

# REGRESSION ANALYSIS

## CHAPTER-1: SIMPLE LINEAR REGRESSION

Books:- Introduction to Linear Regression Analysis (3rd Ed)  
By Montgomery (2001).  
Applied Regression Analysis (3rd Ed) by  
Draper & Smith. (1998).

- Regression analysis is a statistical tool for investigating the relationship between a dependent variable and one or more independent variables. This technique is widely used for prediction and forecasting. Applications: Economics, Management, Life & Biological Sciences, Physical & Chemical science & engineering etc. Scatter plot is an essential tool for checking correlation.
- Simple linear Regression model is a model with single regressor  $x$  that has a linear relationship with a response  $y$ .

Model is:  $y = \beta_0 + \beta_1 x + \epsilon$

Response variable (random variable)       $\downarrow$  intercept       $\downarrow$  regression variable component  
 slope (controlled variable)

We now make some basic assumption on the model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i ; i=1(1)n$$

Given data  $(X_i, Y_i)$ ; checking scatter plot whether linear model is appropriate or not.

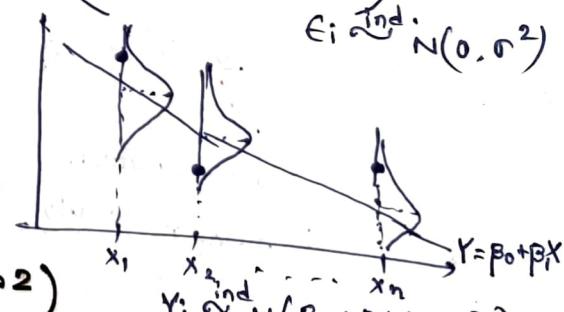
- Assumptions:-
- 1.  $\epsilon_i$  is an RV with mean '0' and s.d.  $\sigma$  (unknown) i.e.,  $E(\epsilon_i) = 0$ ,  $V(\epsilon_i) = \sigma^2$  (Homoscedasticity)
  - 2.  $Cov(\epsilon_i, \epsilon_j) = 0 \Rightarrow \epsilon_i \& \epsilon_j$  are uncorrelated. i.e. (no autocorrelation.)
  - 3.  $\epsilon_i \stackrel{\text{ind.}}{\sim} N(0, \sigma^2)$ .

$$E(Y_i) = E(\beta_0 + \beta_1 X_i + \epsilon_i) = \beta_0 + \beta_1 X_i$$

$$V(Y_i) = V(\beta_0 + \beta_1 X_i + \epsilon_i) = V(\epsilon_i) = \sigma^2$$

$$\epsilon_i \stackrel{\text{ind.}}{\sim} N(0, \sigma^2)$$

$$Y_i \stackrel{\text{ind.}}{\sim} N(\beta_0 + \beta_1 X_i, \sigma^2)$$



The line fitted by least square is the one that makes the sum of squares of all vertical discrepancies as small as possible.

BY TANUJIT CHAKRABORTY

We estimate  $\beta_0$  &  $\beta_1$  so that the sum of squares of all the diff. between the observation  $y_i$  and the fitted line is minimum.

$$S = \text{SSRes} = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

$\hat{\beta}_0$  is minimum.

$\hat{\beta}_0$  is the estimate of  $\beta_0$ , and  $\hat{\beta}_1$  is the estimate of  $\beta_1$ . The least square estimators of  $\beta_0$  and  $\beta_1$  (i.e.,  $\hat{\beta}_0, \hat{\beta}_1$ ) must satisfy

$$\frac{\partial S}{\partial \beta_0} \Big|_{\hat{\beta}_0, \hat{\beta}_1} = -2 \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \quad \left. \begin{array}{l} \\ \end{array} \right\} \text{Normal equations}$$

$$\frac{\partial S}{\partial \beta_1} \Big|_{\hat{\beta}_0, \hat{\beta}_1} = -2 \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) x_i = 0$$

So, the estimators  $\hat{\beta}_0$  and  $\hat{\beta}_1$  are solution of the normal equations.

Solving these two equations:-

Properties of LS fit:

$$\sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \Rightarrow \sum y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum x_i = 0$$

$$\Rightarrow n\hat{\beta}_0 = \sum y_i - \hat{\beta}_1 \sum x_i \quad \left[ \begin{array}{l} \text{where } \bar{x} = \frac{\sum x_i}{n} \\ \bar{y} = \frac{\sum y_i}{n} \end{array} \right]$$

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

$$\sum x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \Rightarrow \sum x_i (y_i - \bar{y} + \hat{\beta}_1 \bar{x} - \hat{\beta}_1 x_i) = 0$$

$$\Rightarrow \sum x_i (y_i - \bar{y}) = \hat{\beta}_1 \sum (x_i - \bar{x}) x_i$$

$$\Rightarrow \hat{\beta}_1 = \frac{\sum (y_i - \bar{y}) x_i}{\sum (x_i - \bar{x}) x_i}$$

Note:  $\sum e_i y_i = 0$  but  $\sum e_i y_i \neq 0$ .

