Regression Analysis:

Chapter-11

1 Regression Analysis is a statistical tool for investing the relationship b/w a dependent variable & one or more independent variables.

Note: Regression Analysis is widely used for prediction and forecasting.

Applications: Economics,

Ex: Suppose you are marketing Analyst for Toys we gather the following data.

($) Advertising sals (X)	Sales (Y)
1	1
2	1
3	2
34	2
5	4

regresser variable

Reponse variable.

Now we want to relationship b/w sales & adverting cost. Since we can decide How much money we want to spend (ie. sort of control variable) for advertising. but sales amount is not a control variable ie. we can not control the sales amount. So sales amount is dependent variable it depends on advertisement. So sales amount is dependent variable and advertising cost is independent variable.

Scanned by CamScanner

Scatter Plot

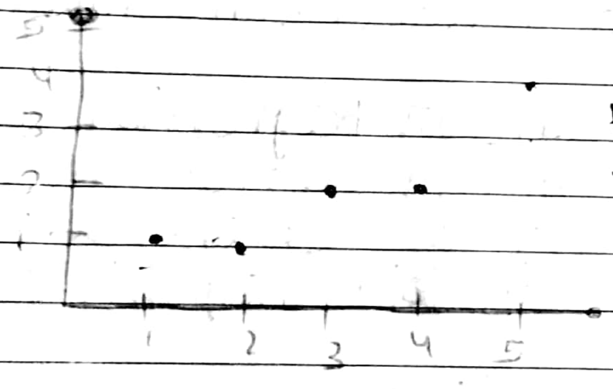
(2) Let X_i stands for regresser variable,
Y_i " " Response variable.

X_i	Y_i
1	1
2	1
3	2
4	2
5	4

(~~here~~ Let we have 5 observation
ie. $(x_1, y_1) = (1,1)$
$(x_2, y_2) = (2,1)$
$(x_3, y_3) = (3,2)$
$(x_4, y_4) = (4,2)$
$(x_5, y_5) = (5,4)$)

{ A scatter plot is a mathematical diagram to display values or two ~~vals~~ variables for a set of data }

{ A Scatter plot are used to investigate the possible relationship b/w the variables. }.

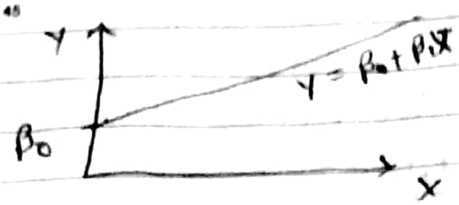


this scatter plot indicate linear relationship b/w regresser variable or response variable.

If Scatter plot indicates linear relationship b/w the variables then we go for linear modal.
If Scatter ~~pol~~ plot indicates non-linear relationship b/w x and y then we go for quadratic or cubic or higher order polynomial model.

Scanned by CamScanner

3



Linear relationship b/w x & y.

β_1 – slop

β_0 – y-interrept.

Note:- ~~for~~ a given x does not always give the same value for y.

* a random sample of size n by the set $\{(x_i, y_i) | i = 1, .. n\}$ If additional samples were taken using exactly the same values of x, we should expect the y values to vary. Hence, the value y_i in the ordered pair (x_i, y_i) is a value of some random variable Y_i.

The Simple Linear Regression Model:- (SLR):-

Simple linear regression model is a model with a single regresser x that has a linear relationship with response variable y.

The simple linear regression model is

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

where ε → is random error component

β_0 → y-intercept

β_1 → slop

X → Regresser Variable.

Y → Response variable.

NOTES: Where β_0 & β_1 are the parameter and they are typically unknown. ~~to be est~~ are called regression cofficient.

2011

5	5	Wk
5	6	45
12	13	46
19	20	47
26	27	48
		49

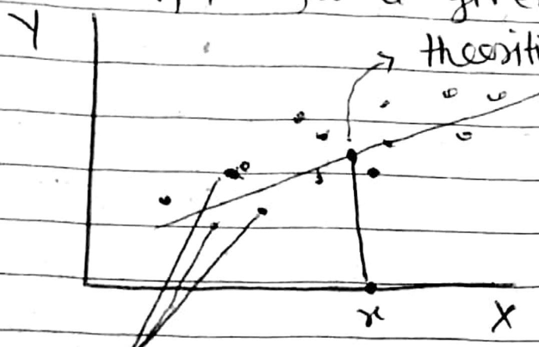
Scanned by CamScanner

* We will assume a linear relationship b/w Y and X

(4) $E[Y/X] = \mu_{Y/x} = \beta_0 + \beta_1 x$

↳ true mean of Y for a given value of x.

