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 variable. This technique is widely used for prediction and forecasting. **Applications:-** Economics, Management, Life & Biological Science, 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: $y = \beta_0 + \beta_1 x + \epsilon$

- y: Response variable (random variable)
- β_0: intercept
- β_1: slope
- x: regression variable (controlled variable)
- ϵ: random error component

We now make some basic assumption on the model

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

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

Assumptions:-

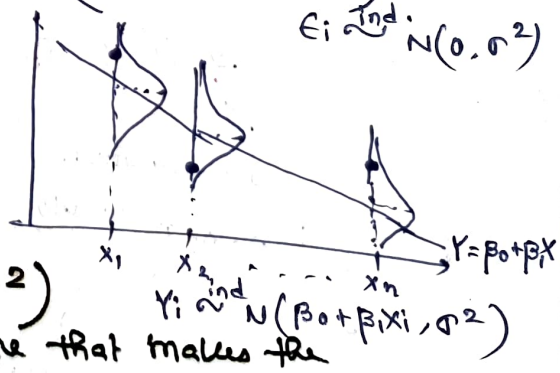
1. ϵ_i is an RV with mean '0' and s.d. σ (unknown). i.e., $E(\epsilon_i) = 0$, $V(\epsilon_i) = \sigma^2$ (Homoscedasticity)
2. $Cov(\epsilon_i, \epsilon_j) = 0 \Rightarrow \epsilon_i$ & ϵ_j are uncorrelated. i.e. (no autocorrelation.)
3. $\epsilon_i \overset{ind.}{\sim} N(0, \sigma^2)$.

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

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

$$\epsilon_i \overset{indep.}{\sim} N(0, \sigma^2)$$

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



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 β_0 & β_1 so that the sum of squares of all the diff. between the observation y_i and the fitted line is minimum.

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

$\hat{\beta}_0$ is the estimate of β_0 and $\hat{\beta}_1$ is the estimate of β_1.

The least square estimators of β_0 and β_1 (i.e., $\hat{\beta}_0, \hat{\beta}_1$) must satisfy

$$\left.\frac{\partial S}{\partial \beta_0}\right|_{\hat{\beta}_0, \hat{\beta}_1} = -2\sum(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

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

} Normal equations

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

Solving these two equations :-

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

$$\Rightarrow n\hat{\beta}_0 = \sum y_i - \hat{\beta}_1 \sum x_i$$

[where $\bar{x} = \frac{\sum x_i}{n}$, $\bar{y} = \frac{\sum y_i}{n}$]

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

Properties of LS fit:

(Sum of residuals in a reg. model = 0) $\rightarrow \sum e_i = 0$ — (A)

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

(Sum of product of residual with regressor = 0) $\rightarrow \sum e_i x_i = 0$ — (B)

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

$$S_{xy} = \sum (y_i - \bar{y})(x_i - \bar{x})$$

$$S_{xx} = \sum (x_i - \bar{x})^2$$

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

Note: $\sum e_i \hat{y}_i = 0$ but $\sum e_i y_i \neq 0$.

[Since $\sum (y_i - \bar{y}) \bar{x} = \bar{x} \sum y_i - \sum \bar{x}\bar{y} = n\bar{x}\bar{y} - n\bar{x}\bar{y} = 0$.]

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

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

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

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

[Since $\sum (x_i - \bar{x}) \bar{y} = 0$.]

fitted model of observation y_i: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

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

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

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

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

Statistical properties of LS estimators: 1.

Ans: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$

$\bar{y} = \beta_0 + \beta_1 \bar{x} + \bar{\epsilon}$; where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\bar{\epsilon} = \frac{1}{n}\sum \epsilon_i$

$\hat{\beta}_1$ is a UE of β_1, $\hat{\beta}_0$ is a UE of β_0.

