LECTURE NOTES ON INTRODUCTORY ECONOMETRICS

Panit Wattanakoon
Faculty of Economics, Thammasat University

May, 2014

Acknowledgement: I appreciate Mr.Phattaris Sittikul's excellent work on the preparation of this draft.

Contents

Chapter 1

INTRODUCTION

1.1 IMPORTANCE AND NECESSITY OF ECONOMETRIC STUDY

Economics is the social science that studies and researches into how the limited resource is efficiently allocated so as to satisfy the unlimited want of human. Economists, thus, have formed the theories to explain economic phenomena and to present the solution to economic problems that probably occur in the economic system.

Nonetheless, economic theories, able to describe the dynamic of economic behaviour only, may not be sufficient to suggest the policy dealing with specific issues. For example, the theories may be incapable of identifying how much government spending is appropriate for expansionary policy so that the economy will be stable, namely the economy will expand so properly that no financial bubble results and no additional burden has to be shouldered by the government itself. Hence, economists initiated application of statistics and mathematics to economic theory and defined this field of study as "*econometrics*".

Econometrics is the adaptation of statistical and mathematical tools for calculating and estimating the value of the variables in an economic model invented from the use of mathematical economics. Accordingly, econometrics is useful as the mean through which the economic theories can be applied to the real world phenomena.

For illustration, suppose company A selling goods X attempts to develop the profit-maximizing strategy. The data necessary to be studied is the demand for goods X itself. Generally, economic demand theory states that, as the price of goods (P_X) increases, the demand for goods (Q_X^d) would decrease. In term of mathematical economics, the model describing this situation may be written as;

$$Q_X^d = a + bP_X$$

Then, economists would gather the information about the price and quantity of goods X and, through econometrics, use that information to estimate the value of coefficient a and b such that the company A can realize the demand for goods X. This approximation of demand function can be utilized further to calculate the price-elasticity of goods X which is helpful for developing suitable pricing strategy in the future.

1.2 PROCEDURE FOR ECONOMETRIC-MODEL ANALYSIS

There are seven steps for the analysis of econometric model[1]. In this section, Keynesian theory of consumption is used as an example to illustrate how to apply econometrics to establish the regression function describing the relationship between income and consumption expenditure.

1. SELECTING THE THEORY OR HYPOTHESIS OF INTEREST

John Maynard Keynes, the English economist, stated that the consumption of people would increase as their income increases. Yet, the increase in consumption expenditure would not increase in the same amount as the increase in income. In other word, marginal propensity to consume, the proportion of consumption expenditure per 1 Baht increase in income, would be greater than 0 but lower than 1.

2. IDENTIFYING MATHEMATICAL MODEL OF THE THEORY

The above theory portrays the relationship between consumption expenditure and income; however, neither the rate of change nor the mathematical model is specified. In term of mathematics, the model representing that theory may be written as;

$$C = a + bY \text{ and } 0 < b < 1 \tag{1.1}$$

where C is the consumption expenditure, Y is income, and a and b are parameters of the models which are intercept and slope coefficients respectively.

This model is generally classified as **single-equation model** and, if the model consists of more than one equation, it is called **multiple-equation model**. From the above model, marginal propensity to consume (MPC) is measured by the coefficient b. Also, describing the Keynesian theory of consumption, the model identifies the relationship between income, which is an **independent variable** for this model, and consumption expenditure, which is a **dependent variable**.

[1]Gujarati, Damodar N. and Dawn C. Porter, *Basic Econometrics*, McGraw Hill, Singapore, fifth edition, 2009, p.3

3. IDENTIFYING ECONOMETRIC MODEL OF THE THEORY

The next step is to identify the econometric model for consumption expenditure, which can be written as;

$$C = a + bY + u_t \tag{1.2}$$

where u_t is known as a disturbance or random term which is a random variable.

Normally, in basic econometrics, the parameters a and b will be estimated through the use of linear regression model, which determines the effectiveness of the consumption theory in explaining the real phenomena in economic system with the disturbance term representing other factors not included in this model.

4. GATHERING DATA

For the data of consumption and national income in Thailand, Office of the National Economic and Social Development Board (NESDB) takes responsibility as the provider. The sample data is shown in Table 1.1.

Table 1.1: The data of consumption expenditure (C) and national income (Y) at constant price (trillion bath)

Year	C	Y	Year	C	Y
1990	1.88	3.36	2001	3.01	5.42
1991	1.98	3.65	2002	3.20	5.76
1992	2.17	3.98	2003	3.44	6.17
1993	2.35	4.33	2004	3.69	6.56
1994	2.53	4.68	2005	3.85	6.84
1995	2.74	5.06	2006	3.96	7.17
1996	2.88	5.34	2007	4.00	7.57
1997	2.84	5.20	2008	4.12	7.69
1998	2.55	4.80	2009	4.06	7.62
1999	2.66	5.02	2010	4.27	8.18
2000	2.84	5.24	2011	4.34	8.21

Source: NESDB. at June 2013

5. ESTIMATING ECONOMETRIC MODEL

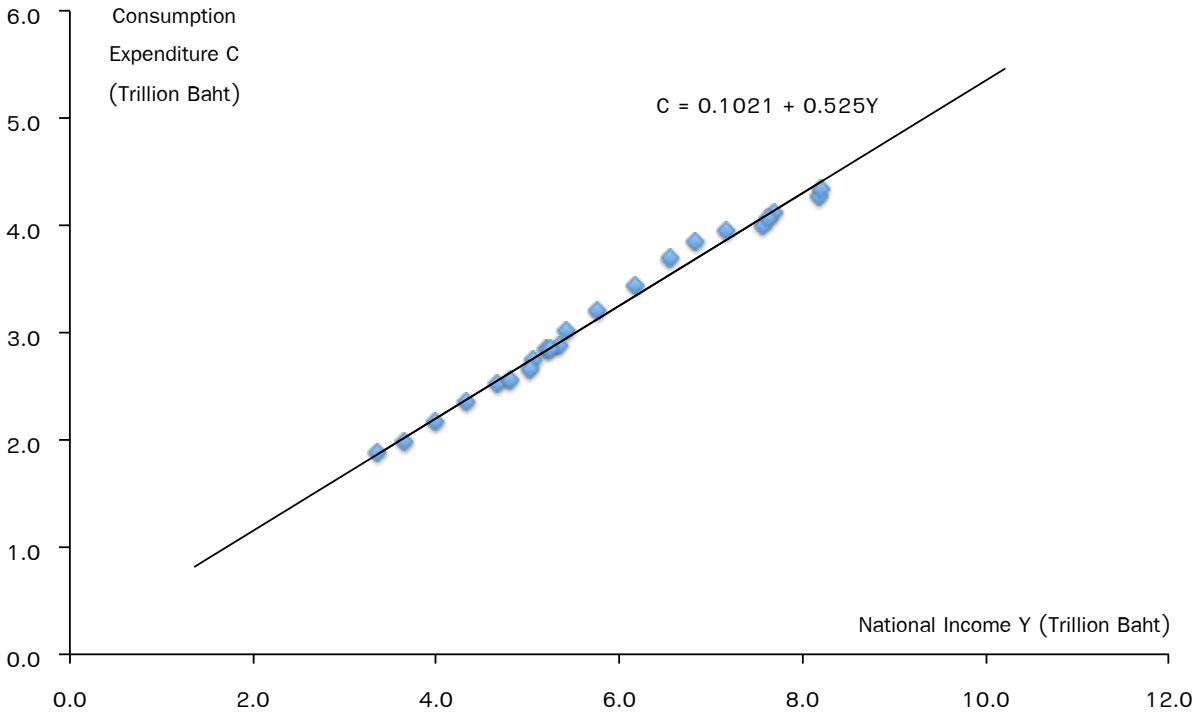
After receiving all the data, the econometric technique is used for *regression analysis* to detect the characteristics of these two groups of data and to estimate the value of parameters a and b. According to the Table 1.1., the following estimates of a and b are obtained,

$$\hat{C}_t = 0.1021 + 0.525Y_t \tag{1.3}$$

where $\hat{C}_t$ is the estimated consumption expenditure. Figure 1.1 depicts the consumption expenditure as the function of national income.

Equation (1.3) indicates that, in accordance with the data from 1990 to 2011, as the real national income increases by 1 Baht, the consumption expenditure in the system increases by, approximately, 0.525 Baht.

Figure 1-1: Consumption Expenditure and National Income from 1990 to 2011 (Trillion Baht)



6. TESTING THE HYPOTHESIS

After acquiring equation (1.3), the next step is to perform hypotheses testing of whether values of parameters a and b significantly equal 0.1021 and 0.525 respectively.

Starting with the sign and magnitude, it can be seen that the estimate of b is positive and below 1, which corresponds to the stated Keynesian theory of consumption. Furthermore, the estimate of a is positive, coinciding with the theory since, even without income, people still need to make consumption expenditure to survive. This expenditure, for instance, may be received from liquidating an asset or borrowing.

Next, the ability of independent variable (national income) to explain the dependent variable (consumption expenditure) will be examined, which will be discussed later in chapter 3 and 4. Moreover, the reliability of the model has to be tested so as to find out if the model faces any problems, which will be clarified from chapter 6 to 9, and should the model confront those problems, which remedial measures have to be taken.

7. APPLYING ECONOMETRIC MODEL TO PREDICTION OR POLICY MAKING

Suppose that the econometric model stated in the previous step passes the statistical tests, the parameters b in the model can be used to calculate a *multiplier*. As an illustration, for the closed economy without government, the multiplier is derived as,

$$Multiplier = \frac{1}{1 - MPC} \tag{1.4}$$

From the model where $MPC = 0.525$, the value of the multiplier will be 2.11 which means, as private or government sector raises the expenditure by 1 Baht, the national income will rise by 2.11 Baht. Hence, with this information, the policy would be established efficiently and effectively.

Besides, policy-maker might set the specific level of consumption expenditure, like 4.5 trillion Baht, in order to achieve some objectives such as unemployment rate reduction. Then, the government has to find the necessary national income that satisfies that level of consumption expenditure. Thanks to the model used previously, the required amount of national income is 8.4 trillion Baht. Government, accordingly, uses this number as the targeted level of national income to accomplish those objectives.

$$\begin{aligned} \hat{C}_t &= 0.1021 + 0.525Y_t \\ 4.5 &= 0.1021 + 0.525Y_t \\ Y_t &= \tfrac{4.5-0.1021}{0.525} \\ \therefore \quad Y_t &= 8.4 \end{aligned}$$

1.3 DIFFERENCE BETWEEN REGRESSION, CAUSATION AND CORRELATION

Regression analysis deals with the statistical relationship between independent and dependent variables with the hypothesis that the independent variables can explain dependent variables. Nonetheless, this relationship does not always imply the causation of one or more variables on the other. To discover the causation, only the data and statistical technique are not sufficient. As a result, economists have to rely on additional information such as economic theory and common sense.

For instance, generally, it is reasonable to believe that rice production per rai depends on or is caused by the level of rainfall. Regression analysis can be used to derive the relationship between these two variables. In the other way around, it is impossible that the level of rainfall relies on rice production; even though, the statistical relationship (in this way) can be found through regression analysis. Hence, the statistical relationship is not necessarily the same as causation and the correct regression analysis should be based on the theory; otherwise, the analysis will be performed in vain.

For correlation and regression analysis, although both techniques are used to obtain the relationship between two variables, the interpretations of each analysis are different. Correlation analysis measures the relationship of two variables in both directions. Statisticians, for example, may want to identify the relationship between weight and height of specific group of people, or the relationship between the students' scores in physics and mathematics class. On the other hand, regression analysis estimates or forecasts the relationship from one variable to the other without reverse relationship, such as to forecast the scores in physics class using the scores in mathematics class.

Another interesting difference is the fact that, for correlation analysis, both variables will be random or stochastic. For regression analysis, an independent variable will be fixed or non-stochastic, which is acquired through repeated sampling, and dependent variable will be random conditional on each value of independent variable.

1.4 CLASSIFICATIONS OF DATA FOR ECONOMETRIC ANALYSIS

Data in economics is classified into many categories. Normally, economist will use the theory together with many sets of data for evaluation and analysis to describe various situations such as how the growth rate of economy affects the change in SET index, how the oil price in Singapore influences the inflation rate of Thailand. Also, some types of data are recorded in different time intervals. As illustration, NESDB announces the value of product in Thailand quarterly and annually while SET index is recorded daily.

Due to various characteristics, the data has been classified into 4 types such that econometricians can realize the distinct property of each type of data for using in the analysis. The 4 types of data are;

1. Time series data is the data collected in every time interval such as annually, monthly, and weekly. Table 1.2 is the example of time series data.

Table 1-2:Time series data

Year	GDP	C	I	G	X - M
1990	4.2	2.0	3.0	1.0	-2.2
1991	4.6	2.1	2.1	1.4	-1.0
⋮	⋮	⋮	⋮	⋮	⋮
2012	8.4	2.3	4.0	2.0	0.1

2. Cross-sectional data is the data collected from the group of sample at the same point in time like the census of population on various issues from the sample (Table 1-3)

Table 1-3: Cross-sectional data

Sample Year 2013	Income	Education	Gender	Marital Status	Expenditure
Mr A. (1)	15,000	Bachelor's degree	Male	Married	10,000
Mrs B. (2)	35,000	Master's degree	Female	Single	30,000
Mr C. (3)	10,000	Technical Certificate	Male	Single	15,000
⋮	⋮	⋮	⋮	⋮	⋮
Mrs Z (500)	100,000	Doctoral degree	Female	Married	20,000

3. Pooled data is the data collected from the different groups of sample in each time interval (Table 1.4)

Table 1.4: Pooled data

Sample	Income	Education	Gender	Marital Status	Expenditure
Year 1990					
Mr A. (1)	10,000	Secondary education	Male	Single	2,000
Mrs B. (2)	20,000	Bachelor's degree	Female	Single	18,000
⋮	⋮	⋮	⋮	⋮	⋮
Mr D. (500)	350,000	Doctoral degree	Male	Married	400,000
Year 1991					
Mrs Z. (1)	20,000	Master's degree	Female	Single	1,000
Mrs B. (2)	22,000	Bachelor's degree	Female	Single	20,000
⋮	⋮	⋮	⋮	⋮	⋮
Mr D. (500)	350,000	Doctoral degree	Male	Married	400,000
⋮	⋮	⋮	⋮	⋮	⋮
Year 2012					
Mr E (1)	40,000	Bachelor's degree	Male	Married	35,000
Mrs F (2)	70,000	Bachelor's degree	Female	Married	40,000
⋮	⋮	⋮	⋮	⋮	⋮
Mrs G (500)	80,000	Bachelor's degree	Female	Married	100,000

4. Panel data is the data collected from the same groups of sample in each time interval (Table 1.5)

Table 1.5: Panel data

Sample	Income	Education	Gender	Marital Status	Expenditure
Year 1990					
Mr A. (1)	10,000	Secondary education	Male	Single	2,000
Mrs B. (2)	20,000	Bachelor's degree	Female	Single	18,000
⋮	⋮	⋮	⋮	⋮	⋮
Mr D. (500)	350,000	Doctoral degree	Male	Married	400,000
Year 1991					
Mr A. (1)	12,000	Secondary education	Male	Single	4,000
Mrs B. (2)	22,000	Bachelor's degree	Female	Single	20,000
⋮	⋮	⋮	⋮	⋮	⋮
Mr D. (500)	360,000	Doctoral degree	Male	Married	420,000
⋮	⋮	⋮	⋮	⋮	⋮
Year 2013					
Mr A. (1)	15,000	Bachelor's degree	Male	Single	10,000
Mrs B. (2)	35,000	Master's degree	Female	Single	30,000
⋮	⋮	⋮	⋮	⋮	⋮
Mr D. (500)	800,000	Doctoral degree	Male	Married	780,000

SUMMARY

1. Econometrics is the synergy between mathematics, statistics and economics used to verify the economic theories with the empirical evidence. Also, econometrics can be applied in various practical means.

2. The procedure of econometric analysis consists of 7 steps, starting with selecting the theory or hypothesis of interest; then, identifying the mathematical model of the theory, identifying the econometric model of the theory, gathering data, estimating economic model, testing hypothesis of whether there is statistically significant and, after passing the test, applying the model to answer economic problems.

3. Regression analysis does not necessarily imply causation. Correlation analysis is different from regression analysis in the statistical interpretation. In other words, correlation analysis considers the relationship of two variables in both directions; while regression analysis considers the relationship only in one direction, namely only from one variable to the other without reverse relationship.

4. Data can be classified into 4 categories which are time series data, cross-sectional data, pooled data and panel data.

Chapter 2

STATISTICAL REVIEW

To be prepared for econometric study, your necessary statistical knowledge will be refreshed throughout this chapter. Firstly, the definition and useful properties of summation are brought up. Then, the concept of probability and random variable are discussed, followed with the various probability distributions. Finally, we talk about the concept of estimator and the properties required for a good estimator.

2.1 DEFINITION AND PROPERTIES OF SUMMATION

The notation $\sum$ (sigma), in mathematical term, denotes the **summation** of a variable X from the first X (the number assigned below $\sum$) to the $n^{th}X$ (the number assigned above $\sum$). In term of equation, thus,

$$\sum_{i=1}^{n} X_i = X_1 + X_2 + \ + X_n \tag{2.1}$$

The noteworthy properties of summation include,

1. $\sum_{i=1}^{n} k = nk$

2. $\sum_{i=1}^{n} kX_i = k\sum_{i=1}^{n} X_i$

3. $\sum_{i=1}^{n}(a + bX_i) = na + b\sum_{i=1}^{n} X_i$

4. $\sum_{i=1}^{n}(X_i + Y_i) = \sum_{i=1}^{n} X_i + \sum_{i=1}^{n} Y_i$

where a, b and k are constant.

Multiple summation is the summation of variable that is in the form of matrix, shown as,

$$\sum_{i=1}^{n}\sum_{j=1}^{m} X_{ij} = X_{11} + X_{12} + \ + X_{1m} + X_{21} + X_{22} + \ + X_{2m} + ... + X_{nm} \tag{2.2}$$

where

$$\mathbf{X} = \begin{bmatrix} X_{11} & X_{12} & \dots & X_{1m} \\ X_{21} & X_{22} & \dots & X_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \dots & X_{nm} \end{bmatrix}_{nXm}$$

The significant properties of multiple summations include,

1. $\sum_{i=1}^{n}\sum_{j=1}^{m} X_{ij} = \sum_{j=1}^{m}\sum_{i=1}^{n} X_{ij}$

2. $\sum_{i=1}^{n}\sum_{j=1}^{m} X_i Y_j = \sum_{i=1}^{n} X_i \times \sum_{j=1}^{n} Y_j$

3. $\sum_{i=1}^{n}\sum_{j=1}^{m}(X_{ij} + Y_{ij}) = \sum_{i=1}^{n}\sum_{j=1}^{m} X_{ij} + \sum_{i=1}^{n}\sum_{j=1}^{m} Y_{ij}$

4. $(\sum_{i=1}^{n} X_i)^2 = \sum_{i=1}^{n} X_i^2 + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} X_i X_j$

2.2 SAMPLE SPACE, SAMPLE POINT AND EVENTS

Sample space is the set of all possible results of an experiment. For example, if you toss the coin twice, all feasible outcomes are composed of head twice, head followed by tail, tail followed by head, and tail twice. Let H denotes head and T denotes tail. The sample space can be written as,

$$SS = \{HH, HT, TH, TT\}$$

Sample Point is the member of sample space, eg. the event that head occurs twice from tossing a coin twice. Specifically, sample point is,

$$SP = HH \text{ or } HT \text{ or } TH \text{ or } TT$$

Events are the set of some consequences of the experiment such as the events that head occurs twice. Specifically, events are the subset of sample space.

$$A = \text{the event that head occurs twice} = \{HH\}$$

Events are **mutually exclusive**, if the occurrence of one event makes no other events in sample space possible. As an illustration, for the experiment of tossing two coins once, let C be the event that both turn head and D be the event that both turn tail. Since C and D cannot happen at the same time, these two events are said to be mutually exclusive. Another example is the experiment of drawing one card from the standard 52-card deck, let E be the event that the rank of card is King and F be the event that suit of card is Clubs. As the event E and F can occur simultaneously, namely the King of Clubs, the two events are not mutually exclusive.

Events are **collectively exhaustive** if they cover all possible outcomes in the sample space. With the experiment of tossing the coin twice, let A be the event that head appears twice, B be the event that tail appears twice, and C be the event that head and tail each appear once. In this case, A, B and C are collectively exhaustive since all events cover all possible results from sample space; that is, HH, HT, TH and TT.

2.3 PROBABILITY AND RANDOM VARIABLE

Probability is the possibility that any event will occur, given some specific sample space.

Let A be the event occurring in the given sample space and $P(A)$ be the probability that A will happen. Then, $P(A)$ is defined as;

$$P(A) = \frac{\text{times the event A will occur}}{\text{the amount of all possible outcomes in sample space}} \tag{2.3}$$

For instance, to draw one card from the standard 52-card deck, let A be the event that the rank of card is 2. Times the event will occur is 4 and the amount of all possible outcomes is 52; hence, the probability of A is $\frac{4}{52}$ or $\frac{1}{13}$.

Some properties of probability are;

1. $0 \leq P(A) \leq 1$

2. If A, B and C are composed to be exhaustive set, then,

$$P(A) + P(B) + P(C) = 1$$

3. If A, B and C are mutually exclusive, then,

$$P(A + B + C) = P(A) + P(B) + P(C)$$

BRAIN-TEASER

"Let's Make a Deal" or "Monty Hall Problem"

An American television-show lets the guests choose one from the three panels, one of which the Prize, like car and television set, is behind; and the other of which the Zonks (booby prize) are behind.

Once the guests choose the panel, the host of the show will turn one of the Zonk-panels and ask whether the guest would like the choose panels again from the two left.

The question is *"Would the decision to reselect the panels enhance the probability of winning the price?"*

Random Variable Suppose that the results of an experiment are in the form of value, the variable, whose value is determined by one of those results, is known as random variable. Random variable can be either **discrete** or **continuous value**.

For discrete random variable, the example is the sum of the values on the face of two dice, when rolling two dice once. In other word, the obtained sum will range from 2 to 12, and it is impossible to get 2.5 or 3.5.

For continuous random variable, the example is the height of the high-school student, constricted to the range from 160 to 180 centimetres. It can be seen that the value of the height need not be the integers and can take the value of 160.5 or 160.52 centimetres.

These two distinct characteristics of random variable enable us to classify them into different probability density functions, which would be stated in Section 2.4.

2.4 PROBABILITY DENSITY FUNCTION

As the value of random variable depends on an experiment, the **probability density function** would portray the overall image of possible random results. The type of the probability density function relies on the characteristic of the random variable. In this section, many important types are discussed.

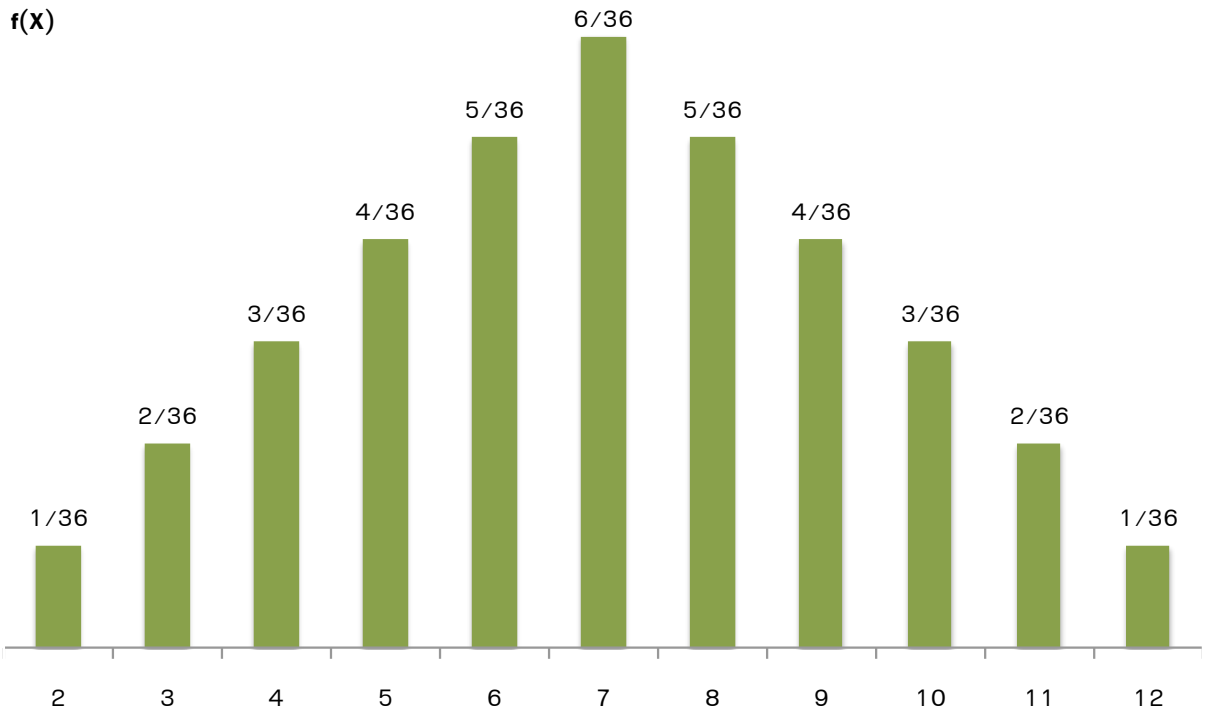
2.4.1 Probability Density Function for Discrete Random Variable

Let X be the discrete random variable with the value $X1, X2, , Xn$ and we get,

$$\begin{array}{rclcrcl} f(X) & = & P(X = X_i) & \text{for} & i & = & 1, 2, ..., n \\ f(X) & = & 0 & \text{for} & X & \neq & X_i \end{array}$$

Example: Let X be random variable of the sum of values on the face of two dices. The value might be 2, that is the value from both rolling round is 1, or 12, that is the value from both rolling round is 6. The Figure 2-1 summarizes all possible results#

Figure 2-1: Probability Density function of the Sum of Values on the Side of the Dice, Obtained from Rolling the Dice Twice



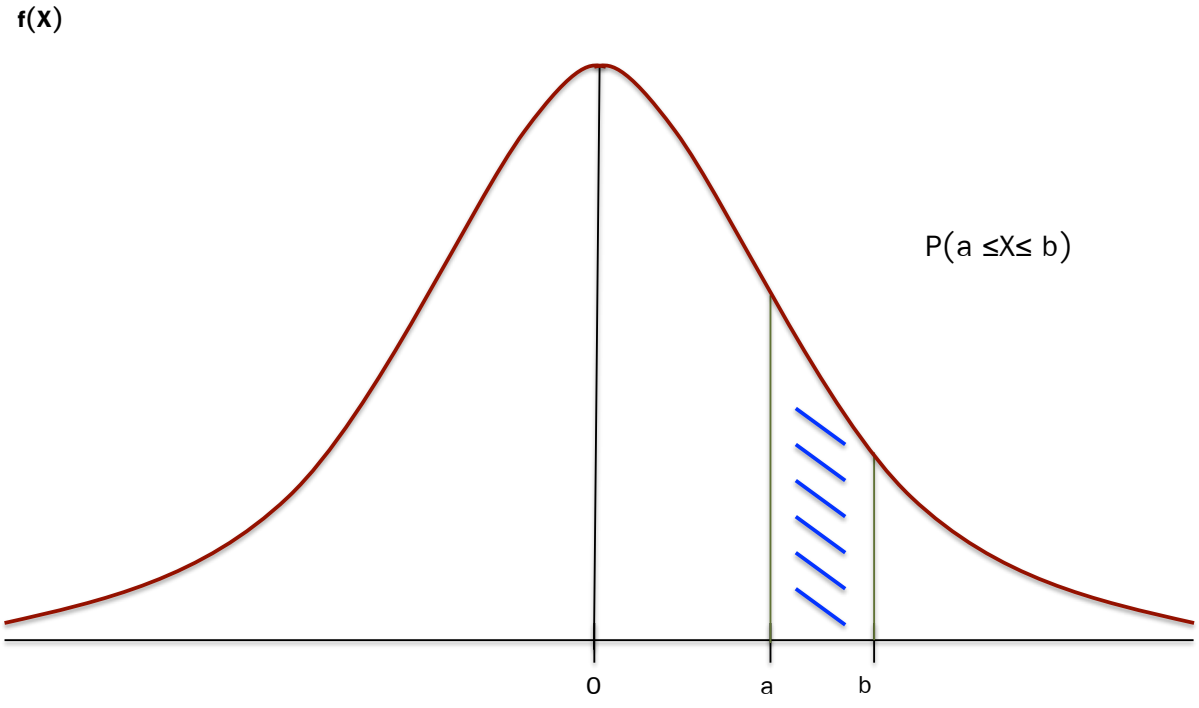
2.4.2 Probability Density Function for Continuous Random Variable

Let X be the continuous random variable. The probability density function of X satisfies the three following conditions.

1. $f(X) \geq 0$
2. $\int_{-\infty}^{\infty} f(X)dx = 1$
3. $\int_{a}^{b} f(X)dx = P(a \leq X \leq b)$

Figure 2-2 exhibits the probability density function for the continuous random variable, where the area under the curve represents the probability that the variable will lay on that range. Specifically, $P(a \leq X \leq b)$ means the probability that X will take the value between a and b.

Figure 2-2: Probability Density Function for Continuous Random Variable



2.4.3 Joint Probability Density Function

In this subsection, only **joint probability density function** for discrete variable is discussed. Let X and Y be discrete random variables. The joint probability density function, identifying the probability that X and Y happen simultaneously, is written as,

$$f(X,Y) = P(X = x \; and \; Y = y)$$

Example: The following table explains the joint probability density function.

Table 2-1: The table illustrating the joint probability density function of X and Y

		X		
		-1	0	1
	1	0.11	0.08	0.05
Y	2	0.09	0.05	0.03
	3	0.35	0.07	0.17

According to the table, the probability that random variable X will be 0 and random variable Y will be 3 is 0.07 or 7 percent. In mathematical term, it can be written as $f(X = 0, Y = 3) = 0.07$.

2.4.4 Marginal Probability Density Function

The above joint probability density function $f(X,Y)$ shows the joint distribution of two variables. On the other hand, **marginal probability density function** with respect to joint probability function, displays the probability density function of single variable like $f(X)$, $f(Y)$, which can be derived from;

$$\begin{array}{rcll} f(X) & = & \sum_Y f(X,Y) & \text{called marginal PDF of X} \\ f(Y) & = & \sum_X f(X,Y) & \text{called marginal PDF of Y} \end{array}$$

where $\sum_Y$ or $\sum_X$ means the summation of probability over all values of X and Y respectively.

Example: According to Table 2-1 above, marginal PDF of X is obtained from

$$
\begin{aligned}
f(X=-1) &= \textstyle\sum_Y f(X=-1,Y) \\
&= f(X=-1,Y=1)+f(X=-1,Y=2)+f(X=-1,Y=3) \\
&= 0.11+0.09+0.35 \\
&= 0.55 \\
f(X=0) &= \textstyle\sum_Y f(X=0,Y) \\
&= f(X=0,Y=1)+f(X=0,Y=2)+f(X=0,Y=3) \\
&= 0.08+0.05+0.07 \\
&= 0.20 \\
f(X=1) &= \textstyle\sum_Y f(X=1,Y) \\
&= f(X=1,Y=1)+f(X=1,Y=2)+f(X=1,Y=3) \\
&= 0.05+0.03+0.17 \\
&= 0.25
\end{aligned}
$$

and marginal PDF of Y is obtained from

$$
\begin{aligned}
f(Y=1) &= \textstyle\sum_X f(X,Y=1) \\
&= f(X=-1,Y=1)+f(X=0,Y=1)+f(X=1,Y=1) \\
&= 0.11+0.08+0.05 \\
&= 0.24 \\
f(Y=2) &= \textstyle\sum_X f(X,Y=2) \\
&= f(X=-1,Y=2)+f(X=0,Y=2)+f(X=1,Y=2) \\
&= 0.09+0.05+0.03 \\
&= 0.17 \\
f(Y=3) &= \textstyle\sum_X f(X,Y=3) \\
&= f(X=-1,Y=3)+f(X=0,Y=3)+f(X=1,Y=3) \\
&= 0.35+0.07+0.17 \\
&= 0.59
\end{aligned}
$$

According to the calculation above, the result can be summarized into Table 2-2.

Table 2-2: Table demonstrating joint probability of random variable X and Y

		X			
		-1	0	1	
Y	1	0.11	0.08	0.05	$f(Y=1)=0.24$
	2	0.09	0.05	0.03	$f(Y=2)=0.17$
	3	0.35	0.07	0.17	$f(Y=3)=0.59$
		$f(X=-1)=0.55$	$f(X=0)=0.20$	$f(X=1)=0.25$	$f(X)=1$ $f(Y)=1$

2.4.5 Conditional Probability Density Function

Conditional probability density function is the probability of one event given that some events have already occurred. The function is written as,

$$f(X|Y) = P(X = x|Y = y)$$

This function can be obtained from the joint probability density function through,

$$f(X|Y) = \frac{f(X,Y)}{f(Y)}$$

Example: According to Table 2.1, find $f(X = 1|Y = 2)$ and $f(Y = 2|X = 0)$

$$\begin{aligned} f(X = 0|Y = 1) &= \frac{f(X=0,Y=1)}{f(Y=1)} \\ &= \frac{0.08}{0.24} \\ &= 0.33 \\ f(Y = 2|X = 0) &= \frac{f(Y=2,X=0)}{f(X=0)} \\ &= \frac{0.05}{0.20} \\ &= 0.25\# \end{aligned}$$

Example: Let event A be tossing the dice once and the point is odd number and B be the tossing the dice once and the point is at least 5. Find the probability that the point coming up is odd given that the point has to be at least 5.

A and B will occur simultaneously if the point from tossing the dice is 5; so, the joint probability of A and B is $\frac{1}{6}$. The probability that B occurs is $\frac{2}{6}$. Hence, the conditional probability of A given B is

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)} = \frac{\frac{1}{6}}{\frac{2}{6}} = \frac{1}{2}\#$$

2.4.6 Statistical Independence

Two random variables are **independent** if the resulting value of one variable does not affect the resulting value of the other; namely,

$$f(X,Y) = f(X)f(Y)$$

Example: Consider Mr. Ake's expenditure for a meal and the Miss Somsri's expenditure for a dessert. Given that they do not know each other, the realization of Mr. Ake's expenditure does not imply the realization of Miss Somsri's expenditure. We can, thus, conclude that the expenditures of these two people are independent#

Example: Consider drawing cards sequentially from the standard 52-card deck without putting it back into the deck. Once the first card is drawn, the probability of drawing the second card will be influenced because the amount of cards in the deck is reduced. In this case, it can be concluded that drawing the first and second card are not independent#

2.5 EXPECTATION, VARIANCE, COVARIANCE AND CORRELATION

2.5.1 Mean or Expected Value

Because the value of random variable hinges on the value of random results of experiment which cannot be determined certainly, statisticians have invented the measures of central tendency of the random variable. One of them is **expected value**.

For discrete random variable, the expected value is calculated by;

$$E(X) = \sum_{i=i}^{n} X_i f(X_i) = X_1 f(X_1) + X_2 f(X_2) + \ldots + X_n f(X_n)$$

For continuous random variable, the expected value is calculated by,

$$E(X) = \int_a^b X f(X) dX$$

where;

$E(X)$ is the measure of central tendency of random variable, resulting from repeated trial of experiment.

$\sum_{i=i}^{n} X_i f(X_i)$ is the average of random variable weighted by the probability corresponding to each value.

a and b are the lowest and highest constant possible respectively.

Example: Find the expected value of rolling two dice once (Figure 2-1)

$$\begin{aligned} E(X) &= \textstyle\sum_{i=2}^{n} X_i f(X_i) \\ &= X_2 f(X_2) + X_3 f(X_3) + \ldots + X_{12} f(X_{12}) \\ &= 2\tfrac{1}{36} + 3\tfrac{2}{36} + 4\tfrac{3}{36} + 5\tfrac{4}{36} + 6\tfrac{5}{36} + 7\tfrac{6}{36} + 8\tfrac{5}{36} + 9\tfrac{4}{36} + 10\tfrac{3}{36} + 11\tfrac{2}{36} + 12\tfrac{1}{36} \\ &= 7 \end{aligned}$$

Hence, the expected value is 7#

Crucial properties of expected value include,

1. $E(b) = b$

2. $E(aX + b) = aE(X) + b$

3. $E(XY) = E(X)E(Y)$; given that X and Y are independent

4. $E(g(X)) = \sum_x g(X)f(X)$

where a and b are constant.