i.e. Expectation of Y for a given value of Y

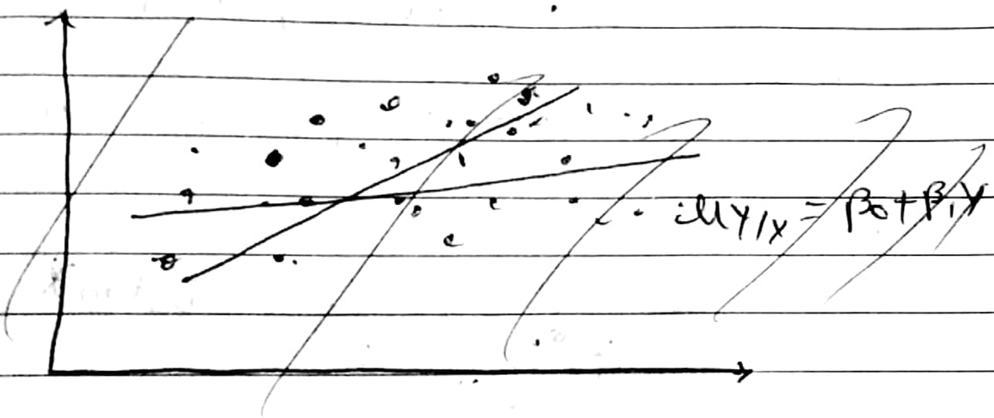


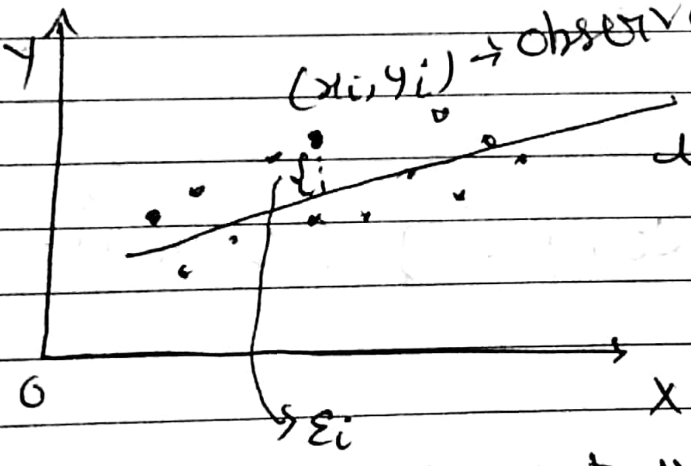
theoritical, ~~mean~~ Expected value of Y for given x.

$\mu_{Y/x} = \beta_0 + \beta_1 X$

true relationship b/w x & Y. (i.e. true regression line)

The observed value of Y vary about the this line





$\mu_{Y/x} = \beta_0 + \beta_1 X.$

here ε_i ~~at 1~~ represents that Y values will vary about the ~~line~~ line and value of Y do not fall precisely on the line $\mu_{Y/x} = \beta_0 + \beta_1 X$.

DECEMBER							2011
WK	M	T	W	T	F	S	S
49				1	2	3	4
50	5	6	7	8	9	10	11
51	12	13	14	15	16	17	18
52	19	20	21	22	23	24	25
53	26	27	28	29	30	31	

Scanned by CamScanner

* We will use sample data to obtain the estimated regression line(fitted Regression line).

$\hat{y} = b_0 + b_1 X$

√ ~~&~~ Where b_0 is estimation of β_0 (ie β_0 estimates b_0)

b_1 is estimation of β_1 (ie β_1 estimates b_1).

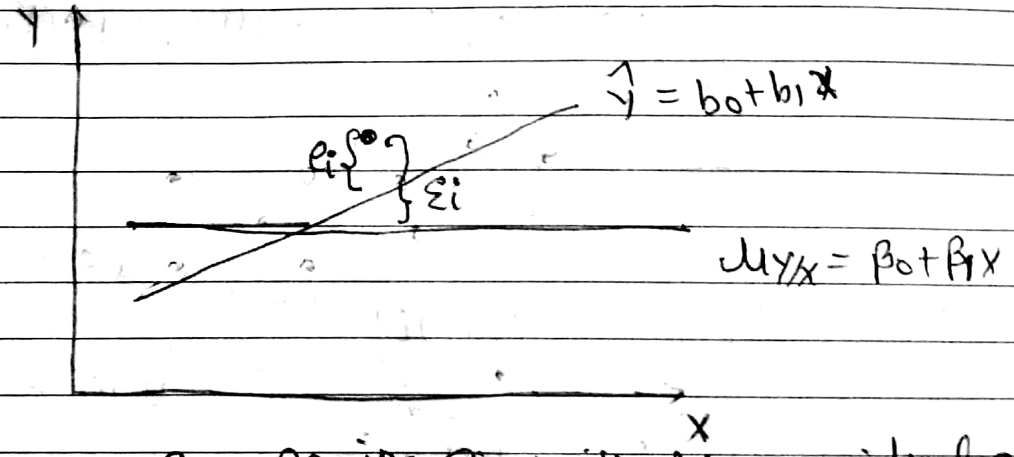
where $\hat{y}$ represents predicted value of Y for a given ~~x~~ value of X.

