beta}} = \mathbf{A}\mathbf{y} = \mathbf{A}\mathbf{X}\boldsymbol{\beta} + \mathbf{A}\mathbf{u}$$

where $\mathbf{A}$ is a weight matrix depending only on $\mathbf{X}$ not on $\mathbf{y}$.

For $\tilde{\boldsymbol{\beta}}$ to be conditionally unbiased, we must have LaTeX

$$E(\tilde{\boldsymbol{\beta}}|\mathbf{X}) = \mathbf{A}\mathbf{X}\boldsymbol{\beta} + \mathbf{A}E(\mathbf{u}|\mathbf{X}) = \boldsymbol{\beta}$$

which only holds when $\mathbf{A}\mathbf{X} = \mathbf{I}_{k+1}$ and the first Gauss-Markov condition holds.

The OLS estimator $\hat{\boldsymbol{\beta}}$ is a linear conditionally unbiased estimator with $\mathbf{A} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. Obviously, $\mathbf{A}\mathbf{X} = \mathbf{I}_{k+1}$ is true for $\hat{\boldsymbol{\beta}}$.

The conditional covariance matrix of $\hat{\boldsymbol{\beta}}$

The conditional variance matrix of $\hat{\boldsymbol{\beta}}$ can be derived as follows LaTeX

$$\begin{aligned}
\text{Var}(\hat{\boldsymbol{\beta}}|\mathbf{X}) &= E\left[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'|\mathbf{X}\right] \\
&= E\left[\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{u}\left(\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{u}\right)'|\mathbf{X}\right] \\
&= E\left[\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{u}\mathbf{u}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}|\mathbf{X}\right] \\
&= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'E(\mathbf{u}\mathbf{u}'|\mathbf{X})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}
\end{aligned}$$

Then, by the second Gauss-Markov condition, we have LaTeX

$$\text{Var}(\hat{\boldsymbol{\beta}}|\mathbf{X}) = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'(\sigma_u^2\mathbf{I}_n)\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = \sigma_u^2(\mathbf{X}'\mathbf{X})^{-1}$$

The **homoskedasticity-only** covariance matrix of $\hat{\boldsymbol{\beta}}$ is LaTeX

$$\text{Var}(\hat{\boldsymbol{\beta}}|\mathbf{X}) = \sigma_u^2(\mathbf{X}'\mathbf{X})^{-1} \tag{20}$$

If the homoskedasticity assumption does not hold, denote the covariance matrix of $\mathbf{u}$ as

$$\text{Var}(\mathbf{u}|\mathbf{X}) = \boldsymbol{\Omega}$$

Heteroskedasticity means that the diagonal elements of $\mathbf{\Omega}$ can be different (i.e. $\text{Var}(u_i|\mathbf{X}) = \sigma_i^2$ for $i = 1, \ldots, n$), while the off-diagonal elements are zeros, that is LaTeX

$$\mathbf{\Omega} = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{pmatrix}$$

Define $\mathbf{\Sigma} = \mathbf{X}'\mathbf{\Omega}\mathbf{X}$. Then the **heteroskedasticity-robust covariance matrix** of $\hat{\boldsymbol{\beta}}$ is LaTeX

$$\text{Var}_{\text{h}}(\hat{\boldsymbol{\beta}}|\mathbf{X}) = \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{\Sigma} (\mathbf{X}'\mathbf{X})^{-1} \tag{21}$$

7.3 The asymptotic normal distribution

In large samples, the OLS estimator $\hat{\boldsymbol{\beta}}$ has the multivariate normal asymptotic distribution as

$$\hat{\boldsymbol{\beta}} \xrightarrow{d} N(\boldsymbol{\beta}, \mathbf{\Sigma}_{\hat{\boldsymbol{\beta}}}) \tag{22}$$

where $\mathbf{\Sigma}_{\hat{\boldsymbol{\beta}}} = \text{Var}(\hat{\boldsymbol{\beta}}|\mathbf{X})$ for which use Equation (20) for the homoskedastic case and Equation (21) for the heteroskedastic case.

The proof of the asymptotic normal distribution and the multivariate central limit theorem are given in Chapter 18.