Conditional expectation value is the expectation value of random variable under some conditions such as expectation value of X conditional on Y or $E(X|Y = 5)$

Let $f(X, Y)$ be the joint probability function of X and Y. The expectation of X conditional on some value of Y is defined as,

For discrete random variable $E(X|Y = y) = \sum_X X_i f(X|Y = y)$

For continuous random variable $E(X|Y = y) = \int_{-\infty}^{\infty} X_i f(X|Y = y)$

Example

$$\begin{aligned} E(Y|X = 1) &= \sum_Y Y_i f(Y|X = 1) \\ &= 1f(Y = 1|X = 1) + 2f(Y = 2|X = 1) + 3f(Y = 3|X = 1) \\ &= 1(\tfrac{0.05}{0.25}) + 2(\tfrac{0.03}{0.25}) + 3(\tfrac{0.17}{0.25}) \\ &= 2.48\# \end{aligned}$$

2.5.2 Variance

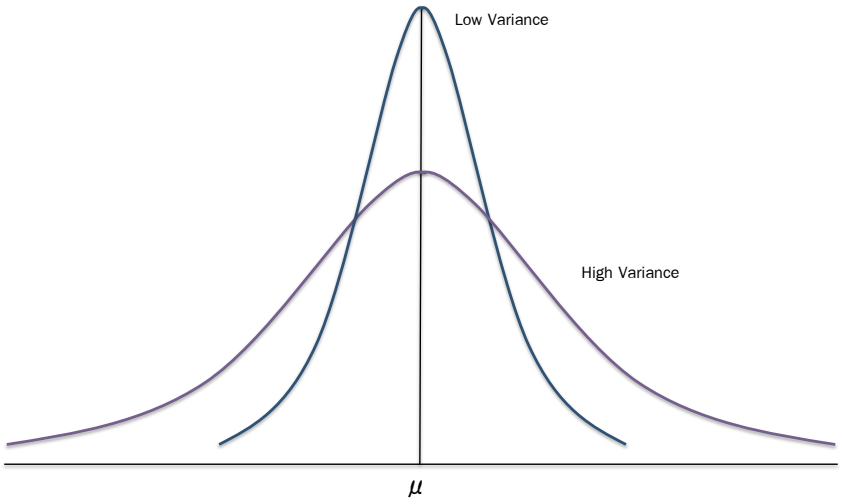
Variance is the measure of dispersion of the value of variable around the expected value. The higher the variance, the more dispersing the random variable (Figure 2-3). If X is the random variable with expected value μ, we get;

$$Var(X) = \sigma_x^2 = E[X - E(X)]^2 = E(X)^2 - \mu^2 \tag{2.4}$$

From,

$$\begin{aligned} Var(X) &= \sigma_x^2 \\ &= E[X - E(X)]^2 \\ &= E[X^2 - 2XE(X) + (E(X))^2] \\ &= E(X^2) - 2(E(X))^2 + (E(X))^2 \\ &= E(X)^2 - \mu^2 \end{aligned}$$

Figure 2-3: Distribution of Random Variables with Different Variance



Important properties of expected value include;

1. $E(X-\mu)^2 = E(X^2) - \mu^2$
2. $Var(b) = 0$
3. $Var(aX+b) = a^2 Var(X)$
4. $Var(X+Y) = Var(X) + Var(Y)$; given that X and Y are independent
5. $Var(aX+bY) = a^2 Var(X) + b^2 Var(Y)$

where a and b are constant.

2.5.3 Covariance

Covariance can indicate the relationship between any two variables such as X and Y. If the covariance is positive, it implies that as X changes, Y will change in the same direction. On the other hand, if the covariance is negative, it implies that as X changes, Y will change in the different direction. Lastly if covariance is zero, it implies no relationship between X and Y. The covariance can be calculated by;

$$cov(X,Y) = E[(X-\mu_X)(Y-\mu_Y)] = E(XY) - \mu_X\mu_Y$$

The essential properties of covariance are;

1. If X and Y are independent,

$$\begin{aligned} cov(X) &= E(XY) - \mu_X \mu_Y \\ &= \mu_X \mu_Y - \mu_X \mu_Y \\ &= 0 \end{aligned}$$

2. $cov(a + bX, c + dY) = bd\ cov(X, Y)$

;where a, b, c and d are constant.

2.5.4 Correlation

Correlation is the measure of covariance in which the value is scaled to range from -1 to 1. For an illustration, if the correlation is equal to 1, it implies that, as X changes, Y changes proportionately in the same direction. Yet, the correlation between X and Y does not necessarily means that X causes Y to change or the other way around. It is merely the measure of relationship between two variables and can be calculated as;

$$\begin{aligned} \rho &= \frac{cov(X,Y)}{\sqrt{var(X)var(Y)}} \\ &= \frac{cov(X,Y)}{\sigma_X \sigma_Y} \end{aligned}$$

2.6 IMPORTANT PROBABILITY DISTRIBUTION

In this section, various types of probability distributions, which are usually applied to economics, are discussed such as *normal distribution*, χ^2 or *Chi-square distribution*, *student's t distribution* and *F distribution*

2.6.1 Normal Distribution

Normal distribution, characterized by the bell-shape, is the probability distribution for continuous random variable. Another interesting property is that the value of mean, median and mode of random variable, which features normal distribution, are the same. The random variable X having normal distribution with mean μ and variance σ^2 is denoted by $X \sim N(\mu, \sigma^2)$ with the function of

$$f(X) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(X-\mu)^2}{2\sigma^2}}$$

Figure 2-4: Normal Distribution

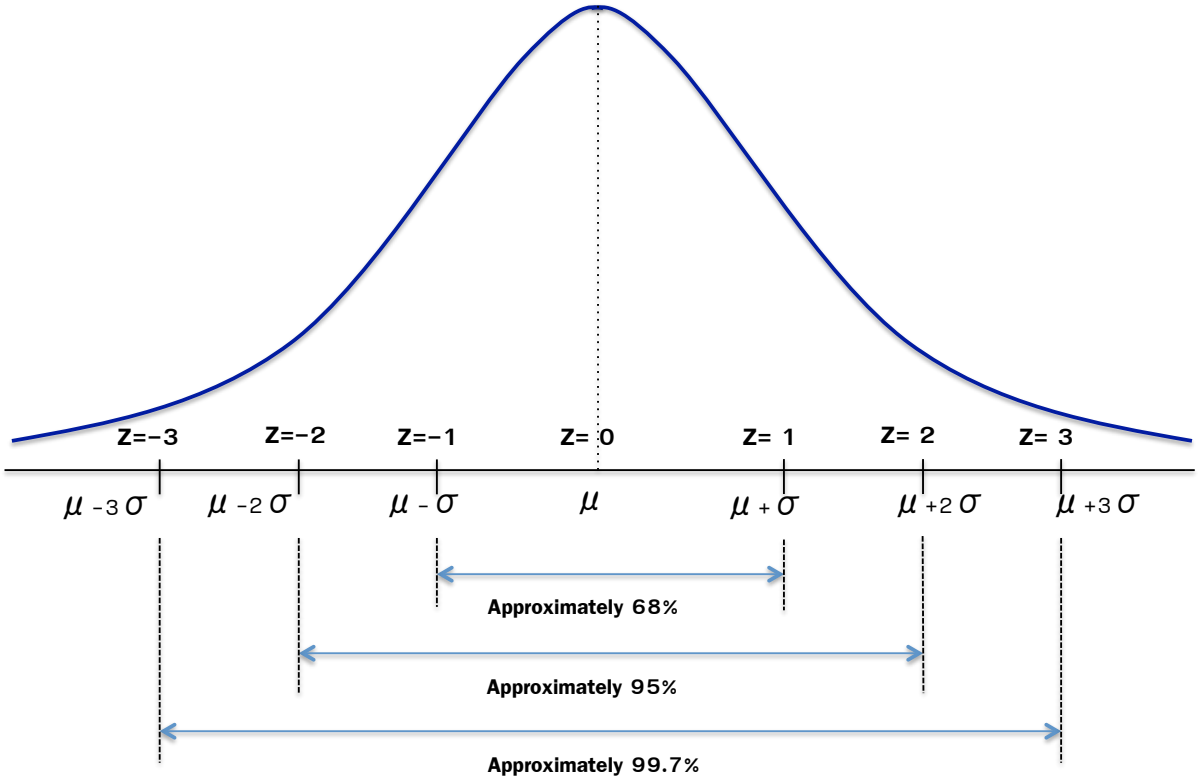


Figure 2-4 depicts the normal distribution that is characterized by bell-shapes and has the following properties.

1. The distribution is symmetric around its mean value.

2. The area under the curve, representing the probability, between $\mu \pm \sigma$, $\mu \pm 2\sigma$ and $\mu \pm 3\sigma$ is approximately 68 percent, 95 percent and 99.7 percent of the total area.

3. The random variable X, having normal distribution with mean μ and variance σ^2, can be converted to standard normal distribution through;

$$Z = \frac{X - \mu}{\sigma}$$

4. Let $X_1 \sim N(\mu_1, \sigma_1^2)$ and $X_2 \sim N(\mu_2, \sigma_2^2)$ and the random variable X_1 and X_2 be independent. Also, let $Y = aX_1 + bX_2$. The random variable Y also has normal distribution, namely $Y \sim N(a\mu_1 + b\mu_2, a^2\sigma_1^2 + b^2\sigma_2^2)$

5. Central limit theorem states that, as the sample size increases (at least 30), the distribution of arithmetic means will be approximately normally distributed; despite the original distribution of population.

6. Mean and variance of random variable normally distributed will be independent from each other. That is, they are not the function of each other.

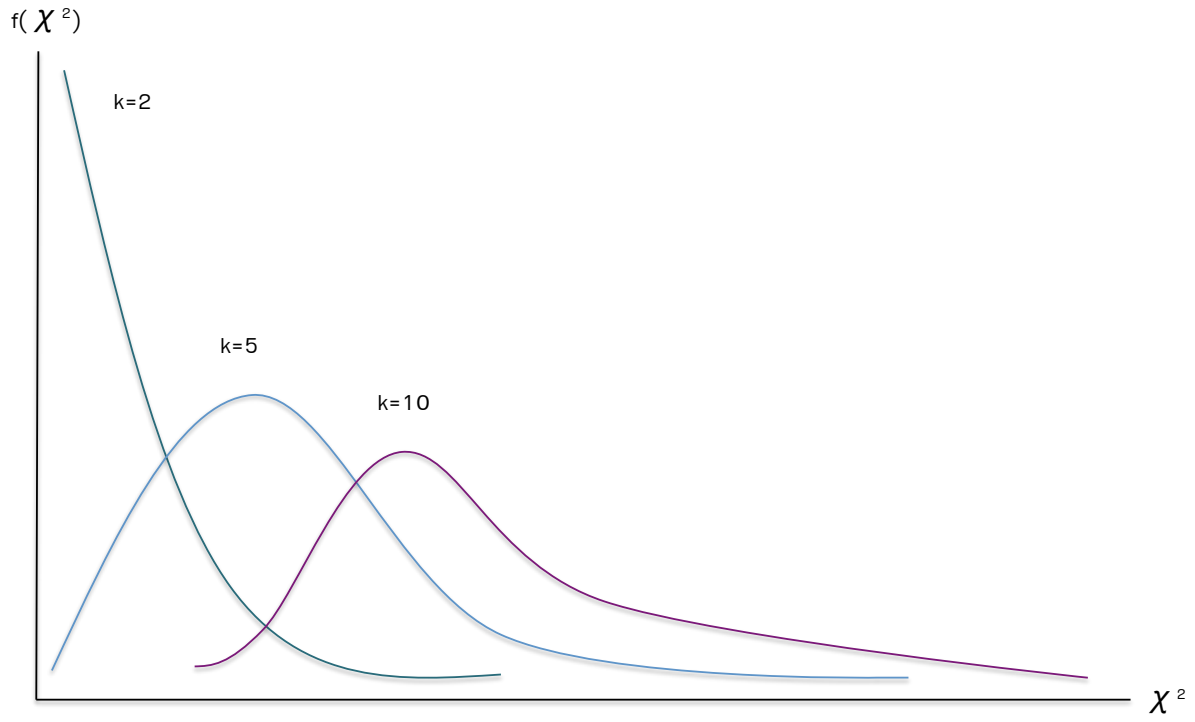
2.6.2 Chi-square Distribution

Let $Z_1, Z_2, , Z_n$ be the random variables that are independent from one another with mean 0 and variance 1 or $Z \sim N(0, 1)$ Then, the variable $\mathbb{Z}$ is defined as;

$$\mathbb{Z} = \sum_{i=1}^{k} Z_i^2$$

The variable $\mathbb{Z}$ has χ^2 distribution with degree of freedom[1] equal to the number of variable Z's constituting $\mathbb{Z}$, which is k. Figure 2-5 portrays χ^2 distribution.

Figure 2-5: Chi-Square χ^2 Distribution



The properties of χ^2 distribution are as following

[1]Degree of freedom is the number of value used to calculate the value final statistic that can vary freely. For example, suppose that the arithmetic mean of 4 value is 100, implying that the sum of 4 value is 400. If 3 values out of 4 are known to be 50, 60 and 90, the last value is restricted to be 200 such that the total sum id 400. In this case, we say the degree of freedom is 3, or generally $n - 1$. Yet, statistically, degree of freedom can be calculated through various fashions.

1. The smaller the degree of freedom (k), the more right-skewed the χ^2 distribution. As the degree of freedom approaches infinity, χ^2 distribution will resemble normal distribution.

2. The mean and variance of χ^2 distribution are k and $2k$ respectively.

2.6.3 Student's t Distribution

Let $Z_1 \sim N(0,1)$ and $Z_2 \sim \chi^2(k)$ and both of them be independent. The variable t is defined as;

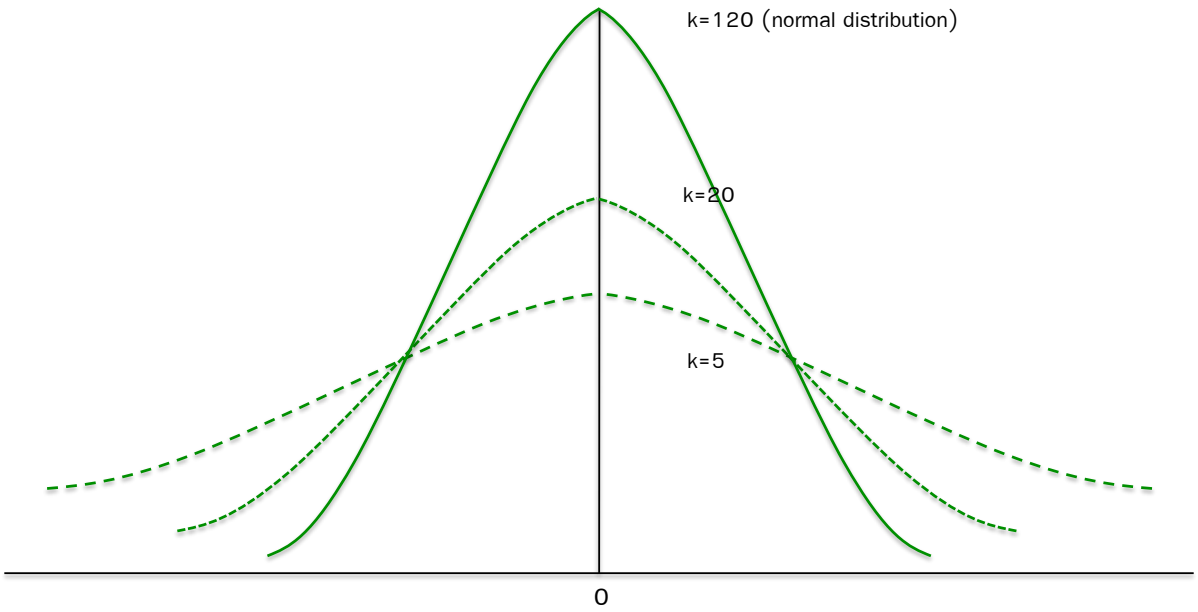
$$t = \frac{Z_1}{\sqrt{\frac{Z_2}{k}}} = \frac{Z_1\sqrt{k}}{\sqrt{Z_2}}, t_k$$

Figure 2-6 exhibits student's t distribution with the following properties.

1. Students t distribution is similar to normal distribution; that is, it is symmetric around the mean value but flatter than the normal distribution. As the degree of freedom increases, students t distribution will approximate normal distribution.

2. The mean and variance of students t distribution are 0 and $\frac{k}{k-2}$ respectively.

Figure 2-6: Student's t Distribution



2.6.4 F Distribution

Let $Z_1 \sim \chi^2(k_1)$ and $Z_2 \sim \chi^2(k_2)$ and both of them be independent, the variable F is defined as;

$$F = \frac{\frac{Z_1}{k_1}}{\frac{Z_2}{k_2}}, F_{k_1,k_2}$$

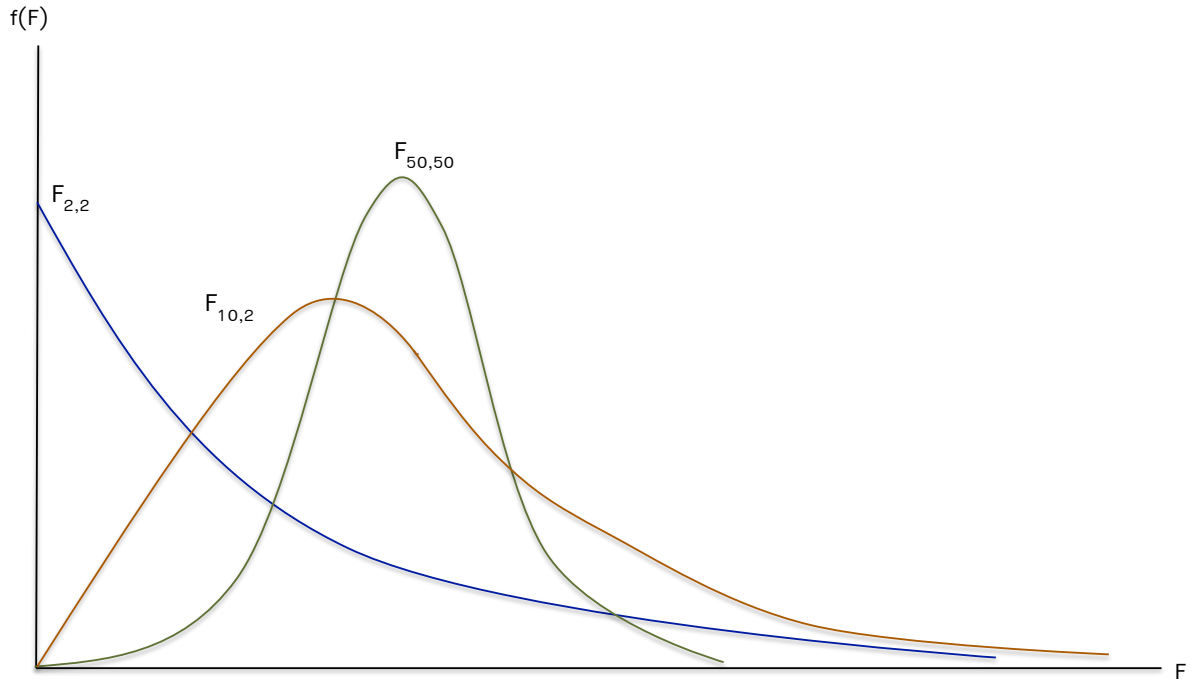
Figure 2.7 illustrates F distribution with the following properties.

1. F distribution is right-skewed distribution. As both degrees of freedom increase, the F distribution will approach normal distribution.

2. The mean and variance of F distribution are $\frac{k_2}{k_2-2}$, which is defined when $k_2 > 2$, and $\frac{2k_2^2(k_1+k_2-2)}{k_1(k_2-2)^2(k_2-4)}$, which is defined when $k_2 > 4$, respectively.

3. $t_k^2 = F_{1,k}$

Figure 2-7: F Distribution



2.7 ESTIMATOR AND ITS DESIRABLE PROPERTIES

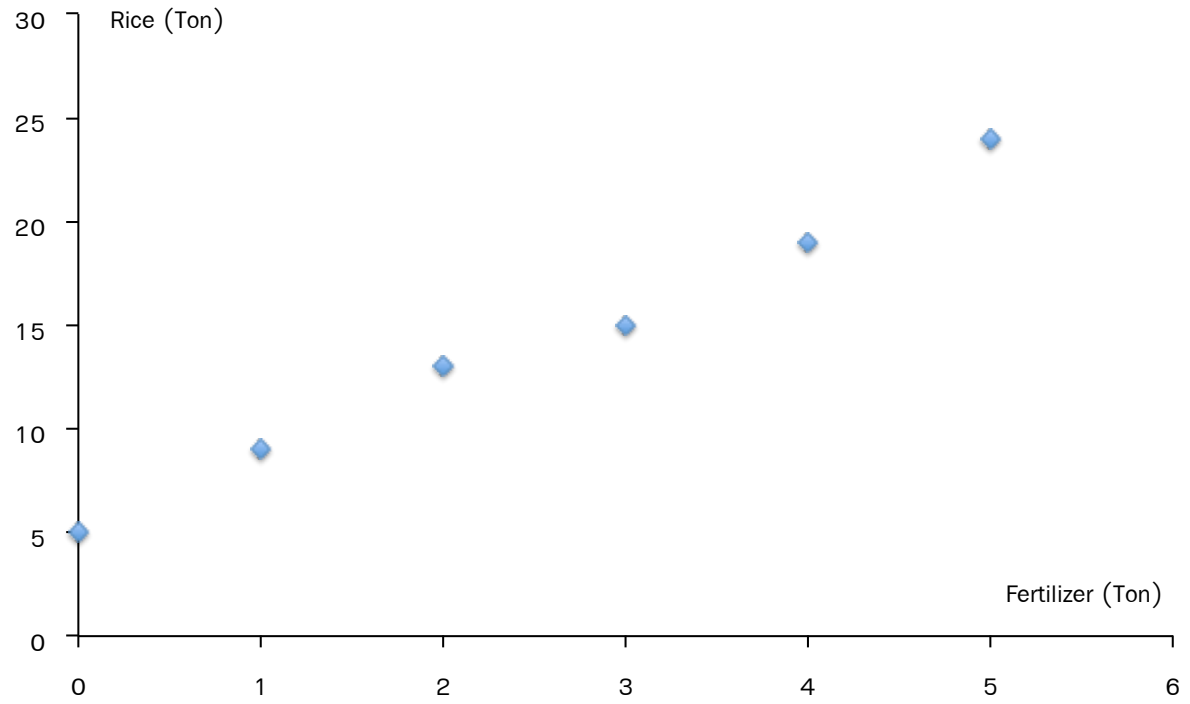
2.7.1 Estimator

Normally, if we want to explain the relationship between X and Y through linear equation, the slope and vertical intercept have to be found which, in turn, use to form linear equation. Unfortunately, this might not always be the case. Consider the data in Table 2-3 that identify the amount of fertilizer used by farmers in one village and the amount of rice product corresponding to each amount of fertilizer. The plot of X and Y is shown in Figure 2-8.

Table 2-3: The amount of fertilizer and the corresponding amount of rice products in ton

X (Fertilizer: Ton)	Y (Rice Product: Ton)
0	5
1	9
2	13
3	15
4	19
5	24

Figure 2-8: The Plot of the Amount of Fertilizer and the Corresponding Amount of Rice product



Suppose the relationship of two variables is explained by;

$$Y_i = a + bX_i$$

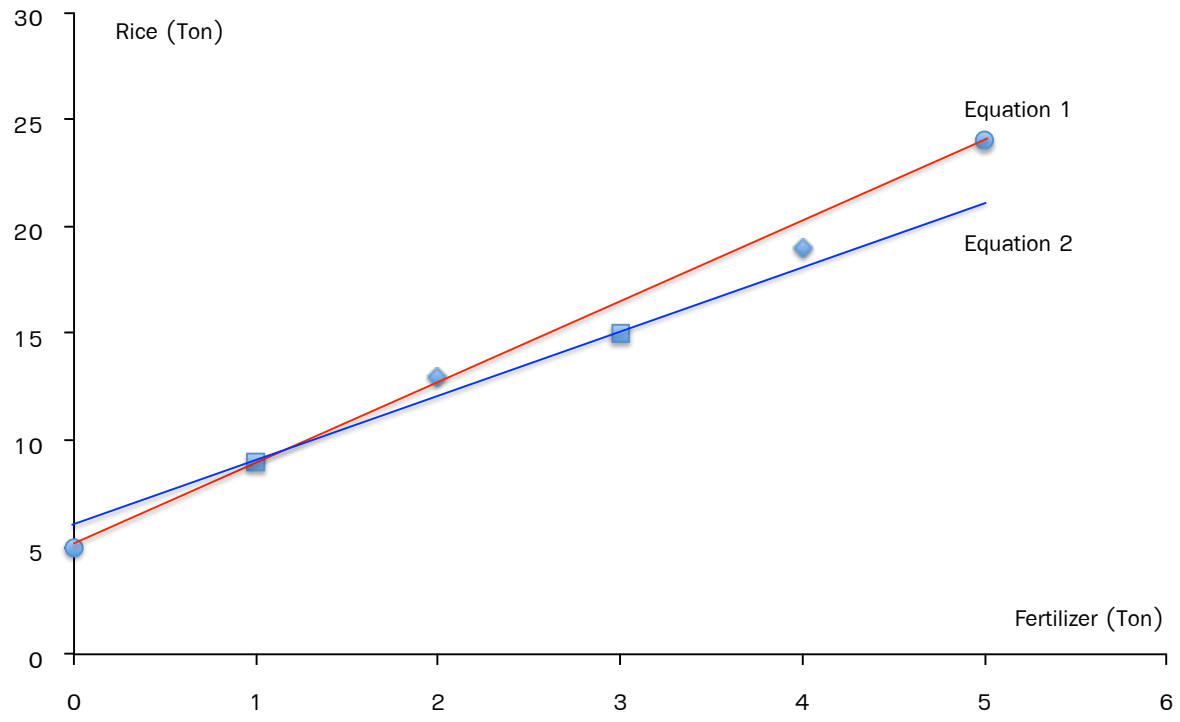
where a and b are population parameter and $\hat{a}$ and $\hat{b}$ are sample statistics or estimators.

When (0,5) and (5,24) are used to obtain vertical intercept and slope, we get;

$$\begin{aligned} & Y & = & \ a + bX_i \\ (0,5) & \ 5 & = & \ a + b(0) \\ (5,24) & \ 24 & = & \ a + b(5) \\ \therefore & \ \hat{a} & = & \ 5 \\ & \ \hat{b} & = & \ 3.8 \end{aligned}$$

According to above computation, the equation 1 is obtained, as depicted in Figure 2-9. Nonetheless, if the data (1, 9) and (3, 15) are used to obtain the linear equation, the equation 2 will result. Consequently, the decision of which pair of data is used determines the values of $\hat{a}$ and $\hat{b}$. The point to be considered is, thus, which sample statistic is the most appropriate.

Figure 2-9: Linear Equations Representing the Relationship between Fertilizer and Rice Product



2.7.2 Desirable Properties

The target of statisticians is to find the sample statistic that identifies the relationship between any variables. Still, the population data is not always available to be used; hence, the sample data is used instead. The method of sampling would yield different set of sample data, which, in turn, results in different values of sample statistics. The difference of the values of parameter and statistic is called *sampling error*.

Accordingly, the best statistic or estimator is the one that generates the lowest sampling error. The procedure of estimation is as following.

Let $\hat{\theta}$ be the estimator of unknown parameter θ, we get

1) The mean of $\hat{\theta} = E(\hat{\theta})$

2) The variance of $\hat{\theta} = E[\hat{\theta} - E(\hat{\theta})]^2$

3) Sampling error $= \hat{\theta} - \theta$

4) $Bias = E(\hat{\theta}) - \theta$

5) Mean Square Error (MSE) $= E[\hat{\theta} - \theta]^2$

Since $\hat{\theta}$ may be greater or lower than θ, the differences of each term might cancel each other when they are summed up to find the average value. Hence, the sum of squared difference is used rather than the sum of difference. Moreover, MSE can be written in another form. To derive;

$$\begin{aligned} MSE &= E[\hat{\theta} - E(\hat{\theta}) + E(\hat{\theta}) - \theta]^2 \\ &= E[(\hat{\theta} - E(\hat{\theta})) + (E(\hat{\theta}) - \theta)]^2 \\ &= E(\hat{\theta} - E(\hat{\theta}))^2 + (E(\hat{\theta}) - \theta)^2 + 2(\hat{\theta} - E(\hat{\theta}))(E(\hat{\theta}) - \theta) \\ &= E[\hat{\theta} - E(\hat{\theta})]^2 + E[E(\hat{\theta}) - \theta]^2 + 2E[(\hat{\theta} - E(\hat{\theta}))(E(\hat{\theta}) - \theta)] \\ &= E[\hat{\theta} - E(\hat{\theta})]^2 + E[E(\hat{\theta}) - \theta]^2 + 0 \\ &= Var(\hat{\theta}) + [Bias(\hat{\theta})]^2 \end{aligned}$$

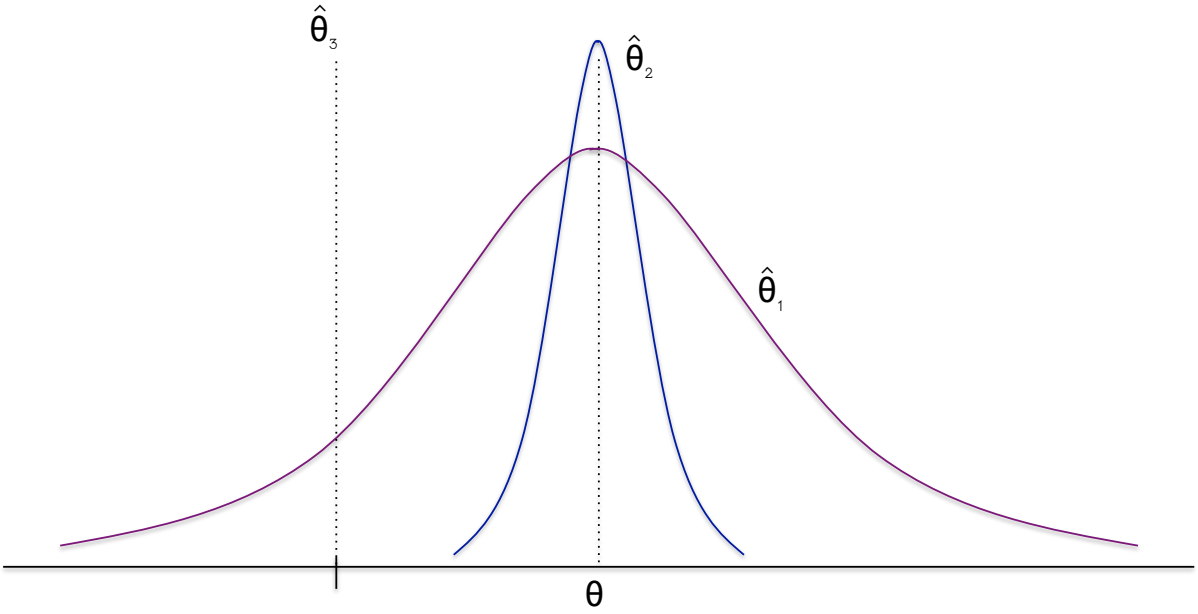
where;

$$\begin{aligned} 2E[(\hat{\theta} - E(\hat{\theta}))(E(\hat{\theta}) - \theta)] &= 2E[(\hat{\theta}E(\hat{\theta}) - \hat{\theta}\theta - [E(\hat{\theta})]^2 + \theta(E(\hat{\theta})] \\ &= 2[E(\hat{\theta})E(\hat{\theta}) - \theta E(\hat{\theta}) - (E(\hat{\theta}))(E(\hat{\theta})) + \theta(E(\hat{\theta})] \\ &= 0 \end{aligned}$$

Therefore, MSE will be minimized if the bias is zero and the variance is low.

The first criterion for considering which estimator is the best is zero bias. Variance, then, is considered. In the Figure 2-10, even the variance of $\hat{\theta_3}$ is zero, its bias (the distance between $\hat{\theta_3}$ and θ) is larger than the bias of $\hat{\theta_1}$ and $\hat{\theta_2}$, which is zero. Hence, only $\hat{\theta_1}$ and $\hat{\theta_2}$ pass the first criterion. According to second criterion, $\hat{\theta_2}$ is better than $\hat{\theta_1}$ thanks to lower variance.

Figure 2-10: The Estimators



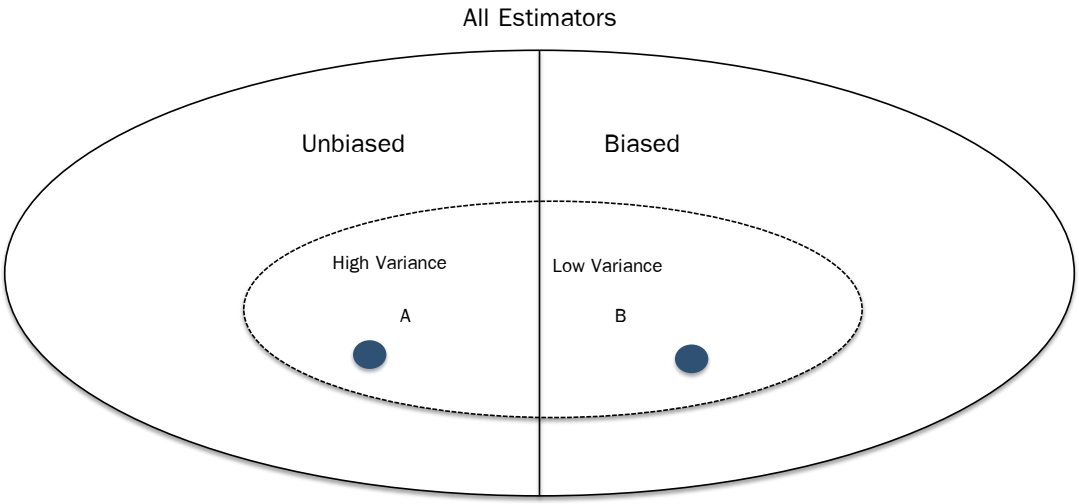
For small sample size, $\hat{\theta}$ is called **efficient estimator** for θ whenever

1. $\hat{\theta}$ must be unbiased, namely $Bias = 0$

2. $\hat{\theta}$ must have minimum variance, namely $Var(\hat{\theta}) \leq Var(\hat{\hat{\theta}})$

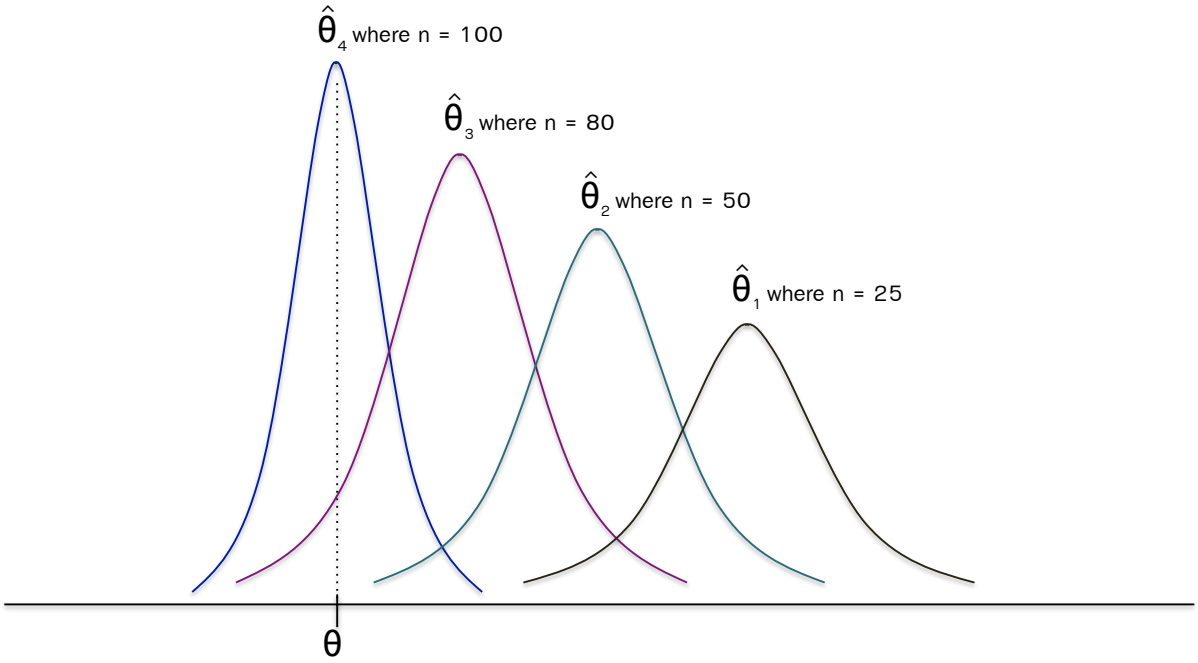
When $\hat{\hat{\theta}}$ is any other unbiased estimators, we can conclude that $\hat{\theta}$ is the minimum variance estimator. Figure 2-11 illustrates the selection of efficient estimator. In spite of low variance, estimator B is biased. In this case, even though estimator A has higher variance but it is unbiased. Thus, estimator A is selected over B.

Figure 2-11: Efficient Estimator



For large sample size, although there is no efficient estimator with the small sample, as the sample size increase, the estimator may become closer to the true parameter. The estimator with this property is said to be **consistent**, as shown in Figure 2-12.