Since

$$\begin{aligned} \sum (y_i - \bar{y}) \bar{x} &= \bar{x} \sum y_i - \bar{x} \sum \bar{y} \\ &= n \bar{x} \bar{y} - n \bar{x} \bar{y} \\ &= 0. \end{aligned}$$

$$\hat{\beta}_1 = \frac{\sum (y_i - \bar{y}) (x_i - \bar{x})}{\sum (x_i - \bar{x}) (x_i - \bar{x})} = \frac{s_{xy}}{s_{xx}}$$

$$= \frac{1}{n} \sum (x_i - \bar{x}) y_i / \frac{1}{n} \sum (x_i - \bar{x})^2$$

$$\Rightarrow \hat{\beta}_1 = \frac{\sum x_i y_i - n \bar{x} \bar{y}}{\sum x_i^2 - n \bar{x}^2}$$

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

$$\Rightarrow \hat{\beta}_0 = \frac{1}{n} \sum y_i - \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_1 = \sum_{i=1}^n c_i y_i \quad \text{where } c_i = \frac{x_i - \bar{x}}{\sum (x_i - \bar{x})^2}$$

$$\text{Since } \sum (x_i - \bar{x})^2 = 0, \quad \text{of observations } y_i: \quad \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Question:

So,  $\hat{\beta}_1, \hat{\beta}_0$  are linear combinations, so its called Linear estimators

To show:  $E(\hat{\beta}_1) = \beta_1$

Statistical properties of LS estimators:

$\hat{\beta}_1$  is a

UE of  $\beta_1$ ,  $\hat{\beta}_0$  is a UE of  $\beta_0$ .

(Unbiased estimator = UE)

Remember:  $y_i$  is the random variable,  $x_i$  is the controlled variable.

$$\bar{y} = \beta_0 + \beta_1 \bar{x} + \bar{e} = \bar{y} = \frac{1}{n} \sum y_i; \text{ where } \bar{x} = \frac{1}{n} \sum x_i \text{ and } \bar{e} = \frac{1}{n} \sum e_i$$

(Continued)

3

$$y_i - \bar{y} = \beta_1(x_i - \bar{x}) + \epsilon_i - \bar{\epsilon}$$

$$E(y_i - \bar{y}) = \beta_1(x_i - \bar{x}) + E(\epsilon_i - \bar{\epsilon}) = \beta_1(x_i - \bar{x}); \text{ since } \epsilon_i \sim N(0, \sigma^2)$$

$$E(\hat{\beta}_1) = E\left[\frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}\right] = \frac{\sum(x_i - \bar{x})E(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{\beta_1 \sum(x_i - \bar{x})^2}{\sum(x_i - \bar{x})^2} = \beta_1$$

Q.D.

$\hat{\beta}_0$  is an  
UE of  $\beta_0$ .

$$E(\hat{\beta}_0) = E(\bar{y} - \hat{\beta}_1 \bar{x}) = E(\beta_0 + \beta_1 \bar{x} - \hat{\beta}_1 \bar{x}) = \beta_0 + (\underbrace{\beta_1 \bar{x} - \hat{\beta}_1 \bar{x}}_{\text{since } E(\hat{\beta}_1) = \beta_1}) = \beta_0$$

$$V(\hat{\beta}_1) = V\left(\frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}\right) = V\left(\frac{\sum(x_i - \bar{x})y_i}{\sum(x_i - \bar{x})^2}\right) = V\left(\frac{\sum c_i y_i}{\sum(x_i - \bar{x})^2}\right) = V\left(\sum c_i y_i\right) \quad (\text{proved})$$

$$\text{where, } c_i = \frac{(x_i - \bar{x})}{\sum(x_i - \bar{x})^2}$$

$$= \sum c_i^2 V(y_i) = \sum c_i^2 \sigma^2 = \sigma^2 \cdot \frac{\sum(x_i - \bar{x})^2}{[\sum(x_i - \bar{x})^2]^2} = \frac{\sigma^2}{S_{xx}}$$

$$V(\hat{\beta}_0) = V(\bar{y} - \hat{\beta}_1 \bar{x}) = V(\bar{y}) + V(\hat{\beta}_1 \bar{x}) - 2\bar{x} \text{Cov}(\bar{y}, \hat{\beta}_1)$$

$$\boxed{V\left(\frac{1}{n} \sum y_i\right) = \frac{1}{n^2} \sum V(y_i) = \frac{\sigma^2}{n^2} = 0} = \frac{\sigma^2}{n} + \bar{x}^2 V(\hat{\beta}_1) - 2\bar{x} \times 0 \\ = \frac{\sigma^2}{n} + \frac{\bar{x}^2 \sigma^2}{S_{xx}} = \sigma^2 \left( \frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right).$$

$$\left[ \text{Cor}(\bar{y}, \hat{\beta}_1) = \text{Cov}\left(\frac{\sum y_i}{n}, \frac{\sum(x_i - \bar{x})y_i}{\sum(x_i - \bar{x})^2}\right) = \frac{\sum(x_i - \bar{x})V(y_i)}{n \sum(x_i - \bar{x})^2} \right]$$

$$= \frac{\sum(x_i - \bar{x})V(y_i)}{n \sum(x_i - \bar{x})^2} = \frac{\sum(x_i - \bar{x})\sigma^2}{n \sum(x_i - \bar{x})^2} = \frac{\sigma^2 \sum(x_i - \bar{x})}{n \sum(x_i - \bar{x})^2}$$

$$= 0.$$

Note:  $V(\hat{\beta}_1), V(\hat{\beta}_0)$  involve  $\sigma^2$ . Now,  
How to estimate  $\sigma^2$ ?