8 The Omitted Variable Bias

8.1 The definition of the omitted variable bias

The **omitted variable bias** is the bias in the OLS esitmator that arises when the included regressors, $\mathbf{X}$, are correlated with omitted variables, $\mathbf{Z}$, where $\mathbf{X}$ may include k regressors, $\mathbf{X}_1, \ldots, \mathbf{X}_k$, and $\mathbf{Z}$ may include l omitted variables, $\mathbf{Z}_1, \ldots, \mathbf{Z}_m$. The omitted variable bias occurs when two conditions are met

1. $\mathbf{X}$ is correlated with some omitted variables in $\mathbf{Z}$.
2. The omitted variables are determinants of the dependent variable $\mathbf{Y}$.

8.2 The reason for the omitted variable bias

Suppose that the true model is

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \mathbf{u} \tag{23}$$

in which the first least squares assumption, $E(\mathbf{u}|\mathbf{X}, \mathbf{Z}) = 0$, holds. We further assume that $\text{Cov}(\mathbf{X}, \mathbf{Z}) \neq 0$

However, we mistakenly exclude $\mathbf{Z}$ in regression analysis and estimate a short model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{24}$$

Since $\boldsymbol{\epsilon}$ represents all other factors that are not in Equation (24), including $\mathbf{Z}$, and $\text{Cov}(\mathbf{X}, \mathbf{Z}) \neq 0$, this means that $\text{Cov}(\mathbf{X}, \boldsymbol{\epsilon}) \neq 0$, which implies that $E(\boldsymbol{\epsilon}|\mathbf{X}) \neq 0$. (Recall that in Chapter 4, we prove that $E(u_i|X_i) = 0 \Rightarrow \text{Cov}(u_i, X_i) = 0$, which implies that $\text{Cov}(u_i, X_i) \neq 0 \Rightarrow E(u_i|X_i) \neq 0$.) Therefore, Assumption #1 does not hold for the short model, which means that the OLS estimator of Equation (24) is biased.

An informal proof of the OLS estimator of Equation (24) is biased is given as follows.

The OLS estimator of Equation (24) is $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$. Plugging $\mathbf{Y}$ with the true model, we have

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \mathbf{u}) = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}\boldsymbol{\gamma} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}$$

Taking the expectation of $\hat{\boldsymbol{\beta}}$, conditioned on $\mathbf{X}$, we have

$$E(\hat{\boldsymbol{\beta}}|\mathbf{X}) = \boldsymbol{\beta} + \underbrace{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}\boldsymbol{\gamma}}_{\text{omitted variable bias}} + \mathbf{0} \tag{25}$$

The second term in the equation above usually does not equal zero unless either

1. $\boldsymbol{\gamma} = \mathbf{0}$, which means that $\mathbf{Z}$ are not determinants of $\mathbf{Y}$ in the true model, or
2. $\mathbf{X}'\mathbf{Z} = 0$, which means that $\mathbf{X}$ and $\mathbf{Z}$ are not correlated.

Therefore, if these two conditions do not hold, $\hat{\boldsymbol{\beta}}$ for the short model is biased. And the magnitude and direction of the bias is determined by $\mathbf{X}'\mathbf{Z}\boldsymbol{\gamma}$.

8.3 An illustration using a linear model with two regressors

Suppose the true model is

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

with $E(u_i|X_{1i}, X_{2i}) = 0$

However, we estimate a wrong model of

$$Y_i = \beta_0 + \beta_1 X_{1i} + \epsilon_i,\ i = 1, \ldots, n$$

In Lecture 5 we showed that β_1 can be expressed as

$$\hat{\beta}_1 = \beta_1 + \frac{\frac{1}{n}\sum_i (X_{1i} - \bar{X}_1)\epsilon_i}{\frac{1}{n}\sum_i (X_i - \bar{X}_1)^2}$$

As $n \to \infty$, the numerator of the second term converges to $\text{Cov}(X_1, \epsilon) = \rho_{X_1\epsilon}\sigma_{X_1}\sigma_\epsilon$ and the denominator converges to $\sigma^2_{X_1}$, where $\rho_{X_1\epsilon}$ is the correlation coefficient between X_{1i} and ϵ. Therefore, we have

$$\hat{\beta}_1 \xrightarrow{p} \beta_1 + \underbrace{\rho_{X_1\epsilon}\frac{\sigma_\epsilon}{\sigma_{X_1}}}_{\text{omitted variable bias}} \tag{26}$$

