Unit - I

*R. 2 mark

Simple regression model with one independent variable:

$\bar{y}$ - without independent variables

(1) ~~The~~ we consider the simple linear regression model, i.e a model, with single independent (Regressor) variable x, that has a relationship with a dependent (Response) a variable y that is a straight line.

$\sum y - \hat{y} = 0$

The simple linear regression model is

$y = \beta_0 + \beta_1 x + \epsilon \longrightarrow (1)$

where β_0 is the intercept and

β_1 is slope

ϵ is a random error component.

Also β_0 and β_1 are unknown constant.

The errors are assumed to have a mean zero and unknown variance σ^2,

Additionaly we usually assume that the errors are uncorrelated this means that the value of one error does not dependent on the value of one error does not dependent on the value of any other error. It is conviniant to the regressor variable x as the controlled by the data analysis (Researcher) and measured and negligible error. while the response variable y is a random variable. i.e., There is a Probability distribution

unimportant

~~of the response~~

(2) for y at each possible variable for x the mean of the distribution is

$E(y/x) = \beta_0 + \beta_1 x \longrightarrow (2)$

And the variance is

$var(Y/x) = var(\beta_0 + \beta_1 x + \epsilon) = \sigma^2 \longrightarrow (3)$

Thus, the mean of y is a linear function of x although the variance of y does not depend on the value of x. Further more, because the errors are uncorrelated therefore the responses of also uncorrelated.

The parameters β_0 and β_1 are usually called regression co-efficient. these co-efficient have a simple often useful interpretation. The slope β_1 is the change in the mean of the distribution of y Produced by a unit change in x. If the range of the data on x includes $x=0$, then the intercept β_0 is the mean of the distribution of the response y when $x=0$. If the range of x does not include zero then β_0 has no Practical interpretation.

Assumption in simple linear regression model:

(3)

Generation of x_i and y_i:

In most social sciences the dependent and independent variables are observed. They are beyind the experimental control of the researcher. the cannot select and vary systameticaly the values of the independent variable x. Hence for researcher both x and y are random.

However in many economic studies x's are fixed, it ~~many~~ is assumed that x is non-random and the values of x have been choosen systametically.

2) variability in x:

It is assumed that all the values of x the independent variable are not the same atleast one value x must differ from other. If this condition is not statisfied then we shall not able to estimate β_0 and β_1..

$xy+\epsilon$

3) Stochastic assumption of ordinary least squares:

$y = a + bx + \epsilon$

a) Randomness of ϵ:

The variable ϵ is a real random variable whose value depend it may be

Positive, negative or zero each value has a assumed by ϵ has certain Probability has in any particular distance and has normal distribution.

b) The variable ϵ_i has a normal distribution:

The value of ϵ_i are normally distributed about their zero mean. i.e., the value of ϵ_i have a bell-shaped symmetrical about their mean zero. This assumption is used in testing the value of the population parameter from the knowledge of the standard error of the estimated co-efficients. Simbalically $\epsilon_i \sim N(0, \sigma_\epsilon^2)$

c) ϵ_i has a zero mean value for each x_i:

This assumption state that, for every observation. The expected value of the distrubance the term. $E(\epsilon_i) = 0$ for all $i = 1, 2 \ldots n$.

This implies that for each value of x_i, ϵ_i may assume that variance value that, sum greater than zero some smaller than zero and some zeros, but on the average $\epsilon = 0$.

5

d) The variance of each ϵ_i is same for all the x_i values (Homo Scadasdicity) :

The various of ϵ_i about its means is constant at all values of x. i.e., for all values of x the ϵ's will show the same disperson around the mean. The variance in the random distrubance constant over the observations in simbal $V(\epsilon) = E(\epsilon_i^2) = \sigma^2$

E) NO auto correlation :

This assumptions requireds that the various distrubance the term are uncorrelated

i.e., $cov(\epsilon_i, \epsilon_j) = 0$

i.e., $E(\epsilon_i \cdot \epsilon_j)$ for $i \neq j$ and $i, j = 1, 2 \ldots n$

where ϵ_i, ϵ_j are two different observation of the distrubance term cov means the covariance.

F) Every distrubance term ϵ_i is independent of the explanatory variable (s) / or zero co-variance between ϵ_i and x_i :

$$cov(\epsilon_i, x_i) = E[(\epsilon_i - E\epsilon_i)][x_i - E(x_i)]$$

$$= E[\epsilon_i (x_i - E(x_i)] \text{ since } E(\epsilon_i) = 0$$

$$= E(\epsilon_i x_i) - E(x_i) E(u_i) \text{ since}$$

$E(x_i)$ is a constant

$$= E[\epsilon_i x_i) \text{ since } E(\epsilon_i) = 0$$

$= 0$ by assumption.

(6) $E(\epsilon_i x_{1i}) = E(\epsilon_i x_{2i}) = 0$ (in the case of multi linear regreesion)

3) The explanatary variables are measured without error:

The random distrubance ϵ_i obsorbes the influence of omited variable and possibly errors of measurement y's (explained variable). i.e., we will assume that the regressor's are error free while the y values may (or) may not include errors of measurement.

4) The explanatary variables are not perfectly linearly correlated (No perfect multi-collinearity):

If there is more than one explanatary variable in the relationship it is assumed that they are not perfectly correlated each other.

For example:

In demand function of a commodity, depends up on the price commodity of A and price of the substituder commodity B. If two price are correlated each other it will be defiault to seperate the influence of two prices on the demand of the commodity.

