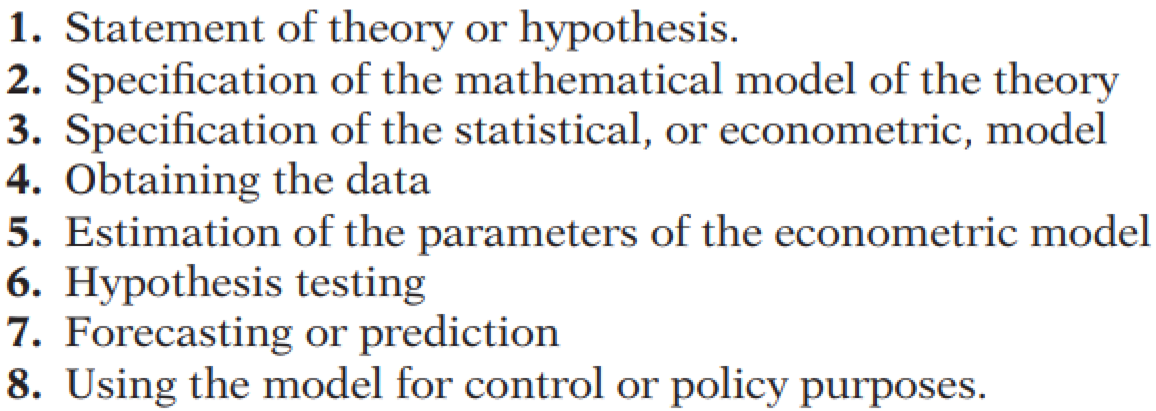
2HSOE52 Introduction to Economics

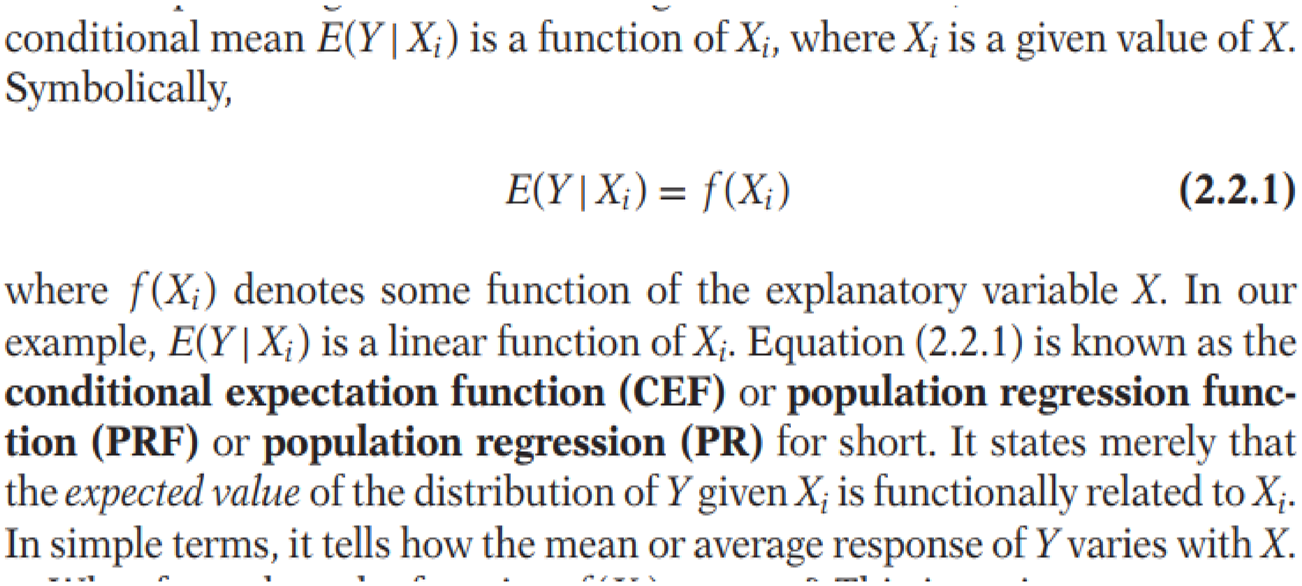
Chapter_3: The Simple Regression Model
: Classical Linear Regression Model (CLRM)

A. Methodology of Econometrics



1. Statement of theory or hypothesis.
2. Specification of the mathematical model of the theory
3. Specification of the statistical, or econometric, model
4. Obtaining the data
5. Estimation of the parameters of the econometric model
6. Hypothesis testing
7. Forecasting or prediction
8. Using the model for control or policy purposes.

B. THE CONCEPT OF POPULATION REGRESSION FUNCTION (PRF)



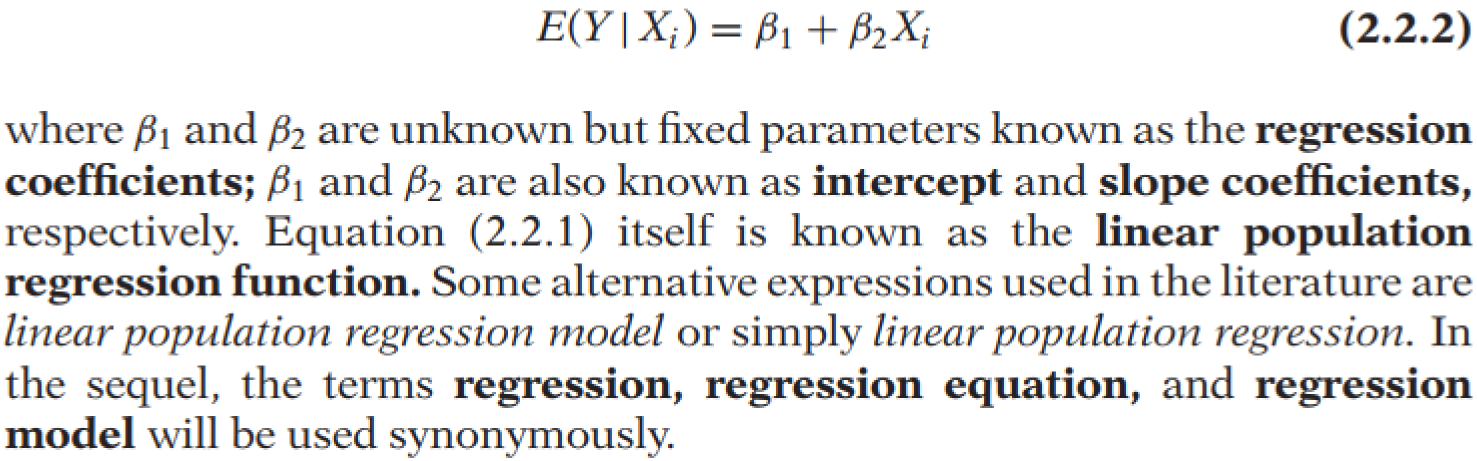
conditional mean $E(Y \mid X_i)$ is a function of X_i, where X_i is a given value of X. Symbolically,

$$E(Y \mid X_i) = f(X_i) \qquad \textbf{(2.2.1)}$$

where $f(X_i)$ denotes some function of the explanatory variable X. In our example, $E(Y \mid X_i)$ is a linear function of X_i. Equation (2.2.1) is known as the **conditional expectation function (CEF)** or **population regression function (PRF)** or **population regression (PR)** for short. It states merely that the *expected value* of the distribution of Y given X_i is functionally related to X_i. In simple terms, it tells how the mean or average response of Y varies with X.

