# * Regression Analysis *

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

Y - response / Dependent / study variable

x - independent / regressor / predictor / explanatory variable

$\beta_0, \beta_1$ - unknown Regression coefficients / parameters

$\varepsilon$ - an observable & we can not control it.

## * Cause & Effect

X - cause &

Y - effect

1) X - I.Q. & Y - marks

X - cause & Y - effect

x - effect & Y - cause.

2) X - Rainfall & Y - yield

Cause - X - Rainfall

Effect - Y - Yield

3) X - #. of literates & Y - #. of criminals

As X increases, Y also increases because of third variable z - population. Such correlation is called "Non - sense correlation".

$Y$ - $X_1, X_2, \ldots X_K$

The general regression model is,

$$y = f(X_1, X_2, \ldots X_K, \beta_1, \ldots \beta_K) + \varepsilon$$

$f$ - well defined fn.

$X_1 \ldots X_K$ - Regressors

$\beta_1 \ldots \beta_K$ - Regression coefficients

If $\varepsilon = 0$, then the model

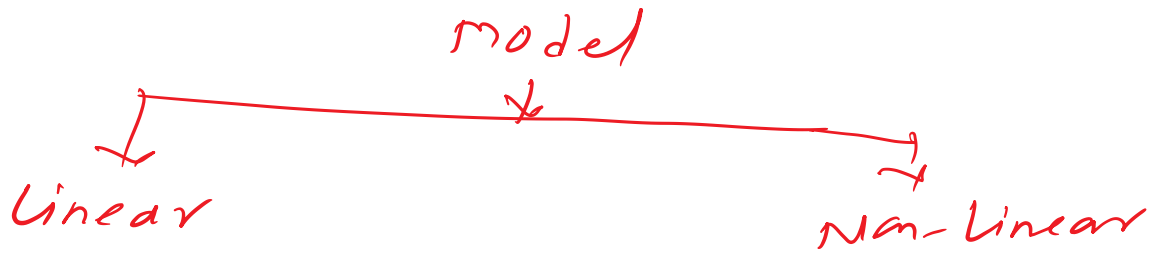
$$y = f(X_1 \ldots X_K, \beta_1 \ldots \beta_K)$$

is a mathematical model

If for $\varepsilon \neq 0$, model is statistical model.

* Types of models :-

$$y = f(X_1, \ldots X_K; \beta_1 \ldots \beta_K) + \varepsilon$$




* It is linear in parameters $\beta_1 \ldots \beta_K$

* Eg. $y = \beta_1 X_1^2 + \beta_2 \sqrt{X_2} + \beta_3 X_3 + \varepsilon$

* It is not linear in parameters $\beta_1 \ldots \beta_K$

* Eg. $y = \beta_1^2 X_1^2 + \beta_2 \sqrt{X_2} + \beta_3 \log X_3 + \varepsilon$

$$Y = \beta_1 X_1^2 + \beta_2 \sqrt{X_2} + \beta_3 X_3 + \varepsilon \quad \text{(I)}$$

$$\frac{\partial Y}{\partial \beta_1} = X_1^2, \quad \frac{\partial Y}{\partial \beta_2} = \sqrt{X_2} \quad \& \quad \frac{\partial Y}{\partial \beta_3} = X_3$$

Hence model (I) is linear

$$Y = \beta_1^2 X_1^2 + \beta_2 \sqrt{X_2} + \beta_3 \log X_3 + \varepsilon \quad \text{(II)}$$

$$\frac{\partial Y}{\partial \beta_1} = 2\beta_1 X_1^2 \quad \frac{\partial Y}{\partial \beta_2} = \sqrt{X_2} \quad \& \quad \frac{\partial Y}{\partial \beta_3} = \log X_3$$

Hence, the model is not linear.
i.e. Non-linear.

In general, linear model is defined as,

$$Y = f(X_1, \dots X_k, \beta_1 \dots \beta_k) + \varepsilon$$

$$= \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k + \varepsilon$$

$X$ & $Y$ - pre-defined variable

Knowledge of model is depend on knowledge of parameters.

Estimation method's :-

is maximum likelihood Estimation method.

2) method of moments

3) Least square method, etc.

do need an knowledge about dist. of y

"Regression" :- To move in the backward direction.

$y$ - Yield, $x_1$ - quality of Fertilizer.

$x_2$ - level of irrigation.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

* Steps in Regression :-

1) statement of the problem/objectives

$\equiv$ find the yield.

2) choice of relevant variables :-

- fertilizer, rainfall, Irrigation, temp, wind speed, .etc.

3) Collection of the Data :-

1. temp - °C

2. temp = $\begin{cases} 1 & > 80^\circ C \\ 0 & o.w. \end{cases}$

17, 19, 31, 82,
0, 0, 1, 1

age = Income =

| $y$ | $x_1$ | $x_2$ | ---- | $x_k$ |
|---|---|---|---|---|
| $y_1$ | $x_{11}$ | $x_{12}$ | ---- | $x_{1k}$ |
| | $x_{21}$ | $x_{22}$ | ---- | |
| | | | ---- | |
| $y_n$ | $x_{n1}$ | $x_{n2}$ | ---- | $x_{nk}$ |

4. Specification of the model :-

$$y = \beta_1 x_1^2 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

or $$y = \beta_1^2 x_1 + \beta_2^2 x_2 + \beta_3 \log x_3 + \varepsilon.$$

$$y = f(x_1, \ldots x_k, \beta_1, \ldots \beta_k) + \varepsilon.$$

5. choice of parameter estimation method (model fitting):-

- MLE
- MME
- OLS
- Ridge.
- LAD.
- ... etc.

6. Fitting of the model:-

$\beta_1, \beta_2, \ldots \beta_k$ — OLS

$$y = f(x_1 \ldots x_k, \hat{\beta}_1, \hat{\beta}_2, \ldots \hat{\beta}_k)$$

$\{(x_i, y_i) : i = 1, 2, \ldots n\}$

$$y = \beta_0 + \beta_1 x + \varepsilon.$$

$\hat{\beta}_0$ & $\hat{\beta}_1$

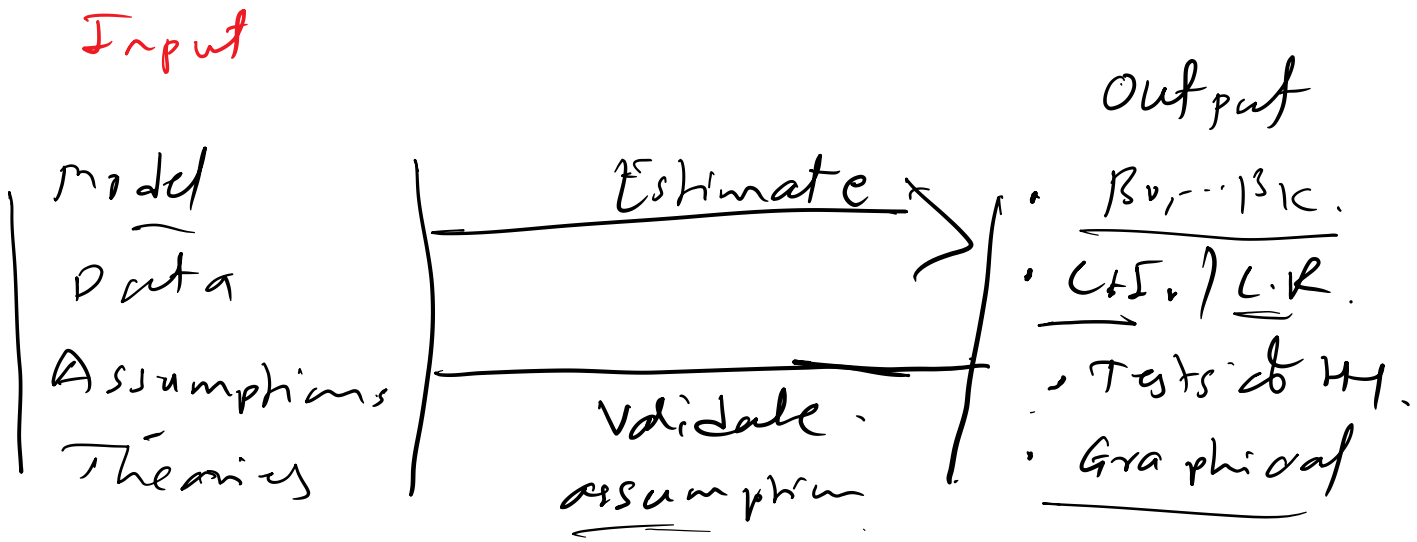
$$y = \hat{\beta}_0 + \hat{\beta}_1 x.$$

fitted value of $y$. — $x \in$ dataset.

predicted value of $y$ — any $x$.

# 7. Model validation :-

* Assumptions.




* Type of Regression :-

1. Univariate — only one response variable.
2. Multivariate — Two or more response variable.
3. Simple — only one regressor
4. Multiple — Two or more regressors.

# * Simple Linear Reg. Analysis :-

$$\underline{y} = \underline{\beta_0} + \underline{\beta_1 x} + \underline{\varepsilon}$$

$y$ - response / study / Dependent variable.

$x$ - independent / regressor /

$\beta_0, \beta_1$ - unknown Reg coeff.

$\varepsilon$ - Error term.

Assumptions :-

i) $\varepsilon$ is r.v. with mean 0 & variance $\sigma^2$.

i.e. $E(\varepsilon_i) = 0$ & $V(\varepsilon_i) = \sigma^2$ $\forall i = 1, 2, \ldots, n$.

ii) $\text{Cov}(\varepsilon_i, \varepsilon_j) = 0 \Rightarrow \varepsilon_i$ & $\varepsilon_j$ are uncorrelated.

i.e. no auto correlation.

iii) $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$

iv) model is linear.

If $x$ is not r.v.

$$E(Y) = E(\beta_0 + \beta_1 X + \varepsilon) \quad (\because Y = \beta_0 + \beta_1 X + \varepsilon)$$

$$= \beta_0 + \beta_1 X + 0$$

$$= \beta_0 + \beta_1 X.$$

&

$$V(Y) = V(a + \varepsilon) \qquad a = \beta_0 + \beta_1 X.$$

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

$$Y \sim N(\beta_0 + \beta_1 X, \sigma^2)$$

$$Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$$

$y_i$ are independent but identical.

If $x$ is r.v.

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

&

$$V(Y|X) = \sigma^2$$

If $\beta_0, \beta_1, \sigma$ are known then model is completely known.

o.w. we need to find out unknown parameters.

Method of Estimation :-

1. least square.
2. m.l.e
3. mm
4. Ridge... etc.

# * Simple Linear Regression *

For the bivariate data.