Estimation of  $\sigma^2$ :

$$SS_{Res} = \sum e_i^2 = \sum (y_i - \hat{y}_i)^2$$

$$= \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2; \text{ since } \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

$$= \sum (y_i - \bar{y} - \hat{\beta}_1 (x_i - \bar{x}))^2$$

$$= \sum (y_i - \bar{y})^2 + \hat{\beta}_1^2 \sum (x_i - \bar{x})^2 - 2\hat{\beta}_1 \sum (x_i - \bar{x}) \times (y_i - \bar{y}) = S_{yy} + \hat{\beta}_1^2 S_{xx} - 2\hat{\beta}_1 S_{xy}; \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$

$$\therefore E\left(\frac{SS_{Res}}{n-2}\right) = \sigma^2$$

$$SS_{Res} = S_{yy} - \hat{\beta}_1^2 S_{xx}$$

$$= S_{yy} + \hat{\beta}_1^2 S_{xx} - 2\hat{\beta}_1^2 S_{xx}$$

To show? → Proof P.T.O.

[4]

$$\begin{aligned}
 E(S_{yy}) &= E(\sum(y_i - \bar{y})^2) = E[\sum y_i^2] - n E[\bar{y}^2] \\
 &= \sum E[y_i^2] - n E[\bar{y}^2] \\
 &= n\sigma^2 + \sum (\beta_0 + \beta_1 x_i)^2 - \sigma^2 - n(\beta_0 + \beta_1 \bar{x})^2 \\
 &= (n-1)\sigma^2 + \beta_1^2 S_{xx} \\
 \left[ \text{since } y_i = \beta_0 + \beta_1 x_i + \epsilon_i ; \quad E(y_i) = \beta_0 + \beta_1 x_i \right. \\
 &\quad \left. V(y_i) = \sigma^2 \right. \\
 E(y_i^2) &= V(y_i) + [E(y_i)]^2 \\
 &= \sigma^2 + (\beta_0 + \beta_1 x_i)^2 \\
 E(\bar{y}^2) &= V(\bar{y}) + [E(\bar{y})]^2 \\
 &= \frac{\sigma^2}{n} + (\beta_0 + \beta_1 \bar{x})^2
 \end{aligned}$$

$$\begin{aligned}
 E(\hat{\beta}_1^2 S_{xx}) &= S_{xx} E(\hat{\beta}_1^2) \\
 &= \sigma^2 + \beta_1^2 S_{xx}
 \end{aligned}$$

$$\begin{aligned}
 \left[ \quad \begin{aligned}
 E(\hat{\beta}_1) &= \beta_1 \\
 V(\hat{\beta}_1) &= \frac{\sigma^2}{S_{xx}} \\
 E(\hat{\beta}_1^2) &= V(\hat{\beta}_1) + [E(\hat{\beta}_1)]^2 \\
 &= \frac{\sigma^2}{S_{xx}} + \beta_1^2
 \end{aligned} \right]
 \end{aligned}$$

$$\begin{aligned}
 E(SS_{Res}) &= E(S_{yy}) - E(\hat{\beta}_1^2 S_{xx}) \\
 &= (n-1)\sigma^2 + \beta_1^2 S_{xx} - \sigma^2 - \beta_1^2 S_{xx} \\
 &= (n-2)\sigma^2
 \end{aligned}$$

$$\therefore E\left(\frac{SS_{Res}}{n-2}\right) = \sigma^2. \quad \text{thus, the UE of } \sigma^2 \text{ is } MS_{Res}.$$

$\downarrow$

$$MS_{Res} = \text{Residual Mean Square}$$

→ What is the distribution of MS residuals?

$$SS_{Res} = \sum_{i=1}^n e_i^2$$

$$e_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

$$e_i \sim N(0, \sigma^2)$$

$$\frac{e_i}{\sigma} \sim N(0, 1)$$

$$\frac{e_i^2}{\sigma^2} \sim \chi^2_1$$

$$e_i = \hat{e}_i \quad \text{indiv.} \quad e_i \sim N(0, \sigma^2)$$

$$E(e_i) = 0 \quad y_i \sim N(\beta_0 + \beta_1 x_i + \sigma^2)$$

$$V(e_i) = \sigma^2$$

$\hat{\beta}_0$  &  $\hat{\beta}_1$  are LSE of  $\beta_0$  &  $\beta_1$ , respectively.

$e_i = y_i - \hat{y}_i$  satisfy

$$e_1 + e_2 + \dots + e_n = 0 \quad \text{(A)}$$

$$e_1 x_1 + e_2 x_2 + \dots + e_n x_n = 0 \quad \text{(B)}$$

direct from Normal equations.

There are  $(n-2)$  degrees of freedom of residuals.

$$\frac{SS_{Res}}{\sigma^2} = \frac{\sum e_i^2}{\sigma^2} \sim \chi^2_{n-2}$$

$$\frac{(n-2) MS_{Res}}{\sigma^2} \sim \chi^2_{n-2}$$

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

$$\frac{SS_{Res}}{\sigma^2} \sim \chi^2_{n-2}$$

Implications: We have the freedom of choosing  $(n-2)$  residuals independently and remaining have to be chosen so that (A) & (B) are satisfied.

### Evaluate Model: Test of Slope Coefficient

Show if there is a linear relationship between  $X$  &  $Y$ ?

$$H_0: \beta_1 = 0 \quad (\text{No linear relationship})$$

$$H_1: \beta_1 \neq 0 \quad (\text{Linear relationship}) \quad [\text{Two sided test}]$$

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x}) y_i}{\sum (x_i - \bar{x})^2} = \sum c_i y_i ; \quad y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$$

Sampling distn. of  $\hat{\beta}_1$ :  $\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{S_{xx}})$ .