b_0, b_1 are the sample statistics that estimates the parameter β_0 & β_1 respectively

Residuals:- (Error in a fit):-

Given a set of regression data $\{(x_i, y_i); i=1,\dots n\}$ and fitted model, $\hat{y}_i = b_0 + b_1 x_i$, then the i^{th} residual e_i is given by

$$e_i = y_i - \hat{y}_i \quad , i = 1,2,\dots,n$$



Comparing ε_i with the residual e_i:-

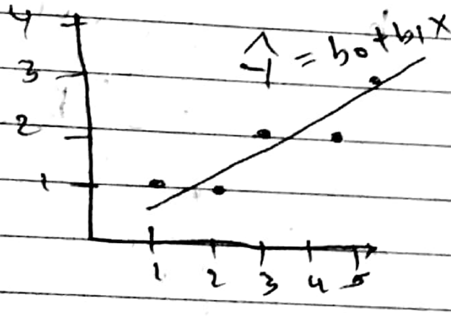
Scanned by CamScanner

⑥ Gauss Markov assumption:-

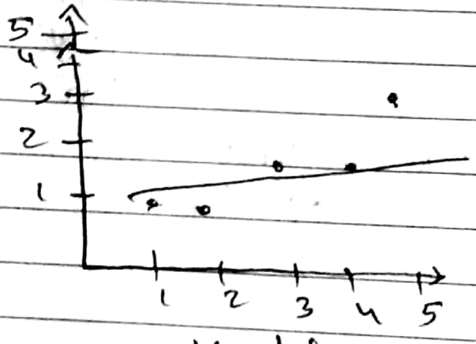
1. ε_i is a random variable with zero mean & variance σ^2 (which is unknown)
   i.e. $E[e_i] = 0$ & $Var[\varepsilon_i] = \sigma^2$
2. ε_i & ε_j are uncorelated. $i \neq j$, so covariance b/w ε_i & ε_j is 0. i.e. $cov(\varepsilon_i, \varepsilon_j) = 0$
3. ε_i is a normally distributed random variable with mean zero and variance σ^2.
   i.e. $\varepsilon_i \sim N(0, \sigma^2)$.

* Since β_0 & β_1 are the unknown parameters so we want to estimates β_0, β_1. for this

Least squares estimations of the parameters:-
The parameter β_0 & β_1 are unknown and must be ~~establis~~ estimated using the data $(x_1, y_1), \dots, (x_n, y_n)$



model 1.



Model

L.S.E it fits the model s.t $\sum_{i=1}^{5} e_i^2$ is minimum.
ie. sum of squares of the residuals is a minimum.

Note ie. residuals sum of squares is often called the sum of squares of errors about the regression line and is denoted by S.S.E.

ie. $SSE = \sum_{i=1}^{n} e_i^2$

NOTES :

DECEMBER 2011

Wk	M	T	W	T	F	S	S
49				1	2	3	4
50	5	6	7	8	9	10	11
51	12	13	14	15	16	17	18
52	19	20	21	22	23	24	25
53	26	27	28	29	30	31	

Scanned by CamScanner

$$SSE = \sum_{i=1}^{n} (e_i^2) = \sum_{i=1}^{n} (y_i - \hat{y}_i) = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$

X

These two equation are called normal curve

Differentiate SSE w.r.t. b_0 & b_1

we have

$$\frac{\partial(SSE)}{\partial b_0} = -2 \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i) = 0$$

$$\frac{\partial(SSE)}{\partial b_1} = -2 \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)\cdot x_i = 0$$

i.e. b_0 & b_1 are the sol. of the normal curve

$$\sum_{i=1}^{n} (y_i - b_0 - b_1 x_i) = 0 \quad -①$$

$$\sum_{i=1}^{n} (y_i - b_0 - b_1 x_i) x_i = 0 \quad -②$$

$$\frac{\sum_{i=1}^{n} y_i}{n} - b_0 \frac{\sum_{i=1}^{n} 1}{n} - b_1 \frac{\sum_{i=1}^{n} x_i}{n} = 0$$