Figure 2-12: Consistent Estimator for Large Sample Size



Chapter 3

SIMPLE REGRESSION MODEL

3.1 METHOD OF ORDINARY LEAST SQUARE

3.1.1 Concept and Assumptions of Model Estimation

Consider the data of weekly income and consumption expenditure of 60 families in one village. The relationship between consumption expenditure (Y) and income (X) can be obtained by letting Y depend on X and be defined as;

$$Y_i = a + bX_i + u_i \tag{3.1}$$

where;

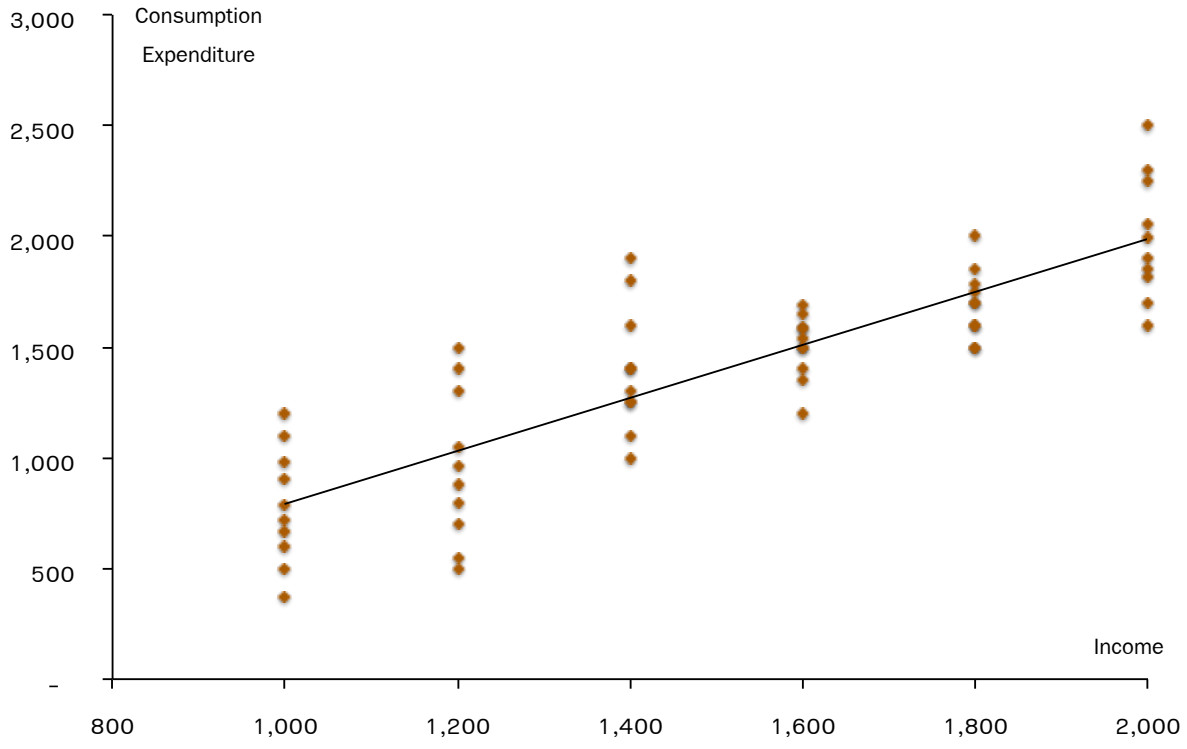
Y is weekly consumption expenditure per family.
X is weekly income per family.
a and b are population parameters
u_i is disturbance term or error term.

Table 3-1 illustrates the data of income and expenditure of 60 families. 10 families receive the weekly income of 1,000 Baht; yet, they have different levels of consumption expenditure ranging from 370 to 1,200 Baht. The condition expectation on the families with 1,000-Baht weekly income is 782 Baht, or $E(Y|X = 100) = 782$. On the other hand, the unconditional expectation for this group of 60 families is 1,386 Baht.

Table 3-1: Data of weekly income and expenditure of 60 families

Income (Baht) X →	1,000	1,200	1,400	1,600	1,800	2,000
Consumption Expenditure (Baht) Y ↓	500	550	1,250	1,540	1,500	1,900
	600	700	1,400	1,690	1,600	2,050
	782	790	1,000	1,500	1,750	1,600
	670	500	1,100	1,590	1,698	1,850
	720	880	1,250	1,580	1,700	1,700
	370	1,400	1,300	1,650	1,850	1,820
	1,200	1,300	1,600	1,500	2,000	1,997
	900	1,500	1,900	1,200	1,780	2,300
	980	963	1,800	1,350	1,600	2,500
	1,100	1,050	1,400	1,400	1,500	2,250
Conditional Expectation $E(Y\|X)$	782	963	1,400	1,493	1,698	1997

Figure 3-1: The Conditional Distribution of Expenditure for Various Levels of Income



Population regression function is acquired through the use of population data to find the linear function explaining the relationship between dependent (Y) and independent variable (X). In Figure 3-1, **population regression line** is drawn to depict this linear relationship of which linear equation [1] is written as;

$$E(Y|X_i) = a + bX_i \tag{3.2}$$

This regression model identifies that the rise in income, on average, will raise the consumption expenditure. However, this expected relationship might not be perfectly applicable to every family. For example, due to an increase in income, some families may increase their expenditure by lower or higher amount than the other families. The above model represents only approximate relationship; hence, there might be some error. This error or disturbance (u_i) can be defined as;

$$u_i = Y_i - E(Y|X_i) \tag{3.3}$$

From equation 3.3, if $E(Y|X_i)$ is a linear equation and y_i is rearranged to be on the left of equation, we get that consumption expenditure (y_i) depends on two factors, which are income (X_i) and disturbance (u_i), and is described by,

$$Y_i = E(Y|X_i) + u_i \tag{3.4}$$

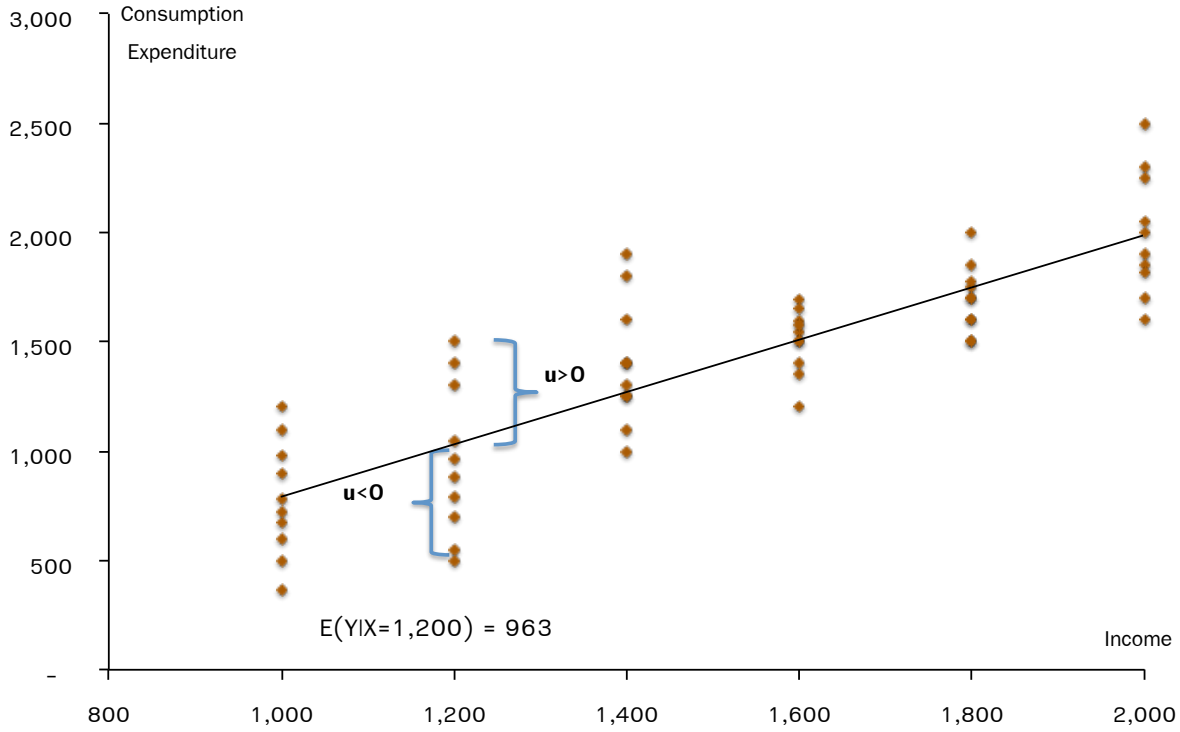
Importantly, disturbance term is random variable since many families have different consumption expenditure, conditional on any specific level of income. The error term, thus, can take many values, which can be positive or negative. Using weekly income of 1,200 Baht as an example, it can be seen that, the error term can take different values (depicted by Figure 3-2) such as;

$$\begin{array}{rrcccl} X = 1,200: & Y_1 & = & 550 & = & a + b(1,200) + u_1 \\ & Y_2 & = & 700 & = & a + b(1,200) + u_2 \\ & & \vdots & & \vdots & \\ & Y_{10} & = & 1050 & = & a + b(1,200) + u_{10} \end{array}$$

where $u_1 \neq u_2 \neq \ldots \neq u_{10}$.

[1] In simple regression analysis, only the regression model that is linear in parameter is considered.

Figure 3-2: Distribution of income, consumption expenditure and disturbance term



Nevertheless, if we assume that population regression line passes through the conditional expectation of Y, the condition expectation of the error term will be zero.

To clarify;

$$
\begin{array}{rrcl}
from & Y_i & = & E(Y_i|X_i) + u_i \\
 & E(Y_i|X_i) & = & E(E(Y_i|X_i)) + E(u_i|X_i) \\
 & & = & E(Y_i|X_i) + E(u_i|X_i) \\
\therefore & E(u_i|X_i) & = & 0
\end{array}
$$

Accordingly, other related effects out of the model are reflected in the error term. The reasons for the existence of disturbance term are as following;

1. Vagueness of theory: In reality, it is difficult to prove the accuracy of economic theory used for testing. Hence, the error term in the model serves as the other factors overlooked by the theory.

2. Unavailability of data: Even though economists realize that there are other factors that should be included in the model, it might be difficult or unavailable to obtain the data of those factors, probably due to problem in collecting process or high cost. The error term, thus, substitutes for those factors.

3. Core variable and peripheral variables: To establish economic model, apart from core independent variable, many trivial variables also possess the explanatory power for the model. Yet, using error term, instead of those peripheral variables, reduces the time and mitigates the burden of collecting the data.

4. Intrinsic randomness in human behaviour: The variety among people is prevalent in the world and their actions are so unanticipated that the error term is required to explain those unpredictable behaviour.

5. Poor proxy variables: Suppose that the economist would like to study the gross domestic product in monthly term; but, only quarterly data is available. Another similar data available in monthly term, like manufacturing production index (MPI), might be used as a proxy. Certainly, this data cannot perfectly represents GDP since a country's production does not rely only on industrial factor. Accordingly, the other effects of GDP that MPI cannot generate are included in the error term.

6. Principle of parsimony: When there are many economic theories with similar ability to explain the real world phenomena, the uncomplicated one is generally preferable and easy to be accepted.

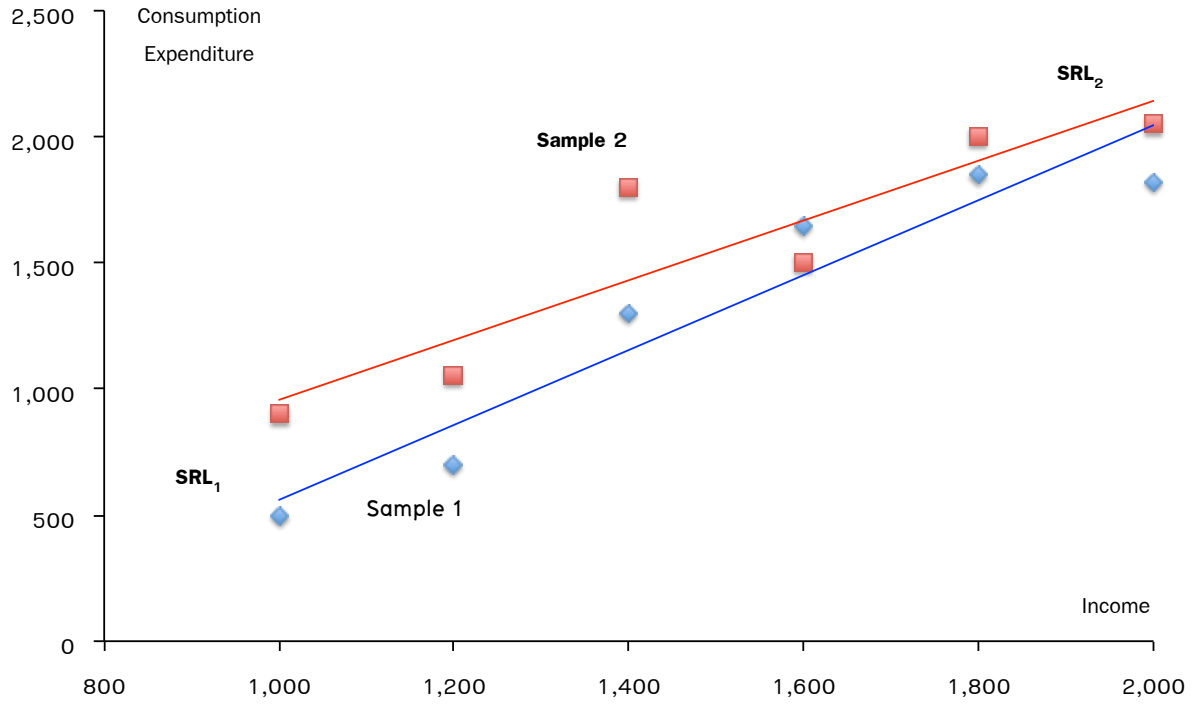
7. Wrong functional form: Although the relationship among variables could be explained in the economic theory, the function that the mathematical economist uses to represent that relation may be wrongly defined. The existence of error term, then, would partially alleviate the effect from this mistake.

Furthermore, in the real world, the population data is usually time-consuming and expensive to obtain. Consequently, only the sample of that population is used to estimate and test the established model. The equation obtained from this process is called **sample regression function**.

Consider the data in Table 3-1. If two sets of 6 data are sampled from the population as in Table 3-2, the scattergram of each data set will be like Figure 3-3.
Table 3-2: Two groups of sample from the population data in Table 3-1

Sample 1		Sample 2	
Y	X	Y	X
500	1000	900	1000
700	1200	1050	1200
1300	1400	1800	1400
1650	1600	1500	1600
1850	1800	2000	1800
1820	2000	2050	2000

Figure 3-3: Scattergram for Each Group of Sample



From Figure 3-3, after sampling the data, we can always obtain the sample regression function from it. $SRL1$ and $SRL2$ depict the sample regression function from different sample. It is, hence, clear that population regression function and sample regression functions from different group of data are not necessarily the same. If we want to acquire the error term for population and sample regression function from the data, we will get;

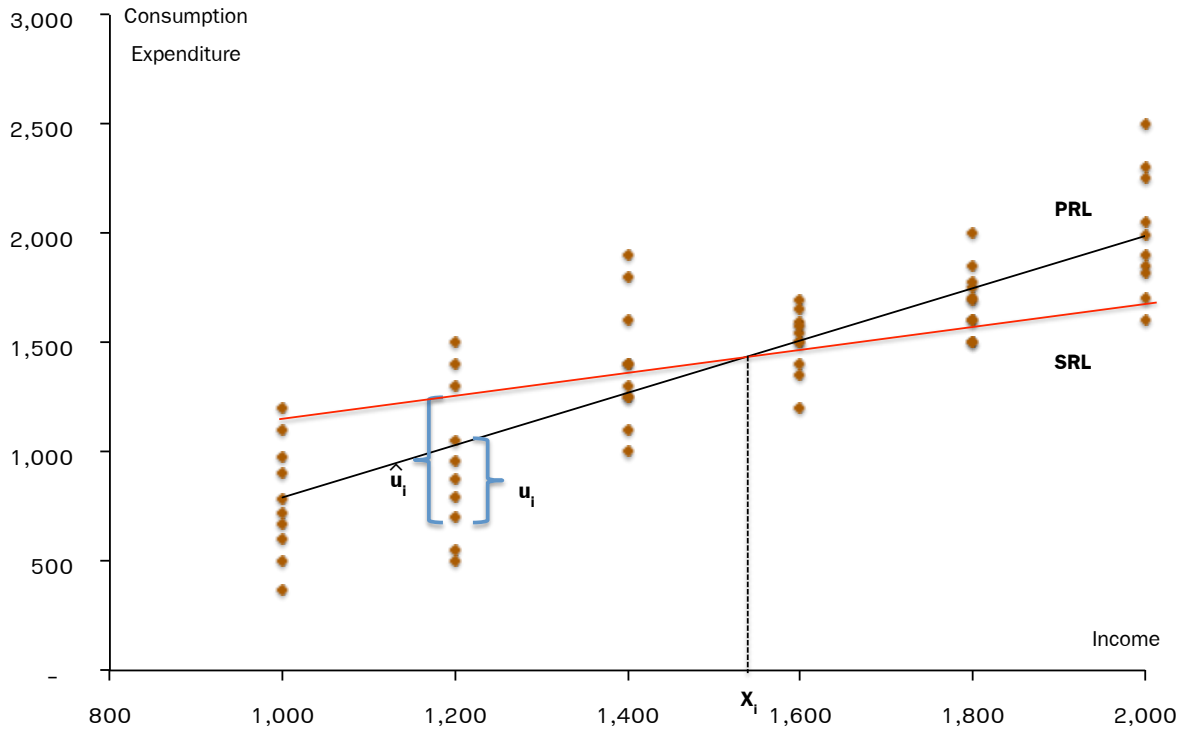
$$\begin{aligned} PRF: \quad Y_i &= a + bX_i + u_i \\ Y_i &= E(Y|X_i) + u_i \end{aligned}$$

$$\begin{aligned} SRF: \quad Y_i &= \hat{a} + \hat{b}X_i + \hat{u}_i \\ \hat{Y}_i &= \hat{a} + \hat{b}X_i \end{aligned}$$

$$Y_i = \hat{Y}_i + \hat{u}_i$$

where u_i and $\hat{u}_i$ are the disturbance terms, namely deviation from the data, of population regression function and sample regression function, respectively. Figure 3-4 illuminates the difference between u_i and $\hat{u}_i$. Also, for income at X_i, $\hat{Y}_i$ obtained from sample regression function is the same as Y_i obtained from population regression function. For income greater than X_i, $\hat{Y}_i$ underestimates Y_i and, for income lower than X_i, $\hat{Y}_i$ overestimates Y_i.

Figure 3-4: Comparison between Population and Sample Regression Function



As shown in Figure 3-4, the sample regression function with the least $\hat{u}_i$ implies that the function has the least error or deviation from the data. In the next section, **ordinary least square**, which is one of the methods used to obtain estimator with the least error, will be discussed.

3.1.2 Ordinary least square (OLS)

To get sample regression function through ordinary least square (OLS) estimation is to find the estimator, which results in the least $\hat{u}_i$, by making the value of ${\hat{u}_i}^2$ as low as possible [2] With the knowledge of calculus, the procedure of minimizing the error term is as follows;

$$\begin{aligned} \hat{Y}_i &= \hat{a} + \hat{b}X_i \\ \hat{u}_i &= Y_i - \hat{Y}_i = Y_i - \hat{a} - \hat{b}X_i \\ {\hat{u}_i}^2 &= (Y_i - \hat{a} - \hat{b}X_i)^2 \\ \sum {\hat{u}_i}^2 &= \sum (Y_i - \hat{a} - \hat{b}X_i)^2 \end{aligned}$$

[2]We consider ${\hat{u}_i}^2$ because $\sum \hat{u}_i$ might be zero, though the deviations are not smallest.

Minimize$_{\hat{a},\hat{b}} \sum \hat{u_i}^2$, we get;

$$\begin{array}{rcccl} \frac{\partial \sum \hat{u_i}^2}{\partial \hat{a}} & = & \frac{\partial \sum (Y_i - \hat{a} - \hat{b}X_i)^2}{\partial \hat{a}} & = & 0 \\ \frac{\partial \sum \hat{u_i}^2}{\partial \hat{b}} & = & \frac{\partial \sum (Y_i - \hat{a} - \hat{b}X_i)^2}{\partial \hat{b}} & = & 0 \end{array}$$

Then, the solutions of this minimizing problem are the estimator $\hat{b}$, with close form solution as equation 3.5, and the estimator $\hat{a}$, with close form solution as equation 3.6.

$$\hat{b} = \frac{n \sum X_i Y_i + \sum X_i \sum Y_i}{n \sum X_i^2 - (\sum X_i)^2} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} \tag{3.5}$$

$$\hat{a} = \frac{\sum X_i^2 \sum Y_i - \sum X_i \sum X_i Y_i}{n \sum X_i^2 - (\sum X_i)^2} = \bar{Y} - \hat{b}\bar{X} \tag{3.6}$$

We call $\hat{a}$ and $\hat{b}$ the OLS estimators with the following properties.

1. **Sample regression line** (**SRL**) passes sample mean of X and Y $(\bar{X}, \bar{Y})$

2. $\sum \hat{u_i} = 0$

3. $\sum \hat{u_i} X_i = 0$

4. $\sum \hat{u_i} \hat{Y_i} = 0$

Nonetheless, OLS estimation requires **classical linear regression model** (**CLRM**) **assumptions**; otherwise the estimators would suffer some problems. In the following chapters, the relaxation of these assumptions will be discussed to study the effects on the estimators after the relaxation.

1. The regression model must be linear in parameter.

2. The independent variable (X) must be fixed, or nonstochastic, in repeated sampling

3. $E(u_i) = 0$

4. $Var(u_i) = E[u_i - E(u_i)]^2 = E(u_i^2) = \sigma_u^2$

5. $cov(u_i, u_j) = E[u_i - E(u_i)][u_j - E(u_j)] = E(u_i u_j) = 0$

6. $cov(X_i, u_i) = E[X_i - E(X_i)][u_i - E(u_i)] = E(X_i u_i) = 0$

7. The amount of data must be greater than the number of parameters in the model; otherwise, the normal equations will be unsolvable.

8. The variance of X must not be zero.

9. There is no specification error.

10. There is no perfect multicollinearity.

Without the violation of above assumptions, it can be proved that the OLS estimators feature the desirable properties stated in chapter 2; that is, it is best (or having minimum variance) and unbiased. Thanks to Gauss-Markov Theorem, it can be concluded that OLS estimators are **BLUE** (**Best Linear Unbiased Estimator**). The variance and standard deviation of OLS estimators are calculated as;

$$Var(\hat{a}) = \frac{\sigma_u^2 \sum X_i^2}{n \sum (X_i - \bar{X})^2} \tag{3.7}$$

$$se(\hat{a}) = \sigma_u \sqrt{\frac{\sum X_i^2}{n \sum (X_i - \bar{X})^2}} \tag{3.8}$$

$$Var(\hat{b}) = \frac{\sigma_u^2}{\sum (X_i - \bar{X})^2} \tag{3.9}$$

$$se(\hat{b}) = \frac{\sigma_u}{\sqrt{\sum (X_i - \bar{X})^2}} \tag{3.10}$$

$$cov(\hat{a}, \hat{b}) = -\bar{X}[Var(\hat{b})] \tag{3.11}$$

From the equations above, it can be seen that the variance and covariance of $\hat{a}$ and $\hat{b}$ depends on the variance of the error term. When the true value of the variance of the error term is unknown, the estimator of that variance has to be obtained through;

$$\hat{\sigma_u}^2 = \frac{\sum \hat{u_i}^2}{n-2} \tag{3.12}$$

The next crucial point is the ability of OLS estimators to explain the movement of the real data. The measure for this ability is **coefficient of determination:** R^2. This coefficient helps identify the extent to which the movement of independent variable affects the movement of dependent one. Zero R^2 means the change in the independent variable cannot explain the behaviour of the dependent variable. On the other hand, $R^2 = 1$ means the movement in the independent variable perfectly explain the movement of the dependent one. R^2 can be calculated as;

$$\begin{aligned}
\hat{u}_i &= Y_i - \hat{Y}_i \\
\hat{u}_i &= Y_i - \bar{Y} - \hat{Y}_i + \bar{Y} \\
\hat{u}_i &= [Y_i - \bar{Y}] - [\hat{Y}_i - \bar{Y}] \\
\hat{u}_i &= y_i - \hat{y}_i
\end{aligned}$$

$$\begin{aligned}
y_i &= \hat{y}_i + \hat{u}_i \\
y_i^2 &= (\hat{y}_i + \hat{u}_i)^2 \\
&= \hat{y}_i{}^2 + \hat{u}_i{}^2 + 2\hat{y}_i e_i
\end{aligned}$$

$$\begin{aligned}
\sum y_i^2 &= \sum \hat{y}_i{}^2 + \sum \hat{u}_i{}^2 + 2\sum \hat{y}_i e_i \text{ where } \sum \hat{y}_i e_i = 0 \\
\therefore \sum y_i^2 &= \sum \hat{y}_i{}^2 + \sum \hat{u}_i{}^2 \\
TSS &= ESS + RSS
\end{aligned}$$

where;

$$\begin{aligned}
\sum y_i^2 &= \sum (Y_i - \bar{Y})^2 = \text{Total Sum of Square: TSS} \\
\sum \hat{y}_i{}^2 &= \sum (\hat{Y}_i - \bar{Y})^2 = \text{Explained Sum of Square: ESS} \\
\sum \hat{u}_i{}^2 &= \sum (Y_i - \hat{Y})^2 = \text{Residual Sum of Square: RSS}
\end{aligned}$$

The definition of R^2 is the ratio of ESS to TSS ($\frac{ESS}{TSS}$) which can be calculated as equation 3.13 and 3.14.

$$R^2 = \frac{ESS}{TSS} = \frac{\sum \hat{y}^2}{\sum y_i^2} \tag{3.13}$$

In another form;

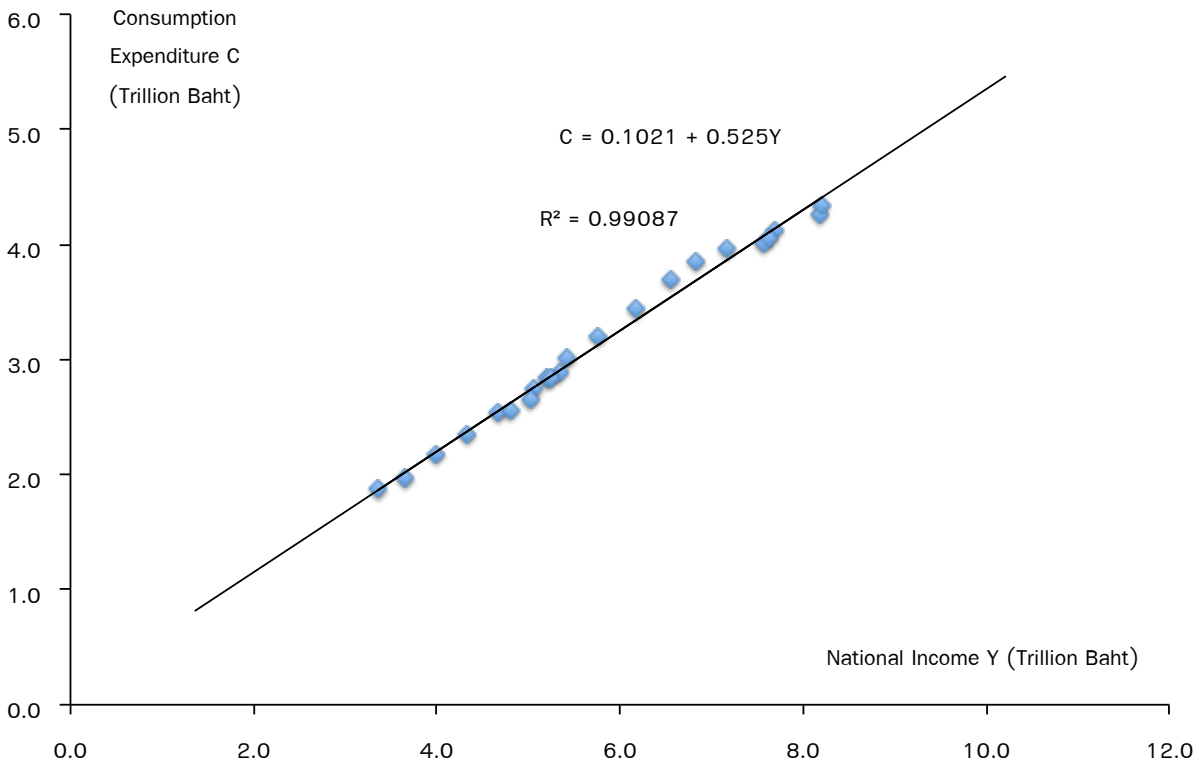
$$\begin{aligned}
TSS &= ESS + RSS \\
\frac{TSS}{TSS} &= \frac{ESS}{TSS} + \frac{RSS}{TSS} \\
1 &= R^2 + \frac{RSS}{TSS}
\end{aligned}$$

$$R^2 = 1 - \frac{\sum \hat{u}_i{}^2}{\sum y_i^2} \tag{3.14}$$

3.1.3 Interpretation and Important Statistics for Simple Regression Model

For sample linear regression, the interpretation of both $\hat{a}$ and $\hat{b}$ will be described through the example of regression function mentioned in Chapter 1, which illustrates the relationship between consumption expenditure and national income of Thailand.

Figure 3-5: Simple Regression Model and Interpretation



For the figure above, it can be found that the slope of the model is 0.525 and the vertical intercept is 0.1021. The interpretation is, as the national income increase by 1 million million Baht. Thai economy will experience the increase in consumption expenditure by approximately 0.525 million million Baht.

Also, if national income of Thailand is zero, Thai economy still has the consumption expenditure of 0.1021 million million baht, which might be received from borrowing or liquidating the properties.

The coefficient of determination or R^2 is equal to 0.99087, which means the change of national income 99.087-percent explains the change in consumption expenditure in Thailand.

3.2 INTERVAL ESTIMATION AND HYPOTHESIS TESTING: THE TEST OF STATISTICAL SIGNIFICANCE

3.2.1 Concept and Additional Assumption

In the previous section, the estimators obtained is specifically called **point estimator** with the belief that they can represent the parameters in the model explaining the relationship of independent and dependent variables. Also, it is proved that the OLS estimators are BLUE.

Nevertheless, the **interval estimation** could be established. It is the estimation of parameter by specifying the possible range in which the parameter most possibly lies. To reach that result, the normality assumption of the error term (u_i) is required. Mathematically,

$$u_i \sim N(0, \sigma_u^2) \quad i = 1, 2, \ldots, n$$

With this assumption, the error term possesses the following properties.

1. Expected value:

$$E(u_i) = 0 \tag{3.15}$$

2. Variance:

$$E[u_i - E(u_i)]^2 = E(u_i^2) = \sigma_u^2 \tag{3.16}$$

3. Covariance between u_i and u_j:

$$E[u_i - E(u_i)][u_j - E(u_j)] = E(u_i u_j) \quad i \neq j \tag{3.17}$$

Also, u_i and u_j are independent from each other. In other word;

$$u_i \sim NID(0, \sigma_u^2) \tag{3.18}$$

where NID stands for **normally and independently distributed**.

3.2.2 Interval Estimation

As the repeated sampling from population may results in different group of sample and different estimators, it can be concluded that estimators are random variable that depends on the sample. Moreover, when considering the two estimators, $\hat{a}$ and $\hat{b}$, both are normally distributed, $\hat{a} \sim (a, \sigma^2_{\hat{a}})$ and $\hat{b} \sim (b, \sigma^2_{\hat{b}})$. If the two estimators are transformed into standard Z-value, we get

$$Z_{\hat{a}} = \frac{\hat{a} - a}{\sigma_{\hat{a}}}$$

$$Z_{\hat{b}} = \frac{\hat{b} - b}{\sigma_{\hat{b}}}$$

Notice that the estimators of their variance are χ^2 distributed.

$$(n-2)\frac{\hat{\sigma^2_u}}{\sigma^2_u} \sim \chi^2_{n-2}$$

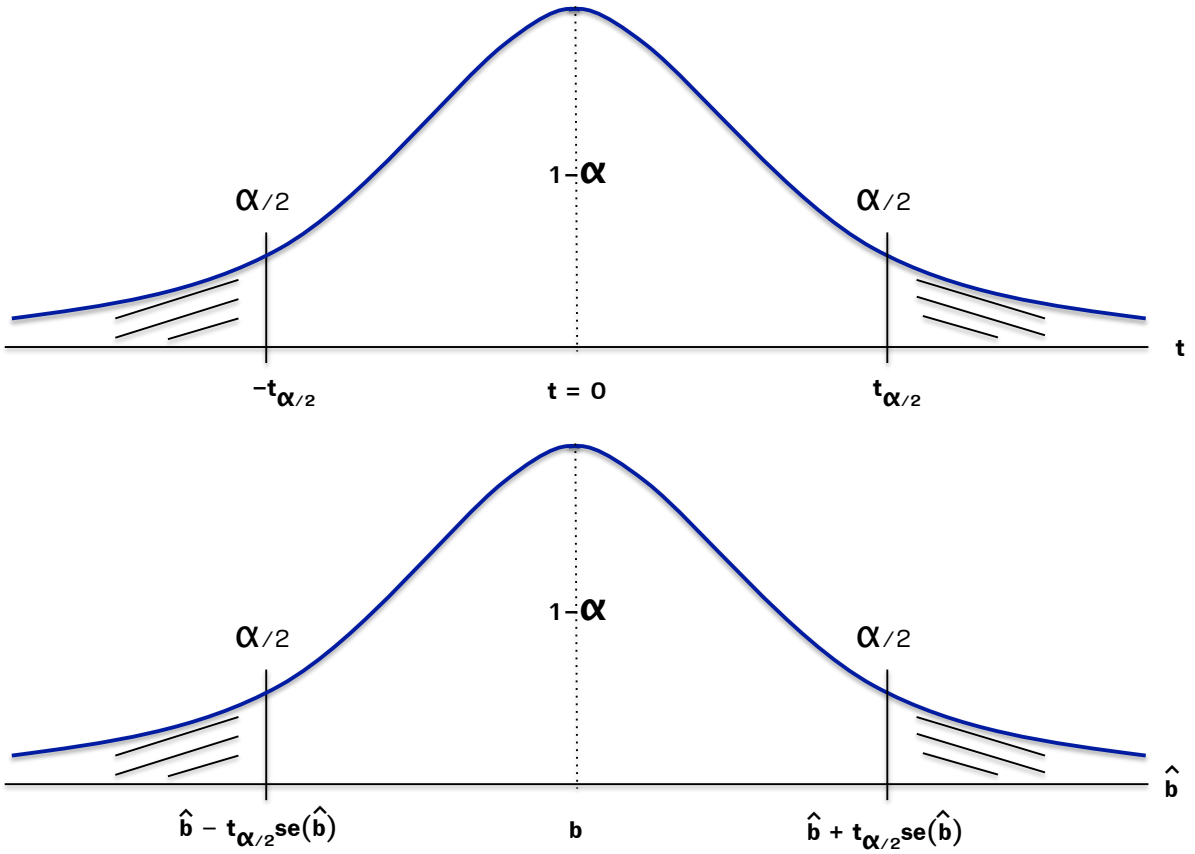
Nonetheless, the variances used to obtain the standard Z-value are unknown parameter; so, the estimator ($\hat{\sigma_u}^2 = \frac{\sum \hat{u_i}^2}{n-2}$). Hence, t-distribution, with n-2 degree of freedom, has to be used instead of standard normal distribution.

$$\hat{t}_{\hat{b}} = \frac{\hat{b} - b}{se(\hat{b})},\ t_{n-2} \tag{3.19}$$

where $\hat{\sigma}^2_{\hat{b}} = \frac{\hat{\sigma}_u}{\sqrt{(X_i - \bar{X})^2}}$.

t-value is distributed as Figure 3-6. The **level of significance** would identify the area under the curve of t-distribution and probability. For example, the level of significance of 0.01, 0.05, and 0.1 imply that the areas under the curve on the left and on the right of diagram are 0.01, 0.05 and 0.1 respectively. Also, this means the areas in the middle $(1 - \alpha)$ are 99 percent, 95 percent and 90 percent respectively.

Figure 3-6: t-distribution at α Level of Significance



When the level of significance of t-distribution is known, **confidence interval** can be formed as;

$$P(-t_{\frac{\alpha}{2}} \leq \hat{t} \leq t_{\frac{\alpha}{2}}) = 1 - \alpha \tag{3.20}$$

We call $t_{\frac{\alpha}{2}}$ the **critical value** at level of significance specified in the t-distribution. Furthermore, when t-value in equation 3.19 is substituted in equation 3.20, we get;

$$P[-t_{\frac{\alpha}{2}} \leq \frac{\hat{b} - b}{se(\hat{b})} \leq t_{\frac{\alpha}{2}}] = 1 - \alpha \tag{3.21}$$

Rearrange the equation 3.21;

$$P[\hat{b} - t_{\frac{\alpha}{2}} se(\hat{b}) \leq b \leq \hat{b} + t_{\frac{\alpha}{2}} se(\hat{b})] = 1 - \alpha \tag{3.22}$$

Consequently, the confidence interval at $100(1-\alpha)\%$ of b is

$$\hat{b} \pm t_{\frac{\alpha}{2}} se(\hat{b}) \tag{3.23}$$

From above confidence interval, if the standard deviation increases, the confidence interval will be widened or the uncertainty of whether the parameter lies in that interval will increase. Additionally, if the level of significance changes from 0.05 to 0.01, the confidence interval will be wider to make sure that the parameter will lie within that interval.

