UNIT-I: SIMPLE LINEAR REGRESSION

LECTURE-1

To understand the concept of regression analysis, it is necessary to understand the concept of correlation first. So, we will revise the concept of correlation.

Correlation: The correlation is a statistical technique that measures the strength of the linear relationship between two variables. In simple words, it is the association between variables.

Pearson's Correlation Coefficient formula:

$$r = \frac{\mathrm{Cov}(X,Y)}{\sqrt{V(X)} * \sqrt{V(Y)}}$$

where, r: Sample correlation coefficient between X and Y
Cov (X, Y): Covariance between X and Y
V(X) and V(Y): Variances of random variables X and Y

Nature of Correlation:

(1) Positive Correlation:
The two variables are said to be positively correlated, when both the variables are moving in same direction; i.e., if one variable increases then other also increases and if one variable decreases then other also decreases.
Examples: (a) Attendance and Marks of student
(b) Exercise time and Calories burned
(c) Demand and Price of a product

(2) Negative Correlation:
The two variables are said to be negatively correlated, when both the variables are moving in opposite direction; i.e., if one variable increases then other variable decreases and vice-versa.
Examples: (a) Speed and Distance
(b) Temperature of place and Demand of the woollen clothes.

Interpretation of r values:

Strength of Association	Coefficient of correlation r (nature)	
	Positive	Negative
No correlation	0	0
Perfect	1	–1
Strong	$0.7 \le r < 1$	$-1 < r \le -0.7$
Moderate	$0.3 \le r < 0.7$	$-0.7 < r \le -0.3$
Weak	$0 < r < 0.3$	$-0.3 < r < 0$

Properties of Correlation Coefficient:
(a) The correlation coefficient r lies between –1 and 1.
(b) The correlation coefficient is not affected by shift of origin and change in scale.
(c) The correlation coefficient is a pure number. It is independent of unit of measurement of the variables.
(d) If x and y are independent then r=0 but converse is not true.

Regression Analysis:

When the correlation coefficient is given between the two variables, we can get information about
1. Whether the two random variables are associated with each other or not.
2. If they are associated then what is the nature of the association between them.
3. Up to what extent they are associated with each other.
But we can not use the correlation coefficient when we want to predict or estimate the values of one variable given the value of other variables.

When we know that two variables are correlated, we try to establish a mathematical equation between them. The mathematical equation is called regression equation. Using the regression equation, we can predict or estimate the values of the unknown variable when values of other variable are known.

Dependent variable:
The variable whose values is to be predicted or estimated from the known values of other variable(s) is called the dependent variable.

Independent variable:
The variable which is used for prediction or estimation of the dependent variable is called independent variable.
Sone common names which can be used alternatively for dependent variable and independent variable:

Dependent variable	Independent variable
Output variable	Input variable
Response variable	Explanatory variable
Regressand	Regressor
Predictand	Predictor
Study variable	

Consider a set of variables:

1) Yield of the crop, Quantity of fertilizer, Quantity of seed, Temperature, Irrigation level, Rain, Soil quality
Here,
Dependent variable: Yield of the crop
Independent variables: Quantity of fertilizer, Quantity of seed, Temperature, Irrigation level, Rain, Soil quality

2) Score in the final exam, Number of lectures attended, Time spent on social media, Duration of the study
Dependent variable: Score in the final exam
Independent variables: Number of lectures attended, Time spent on social media, Duration of the study

Now, we will define Regression analysis:
Regression analysis is the statistical study of estimating or predicting the value of dependent variable from the given value of independent variable.
In regression analysis, the dependent variable is a statistic(random) variable and independent variable can be fixed or random.
We will study only non-random(fixed) independent variable.

Linearity:

A model or an equation is said to be linear if it is linear in its parameters.

Consider the following models for linearity,
$y = \beta_0 + \beta_1 x$
$y = \beta_0 + \beta_1 x^2$
$y = \beta_0 + \beta_1^2 x$

How to check the linearity?
To check the linearity of the model, obtain the partial derivatives of the dependent variable with respect to all the parameters individually. If all the individual partial derivatives are independent of parameters, then the model is said to be linear.
Now we will check the linearity of the above models:
1) $y = \beta_0 + \beta_1 x$
$\frac{\delta y}{\delta \beta_0} = 1$ and $\frac{\delta y}{\delta \beta_1} = x$
Since both the partial derivatives are independent of the parameters, the given model is linear.

2) $y = \beta_0 + \beta_1 x^2$
$\frac{\delta y}{\delta \beta_0} = 1$ and $\frac{\delta y}{\delta \beta_1} = 2x$
Since both the partial derivatives are independent of the parameters, the given model is linear.

3) $y = \beta_0 + \beta_1^2 x$
$\frac{\delta y}{\delta \beta_0} = 1$ and $\frac{\delta y}{\delta \beta_1} = 2\beta_1 x$
Since the partial derivative $\frac{\delta y}{\delta \beta_1}$ depends on parameter β_1, this model is not linear.

REGRESSION ANALYSIS

Suppose the outcome of any process is denoted by a random variable *y*, called as dependent (or study) variable, depends on *k* independent (or explanatory) variables denoted by $X_1, X_2, ..., X_k$. Suppose the behaviour of *y* can be explained by a relationship given by

$y = f(X_1, X_2, ..., X_k, \beta_1, \beta_2, \ldots, \beta_k) + \varepsilon$

where *f* is some well-defined function and $\beta_1, \beta_2, \ldots, \beta_k$ are the parameters which characterize the role and contribution of $X_1, X_2, ..., X_k$, respectively. The term ε reflects the stochastic nature of the relationship between y and $X_1, X_2, ..., X_k$ indicates that such a relationship is not exact in nature. When $\varepsilon = 0$, then the relationship is called the mathematical model otherwise the statistical model. The term "**model**" is broadly used to represent any phenomenon in a mathematical framework.

A model or relationship is termed as linear if it is linear in parameters and nonlinear if it is not linear in parameters. In other words, if all the partial derivatives of y with respect to each of the parameters $\beta_1, \beta_2, \ldots, \beta_k$, are independent of the parameters, then the model is called a **linear model**. If any of the partial derivatives of y with respect to any of the $\beta_1, \beta_2, \ldots, \beta_k$ is not independent of the parameters, then the model is called as nonlinear. Note that the linearity or non-linearity of the model is not described by the linearity or nonlinearity of explanatory variables in the model.

For example

$y = \beta_1 X_1^2 + \beta_2\sqrt{X_2} + \beta_3 \log X_3 + \varepsilon$

is a linear model because $\delta y/\delta\beta_i$,(i =1,2,3) are independent of the parameters β_i ,(*i* =1,2,3). On the other hand,

$y = \beta_1^2 X_1 + \beta_2 X_2 + \beta_3 \log X_3 + \varepsilon$

is a nonlinear model because $\delta y/\delta\beta_1 = 2\beta_1 X_1$ depends on β_1 although $\delta y/\delta\beta_2$ and $\delta y/\delta\beta_3$ are independent of any of the β_1, β_2 or β_3.

When the function *f* is linear in parameters, then $y = f(X_1, X_2, ..., X_k, \beta_1, \beta_2, \ldots, \beta_k) + \varepsilon$

is called a linear model and when the function *f* is nonlinear in parameters, then it is called a nonlinear model. In general, the function *f* is chosen as

$f(X_1, X_2, ..., X_k, \beta_1, \beta_2, \ldots, \beta_k) = \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$

to describe a linear model. Since $X_1, X_2, ..., X_k$ are pre-determined variables and y is the outcome, so both are known. Thus, the knowledge of the model depends on the knowledge of the parameters $\beta_1, \beta_2, \ldots, \beta_k$.

The linear statistical modelling essentially consists of developing approaches and tools to determine $\beta_1, \beta_2, \ldots, \beta_k$ in the linear model

$$y = \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon$$

given the observations on y and $X_1, X_2, \ldots, X_k$.

The regression analysis is a tool to determine the values of the parameters given the data on *y* and $X_1, X_2, \ldots, X_k$. The literal meaning of regression is "to move in the backward direction". Before discussing and understanding the meaning of "backward direction", let us find which of the following statement is correct:

S1: model generates data or
S2: data generates model.

Obviously, S1 is correct. It can be broadly thought that the model exists in nature but is unknown to the experimenter. When some values to the explanatory variables are provided, then the values for the output or study variable are generated accordingly, depending on the form of the function *f* and the nature of the phenomenon. So ideally, the pre-existing model gives rise to the data. Our objective is to determine the functional form of this model. Now we move in the backward direction. We propose to first collect the data on study and explanatory variables. Then we employ some statistical techniques and use this data to know the form of function *f*. Equivalently, the data from the model is recorded first and then used to determine the parameters of the model. The regression analysis is a technique which helps in determining the statistical model by using the data on study and explanatory variables. The classification of linear and nonlinear regression analysis is based on the determination of linear and nonlinear models, respectively.

Consider a simple example to understand the meaning of "regression". Suppose the yield of the crop (y) depends linearly on two explanatory variables, viz., the quality of fertilizer (X_1) and level of irrigation (X_2) as

$$y = \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

There exist the true values of β_1 and β_2 in nature but are unknown to the experimenter. Some values on y

are recorded by providing different values to X_1 and X_2. There exists some relationship between y and X_1, X_2 which gives rise to a systematically behaved data on y, X_1 and X_2. Such a relationship is unknown to the experimenter. To determine the model, we move in the backward direction in the sense that the collected data is used to determine the parameters β_1 and β_2 of the model. In this sense, such an approach is termed as regression analysis.

STEPS IN REGRESSION ANALYSIS:

Regression analysis includes the following steps:

- Statement of the problem under consideration
- Choice of relevant variables
- Collection of data on relevant variables
- Specification of model
- Choice of method for fitting the data
- Fitting of model
- Model validation and criticism
- Using the chosen model(s) for the solution of the posed problem.

These steps are examined below.

1. Statement of the problem under consideration:

The first important step in conducting any regression analysis is to specify the problem and the objectives to be addressed by the regression analysis. The wrong formulation or the wrong understanding of the problem will give the erroneous statistical inferences. The choice of variables depends upon the objectives of the study and understanding of the problem. For example, the height and weight of children are related. Now there can be two issues to be addressed.

(i) Determination of height for a given weight, or

(ii) Determination of weight for a given height.

In case 1, the height is a response variable, whereas weight is a response variable in case 2. The role of explanatory variables is also interchanged in cases 1 and 2.

2. Choice of relevant variables:

Once the problem is carefully formulated and objectives have been decided, the next question is to choose the relevant variables. It has to be kept in mind that the correct choice of variables will determine the statistical inferences correctly. For example, in an agricultural experiment, the yield depends on explanatory

variables like quantity of fertilizer, rainfall, irrigation, temperature etc. These variables are denoted by X_1, X_2, ..., X_k as a set of k explanatory variables.

3. Collection of data on relevant variables:

Once the objective of the study is clearly stated, and the variables are chosen, the next question arises is to collect data on such relevant variables. The data is essentially the measurement on these variables. For example, it is important to know how to record the data on age. For example, the data collection is on age in total complete years or date of birth is to be recorded, which can give the exact age of a specific time. Moreover, it is also important to decide that the data has to be collected on variables as quantitative variables or qualitative variables. For example, if the ages (in years) are 15,17,19,21,23, then these are quantitative values. If the ages are defined by a variable that takes value 1 if ages are less than 18 years and 0 if the ages are more than 18 years, then the earlier recorded data is converted to 1,1,0,0,0. Note that there is a loss of information in converting the quantitative data into qualitative data. The methods and approaches for qualitative and quantitative data are also different. If the study variable is binary, then **logistic regression** is used. If all explanatory variables are qualitative, then **analysis of variance** technique is used. If some explanatory variables are qualitative and others are quantitative, then **analysis of covariance** technique is used. The techniques of analysis of variance and analysis of covariance are the special cases of regression analysis.

Generally, the data is collected on n subjects, then y on data, then y denotes the response or study variable and y_1, y_2, ..., y_n are the n values. If there are k explanatory variables, X_1, X_2, ..., X_k then X_{ij} denotes the ith value of jth variable. The observation can be presented in the following table:

Notation for the data used in regression analysis

Observation	Response y	Explanatory Variables Number			
		X_1	X_2	…	X_k
1	y_1	x_{11}	x_{12}	…	x_{1k}
2	y_2	x_{21}	x_{22}	…	x_{2k}

3	y_3	x_{31}	x_{32}		x_{3k}
n	y_n	x_{n1}	x_{n2}	…	x_{nk}

4. Specification of the model:

The experimenter or the person working in the subject usually help in determining the form of the model. Only the form of the tentative model can be ascertained, and it will depend on some unknown parameters. For example, a general form will be like

$y = f(X_1, X_2, ..., X_k, \beta_1, \beta_2, \ldots, \beta_k) + \varepsilon$

where ε is the random error reflecting mainly the difference in the observed value of *y* and the value of *y* obtained through the model. The form of $f(X_1, X_2, ..., X_k, \beta_1, \beta_2, \ldots, \beta_k)$ can be linear as well as nonlinear depending on the form of parameters $\beta_1, \beta_2, \ldots, \beta_k$. A model is said to be linear if it is linear in parameters. For example,

$y = \beta_1 X_1^2 + \beta_2\sqrt{X_2} + \beta_3 \log X_3 + \varepsilon$ is a linear model

On the other hand,

$y = \beta_1^2 X_1 + \beta_2 X_2 + \beta_3 \log X_3 + \varepsilon$ is a nonlinear model

Many times, the nonlinear models can be converted into linear models through some transformations. So, the class of linear models is wider than what it appears initially.

If a model contains only one explanatory variable, then it is called a **simple regression model.** When there are more than one independent variables, then it is called a **multiple regression model.** When there is only one study variable, the regression is termed as **univariate regression.** When there are more than one study variables, the regression is termed as **multivariate regression.** Note that the simple and multiple regressions are not same as univariate and multivariate regressions. The choice between simple and multiple regression is determined by the number of explanatory variables, whereas the choice between univariate and multivariate regressions is determined by the number of study variables.

5. Choice of method for fitting the data:

After the model has been defined, and the data have been collected, the next task is to estimate the parameters of the model based on the collected data. This is also referred to as **parameter estimation** or

model fitting. The most commonly used method of estimation is called the least-squares method. Under certain assumptions, the least-squares method produces estimators with desirable properties. The other estimation methods are the maximum likelihood method, ridge method, principal components method etc.

6. Fitting of the model:

The estimation of unknown parameters using appropriate method provides the values of the parameter. Substituting these values in the equation gives us a usable model. This is termed as model fitting. The estimates of parameters $\beta_1, \beta_2, \ldots, \beta_k$ in the model $y = f(X_1, X_2, \ldots, X_k, \beta_1, \beta_2, \ldots, \beta_k) + \varepsilon$

are denotes as $\widehat{\beta_1}, \widehat{\beta_2}, \ldots, \widehat{\beta_k}$ which gives the fitted model as

$y = f(X_1, X_2, \ldots, X_k, \widehat{\beta_1}, \widehat{\beta_2}, \ldots, \widehat{\beta_k}) + \varepsilon$

When the value of *y* is obtained for the given values of $X_1, X_2, \ldots, X_k$, it is denoted as $\hat{y}$ and called as fitted value.

The fitted equation is used for prediction. In this case, $\hat{y}$ is termed as the **predicted value.**

There are different methodologies based on regression analysis. They are described in the following table:

Type of Regression	**Conditions**
Univariate	Only one quantitative response variable
Multivariate	Two or more quantitative response variables
Simple	Only one explanatory variable
Multiple	Two or more explanatory variables
Linear	All parameters enter the equation linearly, possibly after transformation of the data
Nonlinear	The relationship between the response and some of the explanatory variables is nonlinear or some of the parameters appear nonlinearly, but no transformation is possible to make the parameters appear linearly
Analysis of variance	All explanatory variables are qualitative variables
Analysis of Covariance	Some explanatory variables are quantitative variables and others are qualitative variables

Logistic	The response variable is qualitative

7. Model criticism and selection

The validity of the statistical method to be used for regression analysis depends on various assumptions. These assumptions become essentially the assumptions for the model and the data. The quality of statistical inferences heavily depends on whether these assumptions are satisfied or not. For making these assumptions to be valid and to be satisfied, care is needed from the beginning of the experiment. One has to be careful in choosing the required assumptions and to decide as well to determine if the assumptions are valid for the given experimental conditions or not? It is also important to determine that the situations in which the assumptions may not meet.

The validation of the assumptions must be made before drawing any statistical conclusion. Any departure from the validity of assumptions will be reflected in the statistical inferences. In fact, the regression analysis is an iterative process where the outputs are used to diagnose, validate, criticize and modify the inputs.

Objectives of regression analysis

The determination of the explicit form of the regression equation is the ultimate objective of regression analysis. It is finally a good and valid relationship between study variable and explanatory variables. Such a regression equation can be used for several purposes. For example, to determine the role of any explanatory variable in the joint relationship in any policy formulation, to forecast the values of the response variable for a given set of values of explanatory variables. The regression equation helps in understanding the interrelationships of variables among them.

SIMPLE LINEAR REGRESSION MODEL

We consider the modelling between the dependent and one independent variable. When there is only one independent variable in the linear regression model, the model is generally termed as a simple linear regression model. When there are more than one independent variables in the model, then the linear model is termed as the multiple linear regression model.

Model:

$y = \beta_0 + \beta_1 x + \varepsilon$

Where, y: Dependent variable (Study/Response/Output variable)
x: Independent variable (Explanatory/Regressor/Input variable)
β_0: Intercept of a regression line
β_1: Slope of a regression line
ε: Error term

Note:

-β_0 & β_1 are parameters of the model. These parameters are usually called as regression coefficients.

-The unobservable error component ε accounts for the failure of data to lie on the straight line and represents the difference between the true and observed realization of y. There can be several reasons for such difference, e.g., the effect of all deleted variables in the model, variables may be qualitative, inherent randomness in the observations etc. We assume that ε is observed as independent and identically distributed random variable with mean zero and constant variance σ^2. Later, we will additionally assume that ε is normally distributed.

Suppose we have n pairs of observations of random variables X and Y,
i.e., (x_i, y_i) , i= 1,2,…,n
For the individual observations, we can write model as
$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, i=1,2, …, n

- β_0: Intercept of a regression line
  β_0 refers to the y-intercept, which means, it is that point where the line would cross over y-axis when x-axis is zero.
- β_1: Slope of a regression line
  For every unit increase in X, the slope tells us how much Y would increase by.
  The slope of regression line tells us the change in dependent variable Y per unit increase in independent variable X.

Examples:

1) Medical researchers often use linear regression to understand the relationship between drug dosage and blood pressure of patients.
Researchers might administer various dosages of a certain drug to patients and observe how their blood pressure responds. They might fit simple linear regression model using dosages as the predictor variable and blood pressure as response variable. The regression model would take the following form:
Blood pressure = $\beta_0 + \beta_1$(Dosage)

The coefficient β_0 would represent the expected blood pressure when dosage is zero.
The coefficient β_1 would represent the average change in blood pressure when dosage is increased by one unit.
If β_1 is negative, it would mean that an increase in dosage is associated with a decrease in blood pressure.
If β_1 is positive, it would mean that an increase in dosage is associated with an increase in blood pressure.
If β_1 is close to zero, it would mean that an increase in dosage is associated with no change in blood pressure.
Depending on the value of β_1, researchers may decide to change the dosage given to a patient.