Population Regression Function (PRF)*E*(*Y* | *Xi*) is a linear function of *Xi*.

Mathematically it is given by,



$$E(Y \mid X_i) = \beta_1 + \beta_2 X_i \qquad \textbf{(2.2.2)}$$

where β_1 and β_2 are unknown but fixed parameters known as the **regression coefficients;** β_1 and β_2 are also known as **intercept** and **slope coefficients,** respectively. Equation (2.2.1) itself is known as the **linear population regression function.** Some alternative expressions used in the literature are *linear population regression model* or simply *linear population regression*. In the sequel, the terms **regression, regression equation,** and **regression model** will be used synonymously.

C. THE MEANING OF THE TERM *LINEAR*

Linearity can be understood in two aspect

- ❑ **Linearity in the Variables**
- ❑ **Linearity in the Parameters**

A linear regression model is liner in parameters .

D: Econometric model of Population Regression

Econometric model of Population Regression Model is given by

$$
\begin{aligned}
Y_i &= E(Y \mid X_i) + u_i \\
&= \beta_1 + \beta_2 X_i + u_i
\end{aligned}
$$

Where E(*Y* | *Xi*) is the **systematic,** or **deterministic,** component, and $u_{i,}$ which is the random, or **nonsystematic, or stochastic** component.

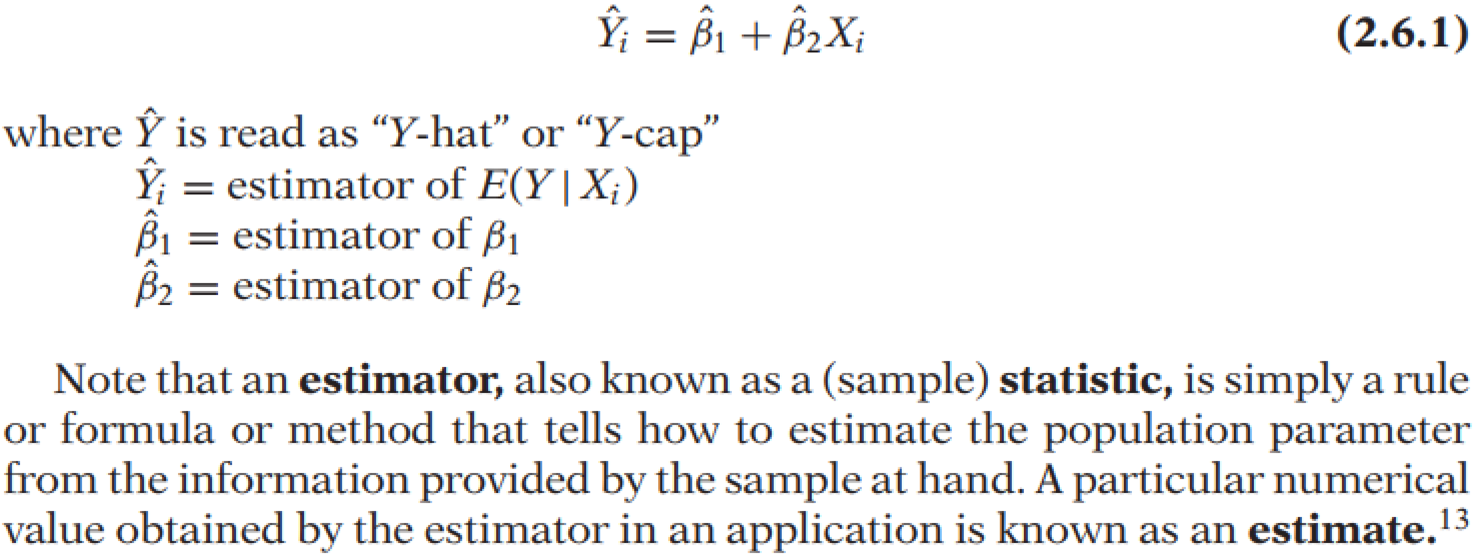
E. THE SIGNIFICANCE OF THE STOCHASTIC DISTURBANCE TERM

The reasons are many

1. ***Vagueness of theory***
2. ***Unavailability of data***
3. ***Core variables versus peripheral variables***
4. ***Intrinsic randomness in human behavior***
5. ***Poor proxy variables***
6. ***Principle of parsimony***
7. ***Wrong functional form***

F. THE SAMPLE REGRESSION FUNCTION (SRF)

Sample regression function (SRF) or sample regression line is the estimate of the PRF which may be written as:



$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i \qquad \textbf{(2.6.1)}$$

