(Mid term) Sta 404 (Regression & Correlation Anyalsis) (Important Notes)

Lecture #01.

* Things to remember:-
* Regression Analysis is a statistical tool for investigation then relationship between a dependent variable & one or more independent variable.
* MCQs;

Regression analysis is widely used for prediction & forecasting.

* Applications:-
  * Economics
  * Engineering
  * Medical fields
  * Management
  * Social Sciences
  * Spoots & physical Sciences
  * Computer Sciences.

Relationship between variables.

* Exact relationship	Non-exact relationship
* Fahrehit ~ celsius	weight ~ Height.
$F = 32 + \frac{9}{5}C$	
area = πr^2.	
* Deterministic Model	Probabilistic Model.
* $Y = a + bX$	$Y = a + bX + e$ → error.

Dependent	Vs	Independent variables.
The dependent variable is the variable in regression that can Not be controlled or Manipulated.		The Independent variables is the variable in regression that can be controlled or Manipulated.

Scatter plots and Relationship.

1) Scatter plot is used to show the relationship between two variables.

2) The Independent and dependent variables can plotted on a graph called a Scatter plot.

3) The Independent variables 'x' is plotted on the horizontal axis. and dependent variables on the vertical axis.

Lecture #02.

Regression line.

$$\hat{y} = a + bx$$ ⇒ equation of regression line.

for finding 'a'

$$a = \bar{y} - b(\bar{x})$$; $x = ?$, $y = ?$

for finding 'b'

$$b = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{n(\Sigma x^2) - (\Sigma x)^2}$$

$n = ?$, $\Sigma xy = ?$, $\Sigma x = ?$ $\Sigma y = ?$, $\Sigma x^2 = ?$

Lecture #03

Important points:-

(1) Simple linear regression is really a Comparison of two Method;

* where the independ variable does not even exist.

* The other uses the best fit regression line.

(2) If there is only one variable the best prediction for other values is the Mean of the "dependent" variable.

③ The difference between the best fit-line and the observed value is called the **residual**. (or error)

④ The residuals are squared and then added together to generate sum of Square residuls/error. SSE.

⑤ Simple linear regression is designed to find the best fiting line through the data that minimize the SSE.

⑥ By Determining the coefficient of determination we can find the estimated regression equation fit our data.

* Formula for coefficient of Determination $= R^2 = \frac{SSR}{SST}$ (COD)

Residual.

* $SST = SSR + SSE$

↓ Sum of Sq. of Total

→ Sum of Sq. Error

"The Simple linear regression Model"

Steps involves;

1. The Mean of the response is a linear function.
2. The errors are Independent.
3. The errors are normally distributed.
4. The errors have equal variances for all x values.

Properties of the least square Regression line..

1. The least square regression line always passes through the Centre of the data.
2. The Sum of deviations of the observed values of y from the LS regression line is always equal to zero. $\Sigma e_i = (y_i - \hat{y}_i) = 0$
3. The Sum of the Squares deviations of the observed values of y from the LS regression line is Minimum.

$$\Sigma e_i^2 = \Sigma (x_i - \hat{y}_i)^2 \Rightarrow \text{Minimum.}$$

④ The line fitted by L.S. Method is "best" fit line.

Lecture # 04.

★ Correlation :-

The word Correlation is made of Co- (together) and Relation.

"Two variables are said to be correlated if they tend to simultaneously vary together in some direction."

★ Types of Correlation :-

★ linear	★ Non-linear.
↓	↓
When graph is plotted like straight line their is linear Co-relation	When graph is Not Straight line.

Correlation Might be

★ Simple

Involve two and ~~or~~ one dependent variable.

★ Multiple :-

one dependent and More

than one independent variables.

* Partial:-

one dependent variable and More than one independent variable but only one independent variable is Considered and other Independent variable are Considered Constant.

Correlation Might be:-

* Positive:- A positive correlation is a relationship between two variables where if one variable increases, the other one also increase. ↑↑

* Negative:- A Negative correlation Means that there is an inverse relationship between two variables. When one variable decreases, the other increases. ↑↓

* No Correlation:-

Their Movement is independent from each other.

zero ↑ —→

Methods of Studying Correlation.

* Scatter plot Method.
* Karl Pearson's Coefficient.
* Spearman's Rank Correlation Coefficient.

Correlation Coefficient - Important points.

(1) The range of the Correlation Coefficient is from -1 to +1.

(2) If there is a strong positive linear relationship between the variable, the value of 'r' will be close to 1.

(3) If there is a strong Negative linear relationship between the variables, the value of 'r' will be close to -1.