2) Businesses often use linear regression to understand the relationship between advertising spending and revenue.
For example, they might fit a linear regression model using advertising spending as the predictor variable and revenue as the response variable. The regression model would take the following form:
Revenue = $\beta_0 + \beta_1$(Advertising spending)

The coefficient β_0 would represent the expected revenue when advertising spending is zero.
The coefficient β_1 would represent the average change in revenue when advertising spending is increased by one unit (e.g., one dollar).
If β_1 is negative, it would mean that more advertising spending is associated with less revenue.
If β_1 is positive, it would mean that more advertising spending is associated with more revenue.
If β_1 is close to zero, it would mean that advertising spending has little effect on revenue.
Depending on the value of β_1, a company may decide to either decrease or increase their advertising spending.

Deterministic model:

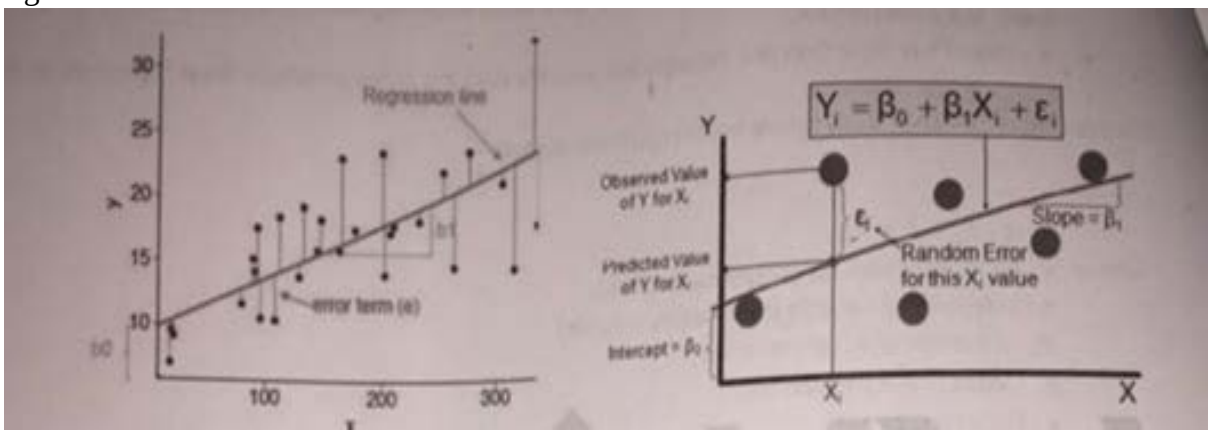
Suppose there is some relationship between dependent variable(Y) and independent variable(X).
Suppose this relationship can be written using equation of a straight line.
$y = \beta_0 + \beta_1 x$
For individual observations, the model can be written as
$y_i = \beta_0 + \beta_1 x_i$, i=1,2,…,n

For example, suppose $\beta_0 = 1$ and $\beta_1 = 2$
$y_i = 1 + 2x_i$
If x_i=1 then using above model, $\hat{y}_i = 3$.
Using the above model, we get exact value of y for the given value of x.
Therefore, this model is called as Deterministic model.
In Deterministic model, the relationship between X and Y is exact. But in real world, this never happens.
Generally, the relationship between X and Y is not exact due to some random error ε say.
If we add this error/disturbance term in the model then model becomes stochastic/probabilistic/statistical model.

$y = \beta_0 + \beta_1 x + \varepsilon$

Residuals:

Suppose (x_i, y_i) are the n pairs of observations on the response and explanatory variables. When we plot these observations then the plot is called as scatter diagram shown in the following figure.



In linear regression analysis, we try to find out a straight-line equation such that line should be as close as possible from all the points.
The points on the regression line corresponding to true value are called the estimated or predicted or fitted values and denoted by $\hat{y}_i$.
The vertical distance from a fitted value $(\hat{y}_i)$ to corresponding true value (y_i) is called error in estimation and is denoted by ε_i.
Therefore, $\varepsilon_i = y_i - \hat{y}_i$, i = 1,2, …, n
The errors are more commonly known as Residuals.

Assumptions of the classical linear regression model (CLRM):

1. The linear regression model is linear in parameters.
2. The values of the regressors i.e., X's are fixed in repeated sampling.
3. The residuals are independently identically normally distributed with mean 0 and constant variance σ^2.
   i.e., $\varepsilon_i \sim$ iid N $(0, \sigma^2)$, i=1,2, …, n
4. The variance of residuals (errors) is the same for any given value of X. This property is known as Homoscedasticity.
   Note:
   Homoscedasticity:
   Homo: Same
   Scedasticity: scattered or variability
   In simple words, Homoscedasticity means equal variance.
5. There must be sufficient variability in the values taken by the regressor's.
6. The number of observations(n) must be greater than number of parameters(k) to be estimated. i.e., n>k
7. The explanatory variables and the residuals are uncorrelated.

8. The regression model is correctly specified.

9. There is no multicollinearity in the data.

Note:

Multicollinearity:

Multicollinearity means the independent variables in the linear regression model are correlated. In such a situation, one independent variable can be used to predict another.

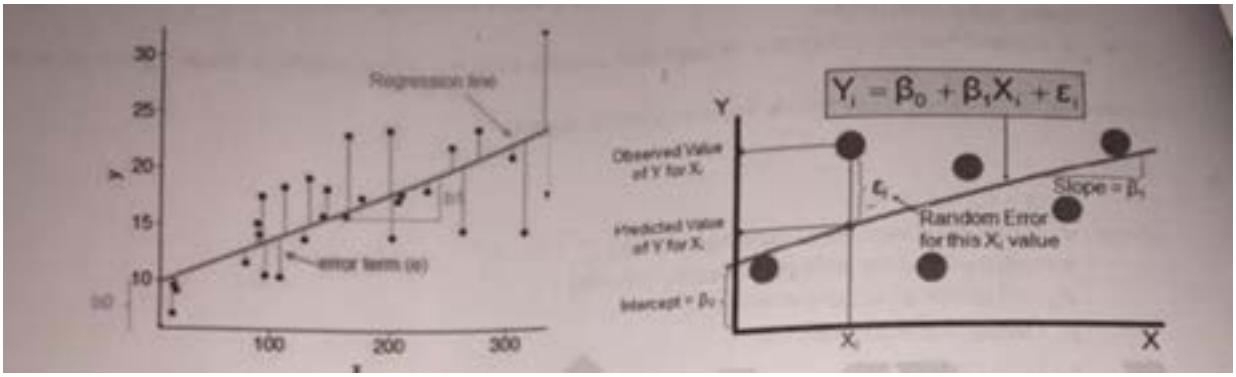
10. There is no autocorrelations in the residuals for the given X's.

Estimation of Parameters:

To know the model exactly, we need to estimate the unknown parameters β_0 and β_1 such that error should be as small as possible. There are various methods to estimate the unknown parameters. We will use least square method.

Ordinary least square method:

The method of least square is used to estimate β_0 and β_1.
In this method, we try to minimize distance between residuals.



Our sample regression model is

$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, i=1,2,…,n

Taking expectation on both the sides,

$\hat{y}_i = \widehat{\beta_0} + \widehat{\beta_1} x_i$, i=1, 2, …, n

Error = $\varepsilon_i = y_i - \hat{y}_i = y_i - \widehat{\beta_0} - \widehat{\beta_1} x_i$

In the least square method, we estimate β_0 and β_1 such that the sum of squares due to the errors are minimized.

Let, $S = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \widehat{\beta_0} - \widehat{\beta_1} x_i)^2$

To obtain the $\widehat{\beta_0}$ and $\widehat{\beta_1}$ of β_0 and β_1 respectively, we differentiate S partially with respected to β_0 and β_1 separately and equate them with zero.

Consider,

$$\frac{\partial s}{\partial \beta_0} = 0 \Rightarrow \frac{\partial}{\partial \beta_0} \sum_{i=1}^{n} [y_i - \beta_0 - \beta_1 x_i]^2 = 0$$

$$\Rightarrow 2 \sum_{i=1}^{n} [y_i - \beta_0 - \beta_1 x_i] \frac{\partial}{\partial \beta_0} [y_i - \beta_0 - \beta_1 x_i] = 0$$

$$\Rightarrow 2 \sum_{i=1}^{n} [y_i - \beta_0 - \beta_1 x_i] (-1) = 0$$

$$\Rightarrow \sum_{i=1}^{n} [y_i - \beta_0 - \beta_1 x_i] = 0$$

$$\Rightarrow \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \hat{\beta}_0 - \hat{\beta}_1 \sum x_i = 0$$

$$\Rightarrow \sum y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum x_i = 0$$

$$\Rightarrow \sum y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum x_i \quad \text{———} (1)$$

Dividing by n, we get

$$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x} \quad \text{———} (2)$$

Now,

$$\frac{\partial s}{\partial \beta_1} = 0 \Rightarrow \frac{\partial}{\partial \beta_1} \sum_{i=1}^{n} [y_i - \beta_0 - \beta_1 x_i]^2 = 0$$

$$\Rightarrow 2 \sum_{i=1}^{n} [y_i - \beta_0 - \beta_1 x_i] \frac{\partial}{\partial \beta_1} [y_i - \beta_0 - \beta_1 x_i] = 0$$

$$\Rightarrow 2 \sum_{i=1}^{n} [y_i - \beta_0 - \beta_1 x_i] (-x_i) = 0$$

$$\Rightarrow \sum_{i=1}^{n} x_i y_i - \hat{\beta}_0 \sum_{i=1}^{n} x_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = 0$$

$$\Rightarrow \sum x_i y_i = \hat{\beta}_0 \sum x_i + \hat{\beta}_1 \sum x_i^2 \quad \text{——— (2)}$$

Equations (1) and (2) are called as Normal equations.

To solve equations (1) and (2) simultaneously, multiply eq^n (1) by $\sum x_i$ and eq^n (2) by n.

$$(\sum x_i)(\sum y_i) = n\hat{\beta}_0 (\sum x_i) + \hat{\beta}_1 (\sum x_i)^2$$

$$- \quad n \sum x_i y_i = n\hat{\beta}_0 \sum x_i + n\hat{\beta}_1 \sum x_i^2$$

$$(\sum x_i)(\sum y_i) - n\sum x_i y_i = \hat{\beta}_1 \left[(\sum x_i)^2 - n\sum x_i^2 \right]$$

Multiply both sides by (-1),

$$n\sum x_i y_i - (\sum x_i)(\sum y_i) = \hat{\beta}_1 \left[n\sum x_i^2 - (\sum x_i)^2 \right]$$

Divide both sides by n^2,

$$\frac{\sum x_i y_i}{n} - \frac{(\sum x_i)}{n}\frac{(\sum y_i)}{n} = \hat{\beta}_1 \left[\frac{\sum x_i^2}{n} - \frac{(\sum x_i)^2}{n^2} \right]$$

$$\Rightarrow \frac{\sum x_i y_i}{n} - \bar{x}\bar{y} = \hat{\beta}_1 \left[\frac{\sum x_i^2}{n} - \bar{x}^2 \right]$$

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

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

i.e. $\boxed{\hat{\beta}_1 = \frac{Cov(x,y)}{V(x)}}$

From eq^n(2),

$$\boxed{\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}}$$

Let, $S_{xx} = V(x)$
$S_{yy} = V(y)$
$S_{xy} = Cov(x, y)$

$$\therefore \boxed{\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}}$$

$\therefore$ Estimates of regression coefficients :-

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$

$$\hat{\beta}_0 = \bar{y} - \bar{x}\hat{\beta}_1.$$

★ To obtain $E(\hat{\beta}_0)$ and $E(\hat{\beta}_1)$:

Linear regression model is : $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

Fitted regression model is : $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

where, $\hat{\beta}_0 = \bar{y} - \bar{x}\hat{\beta}_1$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\sum_{i=1}^{n}(\beta_0 + \beta_1 x_i + \varepsilon_i)$$

$$= \beta_0 + \beta_1 \bar{x} + \bar{\varepsilon}$$

$$y_i - \bar{y} = \beta_0 + \beta_1 x_i + \varepsilon_i - \beta_0 - \beta_1 \bar{x} - \bar{\varepsilon}$$

$$E(y_i - \bar{y}) = \beta_1 (x_i - \bar{x}) + E(\varepsilon_i - \bar{\varepsilon})$$

{ Since y_i is a r.v. & x_i is not a r.v., it's a controlled variable.
Also, we assume that $\varepsilon_i \sim N(0, \sigma^2)$ }

$$\therefore E(y_i - \bar{y}) = \beta_1 (x_i - \bar{x})$$

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

$$= \frac{\sum (x_i - \bar{x}) \beta_1 (x_i - \bar{x})}{\sum (x_i - \bar{x})^2}$$

$$= \beta_1 \frac{\sum (x_i - \bar{x})^2}{\sum (x_i - \bar{x})^2}$$

$$\therefore E(\hat{\beta}_1) = \beta_1$$

$\therefore$ $\hat{\beta}_1$ is an unbiased estimator of β_1.

Now,

$$E(\hat{\beta}_0) = E(\bar{y} - \bar{x}\hat{\beta}_1)$$

$$= E(\bar{y} - \bar{x}\hat{\beta}_1)$$

$$= E(\beta_0 + \beta_1 \bar{x} + \bar{\varepsilon} - \bar{x}\hat{\beta}_1)$$

$$= E(\beta_0) + E(\beta_1 \bar{x}) - E(\bar{x}\hat{\beta}_1)$$

$$\therefore E(\hat{\beta}_0) = \beta_0 + \cancel{\beta_1 \bar{x}} - \cancel{\beta_1 \bar{x}} \quad \ldots \quad \because E(\hat{\beta}_1) = \beta_1$$

$$\therefore E(\hat{\beta}_0) = \beta_0$$

$\therefore$ $\hat{\beta}_0$ is an unbiased estimator of β_0.

★ Variance of estimated Regression Coefficients

We know that.

$\varepsilon_i \sim N(0, \sigma^2)$, $i = 1, 2, \ldots, n$

$\therefore \bar{\varepsilon} \sim N\left(0, \frac{\sigma^2}{n}\right)$

Model is $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

$\therefore$ We can write, $\bar{y} = \beta_0 + \beta_1 \bar{x} + \bar{\varepsilon}$

$\therefore E(\bar{y}) = \beta_0 + \beta_1 \bar{x} + E(\bar{\varepsilon}) = \beta_0 + \beta_1 \bar{x}$

&

$V(\bar{y}) = V(\bar{\varepsilon}) = \frac{\sigma^2}{n}$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{\sum (x_i - \bar{x}) y_i}{\sum (x_i - \bar{x})^2}$$

$\because \sum_{i=1}^{n} (x_i - \bar{x}) \bar{y}$
$= \bar{y} \sum x_i - n\bar{x}\bar{y}$
$= n\bar{x}\bar{y} - n\bar{x}\bar{y} = 0$

Let $c_i = \frac{(x_i - \bar{x})}{\sum (x_i - \bar{x})^2}$

$\therefore \hat{\beta}_1 = \sum c_i y_i$

$$V(\hat{\beta}_1) = \sum_{i=1}^{n} c_i^2 V(y_i) + \sum_{i} \sum_{i \neq j} c_i c_j \, Cov(y_i, y_j)$$

$\because$ y_i's are indeptly distributed $\Rightarrow Cov(y_i, y_j) = 0$

$$\therefore V(\hat{\beta}_1) = \sum_{i=1}^{n} c_i^2 V(y_i)$$

$$= \sum_{i=1}^{n} c_i^2 \sigma^2$$

$$= \sigma^2 \sum_{i=1}^{n} \left[\frac{(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} \right]^2$$

Save the Children

$$\therefore V(\hat{\beta}_1) = \sigma^2 \frac{\sum_{i=1}^{n} (x_i-\bar{x})^2}{\left[\sum (x_i-\bar{x})^2\right]^2}$$

$$= \sigma^2 \frac{\left[\sum_{i=1}^{n} (x_i-\bar{x})^2\right]^1}{\left[\sum_{i=1}^{n} (x_i-\bar{x})^2\right]^{2-1=1}}$$

$$= \frac{\sigma^2}{\sum_{i=1}^{n} (x_i-\bar{x})^2} = \frac{\sigma^2}{S_{xx}}$$

Now,

$$\hat{\beta}_0 = \bar{y} - \bar{x}\hat{\beta}_1$$

$$V(\hat{\beta}_0) = V(\bar{y}) + \bar{x}^2 V(\hat{\beta}_1) - 2\bar{x}\, Cov(\bar{y}, \hat{\beta}_1)$$

$$= \frac{\sigma^2}{n} + \bar{x}^2 \times \frac{\sigma^2}{S_{xx}} - 2\bar{x}\, Cov(\bar{y}, \hat{\beta}_1)$$

Consider,

$$Cov(\bar{y}, \hat{\beta}_1) = E\left[(\bar{y} - E(\bar{y}))(\hat{\beta}_1 - E(\hat{\beta}_1))\right]$$

$$\bar{y} - E(\bar{y}) = \beta_0 + \beta_1\bar{x} + \bar{\varepsilon} - \beta_0 - \beta_1\bar{x} = \bar{\varepsilon}$$

&

$$\hat{\beta}_1 - E(\hat{\beta}_1) = \hat{\beta}_1 - \beta_1 = \sum c_i y_i - \beta_1$$

$$\therefore Cov(\bar{y}, \hat{\beta}_1) = E\left[\bar{\varepsilon}(\sum c_i y_i - \beta_1)\right]$$

$$= 0 \qquad \because E(\bar{\varepsilon}) = 0$$

$$\therefore V(\hat{\beta}_0) = \frac{\sigma^2}{n} + \bar{x}^2 \frac{\sigma^2}{S_{xx}} = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)$$

Save the Children

★ Estimation of σ^2 :

Model is $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

Residual is the difference between the actual value and the predicted value.

$\therefore \varepsilon_i = y_i - \hat{y}_i$

$\therefore$ SSE = sum of square due to error

$$= \sum_{i=1}^{n} \varepsilon_i^2$$

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

$$= \sum_{i=1}^{n} [y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i]^2 \quad \because \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

$$= \Sigma [y_i - \bar{y} + \bar{x}\hat{\beta}_1 - \hat{\beta}_1 x_i]^2 \quad \because \hat{\beta}_0 = \bar{y} - \bar{x}\hat{\beta}_1$$

$$= \Sigma [(y_i - \bar{y}) - \hat{\beta}_1 (x_i - \bar{x})]^2$$

$$= \Sigma (y_i - \bar{y})^2 + \hat{\beta}_1^2 \Sigma (x_i - \bar{x})^2 - 2\hat{\beta}_1 \Sigma (x_i - \bar{x})(y_i - \bar{y})$$

$$= S_{yy} + \hat{\beta}_1^2 S_{xx} - 2\hat{\beta}_1 S_{xy}$$

$$= S_{yy} + \hat{\beta}_1^2 S_{xx} - 2 \frac{S_{xy}}{S_{xx}} \times S_{xy} \quad \left(\because \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}\right)$$

$$= S_{yy} + \hat{\beta}_1^2 S_{xx} - 2 \frac{S_{xy}}{S_{xx}} \times \frac{S_{xy}}{S_{xx}} \times S_{xx}$$

$$= S_{yy} + \hat{\beta}_1^2 S_{xx} - 2\hat{\beta}_1^2 S_{xx}$$

Save the Children

$$\therefore \quad SSE = S_{yy} - \hat{\beta}_1^2 S_{xx}$$

$$\because \quad \varepsilon_i \sim N(0, \sigma^2)$$

$$\therefore \quad \frac{\varepsilon_i}{\sigma} \sim N(0,1)$$

$$\left[\frac{\varepsilon_i}{\sigma}\right]^2 \sim \chi^2_{(1)}$$

$$\sum_{i=1}^{n} \left[\frac{\varepsilon_i}{\sigma}\right]^2 \sim \chi^2_{(n)}$$

$$\therefore \quad \frac{\sum \varepsilon_i^2}{\sigma^2} \sim \chi^2_{(n)}$$

$$\therefore \quad \frac{SSE}{\sigma^2} \sim \chi^2_{(n-2)}$$

The estimate of ε_i depends on two unknown parameters β_0 & β_1. Since we have estimated them, two degrees of freedom are lost. Therefore, degree of freedom for residual is $(n-2)$.

$$\therefore \quad E\left[\frac{SSE}{\sigma^2}\right] = E\left[\chi^2_{n-2}\right] = n-2$$

$$\therefore \quad E\left[\frac{SSE}{n-2}\right] = \sigma^2$$

$$\therefore \quad \hat{\sigma}^2 = \frac{SSE}{n-2} = \frac{\sum \varepsilon_i^2}{n-2}$$

$\therefore \quad \hat{\sigma}^2 = MSSE$

where, MSSE is mean sum of square due to error (Residuals).