5) The relationship should be correctly specified correct specification model):

(7) The model has specification error determing the important explanary variables the mathematical form should be correctly define (linear or non linear no. of variables in the model etc). To explain the relationships alternatively, there is no specification bias, or error in the model used in empirical analysis.

Estimation of the parameters β_0 and β_1:

The method of least squares ~~there~~ is used to estimate parameter β_0 and β_1. i.e., we will, estimate the parameter β_0 and β_1 so that the sum of squares of the difference between the observation y_i and the estimated values $\hat{y}_i$ is a minimum.

Consider the simple linear regression model $y = \beta_0 + \beta_1 x + \epsilon_i$.

$i = 1, 2 \dots n$

$$y = \beta_0 + \beta_1 x + \epsilon_i \longrightarrow (1)$$

The least squares criterian is $\sum \epsilon_i^2$.

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x)^2 \longrightarrow (2)$$

The least square estimators of parameters β_0 and β_1, say $\hat{\beta}_0$ and $\hat{\beta}_1$, must statisfied

(8)

$$\left.\frac{\partial S}{\partial \beta_0}\right|_{\hat{\beta}_0,\hat{\beta}_1} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) = 0$$

and.

$$\left.\frac{\partial S}{\partial \beta_1}\right|_{\hat{\beta}_0,\hat{\beta}_1} = 2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)(-x_i) = 0$$

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

Simplyfing this two equation we get

$$\left.\begin{aligned} n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i &= \sum_{i=1}^{n} y_i \\ \hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 &= \sum_{i=1}^{n} x_i y_i \end{aligned}\right\} \longrightarrow (3)$$

$\sum y_i - \hat{\sum_{i=1}} \beta_0 - \beta_1 \sum x_i = 0$

$\sum y_i = n\beta_0 + \beta_1 \sum x_i$

$a+0$

$\sum x_i y_i = \beta_0 \sum x_i - \beta_1$

$\sum x_i y_i = \beta_0 \sum x_i + \beta_1 \sum x_i^2$

Equation (3) called the least squares normal equation.

The solution to the normal equation is

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \longrightarrow (4)$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i x_i - \left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_i\right)/n}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}} \longrightarrow (5)$$

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

are the averages of y_i and x_i respectively.

Therefore, $\hat{\beta}_0$ and $\hat{\beta}_1$ in equation (4) & (5) are the least square estimators of the β_0 (intercept) and β_1 (slope) respectively.

The fitted simple linear regression model is

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \quad \rightarrow (6)$$

(9) Equation (6) gives a point estimates of the mean of y for the particular x.

Note :

Since the denominator of the equation (5) is the corrected sum of the square x_i and the numerbor is the corrected sum of ~~corres~~ cross product of x_i and y_i. we may write these quantities in a more compact notation as

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

$$S_{xy} = \sum_{i=1}^{n} y_i x_i - \frac{\left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_i\right)}{n} = \sum_{i=1}^{n} y_i (x_i - \bar{x})$$

Thus, the conveniont way to find

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

Properties of least square fit :

The sum of the residuals (error or distrubance) in recurrence model the contain in intercept of always zero, i.e.,

$$\sum_{i=1}^{n} (y_i - \hat{y}_i) = \sum_{i=1}^{n} e_i = 0.$$

The sum of observed value y_i is equals to the fitted values $\hat{y}_i$

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

iii) The least square regression line always passes throw the centraied of the data [The point $(\bar{x}, \bar{y})$]"

(10)

iv) The sum of residuals weighted by the corresponding value of the regressor variable always equal to zero

$$\sum_{i=1}^{n} \epsilon_i x_i = 0$$

v) The sum of residuals weighted by the corresponding fitted value always equal to zero.

$$\sum_{i=1}^{n} \epsilon_i \hat{y}_i = 0$$

~~Consider~~ Estimate the parameter β_0, β_1 for the following data, also varify the least squares Properties

X	Y	x^2	$x_i y_i$	y^2
2	5	4	10	25
4	7	16	28	49
5	8	25	40	64
6	10	36	60	100
8	12	64	96	144
11	18	121	198	324
36	60	266	432	706

$$\bar{x} = \frac{\Sigma x}{n} = \frac{36}{6} = 6$$

$$\bar{y} = \frac{\Sigma y}{n} = \frac{60}{6} = 10$$

11

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i x_i - (\sum y_i)(\sum_{i=1}^{n} x_i)/n}{\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}}$$

$$= \frac{(60) \times (3\cancel{6}) - (60)(\cancel{36})/6}{(266) - (\cancel{266})(36)^2}$$

$$\hat{\beta}_1 = \frac{432 - (60) \times (36)/6}{(266) - \frac{(36)^2}{6}}$$

$$= \frac{432 - 2160/6}{266 - 216}$$

$$= \frac{432 - 360}{266 - 216} = \frac{72}{500}$$

$$\hat{\beta}_1 = 1.44$$

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

$$= 10 - 1.44 \times 6$$

$$= 10 - 8.64$$

$$\hat{\beta}_0 = 1.36$$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

$$= 1.36 + 1.44(2)$$

$$= 4.24$$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 (x)$$

$$= 1.36 + 1.44(4)$$

$$= 7.12$$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 (x)$$

$$= 1.36 + 1.44(5) = 8.56$$

(12)