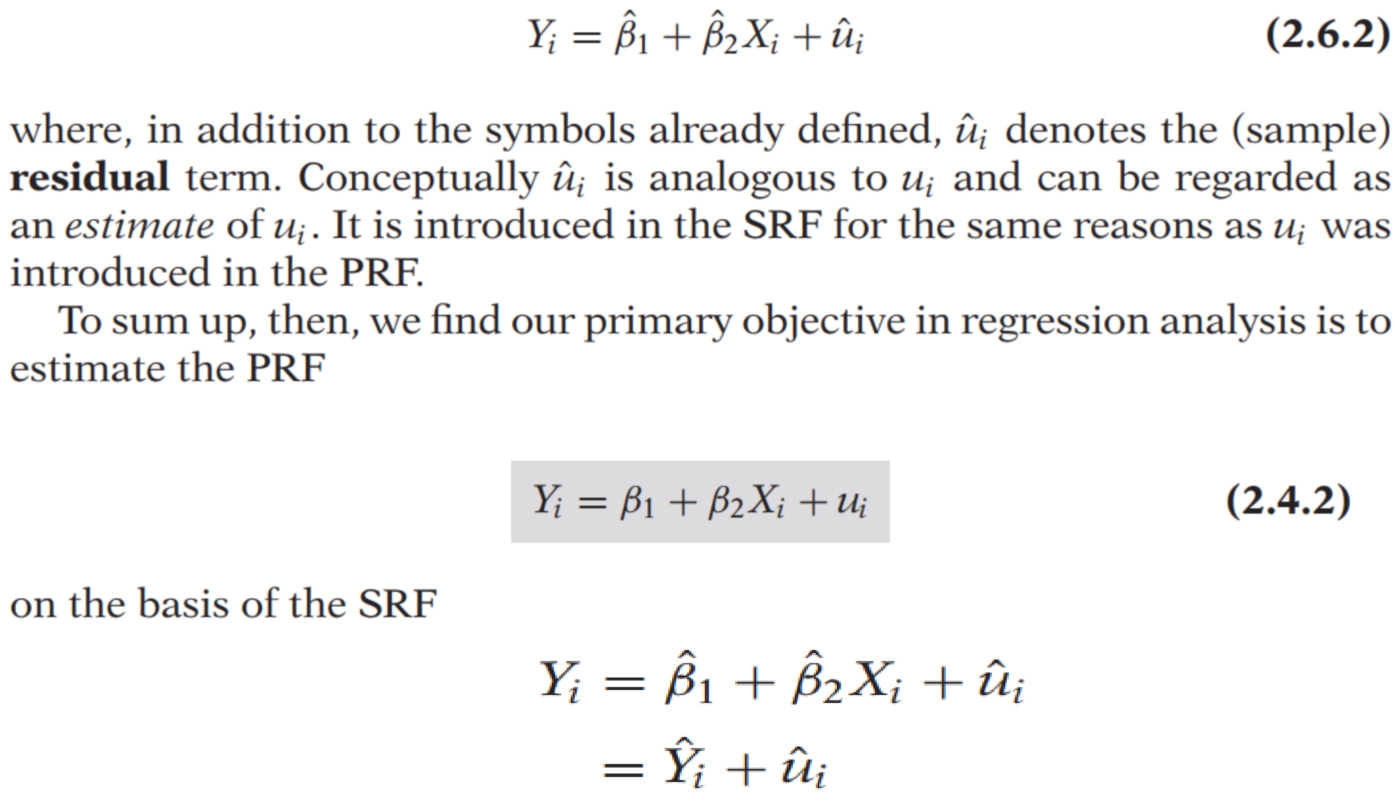
where $\hat{Y}$ is read as "Y-hat" or "Y-cap"

$\hat{Y}_i$ = estimator of $E(Y \mid X_i)$
$\hat{\beta}_1$ = estimator of β_1
$\hat{\beta}_2$ = estimator of β_2

Note that an **estimator,** also known as a (sample) **statistic,** is simply a rule or formula or method that tells how to estimate the population parameter from the information provided by the sample at hand. A particular numerical value obtained by the estimator in an application is known as an **estimate.**[13]

G. Econometric SRF

The SRF in its stochastic form may be written as:



$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + \hat{u}_i \tag{2.6.2}$$

where, in addition to the symbols already defined, $\hat{u}_i$ denotes the (sample) **residual** term. Conceptually $\hat{u}_i$ is analogous to u_i and can be regarded as an *estimate* of u_i. It is introduced in the SRF for the same reasons as u_i was introduced in the PRF.

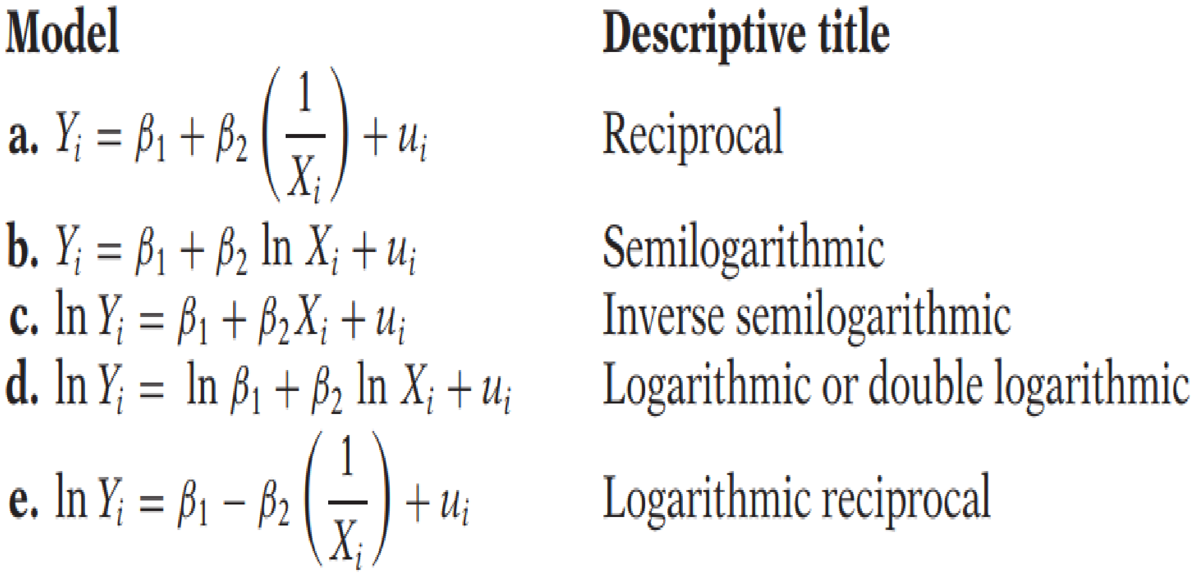
To sum up, then, we find our primary objective in regression analysis is to estimate the PRF

$$Y_i = \beta_1 + \beta_2 X_i + u_i \tag{2.4.2}$$

on the basis of the SRF

$$\begin{aligned} Y_i &= \hat{\beta}_1 + \hat{\beta}_2 X_i + \hat{u}_i \\ &= \hat{Y}_i + \hat{u}_i \end{aligned}$$

H. Some Linear Regression Models



Model	Descriptive title
a. $Y_i = \beta_1 + \beta_2 \left(\frac{1}{X_i}\right) + u_i$	Reciprocal
b. $Y_i = \beta_1 + \beta_2 \ln X_i + u_i$	Semilogarithmic
c. $\ln Y_i = \beta_1 + \beta_2 X_i + u_i$	Inverse semilogarithmic
d. $\ln Y_i = \ln \beta_1 + \beta_2 \ln X_i + u_i$	Logarithmic or double logarithmic
e. $\ln Y_i = \beta_1 - \beta_2 \left(\frac{1}{X_i}\right) + u_i$	Logarithmic reciprocal

I. TWO-VARIABLE REGRESSION MODEL: THE PROBLEM OFESTIMATION : Ordinary Least Square (OLS) Method

We have two-variable PRF:

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

When i=1, 2, 3, …, n. The PRF is not directly observable.

We estimate it from the SRF:

$$\begin{aligned} Y_i &= \hat{\beta}_1 + \hat{\beta}_2 X_i + \hat{u}_i \\ &= \hat{Y}_i + \hat{u}_i \end{aligned}$$

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

Or,

which shows that the *u*ˆ *i* (the residuals) are simply the differences between the actual and estimated *Y* values.

Now given *n* pairs of observations on *Y* and *X*, we would like to determine the SRF in such a manner that it is as close as possible to the actual *Y i.e. PRF*.