★ Properties of Ordinary Least Square (OLS) estimators:

1) The OLS estimators are unbiased.
i.e. $E(\hat{\beta}_0) = \beta_0$ and $E(\hat{\beta}_1) = \beta_1$

2) The OLS estimators of the parameters are linear in dependent variable y.

3) The OLS estimators have minimum variance in a class of all such linear unbiased estimators. Thus, the OLS estimators are efficient and Best Linear Unbiased Estimators. (BLUE).

4) The least square regression line always pass through the point $(\bar{x}, \bar{y})$. i.e. Centroid of the data.

5) The OLS estimators are consistent.

6) The sum of the observed value y_i ~~[illegible]~~ equals the sum of the fitted values $\hat{y}_i$.

i.e. $\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{y}_i$.

Save the Children

What is the Sum of Squares?

Sum of Squares (SS) is a statistical measure that is used to study the variation in the data which enable us to know how well the data can fit the model in regression.

In linear regression there are three types of Sum of Squares:

1. SST: Sum of Squares due to Total
2. SSR: Sum of Square due to Regression
3. SSE: Sum of Square due to Residual/Error

Sum of Square due to Total (SST):

Total sum of squares is a deviation of the values of the dependent variable from its sample mean.

It measures the total variation in the sample.

It is denoted by SST and given by,

$$\mathbf{SST} = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

Where, y_i is i[th] sample observation and $\bar{y}$ is sample mean.

Sum of Squares due to Regression (SSR):

The regression sum of squares tells us how well the regression model explains the data and therefore it is also known as Explained Sum of Squares.

It is the sum of squares of the differences between predicted values and the sample mean of the dependent variable.

It is denoted by SSR. It can also be denoted by SSE because it is also known as explained sum of squares. But we will denote it by SSR.

$$\mathbf{SSR} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

Sum of Squares due to Errors/Residuals (SSE):

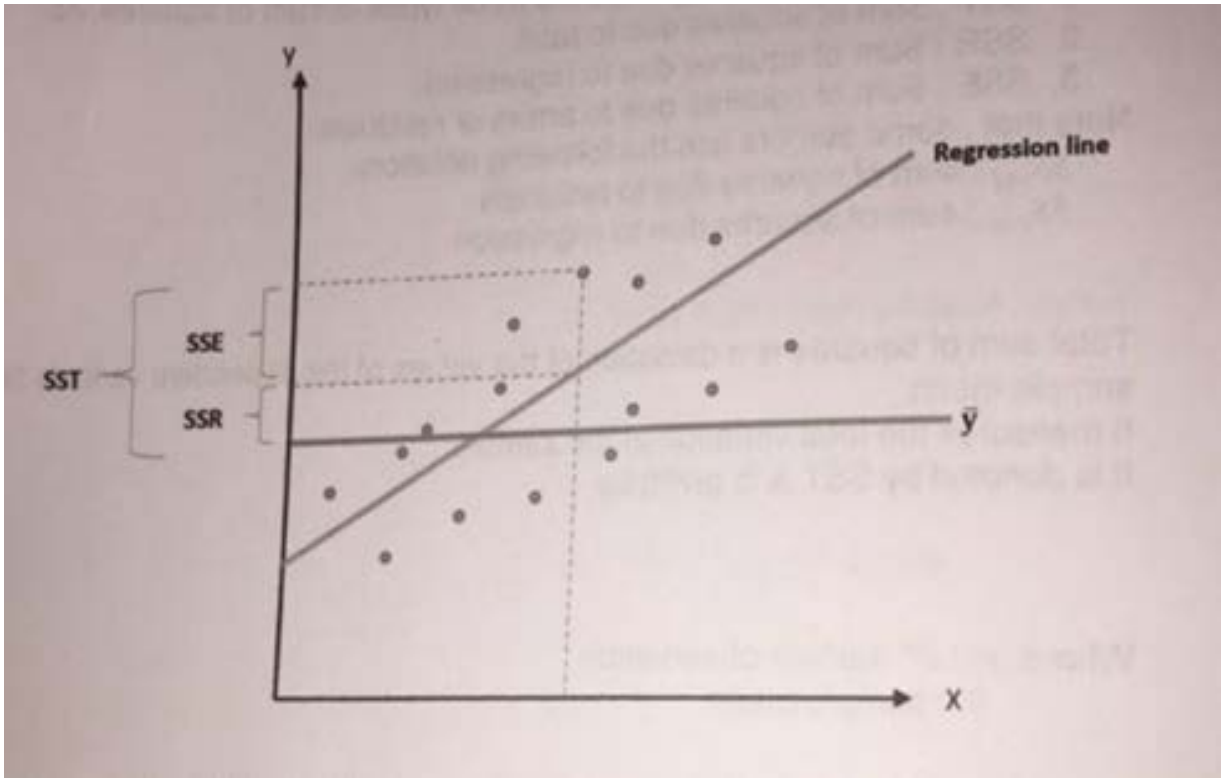
The residual sum of squares measures the variation of modelling error.

It shows how variation in the dependent variable can not be explained by the model. Therefore, sometimes it is called as Unexplained variation.

A lower residual sum of squares indicates that the regression model can better explain the data.

$$\mathbf{SSE} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

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


y
Regression line
SSE
SST
SSR
ȳ
x

We shall now show that,

$$SST = SSR + SSE$$

⇒ Simple Linear Regression Model is :

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Fitted model is

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

where,

$$\hat{\beta}_0 = \bar{y} - \bar{x}\hat{\beta}_1$$

&

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$

Total Sum of Square

$= SST$

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

$$= \sum_{i=1}^{n} \left[y_i - \hat{y}_i + \hat{y}_i - \bar{y}\right]^2$$

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

$$= \sum (y_i - \hat{y}_i)^2 + \sum (\hat{y}_i - \bar{y})^2 + 2\sum (y_i - \hat{y}_i)(\hat{y}_i - \bar{y})$$

$$\therefore SST = SSE + SSR + 2\sum (y_i - \hat{y}_i)(\hat{y}_i - \bar{y})$$

Consider,

$$(y_i - \hat{y}_i) = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$$

$$= y_i - (\bar{y} - \bar{x}\hat{\beta}_1) - \hat{\beta}_1 x_i$$

$$= (y_i - \bar{y}) - \hat{\beta}_1 (x_i - \bar{x})$$

$$(\hat{y}_i - \bar{y}) = \hat{\beta}_0 + \hat{\beta}_1 x_i - \bar{y}$$

$$= \cancel{\bar{y}} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_i - \cancel{\bar{y}}$$

$$= \hat{\beta}_1 (x_i - \bar{x})$$

$$\therefore \; 2 \sum (y_i - \hat{y}_i)(\hat{y}_i - \bar{y})$$

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

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

$$= 2 \hat{\beta}_1 S_{xy} - 2 \hat{\beta}_1^2 S_{xx}$$

$$= 2 \hat{\beta}_1 S_{xy} - 2 \frac{S_{xy}^2}{S_{xx}^2} S_{xx}$$

$$= 2 \hat{\beta}_1 S_{xy} - 2 \frac{S_{xy}}{S_{xx}} \cdot S_{xy}$$

$$= 2 \hat{\beta}_1 S_{xy} - 2 \hat{\beta}_1 S_{xy}$$

$$= 0$$

$$\therefore \; SST = SSR + SSE$$

i.e. $\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2$

Degrees of Freedom:
Degrees of freedom is the number of independent components in the equation.
To understand, consider one example,
Consider one equation a + b + c=10. Here we can choose any values of any two of three variables. Suppose a=5 and b=3.
We can not select the value of c because of the equation.
It means we have the freedom to select the values of any two variables out of three and lost freedom to select the value of third variable because of the constraint of the equation.

The degrees of freedom for regression analysis:

1. Degrees of freedom for SST

   SST = $\boldsymbol{\sum_{i=1}^{n}(y_i - \bar{y})^2}$

   Total number of observations are n.

   One degree of freedom lost because of the constraint $\boldsymbol{\sum_{i=1}^{n}(y_i - \bar{y})^2}$

   Therefore, degrees of freedom for SST are n–1.

2. Degrees of freedom for SSR

   SSR = $\boldsymbol{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}$

   $= \widehat{\beta_1^2} \sum_{i=1}^{n}(x_i - \bar{x})^2$

   $= \widehat{\beta_1^2}$ Sxx

   SSR is completely determined by one parameter $\widehat{\beta_1}$, therefore degree of freedom for SSR is 1.

3. Degrees of freedom for SSE

   SST = SSR + SSE

   df (SST) = df (SSR) + df (SSE)

   n–1 = 1 + df (SSE)

   df (SSE) = n–2

   Degrees of freedom for SSE are n–2.

Estimation of σ^2

$$E\left(\frac{\text{SSRes}}{\text{n}-2}\right) = \sigma^2$$

$$\text{SSRes} = \sum_{\text{i}=1}^{\text{n}} \text{ei}^2 = \sum_{\text{i}=1}^{\text{n}} (\text{yi} - \hat{\text{y}}\text{i})^2$$

$$= \sum_{i=1}^{n} (yi - \widehat{\beta_0} - \widehat{\beta_1}\, Xi)^2$$

Since, $\widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\bar{x}$

$$\text{SSRes} = \sum_{i=1}^{n} (yi - \bar{y} + \widehat{\beta_1}\bar{x} - \widehat{\beta_1}\, Xi)^2$$

$$\text{SSRes} = \sum_{i=1}^{n} (\, yi - \bar{y} - \widehat{\beta_1}(xi - \bar{x}))^2$$

$$\text{SSRes} = \sum_{i=1}^{n} (yi - \bar{y})^2 + \widehat{\beta_1}^2 \sum_{i=1}^{n} (xi - \bar{x})^2 - 2\widehat{\beta_1} \sum_{i=1}^{n} (xi - \bar{x})(yi - \bar{y})$$

$$\text{SSRes} = Syy + \widehat{\beta_1}^2 Sxx - 2\widehat{\beta_1} Sxy$$

We know that

$$\widehat{\beta_1} = \frac{Sxy}{Sxx}$$

$$\text{SSRes} = Syy + \widehat{\beta_1}^2 Sxx - 2\widehat{\beta_1}\widehat{\beta_1} Sxx$$

$$SSRes = Syy - \widehat{\beta_1}^2 Sxx$$

Take expectation on both the sides

$$E(SSRes) = E(Syy) - E(\widehat{\beta_1}^2 Sxx)$$

$$E(Syy) = E[\sum_{i=1}^{n}(yi - \bar{y})^2]$$

$$= E[\sum_{i=1}^{n} yi^2] - nE[\bar{y}^2]$$

Summation and expectation are exchangeable

$$= n\sigma^2 + \sum_{i=1}^{n}(\beta_0 + \beta_1 xi)^2 - \sigma^2 - n(\beta_0 + \beta_1\bar{x})^2$$

$$= n\sigma^2 + \sum_{i=1}^{n}({\beta_0}^2 + {\beta_1}^2 xi^2 + 2\beta_0\beta_1 xi) - \sigma^2 - (n{\beta_0}^2 + n{\beta_1}^2\bar{x}^2 + 2\beta_0\beta_1\bar{x})$$

$= n\sigma^2 + n\beta_0^2 + \beta_1^2(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2) + 2\beta_0\beta_1 n\bar{x} - \sigma^2 - n\beta_0^2 - 2\beta_0\beta_1 n\bar{x}$

$$= (n-1)\sigma^2 + {\beta_1}^2(\sum_{i=1}^{n} xi^2 - n\bar{x}^2)$$

$$= (n-1)\sigma^2 + {\beta_1}^2 Sxx$$

Therefore, $E(Syy) = (n-1)\sigma^2 + \beta_1^2 Sxx$

we have $E\left(\widehat{\beta_1}^2 Sxx\right) = SxxE(\widehat{\beta_1}^2)$

Since $E(\widehat{\beta_1}) = \beta_1$

$$V(\widehat{\beta_1}) = \frac{\sigma^2}{Sxx}$$

$$E\left(\widehat{\beta_1}^2\right) = V(\widehat{\beta_1}) + [E(\widehat{\beta_1})]^2$$

$$= \frac{\sigma^2}{Sxx} + {\beta_1}^2$$

$$E\left(\widehat{\beta_1}^2 Sxx\right) = Sxx(\frac{\sigma^2}{Sxx} + {\beta_1}^2)$$

$$= \sigma^2 + {\beta_1}^2 Sxx$$

Implies

$$E(SSRes) = E(Syy) - E\left(\widehat{\beta_1}^2 Sxx\right)$$

$$= (n-1)\sigma^2 + {\beta_1}^2 Sxx - \sigma^2 - {\beta_1}^2 Sxx$$

$$= (n-2)\sigma^2$$

Therefore,

$$\mathrm{E}\left(\frac{\mathrm{SSRes}}{\mathrm{n}-2}\right) = \sigma^2$$

E(MSSE) = σ^2

Therefore, MSSE is an unbiased estimator of σ^2.

★ Coefficient of Determination :

⇒ Coefficient of determination is the proportion of the variation in the dependent variable (Y) that can be explained by the variation in the indept variable (X).

⇒ It is denoted by R^2 and is given by,

$$R^2 = \frac{\text{Explained variation}}{\text{Total variation}} = \frac{SSR}{SST}$$

$$= \frac{\text{Total variation} - \text{Unexplained variation}}{\text{Total variation}}$$

$$= 1 - \frac{\text{Unexplained variation}}{\text{Total variation}}$$

$$R^2 = 1 - \frac{SSE}{SST}$$

$$\therefore \boxed{R^2 = r^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}}$$

<u>Result:</u> Coefficient of determination is the square of the sample correlation coefficient (r^2).

⇒ $$R^2 = \frac{SSR}{SST}$$

$$= \frac{\Sigma(\hat{y}_i - \bar{y})^2}{\Sigma(y_i - \bar{y})^2}$$

$$R^2 = \frac{\hat{\beta}_1^2 S_{xx}}{S_{yy}}$$

$$= \left[\frac{S_{xy}}{S_{xx}}\right]^2 * \frac{S_{xx}}{S_{yy}}$$

$$= \frac{(S_{xy})^2}{S_{xx} \cdot S_{yy}}$$

$$= \left[\frac{S_{xy}}{\sqrt{S_{xx}} \cdot \sqrt{S_{yy}}}\right]^2$$

$$= \left[\frac{Cov(x,y)}{\sqrt{V(x)}\sqrt{V(y)}}\right]^2$$

$$R^2 = r^2$$

Note: $-1 \leq r \leq 1 \Rightarrow 0 \leq r^2 \leq 1$

$$\therefore \quad 0 \leq R^2 \leq 1$$

⇒ Coefficient of determination is calculated to judge how well the estimated regression equation fits the data.

⇒ The coefficient of determination measures goodness of fit for the estimated regression equation.

⇒ Generally, but not always, a higher value of coefficient of determination value signifies the regression equation is a good fit to the given model.

Save the Children

⇒ To understand the interpretation of R^2, consider the following example.

Suppose, X: Duration of study.
Y: Score in exam

Here, the score in exam is a dependent variable, which depends on the duration of study.
Suppose, correlation coefficient between Y and X is 0.8. Then, coefficient of determination $= R^2 = r^2$
$= (0.8)^2 = 0.64$.

⇒ 64% of the variation in score in exam (Y) can be explained by duration of study (X).
The remaining duration may be due to other independent variables.

Adjusted R^2:

As mentioned earlier, larger the value of R^2 better is the regression model. However, it is not true always.
R^2 is non-decreasing function of the number of explanatory variables or regressor's present in the model. i.e., as the number of regressor's increases, R^2 almost invariably increases and never decreases.
In other words, R^2 always increases by adding independent variables irrespective of the fact that the added independent variables are relevant or not.
Therefore, we go for another statistical measure that penalizes the independent variables which are not correlated to the dependent variable. It is known as Adjusted R^2.

Adjusted R^2 increases only when additional independent variable improves the model.

The Adjusted R^2 adjusts for the number of independent variables in the model. Importantly, its value increases only when the new independent variable improves the model fit more than expected by chance alone. The Adjusted R^2 value actually decreases when the added variable does not improve the model fit.

The following example shows how Adjusted R^2 increases upto a point and then decreases. On the other hand, R^2 increases with each and every additional independent variable.

Consider,
Dependent variable Y: Speed of a car

Independent variables: x_1: Power of Engine
x_2: Weight of car
x_3: Colour of a car

We know that the speed of a car depends on power of engine and weight of a car and not on colour of a car.

Following table shows how the R^2 and Adjusted R^2 affected with addition of the above independent variables one by one.

Sr. No.	Model	R^2	Adjusted R^2
1	$y=\beta_0 + \beta_1 x_1 + \varepsilon$	0.75	0.72
2	$y=\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$	0.85	0.76
3	$y=\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$	0.86	0.74

If we consider only R^2 value to decide the fitness of the model then third model is appropriate. But if we consider Adjusted R^2 then the second model is appropriate. Therefore, along with R^2 value, adjusted R^2 should also be considered.

Adjusted R squared value is denoted by Adjusted R^2 or $\overline{R}^2$ and is given by,

$$\overline{R}^2 = 1 - \frac{SSE/(n-k-1)}{SST/(n-1)} = 1 - \frac{(1-R^2)(n-1)}{(n-k-1)}$$

Where, n: Total number of observations
k: Number of independent variables

How is Adjusted R^2 decreases by adding an irrelevant independent variable?

$$\overline{R}^2 = 1 - \frac{SSE/(n-k-1)}{SST/(n-1)} = 1 - \frac{(1-R^2)(n-1)}{(n-k-1)}$$

If the number of independent variables, k, increases, denominator decreases and hence the quantity $\frac{(1-R^2)(n-1)}{(n-k-1)}$ increases if increase in R^2 is very small. Therefore, Adjusted R^2 decreases.

The quantity $\frac{(1-R^2)(n-1)}{(n-k-1)}$ decreases by adding independent variable only when R^2 value increases significantly.

i.e., when added independent variable is related to dependent variable.

For example,

Suppose, k=2 and $R^2 = 0.8$

Then Adjusted $R^2 = 0.76$

Case (i) Added independent variable is significant (i.e., correlated to the dependent variable)

If added independent variable is correlated to the dependent variable then there will be a significant increase in R^2.