$\hat{y} = 1.36 + 1.44(6)$

$\hat{y} = 10$

$\hat{y} = 1.36 + 1.44(8)$

$= 12.88$

$\hat{y} = 1.36 + 1.44(11)$

$= 17.2$

$\epsilon_i = y - \hat{y}$	ϵ_i^2
$= 5 - 4.24 = 0.76$	0.5776
$= 7 - 7.12 = -0.12$	0.0144
$= 8 - 8.56 = -0.56$	0.3136
$= 10 - 10 = 0$	0
$= 12 - 12.88 = -0.88$	0.7744
$= 18 - 17.2 = 0.8$	0.64
0	2.92

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

$60 = 60$

$SS_{Res} = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2$

$= 706 - 600 - 1.44(72)$

$= 2.32$

$= \frac{2.32}{6-2} = \frac{2.32}{4} = 0.58$

iv) $\epsilon_i x_i = 0$

$0.76 \times 2 = 1.52$

$-0.12 \times 4 = -0.48$

$-0.56 \times 5 = -2.8$

$0 \times 6 = 0$

$-0.88 \times 8 = -7.04$

$0.8 \times 11 = 8.8$

$\underline{0}$

13

v) $\sum_{i=1}^{n} e_i \hat{y}_i = 0$

$= 0.76 \times 4.24 = 3.2224$

$= -0.12 \times 7.12 = -0.8544$

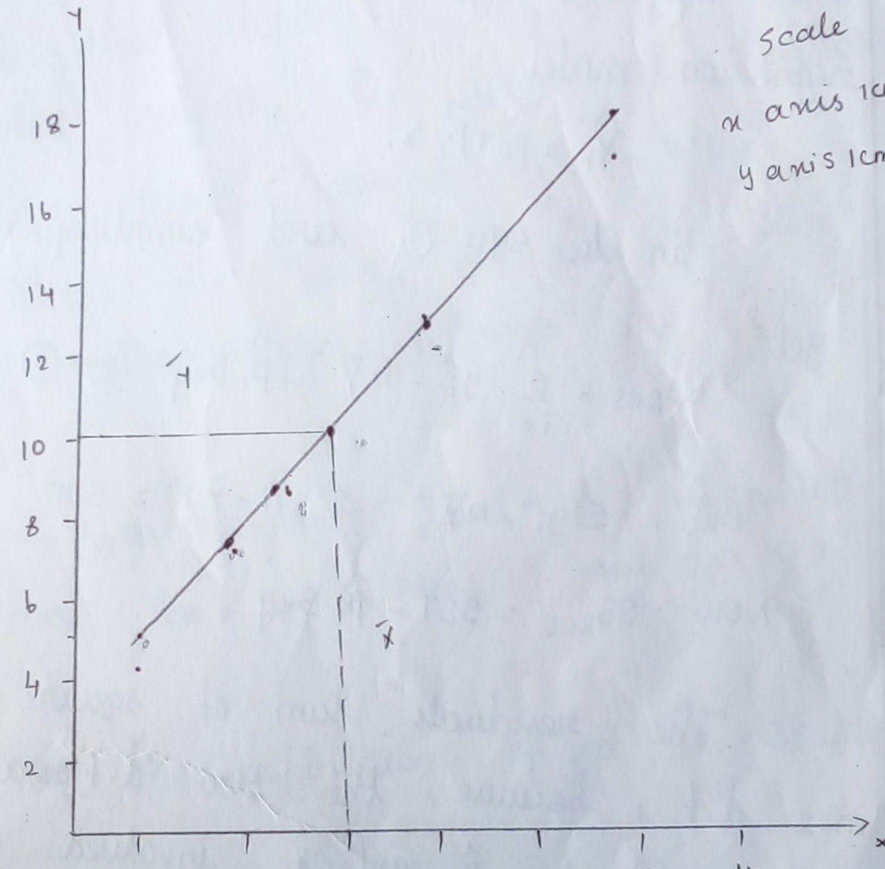
$= -0.56 \times 8.56 = -4.7936$

$= 0 \times 10 = 0$

$= -0.88 \times 12.88 = -11.3344$

$= 0.8 \times 17.2 = 13.76$

0



standard error of the estimator on Regression

In addition to estimating β_0, β_1 and estimate of σ^2 is required to test the hypothesis and construct interval estimates pertinent the regression model. we would like this estimate not to dependent on the adequacy of the fitted model. This is only possible when there are several observation on y for atleast one value of x.

The estimate of σ^2 is obtained from the residuals or error sum of squares.

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

A conviniant computing formula b_i SSRes may be found by substituing the fitted regression model

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

in to equ (1) and simplyfying we get

$$SS_{Res} = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 - \hat{\beta}_1 S_{xy} \longrightarrow (2)$$

But, $\sum y_i^2 - n\bar{y}^2 = \sum (y_i - \bar{y})^2 = SST$

Now $SS_{Res} = SST - \hat{\beta}_1 S_{xy}$

The residuals sum of square has $m-2$ d.f, because the two d.f are associated with estimate $\hat{\beta}_0$ and $\hat{\beta}_1$ involved in obtaining $\hat{y}_i$.

The expected value of SS_{Res} is

$$E(SS_{Res}) = (n-1)\sigma^2$$

So unbiased estimator of σ^2 in $n-2$ d.f

$$\hat{\sigma}^2 = \frac{SS_{Res}}{n-2} = MSS_{Res}$$

The quantity MSS_{Res} is called the

15 residual mean square.

The square root of $\hat{\sigma}^2$ is called standard error of regression ~~are~~ on estimator

Testing the significance of regression coefficient (Hypothesis Testing on the slope and intercept)

Date 10/10/2019

Step 1 :

suppose we wish to test hypothesis that the slope is a equals is a constant say β_{10}. The appropriate hypotheses are $H_0 : \beta_1 = \beta_{10}$ against $H_1 : \beta_1 \neq \beta_{10}$

where we have specified a two sided alternative hypothesis

Step 2 :

Fixing the level of significance, it may be 5 % ~~and~~ 1 % and or 10 %.