Sum of normals is also normal

$$\text{Thus } Z = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\sigma^2}{S_{xx}}}} \sim N(0, 1)$$

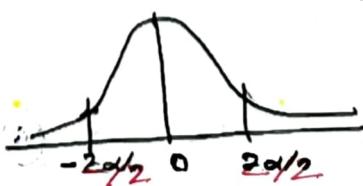
$$E(\hat{\beta}_1) = \beta_1, \quad V(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}$$

(Proved already)

If  $\sigma^2$  is known, we can use  $Z = \frac{\hat{\beta}_1}{\sqrt{\sigma^2/S_{xx}}}$ , under  $H_0: \beta_1 = 0$

to test  $H_0: \beta_1 = 0$ .

Reject  $H_0$  if  $|Z| > Z_{\alpha/2}$ .



Given: UE of  $\sigma^2$  is MSRes.

[6]

Usually  $\sigma^2$  is not known:-

$$\sigma^2 = E(\text{MSRes}) = E\left(\frac{\text{SSRes}}{n-2}\right)$$

, is estimated by this formula.

Test statistic:-  $t = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\beta_1 - \beta_1}{S_{xx}}}}$

$$\frac{\beta_1 - \beta_1}{\sqrt{\sigma^2/S_{xx}}} \sim N(0,1)$$

$$\frac{(n-2) \text{MSRes}}{\sigma^2} \sim \chi^2_{n-2}$$

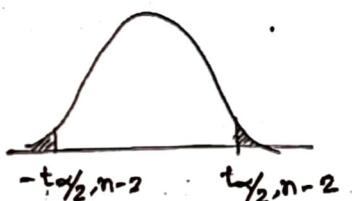
independently

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\sigma^2/S_{xx}}}$$

$\sim t_{n-2}$

$$\frac{\sqrt{(n-2) \text{MSRes}}}{\sqrt{(n-2) \sigma^2}}$$

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\text{MSRes}}{S_{xx}}}} \sim t_{n-2}$$



Test statistic,  $t = \frac{\hat{\beta}_1}{\sqrt{\frac{\text{MSRes}}{S_{xx}}}}$

under  $H_0: \beta_1 = 0$

We reject  $H_0: \beta_1 = 0$  if  $|t| > t_{\alpha/2, n-2}$  (two sided test since  $H_0: \beta_1 = 0$  vs.  $H_1: \beta_1 \neq 0$ )

### ANOVA

Total variation in data =  $\sum_{i=1}^n (y_i - \bar{y})^2$

How much of the variation is explained by the model?

$$Y_i - \bar{Y} = (\hat{Y}_i - \bar{Y}) + (Y_i - \hat{Y}_i)$$

(deviation of the  $i$ th obsrv. from the pred. value)

(deviation of the  $i$ th pred. value from overall mean)

$$\sum (Y_i - \bar{Y})^2 = \sum (\hat{Y}_i - \bar{Y})^2 + \sum (Y_i - \hat{Y}_i)^2 + 2 \sum (\hat{Y}_i - \bar{Y})(Y_i - \hat{Y}_i)$$

$$CPT = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})(Y_i - \hat{Y}_i) = \text{Cross-product term} = 0$$

deviation of  $i$ th observation from the predicted value

$$= \hat{\beta}_1 S_{xy} - \hat{\beta}_1^2 S_{xx}$$

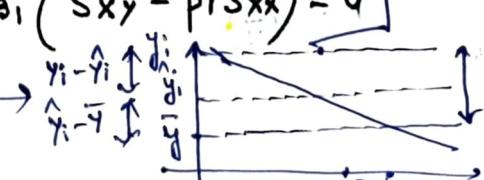
$$= \hat{\beta}_1 (S_{xy} - \hat{\beta}_1 S_{xx}) = 0$$

$$\hat{Y}_i - \bar{Y} = \hat{\beta}_1 (X_i - \bar{x})$$

$$Y_i - \hat{Y}_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)$$

$$= (Y_i - \bar{Y}) - \hat{\beta}_1 (X_i - \bar{x})$$

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



(Pictorial explanation)

$$\left\{ \begin{array}{l} \sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2 \\ \downarrow \quad \quad \quad \downarrow \quad \quad \quad \downarrow \\ S_{ST} = S_{Reg} + S_{Res} \end{array} \right.$$

$$S_{Regression} = \sum (\hat{y}_i - \bar{y})^2 = \sum \hat{\beta}_1^2 (x_i - \bar{x})^2 = \hat{\beta}_1^2 S_{xx}$$

$$S_{Residual} = \sum (y_i - \hat{y}_i)^2 = \sum e_i^2 \sim \chi^2_{n-2}$$

since  $\sum e_i = 0, \sum x_i e_i = 0$  so,  $e_i$ 's are not indep. ( $e_i$ 's are indep.) [(n-2)  $e_i$ 's are indep. but not the remaining d.f. = n-2]

$$S_{ST} = \sum (y_i - \bar{y})^2 \text{ has DF } (n-1)$$

$$\text{since } \sum_{i=1}^n (y_i - \bar{y}) = 0.$$

$$S_{ST} = S_{Reg} + S_{Res}$$

$$DF_T = DF_{Reg} + DF_{Res} \Rightarrow (n-1) = 1 + (n-2)$$

### ANOVA Table:-

Source of Variation	DF	SS	MS	F
Regression	1	$S_{Reg}$	$MS_{Reg} = \frac{S_{Reg}}{1}$	$F = \frac{MS_{Reg}}{MS_{Res}}$
Residual	$n-2$	$S_{Res}$	$MS_{Res} = \frac{S_{Res}}{n-2}$	
Total	$n-1$	$S_{ST}$		

$$E(MS_{Res}) = \sigma^2$$

$$E(MS_{Reg}) = \sigma^2 + \hat{\beta}_1^2 S_{xx}$$

$$\frac{(n-2) MS_{Res}}{\sigma^2} \sim \chi^2_{n-2} \quad \text{under H}_0: \hat{\beta}_1 = 0$$