From Equations (25) and (26), we can summarize some facts about the omitted variable bias:

- Omitt variable bias is a problem irregardless of whether the sample size is large or small. $\hat{\beta}$ is biased and inconsistent when there is omitted variable bias.
- Whether this bias is large or small in practice depends on $|\rho_{X_1\epsilon}|$ or $|\mathbf{X'Z}\boldsymbol{\gamma}|$.
- The direction of this bias is determined by the sign of $\rho_{X_1\epsilon}$ or $\mathbf{X'Z}\boldsymbol{\gamma}$.
- One easy way to detect the existence of the omitted variable bias is that when adding a new regressor, the estimated coefficients on some previously included regressors change substantially.

9 Multicollinearity

9.1 Perfect multicollinearity

Perfect multicollinearity refers to the situation when one of the regressor is a perfect linear function of the other regressors.

- In the terminology of linear algebra, perfect multicollinearity means that the vectors of regressors are linearly dependent.

- That is, the vector of a regressor can be expressed as a linear combination of vectors of the other regressors.

Remember that the matrix of regressors $\mathbf{X}$ can be written in terms of column vectors as LaTeX

$$\mathbf{X} = [\boldsymbol{\iota}, \boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_k]$$

where $\boldsymbol{X}_i = [X_{i1}, X_{i2}, \ldots, X_{in}]'$ is a $n \times 1$ vector of n observations of the i^{th} regressor. $\boldsymbol{\iota}$ is a vector of 1s, representing the constant term.

That the $k+1$ column vectors are linearly dependent means that there exist some $(k+1)\times 1$ nonzero vector $\boldsymbol{\beta} = [\beta_0, \beta_1, \ldots, \beta_k]'$ such that LaTeX

$$\beta_0\boldsymbol{\iota} + \beta_1\boldsymbol{X}_1 + \cdots + \beta_k\boldsymbol{X}_k = 0$$

If $\boldsymbol{X}_i$, for $i = 1, \ldots, n$, are linearly dependent, then it follows

- $\mathbf{X}$ does not have full column rank.
- If $\mathbf{X}$ does not have full column rank, then $\mathbf{X}'\mathbf{X}$ is singular, that is, the inverse of $\mathbf{X}'\mathbf{X}$ does not exist. Therefore, we can state the assumption of requiring no perfect multicollinearity in another way as assuming that $\mathbf{X}$ has full column rank.
- If $\mathbf{X}'\mathbf{X}$ is not invertible, the OLS estimator based on the formula of $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ does not exist.

9.2 Examples of perfect multicollinearity

Remember that perfect multicollinearity occurs when one regressor can be expressed as a linear combination of other regressors. This problem belongs to the logic error when the researcher sets up the regression model. That is, the researcher uses some redundant regressors in the model to provide the same information that merely one regressor can sufficiently provide.

Possible linear combination

Suppose we have a multiple regression model

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

And we want to add a new variable Z into this model. The following practices cause perfect multicollinearity

- $Z = aX_1$ or $Z = bX_2$

- $Z = 1 - aX_1$
- $Z = aX_1 + bX_2$

However, we can add a Z that is not a linear function of X_1 or X_2 such that there is no perfect multicollinearity problem. For example,

- $Z = X_1^2$
- $Z = \ln X_1$
- $Z = X_1 X_2$

9.3 The dummy variable trap

The dummy variable trap is a good case of perfect multicollinearity that a modeler often encounters. Recall that a **binary variable** (or **dummy variable**) D_i, taking values of one or zero, can be used in a regression model to distinguish two mutually exclusive groups of samples, for instance, the male and the female. In fact, dummy variables can be constructed to represent more than two groups and be used in multiple regression to examine the difference between these groups.