$\{(x_i, y_i);\ i = 1, 2, \cdots n\}$. the simple linear regression model is

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \cdots n.$$

$y_i$ - response variable or dependent variable

$x_i$ - regressor or predictor variable or independent variable

$\beta_0, \beta_1$ - regression coefficients

($\beta_0$ - intercept & $\beta_1$ - slope)

$\varepsilon_i$ - Error term

## Assumptions: -

i) $E(\varepsilon_i) = 0$ & $V(\varepsilon_i) = \sigma^2$ $\quad i = 1, 2, \ldots, n$

$\sigma^2$ is unknown

ii) $\text{cov}(\varepsilon_i, \varepsilon_j) = 0 \Rightarrow \varepsilon_i$ & $\varepsilon_j$ are uncorrelated.

i.e. no autocorrelation.

iii) $\varepsilon_i \overset{i.i.d}{\sim} N(0, \sigma^2)$

iv) model is linear

i) If $x$ is not r.v. (constant)

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

$$\therefore \quad E(y_i) = \beta_0 + \beta_1 x_i \qquad (\because E(\varepsilon_i) = 0)$$

& $\quad V(y_i) = \sigma^2 \qquad (\because V(\varepsilon_i) = \sigma^2)$

i.e. $y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$

ii) If $x$ is r.v.

$$E(y_i | x_i) = \beta_0 + \beta_1 x_i$$

&

$$V(y_i | x_i) = \sigma^2$$

i.e. $y_i | x_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$

Example:-

i) X - Sugarcane & Y - Production of sugar

Let simple linear regn model is

$$Y = 10 + 15X$$

ii) X - weight & Y - height

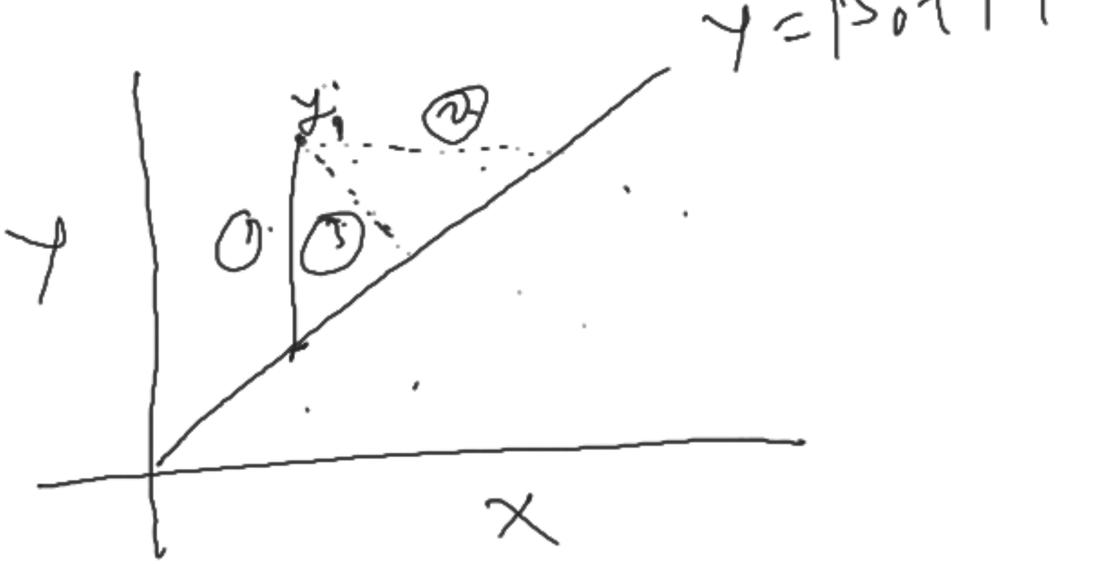
$$Y = 5 + 10X$$

if $X = 0$

$$E(Y) = 5 \quad \text{for } X = 0$$

* Methods for the estimation of unknown parameters $(\beta_0, \beta_1, \sigma^2)$ :-

* Least square method :-




i) Direct Regression method :-

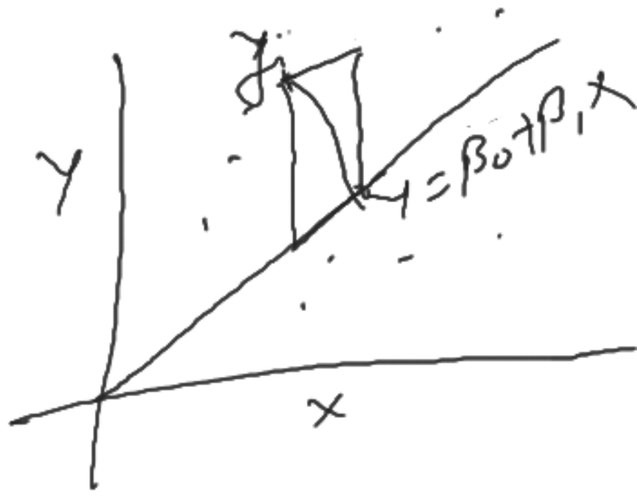
minimize vertical distance

ii) Reverse (or Inverse) Reg.

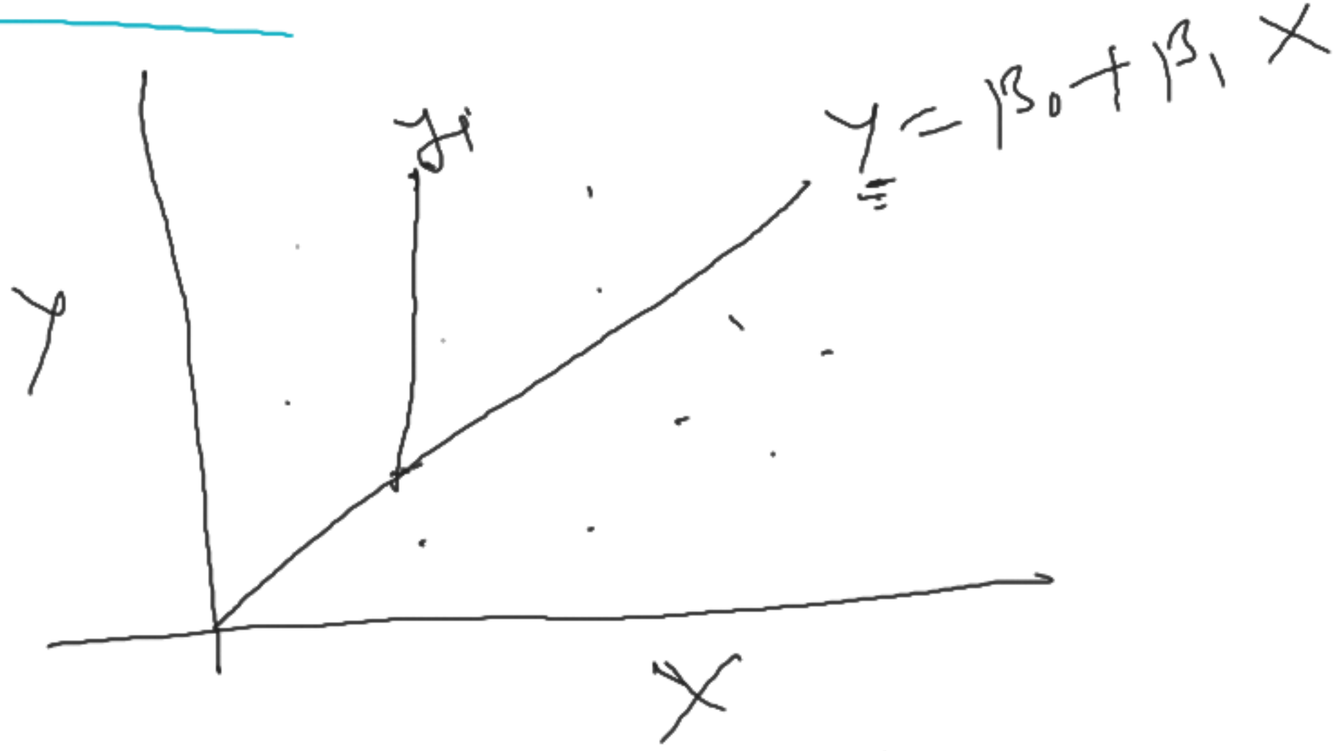
minimize horizontal distance

iii) Orthogonal (or major axis) Reg.

iv) Reduced major axis reg.



* Direct Reg. method :-




Least square estimator :- $\underset{\beta_0, \beta_1}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$

Least absolute deviation :- $\underset{\beta_0, \beta_1}{\operatorname{argmin}} \sum_{i=1}^{n} | y_i - \beta_0 - \beta_1 x_i |$

In general, $\underset{\beta_0, \beta_1}{\operatorname{argmin}} \sum_{i=1}^{n} g(y_i - \beta_0 - \beta_1 x_i)$ , $g(\cdot)$ is function

Suppose

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

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

$$\Rightarrow \sum_{i=1}^{n} y_i = n\beta_0 + \beta_1 \sum x_i \quad \text{——— (I)}$$

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

$$\Rightarrow \sum_{i=1}^{n} y_i x_i = \beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 \quad \text{——— (II)}$$

Least square normal eq$^n$s.

$$\sum_{i=1}^{n} \hat{y}_i = n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i \qquad \text{— (I)}$$

&

$$\sum_{i=1}^{n} \hat{y}_i x_i = \beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 \qquad \text{— (II)}$$

From eq$^n$ (I)

$$n\beta_0 = \sum_{i=1}^{n} \hat{y}_i - \beta_1 \sum_{i=1}^{n} x_i$$

$$\text{i.e. } \beta_0 = \bar{y} - \beta_1 \bar{x} \qquad \text{— (III)}$$

From eq$^n$s (II) & (III)

$$\sum_{i=1}^{n} \hat{y}_i x_i = (\bar{y} - \beta_1 \bar{x}) \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2$$

$$\therefore \sum_{i=1}^{n} y_i x_i = (\bar{y} - \beta_1 \bar{x}) \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2$$

$$= (\bar{y} - \beta_1 \bar{x}) n \bar{x} + \beta_1 \sum x_i^2 \quad \left(\because \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}\right)$$

$$= n \bar{x} \bar{y} - n \beta_1 \bar{x}^2 + \beta_1 \sum_{i=1}^{n} x_i^2$$

$$= n \bar{x} \bar{y} + \beta_1 \left( \sum_{i=1}^{n} x_i^2 - n \bar{x}^2 \right)$$

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

Hence

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad \& \quad \hat{\beta}_1 = \frac{\sum y_i x_i - n \bar{x} \bar{y}}{\sum x_i^2 - n \bar{x}^2}$$

Let $S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} (x_i^2 - 2\bar{x}x_i + \bar{x}^2)$

$$= \sum x_i^2 - 2\bar{x} \sum x_i + n\bar{x}^2$$

$$= \sum x_i^2 - 2\bar{x}\, n\bar{x} + n\bar{x}^2$$

$$= \sum x_i^2 - 2n\bar{x}^2 + n\bar{x}^2$$

$$= \sum x_i^2 - n\bar{x}^2$$

&

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

$$= \sum x_i y_i - n\bar{x}\bar{y}$$

$\therefore$ $cov(x,y) = \frac{S_{xy}}{n}$ & $V(x) = \frac{S_{xx}}{n}$.

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

&

$$\hat{\beta}_1 = \frac{\sum f_i x_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2} = \frac{S_{xy}}{S_{xx}}$$

This is a global optimal soln if

$$\frac{\partial^2 S}{\partial \beta_0^2} > 0, \quad \frac{\partial^2 S}{\partial \beta_1^2} > 0 \quad \& \quad \frac{\partial^2 S}{\partial \beta_0^2} \frac{\partial^2 S}{\partial \beta_1^2} - \left(\frac{\partial^2 S}{\partial \beta_0 \partial \beta_1}\right) > 0$$

since,

$$S(\beta_0, \beta_1) = \sum (y_i - \beta_0 - \beta_1 x_i)^2$$

$$\therefore \frac{\partial S}{\partial \beta_0} = -2\sum (y_i - \beta_0 - \beta_1 x_i) = -2\sum y_i + 2n\beta_0 + 2\beta_1 \sum x_i$$

$$\frac{\partial^2 S}{\partial \beta_0^2} = 2n > 0$$

$$\frac{\partial S}{\partial \beta_1} = -2\sum (y_i \beta_0 - \beta_1 x_i) x_i = -2\sum y_i x_i + 2\beta_0 \sum x_i + 2\beta_1 \sum x_i^2$$

$$\frac{\partial^2 S}{\partial \beta_1^2} = 2\sum x_i^2 > 0$$

$$\frac{\partial^2 S}{\partial \beta_0 \partial \beta_1} = 2\sum x_i$$

$$\therefore \frac{\partial^2 S}{\partial \beta_0^2} \frac{\partial^2 S}{\partial \beta_1^2} - \left(\frac{\partial^2 S}{\partial \beta_0 \partial \beta_1}\right)^2$$

$$= 2n \cdot 2\sum x_i^2 - (2\sum x_i)^2$$

$$= 4n\sum x_i^2 - 4n^2\bar{x}^2 \qquad \left(\because \bar{x} = \frac{\sum x_i}{n}\right)$$

$$= 4n\left(\sum x_i^2 - n\bar{x}^2\right)$$

$$= 4n\, S_{xx} = 4n\sum (x_i - \bar{x})^2 > 0$$

$\therefore (\hat{\beta}_0, \hat{\beta}_1)$ is a global optimum pt.

Alternatively,

Hessian matrix is

$$H = \begin{bmatrix} \frac{\partial^2 S}{\partial \beta_0^2} & \frac{\partial^2 S}{\partial \beta_0 \partial \beta_1} \\ \frac{\partial^2 S}{\partial \beta_0 \partial \beta_1} & \frac{\partial^2 S}{\partial \beta_1^2} \end{bmatrix} = \begin{bmatrix} 2n & 2n\bar{x} \\ 2n\bar{x} & 2\bar{\Sigma} x_i^2 \end{bmatrix}$$

$(\hat{\beta}_0, \hat{\beta}_1)$ is a global optimal soln if $H$ is positive definite matrix.