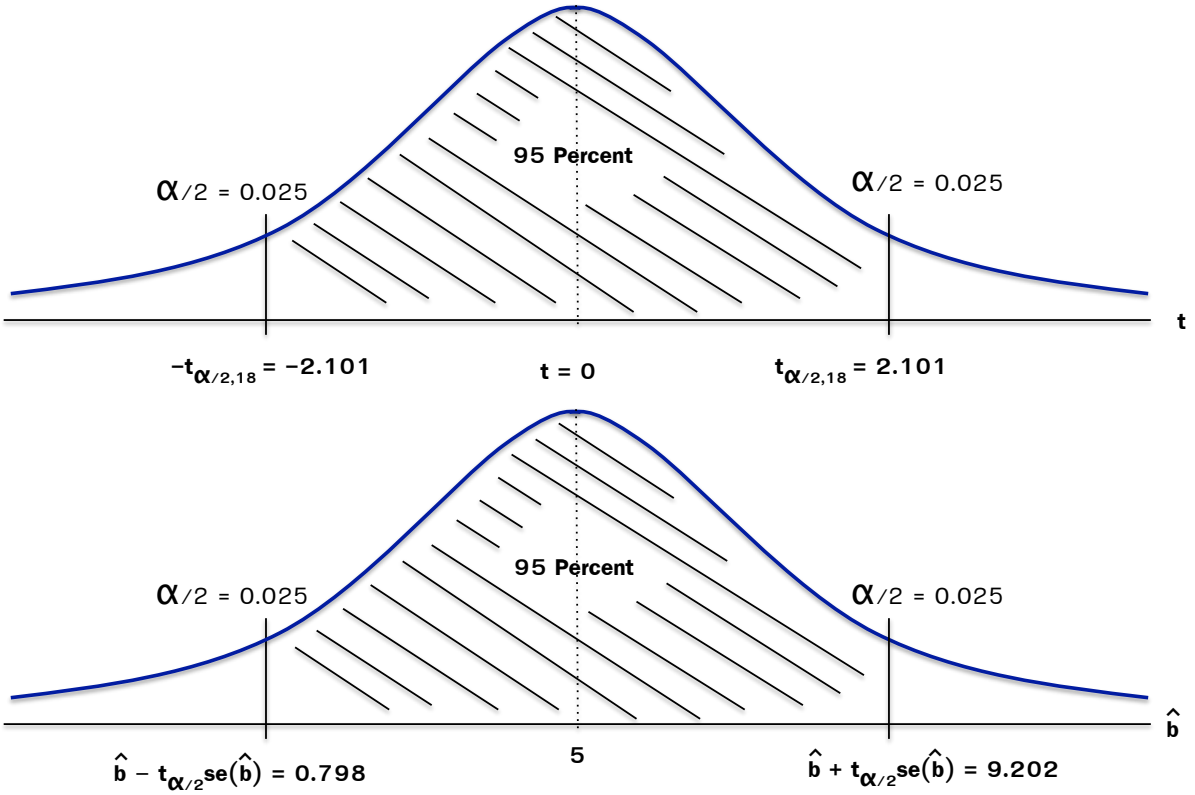
For the interpretation of confidence interval, for example, the confidence level of 95 percent means, 95 out of 100 cases will include the true parameter. This is different from the meaning that any particular confidence interval has 95% probability to include the parameter. The reason is that any particular interval is already fixed; thus, the probability that the parameter will be in that interval is either 1 or 0.

Example: From the model $Y_i = a + bX_i + u_i$. Let $\hat{b} = 5$ and $se(\hat{b}) = 2$. The sample size is 20. Find the 95% confidence interval of b.

$b = \hat{b} \pm t_{\frac{\alpha}{2}=0.025, df=18} se(\hat{b}) = 5 \pm 2.101(2)$

Hence, 95% confidence interval will be between 0.798 and 9.202 as shown in Figure 3-7#

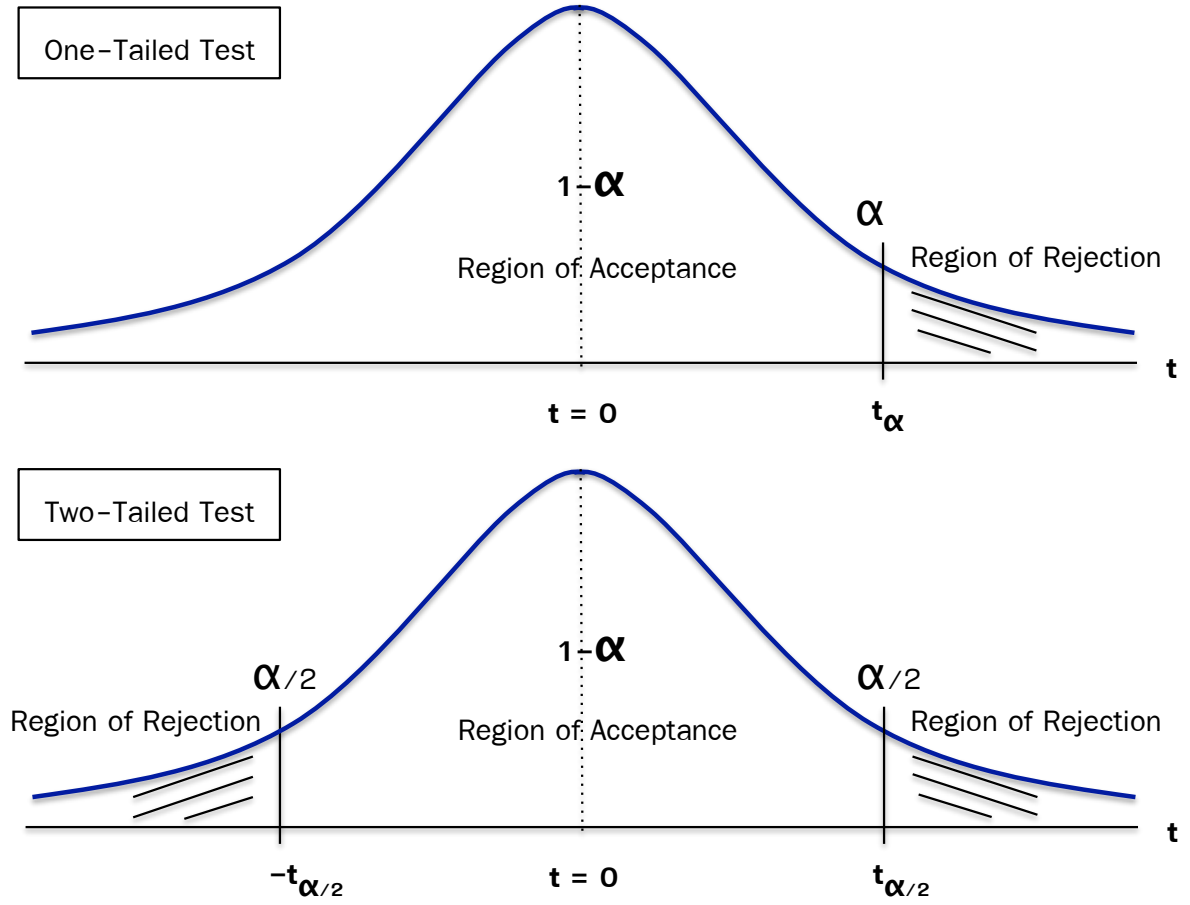
Figure 3-7: 95% Confidence Interval



3.2.3 Hypothesis Testing

Hypothesis testing is the verification of the hypothesis like whether the value of parameter a or b is zero, namely whether the change in independent variable X can explain the change in dependent variable Y. Generally, hypothesis testing may be performed in two fashions which are *one-tailed test* and *two-tailed test*. Figure 3-8 depicts both type of hypothesis testing.

Figure 3-8: One-tailed and Two-tailed Hypothesis Testing



To clarify, **two-tailed hypothesis testing** is used when there is no basis of theory or no specification of relationship between variables before the test. For instance, if we would like to test whether parameter b is statistically significantly different from 0 or not. The hypothesis can be set as;

$$\begin{aligned} H_0 &: \quad b \quad = \quad 0 \\ H_a &: \quad b \quad \neq \quad 0 \end{aligned}$$

where H_0 and H_a are the null hypothesis and alternative hypothesis respectively.

To verify if the parameter b is different from 0 is to consider, at specific confidence level like 95%, whether there is the probability that $b = 0$. If it is different from 0, we can conclude that, the null hypothesis that $b = 0$ is rejected at 95% confidence level. However, if it is not different from 0, we cannot reject the null hypothesis, namely the result is not statistically significant at 0.05.

One-tailed hypothesis testing is proper when there is the basis of economic theory to specify the relationship between two variables and the hypothesis can be set as;

$$\begin{aligned} H_0 : \quad b &\leq 0 \\ H_a : \quad b &> 0 \end{aligned}$$

In this case, it is the test of whether parameter b is less than or equal to zero.

To test for statistical significance, considering Equation 3.21 that identify the confidence interval at the level of significance α, we can identify b^*, which is the value from null hypothesis, and the critical t-value with $n - 2$ degree of freedom which is obtained from the t-table. Then, Equation 3.21 can be rearranged as;

$$P[-t_{\frac{\alpha}{2}} \leq \frac{\hat{b} - b^*}{se(\hat{b})} \leq t_{\frac{\alpha}{2}}] = 1 - \alpha \tag{3.24}$$

Rearrange Equation 3.24 again, we get;

$$P[b^* - t_{\frac{\alpha}{2}} se(\hat{b}) \leq \hat{b} \leq b^* + t_{\frac{\alpha}{2}} se(\hat{b})] = 1 - \alpha \tag{3.25}$$

The result above is the interval that the estimator $\hat{b}$ may lie in at α level of significance. This interval is the identification of region of acceptance and if $\hat{b}$ lie outside this region (or lie in the region of rejection), it can be conclude that the null hypothesis should be rejected. The reason is that, even at probability of α, we still obtain that value.

The procedure of hypothesis testing is as follows;

1. Establish the null and alternative hypothesis.

2. Select the level of significance

3. Calculate the value of $\hat{t}$ from the observed data.

4. Define the region of acceptance and rejection or find the critical value

5. Compare the critical value to the value of $\hat{t}$ and reject the null hypothesis if the absolute value of $\hat{t}$ is greater than the critical value.

Example: From the model $Y_i = a + bX_i + u_i$, let $\hat{b} = 5$ and $se(\hat{b}) = 2$. The sample size is 20. Test the hypothesis of whether X is the statistically significant variable at $\alpha = 0.05$ (Hint: Use the one-tailed hypothesis testing since there is economic theory supporting that b has to be greater than 0)

1. Establish the null and alternative hypothesis

$$\begin{aligned} H_0 : \quad b &\leq 0 \\ H_a : \quad b &> 0 \end{aligned}$$

2. Select the level of significance at $\alpha = 0.05$

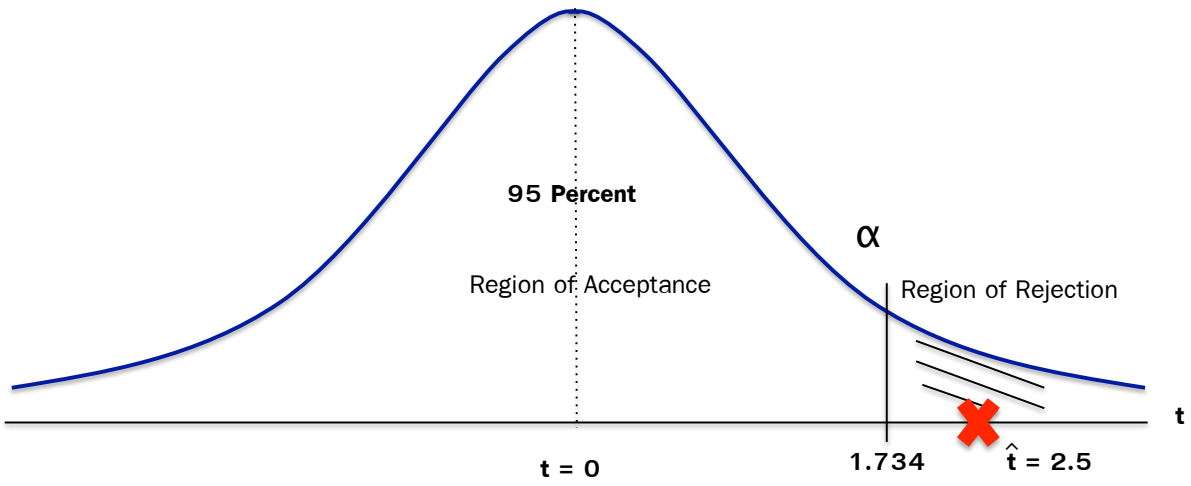
3. Calculate the value of $\hat{t}$ from the observed data

$$\hat{t} = \frac{\hat{b} - b}{se(\hat{b})} = \frac{5 - 0}{2} = 2.5$$

4. The critical t-value at 0.05 level of significance with df $= 18$ $(n - k = 20 - 2)$ is found to be 1.734.

5. Because $\hat{t} >$ critical t or $2.5 > 1.734$, the null hypothesis is rejected because, even at probability of $\alpha = 0.05$, we still observe that b is greater than 0. In other word, there is the relationship between variable X and Y as illustrated in Figure 3-9#

Figure 3-9: One-Tailed Hypothesis Testing



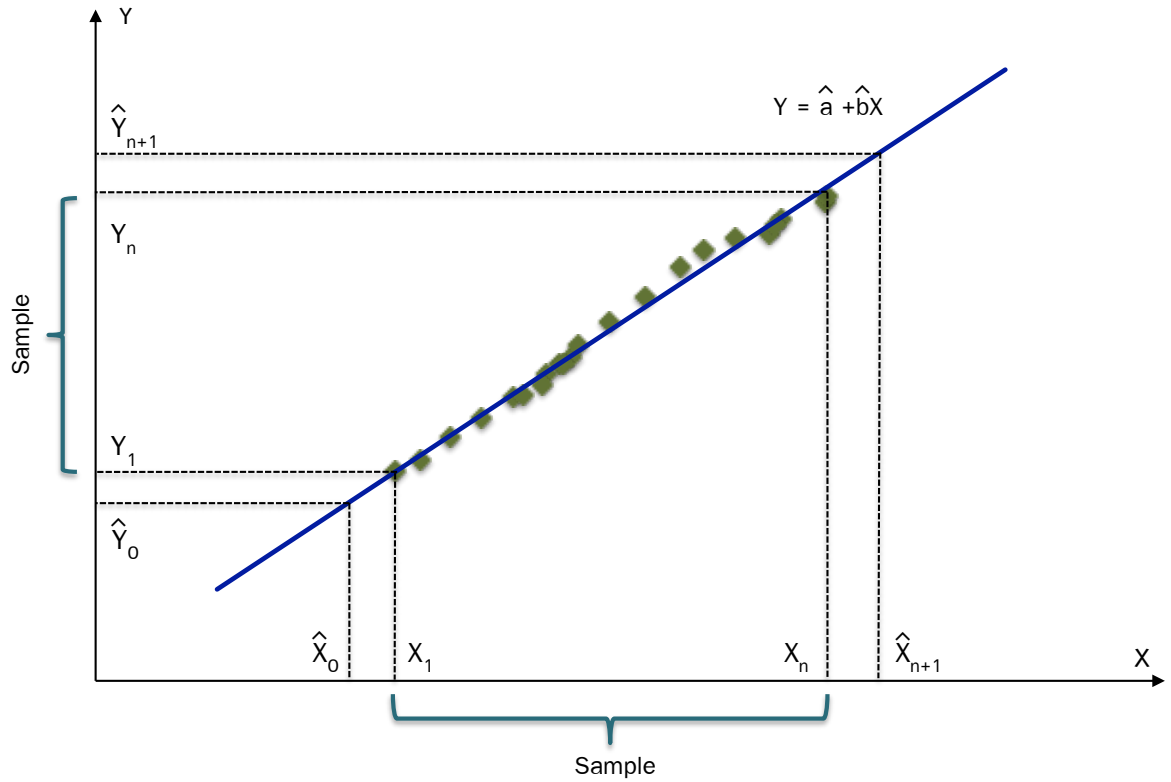
3.2.4 Mean Prediction

The main point of establishing the regression model is using the model to predict or estimate the value further from the currently observed data. As illustration, we may want to estimate the price of common stock of any company such that the future direction and magnitude of price adjustment are achieved.

Normally, the regression function obtained from the process of OLS could be used to predict the value of dependent variable in the future by plugging in the value of independent variable in the function. Figure 3-10 shows the situation in which the regression function is generated from the sample including the data from X_1 to X_n and Y_1 to Y_n. When we want to find the value of Y_0 or Y_{n+1}, the value of X_0 and X_{n+1} could be put into the function to find those value. Yet, we need to assume that this relationship has no *structural break*. Mathematically, it can be written as;

$$\begin{aligned} \text{Model:} \quad & Y_i = a + bX_i + u_i \\ \text{Regression Function:} \quad & \hat{Y}_i = \hat{a} + \hat{b}X_i \\ \text{Estimation:} \quad & E(Y_{n+1}|X_{n+1}) = a + bX_{n+1} \end{aligned}$$

Figure 3-10: Prediction



However, the above prediction may contain some error; thus, to alleviate this problem, statistician applies the confidence interval to predict $E(Y_{n+1}|X_{n+1})$ as;

$$\begin{aligned} \text{Expected Value:} \quad E(Y|X_{n+1}) &= a + bX_{n+1} \\ \text{Regression Function:} \quad Var(\hat{Y}_{n+1}) &= \sigma^2[\tfrac{1}{n} + \tfrac{(X_{n+1}-\bar{X})^2}{(X_i-\bar{X})^2}] \\ \text{t-distibution:} \quad t &= \tfrac{\hat{Y}_{n+1}-(a+bX_{n+1})}{se(\hat{Y}_{n+1})} \end{aligned}$$

The standard normal distribution Z is not the choice because the variance of error term σ_u^2 is unknown. In this case, t-distribution will have $n-2$ degree of freedom and the confidence interval can be written as;

$$P[-t_{\frac{\alpha}{2}} \leq \frac{\hat{Y}_{n+1} - (a + bX_{n+1})}{se(\hat{Y}_{n+1})} \leq t_{\frac{\alpha}{2}}] = 1 - \alpha \tag{3.26}$$

Rearrange Equation 3.26, we get;

$$P[\hat{Y}_{n+1} - t_{\frac{\alpha}{2}} se(\hat{Y}_{n+1}) \leq Y_{n+1} \leq \hat{Y}_{n+1} + t_{\frac{\alpha}{2}} se(\hat{Y}_{n+1})] = 1 - \alpha \tag{3.27}$$

or;

$$P[\hat{a} + \hat{b}X_{n+1} - t_{\frac{\alpha}{2}} se(\hat{Y}_{n+1}) \leq a + bX_{n+1} \leq \hat{a} + \hat{b}X_{n+1} + t_{\frac{\alpha}{2}} se(\hat{Y}_{n+1})] = 1 - \alpha \tag{3.28}$$

Figure 3-11 depicts the interval estimation in which red lines above and under regression line illustrate the interval estimation. Also, it can be seen that, the farther the prediction, the wider the confidence interval. In other word, as X_0 and $X_(n+1)$ deviate from $\bar{X}$, the confidence interval will get wider.

Example: Let $\hat{Y}_{n+1} = 14.4656$, $Var(\hat{Y}_{n+1}) = 0.3826$ or $se(\hat{Y}_{n+1}) = 0.6185$, and $X_{n+1} = 20$, where $n = 13$. Estimate $E(Y_{n+1}|X_{n+1})$ at 0.05 level of significance.

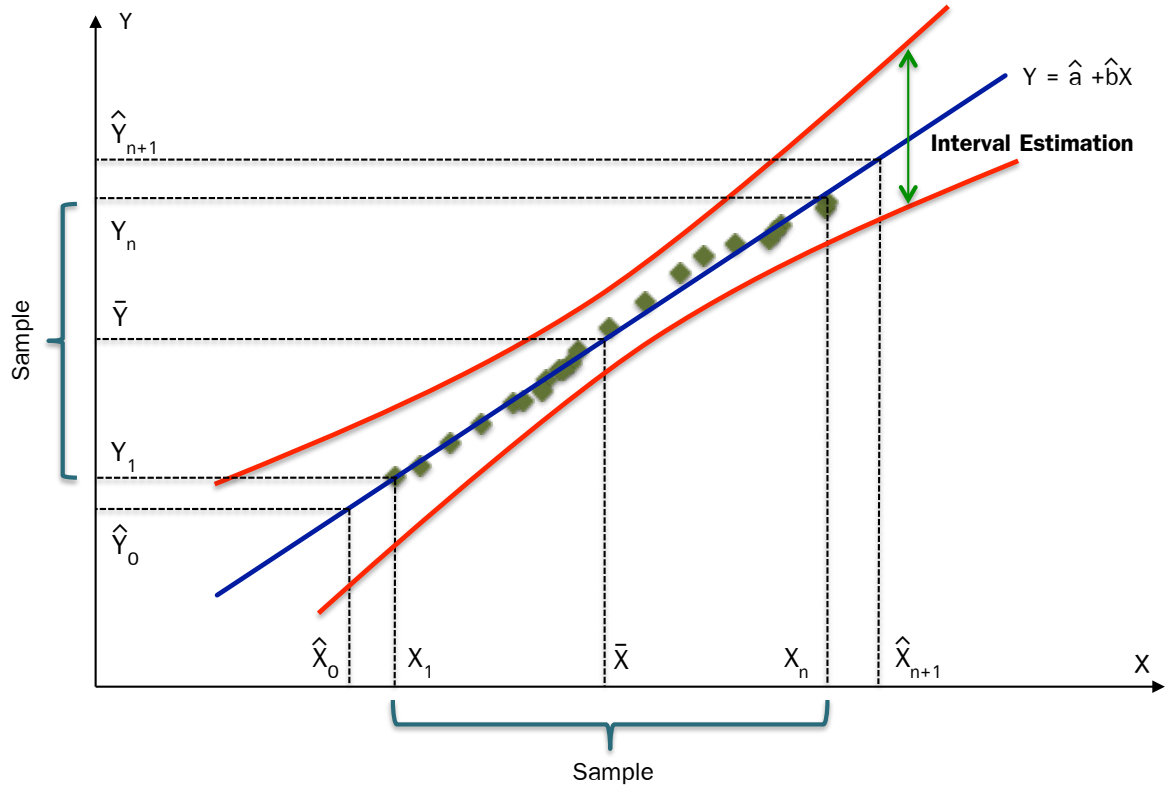
$$P[\hat{Y}_{n+1} - t_{\frac{\alpha}{2}} se(\hat{Y}_{n+1}) \leq E(Y_{n+1}|X_{n+1}) = Y_{n+1} \leq \hat{Y}_{n+1} + t_{\frac{\alpha}{2}} se(\hat{Y}_{n+1})] = 1 - \alpha$$

$$P[14.4656 - 2.201(0.6185) \leq E(Y_{n+1}|X_{n+1}) \leq 14.4656 + 2.201(0.6185))] = 1 - \alpha$$

$$P[13.1043 \leq E(Y_{n+1}|X_{n+1}) \leq 15.826] = 1 - \alpha$$

That is, we can estimate X_{n+1} in the 95% confidence interval from 13.1043 to 15.826#

Figure 3-11: Interval Estimation

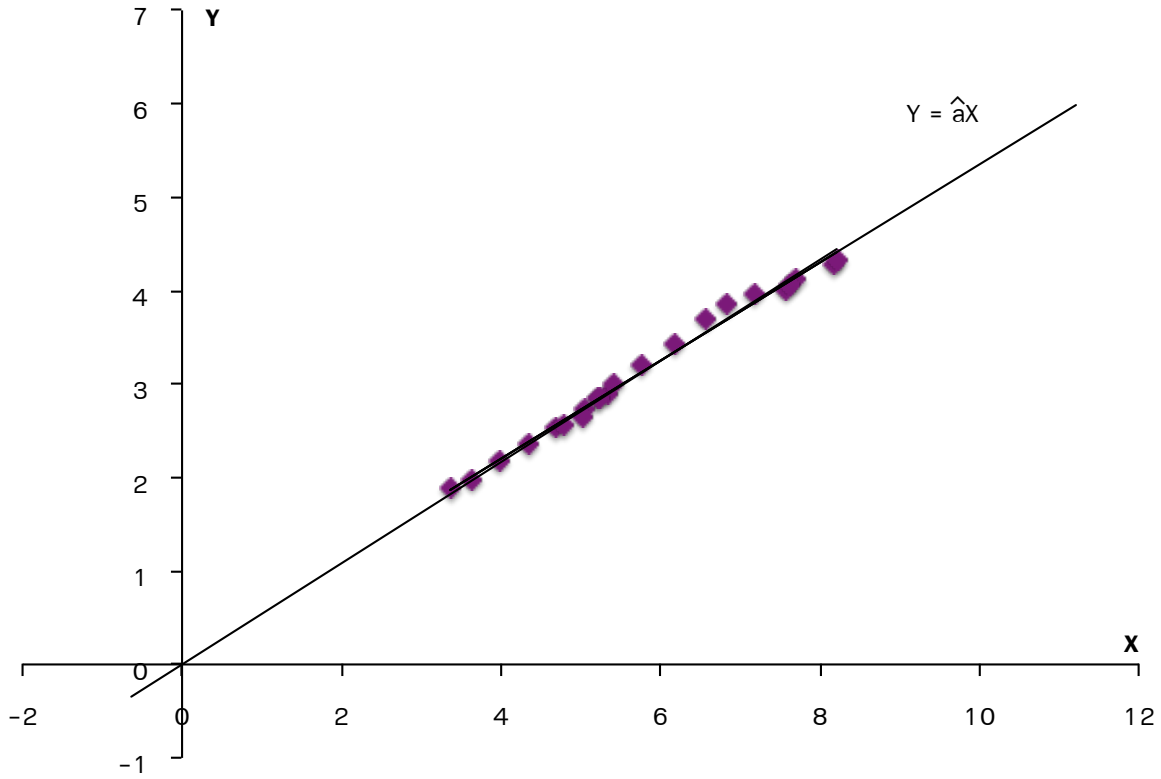


3.3 ADDITIONAL ISSUES OF REGRESSION MODEL

3.3.1 Regression through the origin

Apart from the regression function with vertical intercept, we can also build the model without the vertical intercept, namely with origin as the vertical intercept. Figure 3-12 exhibits the regression function through origin, obtained from OLS process.

Figure 3-12: Regression Function through Origin



Notwithstanding, due to the change in model, the estimation of error term in the function is also changed to;

$$\begin{aligned} \text{Model:} \qquad Y_i &= aX_i + u_i \\ \hat{u}_i &= Y_i - \hat{Y}_i = Y_i - aX_i \\ \sum \hat{u}_i^2 &= \sum (Y_i - \hat{Y}_i)^2 = (Y_i - aX_i)^2 \end{aligned}$$

Then, applying calculus to minimize the sum of error squared, we get the formula for estimator of a as;

$$\hat{a} = \frac{\sum X_i Y_i}{\sum X_i^2} \tag{3.29}$$

Other statistics are illuminated in Equation 3.30 and 3.31.

$$Var(\hat{a}) = \frac{\sigma_u^2}{\sum X_i^2} \tag{3.30}$$

$$Var(u) = \frac{\sum \hat{u}_i^2}{n-1} \tag{3.31}$$

The coefficient of determination, in this case, is not achieved in the same way as R^2 in simple regression model. Owing to the different model, the formulas for TSS, ESS, and RSS change. Hence, the coefficient of determination is calculated as

$$raw\ r^2 = \frac{(\sum X_i Y_i)^2}{\sum X_i^2 \sum Y_i^2} \tag{3.32}$$

Notice the following facts resulting from the regression function through origin.

1. $\sum \hat{u}_i X_i = 0$ but $\sum \hat{u}_i \neq 0$

2. The coefficient of determination might be negative, which conveys no meaningful interpretation.

In principle, if we do not have reasonable theory to back up the use of regression through origin, the simple regression model should be used. The *first reason* is, if the simple regression function is obtained, we could find out first whether the vertical intercept or a could be test for statistical significance and, if it is not, we then apply regression through origin. The *second reason* is that, if we apply regression through origin from the start and, in fact, it has the vertical intercept, that model will suffer the specification error which will be further studied in chapter 9.

3.3.2 Scaling and units of measurement

Scaling and units of measurement is the multiplication of observed data with some constant. The interesting point is that whether that process will affect the characteristics of estimators through OLS process.

Considering Table 3-3, when we multiply the set of data Y and X with w_1 and w_2 respectively, it means the model will be derived from different data. Thus, OLS estimation will result in the different estimates.

Table 3-3: Scaling and Units of Measurement

Model: $Y_i = \alpha + \beta X_i + u_i$	Model: $Y_i^* = \alpha^* + \beta^* X_i^* + u_i^*$
where	where w_1 and w_2 are constant
$x_i = X_i - \bar{X}$	$Y_i^* = w_1 Y_i$
$y_i = Y_i - \bar{Y}$	$X_i^* = w_2 X_i$
OLS:	OLS:
$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$	$\hat{\alpha}^* = \bar{Y}^* - \hat{\beta}^*\bar{X}^*$
$\hat{\beta} = \frac{\sum x_i y_i}{\sum x_i^2}$	$\hat{\beta}^* = \frac{\sum x_i^* y_i^*}{\sum x_i^{*^2}}$
$Var(\hat{\alpha}) = \frac{\sum X_i^2}{n\sum x_i^2}\sigma_u^2$	$Var(\hat{\alpha^*}) = \frac{\sum X_i^{*^2}}{n\sum x_i^{*^2}}\sigma_u^{*^2}$
$Var(\hat{\beta}) = \frac{\sigma_u^2}{\sum x_i^2}$	$Var(\hat{\beta^*}) = \frac{\sigma_u^{*^2}}{\sum x_i^{*^2}}$
$\hat{\sigma}_i^2 = \frac{\sum \hat{u}_i^2}{n-2}$	$\hat{\sigma}_i^{*^2} = \frac{\sum \hat{u}_i^{*^2}}{n-2}$
$R^2 = \frac{\sum \hat{y}_i^2}{\sum y_i^2}$	$R^{*^2} = \frac{\sum \hat{y}_i^{*^2}}{\sum y_i^{*^2}}$

However, it can be proved that;

$$\hat{\beta}^* = \frac{w_1}{w_2}\hat{\beta} \tag{3.33}$$

$$\hat{\alpha}^* = w_1\hat{\alpha} \tag{3.34}$$

$$Var(\hat{\alpha}^*) = w_1^2 Var(\hat{\alpha}) \tag{3.35}$$

$$Var(\hat{\beta}^*) = (\frac{w_1}{w_2})^2 Var(\hat{\beta}) \tag{3.36}$$

$$\hat{\sigma}_u^{*^2} = w_1^2\hat{\sigma}_u^2 \tag{3.37}$$

$$R^{*^2} = R^2 \tag{3.38}$$

3.3.3 Functional form

In previous regression analysis, the linear function is used as the initial model. Nonetheless, the functional form can be adjusted to incorporate elasticity and be suitable with the economic theory.

$$Y_i = f(X_i) + u_i \tag{3.39}$$

Consider the Cobb-Douglass production function with the following functional form.

$$Y_i = f(X_i) = \beta_1 X_1^{\beta_2} e^{u_i}$$

The above equation can be adjusted to linear function by logarithmic transformation as;

$$\begin{aligned} lnY_i &= ln\beta_1 + \beta_2 lnX_i + u_i \\ Y_i' &= \alpha + \beta_2 X_i' + u_i \end{aligned}$$

where;

$$\begin{aligned} Y_i' &= lnY_i \\ X_i' &= lnX_i \\ \alpha &= ln\beta_1 \end{aligned}$$

So, it can be seen that, after adjusting the function to be linear one, we can use OLS to estimate the value of α and β_2.

Example: Specify the Cobb-Douglass production function as;

$$Y_i = AK_i^{\alpha} L_i^{\beta} e^{u_i}$$

Transform the above function to linear one and find the capital and labour elasticity of output.

$$\begin{aligned} lnY_i &= lnA + \alpha lnK_i + \beta lnL_i + u_i \\ Y_i' &= A' + \alpha K_i' + \beta L_i' + u_i \end{aligned}$$

Consider the capital and labour elasticity of output.

$$\begin{aligned} \frac{\partial lnY_i}{\partial K} &= \frac{1}{Y}\frac{\partial Y}{\partial K} \\ \frac{\partial Y}{\partial K} &= \alpha A K_i^{\alpha-1} L_i^{\beta} e^{u_i} \\ &= \frac{\alpha}{K} A K_i^{\alpha} L_i^{\beta} e^{u_i} = \frac{\alpha}{K} Y \end{aligned}$$

$$\begin{aligned}
\therefore \frac{\partial lnY_i}{\partial K} &= \frac{1}{Y}\frac{\alpha}{K}Y = \frac{\alpha}{K} \\
\frac{\partial lnY_i}{\partial L} &= \frac{\beta}{L} \\
\frac{\partial Y_i}{\partial K}\frac{K}{Y} &= \alpha \\
\frac{\partial Y_i}{\partial L}\frac{L}{Y} &= \beta
\end{aligned}$$

Accordingly, if the estimates of α and β are obtained, we will know the capital and labour elasticity of output respectively.

Table 3-4: Other Functional Forms

Model	Equation	$\frac{dY}{dX}$	Elasticity $(\frac{dY}{dX}\frac{X}{Y})$
Linear	$Y_i = \beta_1 + \beta_2 X_i$	β_2	$\beta_2\frac{X}{Y}$
Double log-linear	$lnY_i = \beta_1 + \beta_2 lnX_i$	$\beta_2\frac{Y}{X}$	β_2
Log linear	$lnY_i = \beta_1 + \beta_2 X_i$	$\beta_2 Y$	$\beta_2 X$
Linear-log	$Y_i = \beta_1 + \beta_2 lnX_i$	$\beta_2\frac{1}{X}$	$\beta_2\frac{1}{Y}$
Reciprocal	$Y_i = \beta_1 + \beta_2\frac{1}{X_i}$	$-\beta_2\frac{1}{X^2}$	$-\beta_2\frac{1}{XY}$
Log-reciprocal	$lnY_i = \beta_1 - \beta_2\frac{1}{X_i}$	$\beta_2\frac{Y}{X^2}$	$\beta_2\frac{1}{X}$

Table 3-4 summarizes other functional forms and the application to find the slope and elasticity of that function. For linear form, if the variable X change by 1 unit, Y will change by β_2 unit. In double log-linear form, the coefficient β_2 represents the elasticity of Y to X; that is, if the variable X change by 1 percent, Y will change by β_2 percent.

In log-linear form, if the variable X change by 1 unit, Y will change by $100\beta_2$ percent. On the other hand, in linear-log form, if the variable X change by 1 unit, $ln(X)$ will change by $\frac{1}{X}$. Hence, it can be verified that if X change by 1%, Y will change by $0.01\beta_2$.

Finally, for reciprocal form, if the variable X change, Y will change in opposite direction. Also, if the variable X increase indefinitely, $\beta_2\frac{1}{X}$ will approach zero and Y will approach β_1. This model is mostly used to formulate the model for Phillips Curve.

Chapter 4

MULTIPLE REGRESSION MODEL

4.1 THE ESTIMATION OF MULTIPLE REGRESSION MODEL

From the previous chapter, the analysis of simple regression model involves only a single independent variable (X) to explain the behavior of dependent variable (Y). Nevertheless, practically, the movement of regressand may be well explained by many regressors. Consumption expenditure, for instance, may depends on wealth and the number of family member. In this chapter, additional explanatory variables are included in the model in order to enhance the explanatory power over the interested explained variable. The chapter, first, starts with the estimation of the regression coefficients in the model. Then, the properties of the estimators of those coefficients are discussed. Last but not least, the hypothesis testing for statistical significance is illustrated.

With the basics of the model in Chapter 3 which consists of two variables, namely dependent and independent variables, we would adjust that model to take more independent variables into account. As illustrated in the model comparison below, multiple regression model add X_{3i} into the simple regression model. For this case, it can be seen that there are three coefficients (β_1, β_2 and β_3) instead of two as in the simple model. In general, k denotes the amount of regression coefficients in the model, implying that k in simple regression model and in three-variable model are equal to 2 and 3, respectively.

$$\begin{aligned} \text{Simple Regression Model:} \quad Y_i &= \beta_1 + \beta_2 X_i + u_i \\ \text{Three-Varaible Regression Model:} \quad Y_i &= \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i \\ \text{Multiple Regression Model:} \quad Y_i &= \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \ldots + \beta_k X_{ki} + u_i \end{aligned}$$

Consider three-variable model. The meanings of each coefficient are different. β_1 identifies the value of the dependent variable if the two independent variables are zero. On the other hand, β_2 means that, as X_{2i} changes by one unit, Y would change by β_2 units, ceteris paribus. The same meaning is applicable to β_3; that is, as X_{3i} changes by one unit, Y would change by β_3 units, ceteris paribus. Also, β_2 and β_3 are called **partial regression coefficients**.

To build the multiple regression model, there are some assumptions required such that the consideration of the relationship among variables is possible. The reason is that if that set of assumptions is not satisfied, the estimators obtained from the model will lack desirable properties, that is, it is not BLUE. Those assumptions are as follows.