Suppose that we have a data composed of people of four ethnic groups: White, African American, Hispanic, and Asian. And we want to estimate a regression model to see whether wages among these four groups are different. We may (mistakenly as we will see) set up a multiple regression model as follows LaTeX

$$Wage_i = \beta_0 + \beta_1 White_i + \beta_2 African_i + \beta_3 Hispanic_i + \beta_4 Asian_i + u_i \qquad (27)$$

where $White_i$ is a dummy variable which equal 1 if the i^{th} observation is a white people and equal 0 if he/she is not, similarly for $African_i, Hispanic_i,$ and $Asian_i$.

A concrete example

To be concrete, suppose we have four observations: Chuck, Mike, Juan, and Li, who are White, African American, Hispanic, and Asian, respectively. Then the dummy variables are

$$White = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}, African = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}, Hispanic = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix}, Asian = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}$$

However, when we construct a model like Equation (27), we fall into the dummy variable trap, suffering perfect multicollinearity. This is because this model has a constant term

$\beta_0 \times 1$ which is the sum of all dummy variables. That is,

$$\begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} = White + African + Hispanic + Asian$$

Let see when the observation is Chuck, then the model is

$$Wage = \beta_0 + \beta_1 + u$$

Estimating this model yields $\widehat{\beta_0 + \beta_1}$, from which we cannot get a unique solution for β_1.

To avoid the dummy variable trap, we can either of the following two methods:

1. drop the constant term
2. drop one dummy variable

The difference between these two methods lies in how we interpret the coefficients on dummy variables.

Drop the constant term

If we drop the constant term, the model becomes

$$Wage = \beta_1 White + \beta_2 African + \beta_3 Hispanic + \beta_4 Asian + u \tag{28}$$

For Chuck or all white people, the model becomes

$$Wage = \beta_1 + u$$

Then β_1 is the population mean wage of whites, that is, $\beta_1 = E(Wage|White = 1)$. Similarly, $\beta_2, \beta_3,$ and β_4 are the population mean wage of African Americans, Hispanics, and Asians, respectively.

Drop one dummy variable

If we drop the dummy variable for white people, then the model becomes

$$Wage = \beta_1 + \beta_2 African + \beta_3 Hispanic + \beta_4 Asian + u \tag{29}$$

For white people, the model is

$$Wage = \beta_1 + u_i$$

And the constant term β_1 is just the population mean of whites, that is,

$$\beta_1 = E(Wage|White = 1)$$

So we say that white people serve as a reference case in Model (29).

For African Americans, the model is

$$Wage = \beta_1 + \beta_2 + u$$

From it we have $E(Wage|African = 1) = \beta_1 + \beta_2$ so that

$$\beta_2 = E(Wage|African = 1) - \beta_1 = E(Wage|African = 1) - E(Wage|White = 1)$$

Similarly, we can get that

$$\beta_3 = E(Wage|Hispanic = 1) - E(Wage|White = 1)$$
$$\beta_4 = E(Wage|Asian = 1) - E(Wage|White = 1)$$

Therefore, when we adopt the second method by dropping a dummy variable for the reference case, then the coefficients on other dummy variables represent the difference in the population means between the interested case and the reference case.

9.4 Imperfect Multicollinearity

Definition of imperfect multicollinearity

Imperfect multicollinearity is a problem of regression when two or more regressors are highly correlated. Although they bear similar names, imperfect multicollinearity and perfect multicollinearity are two different concepts.

- Perfect multicollinearity is a problem of modeling building, resulting in a total failure to estimate a linear model.
- Imperfect multicollinearity is usually a problem of data when some regressors are highly correlated.
- Imperfect multicollinearity does not affect the unbiasedness of the OLS estimators. However, it does affect the efficiency, i.e., the variance of the OLS estimators.

An illustration using a regression model with two regressors

Suppose we have a linear regression model with two regressors.

$$\mathbf{Y} = \beta_0 + \beta_1 \mathbf{X}_1 + \beta_2 \mathbf{X}_2 + \mathbf{u} \tag{30}$$

where, for simplicity, $\mathbf{u}$ is assumed to be homoskedastic.