Note:- A matrix $H$ is p.d. if it's all eigen values are greater than zero.

i.e. $|H| > 0$

$$|H| = \left[ 4n \sum x_i^2 - 4n^2 \bar{x}^2 \right]$$

$$= 4n \left[ \sum x_i^2 - n \bar{x}^2 \right]$$

$$= 4n \sum (x_i - \bar{x})^2$$

$$= 4n S_{xx} > 0$$

Hence H is positive definite matrix. ∴

$(\hat{\beta_0}, \hat{\beta_1})$ is a global optimal sol?

# Properties of $\hat{\beta}_0$, $\hat{\beta}_1$ :-

## (1) Unbiasedness :-

We know that,

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

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

$$= \sum x_i y_i - \bar{x} \sum y_i = \sum x_i y_i - n\bar{x}\bar{y}$$

$$\therefore \quad \hat{\beta}_1 = \frac{\sum (x_i - \bar{x}) y_i}{S_{xx}} = \sum c_i y_i \text{ , where } c_i = \frac{x_i - \bar{x}}{S_{xx}}$$

$$\therefore \quad E(\hat{\beta}_1) = E\left(\sum_{i=1}^{n} c_i y_i\right) \qquad , \quad c_i = \frac{x_i - \bar{x}}{S_{xx}}$$

$$= \sum_{i=1}^{n} c_i E(y_i)$$

$$= \sum_{i=1}^{n} c_i (\beta_0 + \beta_1 x_i) \qquad (\because E(y_i) = \beta_0 + \beta_1 x_i)$$

$$= \beta_0 \sum_{i=1}^{n} c_i + \beta_1 \sum c_i x_i \qquad \text{---} (\bar{x})$$

where, $\sum c_i = \sum_{i=1}^{n} \frac{(x_i - \bar{x})}{S_{xx}} = \frac{1}{S_{xx}} \sum (x_i - \bar{x}) = 0$

&

$$\sum c_i x_i = \sum \frac{(x_i - \bar{x}) x_i}{S_{xx}} = \frac{1}{S_{xx}} \sum (x_i - \bar{x}) x_i$$

$$= \frac{1}{S_{xx}} \left[ \sum x_i^2 - n \bar{x}^2 \right] = 1$$

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

$\therefore$ $\hat{\beta}_1$ is an unbiased estimator of $\beta_1$.

---

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

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

$$= \beta_0 + \beta_1 \bar{x} + 0 - \bar{x}\, E(\hat{\beta}_1) \qquad (\because E(\bar{\varepsilon}) = 0)$$

$$= \beta_0 + \beta_1 \bar{x} - \beta_1 \bar{x}$$

$$= \beta_0$$

$\therefore$ $\hat{\beta}_0$ is an unbiased estimator of $\beta_0$

(c) variance :

$$V(\hat{\beta}_1) = V(\Sigma c_i y_i) \quad , \quad \hat{\beta}_1 = \Sigma c_i y_i, \quad c_i = \frac{x_i - \bar{x}}{S_{xx}}$$

since,

$$V\left(\sum_{i=1} a_i y_i\right) = \Sigma a_i^2 V(y_i) + 2 \sum_{i<j} \Sigma a_i a_j \operatorname{cov}(y_i, y_j)$$

$$= \Sigma a_i^2 V(y_i) + \sum_{i \neq j} \Sigma a_i a_j \operatorname{cov}(y_i, y_j)$$

If $y_i$'s, $i = 1, 2, \ldots, n$, are independent r.v.'s then..

$$V(\Sigma a_i y_i) = \Sigma a_i^2 V(y_i)$$

$$\therefore V(\hat{\beta}_1) = \Sigma c_i^2 V(y_i) = \Sigma c_i^2 \sigma^2 = \sigma^2 \Sigma \frac{(x_i - \bar{x})^2}{S_{xx}^2}$$

($\because$ $y_i$'s are independent)

$$\therefore V(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}^2} \sum (x_i - \bar{x})^2 = \frac{\sigma^2}{S_{xx}^2} S_{xx}$$

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

---

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

$$= V(\bar{y}) + \bar{x}^2 V(\hat{\beta}_1) - 2\bar{x}\,\mathrm{cov}(\bar{y}, \hat{\beta}_1)$$

$(\because V(aX + bY) = a^2 V(X) + b^2 V(Y) + 2ab\,\mathrm{cov}(X,Y))$

where

$$V(y_i) = \sigma^2, \quad V(\bar{y}) = \sigma^2/n$$

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

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

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

$$V(\bar{y}) = \sigma^2/n,$$

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

$$Cov(\bar{y}, \hat{\beta}_1) = Cov\left(\frac{\sum y_i}{n}, \frac{\sum (x_i - \bar{x}) y_i}{S_{xx}}\right)$$

$$= \sum_{i=1}^{n} \frac{(x_i - \bar{x})}{n S_{xx}} V(y_i)$$

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

$$= 0$$

$(\because Cov(ax, by) = ab\,Cov(x, y))$

$\because Cov(x+y, u+v) = Cov(x,u) + Cov(y,u) + Cov(x,v) + Cov(y,v)$

$$\therefore V(\hat{\beta}_0) = \frac{\sigma^2}{n} + \bar{x}^2 \frac{\sigma^2}{S_{xx}} \quad -0$$

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

3) $Cov(\hat{\beta}_0, \hat{\beta}_1)$

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

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

$$= 0 - \bar{x} \frac{\sigma^2}{S_{xx}} \qquad (\because Cov(\bar{y}, \hat{\beta}_1) = 0)$$

$$= -\frac{\sigma^2 \bar{x}}{S_{xx}}$$

Remark :-

Gauss- Markov th$^m$ :-

For the linear regression $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ with assumption of $\varepsilon_i$, The OLS estimator is a best linear unbiased estimator (BLUE).

H.W :- Find the least square estimator of $\beta_0$ & $\beta_1$ for the following regression models.

i) $y_i = \beta_0 + \varepsilon_i$

ii) $y_i = \beta_1 x_i + \varepsilon_i$

Also, check the properties of least square estimators.

## Remarks :

$\hat{\beta}_{1yx}$ OLS of reg^ line y on x.
$\hat{\beta}_{1xy}$ OLS of reg^ line x on y
$r = \text{corr}(x, y)$

(1) $\hat{\beta}_{1yx}$, $\hat{\beta}_{1xy}$ & $r$ have same sign.

(2) $r = \sqrt{\hat{\beta}_{1yx}\ \hat{\beta}_{1xy}}$

(3) Both reg. coefficients cannot exceed unity simultaneously.

(4) If $r = 1$ (or $r = -1$) then $\hat{\beta}_{1yx}\ \hat{\beta}_{1xy} = 1$

(5) change of origin (or location) does not affect reg. coeff.

(6) change of scale affect reg. coeff.

(7) The angle bet^ two reg lines y on x & x on y is

$$\theta = \tan^{-1}\left(\frac{1-r^2}{|r|} \cdot \frac{\sigma_x \sigma_y}{\sigma_x^2 + \sigma_y^2}\right)$$

(3) If $r = 1$ or $r = -1$, $\theta = 0$

If $r = 0$, $\theta = \pi/2$.

---

- **Residual :-**
  - It is diff. bet$^n$ $y_i$ & $\hat{y}_i$.
  - $r_i = y_i - \hat{y}_i$
  - It is estimate value of error term.

## Error

i) $\varepsilon_i = Y_i - \beta_0 - \beta_1 x_i$

ii) $\varepsilon_i$'s are independent
& $\varepsilon_i$ are uncorrelated

iii) $\varepsilon_i$'s have same variance.
i.e. constant variance.

iv) $\varepsilon_i \sim$ Normal dist$^n$

## Residual

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

$r_i$'s are independent & correlated.

Variance is not constant

[ if n is large, they are close to independent & constant variance ]

$r_i \sim$ Normal dist$^n$

## Properties :-

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

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

(3) $\sum_{i=1}^{n} x_i e_i = 0$

(4) $\sum \hat{y}_i e_i = 0$

(5) Regression line always passes through the centroid of the data $(\bar{x}, \bar{y})$.

(1) $\sum \hat{y}_i r_i = 0$

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$
$$= \bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_i$$
$$= \bar{y} + \hat{\beta}_1 (x_i - \bar{x})$$

Proof:-

$$\sum \hat{y}_i r_i = \sum \hat{y}_i (y_i - \hat{y}_i)$$

$$= \sum_{i=1}^{n} (\bar{y} + \hat{\beta}_1 (x_i - \bar{x}))(y_i - \bar{y} - \hat{\beta}_1 (x_i - \bar{x}))$$

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

$$= 0 - 0 + \hat{\beta}_1 \sum (x_i - \bar{x})(y_i - \bar{y}) - \hat{\beta}_1^2 \sum (x_i - \bar{x})^2$$

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

$$= \frac{S_{xy}}{S_{xx}} S_{xy} - \frac{S_{xy}^2}{S_{xx}^2} S_{xx} = \frac{S_{xy}^2}{S_{xx}} - \frac{S_{xy}^2}{S_{xx}} = 0$$

* Residual sum of squares:-

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

$$= \sum_{i=1}^{n} \left[ y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \right]^2$$

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

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

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

$$= \underbrace{\sum_{i=1}^{n} (y_i - \bar{y})^2}_{S_{yy}} + \hat{\beta}_1^2 \underbrace{\sum_{i=1}^{n} (x_i - \bar{x})^2}_{S_{xx}} - 2\hat{\beta}_1 \underbrace{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}_{S_{xy}}$$

Let $S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2$, $S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2$, $S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y})$

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

$= S_{yy} + \hat{\beta}_1 (\hat{\beta}_1 S_{xx} - 2 S_{xy})$

$= S_{yy} + \hat{\beta}_1 \left(\frac{S_{xy}}{S_{xx}} S_{xx} - 2 S_{xy}\right)$

$= S_{yy} + \hat{\beta}_1 (S_{xy} - 2 S_{xy})$

$= S_{yy} - \hat{\beta}_1 S_{xy}$ — (I)

$= S_{yy} - \hat{\beta}_1 \cancel{S_{xy}} \cdot \frac{S_{xy}}{S_{xx}} \cdot \frac{S_{xx}}{\cancel{S_{xy}}} = S_{yy} - \hat{\beta}_1^2 S_{xx}$ — (II)

* Estimation of $\sigma^2$:-

Since, $SS_{Res} = S_{YY} - \hat{\beta}_1^2 S_{xx}$

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

$$E(S_{YY}) = E\left(\sum (y_i - \bar{y})^2\right)$$

$$= E\left(\sum y_i^2 - n\bar{y}^2\right)$$

$$= E\left(\sum y_i^2\right) - n E(\bar{y}^2)$$

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

$V(y_i) = E(y_i^2) - E^2(y_i)$

i.e. $\sigma^2 = E(y_i^2) - (\beta_0 + \beta_1 x_i)^2$

i.e. $E(y_i^2) = \sigma^2 + (\beta_0 + \beta_1 x_i)^2$

&

$V(\bar{y}) = E(\bar{y}^2) - E^2(\bar{y})$

i.e. $E(\bar{y}^2) = V(\bar{y}) + E^2(\bar{y})$

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

$$E(S_{xy}) = \sum E(y_i^2) - n E(\bar{y}^2)$$

$$= \sum_{i=1}^{n} \left[ \sigma^2 + (\beta_0 + \beta_1 x_i)^2 \right] - n\left[ \frac{\sigma^2}{n} + (\beta_0 + \beta_1 \bar{x})^2 \right]$$

$$= n\sigma^2 + \sum_{i=1}^{n} (\beta_0 + \beta_1 x_i)^2 - \sigma^2 - n(\beta_0 + \beta_1 \bar{x})^2$$

$$= (n-1)\sigma^2 + \sum_{i=1}^{n} (\beta_0 + \beta_1 x_i)^2 - n(\beta_0 + \beta_1 \bar{x})^2$$

$$= (n-1)\sigma^2 + \sum_{i=1}^{n} (\beta_0^2 + \beta_1^2 x_i^2 + 2\beta_0\beta_1 x_i) - n(\beta_0^2 + \beta_1^2 \bar{x}^2 + 2\beta_0\beta_1 \bar{x})$$

$$= (n-1)\sigma^2 + n\beta_0^2 + \beta_1^2 \sum_{i=1}^{n} x_i^2 + 2\beta_0\beta_1 \sum_{i=1}^{n} x_i - n\beta_0^2 - n\beta_1^2 \bar{x}^2 - 2n\beta_0\beta_1 \bar{x}$$

$$\therefore \quad E(S_{YY}) = (n-1)\sigma^2 + \beta_1^2 \sum x_i^2 - n\beta_1^2 \bar{x}^2$$

$$= (n-1)\sigma^2 + \beta_1^2 \left( \sum x_i^2 - n\bar{x}^2 \right)$$

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

Hence,

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

$$= (n-1)\sigma^2 + \beta_1^2 S_{xx} - S_{xx} E(\hat{\beta}_1^2)$$

Since, $\quad V(\hat{\beta}_1) = E(\hat{\beta}_1^2) - E^2(\hat{\beta}_1)$

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

$$\therefore \quad E(SS_{Res}) = (n-1)\sigma^2 + \beta_1^2 S_{xx} - S_{xx} \left( \frac{\sigma^2}{S_{xx}} + \beta_1^2 \right)$$

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

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

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

Hence, $\frac{SS_{Res}}{n-2}$ is an unbiased estimator $\sigma^2$

$$\hat{\sigma}^2 = \frac{SS_{Res}}{n-2} = MS_{Res}$$ i.e. residual mean square.