1. The model is linear in parameter.

2. Explanatory variables are fixed in repeated sampling.

3. Each explanatory variable is independent of the disturbance term, or

$$cov(u_i, X_{2i}) = cov(u_i, X_{3i}) = 0$$

4. The expectation of disturbance term is zero, or

$$E(u_i|X_{2i}, X_{3i}) = 0$$

5. The variance of disturbance term is constant; namely **homoscedasticity**, or

$$var(u_i) = \sigma_u^2$$

6. There is no **autocorrelation** among disturbance term, or

$$cov(u_i, u_j) = 0$$

7. There is be higher amount of parameter than the amount of estimator, or $n > k$.

8. There is no perfect **multicollinearity** among independent variables.

9. The model is correctly specified; namely no specification error.

In the subsequent chapters, we will consider the effect on the estimators when some of the above assumptions are not satisfied. Chapter 6 will tackle the case when there is multicollinearity among regressors. Chapter 7 will handle the case of heteroscedasticity; that is, when the variance of disturbance term is not constant. Then, Chapter 8 will deal with the problem occurring when there is autocorrelation among disturbance terms. Finally, in Chapter 9, the case when the model is wrongly specified will be discussed.

The next topic to consider is how we can estimate the parameter in the model. If the method of ordinary least square is used to minimize the sum of disturbance term squared ($\sum \hat{u}_i^2$), the procedures are as follows:

$$\begin{aligned} \text{From the model} \quad Y_i &= \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i \\ \text{Consider error term} \quad \hat{u}_i &= Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} - \hat{\beta}_3 X_{3i} \\ \text{Min} \quad \sum \hat{u}_i^2 &= (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} - \hat{\beta}_3 X_{3i})^2 \end{aligned}$$

With calculus, we can find the solution of estimators that make the disturbance term minimum which can be illustrated as

$$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2\bar{X}_2 - \hat{\beta}_3\bar{X}_3 \tag{4.1}$$

$$\hat{\beta}_2 = \frac{\sum_{i=1}^n y_i x_{2i} \sum_{i=1}^n x_{3i}^2 - \sum_{i=1}^n y_i x_{3i} \sum_{i=1}^n x_{2i}x_{3i}}{\sum_{i=1}^n x_{2i}^2 \sum_{i=1}^n x_{3i}^2 - (\sum_{i=1}^n x_{2i}x_{3i})^2} \tag{4.2}$$

$$\hat{\beta}_3 = \frac{\sum_{i=1}^n y_i x_{3i} \sum_{i=1}^n x_{2i}^2 - \sum_{i=1}^n y_i x_{2i} \sum_{i=1}^n x_{2i}x_{3i}}{\sum_{i=1}^n x_{2i}^2 \sum_{i=1}^n x_{3i}^2 - (\sum_{i=1}^n x_{2i}x_{3i})^2} \tag{4.3}$$

In this case, x and y denote the deviation of data from its mean. Furthermore, the variance of each estimator can be achieved by

$$var(\hat{\beta}_1) = [\frac{1}{n} + \frac{\bar{X}_{2i}^2 \sum_{i=1}^n x_{3i}^2 + \bar{X}_{3i}^2 \sum_{i=1}^n x_{2i}^2 - 2\bar{X}_{2i}\bar{X}_{3i} \sum_{i=1}^n x_{2i}x_{3i}}{\sum_{i=1}^n x_{2i}^2 \sum_{i=1}^n x_{3i}^2 - (\sum_{i=1}^n x_{2i}x_{3i})^2}]\sigma_u^2 \tag{4.4}$$

$$var(\hat{\beta}_2) = \frac{\sum_{i=1}^n x_{3i}^2}{\sum_{i=1}^n x_{2i}^2 \sum_{i=1}^n x_{3i}^2 - (\sum_{i=1}^n x_{2i}x_{3i})^2}\sigma_u^2 = \frac{\sigma_u^2}{\sum_{i=1}^n x_{2i}^2(1 - r_{23}^2)} \tag{4.5}$$

$$var(\hat{\beta}_3) = \frac{\sum_{i=1}^n x_{2i}^2}{\sum_{i=1}^n x_{2i}^2 \sum_{i=1}^n x_{3i}^2 - (\sum_{i=1}^n x_{2i}x_{3i})^2}\sigma_u^2 = \frac{\sigma_u^2}{\sum_{i=1}^n x_{3i}^2(1 - r_{23}^2)} \tag{4.6}$$

where

$$\widehat{var(u_i)} = \hat{\sigma}_u^2 = \frac{\sum_{i=1}^n \hat{u}_i^2}{n-3} \tag{4.7}$$

$$\sum_{i=1}^n \hat{u}_i^2 = \sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n y_i^2 - \beta_2 \sum_{i=1}^n y_i x_{2i} - \beta_3 \sum_{i=1}^n y_i x_{3i} \tag{4.8}$$

$$r_{23} = \frac{(\sum_{i=1}^n x_{2i}x_{3i})^2}{\sum_{i=1}^n x_{2i}^2 \sum_{i=1}^n x_{3i}^2} \tag{4.9}$$

The above estimators possess the following properties.

1. $\hat{u}_i$ is independent of X_{2i} X_{3i} and $\hat{Y}_i$

2. According to the above assumptions, it can be proved that the estimators obtained through ordinary least square will be best, linear and unbiased (BLUE).

3. As the correlation between two regressors X_{2i} and X_{3i} (r_{23}) approaches 1, the variance of the estimator of coefficient associated with both will increase, which, in turn, may result in an error of hypothesis testing.

4. If the variance of disturbance term (σ_u^2) is higher, or the sum of data in term of deviation from mean ($\sum x_{2i}^2$ or $\sum x_{3i}^2$) is lower, the variance of the estimators will escalate, negatively influencing the result from the hypothesis testing.

It is essential to realize the difference between *simple correlation* and *partial regression coefficient*. **Simple correlation** identifies the relationship between two variables; while **partial regression coefficient** determines the change in regressand as the regressor changes, given other things remain constant. Hence, the relationship obtained from simple correlation might include the influence of other variables.

Another crucial issue is **coefficient of determination:** R^2 of the multiple regression model. R^2 can be used to measure the ability of the model to explain the relationship between regressor and regressand because

$$R^2 = \frac{ESS}{TSS} = \frac{\sum \hat{y}^2}{\sum y_i^2} = 1 - \frac{\sum \hat{u}_i^2}{\sum y_i^2}$$

The inclusion of additional independent variable helps reduce (or at least not increase) the residuals of the model. As the error term incorporates all influence of other variables excluded from the model, some economists might overly include the additional explanatory variables in the model to get higher level of R^2. To prevent the problem, **adjusted coefficient of determination: adjusted** R^2, is invented and calculated by

$$\bar{R}^2 = 1 - \frac{\frac{\sum \hat{u}_i^2}{n-k}}{\frac{\sum y_i^2}{n-1}}$$

where n denotes the amount of data and k denotes the amount of parameter estimated in multiple regression model. In the formula, it is obvious that the sum of error term squared is divided by its degree of freedom to make it comparable with the variance of Y. In other word, the more explanatory variables included in the model, the more the model lose the degree of freedom; which, in turn, makes adjusted R^2 lower. The competition to add more independent variables will be in vain. Also, it can be shown that

$$\bar{R}^2 = \frac{1-k}{n-k} + \frac{n-1}{n-k} R^2 \text{ or } \bar{R}^2 \leq R^2$$

To compare R^2 from various models, we have to consider whether the number of observation used is equal and whether they share the same form of dependent variable. Otherwise, the comparison will be invalid. For example, if it is linear model where the dependent and independent variables are at level and the other model is log-linear model where the dependent and independent variables are logarithmically adjusted, we cannot compare the R^2 in one model with the other since the dependent variables are in the different form.

Accordingly, economists should be aware of the fact that, it is not necessary for the model to have the highest R^2 but, instead, to have valid statistical inference for estimators and logical statistical relationship between the variables in the model. That is, the estimators are efficient and the hypothesis testing can be correctly conducted. Consequently, the model with low R^2 is not necessarily undesirable.

4.2 HYPOTHESIS TESTING

4.2.1 Normality Assumption of the Random Disturbance Term

For hypothesis testing, as in Chapter 3, additional assumption has to be made to ease the process of hypothesis testing. That assumption is normal distribution of random disturbance term with the mean of zero and the constant variance of σ_u^2. Furthermore, since we do not know the true value of the variance of disturbance term, the t-distribution is applied to hypothesis testing. Specifically, the three estimators from the multiple regression model feature t-distribution in which the t-value for each estimator is calculated by

$$\hat{t} = \frac{\hat{\beta}_1 - \beta_1}{se(\hat{\beta}_1)} \tag{4.10}$$

$$\hat{t} = \frac{\hat{\beta}_2 - \beta_2}{se(\hat{\beta}_2)} \tag{4.11}$$

$$\hat{t} = \frac{\hat{\beta}_3 - \beta_3}{se(\hat{\beta}_3)} \tag{4.12}$$

and the degree of freedom of $n - 3$ since, to obtain the estimators, we lose 3 observation to estimate $\hat{\beta}_1$ $\hat{\beta}_2$ and $\hat{\beta}_3$ (for the case where there are three estimators in the model). In general, the degree of freedom in the multiple regression model is $n - k$.

4.2.2 Hypothesis Testing in Multiple Regression Model

For the hypothesis testing in multiple regression model, there are many types of test to be conducted; namely,

1. Hypothesis testing of an individual coefficient
2. Hypothesis testing of overall significance of the model
3. Hypothesis testing of equality between any pair of partial regression coefficients
4. Hypothesis testing of linear restriction of partial regression coefficients
5. Hypothesis testing of the structure of the model or **Chow test**

Various types of test are performed for different purpose; hence, economists have to realize those differences such that the right tool is applied to the problem. The following subsections discusses each test sequentially.

4.2.3 Hypothesis Testing of Individual Regression Coefficient

To test the hypothesis about the individual partial regression coefficient, t-value has to be calculated under null and alternative hypothesis that

$$H_0 : \beta_2 = 0 \text{ and } H_a : \beta_2 \neq 0$$

and, then, compare the acquired t-value with the critical t-value corresponding to degree of freedom and desired level of significance in the t-table. If the absolute of t-value is greater than the critical t-value, we can reject the null hypothesis and conclude that $\beta_2 \neq 0$. Moreover, we can set the hypotheses for other estimators including β_1, β_3 up to β_k by the same procedure.

4.2.4 Hypothesis Testing of Overall Significance

General F Test

Apart from the hypothesis about the individual partial coefficient, we can also test the hypothesis for statistical overall significance by setting the following null and alternative hypothesis that

$H_0 :$ $\beta_2 = \beta_3 = 0$
$H_a :$ β_2 or β_3, at least one of them, is not equal to zero

The above null hypothesis specifies that both partial regression coefficients are equal to zero or the second and the third regressor do not explain the behaviour of regressand. We can use **analysis of variance** to test the above hypothesis. From,

$$\begin{aligned} TSS &= ESS + RSS \\ \sum y_i^2 &= \sum \hat{y}_i^2 + \sum \hat{u}_i^2 \\ \sum y_i^2 &= \hat{\beta}_2 \sum y_i x_{2i} + \hat{\beta}_3 \sum y_i x_{3i} + \sum \hat{u}_i^2 \end{aligned}$$

If we want to find the mean sum of square of TSS, ESS and RSS, the degree of freedom associated with each sum of square has to be found. The degree of freedom of TSS is $n-1$ since we lose one degree of freedom in computing the mean of Y. Also, $n-k$ is the degree of freedom of RSS as we lose k degree of freedom to estimate k parameters. For the case stated above, k is equal to 3 since we have to estimate β_1 β_2 and β_3. Lastly, the degree of freedom of ESS can be calculated by subtracting RSS degree of freedom from ESS degree of freedom. Hence, ESS degree of freedom is k-1 or 2 in the above case. All results are concluded in Table 4-1.

Table 4-1: Analysis of variance from regression model with two regressors

Source of Deviation	Sum of Square (SS)	df	Mean Sum of Square (MS)
ESS	$\hat{\beta}_2 \sum y_i x_{2i} + \hat{\beta}_3 \sum y_i x_{3i}$	2	$\frac{ESS}{2}$
RSS	$\sum \hat{u}_i^2$	$n-3$	$\frac{RSS}{n-3}$
TSS	$\sum y_i^2$	$n-1$	

F-value ($\hat{F}$) is the ratio between the MS of ESS and the MS of RSS and can be computed by

$$\hat{F} = \frac{ESS/k-1}{RSS/n-k} \tag{4.13}$$

According to the assumption about the distribution of the error term (in the case of two regressors), we can find that

$$E(\frac{RSS}{df}) = E(\frac{\sum \hat{u}_i^2}{n-3}) = E(\hat{\sigma}_u^2) = \sigma_u^2 \tag{4.14}$$

Additionally, if the null hypothesis is true ($\beta_2 = \beta_3 = 0$), it can be shown that

$$E(\frac{ESS}{df}) = E(\frac{\hat{\beta}_2 \sum y_i x_{2i} + \hat{\beta}_3 \sum y_i x_{3i}}{2}) = \sigma_u^2 \tag{4.15}$$

If the hypothesis is true, Equation 4.14 and 4.15 illustrate that MS of TSS and RSS are equal to the true value of variance of the disturbance term. Nevertheless, if the hypothesis is not true, MS of two sum of square will not be equal to the variance of the disturbance term. In other word, the ESS will be so high that the Equation 4.15 is not valid.

The calculation of $\hat{F}$, thus, identify the difference between the two sources of variance in the model (ESS and RSS). If the values of two variances are equal, it implies that the equation of 4.14 and 4.15 or the null hypothesis is true. That is, two independent variables have no relationship with the dependent one. On the other hand, if the difference of two variances is statistically significantly large, it implies that the null hypothesis can be rejected.

Then we compare the obtained F-value with the critical F-value corresponding to degree of freedom of $k-1$ (ESS) and $n-k$ (RSS) and desired level of significance (α) in the F-table. If $\hat{F}$ is greater than $F_{\alpha,k-1,n-k}$, we can reject the null hypothesis. That is, there is at least one parameter not equal to zero.

Besides, we can calculate the value of $\hat{F}$ from the following formula.

$$\hat{F} = \frac{R^2/(k-1)}{(1-R^2)/(n-k)} \tag{4.16}$$

Test on Additional Varabiles

In addition to testing the statistical significance of model overall, we can use F-test to test the statistical significance of the incremental contribution of explanatory variable such as

$$\begin{aligned} \text{Model 1:} \quad C_t &= a + bY_t + u_t \\ \text{Model 2:} \quad C_t &= a + bY_t + dW_t + u_t \end{aligned}$$

Considering the two above models, it is obvious that the dependent variable is explained by the different independent variables. In the first model, income (Y) is the only explanatory variable for the consumption. In the second model, yet, wealth (W) and income (Y) help explain the behavior of consumption expenditure. The question is whether this additional variable, namely wealth, contribute much enough explanatory power to the model. In general, these models will be built as follows:

$$\text{Restricted Model:} \quad Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \ldots + \beta_r X_{rt} + u_t$$

$$\begin{aligned} \text{Unrestricted Model:} \quad Y_t &= \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \ldots + \beta_r X_{rt} + \beta_{r+1} X_{r+1,t} \\ &\quad + \ldots + \beta_k X_{kt} + u_t \end{aligned}$$

Restricted model is the model before the additional explanatory variables are included and it contains r parameters to be estimated. **Unrestricted model** is the model after the additional explanatory variables are included and those variables include $X_{r+1,t}$ $X_{r+2,t}$ up to X_{kt}.

Due to this construction of model, we can set the hypothesis of whether marginal explanatory variables included are equal to zero by

$$\begin{aligned} H_0 &: \ \beta_{r+1} = \beta_{r+2} = \ldots = \beta_k = 0 \\ H_a &: \ \beta_{r+1} \text{ or } \beta_{r+2} \text{ up to } \beta_k, \text{ at least one of them, is not zero} \end{aligned}$$

If the null hypothesis is true, it means the marginal regressors added to the model lack the explanatory power over the regressand. On the other hand, if we can reject the null hypothesis, it means that at least one variable added to the model possess the explanatory power. To reach either of above result, we need to compute the value of $\hat{F}$ to consider the variance of the restricted and unrestricted model and find out whether

there is statistically significant difference between them. If it is so, the null hypothesis can, then, be rejected. The value of $\hat{F}$ in this case is calculated by

$$\hat{F} = \frac{(ESS_U - ESS_R)/\text{no. of new regressors}}{RSS_U/(n - \text{no. of parameters in the new model})} = \frac{(ESS_U - ESS_R)/(k-r)}{RSS_U/(n-k)} \tag{4.17}$$

where
n is the number of observation used to estimate the model
k is the number of parameter to be estimated in the unrestricted model
r is the number of parameter to be estimated in the restricted model

Also, the value of $\hat{F}$ can be calculated by using coefficient of determination as

$$\hat{F} = \frac{ESS_U - ESS_R/(k-r)}{RSS_u/n-k} = \frac{(R_U^2 - R_R^2)/(k-r)}{(1-R_U^2)/(n-k)} \tag{4.18}$$

where

R_U^2 is coefficient of determination of unrestricted model
R_R^2 is coefficient of determination of restricted model

Then we compare the calculated F-value ($\hat{F}$) with the critical F-value ($F_{\alpha,k-r,n-k}$) corresponding to the first and second degree of freedoms of $k-r$ and $n-k$, respectively and desired level of significance in the F-table. If $\hat{F}$ is greater than $F_{\alpha,k-r,n-k}$, we can conclude that the additional variables included in the model statistically significantly posses the explanatory power.

4.2.5 Hypothesis Testing of Equality between Two Partial Coefficients

To test the hypothesis of equality between any pair of partial coefficients, the null and alternative hypotheses are set as

H_0 : $\beta_2 = \beta_3$ or $\beta_2 - \beta_3 = 0$
H_a : $\beta_2 \neq \beta_3$ or $\beta_2 - \beta_3 \neq 0$

According to the set of hypothesis, we can adjust the distribution of the difference to be t-distribution by,

$$\hat{t} = \frac{\hat{\beta}_2 - \hat{\beta}_3 - (\beta_2 - \beta_3)}{se(\hat{\beta}_2 - \hat{\beta}_3)} \tag{4.19}$$

where

$$se(\hat{\beta}_2 - \hat{\beta}_3) = \sqrt{var\hat{\beta}_2 + var\hat{\beta}_3 - 2cov(\hat{\beta}_2, \hat{\beta}_3)}$$

Nevertheless, when we specify the null hypothesis that the difference of two coefficients is zero, we can reduce the term in Equation 4.19 to

$$\hat{t} = \frac{\hat{\beta}_2 - \hat{\beta}_3 - (0)}{se(\hat{\beta}_2 - \hat{\beta}_3)} = \frac{\hat{\beta}_2 - \hat{\beta}_3}{se(\hat{\beta}_2 - \hat{\beta}_3)}$$

To reject or not to reject the null hypothesis, the criterion used is similar to other t-tests. That is, if $|\hat{t}|$ is greater than the critical t-value with the degree of freedom $n-k$, we can reject the null hypothesis and conclude that the two coefficients are not equal.

4.2.6 Hypothesis Testing of Linear Restriction of Parameters in the Model

To test the linear restriction of parameters in the model, first, we can categorize the test into two types. The first type applies t-distribution while the second one uses F-distribution for testing the model with more than two variables.

For the test the linear restriction of model with two regressors, we can use the test of equality used in previous subsection with little adjustment which is

$$\begin{aligned} H_0 &: \quad \beta_2 + \beta_3 = 1 \\ H_a &: \quad \beta_2 + \beta_3 \neq 1 \end{aligned}$$

The value of $\hat{t}$ is computed by

$$\hat{t} = \frac{\hat{\beta}_2 + \hat{\beta}_3 - (\beta_2 + \beta_3)}{se(\hat{\beta}_2 + \hat{\beta}_3)} \tag{4.20}$$

where

$$se(\hat{\beta}_2 + \hat{\beta}_3) = \sqrt{var\hat{\beta}_2 + var\hat{\beta}_3 + 2cov(\hat{\beta}_2, \hat{\beta}_3)}$$

In this particular case, we can specify that the value of $\hat{t}$ can be calculated, given the null hypothesis, by

$$\hat{t} = \frac{\hat{\beta}_2 + \hat{\beta}_3 - 1}{se(\hat{\beta}_2 + \hat{\beta}_3)}$$

The criterion for rejecting null hypothesis is the same as the previous subsection.

For the test the linear restriction of model with more than two regressors, to simplify the problem, we use the model with two regressors. Yet, the alternative way of test, namely F-test, will be shown. Also, this can be further adapted to the case where there is more than two regressors.

Consider Cobb-Douglas production function.

$$\begin{aligned} Y_t &= AL_t^{\beta_2} K_t^{\beta_3} e^{u_t} \\ lnY_t &= lnAL_t^{\beta_2} K_t^{\beta_3} e^{u_t} \\ lnY_t &= \beta_1 + \beta_2 lnL_t + \beta_3 lnK_t + u_t \end{aligned}$$

Sometimes, we would like to know whether the production function features the constant return to scale; namely, whether the sum of β_2 and β_3 is equal to one.

From the above model, we might impose some restrictions to adjust the function which is

$$\begin{aligned} lnY_t &= \beta_1 + \beta_2 lnL_t + \beta_3 lnK_t + u_t \\ \text{Let}\beta_2 &= 1 - \beta_3 \\ lnY_t &= \beta_1 + (1-\beta_3) lnL_t + \beta_3 lnK_t + u_t \\ lnY_t &= \beta_1 + lnL_t + \beta_3 (lnK_t - lnL_t) + u_t \\ lnY_t - lnL_t &= \beta_1 + \beta_3 (lnK_t - lnL_t) + u_t \end{aligned}$$

$$ln\frac{Y_t}{L_t} = \beta_1 + \beta_3 (ln\frac{K_t}{L_t}) + u_t \tag{4.21}$$

Equation 4.21 is called restricted least square. That is, we transform the model with three parameters to the model with two parameters by imposing some restrictions. Hence, we can set the hypothesis as

$$\begin{aligned} H_0 &: \ \beta_2 + \beta_3 = 1 \\ H_a &: \ H_0 \text{ is false} \end{aligned}$$

The value of $\hat{F}$ is computed by,

$$\hat{F} = \frac{(RSS_R - RSS_U)/m}{RSS_U/n-k} = \frac{R_U^2 - R_R^2/m}{1 - R_U^2/n-k} \tag{4.22}$$

where
m is the number of restriction used in the model
k is the number of parameter to be estimated in the unrestricted model
n is the number of data
R_U^2 is the coefficient of determination in the unrestricted model
R_R^2 is the coefficient of determination in the restricted model

Nonetheless, if the regressants of the two models are in the different form, we cannot compute F-value from the formula that uses R^2 since the meanings of R^2 will be different and incomparable.

The value of $\hat{F}$ obtained would identify the ratio of the difference of unexplained part in both models to the unexplained part of unrestricted model. If the ratio is statistically significantly small (or lower the critical F-value obtained from the F-table at $F_{\alpha,m,n-k}$), the null hypothesis cannot be rejected. In other word, the restrictions imposed can be practically used.

4.2.7 Hypothesis Testing of Structure of the Regression Model: The Chow Test

To build the regression model, sometimes, we experience the **structural change**, the situation in which the data we study may change according to various external factors. For instance, suppose we want to study the relationship between savings and national income in one country between 1971 and 2013. During that period, terrorism occurred in 2001, altering the saving and consumption behaviour of the people in the country. To verify the theory, we can test whether the set of data statistically significantly faces the structural change. In other word, we can test to decide whether we should separate the model to explain the behaviour form 1971 to 2001 and the behaviour from 2002 to 2013.

The test of structural change can help answer the above question. Normally, when we have two set of data and wonder whether we should construct the model separately for each set of data or treat it as the same set of data.

$$\begin{aligned} n_1 = 31\ (1971-2001): &\quad S_t &= \lambda_1 + \lambda_2 Y_t + u_t \\ n_2 = 12\ (2002-2013): &\quad S_t &= \delta_1 + \delta_2 Y_t + u_t \\ n_1 + n_2 = 43\ (1971-2013): &\quad S_t &= \gamma_1 + \gamma_2 Y_t + u_t \end{aligned}$$

From the model above, the crucial question is if there is the difference between the first model and second model that use the data from 1971 to 2001 and from 2002 to 2013 respectively. If there is no difference, we should construct the model by using all data from 1971 to 2013. We call the first and second models unrestricted model and call the third model restricted model.

Here, we can set the null hypothesis to be "there is no structural change" or "the parameters in the three models are the same". The alternative hypothesis can be "there is structural change" or "the parameters in the three models are not the same".

$$\begin{aligned} H_0: &\quad \lambda_1 = \gamma_1 = \delta_1 \text{ and } \lambda_2 = \gamma_2 = \delta_2 \\ H_a: &\quad \lambda_1 \neq \gamma_1 \neq \delta_1 \text{ and } \lambda_2 \neq \gamma_2 \neq \delta_2 \end{aligned}$$

We can compute the value of $\hat{F}$ by

$$\hat{F} = \frac{(RSS_R - RSS_U)/k}{RSS_U/n_1 + n_2 - 2k} = \frac{(RSS_R - RSS_1 - RSS_2)/k}{RSS_1 + RSS_2/n_1 + n_2 - 2k} \tag{4.23}$$

where

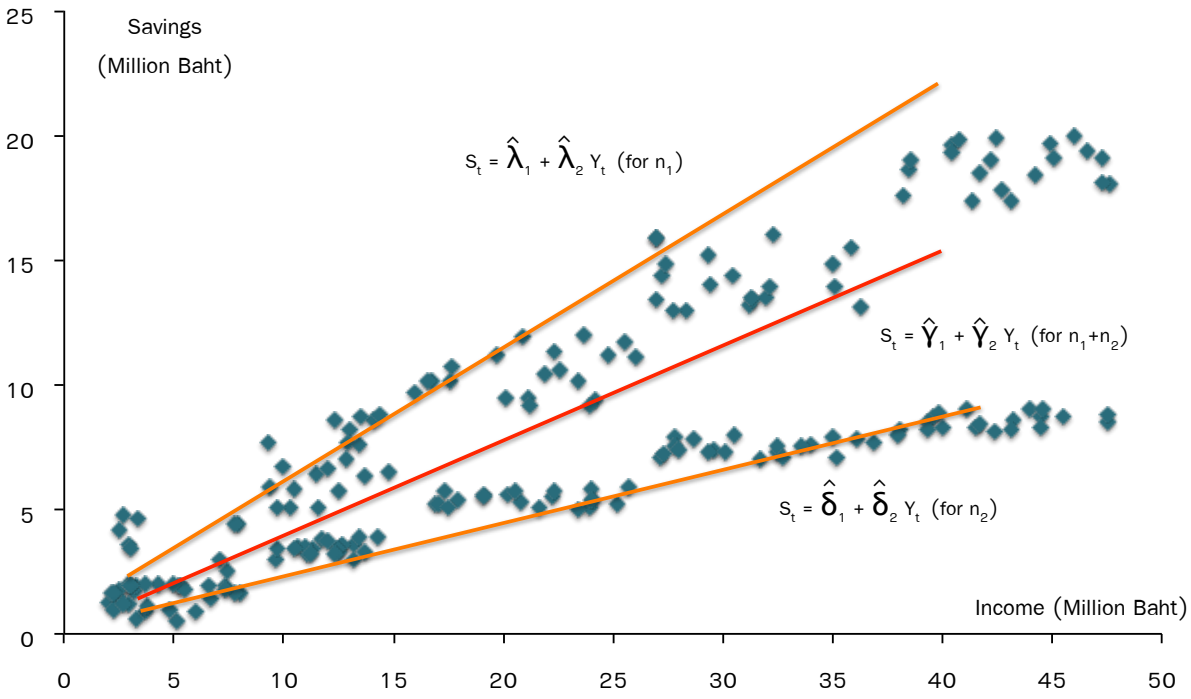
$RSS_U = RSS_1 + RSS_2$
k is the number of parameter in unrestricted model
n_1 is the number of the data in the first set
n_2 is the number of the data in the second set

RSS_R is the residual sum of square of the restricted model
RSS_1 is the residual sum of square of the first restricted model
RSS_2 is the residual sum of square of the second restricted model
RSS_U is the sum of RSS of the first and second unrestricted models

If the value of $\hat{F}$ is greater than F_{α,k,n_1+n_2-2k}, that means there is the statistically significant difference between unexplained part of the two model. Hence, we reject the null hypothesis and conclude that there is the structural change; namely, we should separate the data into two set and use separate model for each.

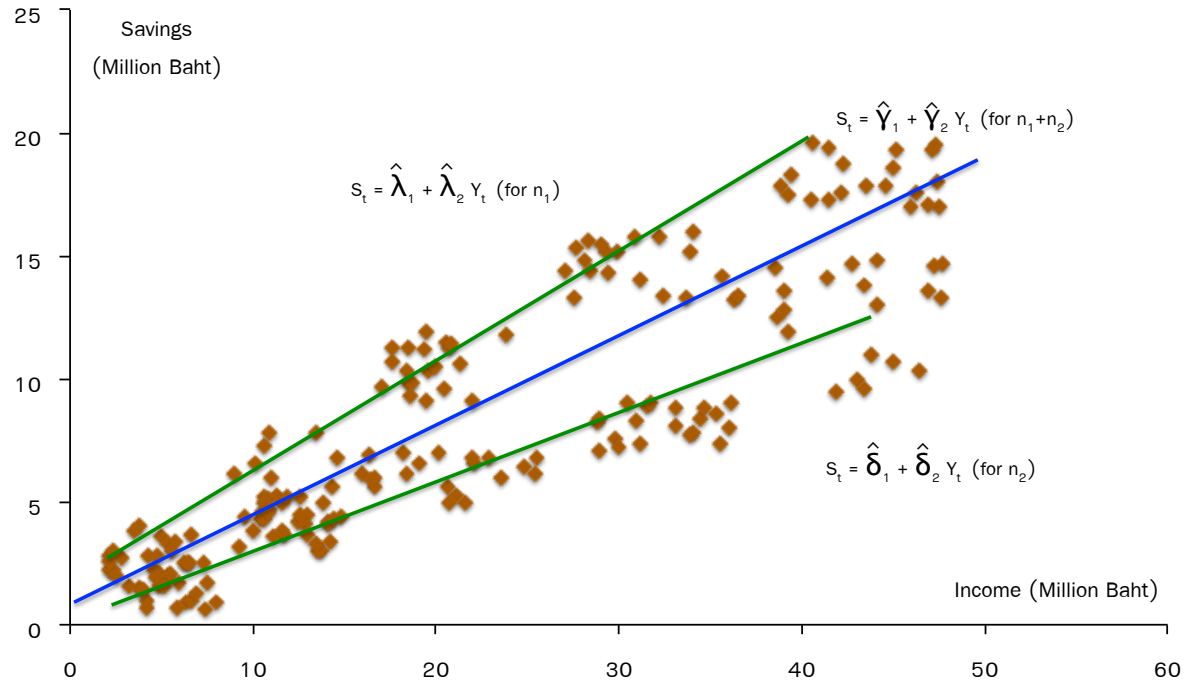
Figure 4-1 is the example of the scatter of data. It seems that the data separately scatter into two groups. With naked eye, it is difficult to identify whether we should separately build the model for two different groups of data. In this case, the calculation of statistic value and the comparison between the test statistic and the critical value would determine which model should be used.

Figure 4-1: Savings and national income from 1970 to 2013: the null hypothesis is rejected



However, when the scatter appears to be Figure 4-2, we might reconsider whether there is structural change. If $\hat{F}$ is less than F_{α,k,n_1+n_2-2k}, that means the unexplained parts in the two model are not statistically significantly different. Hence, we cannot reject the null hypothesis and conclude that there is no structural change; namely a single model should be used to incorporate the whole data set.

Table 4-2: Savings and national income from 1970 to 2013: the null hypothesis is not rejected



Chapter 5

Dummy Variable

5.1 THE IMPORTANCE OF DUMMY VARIABLE IN REGRESSION MODEL

In the previous chapters, the data used to build the regression model is **quantitative data**, the numerical data able to determine the magnitude of that data, such as price, income, interest rate and so forth. Nevertheless, it is possible that the data of interest, such as gender, nationality, religion and so forth, cannot be measured in numerical term. The data with this characteristic is called **quantitative data**. For instance, economists might be interested in the relationship between consumption expenditure and the state of marriage. In this chapter, the qualitative data is used to construct the regression model, and we call the variable representing qualitative data **dummy variable** or **qualitative variable**.

Generally, the property of independent variable would influence the dependent one. To deal with the qualitative regressand, the dummy variable is specified to take on the value of either 0 or 1. Specifically, if the data features the characteristic of interest, the value of the dummy variable is one; otherwise, the value is zero. Consider Table 5-1 which exhibits the data of consumption expenditure, national income and dummy variable (taking on the value of 0, if there no war in that year, and 1 if otherwise). From the Table 5-1, the war period is from 1941 to 1945.

As we stipulate the value of dummy variable to be either 0 or 1, this set of data can be used to construct the regression model and estimate the parameters to determine the relationship between existence of war and the dependent variable.

Table 5-1: Consumption expenditure and national income of one country (million U.S. dollars)

Year	Consumption Expenditure (C)	National Income (Y)	Dummy Variable (D)
1930	5.1	10.2	0
1931	5.4	11.2	0
⋮	⋮	⋮	⋮
1940	6.8	13.2	0
1941	2.4	5.2	1
1942	2.3	5.5	1
1943	2.7	5.3	1
1944	2.8	6.2	1
1945	2.6	4.6	1
1946	7.3	13.3	0
1947	7.5	14.5	0
⋮	⋮	⋮	⋮
1960	9.8	17.9	0

5.2 THE INTERPRETATION OF DUMMY VARIABLE

When the dummy variable is applied to the data set to build the model, the effect on the regression model is different from the normal model. That is, we can create a single model and the results obtained will be separated between the case where the situation of interest occurs and the case where that situation does not occur. To be specific, the case with the occurrence of that situation might possess the different value of vertical intercept or different value of slope or both from the case without that situation.

To illustrate how the application of dummy variable to the model affects the value of **vertical intercept**, we can set the initial models as

$$\begin{aligned} \text{Non-war period} \quad C_t &= \alpha_1 + \beta Y_t + u_t \\ \text{War-period} \quad C_t &= \alpha_2 + \beta Y_t + u_t \end{aligned}$$

The above models depict the situation where the two sets of data are separately considered. The application of dummy variable enables us to build the single model and attain the different results according which one we are interested in. That is, if we are interested in the war period, we assume the number of variable to be 1; and if we are interested in the non-war period, we assume the number of variable to be 0. The model could be formed as

$$D_t = \begin{cases} 1 & \text{War period} \\ 0 & \text{Non-war period} \end{cases}$$

$$\begin{aligned} C_t &= \alpha + \gamma D_t + \beta Y_t + u_t \\ \text{OLS:} \quad \hat{C}_t &= \hat{\alpha} + \hat{\gamma} D_t + \hat{\beta} Y_t \\ \text{Non-war period or D = 0:} \quad \hat{C}_t &= \hat{\alpha} + \hat{\beta} Y_t \\ \text{War period or D = 1:} \quad \hat{C}_t &= \hat{\alpha} + \hat{\gamma} + \hat{\beta} Y_t \end{aligned}$$

It can be seen that $\hat{\gamma}$ is the estimator of the parameter associated with dummy variable that helps separate the two cases. The vertical intercept will be $\hat{\alpha} + \hat{\gamma}$ in the war period; whereas, $\hat{\alpha}$ in the non-war period.

On the other hand, to form the model of which the **slope** is affected by the application of dummy variable, we can set the initial models as,

$$\begin{aligned} \text{Non-war period} \quad C_t &= \alpha + \beta_1 Y_t + u_t \\ \text{War period} \quad C_t &= \alpha + \beta_2 Y_t + u_t \end{aligned}$$

In this case, the application of dummy variable could help establish the single model that incorporates the two cases above and the results will be different slopes for different situations. That single model is formed by

$$\begin{aligned} C_t &= \alpha + \beta Y_t + \delta (D_t Y_t) + u_t \\ \text{OLS:} \quad \hat{C}_t &= \hat{\alpha} + \hat{\beta} Y_t + \hat{\delta} D_t Y_t \\ \text{Non-war period or D = 0:} \quad \hat{C}_t &= \hat{\alpha} + \hat{\beta} Y_t \\ \text{War period or D = 1:} \quad \hat{C}_t &= \hat{\alpha} + (\hat{\beta} + \hat{\delta}) Y_t \end{aligned}$$

If it is the war period, the model will result in the slope of $\hat{\alpha} + \hat{\delta}$ and if it is not the war period, the slope will be $\hat{\alpha}$. Hence, the inclusion of the term $\delta(D_t Y_t)$ enables the model to separate the cases of interest through different term of slope.

Moreover, if we want the application of dummy variable to separate the two cases trough both different intercepts and slopes, we may start the initial model as

$$\begin{aligned} \text{Non-war period} \quad C_t &= \alpha_1 + \beta_1 Y_t + u_t \\ \text{War period} \quad C_t &= \alpha_2 + \beta_2 Y_t + u_t \end{aligned}$$

The model with dummy variable to reflect the difference between two cases can be built as

$$\begin{aligned} C_t &= \alpha + \gamma D_t + \beta Y_t + \delta (D_t Y_t) + u_t \\ \text{OLS:} \quad \hat{C}_t &= \hat{\alpha} + \hat{\gamma} D_t + \hat{\beta} Y_t + \hat{\delta} D_t Y_t \\ \text{Non-war period or D = 0:} \quad \hat{C}_t &= \hat{\alpha} + \hat{\beta} Y_t \\ \text{War period or D = 1:} \quad \hat{C}_t &= \hat{\alpha} + \hat{\gamma} + (\hat{\beta} + \hat{\delta}) Y_t \end{aligned}$$

After the construction of model as stated above, the next step is to determine whether the estimators in the model are statistically significant and how much the dummy variable contributes to the model in term of explanatory power. The hypothesis test will be applied to both γ, which identify the difference in vertical intercept, and δ, which identify the difference in slope.