For this we have to chose that pair of estimators $\hat{\beta}_1$ and $\hat{\beta}_2$ for which sum squared residuals (RSS) $\sum \hat{u}_i^2$ is minimum (Ordinary Least Square Criterion).

Now, RSS=

$$\begin{aligned}\sum \hat{u}_i^2 &= \sum (Y_i - \hat{Y}_i)^2 \\ &= \sum (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)^2\end{aligned}$$

For Minimum RSS, the first order condition requires that

$$\frac{\partial \sum \hat{u}_i^2}{\partial \hat{\beta}_1} = 0 \longrightarrow \textcircled{A}$$

$$\frac{\partial \sum \hat{u}_i^2}{\partial \hat{\beta}_2} = 0 \longrightarrow \textcircled{B}$$

Now, eqn $\textcircled{A} \Rightarrow$

$$\frac{\partial \sum \hat{u}_i^2}{\partial \hat{\beta}_1} = \frac{\partial}{\partial \hat{\beta}_1} \sum (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)^2 = 0$$

$$\Rightarrow 2\sum (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)(-1) = 0$$

$$\Rightarrow \sum (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i) = 0$$

$$\Rightarrow \sum Y_i - \sum \hat{\beta}_1 - \hat{\beta}_2 \sum X_i = 0$$

$$\Rightarrow \frac{\sum Y_i}{n} - \frac{\sum \hat{\beta}_1}{n} - \frac{\hat{\beta}_2}{n} \sum X_i = 0 \left[\text{Dividing both side by } n \right]$$

$$\Rightarrow \bar{Y} - \hat{\beta}_1 - \hat{\beta}_2 \bar{X} = 0$$

$$\Rightarrow \bar{Y} - \hat{\beta}_1 - \hat{\beta}_2 \bar{X} = 0$$

$$\Rightarrow \hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X} \longrightarrow \textcircled{c}$$

Now Eqn $\textcircled{B}$ $\Rightarrow$

$$\frac{\partial \sum u_i^2}{\partial \hat{\beta}_2} = \frac{\partial}{\partial \hat{\beta}_2} \sum (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)^2 = 0$$

$$\Rightarrow 2\sum (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)(-X_i) = 0$$

$$\Rightarrow \sum (Y_i X_i - \hat{\beta}_1 X_i - \hat{\beta}_2 X_i^2) = 0$$

$$\Rightarrow \sum Y_i X_i - \hat{\beta}_1 \sum X_i - \hat{\beta}_2 X_i^2 = 0$$

$$\Rightarrow \sum Y_i X_i - (\bar{Y} - \hat{\beta}_2 \bar{X}) \sum X_i - \sum \hat{\beta}_2 X_i^2 = 0 \quad \text{(using eqn}_c\text{)}$$

$$\Rightarrow \sum Y_i X_i - \bar{Y} \sum X_i - \hat{\beta}_2 \bar{X} \sum X_i - \hat{\beta}_2 \sum X_i^2 = 0$$

$$\Rightarrow \sum x_i y_i - \bar{y}\sum x_i = \hat{\beta}_1 \left(\sum x_i^2 - \bar{x}\sum x_i\right)$$

$$\Rightarrow \hat{\beta}_1 = \frac{\sum x_i y_i - \bar{y}\sum x_i}{\sum x_i^2 - \bar{x}\sum x_i}$$

$$= \frac{\sum x_i y_i - \bar{y}\sum x_i - n\bar{x}\bar{y} + n\bar{x}\bar{y}}{\sum x_i^2 - 2\bar{x}\sum x_i + \bar{x}\sum x_i}$$

$$= \frac{\sum x_i y_i - \bar{y}\sum x_i - \bar{x}\sum y_i + n\bar{x}\bar{y}}{\sum x_i^2 - 2\bar{x}\sum x_i + \bar{x}\cdot n\bar{x}}$$

$$= \frac{\sum (x_i y_i - y x_i - x y_i + \bar{x}\bar{y})}{\sum (x_i^2 - 2\bar{x}x_i + \bar{x}^2)}$$

$$= \frac{\Sigma[X_i(Y_i-\bar{Y}) - \bar{X}(Y_i-\bar{Y})}{\Sigma(X_i-\bar{X})^2}$$

$$\therefore \hat{\beta}_2 = \frac{\Sigma(X_i-\bar{X})(Y_i-\bar{Y})}{\Sigma(X_i-\bar{X})^2} \rightarrow (1)$$

Let $X_i - \bar{X} = x_i$, $Y_i - \bar{Y} = y_i$

Then eqn (1) $\Rightarrow$

$$\hat{\beta}_2 = \frac{\Sigma x_i y_i}{\Sigma x^2}$$ which is the OLS estimate of β.

Further, Eqn (1) $$\hat{\beta}_2 = \frac{\frac{1}{n}\Sigma(X_i - X)(Y_i-\bar{Y})}{\frac{1}{n}\Sigma(X_i-X)^2}$$

$$= \frac{\text{CoV}(X\,Y)}{\text{Var}(X)}$$

putting $\hat{\beta}_2 = \frac{\sum x_i y_i}{\sum x_i^2}$ in eq^n $\textcircled{11}$. we get.

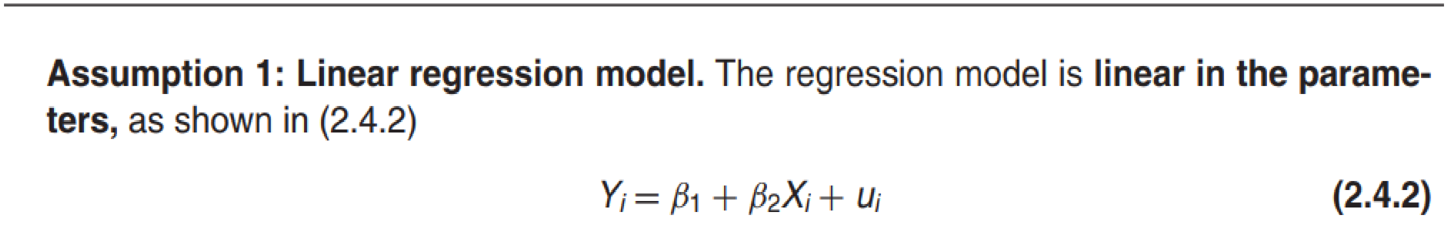
$$\hat{\beta}_1 = \bar{Y} - \frac{\sum x_i y_i}{\sum x_i^2}\bar{X}$$

$$= \frac{Y\sum x_i^2 - \bar{x}\sum x_i y_i}{\sum x_i^2},$$

which is the OLS estimator $\hat{\beta}_1$.