(4) When there is no linear relationship between the variables, the value of r will be close to 0.

★ How to Calculate Correlation Coefficient - r.

★ Formula :-

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n(\Sigma x^2) - (\Sigma x)^2][n(\Sigma y^2) - (\Sigma y)^2]}}$$

★ There are no units associated with r.

★ the value of r will remain unchanged. If the x and y values are switched.

★ "Correlation is Not Causation"..... which says that a correlation does not mean that one thing causes the other. (There could be other reasons the data has a good correlation).

Lecture #05.

Rank Correlation.

Developed By:-

Rank Correlation Method was developed by the British Psycologist Charles Edward Spearman in 1904.

Rank Correlation Coefficient formula.

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

where;

r_s = Coefficient of rank correlation.

d = difference.

n = no. of data.

It deals with Measuring Correspondence between two rankings and assessing the significance of this Correspondence.

Features.

* The range of rank Correlation is -1 to +1.

The rank Correlation Can be Interpreted in the Same way as Karl Pearson's Correlation Coefficient.

Rank Correlation Coefficient is a distribution-Free Measure Since no Strict assumption is Made about the population from which it is drawn.

Spearman's formula is the only formula available to find the Correlation between rank.

Types of Rank Methods.

Ranked data :- where ranks are given.

* Not Ranked data :- where ranks are Not given.

Ranks are tied:- where repeated ranks occurs.

Lecture # 06.

Multiple Linear Regression Analysis.

Difference Between Simple and Multiple Regression.

Simple Linear Regression	Multiple linear Regression.
In Simple linear regression, the regression equation Contains one Independent Variable x and one dependent Variable y and it is written as; $\hat{Y} = a + bX$	In Multiple linear regression equation with two Independent variables (X_1 and X_2) and one dependent variables has the form; $\hat{Y} = a + b_1 X_1 + b_2 X_2$

The general form of the Multiple regression equation with K - Independent Variable is:-

$$\hat{y} = a + b_1 X_1 + b_2 X_2 + \cdots + b_K X_K$$

* Y is Called Dependent variable.
* The X's are the Independent variables.
* The b's are Called partial regression coefficients.
* 'k' is number of Independent variables.

In Mutiple Regression;
the variables X_1 & X_2 are Not correlated with each other.

Multiple Regression Analysis.

Normal Equation:-

$$\hat{Y} = a + b_1 X_1 + b_2 X_2$$

* $\sum Y = na + b_1 \sum X_1 + b_2 \sum X_2$ ---- (1)
* $\sum X_1 Y = a \sum X_1 + b_1 \sum X_1^2 + b_2 \sum X_1 X_2$ ---- (2)
* $\sum X_2 Y = a \sum X_2 + b_1 \sum X_1 X_2 + b_2 \sum X_2^2$ ---- (3)

Method to Solve Problems By using Normal Equations:-

1. Take Multiple Model
2. Make the 3 Normal equations.
3. Calculate the required values in the table.
4. Reduce the 3 Normal equations.
5. Reduce the 3 normal equation to the 2 equations
6. Solve the New two equations to get unknown Constants, step by step.
7. Put the Constants in the Model.

* Method to find Coefficient values by using FORMula.

$$\hat{Y} = a + b_1 X_1 + b_2 X_2.$$

X_1 → Area

X_2 → No. of data

* for Calculating value of 'a'

$$a = \bar{Y} - b_1 \bar{X}_1 - b_2 \bar{X}_2$$

* For Calculating value of b_1:-

$$b_1 = \frac{(\Sigma x_1 y)(\Sigma x_2^2) - (\Sigma x_2 y)(\Sigma x_1 x_2)}{(\Sigma x_1^2)(\Sigma x_2^2) - (\Sigma x_1 x_2)^2}$$

* For Calculating value of b_2;

$$b_2 = \frac{(\Sigma x_2 y)(\Sigma x_1^2) - (\Sigma x_1 y)(\Sigma x_1 x_2)}{(\Sigma x_1^2)(\Sigma x_2^2) - (\Sigma x_1 x_2)^2}$$

For formula Method first you have to Convert data into deviation form.

orginal data			Data in Deviation form.		
Y	X_1	X_2	$y = Y - \bar{Y}$	$x_1 = X_1 - \bar{X}_1$	$x_2 = X_2 - \bar{X}_2$
30	10	15	12.2	4	4.6
22	5	8	4.2	-1	-2.4
16	10	12	-1.8	4	1.6
[illegible]	3	7	-10.8	-3	-3.4
7	2	10	-3.8	-4	-0.4
14					
T [illegible]	30	52	0	0	0
M 17.8	6	10.4			