&

$\hat{\sigma}$ - standard error of reg.

Other method :-

Thm :- Suppose $Y \sim N_p(\mu, \Sigma)$. $A$ is a symmetric matrix of constants with rank $r$. If $\lambda = \frac{1}{2}\mu' A \mu$ then $Y'AY \sim \chi^2_{(r,\lambda)}$ iff $A\Sigma$ is idempotent matrix.

Cochran's Thm :-

Let $x_1 \ldots x_n$ be a random sample from $N(0, \sigma^2)$ If $\sum_{i=1}^{n} x_i^2 = Q_1 + Q_2 + \cdots + Q_k$ where, $Q_j$ is a quadratic form in $x_1 \ldots x_n$ with $r_j$ d.f then r.v.'s $Q_1 \ldots Q_k$ are mutually independent If $Q_j/\sigma^2$ is $\chi^2_{r_j}$ iff $\sum r_i = n$

Using th$^{m}$ 1 & th$^{m}$ 2. we can prove that,

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

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

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

Other method :-

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

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

$$\therefore \left(\frac{\varepsilon_i}{\sigma}\right)^2 \sim \chi_1^2$$

$$\therefore \sum_{i=1}^{n} \frac{(y_i - \beta_0 - \beta_1 x_i)^2}{\sigma^2} \sim \chi_n^2$$

$$\therefore \sum_{i=1}^{n} \frac{(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{\sigma^2} \sim \chi_{n-2}^2$$

$$\therefore \frac{SSReg}{\sigma^2} \sim \chi_{n-2}^2 \quad \text{i.e} \quad E\left(\frac{SSReg}{n-2}\right) = \sigma^2$$

- Estimate of variance of $\hat{\beta}_0$ & $\hat{\beta}_1$ :-

$$\hat{V}(\hat{\beta}_0) = \hat{\sigma}^2 \left( \frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right), \qquad \hat{\sigma}^2 = \frac{SS_{Res}}{n-2} = MS_{Res}$$

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

## Testing of Hypothesis :-

$$H_0 : \beta_1 = \beta_{10} \quad v/s \quad H_1 : \beta_1 \neq \beta_{10}$$

1) $\sigma^2$ is know

Since, $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$,

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

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

under $H_0 : \beta_1 = \beta_{10}$

$$Z_{cal} = \frac{\hat{\beta}_1 - \beta_{10}}{\sqrt{\frac{\sigma^2}{S_{xx}}}} \sim N(0,1)$$

We reject $H_0$ if

$$|Z_{cal}| > Z_{\alpha/2}.$$

o.w. accept $H_0$.

ii) $\sigma^2$ is unknown:

$$t_{cal} = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{SS_{Res}}{(n-2)S_{xx}}}} = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{MS_{Res}}{S_{xx}}}}$$

$\left(\because MS_{Res} = \frac{SS_{Res}}{n-2}\right)$

under $H_0$,

$$t_{cal} = \frac{\hat{\beta}_1 - \beta_{10}}{\sqrt{\frac{MS_{Res}}{S_{xx}}}} \sim t_{n-2}$$

- If $X$ & $Y$ are independent r.v.'s & $X \sim N(0,1)$, $Y \sim \chi^2_n$

then, $\dfrac{X}{\sqrt{Y/n}} \sim t_n$

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

$\therefore X = \dfrac{\hat{\beta}_1 - \beta_1}{\sqrt{\sigma^2/S_{xx}}} \sim N(0,1)$

$\dfrac{SS_{res}}{\sigma^2} \sim \chi^2_{n-2}$

$\therefore Y = \dfrac{MS_{res}(n-2)}{\sigma^2} \sim \chi^2_{n-2}$.

$$\frac{X}{\sqrt{Y/df}} = \frac{\dfrac{\hat{\beta}_1 - \beta_1}{\sqrt{\sigma^2/S_{xx}}}}{\sqrt{\dfrac{MS_{res}(n-2)}{\sigma^2(n-2)}}} = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\dfrac{MS_{res}}{S_{xx}}}}$$

We reject $H_0$ if $|t_{cal}| > t_{\alpha/2, n-2}$.

o.w. we accept $H_0$.

$\alpha$ - l.o.s.

---

b) Testing of hypothesis for $\beta_0$ :-

$$H_0: \beta_0 = \beta_{00} \quad vs \quad H_1: \beta_0 \neq \beta_{00}$$

i) $\sigma^2$ is known :-

$$Z_{cal} = \frac{\hat{\beta_0} - \beta_0}{\sqrt{\sigma^2\left(\frac{1}{n} + \frac{\bar{X}^2}{S_{xx}}\right)}} \sim N(0,1)$$

under $H_0$,

$$Z_{cal} = \frac{\hat{\beta_0} - \beta_{00}}{\sqrt{\sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}} \sim N(0,1)$$

we reject $H_0$ if $|Z_{cal}| > Z_{\alpha/2}$. o.w. accept $H_0$.

ii) $\sigma^2$ unknown,

$$t_{cal} = \frac{\hat{\beta}_0 - \beta_0}{\sqrt{\frac{SS_{Res}}{n-2}\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}}$$

$$= \frac{\hat{\beta}_0 - \beta_0}{\sqrt{MS_{Res}\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}} \sim t_{n-2}.$$

under $H_0$.

$$t_{cal} = \frac{\hat{\beta}_0 - \beta_{0H_0}}{\sqrt{MS_{Res}\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}} \sim t_{n-2}.$$

we reject $H_0$ if $|t_{cal}| > t_{\alpha/2, n-2}$, o.w. accept $H_0$.

c) Test for $\sigma^2$ :-

$$H_0 : \sigma^2 = \sigma_0^2 \quad \text{v/s} \quad H_1 : \sigma^2 \neq \sigma_0^2$$

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

$$C = \frac{SS_{Res}}{\sigma^2}$$


under $H_0$,

$$C = \frac{SS_{Res}}{\sigma_0^2} \sim \chi^2_{n-2}.$$

we reject $H_0$ if $C < \chi^2_{n-2,1-\alpha/2}$ or $C > \chi^2_{n-2,\alpha/2}$

o.w. accept $H_0$.

d) Linearity :-

$H_0: \beta_1 = 0$ v/s $H_1: \beta_1 \neq 0$

(no linear rel$^n$) (linear rel$^n$)

i) $\sigma^2$ known

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

under $H_0$

$$Z_{cal} = \frac{\hat{\beta}_1}{\sqrt{\sigma^2 / S_{xx}}} \sim N(0,1)$$

we reject $H_0$ if $|Z_{cal}| > Z_{\alpha}$ o.w. accept $H_0$.

ii) $\sigma^2$ Unknown

$$t_{cal} = \hat{\beta}_1 - \beta_1 \Big/ \sqrt{\frac{MSRes}{S_{xx}}}$$

under $H_0$ $t_{cal} = \hat{\beta}_1 \Big/ \sqrt{\frac{MSRes}{S_{xx}}} \sim t_{n-2}$

we reject $H_0$ if $|t_{cal}| > t_{\alpha, n-2}$ o.w. accept $H_0$.

# Confidence Interval :-

## C.I. for $\beta_1$ :-

### i) $\sigma^2$ known.

$100(1-\alpha)\%$ C.I. for $\beta_1$ is

$$P\left(-Z_{\alpha/2} \le Z_{cal} \le Z_{\alpha/2}\right) = 1-\alpha$$

$$\therefore P\left(-Z_{\alpha/2} \le \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\sigma^2}{S_{xx}}}} \le Z_{\alpha/2}\right) = 1-\alpha.$$

$$\therefore P\left(-Z_{\alpha/2}\sqrt{\frac{\sigma^2}{S_{xx}}} \le \hat{\beta}_1 - \beta_1 \le Z_{\alpha/2}\sqrt{\frac{\sigma^2}{S_{xx}}}\right) = 1-\alpha$$

$$P\left(-Z_{\alpha/2}\sqrt{\frac{\sigma^2}{S_{xx}}} \le \beta_1 - \hat{\beta}_1 \le Z_{\alpha/2}\sqrt{\frac{\sigma^2}{S_{xx}}}\right) = 1-\alpha.$$

$$\therefore P\left(\hat{\beta}_1 - Z_{\alpha/2}\sqrt{\frac{\sigma^2}{S_{xx}}} \le \beta_1 \le Z_{\alpha/2}\sqrt{\frac{\sigma^2}{S_{xx}}} + \hat{\beta}_1\right) = 1-\alpha.$$

$\therefore$ $100(1-\alpha)\%$ C.I. for $\beta_1$ is

$$\left(\hat{\beta}_1 - Z_{\alpha/2}\sqrt{\frac{\sigma^2}{S_{xx}}},\ \hat{\beta}_1 + Z_{\alpha/2}\sqrt{\frac{\sigma^2}{S_{xx}}}\right)$$

i.e $\hat{\beta}_1 \pm Z_{\alpha/2}\sqrt{\frac{\sigma^2}{S_{xx}}}$.

ii) $\sigma^2$ unknown :-

$100(1-\alpha)\%$ C.I for $\beta_1$ is

$$P(-t_{\alpha/2,n-2} \le t_{cal} \le t_{\alpha/2,n-2}) = 1-\alpha$$

i.e. $$P\left(-t_{\alpha/2,n-2} \le \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{MS_{Res}}{S_{xx}}}} \le t_{\alpha/2,n-2}\right) = 1-\alpha.$$

i.e. $$P\left(-t_{\alpha/2,n-2}\sqrt{\frac{MS_{Res}}{S_{xx}}} \le \hat{\beta}_1 - \beta_1 \le t_{\alpha/2,n-2}\sqrt{\frac{MS_{Res}}{S_{xx}}}\right)$$

$$= 1-\alpha.$$

i.e. $$P\left(-t_{\alpha/2,n-2}\sqrt{\frac{MS_{Res}}{S_{xx}}} \le \beta_1 - \hat{\beta}_1 \le t_{\alpha/2,n-2}\sqrt{\frac{MS_{Res}}{S_{xx}}}\right) = 1-\alpha.$$

$\therefore$ C.I. for $\beta_1$ is

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

b) C.I. for $\beta_0$ :-

i) $\sigma^2$ is known :-

$(1-\alpha)\%$ C.I. for $\beta_0$ is

$$P\left(-Z_{\alpha/2} \leq Z_{cal} \leq Z_{\alpha/2}\right) = 1-\alpha.$$

i.e. $P\left(-z_{\alpha/2} \le \frac{\hat{\beta}_0 - \beta_0}{\sqrt{\sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}} \le z_{\alpha/2}\right) = 1-\alpha.$