By the FWL theorem, estimating Equation (30) will get the same OLS estimators of β_1 and β_2 as estimating the following model,

$$\mathbf{y} = \beta_1 \mathbf{x}_1 + \beta_2 \mathbf{x}_2 + \mathbf{v} \tag{31}$$

where $\mathbf{y} = \mathbf{Y} - \bar{Y}\boldsymbol{\iota}$, $\mathbf{x}_1 = \mathbf{X}_1 - \bar{X}_1\boldsymbol{\iota}$, and $\mathbf{x}_2 = \mathbf{X}_2 - \bar{X}_2\boldsymbol{\iota}$, that is, $\mathbf{y}, \mathbf{x}_1,$ and $\mathbf{x}_2$ are in the form of the deviation from the mean. And denote $\mathbf{x} = [\mathbf{x}_1 \ \mathbf{x}_2]$ as the matrix of all regressors in Model (31).

Suppose that X_1 and X_2 are correlated so that their correlation coefficient $|\rho_{12}| > 0$. And the square of the sample correlation coefficient is

$$r_{12}^2 = \frac{\left(\sum (X_1 - \bar{X}_1)(X_2 - \bar{X}_2)\right)^2}{\sum (X_1 - \bar{X}_1)^2 \sum (X_2 - \bar{X}_2)^2} = \frac{\left(\sum x_1 x_2\right)^2}{\sum x_1 \sum x_2} \tag{32}$$

The OLS estimator of Model (31) is

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{x}'\mathbf{x}\right)^{-1} \mathbf{x}'\mathbf{y} \tag{33}$$

with the homoskedasticity-only covariance matrix as

$$\text{Var}(\hat{\boldsymbol{\beta}}|\mathbf{x}) = \sigma_u^2 \left(\mathbf{x}'\mathbf{x}\right)^{-1} \tag{34}$$

- $\hat{\boldsymbol{\beta}}$ is still unbiased since the assumption of $E(\mathbf{u}|X) = 0$ holds and so does $E(\mathbf{v}|\mathbf{x}) = 0$.
- The variance of $\hat{\beta}_1$, which is the first diagonal element of $\sigma_u^2 \left(\mathbf{x}'\mathbf{x}\right)^{-1}$, is affected by r_{12}. To see this, we write $\text{Var}(\hat{\beta}_1|\mathbf{x})$ explicitly as

$$\begin{aligned}
\text{Var}(\hat{\beta}_1|\mathbf{x}) &= \frac{\sigma_u^2 \sum_i x_2^2}{\sum_i x_1^2 \sum_i x_2^2 - (\sum_i x_1 x_2)^2} \\
&= \frac{\sigma_u^2 \sum_i x_2^2}{\sum_i x_1^2 \sum_i x_2^2 \left(1 - \dfrac{(\sum_i x_1 x_2)^2}{\sum_i x_1^2 \sum_i x_2^2}\right)} \\
&= \frac{\sigma_u^2}{\sum_i x_1^2} \frac{1}{(1 - r_{12}^2)}
\end{aligned}$$

Therefore, when X_1 and X_2 are highly correlated, that is r_{12}^2 gets close to 1, then

$\text{Var}(\hat{\beta}_1|\mathbf{x})$ becomes very large.

- The consequence of multicollinearity is that it may lead us to wrongly fail to reject the zero hypothesis in the t-test for a coefficient.
- The variance inflation factor (VIF) is a commonly used indicator for detecting multicollinearity. The definition is

$$\text{VIF} = \frac{1}{1 - r_{12}^2}$$

The smaller VIF is for a regressor, the less severe the problem of multicollinearity is. However, there is no widely accepted cut-off value for VIF to detect multicollinearity. $VIF > 10$ for a regressor is often seen as an indication of multicollinearity, but we cannot always trust this.

Possible remedies for multicollinearity

- Include more sample in hope of the variation in $\mathbf{X}$ getting widened, i.e., increasing $\sum_i (X_{1i} - \bar{X}_1)$.
- Drop the variable(s) that is highly correlated with other regressors. Notice that by doing this we are at the risk of suffering the omitted variable bias. There is always a trade-off between including all relevant regressors and making the regression model *parsimonious*.[3]

[3]The word "parsimonious" in Econometrics means that we always want to make the model as concise as possible without any redundant variables included.