Step 3 :

By the definition of a t - statistic

$$|t_0| = \frac{\hat{\beta}_1 - \beta_{10}}{\sqrt{MS_{Res} / S_{xx}}}$$

R. Renba

Step 4 :

The d.f associated with t_0 are the no. of d.f associated with MS_{Res}.

Thus the test statistic follows $t_{\alpha/2, n-2}$ distribution.

Step 5 :

$|t_0|$ is greater than $t_{\alpha/2, n-2}$ we reject the Ho.

Alternatively, a P value approach

could also we used for decision making

Note:

The dinaminator of the test statistic t_0 is often called the estimated standard error or simply, the standard error of slope i.e $S.E(\hat{\beta}_1) = \sqrt{\frac{MS_{Res}}{S_{xx}}}$

Therefore, we often see $|t_0|$ is writen us

$$|t_0| = \frac{\hat{\beta}_1 - \beta_{10}}{S.E(\hat{\beta}_1)}$$

$$|t_0| = \frac{\hat{\beta}_1 - \beta_{10}}{\sqrt{MS_{Res}/S_{xx}}}$$

$$MS_{Res} = \frac{SS_{Res}}{n-2} = \frac{2.32}{6-2} = \frac{2.32}{4}$$

$$= 0.58$$

$$= \frac{1.44 - 0}{\sqrt{0.58/50}}$$

$$S_{xx} = \Sigma x_i^2 - \frac{(\Sigma x_i)^2}{n}$$

$$= \frac{1.44}{\sqrt{0.0116}}$$

$$= 266 - \frac{(36)^2}{6}$$

$$= \frac{1.44}{0.1077}$$

$$= 266 - \frac{1296}{6}$$

$$= 266 - 216$$

$$|t_0| = 13.37006$$

$$= 50$$

Table value: d.f = 4

5% level of significance = 2.78

1% level of significance = 4.60

10 % level of significance : 2.13

Inference :

F cal > F tab value. so we reject the Ho.

Problem : 2

Observat-ion	Shear strength (psi) (y_i)	Age of propellent (weeks) (x_i)
1	2158.70	15.50
2	1678.15	23.75
3	2316.00	8.00
4	2061.30	17.00
5	2207.50	5.50
6	1708.30	19.00
7	1784.70	24.00
8	2575.00	2.50
9	2357.90	7.50
10	2256.70	11.00
11	2165.20	13.00
12	2399.55	3.75
13	1779.80	25.00
14	2336.75	9.75
15	1765.30	22.00
16	2053.50	18.00
17	2414.40	6.00
18	2200.50	12.50
19	2654.20	2.00
20	1753.70	21.50

(18)

a) Fit a simple linear regression model to the data.

b) Make a scatter plot of the data.

c) calculate R^2.

d) Test the hypothesis $H_0: \beta_1 = 0$

e) Test the least square's Properties.

Sol:

a)

x	y	x^2	xy	y_i^2
15.50	2158.70	240.25	33459.85	4659,985.69
23.75	1678.15	564.0625	39856.0625	2816187.423
8.00	2316.00	64	18528	5363856
17.00	2061.30	289	35042.1	4248957.69
5.50	2207.50	30.25	12141.25	4873056.25
19.00	1708.30	361	32457.7	2918288.89
24.00	1784.70	576	42832.8	3185154.09
2.50	2575.00	6.25	6437.5	6630625
7.50	2357.90	56.25	17684.25	5559692.41
11.00	2256.70	121	24823.7	5092694.89
13.00	2165.20	169	28147.6	4688091.04
3.75	2399.55	14.0625	8998.3125	5757840.203
25.00	1779.80	625	44495	3167688.04
9.75	2336.75	95.0625	22783.3125	5460400.563
22.00	1765.30	484	38836.6	3116284.09
18.00	2053.50	324	36963	4216862.25
6.00	2414.40	36	14486.4	5829327.36
12.50	2200.50	156.25	27506.25	4842200.25
2.00	2654.20	4	5308.4	7044777.64
21.50	1753.70	462.25	37704.55	3075463.69
267.25	42627.15	4677.6875	528492.6875	92547433.46

(19)

$\Sigma x = 267.25$, $\Sigma y = 42627.15$, $\Sigma (x_i)^2 = 71422.562$

$\Sigma x_i^2 = 4677.6875$ $\quad \Sigma xy = 528492.6375$

$$S_{xx} = \Sigma x_i^2 - \frac{\Sigma (x_i)^2}{n}$$

$$= 4677.6875 - \frac{(267.25)^2}{20}$$

$$= 4677.6875 - \frac{71422.5625}{20}$$

$$= 4677.6875 - 3571.128125$$

$$S_{xx} = 1106.5594$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i x_i - \frac{(\Sigma y_i)(\Sigma x_i)}{n}}{\Sigma x_i^2 - \frac{\Sigma x_i^2}{n}}$$

$$= \frac{528492.6375 - \frac{(42627.15)(267.25)}{20}}{1106.5594}$$

$$= \frac{528492.6375 - 569605.2919}{1106.5594}$$

$$= \frac{-41112.65438}{1106.5594}$$

$$R^2 = \frac{SS_T - SS_{Res}}{SS_T}$$

$$\hat{\beta}_1 = -37.1535$$

$$\frac{SSR}{SST}$$

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

$$\bar{y} = \frac{\Sigma y}{n}$$

$$= \frac{42627.15}{20} = 2131.3575$$

y_i^2
4659,985.69
2816187.423
5363856
4248957.69
4873056.25
2918288.89
3185154.09
6630625
5559692.41
5092694.89
4688091.04
5757840.203
3167688.04
5460400.563
3116284.09
4216862.25
5829327.36
4842200.25
7044777.64
3075463.69
92547433.46