Hence, $100(1-\alpha)\%$ C.I. for $\beta_0$ is

$$\left(\hat{\beta}_0 - z_{\alpha/2}\sqrt{\sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)},\ \hat{\beta}_0 + z_{\alpha/2}\sqrt{\sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}\right)$$

ii) $\sigma^2$ is unknown :-

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

$$\overline{\qquad\qquad se(\hat{\beta}_0)\qquad\qquad}$$

# Confidence Interval :-

| Parameter | $\sigma^2$ | C.I |
|---|---|---|
| $\beta_0$ | known | $\hat{\beta}_0 \mp z_{\alpha/2} \, sd(\hat{\beta}_0)$ |
| | unknown | $\hat{\beta}_0 \mp t_{\alpha/2, n-2} \, se(\hat{\beta}_0)$ |
| $\beta_1$ | known | $\hat{\beta}_1 \mp z_{\alpha/2} \, sd(\hat{\beta}_1)$ |
| | unknown | $\hat{\beta}_1 \mp t_{\alpha/2, n-2} \, se(\hat{\beta}_1)$ |

c) C.I. for $\sigma^2$ :-

100(1−α)% C.I. for $\sigma^2$ is

$$P\left(\chi^2_{1-\alpha/2,\,n-2} \le \frac{SS_{Res}}{\sigma^2} \le \chi^2_{\alpha/2,\,n-2}\right) = 1-\alpha.$$

i.e. $$P\left(\frac{1}{\chi^2_{1-\alpha/2,\,n-2}} \ge \frac{\sigma^2}{SS_{Res}} \ge \frac{1}{\chi^2_{\alpha/2,\,n-2}}\right) = 1-\alpha.$$

i.e. $$P\left(\frac{SS_{Res}}{\chi^2_{\alpha/2,\,n-2}} \le \sigma^2 \le \frac{SS_{Res}}{\chi^2_{1-\alpha/2,\,n-2}}\right) = 1-\alpha.$$

Hence, C.I. for $\sigma^2$ is

$$\left(\frac{SS_{Res}}{\chi^2_{\alpha/2,\,n-2}}, \frac{SS_{Res}}{\chi^2_{1-\alpha/2,\,n-2}}\right) \text{ or } \left[\frac{n-2\,MS_{Res}}{\chi^2_{\alpha/2,\,n-2}}, \frac{n-2\,MS_{Res}}{\chi^2_{1-\alpha/2,\,n-2}}\right]$$

d) C.I for $E(y|x_0)$

$$\hat{E}(y|x_0) = \hat{\mu}_{y|x_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

$$\therefore V(\hat{\mu}_{y|x_0}) = V(\hat{\beta}_0 + \hat{\beta}_1 x_0)$$

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

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

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

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

$$\therefore \quad V(\hat{\mu}_{y|x_0}) = \sigma^2 \left( \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right)$$

$\sigma$ unknown

$$\frac{\hat{\mu}_{y|x_0} - E(y|x_0)}{\sqrt{MS_{res}\left( \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right)}} \sim t_{n-2}$$

$\therefore$ C.I is

$$\hat{\mu}_{y|x_0} \pm \sqrt{MS_{res}\left( \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right)}$$

# * Analysis of variance :-

We know that,

$$r_i = y_i - \hat{y}_i$$

$$= y_i - \bar{y} - \hat{y}_i + \bar{y}$$

$$= (y_i - \bar{y}) - (\hat{y}_i - \bar{y})$$

$$\therefore \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2 + \sum_{i=1}^{n} (\hat{y}_i - \bar{y}) - 2 \sum_{i=1}^{n} (y_i - \bar{y})(\hat{y}_i - \bar{y})$$

where:-

$$\sum_{i=1}^{n}(y_i-\bar{y})(\hat{y}_i-\bar{y}) = \sum_{i=1}^{n}(y_i-\bar{y})\hat{\beta}_1(x_i-\bar{x})$$

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

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

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

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

$$= \hat{\beta}_1^2 \sum_{i=1}^{n}(x_i-\bar{x})^2$$

$$= \sum_{i=1}^{n}[\hat{\beta}_1(x_i-\bar{x})]^2$$

$$= \sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2 \qquad (\because \hat{y}_i-\bar{y} = \hat{\beta}_1(x_i-\bar{x}))$$

Hence,

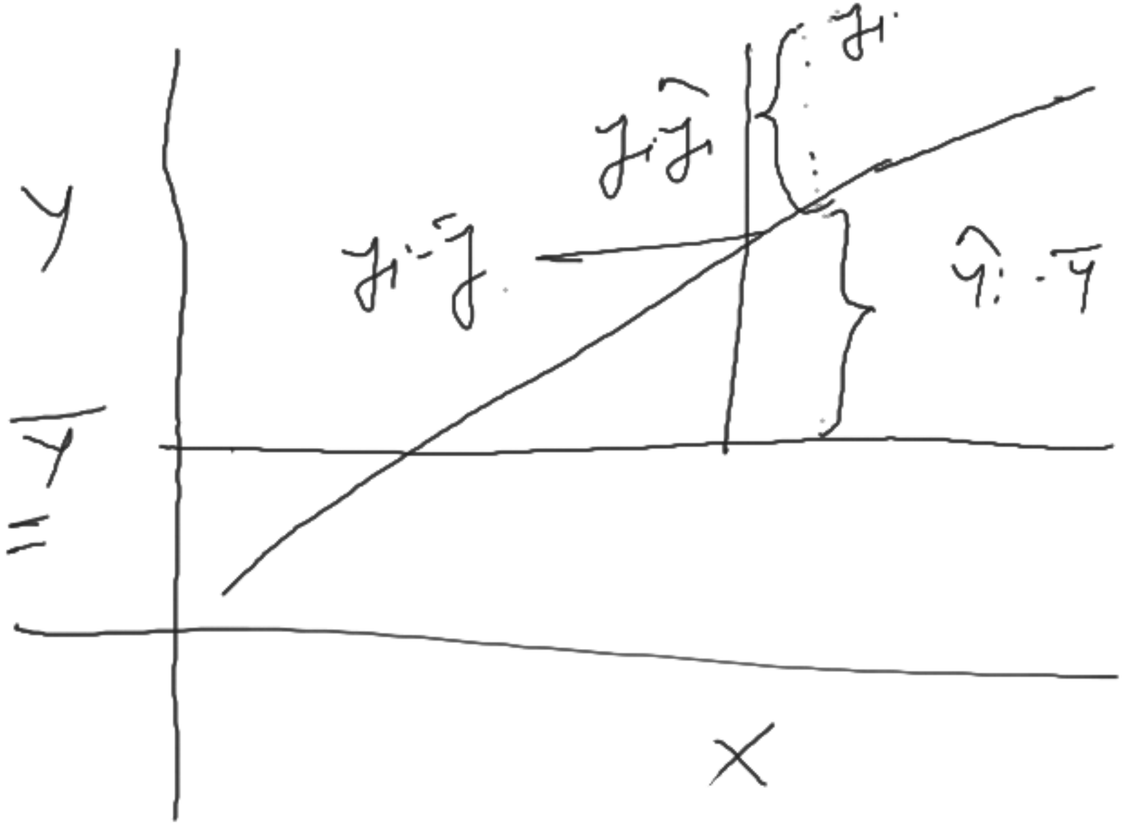
$$\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 = \sum(y_i-\bar{y})^2+\sum(\hat{y}_i-\bar{y})^2-2\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$$

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

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

| | | | |
|---|---|---|---|
| TSS or SST or SY | total sum of squares or sum of square about mean or corrected sum of square. | reg sum of square. SSR or SSReg | residual sum of squa. SSRes SSE = |

| Source of variation | d.f. |
| --- | --- |
| $SS_T = \sum_{i=1}^{n} (y_i - \bar{y})^2$ | $n-1$ |
| $SS_R = \sum (\hat{y}_i - \bar{y})^2$ | $1$ |
| $SS_{Res} = \sum (y_i - \hat{y}_i)^2$ | $n-2$ |



# Degree of freedom :-

7 chocolates : A, B, C, D, E, F, G

| Day | chocolate | choice |
|---|---|---|
| (1) | D | 7 |
| (2) | E | 6 |
| (3) | B | 5 |
| (4) | F | 4 |
| (5) | G | 3 |
| (6) | C | 2 |
| (7) | = | (1) |

d.f

(6)

$x_1, x_2, \cdots x_{10}$

if $\sum x_i = 100$

(or $\bar{x} = 10$)

D.f = 9

Degree of freedom

= No. of obs$^n$ — known rel$^n$ bet$^n$ them

(or No. of constraints or No of estimated parameters)

Examples :-

i) Data : $x_1$ ---- $x_{10}$ & constraint $\sum x_i = 10$

D.f = No. of obs$^n$ - No. of constraint = 10 - 1 = 9

ii) Residual sum of square = $\sum (y_i - \hat{y}_i)^2 = \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$

D.f = No. of obs$^n$ - No. of constraint (normal eq$^n$s) (No. of parameters estimated)

= $n - 2$

# ANOVA table :-

| Source of variation | Df | S.S. | MSS | F |
|---|---|---|---|---|
| Reg. | 1 | $SS_{Reg}$ | $MS_{Reg} = \frac{SS_{Reg}}{1}$ | $\frac{MS_{Reg}}{MSE}$ |
| Residual | n-2 | $SS_{Res}$ | $MSE = \frac{SS_{Res}}{n-2}$ | |
| Total | n-1 | $SS_T$ | | |

Expectation of mean sum of square :-

$$E(MS_{Reg}) = E\left(\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2\right)$$

$$= E\left(\hat{\beta}_1^2 \sum_{i=1}^{n}(x_i - \bar{x})^2\right)$$

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

$$= S_{xx}\, E\left(\hat{\beta}_1^2\right) \qquad \left(\because V(\hat{\beta}_1) = E(\hat{\beta}_1^2) - E^2(\hat{\beta}_1)\right)$$

$$= S_{xx}\left(\frac{\sigma^2}{S_{xx}} + \beta_1^2\right) \qquad \left(\because \frac{\sigma^2}{S_{xx}} = E(\hat{\beta}_1^2) - \beta_1^2\right)$$

$$= \sigma^2 + \beta_1^2 S_{xx} > 0$$

4

$$E(MSE) = E\left(\frac{SSRes}{n-2}\right) = \sigma^2 > 0$$

$$\frac{(n-2)\,MSE}{\sigma^2} \sim \chi^2_{n-2} \quad \& \quad \frac{MSReg}{\sigma^2} \sim \chi^2_1$$

Test :-

$H_0: \beta_1 = 0$ V/S $H_1: \beta_1 \neq 0$

under $H_0$:

$$F_{cal} = \frac{\frac{MSReg}{\sigma^2}}{\frac{(n-2)MSE}{\sigma^2(n-2)}} =$$

$$= \frac{MSReg}{MSE} \sim F_{1,n-2}$$

$X \sim \chi^2_{n_1}$

$Y \sim \chi^2_{n_2}$

$$F = \frac{X/n_1}{Y/n_2}$$

We reject $H_0$ if $F_{cal} > F_{\alpha, 1, n-2}$

o.w. accept $H_0$.

---

## Coefficient of Determination:

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

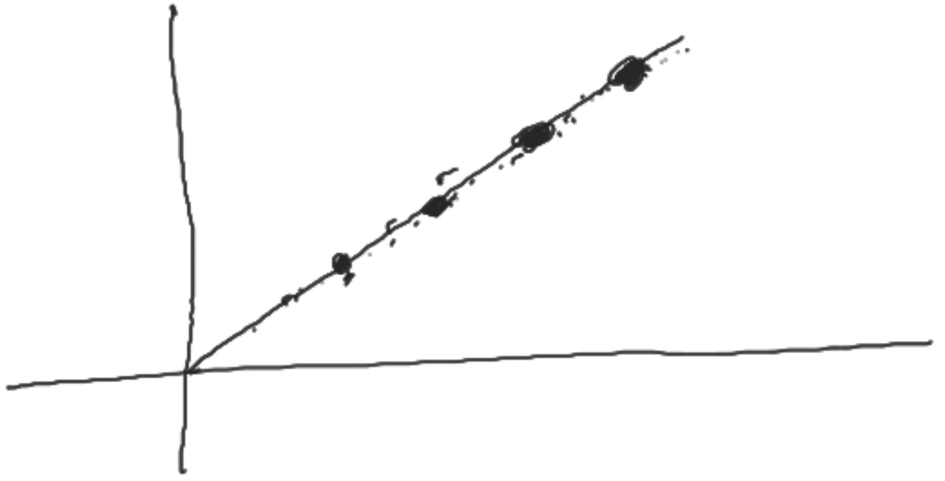
$$R^2 = \frac{SS_{Reg}}{SS_T} = \frac{SS_T - SS_{Res}}{SS_T} = 1 - \frac{SS_{Res}}{SS_T}$$

$R^2$ measures the prop$^n$ of variation in response variable that is explained by the reg$^n$ model.