(Unbised estimator = UE)

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

(Continued)

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

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

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

Since $E(\hat{\beta}_1) = \beta_1$

Q.E. $\hat{\beta}_0$ is an UE of β_0.

$$E(\hat{\beta}_0) = E(\bar{y} - \hat{\beta}_1\bar{x}) = E(\beta_0 + \beta_1\bar{x} - \hat{\beta}_1\bar{x}) = \beta_0 + \beta_1\bar{x} - \beta_1\bar{x} = \beta_0$$ (Proved)

... Thus, $\hat{\beta}_1$ and $\hat{\beta}_0$ are unbiased.

$$V(\hat{\beta}_1) = V\left(\frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2}\right) = V\left(\frac{\Sigma(x_i - \bar{x})y_i}{\Sigma(x_i - \bar{x})^2}\right) = V\left(\sum_{i=1}^{n} c_i y_i\right)$$

where, $c_i = \dfrac{(x_i - \bar{x})}{\Sigma(x_i - \bar{x})^2}$

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

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

$$V\left(\frac{1}{n}\Sigma y_i\right) = \frac{\Sigma V(y_i)}{n^2} = \frac{\Sigma\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

$$= \frac{\sigma^2}{n} + \bar{x}^2 V(\hat{\beta}_1) - 2\bar{x} \times 0$$

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

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

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

$$\left. = 0. \right]$$

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

Estimation of σ^2:-

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

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

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

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

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

$$= S_{yy} + \hat{\beta}_1^2 S_{xx} - 2\hat{\beta}_1 S_{xy} \quad ; \quad \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$

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

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

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

To show? → Proof P.T.O.

$$E(S_{yy}) = E\left(\sum (y_i - \bar{y})^2\right) = E\left[\sum y_i^2\right] - nE\left[\bar{y}^2\right]$$

$$= \sum E\left[y_i^2\right] - nE\left[\bar{y}^2\right]$$

$$= \sum E(y_i^2) - nE(\bar{y}^2)$$

$$= n\sigma^2 + \sum (\beta_0 + \beta_1 x_i)^2 - \sigma^2 - n(\beta_0 + \beta_1 \bar{x})^2$$

$$= (n-1)\sigma^2 + \beta_1^2 S_{xx}$$

[since $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$; $E(y_i) = \beta_0 + \beta_1 x_i$

$V(y_i) = \sigma^2$

$E(y_i^2) = V(y_i) + [E(y_i)]^2$

$= \sigma^2 + (\beta_0 + \beta_1 x_i)^2$

$E(\bar{y}^2) = V(\bar{y}) + [E(\bar{y})]^2$

$= \frac{\sigma^2}{n} + (\beta_0 + \beta_1 \bar{x})^2$]

$$E(\hat{\beta}_1^2 S_{xx}) = S_{xx} E(\hat{\beta}_1^2)$$

$$= \sigma^2 + \beta_1^2 S_{xx}$$

[$E(\hat{\beta}_1) = \beta_1$

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

$E(\hat{\beta}_1^2) = V(\hat{\beta}_1) + [E(\hat{\beta}_1)]^2$

$= \frac{\sigma^2}{S_{xx}} + \beta_1^2$]

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

$$= (n-1)\sigma^2 + \beta_1^2 S_{xx} - \sigma^2 - \beta_1^2 S_{xx}$$

$$= (n-2)\sigma^2$$

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

$= MS_{Res}$ = Residual Mean Square

→ What is the distribution of MS residuals?

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

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

$e_i = \hat{\epsilon}_i$ indet. $\epsilon_i \sim N(0, \sigma^2)$

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

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

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

$$\frac{e_i^2}{\sigma^2} \sim \chi_1^2$$

$\hat{\beta}_0$ & $\hat{\beta}_1$ are LSE of β_0 & β_1 respectively.

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

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

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

} direct from Normal equations.

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

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

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

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

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

$$\Rightarrow \boxed{\frac{SS_{Resi}}{\sigma^2} \sim \chi^2_{n-2}}.$$

Evaluate Model: Test of Slope Coefficient

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

$H_0 : \beta_1 = 0$ (No linear relationship) → [means $y_i = \beta_0 + \epsilon$, x plays no role]

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

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

[Sum of normals is also normals]

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

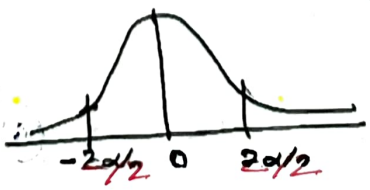
[$E(\hat{\beta}_1) = \beta_1$, $V(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}$ (Proved already)]

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

If σ^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 σ^2 is MS_{Res}.