20

$\bar{x} = \frac{\Sigma x}{N} = \frac{267.25}{20} = 13.3625$

$\hat{\beta}_0 = 2131.3575 - (-37.1535) \times (13.3625)$

$= 2131.3575 + 496.4636438$

$= 2627.821144$

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

$\hat{y} = 2627.821144 + (-37.1535)x$

Fitted value :

$\hat{y}_1 = 2627.821144 - 37.1535 (15.50)$

$= 2627.821144 - 575.87925$

$= 2051.9418$

$\hat{y}_2 = 2627.821144 - 37.1535 (23.75)$

$= 2627.821144 - 882.395625$

$= 1745.4255$

$\hat{y}_3 = 2627.821144 - 37.1535 (8)$

$= 2627.821144 - 297.228$

$= 2330.5931$

$\hat{y}_4 = 2627.821144 - 37.1535 (17)$

$= 2627.821144 - 631.6095$

$= 1996.2116$

$\hat{Y}_5 = 2627.821144 - 37.1535\,(5.50)$

$= 2627.821144 - 204.3443$

$= 2423.4769$

$\hat{Y}_6 = 2627.821144 - 37.1535\,(19)$

$= 2627.821144 - 705.9165$

$= 1921.9046$

$\hat{Y}_7 = 2627.821144 - 37.1535\,(24)$

$= 1736.138$

$\hat{Y}_8 = 2627.821144 - 37.1535\,(2.50)$

$= 2534.9374$

$\hat{Y}_9 = 2627.821144 - 37.1535\,(7.50)$

$= 2349.1699$

$\hat{Y}_{10} = 2627.821144 - 37.1535\,(11)$

$= 2219.1326$

$\hat{Y}_{11} = 2627.821144 - 37.1535\,(13)$

$= \cancel{24}\ 2144.8256$

$\hat{Y}_{12} = 2627.821144 - 37.1535\,(3.75)$

$= 2488.4955$

$\hat{Y}_{13} = 2627.821144 - 37.1535\,(25.00)$

$= 1698.9836$

$\hat{Y}_{14} = 2627.821144 - 37.1535\,(9.75)$

$= 2265.5755$

$\hat{Y}_{15} = 2627.821144 - 37.1535\,(22)$

$= \cancel{2590.6676}\ 1810.4441$

$\hat{Y}_{16} = 2627.821144 - 37.1535\,(18)$

$= 1959.0581$

$\hat{Y}_{17} = 2627.821144 - 37.1535\,(6)$

$= 2404.9001$

$\hat{Y}_{18} = 2627.821144 - 37.1535\,(12.50)$

$= 2163.4024$

$\hat{Y}_{19} = 2627.821144 - 37.1535\,(2.00)$

$= 2553.5141$

$\hat{Y}_{20} = 2627.821144 - 37.1535\,(21.50)$

$= 1829.0208$

$\Sigma \hat{Y} = 42627.1512$

Error :

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

$\epsilon_1 = 2158.70 - 2051.9418$

$= 106.7582$

$\epsilon_2 = 1678.15 - 1745.4255$

$= -67.2755$

$\epsilon_3 = 2316.00 - 2330.5931$

$= -14.5931$

$\epsilon_4 = 2061.30 - 1996.2116$

$= 65.0884$

$\epsilon_5 = 2207.50 - 2423.4769$

$= -215.9769$

$\epsilon_6 = 1708.80 - 1921.9046$

$= -213.1046$

$\epsilon_7 = 1784.70 - 1736.138$

$= \cancel{12.562}\ 48.562$

$\epsilon_8 = 2575.00 - 2534.9374$

$= 40.0626$

$\epsilon_9 = 2357.90 - 2349.1699$

$= 8.7301$

$\epsilon_{10} = 2256.70 - 2219.1326$

$= 37.5674$

$\epsilon_{11} = 2165.20 - 2144.8256$

$= 20.3744$

$\epsilon_{12} = 2399.55 - 2488.4955$

$= -88.9455$

$\epsilon_{13} = 1779.80 - 1698.9836$

$= 80.8164$

$\epsilon_{14} = 2336.75 - 2265.5755$

$= 71.1745$

$\epsilon_{15} = 1765.30 - 1810.4441$

$= \cancel{1574.85}\ -45.1441$

(24)