Suppose $R^2 = 0.87$ then using formula, Adj. $R^2 = 0.815$

Case (ii) Added independent variable is not significant (i.e., Added independent variable is not correlated with the dependent variable)

Then there will be increased in R^2 but not significant.

Suppose $R^2 = 0.81$ then using formula, Adj. $R^2 = 0.72$

From the above two cases, it is clear that the R^2 value increases by adding independent variable irrespective of the fact that whether it is relevant or not.
On the other side, Adj. R^2 increases only when added independent variable is relevant otherwise it decreases.

★ Testing of hypothesis & Confidence Interval:-

1] Testing the overall significance:

Hypothesis to be tested:
$H_0: \beta_1 = 0$ against $H_1: \beta_1 \neq 0$

The null hypothesis implies that there is no linear relationship between X and Y.
The alternative hypothesis is that there is a relationship, positive or negative, between X and Y.

Test Statistic:

$$t = \frac{\hat{\beta}_1}{\sqrt{\hat{V}(\hat{\beta}_1)}} = \frac{\hat{\beta}_1}{\hat{S.E.}(\hat{\beta}_1)}$$

$$= \frac{\hat{\beta}_1}{\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}} = \frac{\hat{\beta}_1}{\sqrt{\frac{MSSE}{S_{xx}}}}$$

i.e.

$$t = \frac{\hat{\beta}_1}{\sqrt{\frac{MSSE}{S_{xx}}}} \sim t_{n-2}$$

Decision Cretiria:-
Reject H_0 if $|t_{cal}| > t_{\frac{\alpha}{2}, n-2}$

Save the Children

There is one more approach to test the overall significance of regression i.e. ANOVA.

ANOVA Table

Source of Variation	df	SS	MSS	Fcal	Ftab
Regression	1	SSR	MSSR	$\frac{MSSR}{MSSE}$	$F_{1,n-2,\alpha}$
Error (Residual)	$n-2$	SSE	MSSE		
Total	$n-1$	SST			

where,

$$SST = \Sigma(y_i - \bar{y})^2 = \Sigma y_i^2 - n\bar{y}^2$$

$$SSR = \Sigma(\hat{y}_i - \bar{y})^2 = \hat{\beta}_1 S_{xx} = r^2 S_{yy}$$

$$SSE = SST - SSR$$

MSS : Mean Sum of Square, is a ratio of sum of square to the corresponding df.

$$MSSR = \frac{SSR}{1}, \quad MSSE = \frac{SSE}{n-2}$$

Test statistic :

$$F_{cal} = \frac{MSSR}{MSSE} \sim F_{1,n-2,\alpha}$$

Decision criteria :

Reject H_0 if $F_{cal} > F_{tab}$.

Note :

Consider the following two possibilities :

$H_0 : \beta_1 = 0$ v/s $H_1 : \beta_1 \neq 0$

(i) Null hypothesis is accepted.

i.e. $\beta_1 = 0$ is true.

Our model is, $y = \beta_0 + \beta_1 x + \epsilon$

for $\beta_1 = 0$, $y = \beta_0 + \epsilon$

$\therefore$ when $\beta_1 = 0$, the response variable X has no role to play in the regression model.

when $\beta_1 = 0$, it means the slope of the regression line is zero. It means the regression line is parallel to the X-axis as shown in fig (1).

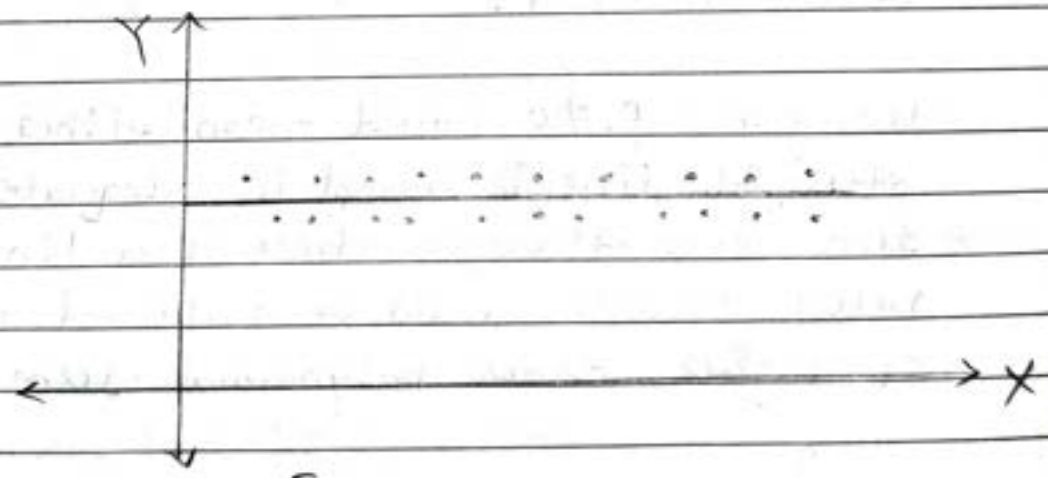
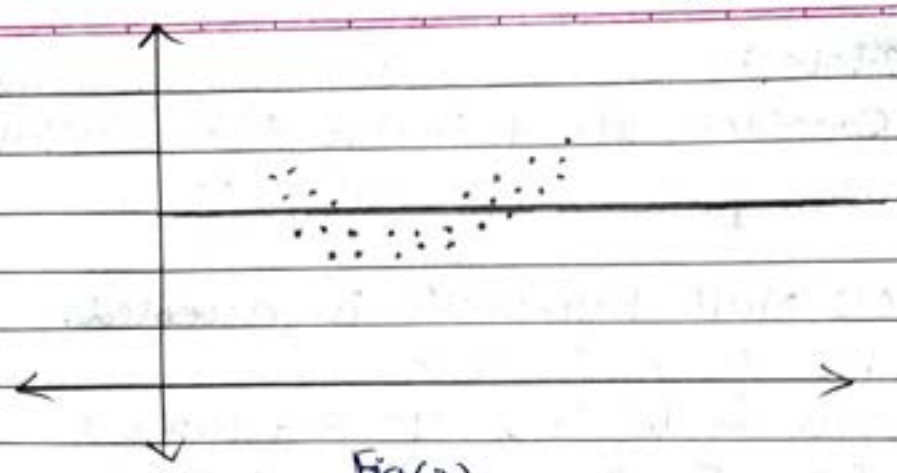


Fig (1)

This shows that, for any value of x, the value of y remain in particular ~~eq~~ range.

i.e. X is not contributing in the model building.



Fig(2)

There may be some relationship between X and Y but it is not linear as shown in fig.(2). Therefore, $\beta_1 = 0$ is accepted means that there is no linear relationship between X and Y.

(ii) Null hypothesis is rejected. i.e. $\beta_1 \neq 0$. $\beta_1 \neq 0$ implies that X is helping in explaining the variation in Y.

However, $\beta_1 \neq 0$ could mean either that the straight line model is adequate (Fig.(3)) or that even though there is a linear effect of X, better results could be obtained with the addition of higher order polynomial term in X. (Fig.(4)).

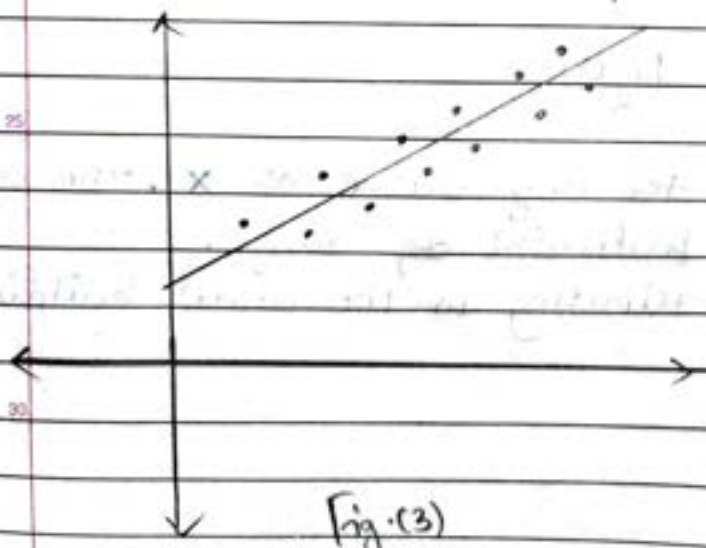


Fig.(3)

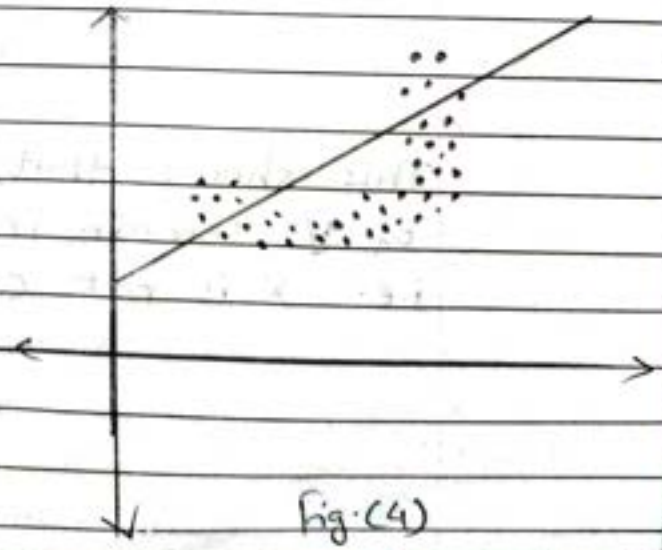


Fig.(4)

2] Test of hypothesis and Confidence interval for the slope parameter (β_1) of the model.

Case (i) when σ^2 is known.
(ii) when σ^2 is unknown.

Hypothesis to be tested :

$H_0 : \beta_1 = \beta_1^*$ where β_1^* is a specified value of β_1.
$H_1 : \beta_1 > \beta_1^*$
$\beta_1 < \beta_1^*$
$\beta_1 \neq \beta_1^*$

Case (i) when σ^2 is known
We know that,

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}, \quad E(\hat{\beta}_1) = \beta_1, \quad V(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}$$

Fitted model is. $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$
$\hat{\beta}_1$ is a linear function of y_i & we know that $y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$

$\therefore$ $\hat{\beta}_1$ is also a Normal variate.

$$\therefore \hat{\beta}_1 \sim N\left[\beta_1, \frac{\sigma^2}{S_{xx}}\right]$$

$$\therefore Z = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\sigma^2}{S_{xx}}}} \sim N(0,1)$$

Test statistic :

$$Z_{cal} = \frac{\hat{\beta}_1 - \beta_1^*}{\sqrt{\frac{\sigma^2}{S_{xx}}}} \sim N(0,1) \text{ under } H_0.$$

Decision criteria :-

(i) $H_1 : \beta_1 > \beta_1^*$

Reject H_0 if $Z_{cal} > Z_{tab} = Z_\alpha$

(ii) $H_1 : \beta_1 < \beta_1^*$

Reject H_0 if $Z_{cal} < -Z_\alpha$

(iii) $H_1 : \beta_1 \neq \beta_1^*$

Reject H_0 if $|Z_{cal}| > Z_{\alpha/2}$

★ Confidence Interval :

Consider $100(1-\alpha)\%$ C.I. for β_1

we know that, $\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right)$

$$\therefore \quad Z = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\sigma^2}{S_{xx}}}} \sim N(0,1)$$

$\therefore \; P[a < z < b] = (1-\alpha)$

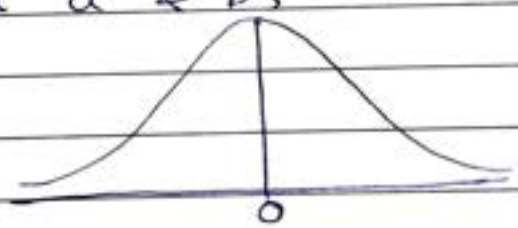
$$P\left[a < \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\sigma^2}{S_{xx}}}} < b\right] = 1-\alpha$$

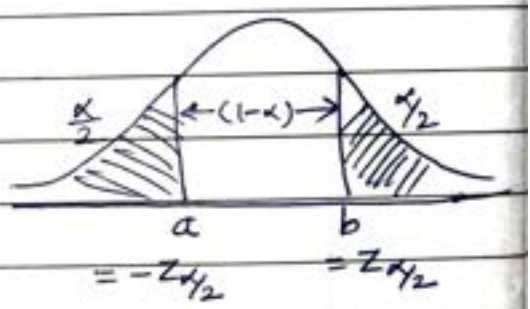
$$P\left[a\sqrt{\frac{\sigma^2}{S_{xx}}} < \hat{\beta}_1 - \beta_1 < b\sqrt{\frac{\sigma^2}{S_{xx}}}\right] = 1-\alpha$$

$$P\left[-\hat{\beta}_1 + a\sqrt{\frac{\sigma^2}{S_{xx}}} < -\beta_1 < -\hat{\beta}_1 + b\sqrt{\frac{\sigma^2}{S_{xx}}}\right] = 1-\alpha$$

$$\therefore P\left[\hat{\beta}_1 - a\sqrt{\frac{\sigma^2}{S_{xx}}} > \beta_1 > \hat{\beta}_1 + b\sqrt{\frac{\sigma^2}{S_{xx}}}\right] = 1-\alpha$$

$\therefore$ To find a & b,





$\therefore$ $100(1-\alpha)\%$ C.I. for β_1 is

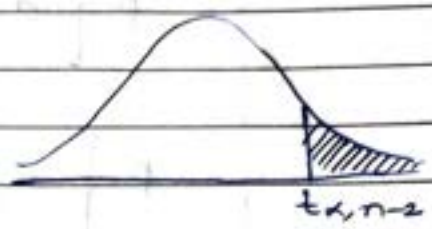
$$\left(\hat{\beta}_1 - Z_{\alpha/2}\sqrt{\frac{\sigma^2}{S_{xx}}}, \; \hat{\beta}_1 + Z_{\alpha/2}\sqrt{\frac{\sigma^2}{S_{xx}}}\right)$$

Case (ii) when σ^2 is unknown.
[which is most common situation in real world]

Test statistic $= \dfrac{\hat{\beta}_1 - \beta_1^*}{\sqrt{\dfrac{MSSE}{S_{xx}}}} \sim t_{n-2}$ under H_0.

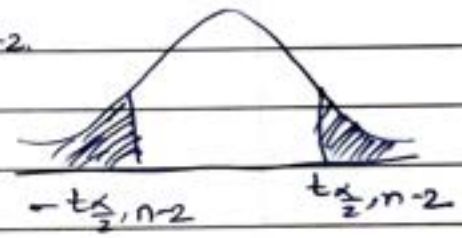
Decision criteria:

(i) $H_1 : \beta_1 > \beta_1^*$
we reject H_0 if $t_{cal} > t_{\alpha, n-2}$



(ii) $H_1 : \beta_1 < \beta_1^*$
we reject H_0 if $t_{cal} < -t_{\alpha, n-2}$

(iii) $H_1 : \beta_1 \neq \beta_1^*$
we reject H_0 if $|t_{cal}| > t_{\frac{\alpha}{2}, n-2}$



* Confidence interval for β_1 when σ^2 is unknown.

$100(1-\alpha)\%$ C.I. for β_1 when σ^2 is unknown.

$$t = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\dfrac{MSSE}{S_{xx}}}} \sim t_{n-2}$$

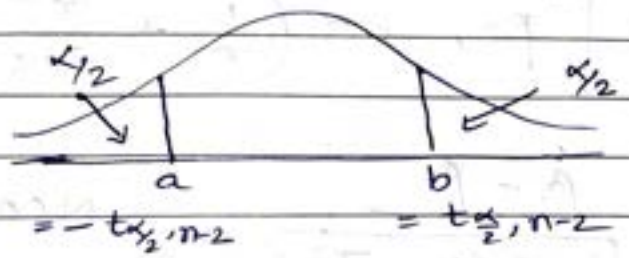
$$P[a < t < b] = (1-\alpha)$$

$$P\left[a < \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{MSSE}{S_{xx}}}} < b\right] = (1-\alpha)$$

$$P\left[a\sqrt{\frac{MSSE}{S_{xx}}} < \hat{\beta}_1 - \beta_1 < b\sqrt{\frac{MSSE}{S_{xx}}}\right] = (1-\alpha)$$

$$P\left[-\hat{\beta}_1 + a\sqrt{\frac{MSSE}{S_{xx}}} < -\beta_1 < -\hat{\beta}_1 + b\sqrt{\frac{MSSE}{S_{xx}}}\right] = (1-\alpha)$$

$$\therefore P\left[\hat{\beta}_1 - b\sqrt{\frac{MSSE}{S_{xx}}} < \beta_1 < \hat{\beta}_1 - a\sqrt{\frac{MSSE}{S_{xx}}}\right] = (1-\alpha)$$



$$\therefore P\left[\hat{\beta}_1 - t_{\frac{\alpha}{2}, n-2}\sqrt{\frac{MSSE}{S_{xx}}} < \beta_1 < \hat{\beta}_1 + t_{\frac{\alpha}{2}, n-2}\sqrt{\frac{MSSE}{S_{xx}}}\right] = (1-\alpha)$$

$\therefore$ $100(1-\alpha)\%$ C.I. for β_1 when σ^2 is unknown :

$$\left(\hat{\beta}_1 - t_{\frac{\alpha}{2}, n-2}\sqrt{\frac{MSSE}{S_{xx}}}, \hat{\beta}_1 + t_{\frac{\alpha}{2}, n-2}\sqrt{\frac{MSSE}{S_{xx}}}\right.$$

3] Testing of hypothesis and confidence interval estimation for the intercept term (β_0) :-

Hypothesis tested :

$H_0 : \beta_0 = 0$ vs $H_1 : \beta_0 \neq 0$

Case (1) when σ^2 is known.

We know that, $E(\hat{\beta}_0) = \beta_0$, $V(\hat{\beta}_0) = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right]$

Also, $\hat{\beta}_0$ is a linear f^n of Normally distributed dependent variable. Therefore, $\hat{\beta}_0$ is also Normally distributed.

$$\therefore \hat{\beta}_0 \sim N \left[\beta_0, \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right) \right]$$

$$\therefore Z = \frac{\hat{\beta}_0 - \beta_0}{\sqrt{\sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right)}} \sim N(0,1)$$

Test statistic :

$$Z_{cal} = \frac{\hat{\beta}_0}{\sqrt{\sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right)}} \sim N(0,1)$$

Decision criteria:

$H_0 : \beta_0 = 0$ is rejected if $|Z_{cal}| > Z_{\alpha/2}$

Save the Children

* $100(1-\alpha)\%$ C.I. for β_0 when σ^2 is known:

$$\left[\hat{\beta}_0 - Z_{\alpha/2}\sqrt{\sigma^2\left(\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right)},\ \hat{\beta}_0 + Z_{\alpha/2}\sqrt{\sigma^2\left(\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right)}\right]$$

Case (ii) when σ^2 is unknown.

$$\text{Test statistic} = \frac{\hat{\beta}_0 - \beta_0}{\sqrt{MSSE\left(\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right)}} \sim t_{n-2}$$

Decision criteria:

Reject H_0, when $|t_{cal}| > t_{\frac{\alpha}{2}, n-2}$

* $100(1-\alpha)\%$ C.I. for β_0 when σ^2 is unknown:

$$\left[\hat{\beta}_0 - t_{n-2,\frac{\alpha}{2}}\ \hat{S.E.}(\hat{\beta}_0),\ \hat{\beta}_0 + t_{n-2,\alpha/2}\ \hat{S.E.}(\hat{\beta}_0)\right]$$

where, $\hat{S.E.}(\hat{\beta}_0) = \sqrt{MSSE\left(\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right)}$

UNIT-2

MULTIPLE LINEAR REGRESSION

⇒ A linear regression model that involves more than one explanatory (independent) variables is called a multiple linear regression model.