Case 1: Vertical Intercept
$H_0: \ \gamma = 0$
$H_a: \ \gamma \neq 0$

Case 2: Slope
$H_0: \ \delta = 0$
$H_a: \ \delta \neq 0$

The tests above are conducted separately. For the case of γ, if the null hypothesis cannot be rejected, it means the existence of war does not statistically significantly result in the difference in vertical intercepts. On the other hand, for the case of δ, if the null hypothesis can be rejected, it means the existence of war statistically significantly results in the difference in the slopes of consumption expenditure with respect to national income.

5.3 APPLICATION OF DUMMY VARIABLE IN ECONOMICS

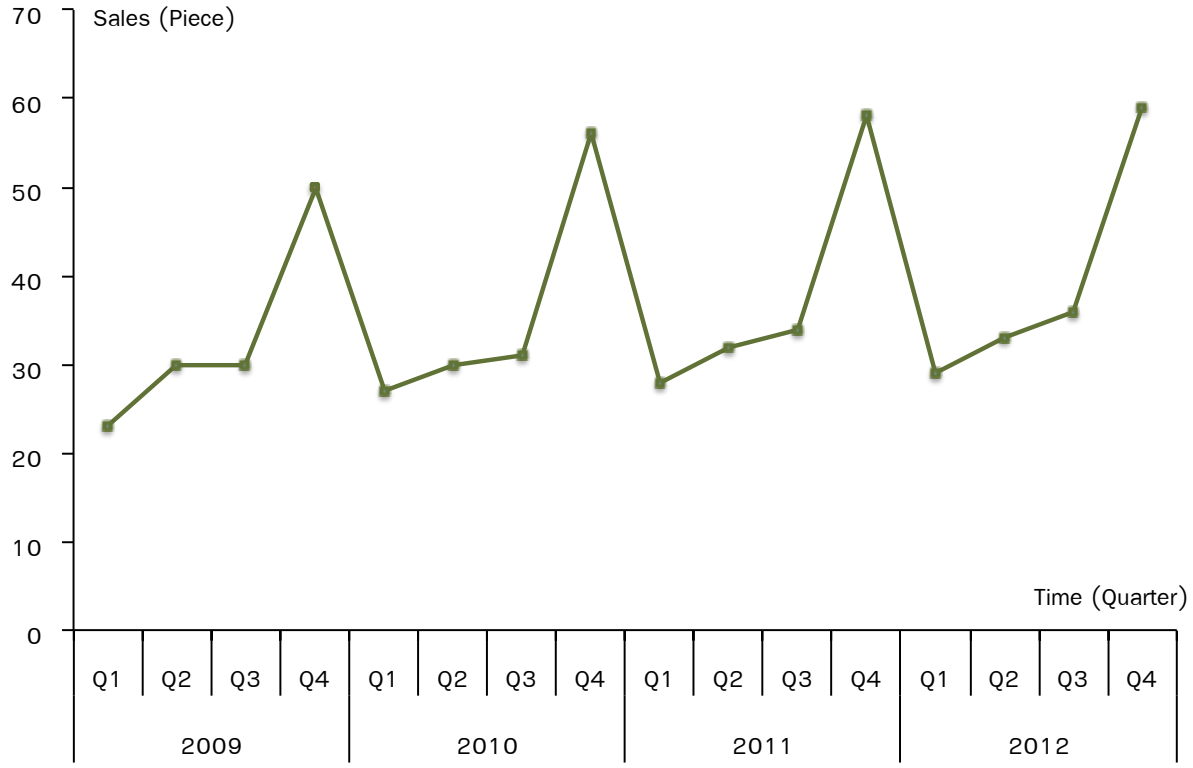
5.3.1 Seasonal Problem

Sometimes the set of data changes according to the season. For instance, the sale of the company in the first, second and third quarter might be similar; yet, the sale in the fourth quarter is evidently higher probably due to the end of the year period in which people purchase the gifts for one another. Table 5-2 illustrates the case in which the data set suffers the seasonal problem and it is depicted in Figure 5-1.

Table 5-2: The quarterly sale data of one department store

Quarter	Sale (Million Baht)	Quarter	Sale (Million Baht)
2009Q1	23	2011Q1	28
2009Q2	30	2011Q2	32
2009Q3	30	2011Q3	34
2009Q4	50	2011Q4	58
2010Q1	27	2012Q1	29
2010Q2	30	2012Q2	33
2010Q3	31	2012Q3	36
2010Q4	56	2012Q4	59

Figure 5-1: The quarterly sale of one department store



With the application of dummy variable, we may separate the data set into two parts. Particularly, we may let the dummy variable (Q_4) to take on the value of 0 if the sale is in the period from quarter 1 to 3, and take on the value of 1 if the sale is in quarter 4. The model could be established as follows

$$Q_4 = \begin{cases} 1 & \text{Sale in the forth quarter} \\ 0 & \text{Sale in other quarters} \end{cases}$$

$$\begin{aligned} & & S_t &= \alpha + \beta X_t + \gamma Q_4 + \delta Q_4 X_t + u_t \\ & OLS: & \hat{S}_t &= \hat{\alpha} + \hat{\beta} X_t + \hat{\gamma} Q_t + \hat{\delta} D_t X_t \\ \text{Other quarters where} & D = 0: & \hat{S}_t &= \hat{\alpha} + \hat{\beta} X_t \\ \text{Quarter 4 where} & D = 1: & \hat{S}_t &= \hat{\alpha} + \hat{\gamma} + (\hat{\beta} + \hat{\delta}) X_t \end{aligned}$$

5.3.2 Interaction Effect of Dummy Variables

Previously, only one dummy variable is included in the model. If there are two dummy variables to explain the behaviour of the dependent variable, different form of model may be set up. Suppose we are considering the consumption expenditure in the country and would like to use gender and the state of marriage of sample as two dummy variables to explain the consumption behaviour. From this example, the model can be built as

$$C_i = \alpha_1 + \alpha_2 D_{2i} + \alpha_3 D_{3i} + \beta Y_i + u_i \tag{5.1}$$

$$C_i = \alpha_1 + \alpha_2 D_{2i} + \alpha_3 D_{3i} + \alpha_4 D_{2i} D_{3i} + \beta Y_i + u_i \tag{5.2}$$

where

C_i is the consumption expenditure per capita
Y_i is the income per capita

$$D_{2i} = \begin{cases} 1 & \text{Female} \\ 0 & \text{Male} \end{cases}$$

$$D_{3i} = \begin{cases} 1 & \text{Married} \\ 0 & \text{Single} \end{cases}$$

The difference between Equation 5.1 and 5.2 is that, in Equation 5.2, there is **interaction dummy** to take into account the fact that married woman may have consumption different behaviour from other groups of people. Usually, this fact is reasonable since, as the married woman, they tend to incur additional expenditures like child care and medical care. Comparing to Equation 5.1, there is no variable to obviously incorporate this joint effect of two characteristics of interest. In equation 5.1, we can categorize our sample into 4 groups: married woman, married man, single woman and single man.

$$\begin{aligned} E(C_i|D_{2i}=1, D_{3i}=1) &= \alpha_1 + \alpha_2 + \alpha_3 + \beta Y_i \\ E(C_i|D_{2i}=0, D_{3i}=1) &= \alpha_1 + \alpha_3 + \beta Y_i \\ E(C_i|D_{2i}=1, D_{3i}=0) &= \alpha_1 + \alpha_2 + \beta Y_i \\ E(C_i|D_{2i}=0, D_{3i}=0) &= \alpha_1 + \beta Y_i \end{aligned}$$

It can be seen that, the effect of gender is determined by the term α_2 (without consideration on state of marriage). Also, the effect of state of marriage is determined by the term α_3 (without consideration on gender). The fact the married woman has different expenditure from the single man is captured by $\alpha_2 + \alpha_3$. On the other hand, consider Equation 5.2 in which the interaction dummy is included.

$$\begin{aligned} E(C_i|D_{2i}=1, D_{3i}=1) &= \alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 + \beta Y_i \\ E(C_i|D_{2i}=0, D_{3i}=1) &= \alpha_1 + \alpha_3 + \beta Y_i \\ E(C_i|D_{2i}=1, D_{3i}=0) &= \alpha_1 + \alpha_2 + \beta Y_i \\ E(C_i|D_{2i}=0, D_{3i}=0) &= \alpha_1 + \beta Y_i \end{aligned}$$

We can find that the inclusion of $\alpha_4 D_{2i} D_{3i}$ to the model capture the interaction effect of two characteristics. That is, the married woman would have different consumption expenditure from single woman by $\alpha_3 + \alpha_4$, from married man by $\alpha_2 + \alpha_4$, and from single man by $\alpha_2 + \alpha_3 + \alpha_4$. In the first model, the married woman would have different consumption expenditure from single woman by α_3, from married man by α_2, and from single man by $\alpha_2 + \alpha_3$.

5.3.3 Hypothesis Testing of Structural Change: Dummy Variable and Chow test

From the hypothesis testing of the structural change in the model of Chapter 4, practically, we can also apply dummy variable to test the existence of that change. Furthermore, we may reach additional results on whether the change results from difference in the vertical intercept or in the slope. Given two set of data, the model can be constructed as

$$\begin{aligned} n_1: \quad Y_t &= \alpha_1 + \alpha_2 X_t + u_t \\ n_2: \quad Y_t &= \beta_1 + \beta_2 X_t + u_t \end{aligned}$$

To answer the question of whether we should treat the two set of data separately, we can test the hypothesis by using Chow test. That is, we could use ordinary least square to estimate three following models.

$$\begin{aligned} n_1: \quad Y_t &= \alpha_1 + \alpha_2 X_t + u_t \\ n_2: \quad Y_t &= \beta_1 + \beta_2 X_t + u_t \\ n_1 + n_2: \quad Y_t &= \gamma_1 + \gamma_2 X_t + u_t \end{aligned}$$

and set the hypothesis as

$$\begin{aligned} H_0: \quad & \alpha_1 = \beta_1 = \gamma_1 \quad \alpha_2 = \beta_2 = \gamma_2 \\ H_a: \quad & \alpha_1 \neq \beta_1 \neq \gamma_1 \quad \alpha_2 \neq \beta_2 \neq \gamma_2 \end{aligned}$$

Then, we compute the value of $\hat{F}$ and compare it with critical F-value, we can reach the conclusion of whether there is any structural change. Nonetheless, the application of dummy variable to hypothesis testing of the structural change will ease the model construction. Let dummy variable D_t take on the value of 1 if the data is in the second group (n_2) and take on the value of 0 if the data is in the first group (n_1). The model can be established as

$$D_t = \begin{cases} 1 & \text{Two sets of data are different} \\ 0 & \text{Two sets of data are the same} \end{cases}$$

$$\begin{aligned} & Y_t = \gamma_1 + \gamma_2 X_t + \delta_1 D_t + \delta_2 D_t X_t + u_t \\ OLS: \quad & \hat{Y}_t = \hat{\gamma}_1 + \hat{\gamma}_2 X_t + \hat{\delta}_1 D_t + \hat{\delta}_2 D_t X_t \\ D = 0: \quad & \hat{Y}_t = \hat{\gamma}_1 + \hat{\gamma}_2 X_t \\ D = 1: \quad & \hat{Y}_t = \hat{\gamma}_1 + \hat{\delta}_1 + (\hat{\gamma}_2 + \hat{\delta}_2) X_t \end{aligned}$$

The advantage is that the test of single model will leave higher degree of freedom to us than the separate test of two data sets. Also, if δ_1 is statistically significantly different from zero, we can conclude that the structural change is rooted in the difference of vertical intercept. On the other hand, if δ_2 is statistically significantly different from zero, we can conclude that the structural change is originated by the difference of slope. Moreover, if both δ_1 and δ_1 are statistically significantly different from zero, we can conclude that the structural change is caused by the difference of both intercept and slope.

Chapter 6

MULTICOLLINEARITY

6.1 CHARACTERISTICS OF MULTI-COLLINEARITY

To make estimators BLUE and able to convey the meaningful relationship between independent and dependent variables, one of the required assumptions in Chapter 4 is no *perfect multicollinearity*. To clarify, first begin with the multiple regression model as Equation 6.1, in general, where $X_1 = 1$ for all observations to enable the intercept term to enter the model.

$$Y_i = \beta_1 X_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \beta_k X_{ki} + u_i \tag{6.1}$$

If the independent variables above can be algebraically formed as Equation 6.2, they are said to have exact linear relationship or **perfect multicollinearity**. That is, we can acquire the value of any independent variable in the model through the linear combination of other independent variables. For instance, if we want to find the value of X_2, we can apply the addition, subtraction, multiplication and division among other independent variables.

$$\lambda_1 X_{1i} + \lambda_2 X_{2i} + \lambda_3 X_{3i} + \cdots + \lambda_k X_{ki} = 0 \tag{6.2}$$

where $\lambda_1, \lambda_2, \ldots, \lambda_k$ are not equal to zero simultaneously.

On the other hand, if the formation of independent variables follows Equation 6.3, rather than Equation 6.2, they are said to have **imperfect multicollinearity**. Specifically, we cannot obtain any independent variable in the model from the linear combination of other independent variables.

$$\lambda_1 X_{1i} + \lambda_2 X_{2i} + \lambda_3 X_{3i} + \cdots + \lambda_k X_{ki} + \nu_i = 0 \tag{6.3}$$

where ν_1 is the random disturbance term.

According to Equation 6.2 and 6.3, consider Table 6-1 which illustrates the collection of 5 observations for each independent variable (X_2 and X_3). It is obvious that we can multiply X_2 by the constant term to transform it into X_3. In this case, we can establish the relationship, as in Equation 6.2, between these two independent variables by letting $\lambda_1 = -3$ and $\lambda_2 = 1$.

$$-3X_{2i} + X_{3i} = 0$$

Table 6-1: Perfect multicollinearity in explanatory variables

Observation	X_{2i}	X_{3i}	$3X_{2i}$	$\nu_i = -3X_{2i} + X_{3i}$
1	5	15	15	0
2	14	42	42	0
3	7	21	21	0
4	31	93	93	0
5	25	75	75	0

For the case of imperfect multicollinearity, consider Table 6-2. It can be found that we cannot form the relationship as Equation 6.2 due to the difference between independent variables (X_2 and X_3). Even we multiply X_2 by -3, the random disturbance term (ν_i) still exists. In Table 6-2, after the forth observation of X_2 is multiplied by 3, it is still different from 89 by 4. Hence, the relationship between these two independent variables can be written as

$$-3X_{2i} + X_{3i} + \nu_i = 0$$

Table 6-2: Imperfect multicollinearity in explanatory variables

Observation	X_{2i}	X_{3i}	$3X_{2i}$	$\nu_i = -3X_{2i} + X_{3i}$
1	5	16	15	1
2	14	45	42	3
3	7	18	21	-3
4	31	89	93	-4
5	25	75	75	0

The relationship discussed in this chapter involves only the linear one. Although the independent variable is squared or cubed, as in Equation 6.4, it does not always mean that the model constructed from these variables will suffer the perfect or imperfect multicollinearity. The important factor to consider is whether X_i and X_i^3 can be written in the form of Equation 6.2 or 6.3.

$$Y_i = \beta_1 + \beta_2 X_i + \beta_3 X_i^3 + u_i \tag{6.4}$$

In the following sections of this chapter, the consequence of the perfect or imperfect linear relationship of independent variables will be discussed. However, to completely understand the characteristics of multicollinearity, it is essential to know the sources of the problem. In principle, multicollinearity is originated from:

1. *Method of data collection used in the regression model*: sometimes researchers collect the data in the limited amount, causing the sample to concentrate in some group of population rather than to represent the population as a whole.

2. *Restriction imposed in the model*: in the study of relationship between a single dependent variable and multiple independent variables, possibly the linear relationship exists among those independent variables. To illustrate, suppose we study the dependence of the sale of goods Y on the prices of goods X and Y, where X is used to produce Y. With this relationship, when the price of goods X increase, it almost certainly raises the price of goods Y. Hence, there seems to be highly linear relationship between these two variables.

3. *Application of polynomial to the model*: such as Equation 6.4, the X_i^3 cubed is included as another variable. If the data used in the study is restricted within the narrow range, the value of two variable, namely X_i and X_i^3, might not be notably different, resulting in the liner relationship between them.

4. *Over-determination of the model*: some models have higher amount of parameters than the amount of observation collected. The evident example of these models is in the medical or human behavior field in which the amount of patients or volunteers is less than the independent variables. Usually, researchers have to discard some variables to make the study possible.

5. *Common trend of independent variables*: the time series data of revenue, expenditure and population seems to move together because, as the time passes, they tend to increase collectively. Thus, the linear relationship among them might occur.

6.2 CONSEQUENCE OF MULTICOLLINEARITY

According to Chapter 4, if we have the regression model as shown in Equation 6.5, we can employ the ordinary least square to estimate β_2 and β_3. By the method of calculus, minimizing the sum of disturbance term squared, we obtain the close solution for the estimators (β_2 and β_3), illustrated in Equation 4.2 and 4.3.

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + \hat{u}_i \tag{6.5}$$

$$\hat{\beta}_2 = \frac{(\sum y_i x_{2i})(\sum x_{3i}^2) - (\sum y_i x_{3i})(\sum x_{2i} x_{3i})}{(\sum x_{2i}^2)(\sum x_{3i}^2) - (\sum x_{2i} x_{3i})^2} \tag{4.2}$$

$$\hat{\beta}_3 = \frac{(\sum y_i x_{3i})(\sum x_{2i}^2) - (\sum y_i x_{2i})(\sum x_{2i} x_{3i})}{(\sum x_{2i}^2)(\sum x_{3i}^2) - (\sum x_{2i} x_{3i})^2} \tag{4.3}$$

where
$y_i = Y_i - \bar{Y}$,
$x_{2i} = X_{2i} - \bar{X}_{2i}$,
$x_{3i} = X_{3i} - \bar{X}_{3i}$

Nonetheless, if the explanatory variables, X_2 and X_3, suffer perfect multicollinearity, namely $X_{3i} = \lambda X_{2i}$ where λ is the constant greater than zero, from the relationship stated, we can substitute $x_{3i} = \lambda x_{2i}$ in Equation 4.2 and 4.3 and get

$$\hat{\beta}_2 = \frac{\lambda^2[(\sum y_i x_{2i})(\sum x_{3i}^2) - (\sum y_i x_{2i})(\sum x_{2i}^2)]}{\lambda^2[(\sum x_{2i}^2)(\sum x_{2i}^2) - (\sum x_{2i})^2]} = \frac{0}{0} \tag{6.6}$$

$$\hat{\beta}_3 = \frac{\lambda^2[(\sum y_i x_{2i})(\sum x_{3i}^2) - (\sum y_i x_{2i})(\sum x_{3i} x_{3i})]}{\lambda^2[(\sum x_{2i}^2)(\sum x_{2i}^2) - (\sum x_{2i})^2]} = \frac{0}{0} \tag{6.7}$$

From Equation 6.6 and 6.7, mathematically, we cannot acquire the estimates of β_2 and β_3 since, in the system of real number, the division by zero is not defined.

On the other hand, consider the case where there is imperfect multicollinearity among regressors as in 6.3. For the model with two regressors, let $X_{3i} = \lambda X_{2i} + \nu_i$ and substitute $x_{3i} = \lambda x_{2i} + \nu_i$ into Equation 4.2, the estimator of β_2 can be obtained by Equation 6.8 which is different from the case with perfect multicollinearity problem. The same is true for both β_1 and β_3.

$$\hat{\beta}_2 = \frac{(\sum y_i x_{2i})(\lambda^2 \sum x_{2i}^2 + \sum \nu_i^2) - (\lambda \sum y_i x_{2i} + \sum y_i \nu_i)(\lambda \sum x_{2i}^2)}{(\sum x_{2i}^2)(\lambda^2 \sum x_{2i}^2 + \sum \nu_i^2) - (\lambda \sum x_{2i}^2)^2} \neq \frac{0}{0} \tag{6.8}$$

by letting

$X_{3i} = \lambda X_{2i} + \nu_i$
$\bar{X}_{3i} = \lambda \bar{X}_{2i}$
$X_{3i} - \bar{X}_{3i} = \lambda(X_{2i} - \bar{X}_{2i}) + \nu_i$

$x_{3i} = \lambda x_{2i} + \nu_i$
where $\lambda \neq 0$ and $\sum x_i \nu_i = 0$

Theoretically, the estimators obtained from OLS will be BLUE. To be specific, the estimators of the model that suffers imperfect multicollinearity problem will be unbiased and have minimum variance. Yet, if the model suffers high degree of multicollinearity, the variance of the estimators would be so high that the regression analysis may be negatively influenced. The variance of both estimators and the covariance between them can be calculated by Equation 6.9 to 6.11 [1].

$$var(\hat{\beta}_2) = \frac{\sigma^2}{\sum x_{2i}^2(1 - r_{23}^2)} \tag{6.9}$$

$$var(\hat{\beta}_3) = \frac{\sigma^2}{\sum x_{3i}^2(1 - r_{23}^2)} \tag{6.10}$$

$$cov(\hat{\beta}_2, \hat{\beta}_3) = \frac{-r_{23}\sigma^2}{(1 - r_{23}^2)\sqrt{\sum x_{2i}^2 x_{3i}^2}} \tag{6.11}$$

where r_{23} is the correlation coefficient between regressors X_2 and X_3 and can be computed by Equation 4.9 and the value ranges from -1 to 1.

$$r_{23} = \frac{(\sum_{i=1}^{n} x_{2i}x_{3i})^2}{\sum_{i=1}^{n} x_{2i}^2 \sum_{i=1}^{n} x_{3i}^2} \tag{4.9}$$

According to Equation 6.9 and 6.10, it can be seen that the higher the correlation coefficient, the higher the variance of estimators. To ease the analysis, redefine Equation 6.9 by **Variance Inflation Factor (VIF)** by letting,

$$\text{VIF} = \frac{1}{1 - r_{23}^2} \tag{6.12}$$

Substitute VIF into Equation 6.9 and 6.10, we get

$$var(\hat{\beta}_2) = \frac{\sigma^2}{\sum x_{2i}^2}\text{VIF} \tag{6.13}$$

$$var(\hat{\beta}_3) = \frac{\sigma^2}{\sum x_{3i}^2}\text{VIF} \tag{6.14}$$

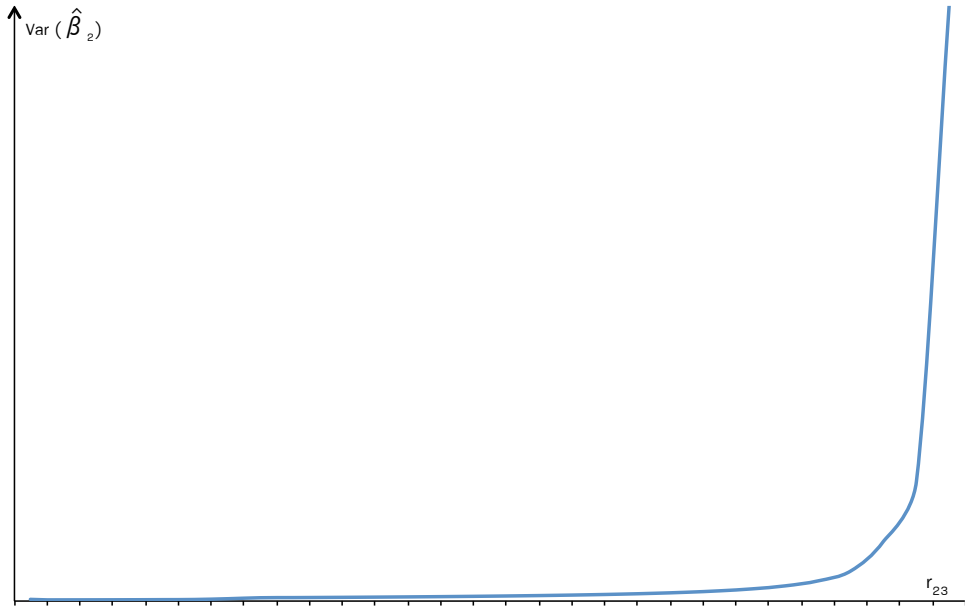
As $r_{23} \to 1$, or the correlation approaches one, VIF $\to \infty$ and the variance will be higher and approach infinity. On the contrary, as $r_{23} \to 0$, or the correlation coefficient approaches zero (namely, no linear relationship), VIF $\to 1$ and the variance will be lower. Consider Table 6-3 and Figure 6-1, it can be seen that the higher the correlation , the higher the VIF and the higher the variance of estimators.

[1] from Chapter 4

Table 6-3: The consequence of an increase in the correlation coefficient on the variance of estimators

r_{23}	VIF	$var(\hat{\beta}_2)$	$var(\hat{\beta}_3)$
		Let	
0.00	1	$\frac{\sigma^2}{\sum x_{2i}^2}$VIF = B	$\frac{\sigma^2}{\sum x_{2i}^2}$VIF = C
0.50	1.33	1.33B	1.33C
0.70	1.96	1.96B	1.96C
0.80	2.78	2.78B	2.78C
0.90	5.76	5.76B	5.76C
0.97	16.92	16.92B	16.92C
0.99	50.25	50.25B	50.25C

Figure 6-1: The consequence of an increase in the correlation coefficient on the variance estimators



When the variance rises due to the multicollinearity among independent variables, the standard deviation will certainly rise and at least two negative effects will result. First, the interval estimation will be impaired because the confidence interval will be widen, which can be seen in Equation 3.23 and Table 6-4 (for 95 percent confidence interval and large amount of observations). The other negative effect is on hypothesis test since the t-statistic, as in Equation 6.15, will be lower and might result in misleading conclusion from hypothesis test.

Table 6-4: The consequence of an increase in the correlation coefficient on 95 percent confidence interval

r_{23}	95 Confidence interval of β_2
0.00	$\hat{\beta}_2 \pm 1.96\sqrt{\frac{\sigma^2}{\sum x_{2i}^2}}$
0.50	$\hat{\beta}_2 \pm 1.96\sqrt{\frac{\sigma^2}{\sum x_{2i}^2}1.33}$
0.70	$\hat{\beta}_2 \pm 1.96\sqrt{\frac{\sigma^2}{\sum x_{2i}^2}1.96}$
0.80	$\hat{\beta}_2 \pm 1.96\sqrt{\frac{\sigma^2}{\sum x_{2i}^2}2.78}$
0.90	$\hat{\beta}_2 \pm 1.96\sqrt{\frac{\sigma^2}{\sum x_{2i}^2}5.76}$
0.97	$\hat{\beta}_2 \pm 1.96\sqrt{\frac{\sigma^2}{\sum x_{2i}^2}16.92}$
0.99	$\hat{\beta}_2 \pm 1.96\sqrt{\frac{\sigma^2}{\sum x_{2i}^2}50.25}$

$$\hat{t} = \left(\frac{\hat{\beta}_2 - \beta_2}{se(\hat{\beta}_2)\uparrow}\right)\downarrow \tag{6.15}$$

In any models, we might have the high value of coefficient of determination (R^2) and statistically overall significant of the model from F-test. The implication is that the model possesses the explanatory power over the dependent variable. However, it is possible that we might not get the statistically significant result from the test of individual coefficients. That is, some coefficient is not significantly different from zero which means the variable associated with that coefficient lacks explanatory variable because the t-statistic is lower due to multicollinearity problem. This situation is called **conflicting test**, namely the result from t-test contradicts with the one from F-test.

To conclude, if there is perfect multicollinearity among independent variables, we are unable to estimate the parameters in the model. Also, the variance of estimators will approach infinity. Furthermore, if there is imperfect multicollinearity, OLS is still applicable to estimate the parameters. Yet, it has to be aware that the variance of estimators might be so high that some aspects of regression analysis, such as interval estimation and hypothesis test, are negatively influenced.

6.3 DETECTION OF MULTICOLLINEARITY

We have already discussed the consequence of both perfect and imperfect multicollinearity among regressors. For the regression analysis, the harmful problem is the situation when there is perfect multicollinearity which will invalidate the estimation of the model in order to explain the true relationship in the population. The case of perfect multicollinearity, thus, can be easily detected.

For the imperfect multicollinearity, if the degree of multicollinearity is not immense, the estimators are still BLUE. Yet, if the degree is huge, the problem will become damaging. Statistically, the extent of multicollinearity can be tested through various approaches. Some of them are discussed here.

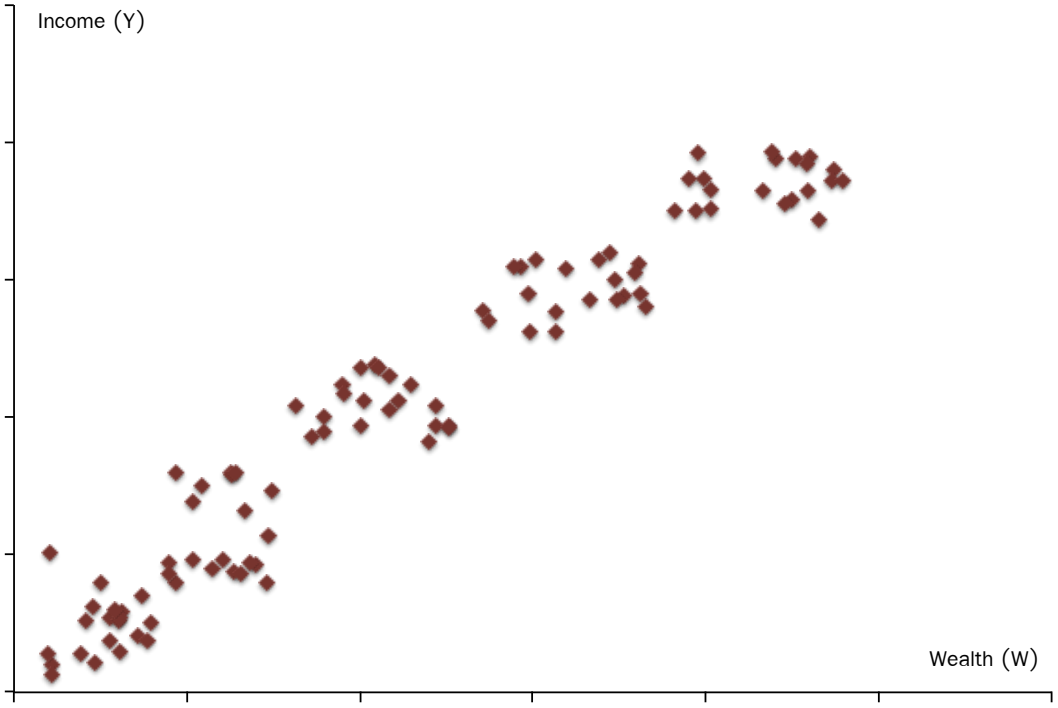
1. *There is conflicting test between t- and F-test*: if we find that the conclusion derived from the two tests are inconsistent, specifically R^2 is high and F-test results in statistical overall significance; whereas, at least, one null hypothesis of some t-tests cannot be rejected, it is reasonable to suspect the multicollinearity problem.

2. *Correlation of regressors is greater than 0.8*: the higher the correlation, the higher the variance of estimators.

3. *Variance inflation factor is greater than 10*: when the regressors face the multicollinearity problem, the value of VIF might be so high that the resulting high variance of estimators adversely affects the regression analysis.

4. *Scatter plot of two regressors is relatively linear*: when we plot the value of on regressor against another and we find that both of them tend to change in the same way, this fact might suggest the existence of multicollinearity. Figure 6-2 depicts the case where income and wealth, which is usually perceived to explain consumption expenditure, are prone to move together.

Figure 6-2: Scatter plot between two regressors, namely income and wealth, showing the linear relationship between both of them



6.4 REMEDIAL MEASURE FOR MULTICOLLINEARITY

In principle, the problem of multicollinearity among explanatory variables is not actually serious as we still have BLUE estimators. Notwithstanding, the problem become more severe as the degree of multicollinearity rises and can be alleviated through:

1. *Do nothing*: if the degree of multicollinearity is low, the model is still valid as the BLUE property of estimators is attained.

2. *Apply prior relationship among explanatory variables*: consider the model

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$$

If we know before that the linear relationship between explanatory variables X_2 and X_3 can be written as $\beta_3 = 0.7\beta_2$, we can use this fact to eliminate the problem by

$$\begin{aligned} Y_i &= \beta_1 + \beta_2 X_{2i} + 0.7\beta_2 X_{3i} + u_i \\ &= \beta_1 + \beta_2 (X_{2i} + 0.7X_{3i}) + u_i \\ &= \beta_1 + \beta_2 X * + u_i \end{aligned}$$

where $X^* = X_{2i} + 0.7X_{3i}$

3. *Discard some explanatory variables*: the removal of the variables could mitigate the problem; but, another problem, namely **misspecification** problem, might occur instead. For example, suppose we want to construct the model where the production is the explained variables; and labour and capital are the explanatory ones. If there is linear relationship between labour and capital, the elimination of one variable might assuage the multicollinearity problem, but might be contrary to economic reasoning. Hence, the decision of which variables will be disposed of should be based on economic theory.

4. *Collect more observation*: this practice will increase $\sum x^2$ which is the component of the variances [2]. Accordingly, the variances will be lower despite high correlation among explanatory variables.

$$var(\hat{\beta}_2) = \frac{\sigma^2}{\sum x^2 (1 - r_{23}^2)}$$

5. *Transform the variables*: although there is linear relationship among explanatory variables, it is not necessary that the *first difference* or *ratio transformation* of the variables will have that relationship.

[2] As the data set gets larger, the sample statistic will approach the population parameter. Consequently, we can reasonably state that mean of X is almost stable under the larger data set. In this case, the increase in the size of data set is likely to increase the sum of the square of deviation from mean

For the first difference of variables, consider the model in period t and t-1

$$\begin{aligned} Y_t &= \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + u_t \\ Y_{t-1} &= \beta_1 + \beta_2 X_{2,t-1} + \beta_3 X_{3,t-1} + u_{t-1} \end{aligned}$$

$$\begin{aligned} Y_t - Y_{t-1} &= \beta_2(X_{2t} - X_{2,t-1}) + \beta_3(X_{3t} - X_{3,t-1}) + \nu_t \\ \Delta Y_t &= \beta_2 \Delta X_2 + \beta_3 \Delta X_3 + \nu_t \\ \text{where} & \\ \nu_t &= u_t - u_{t-1} \\ \Delta Y_t &= Y_t - Y_{t-1} \\ \Delta X_2 &= X_{2t} - X_{2,t-1} \\ \Delta X_3 &= X_{3t} - X_{3,t-1} \end{aligned}$$

This transformation perhaps results in no linear relationship among new regressors. Unfortunately, another serious econometric problem might take place which is the **autocorrelation** problem which will be discussed in Chapter 8.

For the ratio transformation of variables, consider the model

$$\begin{aligned} Y_t &= \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + u_t \end{aligned}$$

$$\begin{aligned} \frac{Y_t}{X_{3t}} &= \beta_1 \frac{1}{X_{3i}} + \beta_2 \frac{X_{2t}}{X_{3i}} + \beta_3 \frac{X_{3t}}{X_{3i}} + \frac{u_t}{X_{3i}} \\ \frac{Y_t}{X_{3t}} &= \beta_1 \frac{1}{X_{3i}} + \beta_3 + \beta_2 \frac{X_{2t}}{X_{3i}} + \frac{u_t}{X_{3i}} \\ Y_t^* &= \beta_1^* + \beta_2 X_{2i}^* + u_t^* \\ \text{where} & \\ Y_t^* &= \frac{Y_t}{X_{3t}} \\ \beta_1^* &= \beta_1 \frac{1}{X_{3i}} + \beta_3 \\ X_{2i}^* &= \frac{X_{2t}}{X_{3i}} \\ u_t^* &= \frac{u_t}{X_{3i}} \end{aligned}$$

With this remedial measure, we can reduce the degree of multicollinearity since there is one explanatory variable left in the model. However, when we consider the random disturbance term in this new model, it is possible that the variance of the disturbance term might not be constant, namely **heteroscedasticity**, which will be discussed in the next chapter.

Chapter 7

HETEROSCEDASTICITY

7.1 CHARACTERISTICS OF HETEROSCEDASTICITY

From Chapter 6, as the assumption of multicollinearity among independent variables is breached, if the degree of multicollinearity is not severely high, economists can ignore the problem. In this chapter, another assumption for the best linear unbiased estimator (BLUE) of regression model trough ordinary least square (OLS) method is discussed. The assumption is **homoscedasticity** or

$$E(u_i^2) = \sigma_u^2 \quad i = 1, 2, ..., n \tag{7.1}$$

Depicted by Figure 7-1, the variances of disturbance terms given any independent variable are constant and equal. Specifically, conditional on X_1, the variance is equal to σ_u^2 which is the same as the variance of disturbance term conditional on X_2 and on the other values of independent variable.