$\epsilon_{16} = 2053.50 - 1959.0581$

$= 94.4419$

$\epsilon_{17} = 2414.40 - \cancel{2463.4024}\ 2404.9001$

$= 9.4999$

$\epsilon_{18} = 2200.50 - 2163.4024$

$= 37.0976$

$\epsilon_{19} = 2654.20 - 2553.5141$

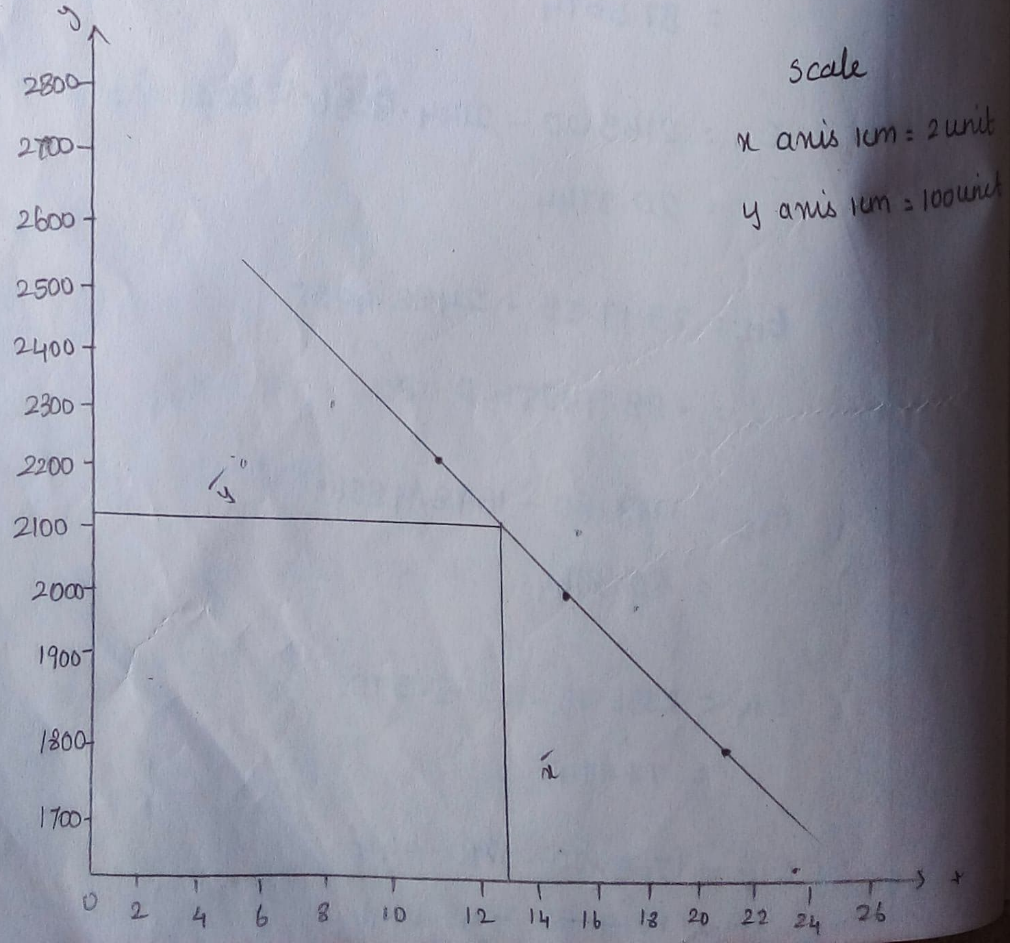
$= 100.6859$

$\epsilon_{20} = 1753.70 - 1829.0208$

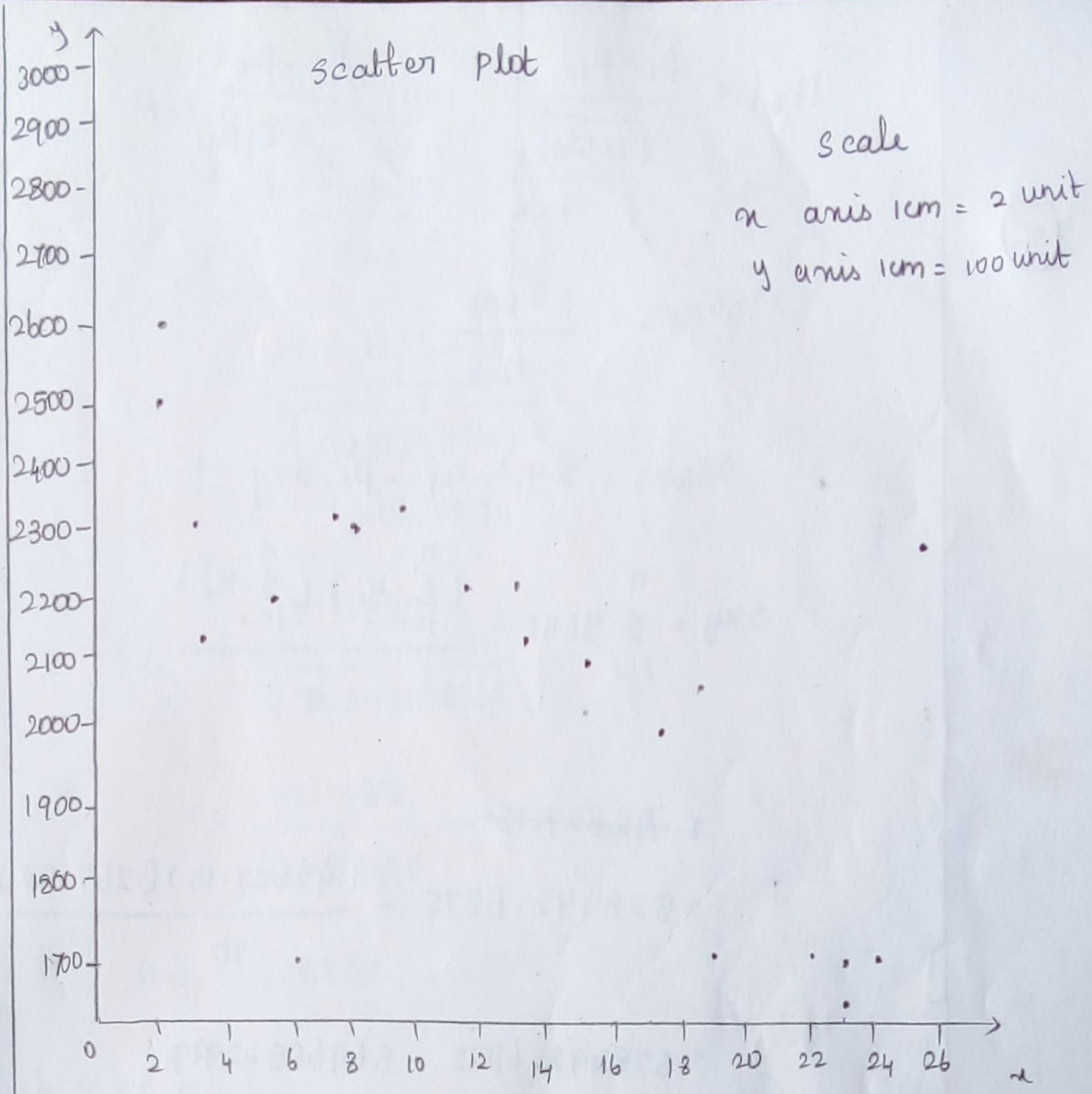
$= -75.3208$

$\Sigma \epsilon_i = 0.4988 \simeq 0$

$\epsilon_i = 0$



b) scatter plot



25

d) Test the hypothesis $\beta_1 = 0$.

Step 1 :

Null hypothesis :

$$H_0 : \beta_1 = 0$$

Alternative hypothesis :

$$H_1 : \beta_1 \neq 0.$$

Step 2 :

Level of significance :

$$\alpha = 5\%.$$

Step 3 :

Test statistic :

$$|t_0| = \frac{\hat{\beta}_1 - \beta_{10}}{\sqrt{\frac{MSS_{Res}}{S_{xx}}}} \quad (\text{or}) \quad \frac{\hat{\beta}_1 - \beta_{10}}{S.E(\hat{\beta}_1)}$$

$$MSS_{Res} = \frac{SS_{Res}}{n-2}$$

$$SS_{Res} = \Sigma y_i^2 - n\bar{y}^2 - \hat{\beta}_1 S_{xy}$$

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

~~= 42627.15 -~~

$$= 528492.6375 - \frac{(42627.15)(267.25)}{20}$$

$$= 528492.6375 - 569605.2919$$

$$S_{xy} = -41112.65$$

$$SS_{Res} = \Sigma y_i^2 - n\bar{y}^2 - \hat{\beta}_1 S_{xy}$$

$$= 92547433.46 - 20(2131.3575)^2 - (-37.1535) \times (-41112.65)$$

$$= 92547433.46 - 90853695.86 - 1527478.842$$

$$= 166258.758$$

$$MSS_{Res} = \frac{SS_{Res}}{n-2} = \frac{166258.758}{20-2}$$

$$= 9236.597667$$

$$|t_0| = \frac{\hat{\beta}_1 - \beta_{10}}{\sqrt{\frac{MSS_{Res}}{S_{xx}}}}$$

$$= \frac{-37.1535 - 0}{\sqrt{\frac{9236.597667}{1106.5594}}}$$