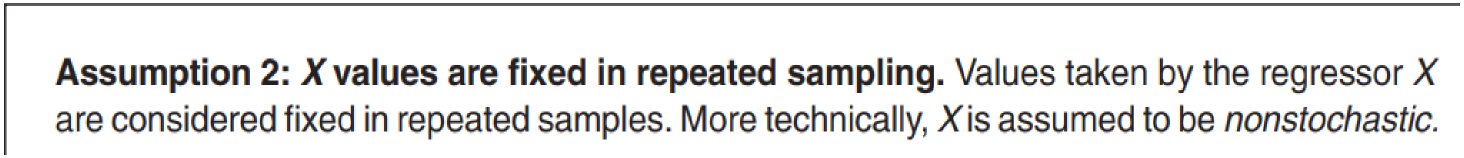
J. THE CLASSICAL LINEAR REGRESSION MODEL: THE ASSUMPTIONS UNDERLYING THE METHOD OF LEAST SQUARES

The Gaussian, standard, or classical linear regression model (CLRM), which is the cornerstone of most econometric theory, makes 10 assumptions .

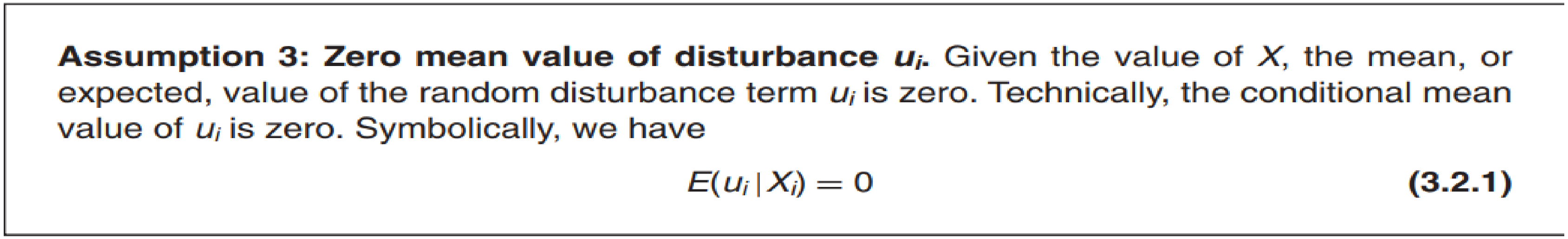


Assumption 1: Linear regression model. The regression model is **linear in the parameters,** as shown in (2.4.2)

$$Y_i = \beta_1 + \beta_2 X_i + u_i \qquad \textbf{(2.4.2)}$$

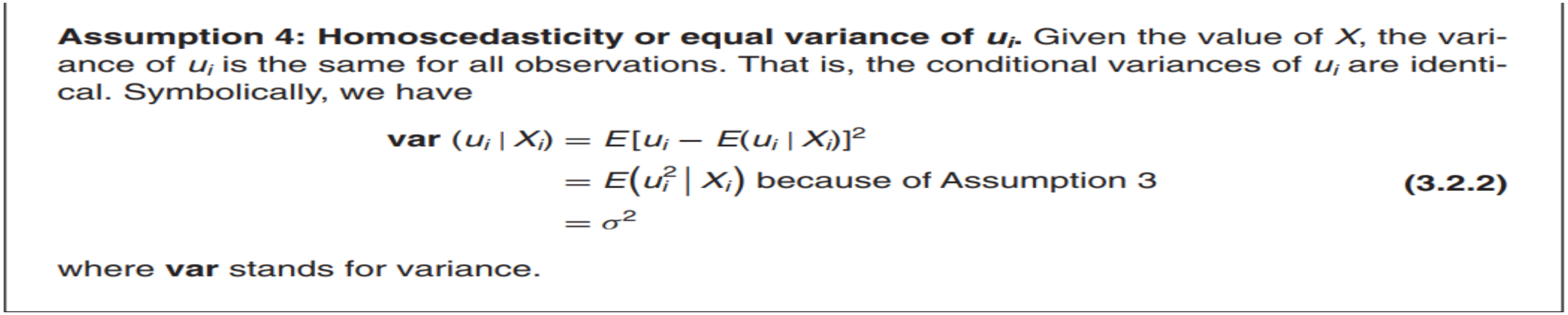


Assumption 2: *X* values are fixed in repeated sampling. Values taken by the regressor *X* are considered fixed in repeated samples. More technically, *X* is assumed to be *nonstochastic.*



Assumption 3: Zero mean value of disturbance u_i. Given the value of X, the mean, or expected, value of the random disturbance term u_i is zero. Technically, the conditional mean value of u_i is zero. Symbolically, we have

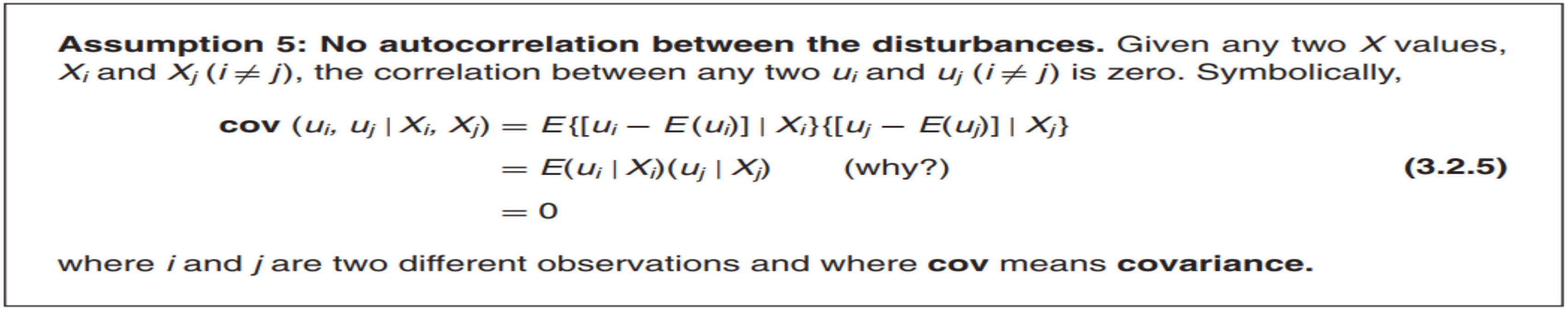
$$E(u_i \mid X_i) = 0 \tag{3.2.1}$$



Assumption 4: Homoscedasticity or equal variance of u_i. Given the value of X, the variance of u_i is the same for all observations. That is, the conditional variances of u_i are identical. Symbolically, we have

$$\begin{aligned} \textbf{var}\ (u_i \mid X_i) &= E[u_i - E(u_i \mid X_i)]^2 \\ &= E\big(u_i^2 \mid X_i\big) \text{ because of Assumption 3} \\ &= \sigma^2 \end{aligned} \tag{3.2.2}$$