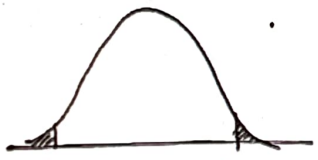
Usually σ^2 is not known:- $\sigma^2 = E(MS_{Res}) = E\left(\frac{SS_{Res}}{n-2}\right)$ → is estimated by ↓ this formula.

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

$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\sigma^2/S_{xx}}} \sim N(0,1)$ and $\frac{(n-2)MS_{Res}}{\sigma^2} \sim \chi^2_{n-2}$ — independently

Let $X \sim N(0,1)$, $Y \sim \chi^2_n$ } indep.
Then $\frac{X}{\sqrt{\frac{Y}{n}}} \sim t_n$
(From sampling distribution theory)

So, $\frac{\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\sigma^2/S_{xx}}}}{\sqrt{\frac{(n-2)MS_{Res}}{(n-2)\sigma^2}}} \sim t_{n-2}$



So, $\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{MS_{Res}}{S_{xx}}}} \sim t_{n-2}$.

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

Identity $Y_i - \bar{Y} = (\hat{Y}_i - \bar{Y}) + (Y_i - \hat{Y}_i)$
(deviation of the i^{th} obsn. 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 $= \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})(Y_i - \hat{Y}_i)$ = Cross-product term $= 0$

(deviation of i^{th} observation from the predicted value)

$= \sum \hat{\beta}_1 (x_i - \bar{x})\left[(Y_i - \bar{Y}) - \hat{\beta}_1(x_i - \bar{x})\right]$

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

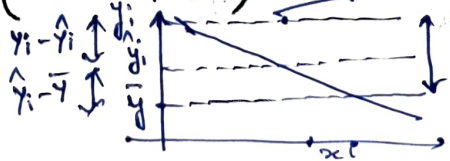
$= \hat{\beta}_1 (S_{xy} - \hat{\beta}_1 S_{xx}) = 0$

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

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

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

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



(Pictorial explanation)

$$\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2$$

$$SS_T = SS_{Reg} + SS_{Res}$$

$$SS_{Regression} = \sum_1^n (\hat{y}_i - \bar{y})^2 = \sum \hat{\beta}_1^2 (x_i - \bar{x})^2 = \hat{\beta}_1^2 S_{xx}.$$

$$SS_{Residual} = \sum_1^n (y_i - \hat{y}_i)^2 = \sum e_i^2 \sim \chi^2_{n-2}$$

since $\sum e_i = 0$, $\sum x_i e_i = 0$ so, e_i's are not indep. (ε_i's are indep)

[$(n-2)$ e_i's are indep. but not the remaining d.f. $= n-2$]

$$SS_T = \sum_1^n (y_i - \bar{y})^2 \text{ has DF } (n-1)$$

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

$$SS_T = SS_{Reg} + SS_{Res}$$

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

ANOVA Table:-

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

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

$$E(MS_{Reg}) = \sigma^2 + \beta_1^2 S_{xx}$$

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

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

$\Big\rangle$ indep under $H_0: \beta_1 = 0$

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

To test $H_0: \beta_1 = 0$

We compute F and reject H_0 if

$$F > F_{\alpha;1,n-2}$$

Let $X \sim \chi^2_m$, $Y \sim \chi^2_n$ } indep

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

Coefficient of Determination:-

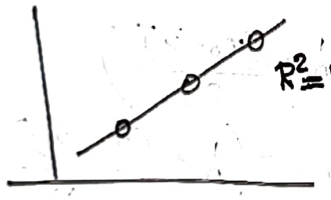
$$R^2 = \frac{SS_{Reg}}{SS_T} = 1 - \frac{SS_{Res}}{SS_T}$$