$$\frac{MS_{Reg}}{\sigma^2} \sim \chi^2_1$$

$$F = \frac{MS_{Reg}}{MS_{Res}} \sim F_{1, n-2}$$

To test  $H_0: \beta_1 = 0$

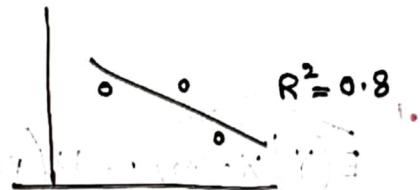
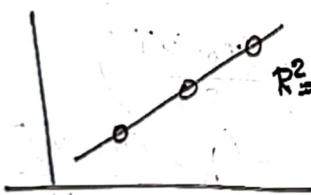
We compute  $F$  and reject  $H_0$  if  
 $F > F_{\alpha/2, 1, n-2}$

### Coefficient of Determination:-

$$R^2 = \frac{SS_{Reg}}{SS_T} = 1 - \frac{SS_{Res}}{SS_T}$$

$$0 \leq R^2 \leq 1$$

$R^2 = 1$ ,  $SS_{Reg} = SS_T$ , i.e.,  $SS_{Res} = 0$ , if the fitted model explains all the variability in  $y$



$$\underline{R^2 = 0}, \quad SS_{Res} = SS_T \Rightarrow \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n (Y_i - \bar{Y})^2$$

$$\Rightarrow \hat{Y}_i = \bar{Y} \quad \begin{array}{c} y \\ (X_i, Y_i) \\ \hat{y} = \bar{y} \end{array}$$

fitted model:  $\hat{Y} = \bar{Y}$   
i.e., when there is no relationship

between  $y$  and  $x$ . (i.e.,  $\beta_1 = 0$ )

### Confidence Interval for $\beta_1$ :

$$\text{LSE of } \beta_1 \text{ is } \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \quad E(\hat{\beta}_1) = \beta_1 \text{ and } V(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}$$

$$= \sum c_i Y_i$$

$$\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{S_{xx}})$$

$$\text{so, } \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\sigma^2}{S_{xx}}}} \sim N(0, 1)$$

for  $\sigma^2$  unknown

[By CLT]

$$t = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{MS_{Res}}{S_{xx}}}} \sim t_{n-2}$$

$$\therefore P \left\{ -t_{\alpha/2, n-2} \leq \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{MS_{Res}}{S_{xx}}}} \leq t_{\alpha/2, n-2} \right\} = 1 - \alpha; \quad \alpha = 0.05$$

Let  $X \sim X_m^2 > \text{ind}$   
 $Y \sim X_n^2 > \text{ind}$

Then  $\frac{X/m}{Y/n} \sim F_{m,n}$

" $R^2$ "; A measure of  
"Goodness of Fit":-

$R^2$  measures the proportion of variability in response variable that is explained by the regression model.

9

100(1- $\alpha$ )% CI for  $\beta_1$  is

$$\hat{\beta}_1 - t_{\alpha/2, n-2} \sqrt{\frac{MSE_{Res}}{S_{xx}}} \leq \beta_1 \leq \hat{\beta}_1 + t_{\alpha/2, n-2} \sqrt{\frac{MSE_{Res}}{S_{xx}}}$$

Interval Estimation of Mean Response  $E(Y)$  for given  $X = x_0$

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

$$E(Y|X=x_0) = \beta_0 + \beta_1 x_0$$

An unbiased estimator of  $E(Y|X=x_0)$  is

$$\hat{E}(Y|X=x_0) = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

$$\sqrt{(\hat{\beta}_0 + \hat{\beta}_1 x_0)} = \sqrt{(\bar{y} + \hat{\beta}_1(x_0 - \bar{x}))} : \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$= \sqrt{(\bar{y})} + \sqrt{(\hat{\beta}_1(x_0 - \bar{x}))} + 2 \underbrace{\text{cov}(\bar{y}, \hat{\beta}_1(x_0 - \bar{x}))}_{=0}$$

$$= \frac{\sigma^2}{n} + (x_0 - \bar{x})^2 \frac{\sigma^2}{S_{xx}}$$

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

$$\hat{E}(Y|X=x_0) \sim N\left(\beta_0 + \beta_1 x_0, \sigma^2 \left( \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right)\right)$$

$$\hat{E}(Y|X=x_0) - E(Y|X=x_0)$$

$$\sim t_{n-2}$$

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

100(1- $\alpha$ )% CI on  $E(Y|X=x_0)$  is

$$\left[ \hat{E}(Y|x_0) \pm t_{\alpha/2, n-2} \sqrt{MSE_{Res} \left( \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right)} \right]$$

CI is minimum at  $x = x_0$ . This widens as  $|x_0 - \bar{x}|$  increases.