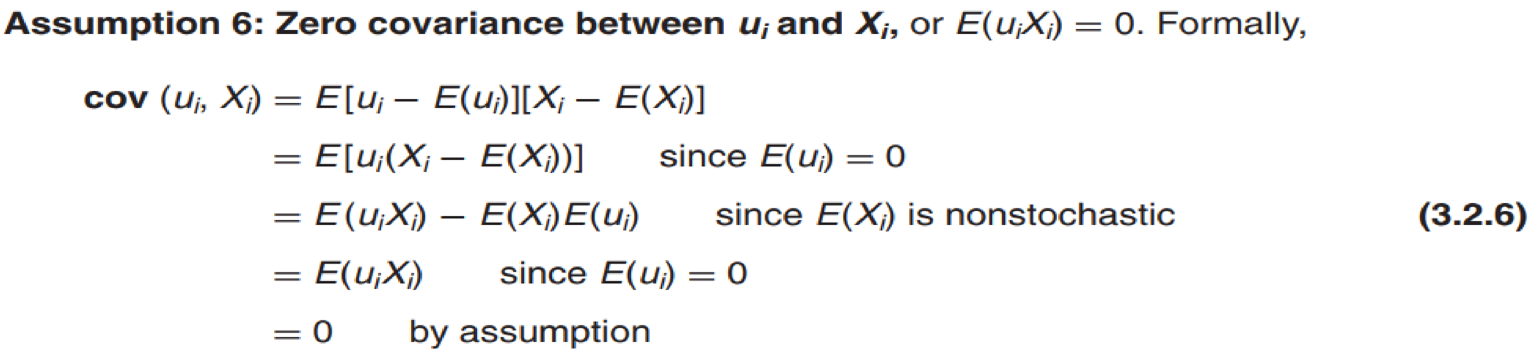
where **var** stands for variance.



Assumption 5: No autocorrelation between the disturbances. Given any two X values, X_i and X_j $(i \neq j)$, the correlation between any two u_i and u_j $(i \neq j)$ is zero. Symbolically,

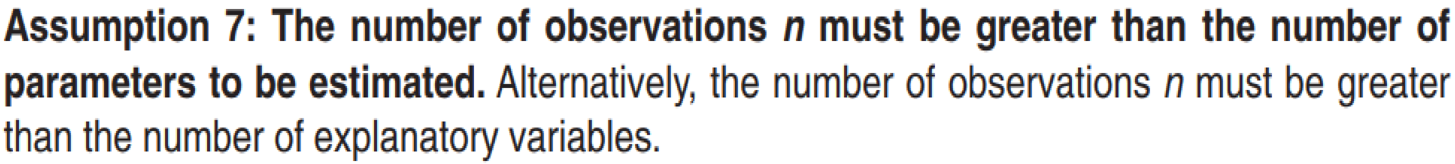
$$\begin{aligned} \textbf{cov}\ (u_i, u_j \mid X_i, X_j) &= E\{[u_i - E(u_i)] \mid X_i\}\{[u_j - E(u_j)] \mid X_j\} \\ &= E(u_i \mid X_i)(u_j \mid X_j) \qquad \text{(why?)} \\ &= 0 \end{aligned} \tag{3.2.5}$$

where i and j are two different observations and where **cov** means **covariance.**

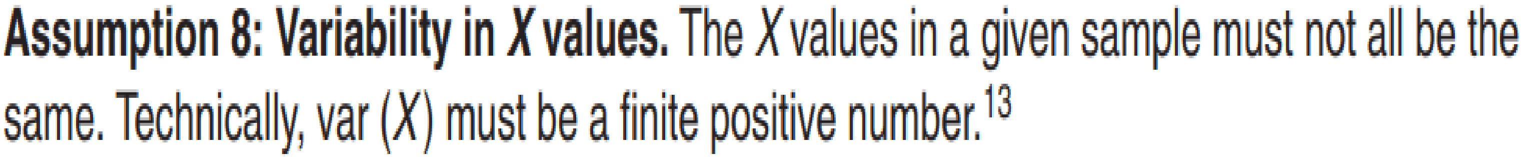


Assumption 6: Zero covariance between u_i and X_i, or $E(u_iX_i) = 0$. Formally,

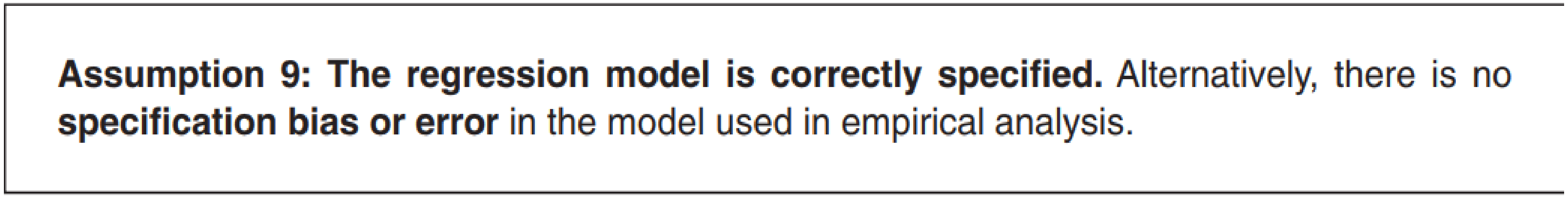
$$\begin{aligned}
\textbf{cov}\,(u_i, X_i) &= E[u_i - E(u_i)][X_i - E(X_i)] \\
&= E[u_i(X_i - E(X_i))] \qquad \text{since } E(u_i) = 0 \\
&= E(u_iX_i) - E(X_i)E(u_i) \qquad \text{since } E(X_i) \text{ is nonstochastic} \qquad \textbf{(3.2.6)} \\
&= E(u_iX_i) \qquad \text{since } E(u_i) = 0 \\
&= 0 \qquad \text{by assumption}
\end{aligned}$$



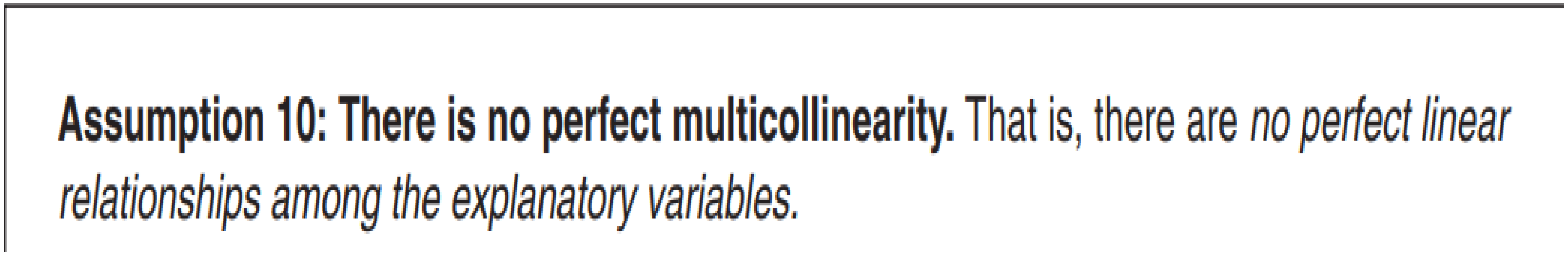
Assumption 7: The number of observations n must be greater than the number of parameters to be estimated. Alternatively, the number of observations n must be greater than the number of explanatory variables.