On the other hand, if this assumption is violated, the problem occurring is called **heteroscedasticity**. That is, conditional on X_1, the variance of disturbance term is σ_1^2; whereas, conditional on X_2, the variance of disturbance term is σ_2^2. In brief, the conditional variance of disturbance term would vary across the values of independent variable, as illustrated, mathematically, by Equation 7.2 and, graphically, by Figure 7-2.

$$E(u_i^2) = \sigma_i^2 \tag{7.2}$$

Figure 7-1: Homoscedasticity

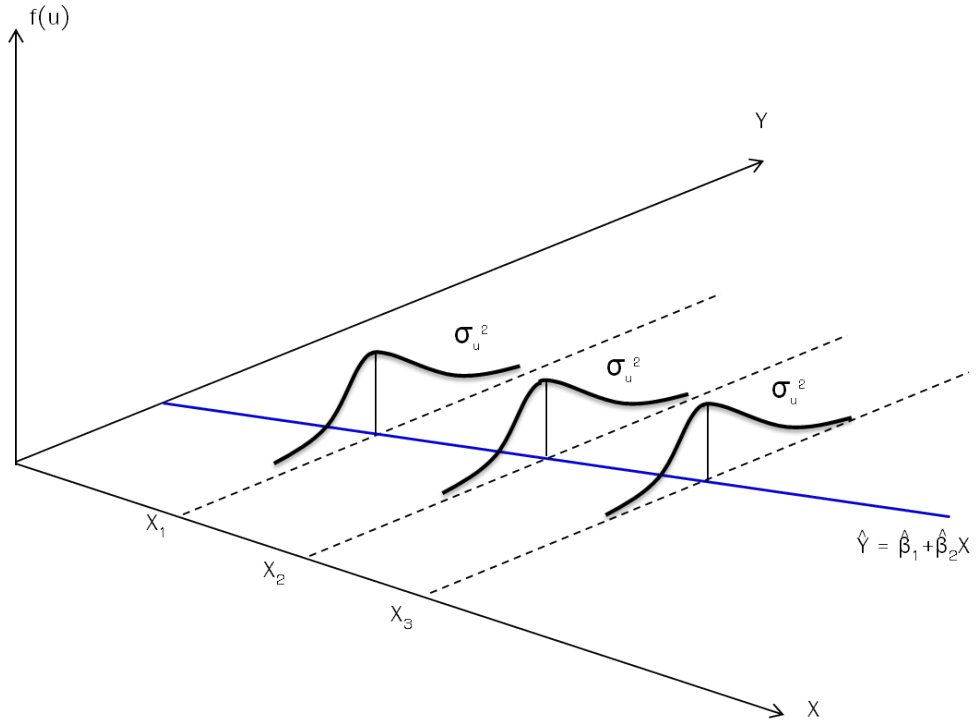
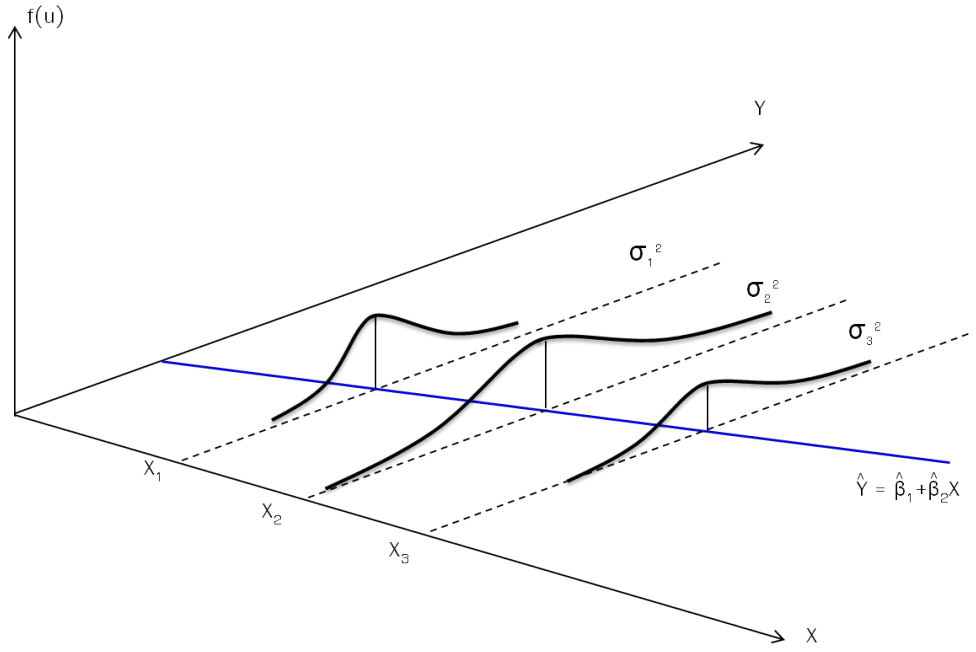


Figure 7-2: Heteroscedasticity



Generally, there is a variety of causes for the existence heteroscedasticity problem in the regression model studied by most economists. Yet, only 6 main sources are discussed here.

1. Normally, error learning is the nature of human. At the initial stage of their work, people probably commit a large number of mistakes. As they carry on working and become more specialized, the amount of errors would be reduced. In this case, it seems that the variance at the initial stage will be high but would decrease as people learn from their errors.

2. Considering the relationship between independent variable X and dependent variable Y, there seems to be possible that as the value of independent variable increases, the variance of the value of dependent one will increases. The feasible reason is the characteristics of those variables such as the relationship between the profit of the company (independent variable) and dividend (dependent variable). As the profit rises, the board of director may have a variety of dividend policy. Some companies may pay a small amount of dividend in order to keep the profit for further development. Some may pay a large amount of dividend to satisfy the shareholders. On the contrary, if the profit is low, the dividend policy will not diverge across the companies since the companies seem to be at the growth stage and decide to keep their profit as retained earnings.

3. The collection of data is another source. As the collection technique employed by the researcher is improved, the collection error would be lower. Contrarily, with the poor collection technique, the data obtained to construct the regression model would probably incur more and more errors, causing the condition variance of disturbance term to vary.

4. The existence of outliers in the independent and dependent variables may make the conditional variance of disturbance term on independent variables volatile. Mostly, if the researchers collect too small amount of data, that set of data tends to include the outliers and undermine the regression analysis. As the amount of data increases, those outliers may become normal relative to other observations.

5. The misspecification of the model could result in the heteroscedasticity problem as well. In some cases, econometricians drop some important and necessary independent variable from the regression model. The disturbance term, then, will incorporate the characteristics of the missing variables, resulting in heteroscedasticity problem. For example, suppose the researchers want to establish the model to explain the relationship of price and quantity of good X, as suggested by the theory of demand. However, if it turns out that good Y is the substitute for good X. This mistake of failing to include price of good Y, which has the explanatory power over the demand for good X, will result in misspecification error. The disturbance term will have the characteristics of good Y, which, in turn, leads to heteroscedasticity. The further details will be provides on specification error in Chapter 9.

6. Heteroscedasticity problem usually happens in the model that applies the **cross-sectional data** which tends to be highly diverse because the data is collected in the same time period. To illustrate, the census acquired from numerous provincial areas may cover the wide range of value and results in the stated problem. On the other hand, for the **time series data**, it is prone to be the collection of the same sample for different period of time. With the same sample, the range of the value covered seems to be narrow. Hence, the stated problem, generally, does not occur with this kind of data.

7.2 CONSEQUENCE OF HETEROSCEDASTICITY

For the estimation of regression model through OLS methods, given the assumption of homoscedasticity, the model can be written as

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

The estimator β_2 can be calculated by

$$\hat{\beta}_2 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2} = \frac{\sum x_i y_i}{\sum x_i^2}$$

In general, the variance of the estimator β_2 is computed by

$$Var(\hat{\beta}_2) = \frac{\sum x_i^2 \sigma_i^2}{(\sum x_i^2)^2} \tag{7.3}$$

Given the assumption of homoscedasticity, σ_i^2 is the same across the value of i; namely, $\sigma_i^2 = \sigma_j^2, \forall i \neq j$. In Chapter 3, we conclude that the estimator will be BLUE and the variance of the estimator β_2 can be written as

$$Var(\hat{\beta}_2) = \frac{\sigma_u^2}{\sum x_i^2}$$

However, if the assumption of homoscedasticity is violated, or there is the problem of heteroscedasticity, the variance of the estimator β_2 will be as Equation 7.4 and the estimators will no longer have the characteristics of minimum variance but it is still unbiased.

$$var(\hat{\beta}_2) = \frac{\sum x_i^2 \sigma_i^2}{(\sum x_i^2)^2} \tag{7.4}$$

7.3 DETECTION OF HETEROSCEDASTICITY

The problem of heteroscedasticity is fairly severe since it causes the estimators, which identify the relationship of regressor and regressand, to lose the minimum variance or best property despite its unbiased property. Accordingly, the statistical inference studied in the previous chapters, such as confidence interval and hypothesis test, is not applicable. It is, thus, essential to detect the problem of heteroscedasticity. Again, there are a large number of methods for detection suggested by econometricians; yet, 4 approaches are discussed here.

1. *Finding the relationship of regressor and random disturbance term by graph*: the nature of the problem is that the conditional variance of disturbance term is not constant. Hence, if we can create the diagram that depicts the relationship between the observations of regressor and the disturbance term squared, which is the estimator of variance, we will be able to identify whether the conditional variance is constant across the observations of regressor. Consider Figure 7-3 and 7-4. When the regressor X has the positive relationship with the estimators of variance, it implies that the variance of disturbance term seems to move in the same direction with the regressor. From Figure 7-3, econometricians should be aware that, in this model, the evidence for heteroscedasticity seems to be eminent. On the contrary, from Figure 7-4, no relationship between regressor and estimator of variance is detected. However, graphical method is merely the initial step for the detection process. Further reliable statistical tests should be performed as well.

Figure 7-3: The noticeable relationship between u_i^2 and regressor X of the model suffering heteroscedasticity problem

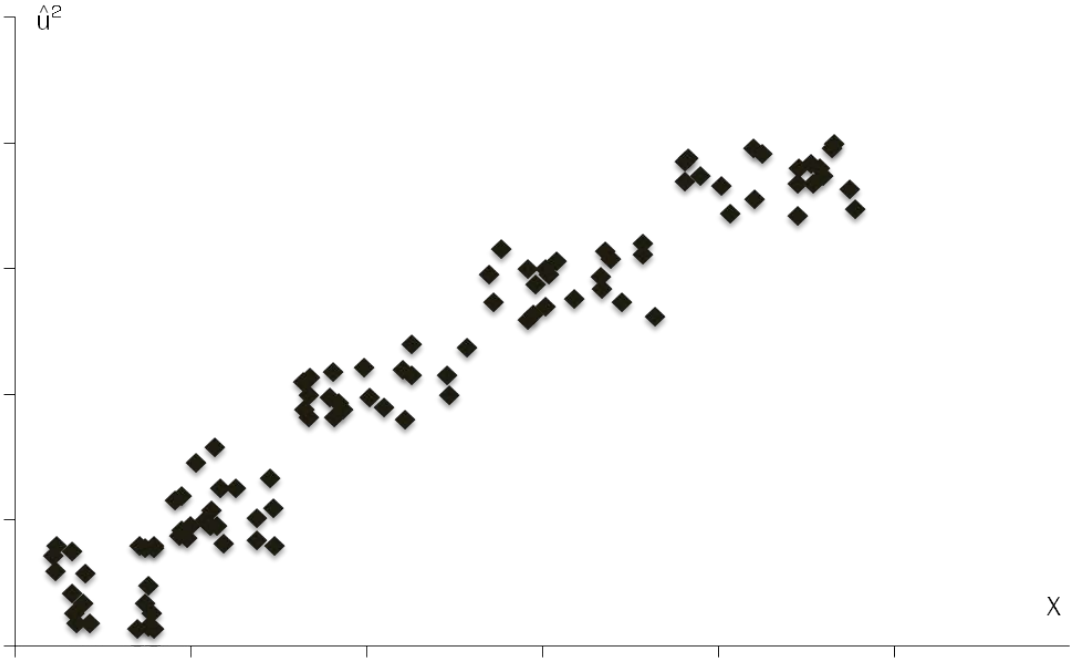
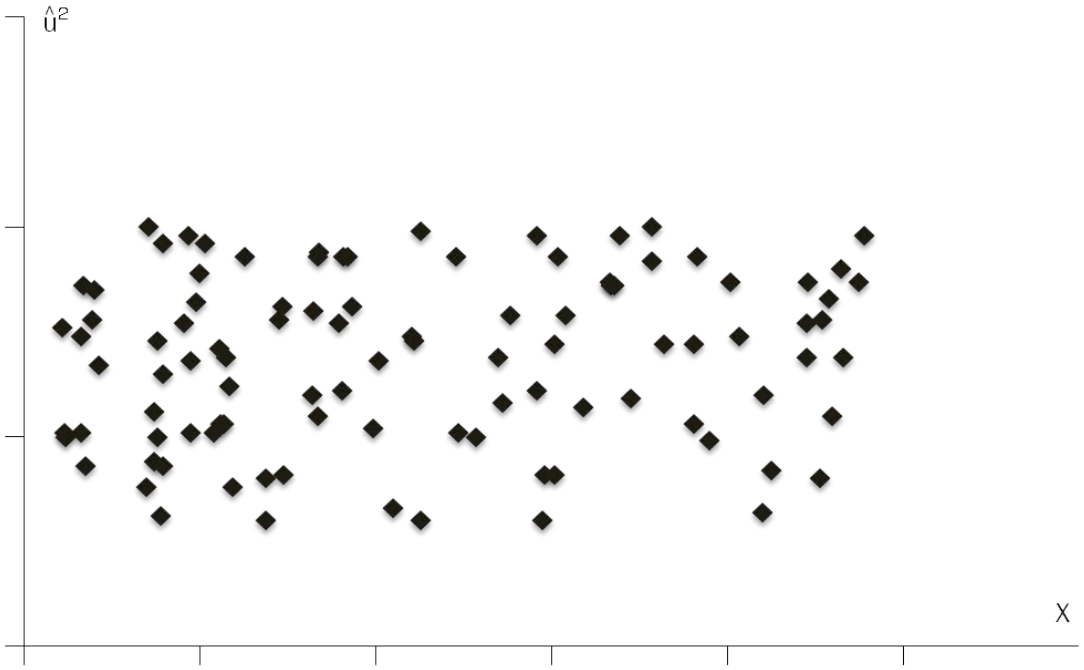


Figure 7-4: The ambiguous relationship between u_i^2 and regressor X of the model not suffering heteroscedasticity problem



2. **Park Test**: the rationales is that, if we can construct the regression model that enables the regressor X to explains the volatility of the variance of disturbance term, that means the variance of disturbance term is not constant and depends on the regressor. The model could be constructed as Equation 7.5 and we can transform the model into the linear one as Equation 7.6.

$$\sigma_i^2 = \sigma_u^2 X_i^\beta e^{\nu_i} \tag{7.5}$$

$$\ln \sigma_i^2 = \ln \sigma_u^2 + \beta \ln X_i + \nu_i \tag{7.6}$$

Practically, we would never know the true variance of the disturbance term in the model; so, we use $\hat{u}_i^2$ as the estimator and form the model similar to Equation 7.6, which is shown in Equation 7.7.

$$\ln \hat{u}_i^2 = \ln \sigma_u^2 + \beta \ln X_i + \nu_i = \alpha + \beta \ln X_i + \nu_i \tag{7.7}$$

After the establishment of the model in Equation 7.7, we, then, can perform the hypothesis test to examine whether the regressor X could explain the change in the regressand $\ln \hat{u}_i^2$. In this case, we test for the statistically significance of the coefficient associated with β by finding the t-statistic. If the regressor X is statistically significantly able to describe the regressand $\ln u_i^2$, we can conclude that the model faces the problem of heteroscedasticity. In short, the procedure for Park test is as follows.

Step 1: establish the model of interest to find the relationship of regressor X and regressand Y

Step 2: calculate $\hat{u}_i^2 = (Y_i - \hat{Y}_i)^2$ from the regression model in Step 1 to be the estimator of variance in Equation 7.6 and 7.7.

Step 3: establish the model as in Equation 7.7 and perform the hypothesis test for the relationship between the regressor and the variance of disturbance term. The hypothesis can be set as

$$\begin{aligned} H_o: &\quad \beta = 0 \\ H_a: &\quad \beta \neq 0 \end{aligned}$$

If the null hypothesis is rejected, we can conclude that, the regressor X possess the explanatory power over the variance of disturbance term. In other word, the model suffers the heteroscedasticity prolem. Contrarily, if the null hypothesis is not rejected, it implies no heteroscedasticity problem in the model

3. **Breusch-Pagan Test** or **LM Test**: consider the multiple regression model in Equation 7.8 and suppose that the variance of disturbance term has the linear relationship with the regressor as in Equation 7.9. To satisfy the homoscedasticity assumption such that the estimator is BLUE, all partial regression coefficients in Equation 7.9 must be zero.

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \beta_k X_{ki} + u_i \tag{7.8}$$

$$Var(u|X_2, \ldots, X_k) = E(u^2|X_2, \ldots, X_k) = \alpha_1 + \alpha_2 X_{2i} + \alpha_3 X_{3i} + \cdots + \alpha_k X_{ki} \tag{7.9}$$

The procedure of Breusch-Pagan test for heteroscedasticity is as follows.

Step 1: estimate the model as in Equation 7.8 by the method of OLS and calculate the value of $\hat{u}_i^2$

Step 2: establish the regression model as in Equation 7.10 to find the coefficient of determination $R^2_{\hat{u}_i^2}$

$$\hat{u}_i^2 = \alpha_1 + \alpha_2 X_{2i} + \alpha_3 X_{3i} + \cdots + \alpha_k X_{ki} + u_i \tag{7.10}$$

Step 3: compute the F-statistic by Equation 7.11 and perform hypothesis test to find out whether all the regressors X's can jointly explain the variance of the disturbance term. If they have the explanatory power, we can conclude that the model in Equation 7.8 would suffer heteroscedasticity problem.

H_o : $\alpha_1 = \alpha_2 = \alpha_3 = ... = \alpha_k = 0$
H_a : otherwise

$$\hat{F} = \frac{R^2_{\hat{u}_i^2}/(k-1)}{(1-R^2_{\hat{u}_i^2})/(n-k)} \tag{7.11}$$

Furthermore, LM-statistic (**Lagrange Multiplier**) can be used to determine whether there is heteroscedasticity problem in the model and can be calculated by Equation 7.12.

$$LM = nR^2_{\hat{u}_i^2} \tag{7.12}$$

The LM-statistic has the chi-square distribution with the degree of freedom $k-1$ or χ^2_{df}. We can use LM-statistic to test the following hypothesis.

H_o : Homoscedasticity
H_a : otherwise

If the LM-statistic obtained is greater than the critical value found the chi-square table, we can conclude that the model in Equation 7.8 faces the heteroscedasticity problem. Nevertheless, if the LM-statistic is less than the critical value, we can conclude that no heteroscedasticity problem exists.

4. **White Test**: Consider the multiple regression model as in Equation 7.3.

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i \tag{7.13}$$

White test has the different procedure from the third test only in the aspect of how the hypothesis is set. While Breusch-Pagan test states that the variance of disturbance term has the relationship with regressors, White test will cover the wider relationship between the variance of disturbance term and higher amount of regressors as in Equation 7.14 with the procedure as follows.

$$Var(u_i|X_2, X_3) = E(u_i^2|X_2, X_3) = \alpha_1 + \alpha_2 X_{2i} + \alpha_3 X_{3i} + \alpha_4 X_{2i}^2 + \alpha_5 X_{3i}^2 + \alpha_6 X_{2i} X_{3i} + \nu_i \tag{7.14}$$

Step 1: establish the model as in Equation 7.13 to obtain $\hat{u}_i^2$

Step 2: establish the regression model as Equation 7.15

$$\hat{u}_i^2 = \alpha_1 + \alpha_2 X_{2i} + \alpha_3 X_{3i} + \alpha_4 X_{2i}^2 + \alpha_5 X_{5i}^2 + \alpha_6 X_{2i} X_{3i} + \nu_i \tag{7.15}$$

Step 3: similar to Step-3 of Breusch-Pagan test, calculate F- or LM-statistic and set the null and alternative hypothesis. Then, compare the F- or LM-statistic with the critical value from the statistical table to test for heteroscedasticity.

Through the four methods stated above, the econometricians can test whether the model that is estimated by OLS method suffers the heteroscedasticity problem. The graphical test is nothing but an initial method without any statistical test at any level of significance. The other tests involve the formulation of hypothesis and statistical test at the chosen level of significance. Hence, either Park, Breusch-Pagan or White test concerns the level of significance at 0.01, 0.05 or 0.1, contingent on the situations.

7.4 REMEDIAL MEASURE FOR HETEROSCEDASTICITY

Even though the heteroscedasticity does not make the estimators biased, the variance of the estimators obtained from the regression model will not be minimum. The loss of the best property will impair the statistical inference in which the econometricians may be interested such as the confidence interval and hypothesis test of whether there is statistically significant relationship between explanatory and explained variables.

The remedial measure for heteroscedasticity will enable the researcher to better analyse and apply the statistical tools for the study of relationship between explanatory and explained variables in the model. The measure can be categorized into two cases which are the case where the variance of each disturbance term is known and the case where the variance of each disturbance term is not known.

7.4.1 Know Variance of Each Disturbance Term

If the variance of each disturbance term (that is the variance of u_i for each i) is known, econometricians have invented the method called **generalized least square: GLS** to solve this problem. Through this method, the variance of the estimators in the model will have the minimum variance, which enables further statistical analysis.

Consider the simple regression model with explanatory variable X and explained variable Y with two parameters β_1 and β_2

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

When the variance of each disturbance term (σ_i^2) is known, dividing the entire regression model by the known standard error, we get

$$\frac{Y_i}{\sigma_i} = \beta_1 \frac{1}{\sigma_i} + \beta_2 \frac{X_i}{\sigma_i} + \frac{u_i}{\sigma_i} \tag{7.16}$$

Transform the equation to

$$Y_i^* = \beta_1 X_{0i}^* + \beta_2 X_i^* + u_i^* \tag{7.17}$$

$$\begin{array}{rrcl} \text{where} & Y_i^* & = & \frac{Y_i}{\sigma_i} \\ & X_{0i}^* & = & \frac{1}{\sigma_i} \\ & X_i^* & = & \frac{X_i}{\sigma_i} \\ & u_i^* & = & \frac{u_i}{\sigma_i} \end{array}$$

With the new model specified by the Equation 7.16, considering the variance of disturbance term, we find that

$$\begin{array}{rcll} Var(u_i^*) & = & E(u_i^*) = E[(\frac{u_i}{\sigma_i})^2] & \\ & = & \frac{1}{\sigma_i^2} E(u_i^2); & \text{due to known} \sigma_i^2 \text{which is a constant} \\ & = & \frac{1}{\sigma_i^2} \sigma_i^2; & \text{due to} E(u_i^2) = \sigma_i^2 \\ & = & 1 & \end{array}$$

In this case, the estimation of the above model by OLS method will eliminate the heteroscedasticity problem. Specifically, the variance of each disturbance term will be equal to 1. The application of OLS method to the model from Equation 7.16 is as follows.

$$\begin{array}{rcl} \frac{Y_i}{\sigma_i} & = & \hat{\beta}_1^* \frac{X_{0i}}{\sigma_i} + \hat{\beta}_2^* \frac{X_i}{\sigma_i} + \frac{u_i}{\sigma_i} \\ Y_i^* & = & \hat{\beta}_1^* X_{0i}^* + \hat{\beta}_2^* X_i^* + u_i^* \\ \sum \hat{u_i^*}^2 & = & \sum (Y_i^* - \hat{\beta}_1^* X_{0i}^* - \hat{\beta}_2^* X_i^*)^2 \\ \sum (\frac{\hat{u}_i}{\sigma_i})^2 & = & \sum (\frac{Y_i}{\sigma_i} - \hat{\beta}_1^* \frac{X_{0i}}{\sigma_i} - \hat{\beta}_2^* \frac{X_i}{\sigma_i})^2 \end{array}$$

Thanks to generalized least square, calculus minimization of error term squared results the closed form solutions for the estimators β_1^* and β_2^*. The formula for β_2^* is shown in Equation 7.18.

$$\hat{\beta}_2^* = \frac{(\sum w_i)(\sum w_i X_i Y_i) - (\sum w_i X_i)(\sum w_i Y_i)}{(\sum w_i)(\sum w_i X_i^2) - (\sum w_i X_i)^2} \tag{7.18}$$

The variance of above estimator can be computed by Equation 7.19.

$$Var(\hat{\beta}_2^*) = \frac{\sum w_i}{(\sum w_i)(\sum w_i X_i^2) - (\sum w_i X_i)^2} \tag{7.19}$$

where $w_i = \frac{1}{\sigma_i^2}$

According to classical normal regression model (CLRM) and generalized least square (trough dividing the entire model by the standard deviation of disturbance term), the acquired estimators will be BLUE.

The difference between OLS and GLS is that, for GLS, the estimators obtained will minimize the weighted sum of residual squared or $w_i\hat{u}_i^2$. On the other hand, for OLS, the ones obtained will minimize the sum of residual squared $\hat{u}_i^2$.

For OLS

$$\sum \hat{u}_i^2 = \sum (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)^2$$

For GLS

$$\sum w_i \hat{u}_i^2 = \sum w_i (Y_i - \hat{\beta}_1^* - \hat{\beta}_2^* X_i)^2$$

Another difference is that GLS will assign different weights to each observation of disturbance term ($\hat{u}_i^2$) according to its importance. To be specific, if the error associated with the observation is large (that is, the variance is high), the value of the estimator in the model will greatly deviate from the value of true parameter. That observation should not be much of interest. Due to GLS method, the weight assigned will be inversely proportional to the variance of observation. Thus, that observation will be assigned a small weight of one over its variance ($\frac{1}{\sigma_i^2}$). In brief, the larger weight will be assigned to the observations that concentrate in their mean (namely, lower variance); whereas, smaller weight will be assigned to the observations that deviate from their mean (namely, higher variance).

Generally, for estimation[1] it is desirable to establish the population regression model that describes the true relationship between explanatory and explained variables. Paying more attention to the observation clustering around its (population) mean is preferable to the observation diffusing from its mean. The practice of weight assignment of GLS is a special case of least square which is known as **weighted least squares: WLS**. Contrarily, OLS assigns the same weight for all observations.

7.4.2 Unknown Variance of Each Disturbance Term

In the model suffering heteroscedasticity with unknown variance of each disturbance term, the estimators will lack the desirable property of minimum variance, undermining the process of statistical inference and resulting in misleading conclusion. Consequently, the remedial measure for this problem is crucial.

In the case where the variance of each disturbance term is unknown, there are a large number of measures. Here, two main measures are discussed such that the estimators

[1] In principle, the variance of each random disturbance term is known when we have an access to the whole population data. In that case, to reflect the true relationship on average, the establishment of the population regression model should focus more on the observation close to its mean than the one far from its mean.

from the OLS method become BLUE.

1. *Whites heteroscedasticity-consistent standard errors*: by the method of White to estimate unknown standard deviation, in principle, if the sample size is large enough, this estimator can be used to represent the true standard deviation. Moreover, econometricians can apply this estimator to further statistical test as if there is no heteroscedasticity problem. Nevertheless, if the sample size is not large enough, the estimators through White method will not have t-distribution and result in false statistical conclusion.

After all, it should be aware that if the model dose not suffer heteroscedasticity problem but econometricians still use White's estimator, the conclusion from the statistical analysis will be erroneous. Accordingly, the formal test for heteroscedasticity in the model should be conducted such that the existence of heteroscedasticity is verified.

2. *Some assumptions for the distribution of random disturbance term*: consider simple regression model

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

We can set some assumptions for the distribution of random disturbance term as follows.

Assumption 1: let the variance of random disturbance term be proportional to X_i^2

$$E(u_i^2) = \sigma_u^2 X_i^2$$

With this assumption, it can be found that the variance of disturbance term of the regression model will be constant and can be shown by

$$\begin{aligned} Y_i &= \beta_1 + \beta_2 X_i + u_i \\ \tfrac{Y_i}{X_i} &= \tfrac{\beta_1}{X_i} + \beta_2 + \tfrac{u_i}{X_i} \\ &= \beta_1 \tfrac{1}{X_i} + \beta_2 + \nu_i \end{aligned}$$

$$\begin{aligned} E(\nu_i^2) &= E(\tfrac{u_i}{X_i})^2 = \tfrac{1}{X_i^2} E(u_i^2) \\ &= \sigma_u^2 \end{aligned}$$

Assumption 2: let the variance of random disturbance term be proportional to X_i

$$E(u_i^2) = \sigma_u^2 X_i$$

With this assumption, considering the variance of disturbance term of the regression model, it can be found that the variance will be constant.

$$\begin{aligned} Y_i &= \beta_1 + \beta_2 X_i + u_i \\ \frac{Y_i}{\sqrt{X_i}} &= \frac{\beta_1}{\sqrt{X_i}} + \beta_2\sqrt{X_i} + \frac{u_i}{\sqrt{X_i}} \\ &= \beta_1\frac{1}{\sqrt{X_i}} + \beta_2\sqrt{X_i} + \nu_i \end{aligned}$$

$$\begin{aligned} E(\nu_i^2) &= E(\tfrac{u_i}{\sqrt{X_i}})^2 = \tfrac{1}{X_i}E(u_i^2) \\ &= \sigma_u^2 \end{aligned}$$

Assumption 3: let the variance of random disturbance term be proportional to the mean of explanatory variable squared or $[E(Y_i)]^2$.

$$E(u_i^2) = \sigma_u^2[E(Y_i)]^2$$

$$\begin{aligned} Y_i &= \beta_1 + \beta_2 X_i + u_i \\ E(Y_i) &= \beta_1 + \beta_2 X_i \\ \frac{Y_i}{E(Y_i)} &= \frac{\beta_1}{E(Y_i)} + \beta_2\frac{X_i}{E(Y_i)} + \frac{u_i}{E(Y_i)} \\ &= \beta_1\frac{1}{E(Y_i)} + \beta_2\frac{X_i}{E(Y_i)} + \nu_i \end{aligned}$$

Yet, since we do not know the true value of $E(Y_i)$, we, thus, use $\hat{Y}$ instead.

$$\frac{Y_i}{\hat{Y}_i} = \beta_1\frac{1}{\hat{Y}} + \beta_2\frac{X_i}{\hat{Y}} + \nu_i$$

Consider the variance of random disturbance term. It can be found that the variance is constant.

$$\begin{aligned} E(\nu_i^2) &= E(\tfrac{u_i}{\hat{Y}_i})^2 = \tfrac{1}{\hat{Y}_i{}^2}E(u_i^2) \\ &= \sigma_u^2 \end{aligned}$$

Before the assumption about disturbance term is made, we need to identify whether the disturbance term has the relationship with other variables as in the assumption going to be made. To illustrate, the diagram depicting the relationship between the disturbance term and explanatory variable squared may be constructed to examine the validity of Assumption 1. After we explore that relationship, we, then, set the corresponding assumption to solve heteroscedasticity problem.

Chapter 8

AUTOCORRELATION

8.1 CHARACTERISTICS OF AUTOCORRELATION

In Chapter 6 and 7, we study the problems of multicollinearity and heteroscedasticity with the different effects on the estimators in regression model. The problem of multicollinearity, if not perfect, does not cause the estimators to lose the desirable properties. Those estimators can well represent the true parameters. On the other hand, the problem of heteroscedasticity ruins the minimum variance property of estimators. In this Chapter, another problem of random disturbance term is considered. That problem is **autocorrelation** among disturbance term which violates one of the assumptions for classical linear regression model (CLRM)

The nature of autocorrelation is when there is correlation among disturbance terms or

$$cov(u_i, u_j|X_i, X_j) = E(u_i, u_j) \neq 0 \quad \text{where } i \neq j \tag{8.1}$$

For time series data, when the random disturbance terms are autocorrelated when the data in each period is correlated. For instance, the protest in a country that reduces the amount of export of goods and services in one month may also reduce the export of the following months. Hence, in this case, the random disturbance terms in these periods will be negative to reflect the fact that the amount of export tends to be below the mean.

For cross-sectional data, the problem of autocorrelation may occur. For example, the consumption expenditure of one family may reduce due to the great flood. Also, the flood influences other families in the same way. The consumption expenditure of these families tends to be positively correlated; hence, the random disturbance term from this set of data may also be positively correlated.

Figure 8-1 illustrates the pattern of random disturbance term when the random disturbance term faces autocorrelation problem with the increasing trend. Contrarily, Figure 8-2 depicts the case where the random disturbance term has no obvious systematic pattern, namely no autocorrelation.

Figure 8-1: Autocorrelation among disturbance term with increasing pattern

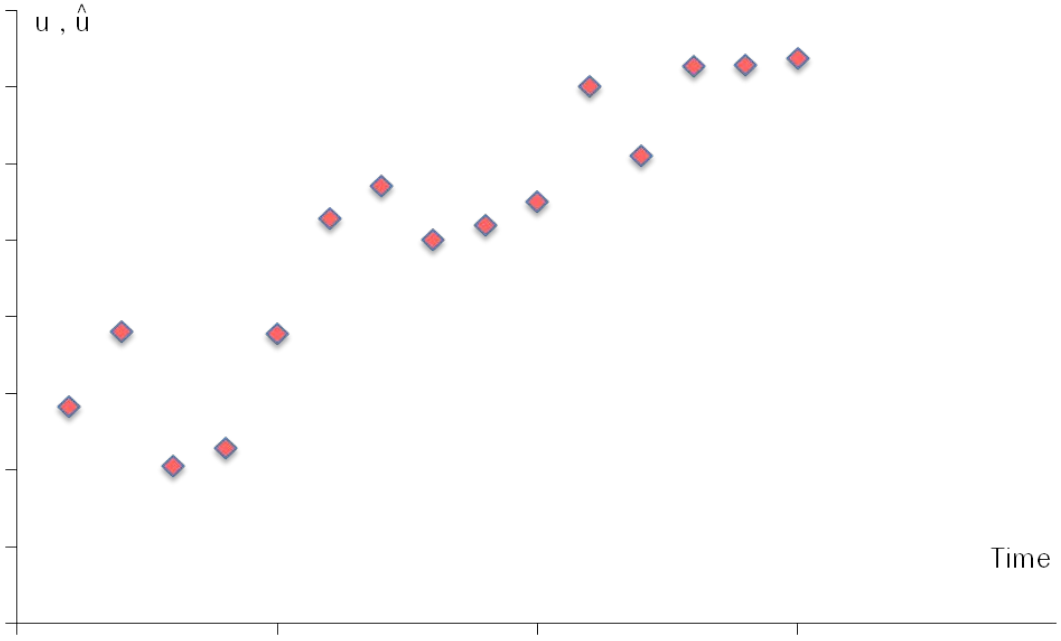
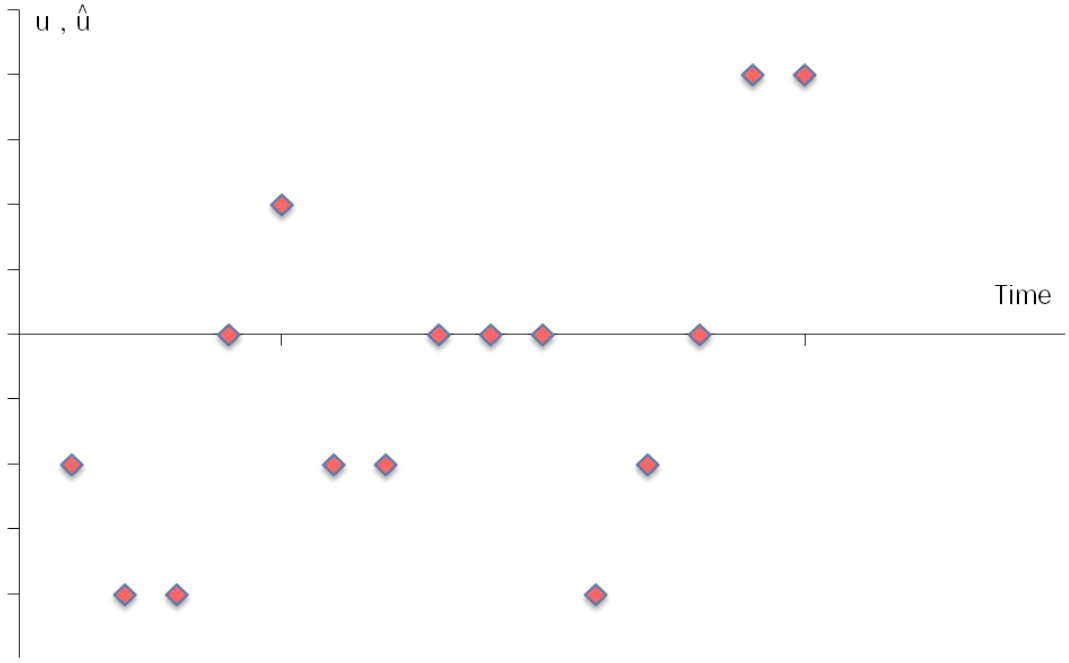


Figure 8-2: No autocorrelation among disturbance term



Autocorrelation among random disturbance terms stems from many factors. The main causes are when the model or data used in the model have the following properties.

1. Usually, the autocorrelation problem is more frequently found in the model where time series data is used than where cross-sectional one is used. The reason is that cross-sectional data involves a greater variety of observations which tend to be independent from one another. The consumption expenditure of people in an entire country, for instance, is diverse. Any factors liable to cause an error may be negligible when the data of the entire country is employed. On the other hand, for time series data, the same sample is studied across time. Mostly, this fact results in the relationship among observations. To illustrate, macroeconomic data may indicate a positive sign in the recovery period and this trend may be prolonged until any external shock leading to economic crisis.