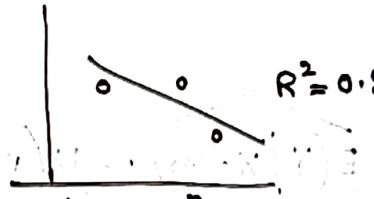
$$0 \le R^2 \le 1$$

'R²': A measure of "Goodness of Fit":-

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

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

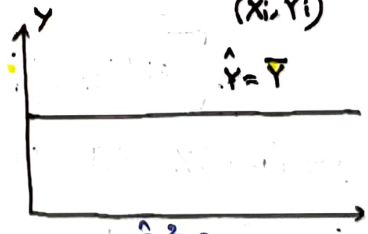




$R^2 = 0$, $SS_{Res} = SS_T \Rightarrow \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$

$\Rightarrow \hat{Y}_i = \bar{Y}$

Fitted model: $\hat{Y} = \bar{Y}$



i.e., when there is no relationship between Y and X. (i.e., $\hat{\beta}_1 = 0$)

[$\because R^2 = \frac{SS_{Reg}}{SS_T} = \frac{\hat{\beta}_1^2 S_{xx}}{SS_T} = 0 \Leftrightarrow \hat{\beta}_1 = 0$]

Confidence Interval for β_1:-

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

LSE of β_1 is $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$

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

$= \sum c_i Y_i$

[$\because$ Since Y_i follows normal, so does $\hat{\beta}_1$]

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

So, $\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\sigma^2}{S_{xx}}}} \sim N(0,1)$ [By CLT]

For σ^2 unknown, $t = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{MS_{Res}}{S_{xx}}}} \sim t_{n-2}$.

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

$100(1-\alpha)\%$ CI for β_1 is

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

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

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

$$V(\hat{\beta}_0 + \hat{\beta}_1 x_0) = V(\bar{y} + \hat{\beta}_1(x_0 - \bar{x})) \quad ; \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$$

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

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

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

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

$$\frac{\widehat{E(Y|X=x_0)} - E(Y|x=x_0)}{\sqrt{MS_{Res}\left(\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}\right)}} \sim t_{n-2}$$

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

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

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 regressor $x = x_0$

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

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

$$\psi = y_0 - \hat{y}_0 \quad , \quad E(\psi) = 0 \qquad [\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0]$$

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

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

$$\frac{\psi - 0}{V(\psi)} \sim t_{n-2}.$$

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

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

[since σ^2 is an unbiased estimator for MS_{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 $(i = 1, 2, \ldots, n)$. S.T. $\sum_{i=1}^{n} x_i e_i = 0$. What happens to $\sum_{i=1}^{n} e_i$?

Solution:-

(a) $$\text{Cov}(\hat{\alpha}, \hat{\beta}) = \text{cov}\left(\frac{1}{n}\Sigma y_i - \hat{\beta}\bar{x}, \frac{\Sigma(x_i - \bar{x})y_i}{\Sigma(x_i - \bar{x})^2} \right)$$

$$= \text{cov}\left(\frac{1}{n}\Sigma y_i, \frac{\Sigma(x_i - \bar{x})y_i}{\Sigma(x_i - \bar{x})^2} \right) = \text{cov}(\bar{y}, \hat{\beta})$$

$$= \frac{\Sigma(x_i - \bar{x})V(y_i)}{n\Sigma(x_i - \bar{x})^2} = \frac{\Sigma(x_i - \bar{x})\sigma^2}{n\Sigma(x_i - \bar{x})^2} = 0.$$

(b) $y_i = \alpha + \beta x_i + \epsilon_i$

Given e_i is the residuals in the least square fit.

LS method determines α, β by minimizing

$$SS_{Res} = \Sigma e_i^2 = \Sigma (y_i - \hat{y}_i)^2 = \Sigma(y_i - \hat{\alpha} - \hat{\beta}x_i)^2 \quad \text{---(*)}$$

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

Properties of least square fit:

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

or, if you minimize normal equation

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

(ii) $\Sigma x_i e_i = 0$

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

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

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

Method of Parameter Estimation: MAXIMUM LIKELIHOOD ESTIMATORS

Likelihood Function: The probability (likelihood) of the observed sample given the parameter. The likelihood function is a function of the parameter. Suppose θ is the unknown parameter. We write the likelihood function as $L(\theta | x_1, x_2, \ldots, x_n)$.