Assumption 8: Variability in X values. The X values in a given sample must not all be the same. Technically, var (X) must be a finite positive number.[13]



Assumption 9: The regression model is correctly specified. Alternatively, there is no **specification bias or error** in the model used in empirical analysis.



Assumption 10: There is no perfect multicollinearity. That is, there are *no perfect linear relationships among the explanatory variables.*

K. PRECISION OR STANDARD ERRORS OF LEAST-SQUARES ESTIMATES

The standard errors of the OLS estimates can be obtained as follows:

$$\operatorname{var}(\hat{\beta}_2) = \frac{\sigma^2}{\sum x_i^2}$$

$$\operatorname{se}(\hat{\beta}_2) = \frac{\sigma}{\sqrt{\sum x_i^2}}$$

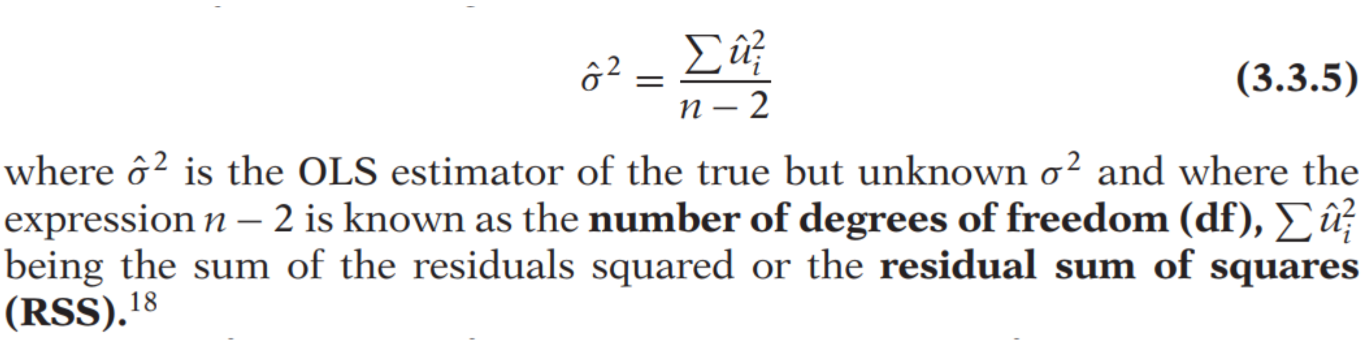
$$\operatorname{var}(\hat{\beta}_1) = \frac{\sum X_i^2}{n \sum x_i^2}\sigma^2$$

$$\operatorname{se}(\hat{\beta}_1) = \sqrt{\frac{\sum X_i^2}{n \sum x_i^2}}\sigma$$

where var = variance and se = standard error and where σ^2 is the constant or homoscedastic variance of u_i of Assumption 4 .

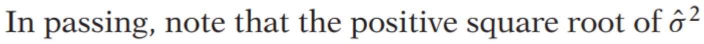
$$\begin{aligned}\operatorname{cov}(\hat{\beta}_1, \hat{\beta}_2) &= -\bar{X}\operatorname{var}(\hat{\beta}_2) \\ &= -\bar{X}\left(\frac{\sigma^2}{\sum x_i^2}\right)\end{aligned}$$

σ^2 is unknown. σ^2 itself is estimated by the following formula

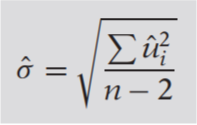


$$\hat{\sigma}^2 = \frac{\sum \hat{u}_i^2}{n-2} \qquad \textbf{(3.3.5)}$$

where $\hat{\sigma}^2$ is the OLS estimator of the true but unknown σ^2 and where the expression $n-2$ is known as the **number of degrees of freedom (df),** $\sum \hat{u}_i^2$ being the sum of the residuals squared or the **residual sum of squares (RSS).**[18]



In passing, note that the positive square root of $\hat{\sigma}^2$

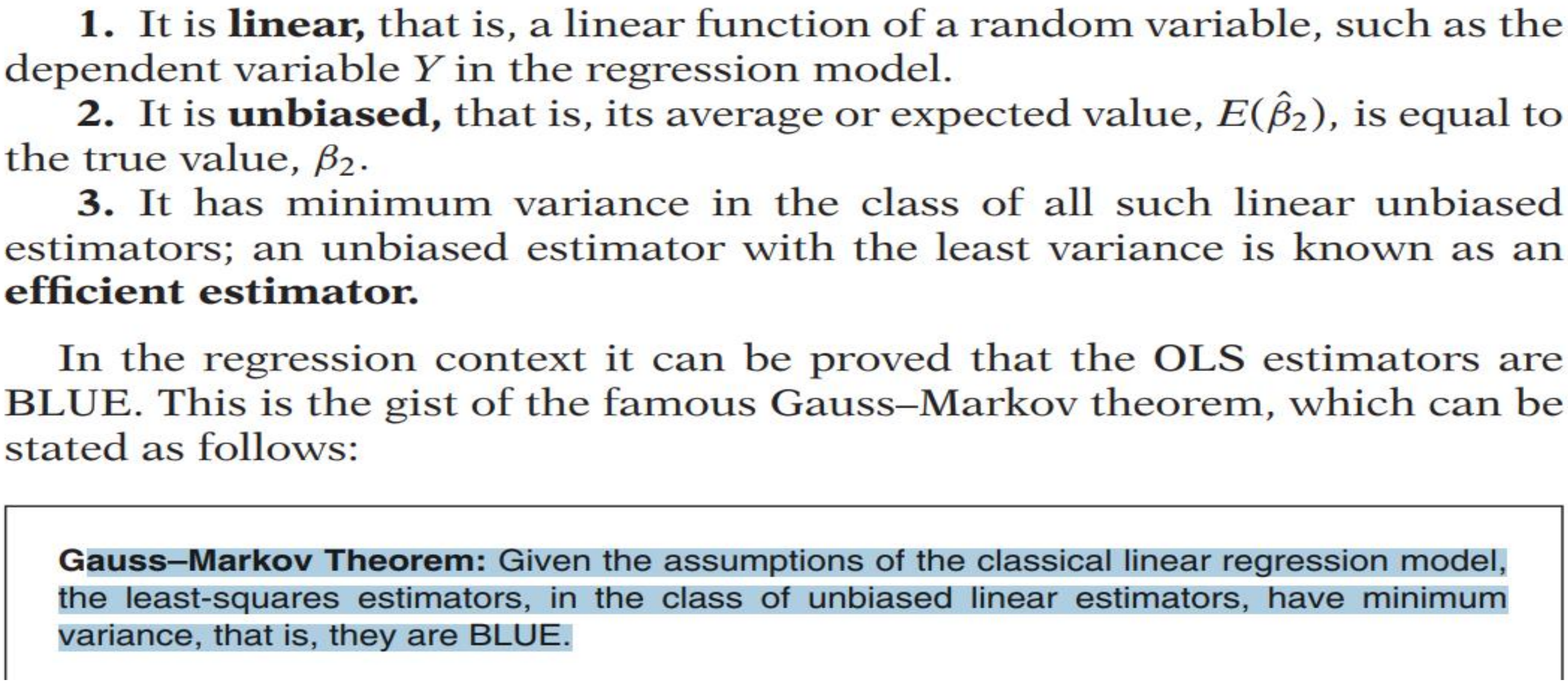


$$\hat{\sigma} = \sqrt{\frac{\sum \hat{u}_i^2}{n-2}}$$

is known as the **standard error of estimate** or **the standard error of the regression (se).** It is simply the standard deviation of the *Y* values about the estimated regression line