⇒ A simple linear regression model studied in previous unit is often inadequate in practice.

Consider the following examples,

1. Suppose Y: Speed of a car
   X: Power of an engine
   Then the simple linear regn model : $Y = \beta_0 + \beta_1 X + \epsilon$

   Though, the speed of a car depends heavily on the power of an engine, it is not only factor that affects the speed of a car.
   The other factors such as weight of a car, size of tyres, pressure in tyres, suspension can also affect.

2. Performance in the exam (dependent variable) can be affected not only by the no. of lectures you have attended but also by duration of revision, No. of hours watching TV, Amount of time on mobile, liking towards the subject, IQ and so on.

3. Yield of a crop depends on fertilizer, temperature, rain, sunshine, seed quality, soil quality, humidity etc.

Save the Children

☆ Linear model with 2 explanatory variables:

The linear regⁿ model with 2 explanatory variables is given by,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u$$

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$$

The error term is given by,

$$e = Y - \hat{Y}$$

$$= Y - \hat{\beta}_0 - \hat{\beta}_1 X_1 - \hat{\beta}_2 X_2$$

* Derivation of ordinary least square estimator (OLS) of regⁿ coefficients for two explanatory variable model:

Acc. to the principle of least square estimate, we obtain the estimates of the parameters by solving the eqⁿ $\frac{\partial S}{\partial \beta_0} = 0$ & $\frac{\partial S}{\partial \beta_1} = 0$ & $\frac{\partial S}{\partial \beta_2} = 0$

where $S = \Sigma e^2 = \Sigma (Y - \hat{Y})^2$

$$= \Sigma (Y - \hat{\beta}_0 - \hat{\beta}_1 X_1 - \hat{\beta}_2 X_2)^2$$

Consider, $\frac{\partial S}{\partial \beta_0} = 0$

$$\cancel{2} \Sigma (Y - \hat{\beta}_0 - \hat{\beta}_1 X_1 - \hat{\beta}_2 X_2)(\cancel{-1}) = 0$$

$$\Sigma Y - n\hat{\beta}_0 - \hat{\beta}_1 \Sigma X_1 - \hat{\beta}_2 \Sigma X_2 = 0$$

$$\Rightarrow \Sigma Y = n\hat{\beta}_0 + \hat{\beta}_1 \Sigma X_1 + \hat{\beta}_2 \Sigma X_2 \quad \text{——— (i)}$$

Dividing both the sides by n we get,

$$\bar{Y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{X}_1 + \hat{\beta}_2 \bar{X}_2 \quad \text{——— (ia)}$$

Consider, $\frac{\partial S}{\partial \beta_1} = 0$

$$2 \Sigma (Y - \hat{\beta}_0 - \hat{\beta}_1 X_1 - \hat{\beta}_2 X_2) \cdot (-X_1) = 0$$

$$\Sigma X_1 Y = \hat{\beta}_0 \Sigma X_1 + \hat{\beta}_1 \Sigma X_1^2 + \hat{\beta}_2 \Sigma X_1 X_2$$

Substituting $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}_1 - \hat{\beta}_2 \bar{X}_2$ ——— (from (ia))

$$\Sigma X_1 Y = (\bar{Y} - \hat{\beta}_1 \bar{X}_1 - \hat{\beta}_2 \bar{X}_2) n\bar{X}_1 + \hat{\beta}_1 \Sigma X_1^2 + \hat{\beta}_2 \Sigma X_1 X_2$$

Save the Children

$$\Sigma x_1 y = n\bar{x}_1\bar{y} - n\hat{\beta}_1\bar{x}_1^2 - n\hat{\beta}_2\bar{x}_1\bar{x}_2 + \hat{\beta}_1\Sigma x_1^2 + \hat{\beta}_2\Sigma x_1 x_2$$

$$\Sigma x_1 y - n\bar{x}_1\bar{y} = \hat{\beta}_1(\Sigma x_1^2 - n\bar{x}_1^2) + \hat{\beta}_2(\Sigma x_1 x_2 - n\bar{x}_1\bar{x}_2).$$

Consider, $\Sigma(x_1-\bar{x}_1)(y-\bar{y})$

$$= \Sigma x_1 y - \bar{y}\Sigma x_1 - \bar{x}_1\Sigma y + n\bar{x}_1\bar{y}$$

$$= \Sigma x_1 y - n\bar{y}\bar{x}_1 - n\bar{y}\bar{x}_1 + n\bar{x}_1\bar{y}$$

$$= \Sigma x_1 y - n\bar{y}\bar{x}_1$$

now, $\Sigma(x_1-\bar{x}_1)^2 = \Sigma x_1^2 - n\bar{x}_1^2$.

now, $\Sigma(x_1-\bar{x}_1)(x_2-\bar{x}_2)$

$$= \Sigma x_1 x_2 - \bar{x}_2\Sigma x_1 - \bar{x}_1\Sigma x_2 + n\bar{x}_1\bar{x}_2$$

$$= \Sigma x_1 x_2 - n\bar{x}_1\bar{x}_2.$$

$$\therefore \Sigma x_1 y = \hat{\beta}_1\Sigma x_1^2 + \hat{\beta}_2\Sigma x_1 x_2 \quad \text{—— (ii)}$$

where, $x_1 = x_1 - \bar{x}_1$, $x_2 = x_2 - \bar{x}_2$

& $y = y - \bar{y}$

Similarly, $\frac{\partial S}{\partial \beta_2} = 0$

$$\Rightarrow \Sigma x_2 y = \hat{\beta}_1\Sigma x_1 x_2 + \hat{\beta}_2\Sigma x_2^2 \quad \text{—— (iii)}$$

(ii) $\times (\Sigma x_1 x_2)$ — (iii) $\times (\Sigma x_1^2)$

$$(\Sigma x_1 y)(\Sigma x_1 x_2) = \hat{\beta}_1(\Sigma x_1^2)(\Sigma x_1 x_2) + \hat{\beta}_2(\Sigma x_1 x_2)^2$$

$$(\Sigma x_1^2)(\Sigma x_2 y) = \hat{\beta}_1(\Sigma x_1^2)(\Sigma x_1 x_2) + \hat{\beta}_2(\Sigma x_1^2)(\Sigma x_2^2).$$

$$(\Sigma x_1 y)(\Sigma x_1 x_2) - (\Sigma x_1^2)(\Sigma x_2 y) = \hat{\beta}_2\left[(\Sigma x_1 x_2)^2 - (\Sigma x_1^2)(\Sigma x_2^2)\right]$$

$$\hat{\beta}_2 = \frac{(\Sigma x_1 y)(\Sigma x_1 x_2) - (\Sigma x_1^2)(\Sigma x_2 y)}{(\Sigma x_1 x_2)^2 - (\Sigma x_1^2)(\Sigma x_2^2)}$$

Save the Children

Date: / / Page No.

$$\Rightarrow \boxed{\hat{\beta}_2 = \frac{(\Sigma x_1^2)(\Sigma x_2 y) - (\Sigma x_1 y)(\Sigma x_1 x_2)}{(\Sigma x_1^2)(\Sigma x_2^2) - (\Sigma x_1 x_2)^2}}$$

Similarly,

$$\boxed{\hat{\beta}_1 = \frac{(\Sigma x_2^2)(\Sigma x_1 y) - (\Sigma x_2 y)(\Sigma x_1 x_2)}{(\Sigma x_1^2)(\Sigma x_2^2) - (\Sigma x_1 x_2)^2}}$$

from (ia) $\hat{\beta}_0$ can be obtained by substituting $\hat{\beta}_1$ and $\hat{\beta}_2$.

$$\bar{Y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{X}_1 + \hat{\beta}_2 \bar{X}_2$$

$$\boxed{\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}_1 - \hat{\beta}_2 \bar{X}_2}$$

* For two explanatory variables model, Coefficient of determination:

$$R^2 = \frac{SSR}{SST} = \frac{\Sigma(\hat{y}_i - \bar{y})^2}{\Sigma(y_i - \bar{y})^2}$$

$$\hat{y}_i - \bar{y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 - \bar{y}$$

$$= \cancel{\hat{\beta}_0} + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 - \cancel{\hat{\beta}_0} - \hat{\beta}_1 \bar{X}_1 - \hat{\beta}_2 \bar{X}_2$$

$$= \hat{\beta}_1 (X_1 - \bar{X}_1) + \hat{\beta}_2 (X_2 - \bar{X}_2)$$

$$= \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$ where, $x_1 = X_1 - \bar{X}_1$, $x_2 = X_2 - \bar{X}_2$

From eq^n (ii),

$$\Sigma x_1 y = \hat{\beta}_1 \Sigma x_1^2 + \hat{\beta}_2 \Sigma x_2^2$$

$$\hat{\beta}_1 \Sigma x_1 y = \hat{\beta}_1^2 \Sigma x_1^2 + \hat{\beta}_1 \hat{\beta}_2 \Sigma x_2^2 \quad \text{— (a)}$$ — multiply both sides by $\hat{\beta}_1$

From eq^n (iii),

$$\Sigma x_2 y = \hat{\beta}_1 \Sigma x_1 x_2 + \hat{\beta}_2 \Sigma x_2^2$$

$$\hat{\beta}_2 \Sigma x_2 y = \hat{\beta}_1 \hat{\beta}_2 \Sigma x_1 x_2 + \hat{\beta}_2^2 \Sigma x_2^2 \quad \text{— (b)}$$ — multiply both sides by $\hat{\beta}_2$.

$$(\hat{y}_i - \bar{y})^2 = \hat{\beta}_1^2 x_1^2 + \hat{\beta}_2^2 x_2^2 + 2 \hat{\beta}_1 \hat{\beta}_2 x_1 x_2$$

$$\Sigma(\hat{y}_i - \bar{y})^2 = \hat{\beta}_1^2 \Sigma x_1^2 + \hat{\beta}_2^2 \Sigma x_2^2 + 2 \hat{\beta}_1 \hat{\beta}_2 \Sigma x_1 x_2$$

$$SSR = \hat{\beta}_1 \Sigma x_1 y + \hat{\beta}_2 \Sigma x_2 y$$ — from (a) & (b)

Save the Children

M T W T F S

Date: / / Page No.

$$SST = \Sigma(y_i - \bar{y})^2$$

Coefficient of Determination

$$= R^2 = \frac{SSR}{SST}$$

$$= \frac{\hat{\beta}_1 \Sigma x_1 y + \hat{\beta}_2 \Sigma x_2 y}{\Sigma y^2}$$

where, $\Sigma x_1 y = \Sigma X_1 Y - n\bar{X}_1\bar{Y}$

$\Sigma x_2 y = \Sigma X_2 Y - n\bar{X}_2\bar{Y}$

$\Sigma y^2 = \Sigma(Y_i - \bar{Y})^2$

Date: / / Page No.

★ Procedure of testing the overall significance of the model :

Hypothesis to be tested :

$H_0: \beta_1 = \beta_2 = 0$

$H_1: \beta_i \neq 0 \quad , i = 1, 2$

i.e. H_1 : at least one of two variables is significant.

ANOVA table

Source of Variation	df	SS	MSS	Fcal	Ftab
Regression	2	SSR	MSSR	$F_{cal} = \frac{MSSR}{MSSE}$	$F_{tab} = F_{\alpha, 2, n-3}$
Error	$n-3$	SSE	MSSE		
Total	$n-1$	SST			

where, $SST = \Sigma y^2 = \Sigma(y_i - \bar{y})^2$

$SSR = R^2 \Sigma y^2 = R^2 \Sigma(y_i - \bar{y})^2$

where, $R^2 = \dfrac{\hat{\beta}_1 \Sigma x_1 y + \hat{\beta}_2 \Sigma x_2 y}{\Sigma y^2}$

$\Sigma x_1 y = \Sigma x_1 y - n\bar{x}_1 \bar{y}$

$\Sigma x_2 y = \Sigma x_2 y - n\bar{x}_2 \bar{y}$

$\Sigma y^2 = \Sigma(y_i - \bar{y})^2$

$SSE = SST - SSR$

$MSSR = \frac{SSR}{2}$ and $MSSE = \frac{SSE}{n-3}$

Reject H_0 if $F_{cal} > F_{tab} = F_{\alpha, 2, n-3}$

Conclusion :

If H_0 is Rejected $\Rightarrow$ β_1 and β_2 are significant.

If H_0 is not rejected $\Rightarrow$ explanatory variables have no linear influence on the dependent variable.

★ The Incremental or marginal contribution of an explanatory variable :

$\Rightarrow$ We first regress Y on X_1 and assess its significance and then add X_2 to the model to find out whether it contributes anything.
By contribution we mean whether the addition of the variable to the model increases SSR (and thus R^2) significanty in relation to the SSE. This contribution may appropriately be called the incremental, or marginal, contribution of an explanatory variable.

$\Rightarrow$ The topic of incremental contribution is an important one in practice. In most emperical investigations the researcher may not be completely sure whether it is worth adding an X variable to the model knowing that several other X variables are already present in the model. One does not wish to include variable(s) that contribute very little towards SSR.

By the same time, one does not want to exclude variable(s) that substantially increase SSR. But how does one decide whether an X variable significantly reduces SSE ?

The ANOVA technique can be easily extended to answer this question.

* Procedure of testing the significance of incremental contribution of explanatory variable for the two explanatory model :

⇒ Hypothesis to be tested :

H_0 : incremental contribution of X_2 is NOT significant.

v/s

H_1 : incremental contribution of X_2 is significant.

Source of variation	df	SS	MSS	Fcal
Due to X_1	1	S_1		
Due to addition of X_2	1	S_2	$MSS = \frac{S_2}{1}$	$Fcal = \frac{S_2}{S_4/n-3}$
Due to X_1 & X_2	2	S_3		
Error	$n-3$	S_4	MSSE	
Total	$n-1$	S_5		

where,

$S_1 = \hat{\beta}_1^{*2} \Sigma x_1^2$

where, $\beta_1^* = \frac{\Sigma x_1 y}{\Sigma x_1^2}$

$S_3 = SSR = R^2 \Sigma y^2 = R^2 \Sigma (Y_i - \bar{Y})^2$

$S_2 = S_3 - S_1$

$S_4 = SSE = SST - SSR = (1-R^2) \Sigma (Y_i - \bar{Y})^2$

$S_5 = SST = \Sigma y^2 = \Sigma (Y_i - \bar{Y})^2$

Reject H_0 if $Fcal > Ftab$.

$Fcal = \frac{S_2}{S_4/n-3} \sim F_{\alpha,1,n-3}$

* Procedure of testing significance of individual coefficients :

1] Hypothesis to be tested :

$H_0 : \beta_0 = 0$ v/s $H_1 : \beta_0 \neq 0$

Test statistic = $t_{cal} = \dfrac{\hat{\beta}_0 - \beta_0}{S.\hat{E}.(\hat{\beta}_0)} \sim t_{n-3}$

$$t = \frac{\hat{\beta}_0}{S.\hat{E}.(\hat{\beta}_0)} \sim t_{n-3} \text{ under } H_0$$

Reject H_0 if $|t_{cal}| > t_{n-3, \frac{\alpha}{2}}$

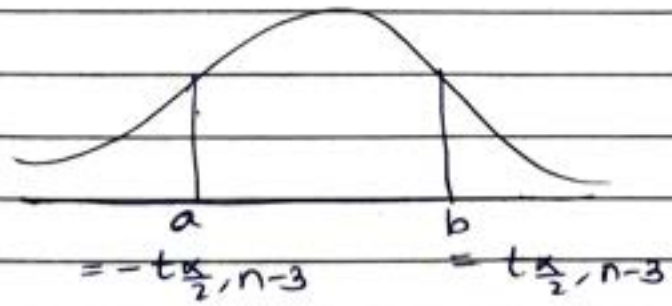
* $100(1-\alpha)\%$ C.I. for individual coefficients :

$P(a < t < b) = (1-\alpha)$

$$P\left[a < \frac{\hat{\beta}_0 - \beta_0}{S.\hat{E}.(\hat{\beta}_0)} < b\right] = 1-\alpha$$

$$P\left[-\hat{\beta}_0 + a.S.\hat{E}.(\hat{\beta}_0) < -\beta_0 < -\hat{\beta}_0 + b.S.\hat{E}.(\hat{\beta}_0)\right] = 1-\alpha$$

$$P\left[\hat{\beta}_0 - b\,S.\hat{E}.(\hat{\beta}_0) < \beta_0 < \hat{\beta}_0 - a\,S.\hat{E}.(\hat{\beta}_0)\right] = 1-\alpha$$



$\therefore$ $100(1-\alpha)\%$ C.I. for β_0 :

$$\left(\hat{\beta}_0 - t_{n-3, \frac{\alpha}{2}} \cdot S\hat{E}.(\hat{\beta}_0),\ \hat{\beta}_0 + t_{n-3, \frac{\alpha}{2}} \cdot S.\hat{E}.(\hat{\beta}_0)\right)$$

Date: / / Page No.

Matrix Theory

⇒ x be any matrix given by,

$$x = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \quad \text{Then,} \quad x' = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix}$$

⇒ $(AB)' = B'A'$

Ex. $A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$, $B = \begin{bmatrix} 5 & 6 \\ 7 & 8 \end{bmatrix}$

$$A*B = \begin{bmatrix} 19 & 22 \\ 43 & 50 \end{bmatrix}$$

$$(AB)' = \begin{bmatrix} 19 & 43 \\ 22 & 50 \end{bmatrix} \quad \text{—— } (*)$$

$$B' = \begin{bmatrix} 5 & 7 \\ 6 & 8 \end{bmatrix}, \quad A' = \begin{bmatrix} 1 & 3 \\ 2 & 4 \end{bmatrix}$$

$$B'A' = \begin{bmatrix} 19 & 43 \\ 22 & 50 \end{bmatrix} \quad \text{—— } (**)$$

$$\Rightarrow \quad y'X\beta = (\beta'X'y)'$$

$y = n\times 1$
$X = n\times p$
$\beta = p\times 1$

$$y'_{1\times n}\, X_{n\times p}\, \beta_{p\times 1} = [\ \ ]_{1\times p} * [\ \ ]_{p\times 1} = [\ \ ]_{1\times 1}$$

$$\beta'X'y = [\ \ ]_{1\times p} * [\ \ ]_{p\times n} * [\ \ ]_{n\times 1}$$

$$= [\ \ ]_{1\times n} * [\ \ ]_{n\times 1}$$

$$= [\ \ ]_{1\times 1}$$

23 Sunday <u>For example</u> :

$$y = \begin{bmatrix} 1 \\ 2 \end{bmatrix}_{2\times 1}, \quad X = \begin{bmatrix} 3 & 4 \\ 5 & 6 \end{bmatrix}_{2\times 2}, \quad \beta = \begin{bmatrix} 7 \\ 8 \end{bmatrix}_{2\times 1}$$

$$y'X\beta = \begin{bmatrix} 1 & 2 \end{bmatrix} \begin{bmatrix} 3 & 4 \\ 5 & 6 \end{bmatrix} \begin{bmatrix} 7 \\ 8 \end{bmatrix} = \begin{bmatrix} 13 & 16 \end{bmatrix} * \begin{bmatrix} 7 \\ 8 \end{bmatrix}$$

$$= 219$$

$$\beta'X'y = \begin{bmatrix} 7 & 8 \end{bmatrix} \begin{bmatrix} 3 & 5 \\ 4 & 6 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \end{bmatrix} = \begin{bmatrix} 53 & 83 \end{bmatrix} * \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

$$= 219$$

★ Multiple Linear Regression Model :

Model :

The multiple linear regression model with 'k' explanatory variables is given by,

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \varepsilon$$

where,

Y : Dependent variable

β_0 : intercept

$\beta_1, \beta_2, \dots, \beta_k$: regression coefficients

$x_1, x_2, \dots, x_k$: independent (explanatory) variables.