Prediction of new observation: (Prediction interval for  $y$ )

$y_0$  corresponds to a specific value of regression  $x = x_0$

$$y_0 = \beta_0 + \beta_1 x_0 + \epsilon \quad E(Y|x_0) = \beta_0 + \beta_1 x_0$$

If  $x = x_0$ , then  $\hat{\beta}_0 + \hat{\beta}_1 x_0$  is point estimator of the response  $y_0$

$$\psi = y_0 - \hat{y}_0, \quad E(\psi) = 0$$

$$[\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0]$$

$$V(\psi) = V(y_0 - \hat{y}_0) = V(y_0) + V(\hat{y}_0) = \sigma^2 + V(\hat{y}_0)$$

$$\frac{\psi - 0}{V(\psi)} \sim t_{n-2} \quad \therefore V(\psi) = \sigma^2 + \sigma^2 \left( \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right)$$

Thus, 100(1- $\alpha$ )% Prediction interval for  $y_0$  is

$$y_0 - t_{\alpha/2, n-2} \sqrt{MSE_{Res} \left[ 1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right]} \leq y_0 \leq \hat{y}_0 + t_{\alpha/2, n-2} \sqrt{MSE_{Res} \left[ 1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right]}$$

[since  $\sigma^2$  is an unbiased estimator for  $MSE_{Res}$ ]

EXERCISE:

(a) In a reg. model  $y_i = \alpha + \beta x_i + \epsilon_i$  if the sample mean  $\bar{x} = 0$ , show that  $\text{cov}(\hat{\alpha}, \hat{\beta}) = 0$ .

(b) Let  $e_i$  be the residuals in the least squares fit of  $y_i$  against  $x_i$  ( $i=1, 2, \dots, n$ ). S.T.  $\sum_{i=1}^n x_i e_i = 0$ . What happens to  $\sum_{i=1}^n e_i$ ?

Solution:-

$$\begin{aligned} \text{(a)} \quad \text{cov}(\hat{\alpha}, \hat{\beta}) &= \text{cov}\left(\frac{1}{n} \sum y_i - \hat{\beta} \bar{x}, \frac{\sum (x_i - \bar{x}) y_i}{\sum (x_i - \bar{x})^2}\right) \\ &= \text{cov}\left(\frac{1}{n} \sum y_i, \frac{\sum (x_i - \bar{x}) y_i}{\sum (x_i - \bar{x})^2}\right) = \text{cov}(\bar{y}, \hat{\beta}) \\ &= \frac{\sum (x_i - \bar{x}) v(\bar{y})}{n \sum (x_i - \bar{x})^2} = \frac{\sum (x_i - \bar{x}) \sigma^2}{n \sum (x_i - \bar{x})^2} = 0. \end{aligned}$$

$$\text{(b)} \quad y_i = \alpha + \beta x_i + \epsilon_i$$

Given  $e_i$  is the residuals in the least square fit.

LS method determines  $\alpha, \beta$  by minimizing

$$SS_{Rao} = \sum e_i^2 = \sum (y_i - \hat{y}_i)^2 = \sum (y_i - \hat{\alpha} - \hat{\beta} x_i)^2$$

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{\sum (x_i - \bar{x}) y_i}{\sum (x_i - \bar{x})^2}, \quad \hat{\alpha} = \bar{y} - \hat{\beta} \bar{x}. \quad \textcircled{1}$$

Properties of least square fit:

$$\text{(i)} \quad \sum e_i = 0 = \sum (y_i - \hat{y}_i) \quad [\because \sum y_i = \sum \hat{y}_i]$$

or, if you minimize normal equation

$$-2 \sum (y_i - \hat{\alpha} - \hat{\beta} x_i) = 0 \Rightarrow \sum e_i = 0.$$

$$\text{(ii)} \quad \sum \alpha_i e_i = 0$$

$$\text{Proof:-} \quad -2 \sum (y_i - \hat{\alpha} - \hat{\beta} x_i) x_i = 0$$

$$\Rightarrow \sum (\hat{y}_i - \hat{y}_i) x_i = 0$$

$$\Rightarrow \sum e_i x_i = 0.$$

INTERESTING RESULT:

$$\hat{\beta} = \text{MLE} = \text{Least square estimates}$$

THEOREM: Maximum Likelihood Estimators (MLEs) for the simple linear regression model parameters are the least square estimators (assuming that the model errors are independently and identically distributed).

PROOF:  $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ ;  $\epsilon_i \stackrel{iid}{\sim} N(0, \sigma^2)$ ; density of  $\epsilon_i$  is

$$f(\epsilon_i) = \frac{1}{\sigma \sqrt{2\pi}} \exp\left[-\frac{1}{2} \cdot \frac{(\epsilon_i - 0)^2}{\sigma^2}\right]$$

$$\text{Likelihood function: } f(\epsilon_1, \dots, \epsilon_n) = \prod_{i=1}^n f(\epsilon_i)$$

$$L = \prod_{i=1}^n \frac{1}{\sigma \sqrt{2\pi}} \exp\left[-\frac{1}{2} \cdot \frac{\epsilon_i^2}{\sigma^2}\right]$$

$$= \left(\frac{1}{\sigma \sqrt{2\pi}}\right)^n \exp\left[-\frac{1}{2} \sum_{i=1}^n \frac{\epsilon_i^2}{\sigma^2}\right]$$

$$= \left(\frac{1}{\sigma \sqrt{2\pi}}\right)^n \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2\right]$$

Objective: Maximize L and estimate  $\beta_0$ ,  $\beta_1$ , and  $\sigma^2$ .

Since  $\log$  is a monotonic function, so taking  $\log L = L^*$  for calculation.

$$L^* = \log L = -\frac{n}{2} \log 2\pi - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$$

$$\frac{\partial L^*}{\partial \beta_0} = -\frac{1}{\sigma^2} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) = 0 \quad (i)$$

$$\frac{\partial L^*}{\partial \beta_1} = -\frac{1}{\sigma^2} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) x_i = 0 \quad (ii)$$

$$\frac{\partial L^*}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2 = 0 \quad (iii)$$

(i) gives  $\tilde{\beta}_0 = \bar{y} - \bar{\beta}_1 \bar{x}$ . similar to OLS estimates.

$$(ii) \text{ gives } \tilde{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

$$(iii) \text{ gives } \tilde{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_i)^2 \quad \text{different from OLS estimate of } \hat{\sigma}^2$$

- Different estimation method may produce different results in parameter estimation of linear regression.  $[\hat{\sigma}^2 = E(\frac{SS_{Res}}{n-2})]$

# Simple Linear Regression

EXAMPLE

company.