Note: Likelihood function is not probability. If we sum (or integrate) $L(\theta | x_1, \ldots, x_n)$ over all possible values of θ, it will not become 1.

Maximum Likelihood Principle: Choose as your estimates those values of the parameter that maximizes likelihood of the observed data.

Log likelihood: Likelihood function is $L(\theta | \underset{\sim}{x}) = \prod_{i=1}^{n} p(x_i, \theta)$. The natural logarithm of the likelihood function. It is often preferable to work with the log likelihood for both practical and theoretical reason. The log likelihood converts the product into sum and hence it's easier to handle. If we take logarithm, it results in a large number (since product of probabilities is a tiny value) and $\log L(\theta | \underset{\sim}{x})$ is always negative.

Implication: A likelihood method is a measure of how well a particular model fits a data. They explain how well a parameter explains the observed data.

Advantages of log likelihood: Loglikelihood increase the numerical stability of the estimates. Likelihood functions are product of marginal probabilities are tend to become very small for large samples. Log likelihoods are large negative numbers and hence their usage improves numerical stability.

Kernel Likelihood: $L(\theta | x) = K(\underset{\sim}{x}) \, p(\theta | \underset{\sim}{x})$; where $K(\underset{\sim}{x})$ is a function of the observed data and does not involve the parameter to be estimated. **Example:** Suppse $X_i \overset{IND.}{\sim} Pois(\lambda)$

$$L(\lambda | x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i, \lambda) = \prod_{i=1}^{n} \frac{e^{-\lambda} \cdot \lambda^{x_i}}{x_i!} = K(\underset{\sim}{x}) \, p(\underset{\sim}{x} | \lambda)$$

where $K(x_1, \ldots, x_n) = \prod_{i=1}^{n} \frac{1}{x_i!}$ and $p(\theta | x_1, \ldots, x_n) = e^{-n\lambda} \cdot \lambda^{\Sigma x_i}$.

INTERESTING RESULT: $\hat{\beta}$ = MLE = 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: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$; $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$; density of ϵ_i is

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

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

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

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

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

Objective: Maximize L and estimate β_0, β_1 and σ^2.

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

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

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

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

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

} Normal Equations

(i) gives $\tilde{\beta}_0 = \bar{y} - \tilde{\beta}\bar{x}$.

(ii) gives $\tilde{\beta}_1 = \dfrac{\sum_1^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_1^n (x_i - \bar{x})^2}$.

} similar to OLS estimates.

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

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

— Different estimation method may produce different results in parameter estimation of linear regression.

Simple Linear Regression
EXAMPLE

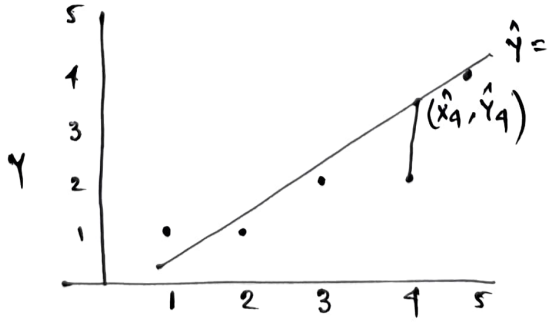
Example-1. You are marketing analyst for Chocolates company. Data is as follows:

Advertisement ($) in 1000' (spending cost) ↓ controlled (X_i)	Sales (units) in 100 ↓ (dependent) (Y_i)
1	1
2	1
3	2
4	2
5	4

(a) What is the relationship between sales and advertisement?
(b) Is the relationship significant at the 5% level? ($\alpha = 0.05$)
(c) Build the ANOVA table to test $H_0: \beta_1 = 0$ vs. $H_1: \beta_1 \neq 0$.
(d) Calculate R^2 and comment on the fit.
(e) Find Confidence Interval (CI) for β_1 where ($Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$)
(f) 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.