Case 1 : $R^2 = 1$

i.e. $R^2 = \frac{SS_{reg}}{SS_T} = 1 \Rightarrow SS_{reg} = SS_T$.

i.e. $SS_{reg} = 0$.



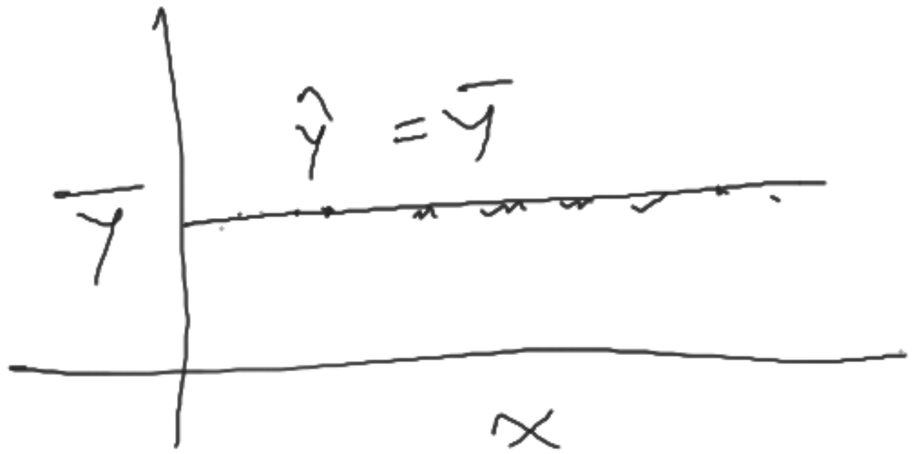
Case 2 : $R^2 = 0$

$R^2 = \frac{SS_{reg}}{SS_T} = 1 - \frac{SS_{res}}{SS_T} = 0 \Rightarrow SS_{reg} = SS_T$

i.e. $\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2$ i.e. $\hat{y}_i = \bar{y}$

Hence, If $R^2 = 0$

then $X$ & $Y$ are uncorrelated.




Remark:-

1) $R^2 = (corr(X, Y))^2$

2) since $-1 \leq corr(X,Y) \leq 1$ $\Rightarrow 0 \leq R^2 \leq 1$

3) $$R^2 = \frac{SS_{Reg}}{SS_T}$$

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

$\therefore$ $SS_{Reg} \leq SS_T \Rightarrow \frac{SS_{Reg}}{SS_T} \leq 1$ — (I)

& $SS_{Reg} \geq 0$ & $SS_T > 0 \Rightarrow \frac{SS_{Reg}}{SS_T} \geq 0$ — (II)

$\therefore$ $0 \leq \frac{SS_{Reg}}{SS_T} \leq 1$

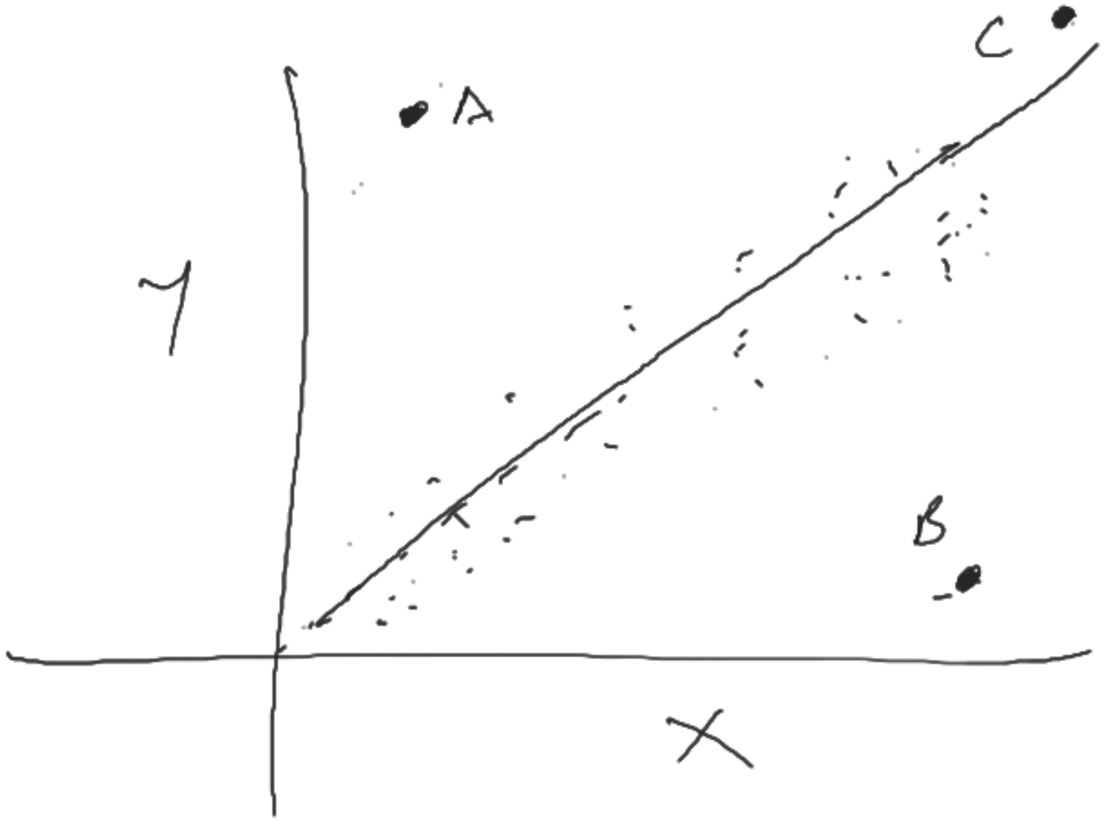
i.e. $0 \leq R^2 \leq 1$.

* Types of outliers:-

~ A - Vertical outlier

~ B - Bad leverage point.

c. Good Leverage point.



# * Model Adequacy checking :-

Assumptions :-

$x$ & $y$ are linearly related

i) $E(\varepsilon_i) = 0$

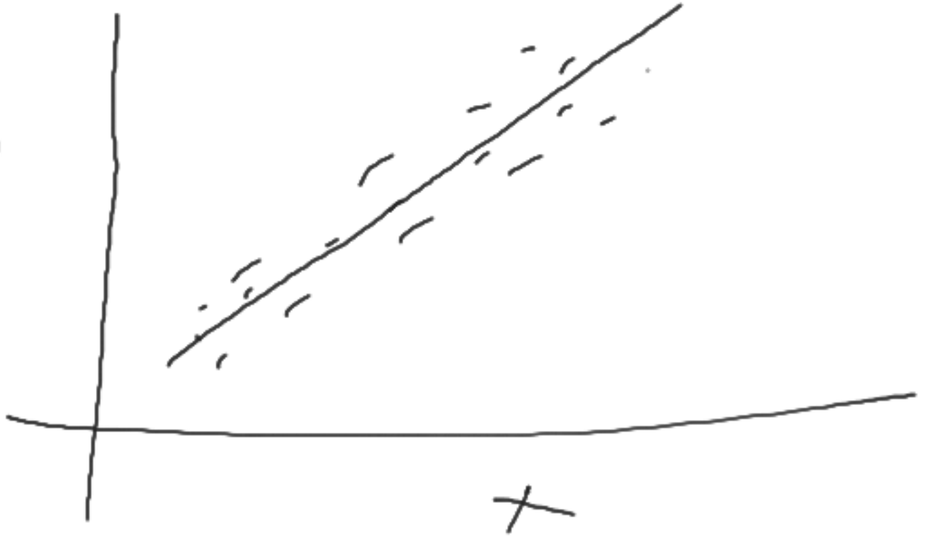
ii) $V(\varepsilon_i) = \sigma^2$

iii) $\varepsilon_i$'s are uncorrelated

iv) $\varepsilon_i \overset{i.i.d}{\sim} N(0, \sigma^2)$

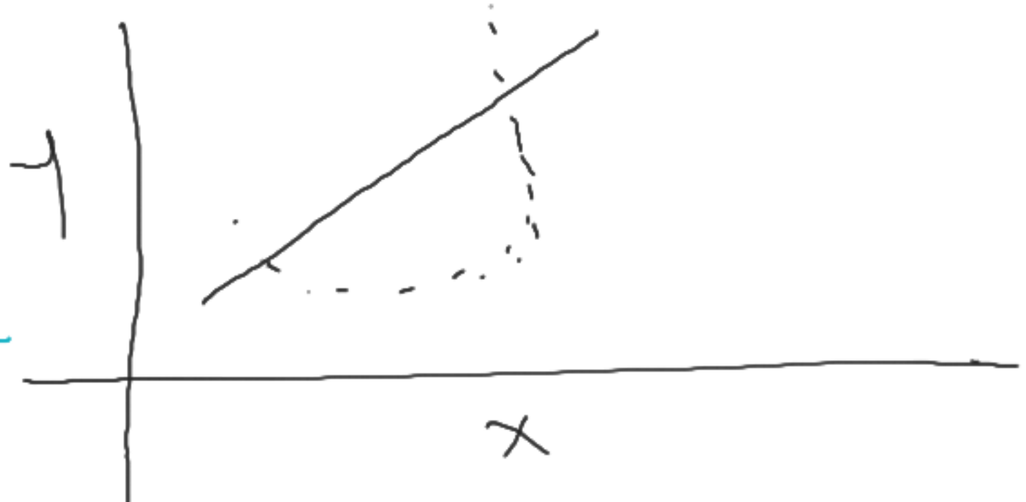
1) Linear relationship bet^n x & y.

Linear





Non-Linear




* Residual Analysis:-

i) Residuals are estimated values of errors.

ii) $E(\varepsilon_i) = 0$

iii) $\hat{\sigma}^2 = MSE = \frac{\sum (y_i - \hat{y}_i)^2}{n-2} = \frac{SS_{Res}}{n-2}$.

## + method for scaling residuals :-

1. standardized residuals :-

$$r_i = y_i - \hat{y}_i$$

$$d_i = \frac{r_i - E(r_i)}{\sqrt{MSE}} = \frac{r_i}{\sqrt{MSE}}, \quad i = 1, 2, \ldots n.$$

* $E(d_i) \approx 0$
* $V(d_i) \approx 1$

if $|d_i| > 3$, then $i^{th}$ obs$^n$ is an outlier.

(i.e. $d_i < -3$ or $d_i > 3$)

2> Studentized residual :-

$$r_i^* = \frac{r_i - E(r_i)}{\sqrt{V(r_i)}}, \quad i = 1, 2, \ldots n.$$

since,

$$V(r_i) = \sigma^2 \left[ 1 - \left( \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}} \right) \right]$$

$$r_i^* = \frac{r_i}{\sqrt{MSE\left(1 - \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}\right)\right)}}, \quad i = 1, 2, \ldots n.$$

Remarks :-

i) if $(x_i - \bar{x})^2$ is large, $r_i^*$ large

ii) if $(x_i - \bar{x})^2$ is small, $r_i^*$ small.

iii) Let $$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}$$

since, $$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

$$\therefore \quad 0 \leq h_{ii} \leq 1$$

$$\therefore \quad 0 \leq 1 - h_{ii} \leq 1$$

Hence. $$V(r_i) = MSE(1 - h_{ii}) = MSE\left[1 - \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}\right)\right] \leq MSE$$

iv) for large $n$, $h_{ii} \to 0$ & $V(r_i) = MSE$.

$$\therefore \quad d_i = r_i^*$$

# PRESS / deleted residual :- {Prediction Error Sum of Squares}

We know that,

$$r_i = y_i - \hat{y}_i$$

$$= y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$$

The PRESS residual is defined as,

$$r_{(i)} = y_i - \hat{y}_{(i)}$$

## Procedure :-

(i) Delete $i^{th}$ observation.

(ii) Fit the regression line based on remaining $(n-1)$ observations.

(iii) calculate Fitted value of $y_i$ ($\hat{y}_{(i)}$) using the fitted regression model.