$$= \frac{-87.1535}{\sqrt{8.347132261}}$$

$$= \frac{-37.1535}{2.889140}$$

$$|t_0| = -12.8597$$

~~Tab~~ step 4:

Table value:

$\cancel{(18, 5)}$ d.f = 18 at 5% level of significance

= 2.10

Inference:

$\cancel{t_{cal} > t_{tab}}$

$t_{cal} > t$ tab value. so we reject

the H_0.

e) Test the least square's properties :

(28)

i). $\sum e_i = 0 = \sum_{i=1}^{n} (y_i - \hat{y}_i) = 42627.15 - 42627.15 = 0$

ii) $\sum y_i = \sum \hat{y}_i$ =

$42627.15 = 42627.15$

iii) $\sum e_i x_i = 0$

$= (106.7582) \times (15.50) + (-67.2755 \times 23.75) + \ldots$

$+ (-75.8208 \times 21.50)$

$= 0$

0.9018

2.4875

4.0875

c) calculate R^2

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2} \sqrt{\sum (y - \bar{y})^2}}$$

=

$15.50 - 13.3625 = 2.1375$

$23.75 - 13.3625 = 10.3875$

$8.00 - 13.3625 = -5.3625$

$17.00 - 13.3625 = 3.6375$

$5.50 - 13.3625 = -7.8625$

$19.00 - 13.3625 = 5.6375$

$24.00 - 13.3625 = 10.6375$

$2.50 - 13.3625 = -10.8625$

$7.50 - 13.3625 = -5.8625$

$11.00 - 13.3625 = -2.3625$

$13.00 - 13.3625 = -0.3625$

-9.6125

A similar Procedure can be used to test hypotheses about the intercept :

Step 1 :

Null hypothesis :

Suppose ~~the~~ we wish to test hypothesis that the intercept is a equals is a constant say β_{00}. The appropriate hypotheses are $H_0 : \beta_0 = \beta_{00}$ against $H_1 : \beta_0 \neq \beta_{00}$

where we have speified a two sided alternative hypothesis.

Step 2 :

Fixing the level of significance it may be 5 % & 1 % & 10 %.