Example-1.

You are marketing analyst for Chocolate company. Data is as follows:

Advertisement (\$ in 1000')	(spending cost)	controlled ( $X_i$ )
1		
2		
3		
4		
5		

Sales (units) in 100  
 ↓  
 (dependent)  
 ( $y_i$ )

1
1
2
2
4

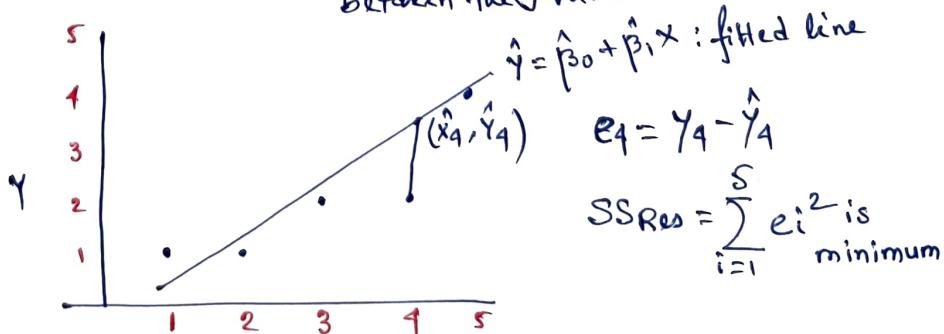
NOTES:  $\beta_0, \beta_1$  are in sales/spending cost and the unit of  $E$  are sales (units) and it has s.d.  $\sigma$ .

- What is the relationship between sales and advertisement?
- Is the relationship significant at the 5% level? ( $\alpha = 0.05$ )
- Build the ANOVA table to test  $H_0: \beta_1 = 0$  vs.  $H_1: \beta_1 \neq 0$ .
- calculate  $R^2$  and comment on the fit.
- find Confidence Interval (CI) for  $\beta_1$  where  $(y_i = \beta_0 + \beta_1 x_i + e_i)$
- Estimate mean sales amount when advertisement cost is \$4 at the 0.05 level.

SOLUTIONS:

(a)

Scatter plot: A scatter plot is a mathematical diagram to display values of two variables for a set of data. It is used to investigate the possible relationship between the variables.

Table:

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

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i - n \bar{x} \bar{y}}{\sum_{i=1}^n x_i^2 - n \bar{x}^2}$$

$$= \frac{37 - 5 \times 3 \times 2}{55 - 5 \times 3^2} = 0.70$$

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

$$\therefore \hat{y} = -0.10 + 0.7x$$

- Interpretation:
- ① Sales volume ( $Y$ ) is expected to increase by 0.7 units for each \$1 increase in Advertising ( $X$ ).
  - ② Average value of sales volume ( $Y$ ) is -0.10 units when advertising ( $X$ ) is 0.  
(Difficult to explain to manager since one expects some sales without advertising)

(b)

Adv. ( $X$ )	$Y$ (Sales)	$\hat{Y} = -0.10 + 0.7X$	$e_i = Y_i - \hat{Y}_i$
1	1	0.6	0.4
2	1	1.3	-0.3
3	2	2	0
4	2	2.7	-0.7
5	4	3.4	0.6

$$H_0: \beta_1 = 0 \text{ vs. } H_1: \beta_1 \neq 0 \quad (\alpha = 0.05) \quad (df = 5 - 2 = 3)$$

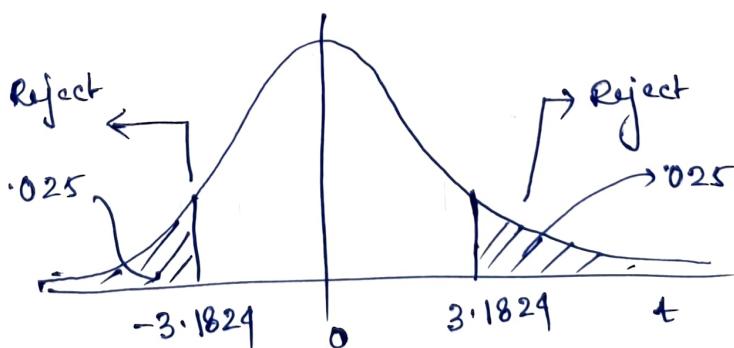
$$SS_{Res} = \sum e_i^2 = (0.4)^2 + (-0.3)^2 + 0^2 + (-0.7)^2 + (0.6)^2 = 1.1.$$

$$MS_{Res} = \frac{SS_{Res}}{n-2} = \frac{1.1}{3} = 0.3666$$

Test statistic:  $t = \frac{\hat{\beta}_1}{\sqrt{\frac{MS_{Res}}{S_{xx}}}} = \frac{0.7}{\sqrt{\frac{0.3666}{10}}} = 3.656$  [since  $S_{xx} = \sum x_i^2 - n\bar{x}^2 = 55 - 5 \times 3^2 = 10$ ]

$$t_{0.05/2, 3} = 3.18 \quad (\text{from statistical table})$$

Since  $t_{\text{cal}} > t_{\text{table}}$ , then  $H_0: \beta_1 = 0$  is rejected at 5% level of significance.



Decision: Reject  $H_0$  at  $\alpha = 0.05$

Conclusion: There is evidence of a linear relationship between  $X$  and  $Y$ .