(iv) calculate $r_{(i)} = y_i - \hat{y}_{(i)}$

Remarks :-

(i) To calculate the PRESS residuals for n observations, n regression lines are required. Instead of this, we can calculate PRESS residuals using the relation bet $r_{(i)}$ & $r_i$

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

(ii) Since, $0 \leq 1 - \left( \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}} \right) \leq 1$

$\therefore$ $r_i \leq r_{(i)}$

(iii) PRESS is used to identify outliers.

## Standardized PRESS residual :-

The PRESS residual is,

$$r_{(i)} = \frac{r_i}{1-h_{ii}}, \quad \text{where } h_{ii} = \left(\frac{1}{n} + \frac{(x_i-\bar{x})^2}{S_{xx}}\right)$$

$$\therefore \quad E(r_{(i)}) = 0$$

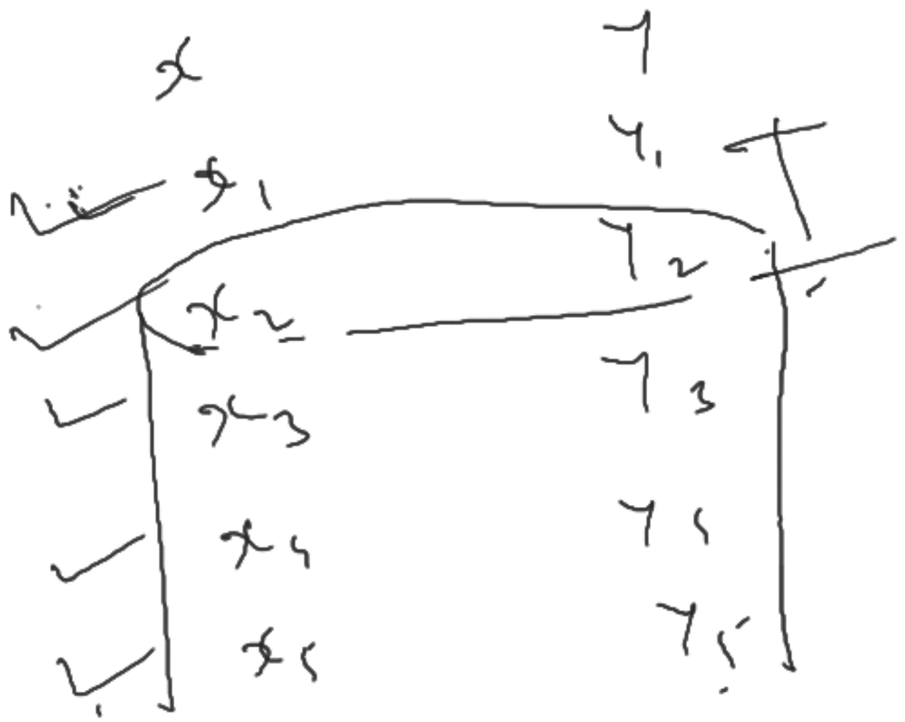
&

$$V(r_{(i)}) = \frac{1}{(1-h_{ii})^2} V(r_i) = \frac{\sigma^2(1-h_{ii})}{(1-h_{ii})^2} = \frac{\sigma^2}{1-h_{ii}}$$

Therefore the standardized PRESS residual is

$$\frac{r_{(i)} - E(r_{(i)})}{\sqrt{V(r_{(i)})}} = \frac{r_i/(1-h_{ii})}{\sqrt{\sigma^2/(1-h_{ii})}} = \frac{r_i}{\sqrt{\sigma^2(1-h_{ii})}}$$

$n = 5$




OLS Estimate

$\hat{\beta}_{10}, \hat{\beta}_{11}$

$(\hat{\beta}_{20}, \hat{\beta}_{21})$

$(\hat{\beta}_{30}, \hat{\beta}_{31})$

$(\hat{\beta}_{50}, \hat{\beta}_{51})$

$r_{(i)} = Y_i - \hat{Y}_{(i)}$

$r_{(i)} = Y_i - (\hat{\beta}_0 + \hat{\beta}_{1(i)} x_i)$

$Y_i - (\hat{\beta}_{20} + \hat{\beta}_{21} x_i)$

$Y_i - (\hat{\beta}_{30} + \hat{\beta}_{31} x_i)$

$Y_i - (\hat{\beta}_{50} + \hat{\beta}_{51} x_i)$

$x(5)$

$$r_{(i)} = \frac{r_i}{1 - h_{ii}} \qquad h_{ii} = \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}\right)$$

$$= \frac{Y_i - \hat{Y}_i}{1 - h_{ii}} = \frac{Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)}{1 - h_{ii}}$$

## R- student residual :-

The R- student residual is defined as,

$$t_i = \frac{r_i}{\sqrt{s_{(i)}^2 (1-h_{ii})}},$$

where,

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

&

$$s_{(i)}^2 = \frac{(n-2)\,MSE - r_i^2/(1-h_{ii})}{n-3}$$

Note :-

$h_{ii} \to 0$ as $n \to \infty$

# Normality of error :-

Let $r_1, \ldots r_n$ are residuals.

Procedure of Normal probability plot :-

i) Arrange residuals in an increasing order $(r_{(1)}, r_{(2)}, \ldots r_{(n)})$

ii) Calculate cumulative probabilities.

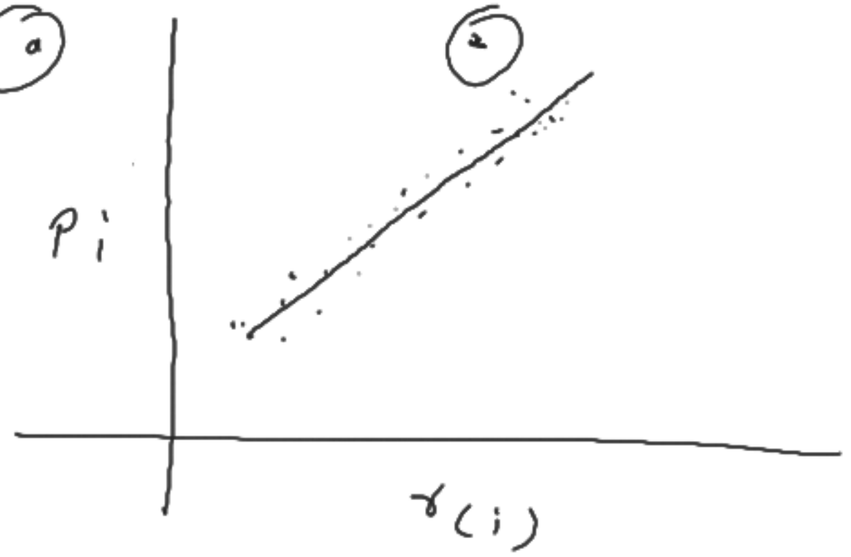
$$p_i = \frac{i - 1/2}{n}, \quad i = 1, 2, \ldots n.$$

iii) Draw the plot of $(r_{(i)}, p_i)$.

or Alternatively, draw the plot of $(r_{(i)}, \phi^{-1}(p_i))$. $(\because E(r_{(i)}) = \phi^{-1}(p_i))$

