eg:- Annual truck repair expenses

Truck No	Age of Truck in yrs (x)	Repair expense during last yr in hundreds of $ (y)
101	5	7
102	3	7
103	3	6
104	1	4

Sol:-

$$b = \frac{\Sigma xy - n\cdot\bar{x}\bar{y}}{\Sigma x^2 - n\bar{x}^2} = \frac{(35+21+18+4) - 4\times3\times6}{(25+9+9+1) - 4\times9}$$

$$= \frac{68-72}{44-36} = \frac{-4}{8} = 0.75$$

$$a = \bar{y} - b\cdot\bar{x} = 6 - 0.75\times3 = 3.75$$

Reg. eqn $\Rightarrow$ $y = 3.75 + 0.75x$

If the age of truck is 6 then estimated value (repair expense) is $\hat{y} = 3.75 + 6\times0.75$

Standard Error

$$S_e = \sqrt{\frac{(y-\hat{y})^2}{n-2}}$$

y = act. value, $\hat{y}$ = estimated value

when we have regression eqn, we will have 'n' pairs of datasets, the regression eqn $y = a+bx$ and a and b are reg. coefficients and are estimated values, so, degrees of freedom n-2. because the 'n' no. of

data points are linked by 2 normal equations. i.e., 2 constraints
So, degrees of freedom is n-2.
Standard Error is measure of reliability of regression eqn.
If SE is higher i.e., dist b/w y and $\hat{y}$ (reg line) then points are more dispersed / they are far away
If SE is 0 i.e., all the points are on reg line. Then we can say that regression line is perfect fit for data points.

Standard Error for same problem,

$$S_e = \sqrt{\frac{\Sigma(y-\hat{y})^2}{n-2}} \quad S_e = \sqrt{\frac{\Sigma y^2 - a\Sigma y - b\Sigma xy}{n-2}}$$ (both can be used.)

$$S_b = \frac{S_e}{\sqrt{\Sigma x^2 - n\bar{x}^2}}$$

$\hat{y} = 3.75 + 0.75x \Rightarrow$ if $x = 5 \Rightarrow \hat{y} = 3.75 + 3.75 = 7.5$

$x = 3 \Rightarrow \hat{y} = 3.75 + 2.25 = 6$

$x = 3 \ \hat{y} \Rightarrow 6$

$x = 1 \Rightarrow \hat{y} = 3.75 + 0.75 = 4.5$

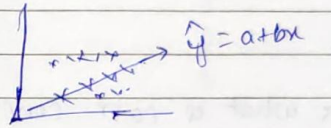
~~Se~~ $= \sqrt{\frac{6.25 + 9 + 12.25}{4-2}}$

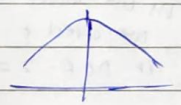
$$S_e = \sqrt{\frac{(7-7.5)^2 + (7-6)^2 + (6-6)^2 + (4-4.5)^2}{4-2}}$$

$$= \sqrt{\frac{0.25 + 1 + 0 + 0.25}{2}} = \sqrt{\frac{1.5}{2}} = \sqrt{0.75}$$

$$= 0.866$$

This is standard error of equation.
If $S_e = 0$ then we would have said line is passing through points. But here value is 0.866 i.e., some points are away from line.





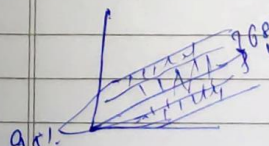
In a normal distribution, there are
68% points in $\mu \pm 1$ sd
95% " " $\mu \pm 2$ sd

If the size of sampling dist > 30 then 68% of sample points then 68% in sample mean ± 1 sd
95% points lie in sample mean ± 3 sd

$$\pm z = \frac{\bar{x} - \mu}{SE}, \quad \mu \pm 3SE \rightarrow 95\%$$

$$\hat{\mu} = \bar{x} \pm z \cdot SE$$
95-99

95% confidence level is wider than 95%.
95% " " " " " 90%.



when we have large (>30) data set then we can tell

Conf. interval $= (\hat{y} \pm z \times SE)$ it follows z-distribution.

for smaller sample size it follows t distribution.

conf. level interval $= (\hat{y} \pm t \times SE)$

if age of truck = 4 yrs. what is corr. conf. int at 90%

$\hat{y} = 3.75 + 4 \times 0.75 = 6.75$

As we have only 4 observations we used t-distribution.

