Regression Analysis - 2

Step 1 :- for finding weather there is relationship b/w exp & Salary we use hypothesis.

H_0 : $Y = \beta_0 + \beta_1 X$

Y ↑ Salary, X ↑ exp

H_0 : $X = 0$ $\quad\quad$ H_1 : $X \neq 0$

Step-2 find linear regression model of data estimating regression coefficients (intercept & slopes)

Step-3 Calculate TSS, ESS, RSS

Step 4 mean Square ~~Error~~ Regression (MSR)

$$= \frac{ESS}{df\_model} = \frac{ESS}{K}$$

↳ also called Avg explained variance per independent feature

Step 5 mean Square Error (MSE)

$$= \frac{RSS}{df\_residual} = \frac{RSS}{n-(K+1)}$$

→ also called avg unexplained variance per dof

Step 5. : calculate f statistics

$$= \frac{MSR}{MSE}$$

$$MSR = \frac{ESS}{k} \therefore \frac{TSS - RSS}{k}$$

$$MSE = \frac{RSS}{n-k-1} = \frac{\Sigma(y_i - \hat{y}_i)^2}{n-k-1}$$

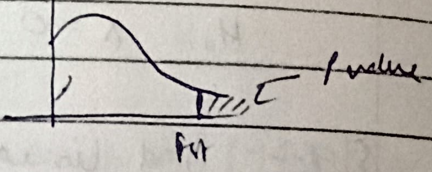
$$TSS = (y_i - \bar{y})^2$$

Step 6 As we got f-statistics now we will find p-value

p-value < 0.05 (reject null hypo)

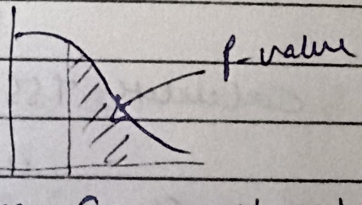
$$F_{stat} \Rightarrow \frac{\frac{ESS}{k}}{\frac{RSS}{n-k-1}}$$

→ suppose $F_{stat} >> 1$ means explained SS is large



p-value will be small, reject H_0

· $F_{stats} << 1$



Pvalue will be large so can't reject H_0

· If we have more than one β_0 β_1 β_2 then rejecting H_0 means atleast one of $\beta_0, \beta_1, \beta_2$ is non-zero

Page No.
Date

Strength of

- Now finding Relationship b/w X & Y

↳ $R^2 = \frac{ESS}{TSS} = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$

- R^2 → ~~is used~~ tells how much % or proportion of variance present in dependent variable can explained by independent var

→ Suppose R_2 score of CGPA | Salary is 0.4

↳ we add a relevant colm → r_2 Score = 0.6

But if we add irrelevant " → r_2 Score = 0.4

But we want r_2 Score to decrease if add irrelevant colm.

→ Adjusted R^2 Score

$$= 1 - \left[\frac{(1-R^2) * (n-1)}{(n-k-1)} \right]$$

- T-Statistics -

T test ⇒ $y_{\alpha} = \beta_0 + m\beta_1$

for slope → Null hypothesis : $\beta_1 = 0$

Alternative " : $\beta_1 \neq 0$

for intercept : $H_0 : \beta_0 = 0$

$H_a = \beta_1 \neq 0$

t-Statistic $= \frac{\beta_1 - 0}{SE(\beta_1)}$ because we assume $\beta_1 = 0$

for $\beta_0 \Rightarrow \dfrac{b_0 - 0}{SE(\beta_0)}$

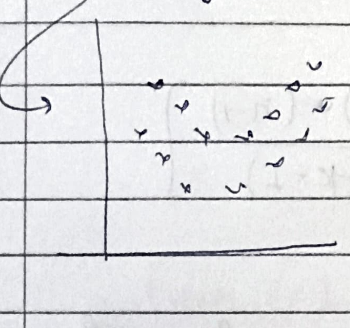
$$SE(\beta_1) = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{(n-2)\sum(x_i - \bar{x})^2}}$$

$$SE(\beta_0) = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{(n-2)}\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2}\right]}$$

- Calculating confidence interval -

$$CI_{b_0} = b_0 \pm t\_value * SE(b_0)$$
$$CI_{b_1} = b_1 \pm t\_value * SE(b_1)$$

- Polynomial Regression -



input X	output Y
35	100

$(1,2)$

Suppose we do degree 2

then

x^0, x^1, x^2			y
1	35	1225	100

$(1,4)$

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2$$

→ we have to choose degree in proper way.

If degree is too high overfitting
" " " less underfitting