L. PROPERTIES OF LEAST-SQUARES ESTIMATORS: THE GAUSS–MARKOV THEOREM

The least-squares estimates possess some ideal or optimum properties. These properties are contained in the well-known **Gauss–Markov theorem.** An estimator, say the OLS estimator $\beta\hat{}_2$, is said to be a best linear unbiased estimator (BLUE) of β_2 if the following hold:



1. It is **linear,** that is, a linear function of a random variable, such as the dependent variable Y in the regression model.

2. It is **unbiased,** that is, its average or expected value, $E(\hat{\beta}_2)$, is equal to the true value, β_2.

3. It has minimum variance in the class of all such linear unbiased estimators; an unbiased estimator with the least variance is known as an **efficient estimator.**

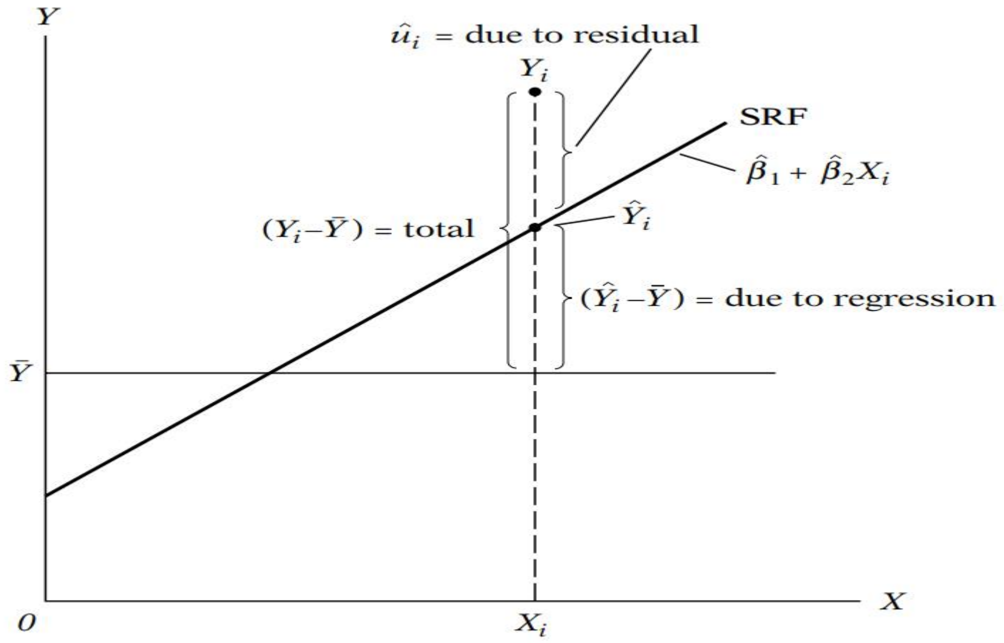
In the regression context it can be proved that the OLS estimators are BLUE. This is the gist of the famous Gauss–Markov theorem, which can be stated as follows:

Gauss–Markov Theorem: Given the assumptions of the classical linear regression model, the least-squares estimators, in the class of unbiased linear estimators, have minimum variance, that is, they are BLUE.

M. THE COEFFICIENT OF DETERMINATION *r* ^{2}: MEASURE OF “GOODNESS OF FIT”

Goodness of fit of the fitted regression line to a set of data shows how “well” the sample regression line fits the data.

Total variation in the observed *Y* values about their mean value can be partitioned into two parts, one attributable to the regression line and the other to random forces because not all actual *Observations* lie on the fitted line.



Y
$\hat{u}_i$ = due to residual
Y_i
SRF
$\hat{\beta}_1 + \hat{\beta}_2 X_i$
$(Y_i - \bar{Y})$ = total
$\hat{Y}_i$
$(\hat{Y}_i - \bar{Y})$ = due to regression
$\bar{Y}$
0
X_i
X

Breakdown of the variation of Y_i into two components.

Thus, **TSS = ESS + RSS**

Where ,

TSS = Total sum of squares = $\sum y_i^2 = \sum(Y_i - \bar{Y})^2$

ESS = Explained sum of squares = $\sum \hat{y}_i^2 = \sum(\hat{Y}_i - \bar{\hat{Y}})^2 = \sum(\hat{Y}_i - \bar{Y})^2 = \hat{\beta}_2^2 \sum x_i^2$

RSS = Residual sum of squares $\sum \hat{u}_i^2 = \sum(Y_i - \hat{Y}_i)^2$

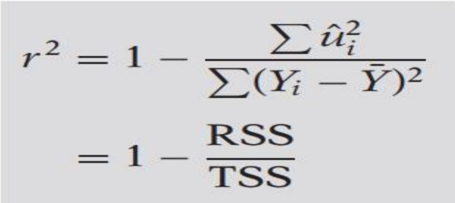
Dividing by TSS on both sides, we obtain

$$\begin{aligned} 1 &= \frac{\text{ESS}}{\text{TSS}} + \frac{\text{RSS}}{\text{TSS}} \\ &= \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2} + \frac{\sum \hat{u}_i^2}{\sum(Y_i - \bar{Y})^2} \end{aligned}$$

We now define r^2 as

$$r^2 = \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2} = \frac{\text{ESS}}{\text{TSS}}$$

or, alternatively, as



$$r^2 = 1 - \frac{\sum \hat{u}_i^2}{\sum (Y_i - \bar{Y})^2}$$

$$= 1 - \frac{\text{RSS}}{\text{TSS}}$$

The quantity *r^2* thus defined is known as the (sample) coefficient of determination and is the most commonly used measure of the goodness of fit of a regression line.

Verbally, r^2 *measures the proportion or percentage of the total variation in Y explained by the regression model.*

Two properties of r^2 are

1. t is a nonnegative quantity
2. Its limits are $0 \leq r^2 \leq 1$

Source:

D N Gujarati: Basic Econometrics

Entire Note here is based on the above source.

Samir K Mahajan