$= 6.75 \pm t_{90} \times 0.866 \Rightarrow t_{90}$ at DOF = 2 $\Rightarrow$

$= [6.75 - 2.920 \times 0.866, 6.75 + 2.920 \times 0.861]$

$[4.22, 9.27]$

For a truck of age 4 yrs, with 90% confidence we can say based on data that we have that repair expense would be b/w (4.22, 9.27)

Here, we also have the standard error for b-coefficient in $y = a + bx$.

$\Rightarrow S_b = \dfrac{S_e \rightarrow \text{SE of regression.}}{\sqrt{\Sigma x^2 - n\bar{x}^2}}$

classmate
Date
Page

Eg:-	Yr	RND Exp (x)	Profit (y)
	2013	5	31
	2014	11	40
	2015	4	30
	2016	5	34
	2017	3	25
	2018	2	20

Sol: y can be expressed as $y = a + bx$

$S_e = \sqrt{\dfrac{(y - \hat{y})^2}{n-2}}$

$b = \dfrac{\Sigma xy - n\bar{x}\bar{y}}{\Sigma x^2 - n\bar{x}^2}$

$a = \bar{y} - b\cdot\bar{x} = 30 -$

$\hat{y} =$

$\bar{b} = (1055 + 640 + 120 + 120 + 25 + 40)$

$- 6 \times 30 \times 5$

$(25 + 121 + 16 + 25 + 9 + 4) - 6 \times 25$

$b = \dfrac{1000 - 900}{200 - 150} = \dfrac{100}{50} = 2$

$\hat{y} = 20 + 2x$

$\hat{y}$ if $x = 5 \Rightarrow 20 + 10 = 30$

$x = 11 \Rightarrow 20 + 22 = 42$

4 → 28

5 → 30

3 → 26

2 → 24

$S_e = \sqrt{\dfrac{(1) + 4 + 4 + 16 + 1 + 16}{4}}$

$= \sqrt{42/4} = \sqrt{10.5}$

$= 3.24$

$$S_b = \frac{S_e}{\sqrt{\Sigma x^2 - n\bar{x}^2}}$$

$$= \frac{3.24}{\sqrt{(25+121+16+2+9+4) - 6\times(5)^2}}$$

$$= \frac{3.24}{\sqrt{200-150}} = \frac{3.24}{\sqrt{50}} = \frac{3.24}{7.071} = 0.458.$$

S_b is dependent on S_e.

Conf. int for b coeff at 90% conf. level.

conf. level int $= (\hat{y} \pm t \times S_b')$

$= (2 \pm 2.132 \times 0.458)$

$= (2 \pm 0.98) = (1.02, 2.98)$

at 90% conf. level.

It is b value i.e., slope.

For every ↑ in R&D exp profit will be increase in (1.02, 2.98)

if we have less conf. level the range will be less.

For simple regression, it depends on 2 normal eqns.

If we consider Se as a whole then it will effect whole p

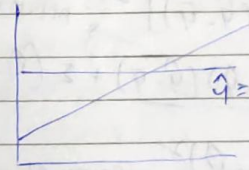
Based on, Reg. coeff 'b' we can have conf. inte. every 1 unit ↑ in x value what will be the ↑ in y.

6/11/21

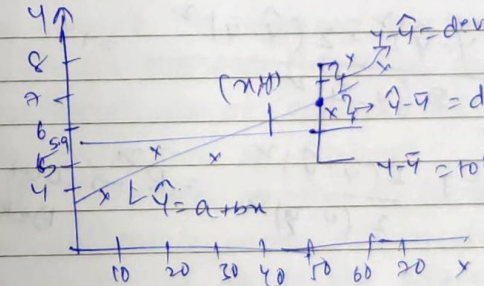
for a certain value of independent variable the confidence interval of dependent variable value can be determined using sample error of equation.

Explaining variation : R^2

A. Breaking down the distances



Total variation.



Total Deviation = Explained Deviation + Unexplained Deviation.

explained → when the point is on the line i.e., it satisfies the regression equation. So, it's been explained by regression.

unexplained → If the line is not passed through point i.e., not explained by regression i.e., at bitary point.