$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$: fitted line

$e_4 = Y_4 - \hat{Y}_4$

$SS_{Res} = \sum_{i=1}^{5} e_i^2$ is minimum

Table:

X_i	Y_i	X_i^2	Y_i^2	$X_i Y_i$
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
15	10	55	26	37

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

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

$\therefore$ $\boxed{\hat{Y} = -0.10 + 0.7X}$.

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

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

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

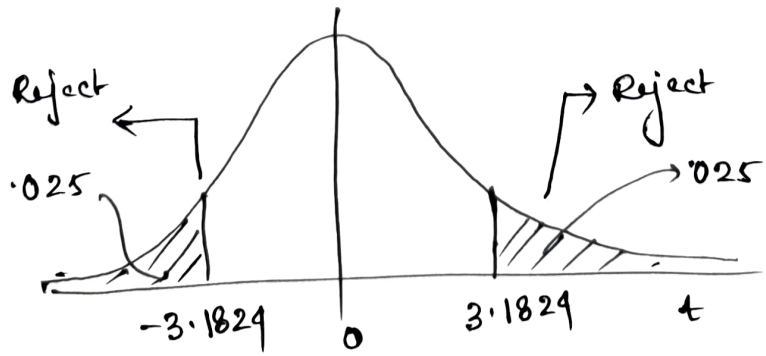
$$MS_{Res} = \frac{SS_{Res}}{n-2} = \frac{1.1}{3} = 0.3666$$

Test statistic:
$$t = \frac{\hat{\beta}_1}{\sqrt{\frac{MS_{Res}}{S_{xx}}}} = \frac{0.7}{\sqrt{\frac{0.3666}{10}}} = 3.655$$

[since $S_{xx} = \sum x_i^2 - n\bar{x}^2 = 55 - 5 \times 3^2 = 10$]

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

Since $t_{cal} > t_{table}$, 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.

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

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

Source of Variation	DF	SS	MS	F_{cal}
Reg	1	4.9	4.9	$4.9/0.367 = 13.61$
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 the same.

NOTE: For simple linear regression, t-test and F-test are same.

$F = t^2$ $\quad F_{tab} = 13.61 = (3.685)^2$

$H_0: \beta_1 = 0$ vs. $H_1: \beta_1 \neq 0$

$$t^2_{n-2} = \frac{\hat{\beta}_1^2 S_{xx}}{MS_{Res}} = \frac{MS_{reg}}{MS_{res}} = F_{1,n-2}.$$

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

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

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

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

$\hat{\beta}_0 + \hat{\beta}_1 x_0$ $\quad$ $t_{0.025,3}$ $\quad$ $\sqrt{0.367\left(\frac{1}{5} + \frac{(4-3)^2}{10}\right)}$

$= -0.1 + 0.7 \times 4$ $\quad$ $= 3.182$ $\quad$ $= 0.33$

$= 2.7$

Thus, $P\left[1.65 \le E(Y|x_0=4) \le 3.75\right] = 0.95.$

Example-2. Consider the simple linear regression model:

$Y = \beta_0 + \beta_1 x + \epsilon$; where β_0 is known. (i.e., value of β_0 is given)

(a) Find the LSE of β_1 for this model.

(b) Find the $100(1-\alpha)\%$ CI for β_1. [Ans: $\hat{\beta}_1 \pm t_{\alpha/2, n-1}\sqrt{\frac{MS_{Res}}{\sum x_i^2}}$]

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

residual error: $e_i = y_i - \hat{y}_i = (y_i - \beta_0 - \hat{\beta}_1 x_i)$

$$S = SS_{Res} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \hat{\beta}_1 x_i)^2$$

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

Here, we get only one normal equation (*), since we have only one unknown parameter. Solving (*), we get the LSE of β_1.

From (*), we have: $\hat{\beta}_1 = \frac{\sum (y_i - \beta_0) x_i}{\sum x_i^2}$ and since β_0 is known

thus, we donot need to estimate β_0.

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

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

[$\hat{\beta}_1$ is unbiased]

$Var(\hat{\beta}_1) = \frac{\sigma^2}{\sum x_i^2}$.

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

Thus, $\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{MS_{Res}}{\sum x_i^2}}} \sim t_{n-1}$.

[When β_0, β_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).