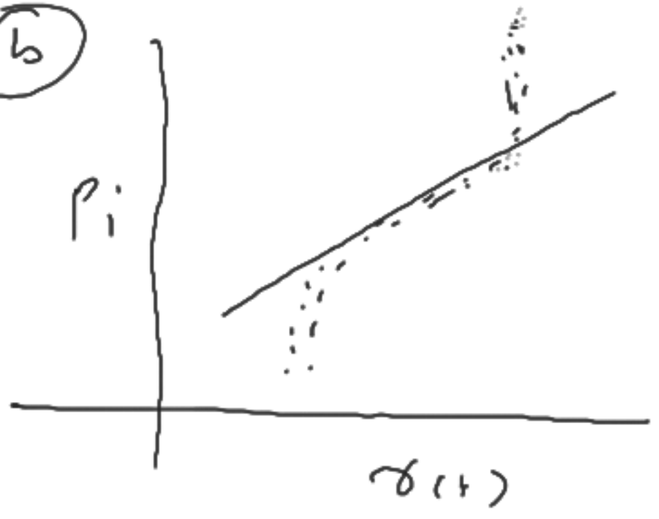
where $\phi(\cdot)$ c.d.f of $N(0,1)$

a



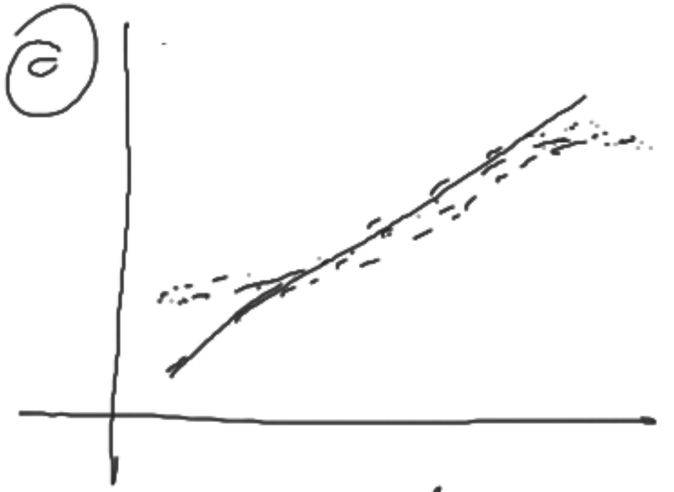

$r_i \sim$ Normal dist$^n$

b



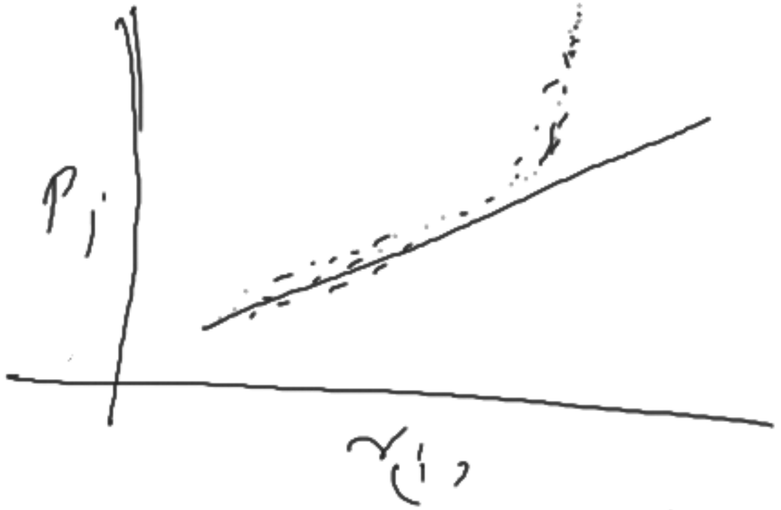

$r_i \sim$ heavy-tailed dist$^n$

c



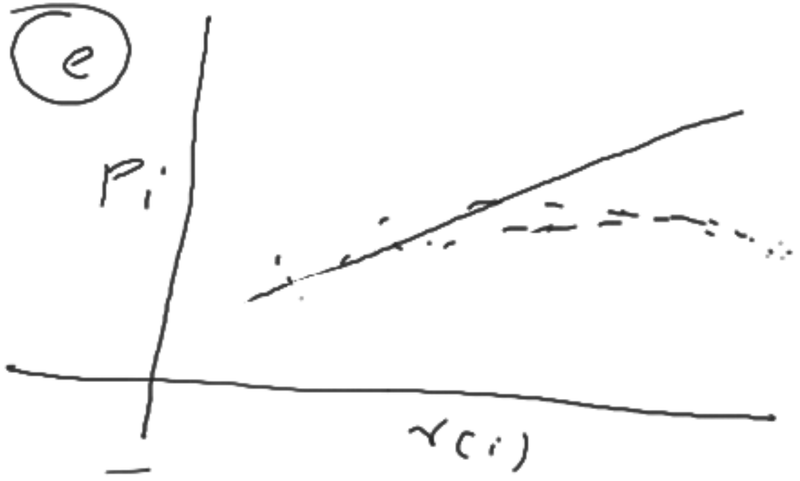
$r_i \sim$ light-tailed dist$^n$

d




$r_i \sim$ positively skewed

e

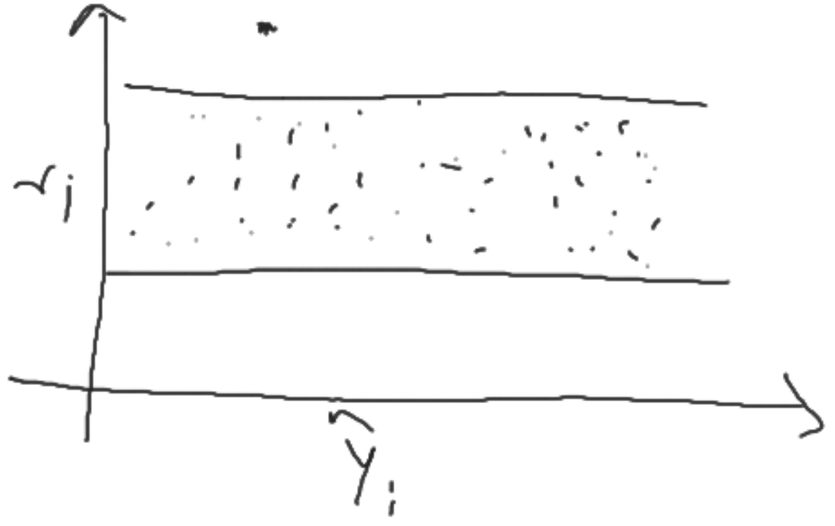



$r_i \sim$ negatively skewed

# Constant variance of Error :-

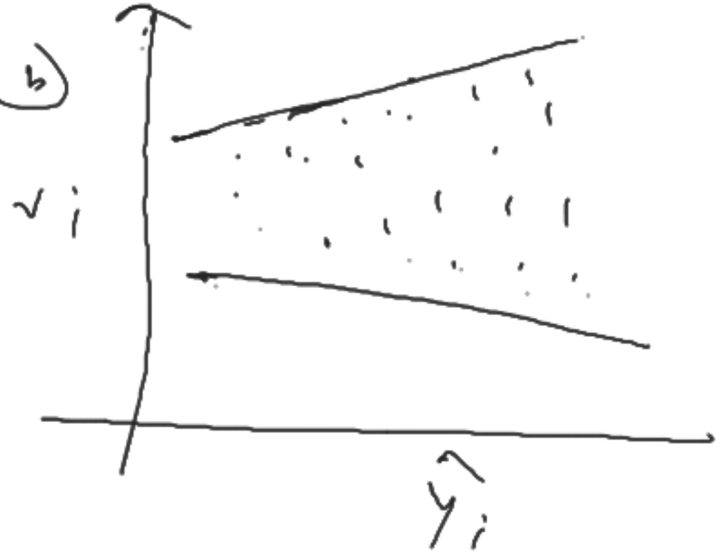
outward opening funnel

(a)




$v(\epsilon_i)$ is constant

(b)

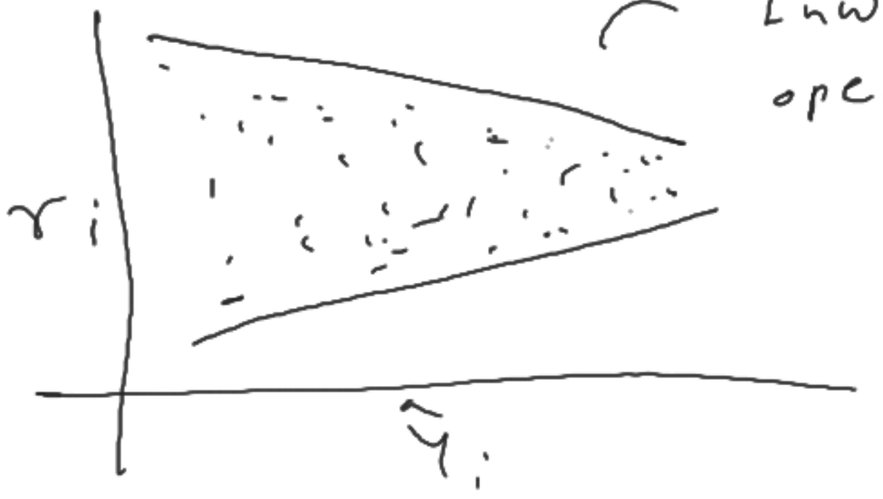



$v(\epsilon_i)$ is not constant

$v(\epsilon_i)$ is increasing

funnel

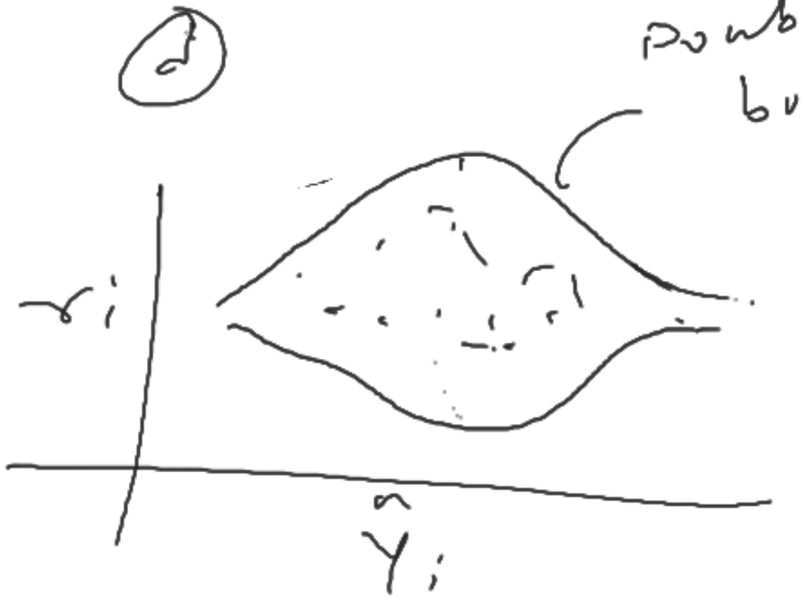
(c)




$V(\varepsilon_i)$ is not constant

$V(\varepsilon_i)$ is a decreasing fn. of $\hat{Y}_i$

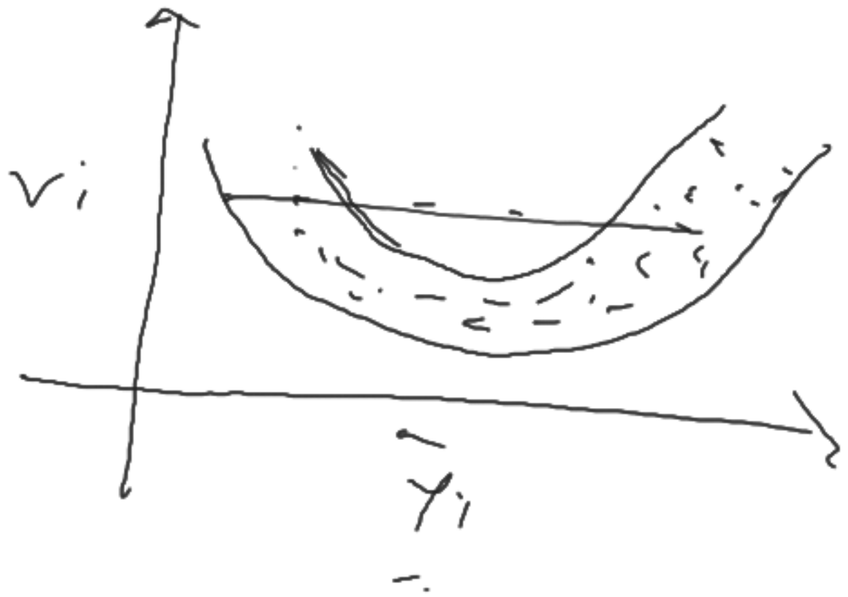
(d)




$V(\varepsilon_i)$ is not constant

In this case, $Y$ may follows Binomial dist

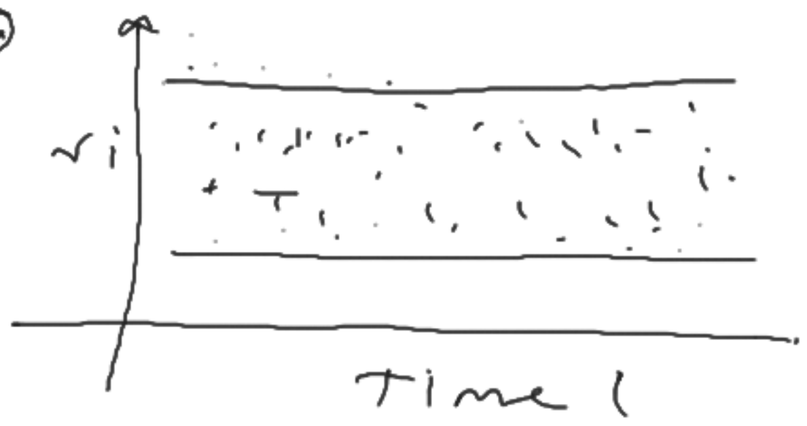
(e)



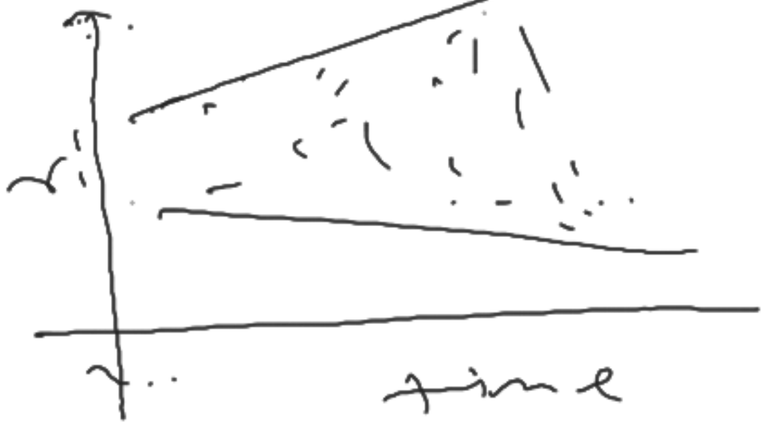

Non-linearity

Errors are uncorrelated:-

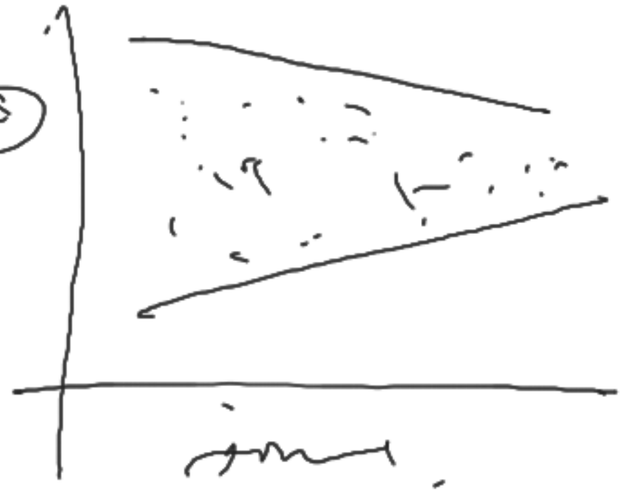
(a)




(b)




(c)

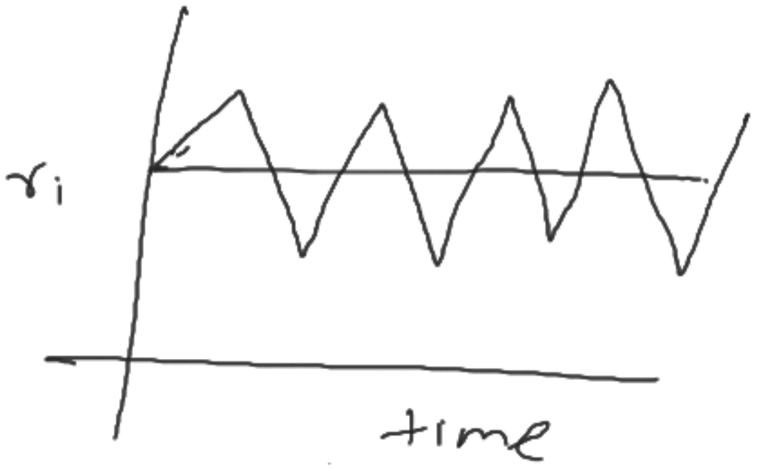


(d) Errors uncorrelated


(e)

(f)


Positive autocorrelation




Negative autocorrelation