Sums of Squares

we can sum this equation across all the y's and square both sides to get:

$$\Sigma (y-\bar{y})^2 = \Sigma [(y-\hat{y}) + (\hat{y}-\bar{y})]^2$$ → acc. to a theorem this prod = 0

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

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

Square of exp. deviation — Square of unexplained deviation

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

$$\Rightarrow \frac{\Sigma (y-\hat{y})^2}{\Sigma (y-\bar{y})^2} + \frac{\Sigma (\hat{y}-\bar{y})^2}{\Sigma (y-\bar{y})^2} = 1$$

$$\frac{\Sigma (\hat{y}-\bar{y})^2}{\Sigma (y-\bar{y})^2} = 1 - \frac{\Sigma (y-\hat{y})^2}{\Sigma (y-\bar{y})^2} = R^2$$ → Coefficient of Determination

Rel. Deviation

How many % of points of plane can be explained by regression eq in scatter plot that is those many points lie on line.

if $R^2 = 0.81$

$= \sqrt{0.81} = \pm 0.9$.

i.e., it can be +ve or -vely correlated.

It depends on the slope of the line.

If the slope is +ve then the coefficient is +ve,

" " " " -ve " " " " -ve.

Variation of Y values around the regression line:

$\Sigma (y-\hat{y})^2$.

Variation of Y values around their own values

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

Coefficient of determination $= r^2$.

(1) Total sum of squares (SST).

The term on the left hand side of this eqn is the sum of the squared distances from all points to $\bar{y}$. we call this the total variation in the y's, or the total sum of squares (SST).

(2) Regression Sum of Squares:

The first term on the right hand side is the sum of the squared distances from the regression line to $\bar{y}$. We call it the Regression sum of Squares or SSR.

(3) Error sum of Squares

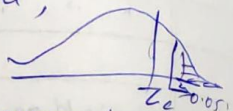
Finally, the last term is the sum of squared distances

from the points to the regression line. Remember, this is the quantity that least squared minimizes.

$TSS = SSR + SSE$

P-Value Approach

The area under the density function right to the value of Z obtained from the sample,



when $p < 0.05$

Alternate/reasearch hypothesis gets accepted at 5% significance level.

when p < level of sig then alt. hypothesis gets accepted.

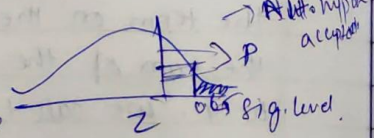
" p > " " " null " " "

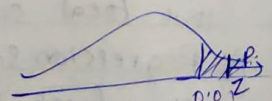
Instead of critical value we are checking P value.

$P = 1 - Z_c$

P is right side of Z

P is compared to significance level.





Multiple regression

$Y = a + b_1 x_1 + b_2 x_2$ ($k = 2$)

Y → dep. var

x_1, x_2 → indep variables.

($Y = a + bx$, $SE = \sum(Y - \hat{Y})^2 / n-2$) single regression.

$SE = \sqrt{\frac{\sum(Y - \hat{Y})^2}{(n - k - 1)}}$ k = no. of independent variable

In this case $k = 2 \Rightarrow n - 2 - 1$

In prev. case indep var $= 1 \Rightarrow k = 1 \Rightarrow n - 1 - 1 = n - 2$

$TSS = SSR + SSE$

Similar to F-test in regression also we have 2 variances i.e., SSR and SSE.

We can also do a F-test to find significance of regression as a whole.

$$F = \frac{SSR/k}{SSE/(n-k-1)}$$

If $P > F$ then we can say reg. eqn is significant.

Pop. reg. eqn $Y = A + B_1 x_1 + B_2 x_2$

$H_0 : B_1 = B_2 = 0$

$H_a : B_1 \neq B_2 \neq 0$

If null hypo is true

x_1 and x_2 cannot explain regression as whole

when alt. hypothesis is true then we can say regression can be explained in terms of independent variables.

$H_0 : B_1 = 0$ $H_a : B_1 \neq 0$

This is regarding one coefficient B.

Then use t-test.

$$t = \frac{B_1 - 0}{SE_B}$$

with the help of t-test we can tell for individual coefficients.

ANOVA in Excel

MS → Mean Squares

SS → Sum of Squares.

In reg,

$$\text{Adjusted } R^2 = 1 - (1 - R^2)\left(\frac{N-1}{N-K-1}\right)$$

N = No. of Data points K = No. of independent variables

When K is small when compared to 1 then $\frac{N-1}{N-K-1} = 1$,

$\Rightarrow$ Adjusted $R^2 = R^2$