2. Model misspecification, where the important regressors are omitted, could bring about the autocorrelation problem. For instance, consider the model explaining the demand for chicken with essential regressors including its price and the price of pork, as the substitute product.

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + u_t$$

where

Y_t = demand for chicken
X_{2t} = price of chicken
X_{3t} = price of pork

Unfortunately, suppose we wrongly specify the model such that the regressor X_{3t} is dropped and the model becomes

$$Y_t = \beta_1 + \beta_2 X_{2t} + \nu_t$$

where $\nu_t = \beta_3 X_{3t} + u_t$. It can be seen that the random disturbance term in this misspecified model (ν_t) incorporates the relationship of demand for chicken and the price of pork. This characteristic could result in significant pattern in disturbance term, leading to autocorrelation problem. The autocorrelation in this case is called **false autocorrelation** since the problem is not originated from the disturbance term itself but model misspecification instead.

3. Model misspecification, where the functional form is incorrect, could give rise to the autocorrelation problem as well. Consider the model of marginal cost which depends on the amount of goods produced.

$$MC_i = \beta_1 + \beta_2 Output_i + \beta_3 Output_i^2 + u_i$$

However, suppose the model is mistakenly specified as

$$MC_i = \alpha_1 + \alpha_2 Output_i + \nu_i$$

In this case, the random disturbance term is $\nu_t = \beta_3 Output_i^2 + u_i$. The result occurring is be similar to the case where the crucial regressors are neglected from the model. That is, a systematic pattern can be observed in the random disturbance term. The resulting autocorrelation is also called false autocorrelation.

4. **Cobweb phenomenon** might be another cause. For instance, some economists believe that supply of agricultural product displays the cobweb pattern. That is, the supplier of agricultural product makes a decision based on the last-year price as the production process takes time to deploy. The farmers have to decide first which types of plant will be produced and then production process will be carried out. Hence, they tend to base their decision on the price in the period when the type of plants is chosen rather than the price when the product is marketed.

If the price of one plant in the last period is high, there will be a great incentive for farmers to produce that plant. The product will, then, flood the market, forcing its price to go down. Contrarily, if the price of that plant in the last period is low, that plant will become unprofitable to produce in the view of farmers. This probably results in deficiency of the product, raising the price of the plant. Accordingly, the current amount of agricultural product will rely on the price last year. With the predictable pattern of regressor and regressand, the random disturbance term may display that systematic pattern, leading to the autocorrelation problem.

5. The transformation of data into first-difference form may inflict the autocorrelation problem on the model. Consider simple regression model. It is obvious that, when the model is established through first-difference transformation, autocorrelation in random disturbance term will result.

$$
\begin{aligned}
&\text{At (level form)} \\
&\qquad Y_t &&= \beta_1 + \beta_2 X_t + u_t \\
&\qquad Y_{t-1} &&= \beta_1 + \beta_2 X_{t-1} + u_{t-1} \\
&\text{At (difference form)} \\
&\qquad Y_t - Y_{t-1} &&= (\beta_1 + \beta_2 X_t + u_t) - (\beta_1 + \beta_2 X_{t-1} + u_{t-1}) \\
&\qquad \Delta Y_t &&= \beta_2 \Delta X_t + \Delta u_t \\
&\qquad \Delta Y_t &&= \beta_2 \Delta X_t + \nu_t \\
&\text{where} \quad \nu_t &&= \Delta u_t = (u_t - u_{t-1}) \\
&\\
&\text{suppose} \quad E(u_t) &&= 0 \\
&\text{thus} \quad E(\nu_t) &&= E(u_t - u_{t-1}) \\
&\qquad\qquad &&= E(u_t) - E(u_{t-1}) \\
&\\
&\qquad Var(\nu_t) &&= Var(u_t - u_{t-1}) \\
&\qquad\qquad &&= Var(u_t) - Var(u_{t-1}) \\
&\qquad\qquad &&= 2\sigma_u^2
\end{aligned}
$$

$$\begin{aligned} cov(\nu_t, \nu_{t-1}) &= E(\nu_t - \nu_{t-1}) \\ &= E[u_t - u_{t-1}][u_{t-1} - u_{t-2}] \\ &= -\sigma_u^2 \end{aligned}$$

8.2 CONSEQUENCE OF AUTOCORRELATION

Since the autocorrelation in random disturbance term violates classical linear regression model assumption, the following result can be shown in this section. Consider the simple linear regression model with X as explanatory variable, Y as explained variable, and the random disturbance term at time t u_t has the relationship with the one-period-lagged disturbance term u_{t-1}.

$$Y_i = \beta_1 + \beta_2 X_i + u_i \tag{8.2}$$

$$u_t = \rho u_{t-1} + \epsilon_t, \ -1 < \rho < 1 \tag{8.3}$$

where ρ is the coefficient of autocovariance which specifies the degree of relationship between the disturbance term at one period and the term at lagged period. Let the value of ρ ranges between -1 and 1. According to Equation 8.3, this kind of relationship is called **first-order autoregressive: AR(1)**, namely the lag period is one. Also, if the maximum lag period is two, we call it AR(2). Generally, with the maximum lag period of p, we can write the autoregressive model as Equation 8.4.

$$u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + ... + \rho_p u_{t-p} + \epsilon_t \tag{8.4}$$

ϵ is the **white noise error term** in the autoregressive model with the following properties.

$$\begin{aligned} E(\epsilon_t) &= 0 \\ var(\epsilon_t) &= \sigma_\epsilon^2 \\ cov(\epsilon_t, \epsilon_{t+s}) &= 0 \end{aligned}$$

First, consider the first-order autoregressive in Equation 3.8. The variance of the disturbance term u_t is equal to $\frac{\sigma_\epsilon^2}{1-\rho^2}$

$$\begin{aligned} E(u_t) &= \rho E(u_{t-1}) + E(\epsilon_t) = 0 \\ Var(u_t) &= E(u_t^2) \\ &= \rho^2 Var(u_{t-1}) + Var(\epsilon_t) \\ \therefore Var(u_t) &= \tfrac{\sigma_\epsilon^2}{1-\rho^2} \\ \text{due to} & \\ Var(u_t) &= Var(u_{t-1}) = \sigma_u^2 \text{ due to homoscedaticity} \\ Var(\epsilon_t) &= \sigma_\epsilon^2 \end{aligned}$$

It can be seen that, if the coefficient of autocovariance ρ is equal to 1 or -1, the variance of the error term will be undefined. We have to specify the value of ρ between this range to make the disturbance term stationary. Otherwise, the disturbance term may deviate from what it should be, namely in non-stationary disturbance term, so that the regression analysis is inapplicable.

Consider simple regression model without first-order autoregressive disturbance term.

$$Y_t = \beta_1 + \beta_2 X_t + u_t$$

Though OLS method, the estimator has the following close form solution.

$$\hat{\beta}_2 = \frac{\sum(X_t - \bar{X})(Y_t - \bar{Y})}{\sum(X_t - \bar{X})^2} = \frac{\sum x_t y_t}{\sum x_t^2}$$

The variance of the estimator can be calculated by the following formula.

$$Var(\hat{\beta}_2) = \frac{\sigma^2}{\sum(X_t - \bar{X})} = \frac{\sigma^2}{\sum x_t^2}$$

Nevertheless, under AR(1) scheme, the variance of the estimator can be computed by the Equation 8.5.

$$Var(\hat{\beta}_2)_{AR(1)} = \frac{\sigma^2}{\sum x_t^2}[1 + 2\rho\frac{\sum x_t x_{t-1}}{\sum x_t^2} + 2\rho^2\frac{\sum x_t x_{t-2}}{\sum x_t^2} + \cdots + 2\rho^{n-1}\frac{\sum x_t x_n}{\sum x_t^2}] \quad (8.5)$$

where

$$\begin{aligned} x_t &= (X_t - \bar{X}) \\ x_{t-1} &= (X_{t-1} - \bar{X}) \\ &\vdots \\ x_n &= (X_n - \bar{X}) \end{aligned}$$

The difference between the two above situation is that the variance under AR(1) scheme will be higher. In Equation 8.5, as ρ is equal to zero, or there is no autocorrelation in disturbance term, Equation 8.5 will converge to the usual formula of the variance of the estimator. Hence, due to autocorrelation problem, the OLS estimators will not possess best property, namely the estimator will not have minimum variance. Still, the OLS estimators will be unbiased.

8.3 DETECTION OF AUTOCORRELATION

When the autocorrelation problem leads to some undesirable properties of the estimators, the conclusion drawn from hypothesis testing may be misleading. Therefore, to prevent this mistake, we should examine whether the model suffers this problem. There are many approaches used to detect this problem and some of them are discussed in this section.

1. *Finding the relationship among disturbance terms by graph*: as depicted in Figure 8-1 and 8-2, if there is autocorrelation, the pattern of disturbance term across time will have systematic form.

2. *t-test for autocorrelation*: we can examine AR(1) model or $u_t = \rho u_{t-1} + \epsilon_t$ and the independent variables in the model have to be **strictly exogenous**. That is,

$$E(u_t|X_{2t}, X_{3t}, ..., X_{kt}) = 0$$

or

$$cov(u_t, X_{jt}) = 0 \text{ where } j = 2, 3, ..., k$$

When the independent variables are strictly exogenous, we can set the following hypothesis as

$$H_o : \rho = 0 \text{ and } H_a : \rho \neq 0$$

with the following test procedure.

Step 1: estimate the model of interest through OLS method to obtain $\hat{u}_t$

Step 2: construct AR(1) model with dependent variable u_t and independent variable u_{t-1} and, then, estimate the model to obtain the estimated value of ρ.

Step 3: calculate t-statistic of the estimator $\hat{\rho}$ and test for statistical significance. If we can reject the null hypothesis, that means there is the first order autocorrelation among disturbance terms.

This approach can be applied to the test for higher order of autoregressive model like $u_t = \rho u_{t-3} + \epsilon_t$ which includes setting hypothesis, computing t-statistic, comparing it with the critical value in statistical table, and drawing the conclusion of whether the null hypothesis should be rejected.

3. *Durbin-Watson test*: DW-statistic test is one of the popular methods used to detect first order autocorrelation problem. Six assumptions are required for DW-test.

Assumption 1: the model has to include the intercept

Assumption 2: explained variable X is non stochastic

Assumption 3: the relationship of the disturbance term has to be generated by AR(1) process.

Assumption 4: the disturbance term u_t is normally distributed.

Assumption 5: the model under examination does not include lagged regressand Y_{t-1} as the explanatory variable. That is, if the model follows equation below, we cannot apply DW-statistic to the test for autocorrelation.

$$Y_t = \beta_1 + \beta_2 X_{2t} + \alpha Y_{t-1} + u_t$$

Assumption 6: There is no missing data.

When all the assumptions are satisfied, DW-statistic can be computed by Equation 8.6.

$$DW = \frac{\sum_{t=2}^{n}(\hat{u}_t - \hat{u}_{t-1})^2}{\sum_{t=1}^{t=n} \hat{u}_t^2} \approx 2(1 - \hat{\rho}) \tag{8.6}$$

$$\hat{\rho} = \frac{\sum \hat{u}_t \hat{u}_{t-1}}{\sum \hat{u}^2} \tag{8.7}$$

where $\hat{\rho}$ ranges from -1 to 1, causing DW-statistic to range from 0 to 4. Consider the following possible cases. If $\hat{\rho}$ approach 0, DW will approach 2, that is no first-order autocorrelation among disturbance terms. If $\hat{\rho}$ approach 1, DW will approach 0, that is positive first-order autocorrelation among disturbance terms. If $\hat{\rho}$ approach -1, DW will approach 4, that is negative first-order autocorrelation among disturbance terms.

After DW-statistic is obtained, we can use it to test the following hypothesis.

$$H_o : \rho = 0 \text{ and } H_a : \rho \neq 0$$

The DW-statistic has to be compared with DW-statistical table invent in 1950 in order to draw the conclusion about the test. The critical value will include d_L and d_u which are the upper and lower bound respectively. The degree of freedom $k - 1$ (which is the amount of explanatory variables excluding intercept term) and level of significance (α) of 0.1, 0.05 and 0.01 will vary according to the circumstance. The comparison of DW-statistic to the critical value provides two beneficial insights.

If it turns out that the estimate of the coefficient of autocovariance is greater than 0, or equivalently DW-statistic is lower than 2, it can be the suspected that the random disturbance terms may have **positive autocorrelation**. Hence, the null and alternative hypothesis can be set as

$$H_o : \rho \leq 0 \text{ and } H_a : \rho > 0$$

We cannot reject the null hypothesis when calculated DW-statistic is greater than d_U. We reject the null hypothesis when calculated DW-statistic is lower than d_L. That is, disturbance term has no positive serial correlation at the level of significance α. Nevertheless, if DW-statistic lies between d_L and d_U ($d_L \leq DW \leq d_U$), we cannot conclude whether the disturbance term has positive serial correlation.

If, on the other hand, it appears that the estimate of the coefficient of autocovariance is less than 0, or equivalently DW-statistic is greater than 2, it can be the suspected that the random disturbance terms may have **negative autocorrelation**. Hence, the null and alternative hypothesis can be set as

$$H_o : \rho \geq 0 \text{ and } H_a : \rho < 0$$

We cannot reject the null hypothesis when calculated DW-statistic is greater than $4 - d_U$. We reject the null hypothesis when calculated DW-statistic is greater than $4 - d_L$. That is, disturbance term has no positive serial correlation at the level of significance α. Nevertheless, if DW-statistic lies between $4 - d_L$ and $4 - d_U$ ($4 - d_U \leq DW \leq 4 - d_L$), we cannot conclude whether the disturbance term has negative serial correlation. Figure 8-3 illustrates the criterion whether reject or not reject the null hypothesis.

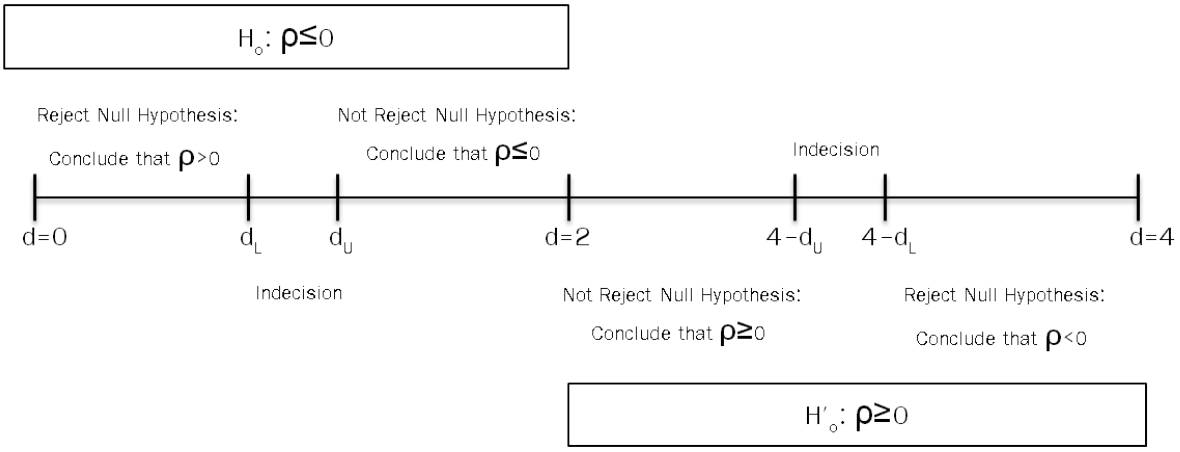
In short, the procedure to test for first order autocorrelation among disturbance terms with DW-statistic involves the following steps.

Step 1: estimate the model of interest to find $\hat{u}_t$

Step 2: calculate DW-statistic by the formula in Equation 8.6

Step 3: compare DW-statistic with the critical d from the table with the criterion shown in Figure 8-3 to achieve the conclusion about the relationship among disturbance terms.

Figure 8-3: Criterion for rejecting or not rejecting the null hypothesis with DW-statistic



4. **Breusch-Godfrey test**: this method can be used to test for any order, from 1 or AR(1) to p or AR(p) of autocorrelation as illustrated in Equation 8.4 from the simple regression model

$$u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + ... + \rho_p u_{t-p} + \epsilon_t \tag{8.4}$$

We can form the hypothesis about the relationship among these disturbance terms as
H_o : $\rho_1 = \rho_2 = ... = \rho_p = 0$
H_a : otherwise
with the following procedure.

Step 1: establish the regression model of interest. The example shown is the simple one.

Step 2: establish another model to obtain the relationship among disturbance terms with $\hat{u}_t$ as the explained variable and X_{2t} and $\hat{u}_{t-1}$ $\hat{u}_{t-2}$ until $\hat{u}_{t-p}$ as the explanatory variables.

$$\hat{u}_t = \alpha_1 + \alpha_2 X_{2t} + \hat{\rho}_1 \hat{u}_{t-1} + \hat{\rho}_2 u_{t-2} + ... + \hat{\rho}_p u_{t-p} + \epsilon_t \tag{8.8}$$

Step 3: if the sample size is large, LM-statistic can be computed by

$$LM = (n-p)R^2 \sim \chi^2_p \tag{8.9}$$

Step 4: compare LM-statistic with the critical value of chi-square table to conclude whether the null hypothesis should be rejected. Specifically, we reject the null hypothesis if $(n-p)R^2$ is greater than the critical chi-square at the chosen level of significance and conclude that there exists the autocorrelation problem.

8.4 REMEDIAL MEASURE FOR AUTOCORRELATION

After the problem is detected, to prevent the problem from making the variance of estimators so unreliable that the conclusion drawn from hypothesis test is misleading, the remedial measure is necessary. In this section, only the remedial measures for the first order autocorrelation are discussed.

1. *Generalized least square (GLS)*: consider the simple regression model with the first order autocorrelation problem AR(1)

$$Y_t = \beta_1 + \beta_2 X_t + u_t$$

$$u_t = \rho u_{t-1} + \epsilon_t$$

We can separate the procedure into the case where the coefficient of autocovariance (ρ) is known and unknown. For the case when the ρ **is known**, we can solve the problem by

$$\begin{aligned} Y_{t-1} &= \beta_1 + \beta_2 X_{t-1} + u_{t-1} \\ \rho Y_{t-1} &= \rho\beta_1 + \rho\beta_2 X_{t-1} + \rho u_{t-1} \end{aligned}$$

$$\begin{aligned} (Y_t - \rho Y_{t-1}) &= \beta_1(1-\rho) + \beta_2(X_t - \rho X_{t-1}) + (u_t - \rho u_{t-1}) \\ (Y_t - \rho Y_{t-1}) &= \beta_1(1-\rho) + \beta_2(X_t - \rho X_{t-1}) + \epsilon_t \end{aligned}$$

$$Y_t^* = \beta_1^* + \beta_2^* X_t^* + \epsilon_t$$

where

$$\begin{aligned} \epsilon_t &= u_t - \rho u_{t-1} \\ Y_t^* &= Y_t - \rho Y_{t-1} \\ X_t^* &= X_t - \rho X_{t-1} \\ \beta_1^* &= \beta_1(1-\rho) \end{aligned}$$

According to the above property of ϵ_t, we find that the classical linear regression model assumption is satisfied. Hence, it can be concluded that, the estimators in the new model generated from the above procedure are best, linear and unbiased estimators (BLUE).

Notwithstanding, due to the above procedure, the regressor and regressand of the new model is in the difference form which means that one observation is lost. Thus, to mitigate the problem, the first observation on X and Y may be transformed to $X_1^* = X_1\sqrt{1-\rho^2}$ and $Y_1^* = Y_1\sqrt{1-\rho^2}$. We call this transformation process Prais-Winsten transformation.

For the case when the ρ **is unknown**, two solutions to the problem is available. The first one is *first difference method*. This approach is usually used when DW-statistic for the model of interest is less than R^2. The first-difference model is constructed as

$$\begin{aligned} Y_t - Y_{t-1} &= \beta_1 - \beta_1 + \beta_2(X_t - X_{t-1}) + (u_t - u_{t-1}) \\ \Delta Y &= \beta_2 \Delta X_t + \epsilon_t \end{aligned}$$

When the above regressor and regressand are obtained, we can apply the OLS method for non-intercept model to estimate this new model which solves the first order autocorrelation problem.

The other method is *to use estimator $\hat{\rho}$ obtained from the calculation of DW-statistic*. From Equation 8.6 and 8.7, when the sample size is large enough, we can compute the value of $\hat{\rho}$ through

$$\hat{\rho} = 1 - \frac{d}{2} \tag{8.10}$$

After we obtain the estimate, we can apply it to the remedial measure stated above when the value of ρ is known.

2. *Heteroscedasticity and autocorrelation-consistent standard error*: when the sample size is large enough, Newey and West suggest the formula for the variance of estimators when the autocorrelation and heteroscedasticity problem occur. With this formula, the standard deviation required in statistical analysis, such as hypothesis test, confidence interval and so forth, will be applicable.

At the present time, most econometric computer packages provide this estimation of variance in the set of statistical results. Although this method does not lessen the degree of autocorrelation problem, the obtained standard error is fixed and applicable for further statistical analysis.

It is noteworthy that the difference between the method of White and Newey-West is recognized. The approach suggested by White can solve the specific problem of heteroscedasticity; whereas the one suggested by Newey-West is designed to tackle the problems of both heteroscedasticity and autocorrelation.

Chapter 9

SPECIFICATION ERROR

9.1 TYPES OF SPECIFICATION ERROR

One of the assumptions under classical linear regression model (CLRM) to obtain the desirable properties of OLS estimators, namely BLUE, is no specification error. In this chapter, many types of specification error are discussed as well as the consequence resulting from these errors.

The possible errors from misspecification are categorized into 4 types which are omission of necessary variables, inclusion of unnecessary variables, adoption of the wrong functional form and error of measurement. Each type of error is discussed sequentially.

Omission of necessary variables is when an economist fails to include the important regressors to the model for studying the economic theory. Suppose the true model follows Equation 9.1.

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i i \tag{9.1}$$

However, the economist mistakenly specifies the model as in Equation 9.2. This is the situation where the necessary regressor is omitted from the model.

$$Y_i = \alpha_1 + \alpha_2 X_{2i} + \nu_i \tag{9.2}$$

Inclusion of unnecessary variables is when some variables with insignificantly explanatory power enter the model. For instance, if the model in Equation 9.1 is the true form where regressors X_2 and X_3 are sufficient to explain the regressand Y. Unluckily, an economist includes the variable X_4 to the model and become Equation 9.3.

$$Y_i = \gamma_1 + \gamma_2 X_{2i} + \gamma_3 X_{3i} + \gamma_4 X_{4i} + u_i' \tag{9.3}$$

Adoption of the wrong functional form is another type of error. Suppose Equation 9.1 is the true functional form. That is, the relationship between regressor and regressand is linear. Nevertheless, if the functional form is specified as Equation 9.4, this kind of problem will occur.

$$Y_i = \lambda_1 + \lambda_2 X_{2i} + \lambda_3 X_{3i} + \lambda_4 X_{2i}^2 + u_i'' \tag{9.4}$$

Last but not least, **error of measurement** is when the proxy of either regressor X or regressand Y or both which may contain the error of measurement. For example, suppose that, theoretically, the variables Y, X_2 and X_3 should be used in the model. Yet, practically, it may be impossible to obtain the value of those variables. Therefore, the variable Y^*, X_2^* and X_3^* are used as proxies for the true model with the belief that $Y_i^* = Y_i + \epsilon_i$ where ϵ_i denotes the measurement error (The same is valid for X_2 and X_3.)

The consequence and the detection of four types of specification error are discussed in the following sections.

9.2 CONSEQUENCE OF SPECIFICATION ERROR

9.2.1 Omission of Necessary Variables

For the omission of necessary variables where the true model is Equation 9.1 but the model is wrongly specified as Equation 9.2, the random disturbance term of 9.2 can be modelled as

$$\nu_i = \beta_3 X_{3i} + u_i \tag{9.5}$$

The disturbance term ν_i will contain the effect of the omitted variable X_3. The estimation of the wrong model is shown in Equation 9.6.

$$Y_i = \hat{\alpha}_1 + \hat{\alpha_2} X_{2i} \tag{9.6}$$

The consequences of the omission of this variable on Equation 9.6 are as follows.

1. If the omitted independent variable X_3 has the relationship with the independent variable X_2 in the model of interest, the estimator $\hat{\alpha}_1$ and $\hat{\alpha_2}$ will be biased and inconsistent. Even though the sample size gets larger, the estimators still face the same problem.

$$\begin{aligned} E(\hat{\alpha}_1) &\neq \beta_1 \\ E(\hat{\alpha}_2) &\neq \beta_2 \end{aligned}$$

2. Although the omitted independent variable X_3 has no relationship with the independent variable X_2, the estimator $\hat{\alpha}_1$ will be biased but the estimator $\hat{\alpha_2}$ will be not.

$$\begin{aligned} E(\hat{\alpha}_1) &\neq \beta_1 \\ E(\hat{\alpha}_2) &= \beta_2 \end{aligned}$$

3. The estimated variance of random disturbance term will not reflect the true value which, in turn, causes the variance of the estimators to be wrongly estimated. The result is the misleading conclusion from hypothesis test and confidence interval.

9.2.2 Inclusion of Unnecessary Variables

For the inclusion of necessary variables, consider again the case where the true model is Equation 9.1 but the model is wrongly specified as Equation 9.3 in which the independent variable X_4, with insufficient explanatory power, is included. When the model in Equation 9.3 is estimated, the result is Equation 9.7.

$$Y_i = \hat{\gamma}_1 + \hat{\gamma_2} X_{2i} + \hat{\gamma}_3 X_{3i} + \hat{\gamma}_4 X_{4i} \tag{9.7}$$

The consequences on the model are as follows.

1. The estimators will be unbiased and consistent. In other word, the estimators according to OLS method will reflect the true parameters but the estimator $\hat{\gamma}_4$, as the sample size get larger, will approach zero. That means the regressor X_4 should not be included in this regression model or

$$\begin{aligned} E(\hat{\gamma}_1) &= \beta_1 \\ E(\hat{\gamma}_2) &= \beta_2 \\ E(\hat{\gamma}_3) &= \beta_3 \\ E(\hat{\gamma}_4) &= \beta_4 = 0 \end{aligned}$$

2. The estimated variance of the random disturbance term will reflect the true value of parameter which, in turn, makes the variance of the estimators reliable and applicable for further statistical analysis without misleading conclusion. Nevertheless, the variance of the estimators in this wrongly specified model will be higher than the one in the correct model.

9.2.3 Adoption of Wrong Functional Form

Consider the true regression model to explain the total cost of production for a single good which depends on the amount of good produced as the polynomial function in Equation 9.8. Let regressand Y be the total cost and regressor X the amount of good produced.

$$Y_i = \beta_1 + \beta_2 X_i + \beta_3 X_i^2 + \beta_4 X_i^3 + u_i \tag{9.8}$$

Nonetheless, suppose the functional form of total cost function is erroneously specified such that X_i^2 and X_i^3 are dropped from the model resulting in the model shown as Equation 9.9.

$$Y_i = \alpha_1 + \alpha_2 X_i + \nu_i \tag{9.9}$$

Consider Equation 9.9. It can be seen that ν_i will cover the squared and cubed term omitted from the original model, namely $\nu_i = \beta_3 X_i^2 + \beta_4 X_i^3 + u_i$. In this case, the consequence will be similar to the case of omission of necessary variables. The effect on both estimators of $\hat{\alpha_1}$ and $\hat{\alpha_2}$ will, hence, be similar to the case studied in Subsection 9.2.1. That is, the estimators will be biased and inconsistent and the statistical analysis will be impaired.

The case of error of measurement will be separately discussed in Section 9.4.

9.3 DETECTION OF SPECIFICATION ERROR

There are many categories of specification error and the influence on the estimators varies according to each category. For the case of inclusion of unnecessary variables with insufficiently explanatory power, the consequence may not be so severe that the estimators in the model are inapplicable. For the omission of the necessary variables and adoption of wrong functional form, the impact may be so immense that the desirable property of the estimators is lost. The detection of specification error, therefore, is essential. There are many methods; yet, only four main ones are discussed here.

1. *Detecting the inclusion of unnecessary variables*: usual hypothesis test for statistical significance can be applied such as t-test and F-test. If the statistical result appears that the independent variables are not necessary, the implication is that the model may suffer specification error.

2. *Examination of residuals*: consider the model representing the total cost of production as Equation 9.8. The true specification of the model is Equation 9.8; but an economist wrongly specifies the model to be either Equation 9.9 or 9.10.

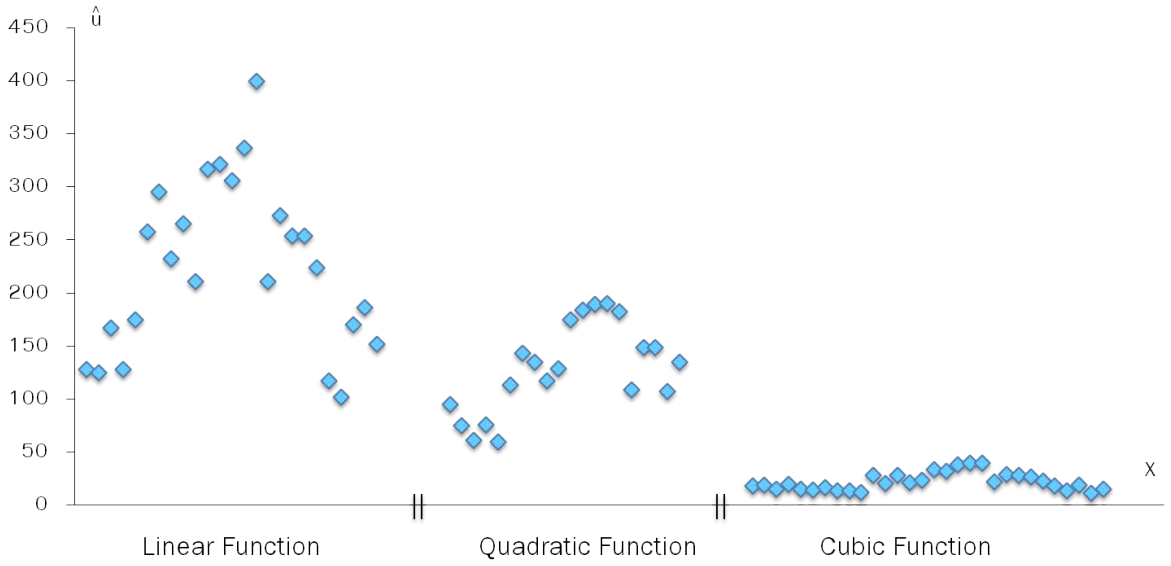
$$Y_i = \alpha_1 + \alpha_2 X_i + \nu_i \tag{9.9}$$

$$Y_i = \alpha_1' + \alpha_2' X_i + \alpha_3' X_i^2 + \nu_i' \tag{9.10}$$

$$Y_i = \beta_1 + \beta_2 X_i + \beta_3 X_i^2 + \beta_4 X_i^3 + u_i \tag{9.8}$$

Examining Equation 9.8, 9.9 and 9.10 in contrast, after the regression model is estimated, we can plot the residuals $\hat{u}_i$ against the independent variables to study the volatility of the residuals. If there seems to be high volatility involved, it can be concluded that the model may be erroneous specified. From Figure 9-1, it can be observed that the model from Equation 9.9 and 9.10 are more likely to undergo the problem of specification error than the one from Equation 9.8.

Figure 9-1: Comparison among different models



3. *Durbin-Watson statistic*: sometimes we deal with the specification error in the case of omission of necessary variables which may result in autocorrelation among disturbance terms. Hence, from Chapter 8, we can apply DW-statistic to detect this problem according to the following steps.

Step 1: estimate the model by ordinary least square method and the obtain the residuals from the model

Step 2: calculate DW-statistic by

$$d = \frac{\sum_{t=2}^{t=n} (\hat{u}_t - \hat{u}_{t-1})^2}{\sum_{t=1}^{t=n} \hat{u}_t^2} \tag{9.11}$$

Step 3: compare the critical d from the statistical to conclude whether the null hypothesis will be rejected or not. If the null hypothesis is rejected, it implies that there is autocorrelation among disturbance terms and, also, specification error.

4. *Ramsey regression specification error test*: this test focuses on the relationship between residuals ($\hat{u}$) and estimated dependent variable ($\hat{Y}$). The idea behind the test is that dependent variable (Y) has the relationship with independent variable (X) rather than the estimated dependent variable ($\hat{Y}$). The procedure for the test is stated as follows.

Step 1: construct the regression model like

$$Y_i = \beta_1 + \beta_2 X_{2i} + ... + \beta_k X_{ki} + u_i$$

Step 2: construct other regression model given that $\hat{Y}^2$ and $\hat{Y}^3$ are additional regressor in this new model. The reason we do not simply use $\hat{Y}$ because there may lead to the problem of perfect multicollinearity between regressor and regressand.

$$Y_i = \beta_1 + \beta_2 X_{2i} + ... + \beta_k X_{ki} + \delta_1 \hat{Y}_i^2 + \delta_2 \hat{Y}_i^3 + u_i \tag{9.12}$$

Step 3: form null and alternative hypothesis and calculate F-statistic from Equation 9.13

H_o : The model is correctly specified
H_a : Otherwise

$$\hat{F} = \frac{R^2_{new} - R^2_{old}/\text{The amount of additional parameter}}{(1 - R^2_{new})/(n - \text{The amount of parameter in the new model})} \tag{9.13}$$

Step 4: compare the F-statistic with the critical F from the table. If the F-statistic is greater than the critical value, the null hypothesis should be rejected and conclude that the additional regressors have significantly explanatory power. That means the original set of regressor does not sufficiently explain the regressand. In other word, the original model is mistakenly specified.

9.4 ERROR OF MEASUREMENT

9.4.1 Error of Measurement in Regressand

Consider simple regression model between regressor X and regressand Y'

$$Y_i' = \beta_1 + \beta_2 X_i + u_i \tag{9.14}$$

The disturbance term (u_i) satisfies CLRM assumptions. Nonetheless, the data of Y' cannot be collected; thus, the regressand Y is used as a proxy such that

$$Y_i = Y_i' + \epsilon_i$$

and ϵ_i is the error of the proxy regressand. Considering the disturbance term of the model in Equation 9.14, we find that

$$\begin{aligned} Y_i &= (\alpha + \beta X_i + u_i) + \epsilon_i \\ &= (\alpha + \beta X_i) + (u_i + \epsilon_i) \\ &= (\alpha + \beta X_i) + \nu_i \end{aligned}$$

the disturbance term ν_i of the model using the proxy Y is equivalent to the sum of the disturbance term from the true model and the error associated with the proxy.

To simplify the analysis of the influence on the estimators of the new model, assume that the following properties of the error term ϵ_t are satisfied.

$$\begin{aligned} E(u_i) &= E(\epsilon_i) = 0 \\ Cov(X_i, \epsilon_i) &= 0 \\ Cov(u_i, \epsilon_i) &= 0 \end{aligned}$$

Thanks to these assumptions, the disturbance term in the new model will, still, satisfy CLRM assumptions. The implication is that the estimators in the new model, even facing the measurement error from the use of proxy, will be unbiased.

Notwithstanding, comparing the variance of the estimator of β in the new model to the one in the old model, we find that

$$\begin{aligned} \text{True model: } Y_i' &= \alpha + \beta X_i + u_i \\ var(\hat{\beta}) &= \frac{\sigma_u^2}{\sum x_i^2} \\ \text{Proxy model: } Y_i &= (\alpha + \beta X_i) + \nu_i \\ var(\hat{\beta}) &= \frac{\sigma_\nu^2}{\sum x_i^2} = \frac{\sigma_u^2 + \sigma_\epsilon^2}{\sum x_i^2} \end{aligned}$$

In spite of unbiasedness property, the variance of the estimator in the true model will be less than the one in the proxy model. Still, this variance is applicable for further statistical analysis like confidence interval, individual hypothesis test (t-test) and overall hypothesis test (F-test). The impact is similar to the case where the model faces the problem of imperfect multicollinearity.

9.4.2 Error of Measurement in Regressor

For the error of measurement in regressor, consider the model in Equation 9.15 where the variable contains an error is the independent one.

$$Y_i = \beta_1 + \beta_2 X_i' + u_i \tag{9.15}$$

Again, suppose that the disturbance term (u_i) satisfies CLRM assumptions and the data of X' cannot be obtained; thus, the regressor X is used as a proxy such that

$$X_i = X_i' + w_i$$

the disturbance term w_i is the error of the proxy regressor. Considering the disturbance, we find that

$$\begin{aligned} Y_i &= \alpha + \beta(X_i - w_i) + u_i \\ &= \alpha + \beta X_i + (u_i - \beta w_i) \\ &= \alpha + \beta X_i + z_i \end{aligned}$$

where the disturbance term z_i of the model using the proxy Y is equivalent to the difference of the disturbance term from the true model (u_i) and the error associated with the proxy (w_i).

To examine whether the estimators in the model have the desirable properties, it can be seen that, even though the same assumptions of the error term are made as the case with the error of measurement in regressand, there is still the covariance between the disturbance term in the proxy model and the regressor.

$$\begin{aligned} cov(z_i, X_i) &= E[z_i - E(z_i)][X_i - E(X_i)] \\ &= E(u_i - \beta w_i)(w_i) \\ &= E(-\beta w_i^2) \\ &= -\beta \sigma_w^2 \end{aligned}$$

In this case, when one of the CLRM assumptions is violated, namely the disturbance term is not independent from the regressor, the estimators obtained from OLS method will be biases and inconsistent, no matter how large the sample size is.