$$\bar{y} - b_0 \frac{n}{n} - b_1 \bar{x} = 0$$

where $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$

$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$

$$\bar{y} - b_0 - b_1 \bar{x} = 0$$

$$\boxed{b_0 = \bar{y} - b_1 \bar{x}}$$

put the value of b_0 in ②

$$\sum_{i=1}^{n} (y_i - \bar{y} + b_1 \bar{x} - b_1 x_i) x_i = 0$$

$$\sum_{i=1}^{n} (y_i - \bar{y}) x_i = \sum_{i=1}^{n} b_1 (x_i - \bar{x}) x_i$$

$$b_1 = \frac{\sum_{i=1}^{n} (y_i - \bar{y}) x_i}{\sum_{i=1}^{n} (x_i - \bar{x}) x_i}$$

$$\boxed{b_1 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}$$

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

$$= \bar{x} \cdot \cancel{n\bar{y}} - \bar{x}\bar{y} \cdot \cancel{n}$$

$$= 0$$

VEMBER 2011

T	W	T	F	S	S	Wk
1	2	3	4	5	6	45
8	9	10	11	12	13	46
5	16	17	18	19	20	47
2	23	24	25	26	27	48

NOTES :

(8) Estimating the Regression Cofficients

Given the sample $\{(x_i, y_i) ; i = 1, 2, \dots, n\}$ the least squares estimates b_0 and b_1 of the regression cofficient β_0 & β_1 are computed from the formulas.

$$\boxed{b_1 = \frac{n \sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i}}$$

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

$$\boxed{b_0 = \frac{\sum_{i=1}^{n} y_i - b_1 \sum_{i=1}^{n} x_i}{n} = \bar{y} - b_1 \bar{x}}$$

$$\rightarrow b_1 = \frac{\sum_{i=1}^{n} x_i y_i - \bar{y} \sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2 - \bar{x} \sum_{i=1}^{n} x_i}$$

$$\boxed{b_1 = \frac{\sum_{i=1}^{n} x_i y_i - n \bar{y} \bar{x}}{\sum_{i=1}^{n} x_i^2 - n \bar{x}^2}}$$

Scanned by CamScanner

Start Lec-8:

Parameter estimation solution Table

x_i	y_i	x_i^2	y_i^2	x_iy_i
1	1	1	1	1
2	1	4	1	2
3	2	9	4	6
4	2	16	4	8
5	4	25	16	20
Sum 15	10	55	26	37

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} = 2 - (0.70)(3) = -0.10$$

$$\bar{Y} = \frac{\sum_{i=1}^{5} y_i}{5} = \frac{10}{5} = 2$$

$$\bar{X} = \frac{\sum_{i=1}^{5} x_i}{5} = \frac{15}{5} = 3$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{X}\bar{Y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = \frac{37 - 5 \times 3 \times 2}{55 - 5 \times 3^2} = 0.70$$

So fitted eqn $\boxed{\hat{Y} = -0.10 + 0.7X}$

Correlation cofficient:- The measure ρ of linear association b/w two variable x & y is estimated by sample correlation cofficient r where

$$r = b_1 \sqrt{\frac{S_{xx}}{S_{yy}}} = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$$

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

$$S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \text{cov}(x,y).$$

$$S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \text{var}(y).$$

Scanned by CamScanner

Q. What is correlation ?

8 A Correlation is the statistical technique Which used for knowing How strong pairs of variables

9 are related.

Ex:- height and weight are related.

10 i.e. taller people is heavier than the shorter people. this relation is not perfect.

11 because people of same height may very. So shorter one may have heavier than the taller.

12:00

Correlation Cofficient:- The main result of the

1 correlation is called the correlation cofficient (or r) It's range is from -1.0 to 1.0. & closer r is to +1 or -1?.

2 In statistics, the correlation cofficient 'r' measures the strength and direction of a linear relation-

3 ship b/w two variables on a scatterplot.

If $r = -1$ means A perfect downhill (negative) linear relationship

4 $r = -0.7$ " A strong " " " "

$r = -0.5$ " " moderate " " " "

$r = -0.3$ " " weak " " " "

5 • $r = 0$ " No relationship.

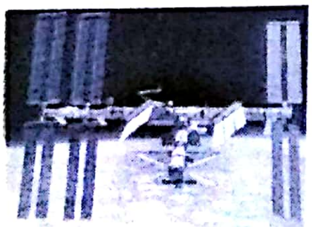
$r = +0.3$ " A weak uphill (positive) linear relationship

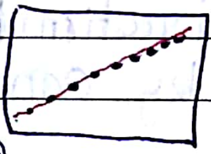
6 $r = +0.5$ " A moderate " " " "

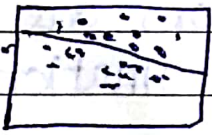
$r = +0.7$ " " strong " " " "

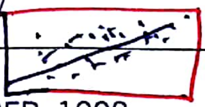
$r =$ Exactly +1 " " perfect " " " "

1 3 5 7 1
w m s p









20th NOVEMBER, 1998
The first module of the International Space Statio