Step 3 :

By the definition of a t - statistic

$$|t_0| = \frac{\hat{\beta}_0 - \beta_{00}}{S.E(\hat{\beta}_0)}$$

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

$$|t_0| = \frac{\hat{\beta}_0 - \beta_{00}}{\sqrt{MS_{Res}\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}}$$

Step 4 :

The d.f associated with t_0 are the no. of d.f associated with MS_{Res}. Thus the test statistic follows $t_{\alpha/2, n-2}$ distribution.

(30)

step 5 :

$|t_0|$ is greater than $t_{\alpha/2}, n-2$ we reject the H_0.

Alternatively, a p value approach could also we used for decision making.

$$|t_0| = \frac{\hat{\beta}_0 - \beta_{00}}{S.E(\hat{\beta}_0)}$$

$$= \frac{1.36}{\sqrt{0.58\left(\frac{1}{6}+\frac{36}{50}\right)}} = \frac{1.36}{\sqrt{0.58 \times (0.1666 + 0.72)}}$$

$$= \frac{1.36}{\sqrt{0.514228}}$$

$$= \frac{1.36}{0.717096}$$

$$|t_0| = 1.8965358$$

Table value :

α =

4 d.f 5% level of significance = 2.78.

Inference.

$t_{cal} < t_{\alpha/2}, n-2$. Accept the H_0.

Date 17.12.2019

(31)

standard error of Prediction :

An important of the application of the regression model is Prediction of New observations y corresponding to a specified level of the regressor variable x. If x_0 is the value of the regressor variable of interest. Then,

$$\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0 \longrightarrow (1)$$

Is the point estimate of the new value of the response Y_0.

Now, considering two obtain an interval estimate of this future observation Y_0.

If we use $\hat{Y}_0$ to Predict Y_0 then the standard error of $Y_0 - \hat{Y}_0$ is the appropriate statistic are which to base a Prediction interval.

∴ The standard error of $Y_0 - \hat{Y}_0$ an a future observation at x_0 is

$$\sqrt{MS_{Res}\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)} \longrightarrow (2)$$

Thus, the $100 \times (1-\alpha)$ percent prediction interval on a future observation x_0 is

$$\hat{Y}_0 - t_{\alpha/2, n-2}\ S.E(\hat{y}_0) \leq Y_0 \leq \hat{Y}_0 + t_{\alpha/2, n-2}\ S.E(\hat{y}_0)$$

34

$$S.E(\hat{y}_0) = \sqrt{MS_{Res}\left(1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right)}$$

$$= \sqrt{0.58\left(1+\frac{1}{6}+\frac{(20-6)^2}{50}\right)}$$

$$= \sqrt{0.58(1+0.1666+3.92)}$$

$$= \sqrt{0.58(5.0866)}$$

$$= \sqrt{2.9502}$$

$x_0-\bar{x} =$

$7.7135 \leq Y_0 \leq 12.286$

$$S.E(\hat{y}_0) = 1.7176$$

$$\hat{Y}_0 - t_{\alpha/2, n-2}\, S.E(\hat{y}_0) \leq Y_0 \leq \hat{Y}_0 + t_{\alpha/2, n-1}\, S.E(\hat{y}_0)$$

$$30.16 - 2.78(1.7176) \leq Y_0 \leq 30.16 + 2.78(1.7176)$$

$$30.16 - 4.774928 \leq Y_0 \leq 30.16 + 4.774928$$

$$25.3850 \leq Y_0 \leq 34.9349$$

(37)

i.e

$$\hat{Y}_0 - t_{\alpha/2, n-2}\sqrt{MS_{Res}\left(1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right)} \le Y_0 \le \hat{Y}_0 + t_{\alpha/2, n-2}\sqrt{MS_{Res}\left(1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right)} \rightarrow (3)$$

The prediction interval in equation (3) is of minimum of with at $x_0 = \bar{x}$ and widens as $(x_0-\bar{x})$ increases.

Problem (1):

$$\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

$$= 1.36 + 1.44\, x_0 \qquad MS_{Res} = 0.58$$

Assume that $x_0 = 13$ $\qquad S_{xx} = 50$

$$\hat{Y}_0 = 1.36 + 1.44(13)$$

$$= 20.08$$

$$S.E(\hat{y}_0) = \sqrt{MS_{Res}\left(1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right)}$$

$$= \sqrt{0.58\left(1+\frac{1}{6}+\frac{(13-6)^2}{50}\right)}$$

$$= \sqrt{0.58\left(1+\frac{1}{6}+\frac{(7)^2}{50}\right)}$$

$$= \sqrt{0.58\left(1+0.1666+\frac{49}{50}\right)}$$

$= \sqrt{0.58\ (1+0.1666+0.98)}$

33

$= \sqrt{0.58\ (2.1466)}$

$= \sqrt{1.245028}$

$S.E(\hat{y}_0) = 1.115808$

$$\hat{Y}_0 - t_{\alpha/2, n-2}\sqrt{MS_{Res}\left(1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right)} \le Y_0 \le \hat{Y}_0 + t_{\alpha/2, n-2}\sqrt{MS_{Res}\left(1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right)}$$

$20.08 - 2.78\ (1.115808) \le Y_0 \le 20.08 + 2.78\ (1.115808)$

$20.08 - 3.101946 \le Y_0 \le 20.08 + 3.101946$

$16.97805 \le Y_0 \le 23.181946$

The standard error of the prediction interval is

$$\sqrt{MS_{Res}\left(1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right)}$$

Problem ①

$\hat{Y}_0 = 1.36 + 1.44\ x_0$

$\hat{Y}_0 = 1.36 + 1.44\ (20)$

$\hat{Y}_0 = 30.16$

$MS_{Res} = 0.58$

$S_{xx} = 50.$