- Suppose, we have n observations on $x_1, x_2, \dots, x_k$

Observation	Y	x_1	x_2	x_3	...	x_k
1	y_1	x_{11}	x_{12}	x_{13}	...	x_{1k}
2	y_2	x_{21}	x_{22}	x_{23}	...	x_{2k}
3	y_3	x_{31}	x_{32}	x_{33}	...	x_{3k}
⋮	⋮	⋮	⋮	⋮		⋮
n	y_n	x_{n1}	x_{n2}	x_{n3}	...	x_{nk}

Let us represent model in a matrix form :

$$\underset{n\times 1}{\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix}} = \underset{n\times(k+1)}{\begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ 1 & x_{31} & x_{32} & \dots & x_{3k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix}} \underset{(k+1)\times 1}{\begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}} + \underset{n\times 1}{\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \vdots \\ \varepsilon_n \end{bmatrix}}$$

Save the Children

Let, $p = K+1$

$\therefore$ Model : $\underset{n\times 1}{Y} = \underset{n\times p}{X} \underset{p\times 1}{\beta} + \underset{n\times 1}{\varepsilon}$

or simply. $Y = X\beta + \varepsilon$

where

$Y_{n\times 1}$: $n\times 1$ column vector of obs^{ns} on the dependent variable Y.

$X_{n\times p}$: $n\times p$ matrix giving n obs^{ns} on K explanatory variables. The 1st column of 1's representing the intercept term. (This matrix is also known as the data matrix)

$\beta_{p\times 1}$: $p\times 1$ column vector of the unknown parameters $\beta_0, \beta_1, \dots, \beta_K$

* OLS Estimation :

Model in matrix form,

$$Y = X\beta + \varepsilon$$
$$\therefore \varepsilon = Y - X\beta$$

In OLS, we try to minimize S.S. due to errors.

Let $S = \sum \varepsilon_i^2$ for scalar

$= \varepsilon'\varepsilon$ for matrix

$$\therefore S = (Y - X\beta)'(Y - X\beta)$$
$$= Y'Y - Y'X\beta - \beta'X'Y + \beta'X'X\beta$$
$$= Y'Y - 2\beta'X'Y + \beta'X'X\beta$$

$$\frac{\partial S}{\partial \beta} = \frac{\partial}{\partial \beta}\left[Y'Y - 2\beta' X'Y + \beta' X' X \beta\right] = 0$$

$$\Rightarrow \quad 0 - 2X'Y + 2X'X\beta = 0$$

$$\Rightarrow \quad X'X\hat{\beta} = X'Y$$

$$\boxed{\hat{\beta} = (X'X)^{-1} X'Y}$$

provided that inverse matrix $(X'X)^{-1}$ exists.

The $(X'X)^{-1}$ matrix will always exist if the regressors are linearly indep^t, that is, if no column of the X matrix is a linear combination of the other columns.

{ If rank of X is $p = k+1$ then all the columns of matrix X are indep^t }

* Fitted model :

$$\hat{Y} = X\hat{\beta}$$
$$= \underline{X(X'X)^{-1}X'}Y$$
$$= HY$$

where, $H = X(X'X)^{-1}X'$

H matrix is called Hat matrix.

★ Statistical Properties of Least-Square Estimator :

1) $\hat{\beta} = (X'X)^{-1} X'Y$ is an u.e. of β.

$$E(\hat{\beta}) = E\left[(X'X)^{-1} X'Y\right]$$

$$= E\left[(X'X)^{-1} X' (X\beta + \varepsilon)\right] \quad \because Y = X\beta + \varepsilon$$

$$= E\left[(X'X)^{-1}(X'X)\beta\right] + E\left[(X'X)^{-1} X' \varepsilon\right]$$

$$E(\hat{\beta}) = \beta \qquad \left\{ \begin{array}{l} \because \varepsilon \sim N(0, \sigma^2) \\ \therefore E(\varepsilon) = 0 \\ \& \ (X'X)^{-1}(X'X) = I \end{array} \right\}$$

2) $$Var(\hat{\beta}) = Var\left[(X'X)^{-1} X'Y\right] \qquad \{Var(aX) = aV(X)a'\}$$

$$= \left[(X'X)^{-1} X'\right] Var(Y) \left[(X'X)^{-1} X'\right]'$$

$$= (X'X)^{-1} X' \underset{\sim}{\sigma^2} X (X'X)^{-1} \qquad \left\{\because (X'X)^{-1} = \left[(X'X)^{-1}\right]'\right\}$$

$$= (X'X)^{-1}(X'X)\sigma^2 (X'X)^{-1}$$

$$Var(\hat{\beta}) = \sigma^2 (X'X)^{-1}$$

★ SST, SSR, SSE

$$SSE = \varepsilon'\varepsilon$$
$$= Y'Y - 2\hat{\beta}'X'Y + \hat{\beta}'X'X\hat{\beta}$$
$$= Y'Y - 2\hat{\beta}'X'Y + \hat{\beta}'\underbrace{X'X(X'X)^{-1}}X'Y \quad \{\because \hat{\beta} = (X'X)^{-1}X'Y\}$$
$$= Y'Y - 2\hat{\beta}'X'Y + \hat{\beta}'X'Y$$
$$= Y'Y - \hat{\beta}'X'Y$$

$$\hat{\sigma}^2 = \frac{SSE}{\text{Error d.f.}} = \frac{SSE}{n-p} = \frac{SSE}{n-k-1} = MSSE$$

$\Rightarrow$ MSSE is an u.e. of σ^2.

$$SST = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum Y_i^2 - n\bar{Y}^2$$
$$= Y'Y - n\bar{Y}^2$$

$$\therefore SSR = SST - SSE$$
$$= Y'Y - n\bar{Y}^2 - Y'Y + \hat{\beta}'X'Y$$
$$= \hat{\beta}'X'Y - n\bar{Y}^2$$

★ Coefficient of determination R^2 in matrix notation:

$$R^2 = \frac{SSR}{SST} = \frac{\hat{\beta}'X'Y - n\bar{Y}^2}{Y'Y - n\bar{Y}^2}$$

Date: / / Page No.

★ Procedure for testing overall significance of the model :

Hypothesis to be tested :

$H_0 : \beta_1 = \beta_2 = \dots = \beta_K = 0$

H_1 : Atleast one variable is significant.

$p = K+1$

$\therefore p-1 = K$

S.V.	d.f.	S.S.	M.S.S.	Fcal
Regression	$p-1$	SSR	MSSR	$F_{cal} = \frac{MSSR}{MSSE}$
Error	$n-p$	SSE	MSSE	
Total	$n-1$	SST		

$$SST = Y'Y - n\bar{Y}^2$$

where, $SSR = \hat{\beta}'X'Y - n\bar{Y}^2$

$$SSE = SST - SSR = Y'Y - \hat{\beta}'X'Y$$

Test statistic $= F_{cal} = \frac{MSSR}{MSSE} \sim F_{p-1, n-p, \alpha}$

We reject H_0, if $F_{cal} > F_{tab}$.

☆ Hypothesis testing about individual regression coefficient in matrix notations :

→ Hypothesis to be tested :

$H_0 : \beta_i = 0 \quad i = 0, 1, 2, \ldots, k$

$H_1 : \beta_i \neq 0$

We have assumed that,

$\varepsilon \sim N(0, \sigma^2 I)$

where,

ε is $n \times 1$ vector of error components.

we can show that,

$\hat{\beta} \sim N(\beta, \sigma^2 (X'X)^{-1})$

Test statistic :

$$t = \frac{\hat{\beta}_i - \beta_i}{\hat{S.E}(\hat{\beta}_i)} \sim t_{n-p} \quad \text{i.e. } t_{n-k-1}$$

Reject H_0 if $|t_{cal}| > t_{\frac{\alpha}{2}, n-p}$

Ex. Consider the data in the table

Y	6	8	1	0	5	3	2	-4	10	-3	5
X_1	1	4	9	11	3	8	5	10	2	7	6
X_2	8	2	-8	-10	6	-6	0	-12	4	-2	-4

Soln:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

$$Y = \begin{bmatrix} 6 \\ 8 \\ 1 \\ 0 \\ 5 \\ 3 \\ 2 \\ -4 \\ 10 \\ -3 \\ 5 \end{bmatrix}_{11\times 1}, \quad X = \begin{bmatrix} 1 & 1 & 8 \\ 1 & 4 & 2 \\ 1 & 9 & -8 \\ 1 & 11 & -10 \\ 1 & 3 & 6 \\ 1 & 8 & -6 \\ 1 & 5 & 0 \\ 1 & 10 & -12 \\ 1 & 2 & 4 \\ 1 & 7 & -2 \\ 1 & 6 & -4 \end{bmatrix}_{11\times 3}$$

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}_{3\times 1}, \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_{11} \end{bmatrix}_{11\times 1}$$

The model can be written in matrix form as

$$Y = X\beta + \varepsilon$$

& the least square estimates of β_0, β_1 & β_2 are

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

$$X'X = \begin{bmatrix} \quad \end{bmatrix}_{3\times 11} \begin{bmatrix} \quad \end{bmatrix}_{11\times 3} = \begin{bmatrix} \quad \end{bmatrix}_{3\times 3}$$

$$X'X = \begin{bmatrix} 11 & 66 & -22 \\ 66 & 506 & -346 \\ -22 & -346 & 484 \end{bmatrix}$$

$$(X'X)^{-1} = \begin{bmatrix} 4.3705 & -0.8495 & -0.4086 \\ -0.8495 & 0.1690 & 0.0822 \\ -0.4086 & 0.0822 & 0.0422 \end{bmatrix}$$

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

$$= \begin{bmatrix} 4.3705 & -0.8495 & -0.4086 \\ -0.8495 & 0.1690 & 0.0822 \\ -0.4086 & 0.0822 & 0.0422 \end{bmatrix} \underset{3\times 11}{\begin{bmatrix} \quad \end{bmatrix}} \underset{11\times 1}{\begin{bmatrix} \quad \end{bmatrix}}$$

$$= \begin{bmatrix} \quad \end{bmatrix} \begin{bmatrix} 33 \\ 85 \\ 142 \end{bmatrix}$$

$$= \begin{bmatrix} 14 \\ -2 \\ -1/2 \end{bmatrix}$$

$\hat{\beta}_0 = 14$, $\hat{\beta}_1 = -2$, $\hat{\beta}_2 = -1/2 = -0.5$

So our fitted model will be

$$Y = 14 + (-2)X_1 - \frac{X_2}{2}$$

AUTOCORRELATION

- We define correlation coefficient as a measure of the strength of the linear relationship betn two variables X and Y. i.e. corr (X, Y)
While the term autocorrelation is defined as the correlation between members of series of obs^{ns} ordered in time. i.e. (X_t, X_{t-1})

- In the regression context, the CLRM assumes that the disturbances term relating to any obs^n is not influenced by the disturbance term relating to any other observation.
Symbolically, $E(u_i u_j) = 0 \quad , i \neq j$

- However, if there is correlation between the disturbance term then $E(u_i u_j) \neq 0 \quad , i \neq j$
Then it is termed as autocorrelation.

- In simple words,
One of the basic assumptions in the CLRM is that the random error components or disturbances are iid distributed. So in the model it is assumed that

$$E(u_t, u_{t-s}) = \begin{cases} \sigma_u^2 & \text{if } s=0 \\ 0 & \text{if } s=0 \end{cases}$$

- In this assumption, when $E(u_t, u_{t-s}) = \sigma_u^2$, $s=0$ is violated, i.e., the variance of disturbance term does not remain constant, then the problem of heteroscedasticity arises.

- When $E(u_t, u_{t-s}) = 0$, $s \neq 0$ is violated, i.e. variance of disturbance term remains constant though the successive disturbance terms are correlated, then such problem is termed as the problem of autocorrelation.

Consequences of using OLS estimators in presence of Autocorrelation:

In presence of autocorrelation the OLS estimators are linear, unbiased, consistent & asymptotically normally distributed.

But they are no longer efficient. i.e. their variances are not minimum.

(1) Because of which confidence intervals are likely to be wider.

(2) We are likely to declare a coefficient statistically insignificant even though in fact it may be

Save the Children

statistically significant

* methods for detecting Autocorrelation —

1) Durbin-Watson (D-W) method.

Durbin & Watson have developed the test for detecting autocorrelation.

D-W's 'd' statistic is defined as

$$d = \frac{\Sigma(e_t - e_{t-1})^2}{\Sigma e_t^2}$$

which is a ratio of sum of squared differences in successive residuals to RSS

$$d = \frac{\Sigma(e_t^2 + e_{t-1}^2 - 2e_t e_{t-1})}{\Sigma e_t^2}$$

$$= \frac{\Sigma e_t^2 + \Sigma e_{t-1}^2 - 2\Sigma e_t e_{t-1}}{\Sigma e_t^2}$$

$$d \doteq \frac{2\Sigma e_t^2 - 2\Sigma e_t e_{t-1}}{\Sigma e_t^2}$$ ($\because \Sigma e_{t-1}^2$ differes only by 1 term $\therefore \Sigma e_{t-1}^2 = \Sigma e_t^2$)

$$= \frac{2(\Sigma e_t^2 - \Sigma e_t e_{t-1})}{\Sigma e_t^2}$$

$$= 2\left(1 - \frac{\Sigma e_t e_{t-1}}{\Sigma e_t^2}\right)$$

$$= 2(1 - \hat{\rho})$$

where, $\hat{\rho} = \frac{\Sigma e_t e_{t-1}}{\Sigma e_t^2}$ is called as <u>sample first order autocorrelation</u>.

If $\hat{\rho}$	d	Conclusion
0	2	no autocorrelation
1	0	perfect +ve correlation in residuals
-1	4	perfect -ve

Conclusion

Thus, if 'd' is closer to zero, there is greater

evidence of +ve autocorrelation.

② If d is closer to 4 there is greater evidence of −ve autocorrelation.

Durbin-Watson have successfully derived lower bound d_L & Upper bound d_U such that computed d lies outside these critical values a decision can be made regarding the presence of +ve or −ve or no autocorrelation as indicated below −

$0 < d \leq d_L$ +ve autocorrelation

$d_L < d \leq d_U$ No decision

$d_U < d < 4 - d_U$ No +ve or −ve autocorrelation

$4 - d_U < d \leq 4 - d_L$ No decision

$4 - d_L < d \leq 4$ −ve autocorrelation.

Ex. The following table gives the values of residuals from a regⁿ eqn with one explanatory variable. Test for the presence of autocorrelation.

(Given) $e_t = Y_t - \hat{Y}_t$	(calculate) $(e_t - e_{t-1})^2$	(cal). e_t^2
1	—	1
−1.75	7.5625	3.0625
2.35	16.81	5.5225
1.65	0.49	2.7225
−0.3	3.8025	0.09
−2.95	7.0225	8.7025
	35.6875	21.1

$$d = \frac{\Sigma(e_t - e_{t-1})^2}{\Sigma e_t^2} = \frac{35.6875}{21.1} = 1.6914$$

for $n=6$, $(k=1)$ → no. of explanatory variables, the critical values from D-W table are, $d_L = 0.61$, $d_U = 1.4$

$$d_U < d < 4 - d_U$$

$$1.4 < 1.6914 < 2.6$$

i.e. $d = 1.6914 \in (d_U = 1.4,\ 4 - d_U = 2.6)$

∴ There is no autocorrelation.

2) RUN TEST

Let e_t $(t = 1, 2, \dots, n)$ be the residuals.

Let n_1: no. of +ve residuals.

n_2: no. of −ve residuals.

R: no. of runs of the residuals.

Under the (null) hypo. that successive residuals are indept.

Then, $E(R) = \frac{2n_1n_2}{n+1}$ & $V(R) = \frac{2n_1n_2(2n_1n_2 - n)}{n^2(n-1)}$

If n_1 and n_2 are greater than 10 then

$$Z = \frac{R - E(R)}{\sqrt{V(R)}} \sim SND$$

If $|Z| > Z_{\alpha/2}$ then reject H_0 and conclude that residuals are not random.

i.e. Residuals exhibit autocorrelation.

As a general rule if there is +ve autocorrelation, the no. of runs will be few.

And if there is -ve autocorrelation, the no. of runs will be many.

Swed and Eisenhart have developed a statistical table that gives critical values of run expected. In the random sequence of n obs^{ns}.

If n_1 & n_2 are smaller than 20 which can be used to detect autocorrelation.

Ex. For the following data test for presence of autocorrelation using RUN test.

Year	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003
(residuals) e_t	-0.095	0.19	0.2	-0.385	-0.115	0.025	0.205	0.185	-0.32	0.11
	1st run	2nd run		3rd run		4th run			5th run	6th run

H_0: There is no autocorrelation.

$n = 10$

n_1 = no. of +ve residuals = 6

$n_2 = 4$

$R = 6$

$\because n_1$ & $n_2 < 20$

$\therefore$ we obtain the critical values of R from the table for $n_1 = 6, n_2 = 4$, from the table

Critical values of R are (2, 9).

$\because R = 6 \in (2, 9)$

$\therefore$ we may accept H_0 & conclude that there is no autocorrelation.

OR

The disturbance terms are random.

Save the Children

★ Generalised Least Square Method :-

- If autocorrelation exist then OLS estimators do not remain efficient.
- To overcome this lack of efficiency, we use GLS method.

- Consider a two variable eqn model,

$$Y_t = \beta_0 + \beta_1 X_t + u_t$$

In presence of autocorrelation we assume that the error term u_t are generated by the 1st autoregressive scheme given by,

$$u_t = \rho u_{t-1} + v_t$$

Case (i) when ρ is known.

We know that $Y_t = \beta_0 + \beta_1 X_t + u_t$ — ①

$\therefore$ $Y_{t-1} = \beta_0 + \beta_1 X_{t-1} + u_{t-1}$

$\therefore$ $\rho Y_{t-1} = \rho\beta_0 + \rho\beta_1 X_{t-1} + \rho u_{t-1}$ — ②

① — ②

$$Y_t - \rho Y_{t-1} = (1-\rho)\beta_0 + (X_t - \rho X_{t-1})\beta_1 + (u_t - \rho u_{t-1})$$

$$Y_t^* = \beta_0^* + \beta_1^* X_t^* + V_t \rightarrow \text{error term}$$

where, $Y_t^* = Y_t - \rho Y_{t-1}$

$X_t^* = X_t - \rho X_{t-1}$

$\beta_0^* = (1-\rho)\beta_0$

This eqn is called as Generalised ~~difference~~ equation.

- Since error term (V_t) satisfies the usual OLS assumption, we can apply OLS method to transformed variables Y_t^* and X_t^* and obtain the estimators

Save the Children

of β_0^* and β_1^* which are BLUE.

NOTE:
GLS is nothing but OLS applied to transformed model which satisfies a classical assumption.

Case (ii) When ρ is unknown.

- If ρ is unknown, it is estimated using its relationship in terms of D-W 'd' statistic discussed earlier.
  ∴ we get: $\rho = 1 - \frac{d}{2}$

- However it should be noted that relationship may not hold for small samples.