[1]

$$(c) \quad SS_{Res} = 1.1 ; \quad SST = \sum_{i=1}^5 (y_i - \bar{y})^2 = \sum (y_i - 2)^2 = 6$$

$$SS_{Reg} = \hat{\beta}_1^2 S_{xx} = (0.7)^2 \times 10 = 4.9$$

Source of Variation	DF	SS	MS	F <sub>cal</sub>
Reg	1	4.9	4.9	$\frac{4.9}{0.367} = 13.61$
Res	3	1.1	$\frac{1.1}{3} = 0.367$	
Total	4	6		

$$F \sim F_{0.5, 1, 3} = 10.13$$

$F_{cal} > F_{tabulated}$ ; thus  $H_0$  is rejected.

Thus, t-test and F-test results are the same.  
For simple linear regression, t-test and F-test are same.

NOTE:

$$F = t^2$$

$$F_{tab} = 13.61 = (3.65)^2$$

$$H_0: \beta_1 = 0$$

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

$$t_{n-2}^2 = \frac{\hat{\beta}_1^2 S_{xx}}{MS_{Res}} = \frac{MS_{Reg}}{MS_{Res}}$$

$$= F_{1, n-2}.$$

(d)  $R^2 = \frac{4.9}{6} = 0.82$

Comment: 82% of the total variability in sales amount can be explained by the amount of money spent on advertisement. There could be some other variable which may be useful to explain the rest of the variability (say, number of sales person) which is the area of MLR.

(e)  $\hat{\beta}_1 - t_{\alpha/2, n-2} \sqrt{\frac{MS_{Res}}{S_{xx}}} \leq \beta_1 \leq \hat{\beta}_1 + t_{\alpha/2, n-2} \sqrt{\frac{MS_{Res}}{S_{xx}}};$

$$\Rightarrow \beta_1 \in \left( 0.7 - \frac{3.182}{(from table)} \times \sqrt{\frac{0.367}{10}}, 0.7 + 3.182 \times \sqrt{\frac{0.367}{10}} \right)$$

$$\Rightarrow \beta_1 \in (0.1, 1.3); i.e., P[0.1 \leq \beta_1 \leq 1.3] = 0.95.$$

$$(f) \quad \widehat{E(Y|x_0)} + t_{\alpha/2, n-2} \sqrt{MS_{Res} \left( \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right)}$$

$$\downarrow \quad \downarrow$$

$$\begin{aligned} & \hat{\beta}_0 + \hat{\beta}_1 x_0 \\ & = -0.1 + 0.7 \times 9 \\ & = 2.7 \end{aligned}$$

$$\begin{aligned} & t_{0.025, 3} \\ & = 3.182 \end{aligned}$$

$$\begin{aligned} & \sqrt{0.367 \left( \frac{1}{5} + \frac{(1-3)^2}{10} \right)} \\ & = 0.33 \end{aligned}$$

$$\text{Then, } P[1.65 \leq E(Y|x_0=4) \leq 3.75] = 0.95.$$

Example-2: Consider the simple linear regression model:

$$Y = \beta_0 + \beta_1 x + \epsilon ; \text{ where } \beta_0 \text{ is known.} \\ (\text{i.e., value of } \beta_0 \text{ is given})$$

(a) Find the LSE of  $\beta_1$  for this model.

(b) Find the  $100(1-\alpha)\%$  CI for  $\beta_1$ . [Ans:  $\hat{\beta}_1 \pm t_{\alpha/2, n-1} \sqrt{\frac{MS_{Res}}{\sum x_i^2}}$ ]

Solution: (a) Fitted model:  $\hat{y} = \beta_0 + \hat{\beta}_1 x$

$$\begin{aligned} \text{residual errors: } e_i &= y_i - \hat{y}_i = (y_i - \beta_0 - \hat{\beta}_1 x_i) \\ S = SS_{Res} &= \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \beta_0 - \hat{\beta}_1 x_i)^2 \end{aligned}$$

$$\frac{\partial S}{\partial \hat{\beta}_1} = 0 \Rightarrow \sum_{i=1}^n 2(y_i - \beta_0 - \hat{\beta}_1 x_i) x_i = 0 \quad (*)$$

Here, we get only one normal equation since we have only one unknown parameter. Solving (\*), we get the LSE of  $\beta_1$ .

From (\*), we have:  $\hat{\beta}_1 = \frac{\sum (y_i - \beta_0) x_i}{\sum x_i^2}$  and since  $\beta_0$  is known

thus, we do not need to estimate  $\beta_0$ .

$$(b) \quad E(\hat{\beta}_1) = \frac{E(\sum (y_i - \beta_0) x_i)}{\sum x_i^2} \quad [\text{Note that } y_i = \beta_0 + \beta_1 x_i + \epsilon_i]$$

$$\therefore E(\hat{\beta}_1) = \frac{E(\sum (\beta_1 x_i + \epsilon_i) x_i)}{\sum x_i^2} = \frac{\beta_1 \sum x_i^2}{\sum x_i^2} = \beta_1 \quad [\because E(\epsilon_i) = 0]$$

$$\text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum x_i^2}, \quad \text{thus, } \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{MS_{Res}}{\sum x_i^2}}} \sim t_{n-1}.$$

i.e.,  $\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{\sum x_i^2})$ .

When  $\beta_0, \beta_1$  both are unknown,  
it's  $t_{n-2}$  due to two  
normal equations]

Since DF of  $SS_{Res}$  is  $n-1$  since we have only one restriction on  $e$ , i.e.,  $\sum e_i x_i = 0$  (in Eqn. (\*) of normal equation).