- Using the estimate of ρ transform a model to obtain Generalised difference eqn.

- Applying OLS transformed model's estimates of the parameters can be obtain.

Ex. The following data shows imports in 1000 million dollars & GNP in '000 million dollars. Test the presence of autocorrelation & fit the appropriate regⁿ model of imports on GNP.

Import (Y)	GNP(X)	$\hat{Y}_t$	$e_t = Y_t - \hat{Y}_t$	$(e_t - e_{t-1})^2$
3.75	21.8	3.63	0.12	—
4.01	22.4	3.80	0.21	0.0081
3.71	22.3	3.77	−0.06	0.0729

$Y_t - \hat{\rho} Y_{t-1}$	$X_t - \hat{\rho} X_{t-1}$
—	—
2.054	11.0291
1.6184	10.6162

By regⁿ mode,

$A = -2.4592$

$B = 0.2794$

The fitted regⁿ model / eqⁿ of imports on GNP using OLS method,

$\therefore \hat{Y} = -2.4592 + 0.2794 X$

Test for autocorrelation using D-W method,

$$d = \frac{\Sigma (e_t - e_{t-1})^2}{\Sigma e_t^2} = 0.9568$$

$$\hat{\rho} = 1 - \frac{d}{2} = 0.5216$$

From the D-W table the values of d_L & d_U for $n=20$, $K=1$ are $d_U = 1.411$, $d_L = 1.201$

$\because d = 0.9568 \in [0 < d \leq d_L]$

$\Rightarrow$ There is positive autocorrelation.

Using GLS method;

$\beta_0^* = A = -1.3903$

$\beta_1^* = B = 0.2950$

The fitted reg^n eq^n of Y_t^* on X_t^* is

$$Y_t^* = -1.3903 + 0.2950 X_t^*$$

॥ श्री ॥

★ Homoscedasticity –

- One of the important assumption of the CLRM is that the variance of each disturbance term conditional on the chosen values of explanatory variables is same which is σ^2 (say).

- This is the assumption of homoscedasticity or equal variance or equal spread.

- In multiple regn model,

$$Y = X\beta + \varepsilon$$

it is assumed that

$$V(\varepsilon) = \sigma^2 I = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & & & \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix}$$

i.e. $V(\varepsilon_i) = \sigma^2$
$Cov(\varepsilon_i \varepsilon_j) = 0 \quad \forall\ i \neq j = 1, 2, \ldots, n$ } (Scalar notations)

- In this case, the diagonal elements of the covariance matrix of ε are the same indicating that the variance of each ε_i is same and off-diagonal elements of the covariance matrix ε are zero indicating that all disturbances are pairwise uncorrelated.

- In many situations, this assumption may not be plausible, and the variances may not remain the same. The disturbances (errors) whose variances are not same / constant across the obsns are called heteroscedastic

disturbance, and this property is known as heteroscedasticity.

In this case,

$V(\varepsilon_i) = \sigma_i^2, \; i = 1, 2, \ldots, n$

and disturbances are pairwise uncorrelated.

So The covariance matrix of disturbances is

$$V(\varepsilon) = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_1^2 & 0 \cdots & 0 \\ \vdots & & & \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix}$$

★ Consequences of using OLS estimators in presence of heteroscedasticity :-

- In presence of heteroscedasticity OLS estimators are unbiased and consistent, but they are not BLUE.
In other words, OLS estimators are not more efficient.

- t & F tests are based on these estimators can be highly misleading and we may not get correct results.

★ ~~Detection~~ Determination of Heteroscedasticity :-

1) Spearman's rank correlation test

Let e_i, $i = 1, 2, \ldots, n$ be the residuals of fitted regn eqn of Y on X.

Let R_1 & R_2 be the ranks given either in ascending order or descending order to $|e_i|$ and X_i respectively.

d_i be the differences in rank.

The spearman's rank correlation coeff. is given by,

$$r = R = 1 - \frac{6 \Sigma d^2}{n(n^2-1)}$$

Let ρ be the popln rank correlation betn u_i & X_i.

Then to test $H_0 : \rho = 0$ Vs $H_1 : \rho \neq 0$, the test statistic is,

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

Under H_0,

$t \sim t_{n-2}$

If $|t| > t_{\alpha/2, n-2}$ then H_0 may be rejected.

Accepting H_1,

which means that popn rank correlation coeff betn u_i & X_i is significant.

i.e. Heteroscedasticity exist.

Ex. The following table gives the values of residuals & variable. Test the presence of Hetero. using spearmans rank method.

e_i	x_i	$\lvert e_i \rvert$	R_1 (rank of $\lvert e_i \rvert$)	R_2 (rank of x_i)	$d^2=(R_1-R_2)^2$
1	5	1	5	3	4
-1.75	10	1.75	3	1	4
2.35	4	2.35	2	4	4
1.65	6	1.65	4	2	4
-0.3	3	0.3	6	5	1
-2.95	2	2.95	1	6	25
					$\Sigma d^2 = 42$

$$r = R = 1 - \frac{6\Sigma d^2}{n(n^2-1)}$$

$$r = R = 1 - \frac{6 \times 42}{6(36-1)}$$

$$r = -0.2$$

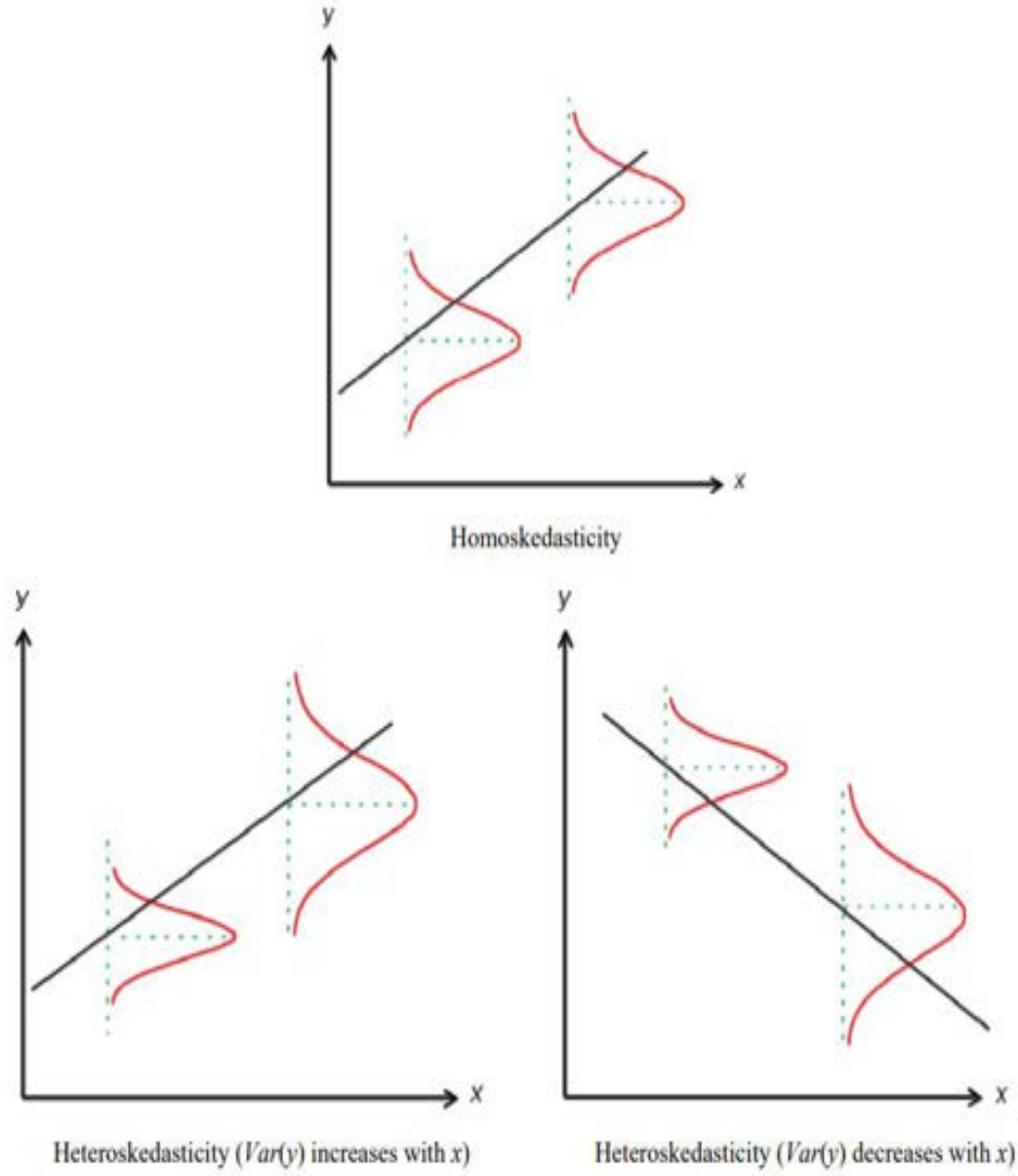
$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{-0.4}{0.9798} = -0.4082$$

$$t_{tab} = t_{\frac{\alpha}{2}, n-2} = t_{0.025, 4} = 2.776$$

$\because |t| < t_{\alpha/2, 4}$

$\therefore$ we accept H_0 i.e. There is no heteroscedasticity.

Graphically, the following pictures depict homoskedasticity and heteroskedasticity.



Examples:

Suppose in a simple linear regression model, x denotes the income and y denotes the expenditure on food. It is observed that as the income increases, the expenditure on food increases because of the choice and varieties in food increase, in general, up to a certain extent. So, the variance of observations on y will not remain constant as income changes. The assumption of homoscedasticity implies that the consumption pattern of food will remain the same irrespective of the income of the person. This may not generally be a correct assumption in real situations. Instead, the consumption pattern changes and hence the variance of y and so

the variances of disturbances will not remain constant. In general, it will be increasing as income increases.

In another example, suppose in a simple linear regression model, x denotes the number of hours of practice for typing and y denotes the number of typing errors per page. It is expected that the number of typing mistakes per page decreases as the person practices more. The homoscedastic disturbances assumption implies that the number of errors per page will remain the same irrespective of the number of hours of typing practice which may not be true in practice.

★ Breusch-Pagan-Godfrey test :-

- This test is used to detect Heteroscedasticity in the linear regn model.

- Consider k explanatory variable linear regn model:

$$Y = X\beta + \varepsilon \quad \text{[in matrix form]}$$

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k + u$$

- Assume that error variance σ_i^2 is a linear function of $Z_1, Z_2, \dots, Z_m$

- Let $\sigma_i^2 = \alpha_0 + \alpha_1 Z_1 + \alpha_2 Z_2 + \dots + \alpha_m Z_m$.

Here, if the model is homoscedastic then $\sigma_i^2 = \alpha_0$
i.e. $\alpha_1 = \alpha_2 = \dots = \alpha_m = 0$

- To check homoscedasticity, we test the hypothesis
$H_0: \alpha_1 = \alpha_2 = \dots = \alpha_m = 0$
i.e. No Heteroscedasticity present.

H_1: at least one $\alpha_i \neq 0$ $(\forall\, i = 1, 2, \dots, n)$

Test procedure :

1) Obtain the OLS estimators of parameters of k explanatory variable linear regn model.

2) Obtain $\hat{y}_i$. Hence find residual $e_i = y_i - \hat{y}_i$

3) Obtain the maximum likelihood estimator of σ^2

$$\hat{\sigma}^2 = \frac{\sum e_i^2}{n}$$

4) Make the transformation.

$$p_i = \frac{e_i^2}{\hat{\sigma}^2}$$

5) Consider linear regⁿ model of p on $Z_1, Z_2, \ldots, Z_m$.

$$p = \alpha_0 + \alpha_1 Z_1 + \alpha_2 Z_2 + \ldots + \alpha_m Z_m + V$$

6) Using regⁿ mode, obtain the estimates of α_i's for the linear regⁿ of p on Z.

7) Obtain $SSR = \hat{\alpha}_i^2 \Sigma z_i^2 = \hat{\alpha}_i^2 \Sigma (Z_i - \bar{Z})^2$

8) Test statistic: $\chi^2 = \frac{ESS}{2} \sim \chi^2_m$ for large sample.

9) If $\chi^2_{cal} > \chi^2_{\alpha, m}$ then reject H_0.

Ex. The following table gives the values of residuals & explanatory variables.

e_i	X_i	$P_i = \frac{e_i^2}{\hat{\sigma}^2}$	$Z = X_i$	$(Z_i - \bar{Z})^2$
−2.1071	44	1.0260	44	
2.9286	35	1.9820	35	
0.9643	30	0.2149	30	
−1.9643	23	0.8916	23	
−1.9286	20	0.8595	20	
2.1071	10	1.0260	10	

Soln: $\hat{\sigma}^2 = \frac{\Sigma e_i^2}{n} = \frac{25.9643}{6} = 4.3274$

Using calculator, $\hat{\alpha}_0 = 0.7390$
$\hat{\alpha}_1 = 0.009662$

$$RSS = \hat{\alpha}_1^2 \Sigma z_i^2 = \hat{\alpha}_1^2 \Sigma (z_i - \bar{z})^2 \qquad (\bar{z} = 27)$$
$$= (0.009662)^2 \times [\Sigma z_i^2 - n\bar{z}^2]$$
$$= (0.009662)^2 \times [5090 - 6 \times (27)^2]$$
$$= 0.06684$$

$$\chi^2_{cal} = \frac{RSS}{2} = \frac{0.06684}{2} = 0.03342$$

$$\chi^2_{tab} = \chi^2_{0.05}(1) = 3.841$$

$$\chi^2_{cal} < \chi^2_{tab}$$

∴ Do not reject H_0 & conclude that Homoscedasticity exist.

☆ Weighted least Square Estimators (WLS) method:

- Since in presence of heteroscedasticity OLS estimators are not more efficient.
i.e. they do not have minimum variance.
∴ We use alternative estimators i.e. WLS estimators.

- In usual OLS method, equal weights are assigned to each obsⁿ.
Since in heteroscedasticity, variance is not constant, the knowledge of unequal variability observed in Y for different values of X, should be used in such a way that obsⁿ with greater variability should be given less weight than those with smaller variability.

- Consider two variable regⁿ model,

$$Y_i = \beta_0 + \beta_1 X_{1i} + u_i \quad \forall\ i = 1, 2, \dots, n$$

It can also be written as

$$Y_i = \beta_0 X_{0i} + \beta_1 X_{1i} + u_i \quad \forall\ i = 1, 2, \dots, n \ \ \& \ X_{0i} = 1$$

Assuming heteroscedastic variances σ_i^2 to be known.
Divide the whole model by σ_i.

$$\therefore \frac{Y_i}{\sigma_i} = \beta_0 \frac{X_{0i}}{\sigma_i} + \frac{\beta_1 X_{1i}}{\sigma_i} + \frac{u_i}{\sigma_i}$$

$$Y_i^* = \beta_0^* X_{0i}^* + \beta_1^* X_{1i}^* + u_i^*$$

where, $Y_i^* = \frac{Y_i}{\sigma_i}$, $X_{0i}^* = \frac{X_{0i}}{\sigma_i}$, $X_{1i}^* = \frac{X_{1i}}{\sigma_i}$, $u_i^* = \frac{u_i}{\sigma_i}$

& β_0^* & β_1^* are parameters of the transformed model.

The variance of the unknown term / error term of the transformed model is,

$$V(u_i^*) = V\left(\frac{u_i}{\sigma_i}\right) = \frac{1}{\sigma_i^2} V(u_i) = \frac{\sigma_i^2}{\sigma_i^2} = 1$$

$\Rightarrow$ Constant

i.e. Variance of the transformed disturbance term u_i^* is homoscedastic.

$\therefore$ Applying OLS procedure to the transformed model is called as weighted least square estimator, which are BLUE.

Multicollinearity:

Multicollinearity -

- The classical linear reg^n model assumes that the explanatory variables are not correlated. i.e. there is no multicollinearity among the explanatory variables that are included in the reg^n model.

- Consider the multiple reg^n model with k-explanatory variables.

$$Y = X\beta + \varepsilon$$

Using OLS, we derived $\hat{\beta} = (X'X)^{-1} X'Y$.

To obtain $\hat{\beta}$, we require the inverse of $(X'X)$ matrix.

If two or more columns of matrix X are perfectly related then there exists a linear relationship between the independent variables and that is called multicollinearity.

- If the multicollinearity is perfect, we do not get the exact estimate of regression coefficient and standard error are infinite.
- If the multicollinearity is less, we can estimate regression coefficients such that in presence of multicollinearity regression coefficients can not be estimated with great precision.

- **We can detect multicollinearity using :**

1) **R^2 and t-ratio :**

- Consider the individual testing of regression coefficient say $H_o : \beta_1 = 0$
  To test it we use t-ratio (t -test).
  i.e. $t = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)}$ and compare it with the table value of t.
  If there exists multicollinearity, then $SE(\hat{\beta}_1)$ increases and thereby making the t-ratio decreases. Therefore, in such cases, one will increasingly accept the null hypothesis that $H_o = \beta_1 = 0$

- If the coefficient of determination = R^2 is high say more than 0.8 then the F test in most of the cases will reject the hypothesis
  $H_o : \beta_1 = \beta_2 = ---- = \beta_k = 0$
  But the individual t tests will show that none or very few regression coefficients are statistically different from zero.

2) **High pair-wise correlations among regressors (Explanatory variables)**

- Another suggested rule of thumb is that if the pair-wise correlation coefficient between two explanatory variables is high, say, in excess of 0.8, then multicollinearity is a serious problem.

- The problem with this criterion is that, although high pair wise correlation may suggest collinearity, it is not necessary that they be high to have collinearity in any specific case.
  That is, high pair-wise correlation is a sufficient condition but not a necessary condition for the existence of multicollinearity because it can exist even though correlations are comparatively low (say, less than 0.5)

3) Variance Inflation Factor (VIF):

- VIF is an indicator of multicollinearity. It is defined as

$$VIF = \frac{1}{(1-r_{ij})^2}$$

where r_{ij} : Coefficient of correlation between i^{th} & j^{th} explanatory variables.

- The larger the value of VIF (exceeding 10) indicates the presence of multicollinearity.

i.e As r_{ij}^2 approaches 1, the VIF approaches ∞. That is, the extent of collinearity increases, the variance of an estimator increases, and in limit, it can become infinity.

As can be readily seen, if there is no collinearity between i^{th} & j^{th} explanatory variables, VIF will be 1.

★ Consequences using OLS estimators in presence of multicollinearity:

1) The OLS estimators have large variances and covariances, making precise estimation difficult.

2) The confidence intervals tend to much wider, leading to acceptance of the "zero null hypo.".

3) The t-ratio of one or more coefficients tends to be statistically insignificant but R^2 the overall measure of goodness of fit, can be very high.

4) The OLS estimators & their standard errors can be sensitive to small changes in data.

★ WHEN MULTICOLLINEARITY IS PRESENT, WE USE RIDGE REGRESSION.

To obtain estimates :

The unweighted least square method minimizes
$\sum u_i^2 = \sum (Y_i - \beta_0 - \beta_1 X_1)^2$

whereas, weighted least square method minimizes the weighted residual sum of squares :

$$\sum \left(\frac{u_i}{\sigma_i}\right)^2 = \sum \left(\frac{Y_i}{\sigma_i} - \beta_0^* \frac{1}{\sigma_i} - \beta_1^* \frac{X_i}{\sigma_i}\right)^2$$

$$S = \sum w_i u_i^2 = \sum w_i (Y_i - \beta_0^* - \beta_1^* X_i)^2$$

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

that is the weights are inversely proportional to the variance of u_i or Y_i conditional upon the given X_i, $var(u_i|X_i) = var(Y_i|X_i) = \sigma_i^2$

$$\frac{\partial S}{\partial \beta_0^*} = 0 \Rightarrow 2 \sum w_i (Y_i - \beta_0^* - \beta_1^* X_i)(-1) = 0$$

$$\sum w_i (Y_i - \beta_0^* - \beta_1^* X_i) = 0$$

$$\sum w_i Y_i - \beta_0^* \sum w_i - \beta_1^* \sum X_i w_i = 0$$

$$\Rightarrow \sum w_i Y_i = \hat{\beta}_0^* \sum w_i + \hat{\beta}_1^* \sum X_i w_i \quad \text{--- (1)}$$

Divide both sides by $\sum w_i$,

$$\frac{\sum w_i Y_i}{\sum w_i} = \hat{\beta}_0^* + \hat{\beta}_1^* \frac{\sum X_i w_i}{\sum w_i}$$

$$\bar{Y}^* = \hat{\beta}_0^* + \hat{\beta}_1^* \bar{X}^*$$

$$\Rightarrow \hat{\beta}_0^* = \bar{Y}^* - \hat{\beta}_1^* \bar{X}^*$$

where, $\bar{Y}^* = \frac{\sum w_i Y_i}{\sum w_i}$, $\bar{X}^* = \frac{\sum w_i X_i}{\sum w_i}$, $w_i = \frac{1}{\sigma_i^2}$

$$\frac{\partial S}{\partial \beta_1^*} = 0 \Rightarrow 2\sum w_i (y_i - \beta_0^* - \beta_1^* x_i)(-x_i) = 0$$

$$\Rightarrow \sum w_i (y_i - \beta_0^* - \beta_1^* x_i)(-x_i) = 0$$

$$\Rightarrow \sum w_i y_i x_i - \beta_0^* \sum w_i x_i - \beta_1^* \sum w_i x_i^2 = 0$$

$$\Rightarrow \sum w_i y_i x_i = \hat{\beta}_0^* \sum w_i x_i + \hat{\beta}_1^* \sum w_i x_i^2 \quad \text{--- (2)}$$

Solve eq^n (1) & (2) Simultaneously,
multiply eq^n (1) by $(\sum w_i x_i)$ & eq^n (2) by $\sum w_i$

$$(\sum w_i)(\sum w_i y_i x_i) = \cancel{\hat{\beta}_0^* (\sum w_i x_i)} + \hat{\beta}_1^* (\sum w_i x_i^2)(\sum w_i)$$

$$(\sum w_i x_i)(\sum w_i y_i) = \cancel{\hat{\beta}_0^* (\sum w_i)(\sum w_i x_i)} + \hat{\beta}_1^* (\sum w_i x_i)^2$$

$$(\sum w_i)(\sum w_i x_i y_i) - (\sum w_i x_i)(\sum w_i y_i) = \hat{\beta}_1^* \left[(\sum w_i x_i^2)(\sum w_i) - (\sum w_i x_i)^2 \right]$$

$$\Rightarrow \hat{\beta}_1^* = \frac{(\sum w_i)(\sum w_i x_i y_i) - (\sum w_i x_i)(\sum w_i y_i)}{(\sum w_i x_i^2)(\sum w_i) - (\sum w_i x_i)^2}$$

Ex. Using following data, estimate the model by WLS method.

X	Y	σ_i^2	$w_i = 1/\sigma_i^2$	$w_i x_i$	$w_i y_i$	$w_i x_i^2$	$w_i x_i y_i$
6	8	1	1	6	8	36	48
12	4	4	0.25	3	1	36	12
16	10	2	0.5	8	5	128	80
18	11	1	1	18	11	324	198
20	6	5	0.2	4	1.2	80	24
			2.95	39	26.2	604	362

$$\hat{\beta}_0^* = \bar{Y}_i^* - \hat{\beta}_1^* \bar{X}_i^*$$

$$\bar{Y}_i^* = \frac{\Sigma w_i y_i}{\Sigma w_i} = \frac{26.2}{2.95} = 8.8814$$

$$\bar{X}_i^* = \frac{\Sigma w_i x_i}{\Sigma w_i} = \frac{39}{2.95} = 13.2203$$

$$\hat{\beta}_1^* = 0.1768$$

$$\hat{\beta}_0^* = 6.5441$$

$\therefore$ model is,

$$\frac{Y_i}{\sigma_i} = \frac{6.5441}{\sigma_i} + 0.1768 \frac{X_i}{\sigma_i}$$

❖ Dealing with Missing Values:

Statisticians are often asked to apply their skills and knowledge into solving real-life problems. Most of the data sets in real life are **incomplete**.

The missing values are common in Real-world data set and it has a profound effect on the final analysis result which may make the conclusion unreliable.

Missing responses represent values of a variable that are unknown, either because respondents provided ambiguous answers or their answers were not properly recorded. The former cause is also known as item nonresponse that occurs because the respondent refuses, or is unable, to answer specific questions or items because of the content, form, or the effort required.

There are different types of missing values.

If the data is missing at random or missing is related to the predictor but the predictor has no relation with the outcome, then the data sample can still represent the population.

However, if data is missing in a pattern that is related to the response it can lead to a significant bias in the model, making the analysis result unreliable.

Many techniques have been proposed to deal with missing values.
They are:

The first and simplest one is to remove the tuple directly.
If a missing value is distributed at random or missingness is related to a predictor and the data set is large enough then the removal of missing data has little effect on the performance of the analysis.
However, if one of the above conditions is not satisfied then simply removing the tuple with missing value is in appropriate.

The second strategy is to fill in or impute the missing data based on the rest of the data.
Generally, there are two approaches, one method simply uses the average of the predictor to fill in the missing values.
Alternatively, we can use learning algorithms such as Bayes's or decision tree to predict the missing value.

❖ Dealing with outliers:

An outlier is defined as an observation point that is distant from the main screen value. The presence of an outlier can break the model analysis ability.

For example, outliers have a strong impact on data scaling and the regression fit. However, it is hard to identify outliers if the data range is not specified.

One of the most efficient ways to identify outliers may be data visualization. The methods can be: 1) Scatter plot 2) Box plot 3) Histogram

By looking at a figure we can point out some suspected observations and check whether these values are scientifically valid.

Outliers should be carefully investigated to see if a reason for their unusual behaviour can be found. Sometimes outliers are "bad" values, occurring as a result of unusual but explainable events. Examples include faulty measurement or analysis, incorrect recording of data, and failure of a measuring instrument. If this is the case, then the outlier should be corrected (if possible) or deleted from the data set. Clearly discarding bad values is desirable because least squares pull the fitted equation toward the outlier as it minimizes the residual sum of squares. However, we emphasize that there should be strong nonstatistical evidence that the outlier is a bad value before it is discarded.

Sometimes we find that the outlier is an unusual but perfectly plausible observation. Deleting these points to "improve the fit of the equation" can be dangerous, as it can give the user a false sense of precision in estimation or prediction. Occasionally we find that the outlier is more important than the rest of the data because it may control many key model properties. Outliers may also point out inadequacies in the model, such as failure to fit the data well in a certain region of x space. If the outlier is a point of particularly desirable response (e.g., low cost, high yield), knowledge of the regressor values when that response was observed may be extremely valuable. Identification and follow-up analyses of outliers often result in process improvement or new knowledge concerning factors whose effect on the response was previously unknown.

Removal of outliers can only be taken when there is a truly valid reason. Instead of removing outliers an alternative way is to transform the data to minimize the effect caused by outliers.

Data transformation:

a. Centering and Scaling:

- Data processing involves transforming data into a stable form for analysis. The most basic and straight forward data transformation techniques are data centering and scaling.
- Many data mining techniques such as Euclidean distance require data to be centered and scaled before entering the model.

- Data centering subtracted the mean value of the predictors from the data and to scale the data each value of the predictor is divided by its standard deviation. Thus, scaled data has unit standard deviation and the centered data has zero mean.

b. Transformation to resolved skewness:

- Data can be positively or negatively skewed while many statistical models make the assumption that the data is normally distributed.
- Applying transformation to resolve skewness helps to improve the data normality. There are quite a lot of transformation methodology that helps to fix the skewness such as replacing the data by log, square root, inverse or box cox transformations, etc.

Selection of Variables in Multiple linear regression:

- You have learnt that the regressor variables included in the multiple regression model are considered to be important since the variability in the response variable/dependent variable is accounted for only by them. So far, we have focused on the techniques to ensure that the functional form of the fitted model is correct and the underlying assumptions are not violated.

- Along with the above considerations and assumptions we have set the following main objectives to obtain an adequate multiple regression model:
  i) to cover as much information as possible in the model by including as many regressor variables as possible, and
  ii) to increase the variability accounted for by the regressor variables by including only a few important regressor variables, so that fitted model is simple and adequate.

- These two objectives contradict each other since adding a large number of regressor variables increases the accounted variability, and selecting only a few of them may reduce the accounted variability and may cause loss of useful information.

- The problem of building an adequate model by striking a balance between both these objectives is not trivial. In multiple linear regression, we determine a group of regressor variables, which are proved to be important significantly, to find an adequate model. In some cases, theoretical considerations and personal experience may be helpful in selecting the important regressor variables to be included in the model. However, in most cases, the appropriate subset of the important variables to be included in the model needs to be examined. The methods of finding an appropriate subset of regressor variables, which adequately explain the variability in the response variable and need to be included in the model are called variable selection methods.

- When a number of independent variables X_1, X_2, …, X_p are available, we may be interested in retaining only a few of them in the model, without sacrificing much of the explanatory power of the model. In this section, we describe some methods that help us in selecting important variables, which adequately explain the variability in the dependent variable Y. The main criterion is that the number of parameters should be as small as possible and all the variables which contribute significantly should be included in the model. When a variable X_i is included in the model, then any other variable, say, X_k, which is highly correlated with X_i, need not be included as its contribution has already been accounted for by X_i. In the same way if X_k is included in the model first, then X_i may not be included as its contribution has already been taken into account by X_k. Thus, the order in which variables enter in the model is important.

Computational Techniques for Variable Selection:

In order to select a subset model several techniques based on computational procedures and algorithm are available. These essentially based on two ideas: select all possible explanatory variables or select the explanatory variables stepwise.

1) Use all possible explanatory variables (all possible regression):

This methodology is based on the following steps:
Fit a model with one explanatory variable,
Fit a model with two explanatory variables,
Fit a model with three explanatory variables and so on,
Choose the suitable criteria for model selection such as R^2 or Adjusted R^2 and evaluate each of the fitted regression equation with selection criteria.
The total number of models to be filled rises with an increase in k. So, such models can be evaluated using model selection criteria with the help of an efficient computation algorithm on computers.

2) Stepwise selection techniques:

This method is based on choosing the explanatory variables in the subset model which can be either adding one variable at a time or deleting one variable at a time.
We now describe three methods for selection of variables and model building: The forward selection method, the backward elimination method and the stepwise selection method.

1) Forward Selection Method:
In this method, the most important variable is included first. The most important variable is the one which accounts for a maximum sum of squares (SS). After including this variable, the next most important variable is chosen which accounts for a maximum additional sum of squares. The contribution is tested at each stage and the most important variable, which accounts for a maximum additional sum of squares is included in the model. When the maximum contribution is not significant, the process is stopped. Otherwise, the process continues by including the variables with significant contributions.
In the forward selection method, variables are added to the model one at a time until the addition of a variable does not significantly improve the model. Thus, variables are added one at a time until further addition of a variable results in the acceptance of the reduced model.

2) Backward Elimination Method:
In this method, all p variables are included in the model. Then the least important variable is deleted if its contribution to the sum of squares is not significant. At the second stage, the next least important variable whose contribution is not significant is deleted. This process goes on till a variable cannot be deleted from the model.
In backward elimination method, we begin with the model that includes all potential regressor or independent variables. Variables are deleted from the model one at a time until further deletion of a variable results in the rejection of the reduced model.

It is considered to be an economical procedure as it attempts to examine only the best regression equation containing a certain number of variables.

3) Stepwise Selection Method:

Neither forward selection nor backward elimination takes into account the effect that the addition or deletion of a variable can have on the contribution of other variables to the model. A variable added early to the model in forward selection method can become unimportant after other variables are added, or variables previously dropped in backward elimination can become important after other variables are dropped from the model. Stepwise regression is a forward selection method that rechecks at each step the importance of all previously included variables. If the partial s.s. for any previously included variables do not meet a minimum criterion to stay in the model, the selection procedure changes to backward elimination and variables are dropped one at a time until all remaining variables meet the minimum criterion. Then forward selection resumes.

The stepwise selection method begins with selecting an equation containing the single best independent variable X. Then, the second most important variable is included in the model. But at this stage, each of the variables selected earlier in the model are tested for their significance. If they are found to be significant, then we try to build an equation with the subsequent inclusion of other regressor variables one at a time. These additional variables are included as long as their inclusion is significant. A variable is excluded from the model if found insignificant. The selection process continues till no additional variable can be included or deleted in the model.

The order of addition is determined by using a test of significance based on partial F-test values. In order to decide the next variable to be included, the highest partial F-value is compared with the selected F-tabulated value (for the corresponding variable to be included). After a variable has been included, the fitted equation is examined to check whether some variable(s) can be deleted.

Stepwise selection method is a modified forward selection method. In the forward selection method, once a variable enters the model, it stays. Unfortunately, it is possible for a variable entering at a later stage to render a previously selected variable unimportant when these variables are themselves closely related. The forward selection method does not consider this possibility. In stepwise selection method, each time a new variable is entered into the model, all previous variables in the model are checked for continued importance.

★ PARTIAL F- STATISTIC :-

- Suppose that we have already p variables in the model $x_1, x_2, \ldots, x_p$ and we are interested in whether we should add one extra variable x^* or not. We can do it by calculating extra sum of square from adding x^* to the model given that $x_1, x_2, \ldots, x_p$ are already in the model.

$H_0: \beta^* = 0$ v/s $H_1: \beta^* \neq 0$

$$SS(x^* | x_1, x_2, \ldots, x_p) = \underbrace{SS(x^*, x_1, x_2, \ldots, x_p)}_{\text{(full model)}} - \underbrace{SS(x_1, x_2, \ldots, x_p)}_{\text{(full model except } x^*)}$$

$$\text{Partial F-statistic} = \frac{SS(x^* | x_1, x_2, \ldots, x_p) / \ell}{MSSE(x^*, x_1, \ldots, x_p)}$$

$$\sim F_\alpha(\ell, n-k-1)$$

Reject H_0 if $F > F_\alpha(\ell, n-k-1)$

1) Forward selection method :-

Algorithm for forward selection :-

step 1 :- Consider only intercept term and insert one variable at a time.

step 2 :- Calculate the simple correlations of x_i's $(i=1,2,\dots,k)$ with y.

step 3 :- Choose x_i which has the largest correlation with y.

step 4 :- Suppose X_1 is the variable which has the highest correlation with Y.

$F(X_1|__)$ = Partial F-value for the model in presence of no regressor variable.

$$F_0 = \frac{MSSR}{MSSE} = \frac{R^2 S_{yy}/k}{(S_{yy} - R^2 S_{yy})/n-k-1}$$

$$= \frac{n-k-1}{k} \frac{R^2}{1-R^2}$$

$$= (n-2) \frac{R^2}{1-R^2}$$

If $F_0 > F_{tab} = F_\alpha(1, n-2)$

i.e. It will be significant then include X_1 into the model.

Now, model will become $Y = \beta_0 + \beta_1 X_1 + \varepsilon$

Step-5:- Partial F-statistic is computed for all of the remaining regressors in the presence of previously selected regressor, the one yielding the highest F is added to the model if $F_{cal} > F_{tab}$.

i.e. $\left.\begin{array}{l} F(X_2|X_1) \\ F(X_3|X_1) \\ \vdots \\ F(X_K|X_1) \end{array}\right\} \rightarrow$ Choose largest partial F-statistic.

Assume that $F(X_2|X_1)$ is largest one.

- If it is insignificant then stop the process and hence our model will be $Y = \beta_0 + \beta_1 X_1 + \varepsilon$

- If it is significant then include X_2 into the model.

So now, our model becomes

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon.$$

Step-6:- Now obtain

$\left.\begin{array}{l} F(X_3|X_1X_2) \\ F(X_4|X_1X_2) \\ \vdots \\ F(X_K|X_1X_2) \end{array}\right\} \rightarrow$ Choose largest partial F-statistic

Assume that $F(X_3|X_1X_2)$ is largest one.

- If it is insignificant then stop the process and hence our model will be $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$.

- If it is significant then include X_3 into the model.

2) Backward Elimination Method :-

- This methodology is contrary to the forward selection procedure. The forward selection procedure starts with no explanatory variable in the model and keeps on adding one variable at a time until a suitable model is obtained.

- The backward elimination methodology begins with all explanatory variables and keeps on deleting one variable at a time until a suitable model is obtained.

steps are :

1) Consider all k explanatory variables and fit the model.

2) Compute partial F-statistic for each explanatory variable.
Select the variable with smallest partial F-statistic denote it by Fmin.

3) Test for significance of Fmin.
If Fmin is not significant remove the variable associated with Fmin from the model and return to step-2.
If Fmin is significant then stop and report the current model.

3) Stepwise regression Procedure :-

- A combination of forward selection & backward elimination procedure is the stepwise regression.

- Suppose we have X_1, X_2, X_3, X_4 variables.
- We check correlations of X_i's with Y.
  $Corr(X_1, Y)$, $Corr(X_2, Y)$, $Corr(X_3, Y)$, $Corr(X_4, Y)$.
- Suppose X_1 is highly correlated and it is significant then model becomes
  $$Y = \beta_0 + \beta_1 X_1 + \varepsilon.$$

- Partial F-statistic is obtained for all of the remaining regressors in the ~~model~~ presence of previously selected regressor

$$\left.\begin{array}{l} F(X_2|X_1) \\ F(X_3|X_1) \\ F(X_4|X_1) \end{array}\right\} \text{Choose largest partial F-statistic.}$$

Suppose $F(X_2|X_1)$ is largest partial F-statistic. and X_2 is significant then we can write
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon.$$

- Now we use backward elimination method.

$$\left.\begin{array}{l} F(X_1|X_2) \\ F(X_2|X_1) \end{array}\right\} F_{min}$$

Suppose $F(X_1|X_2)$ is smallest & significant then model is $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$.
and if F_{min} is insignificant remove one variable.

- It is a modification of forward selection procedure.
- The explanatory variable that was added at an earlier step may now become insignificant due to its relationship with currently present explanatory variables in the model.

- Then stepwise procedure checks on this possibility by performing at each step partial F-statistic, using partial s.s. for each variable currently in the model.
- If there is variable in the model that is no longer significant variable is removed from the model.
- The procedure terminates when no further variable can enter or remove.

- Then again forward method for X_3 & X_4,

$F(X_3 | X_1 X_2)$ } choose largest F-statistic.
$F(X_4 | X_1 X_2)$

$F(X_3 | X_1 X_2)$ is largest F-statistic & significant then model becomes

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$

Again we check backward procedure,

$F(X_1 | X_2 X_3)$ }
$F(X_2 | X_1 X_3)$ } choose Fmin
$F(X_3 | X_1 X_2)$ }

$F(X_1 | X_2 X_3)$ has Fmin & is insignificant then remove variable.
& if it is significant stop process & model will be

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$

Same procedure for X_4, we can apply.