Regression Analysis.

classmate
Date
Page

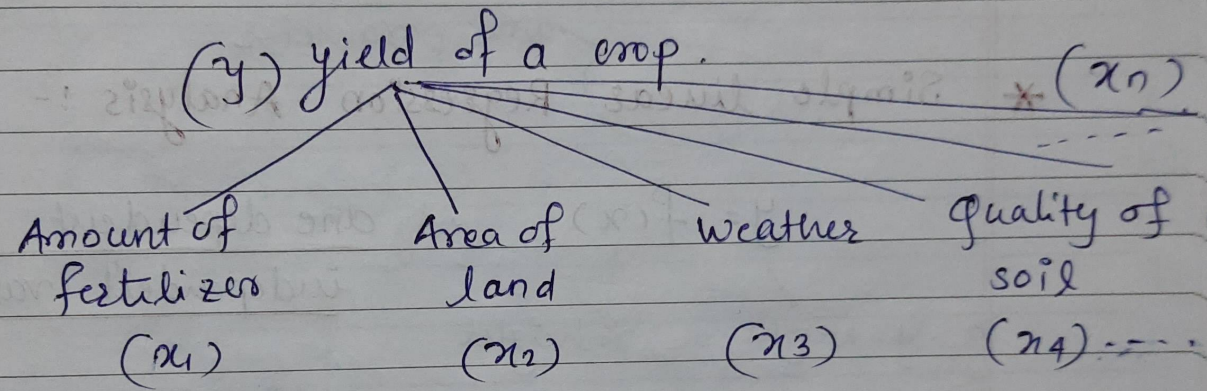
10th

Moudomery - D. C. Peck, E & Vinning, G. G (2006)

Intro to linear Regression Analysis, Wiley.

* linear Regression is technique to model relationship ~~betn ip & op.~~ & investigate

eg.



∴ y depends on $x_1, x_2, \dots x_n$. ~~and~~ all these are independent variables.

∴ There is a relation betn y and $x_1, x_2 \dots x_n$.

∴ $y = f(x_1, x_2, \dots x_n)$

↑

Aim ⇒ find out this f".

eg $y = x_1 + 2x_2 + 3x_3 + \dots + 10x_n$. ← linear Relationship

eg $y = x_1 x_2 + 2x_2^2 + 5x_3 + 6x_1 x_4$ ← Non-linear Relationship.

Target is to find out Relationship based on data.

Model ⇒ $y = f(x_1, x_2, \dots x_n)$

$x_1, x_2, \dots x_n$ ⇒ Independent variables / Regressions

y ⇒ Response variable / o/p var / dependent variable.

* Simple linear (wrt parameter) Regression Analysis :-

$y = f(x)$ only one dependent & one independent variable.

eg y = weight, x = height

(linear in terms of β_0 & β_1)

Imp → * linear wrt parameters, not variables.

$y = \beta_0 + \beta_1 x$

$y = \beta_0 + \beta_1 x^2$

$y = \beta_0 + \beta_1 e^x$.

eg $y = \beta_0 + \beta_1 x^2$

if $x^2 = z$ then

$y = \beta_0 + \beta_1 z$

eg $y = e^{\beta_0} + e^{\beta_1 x}$

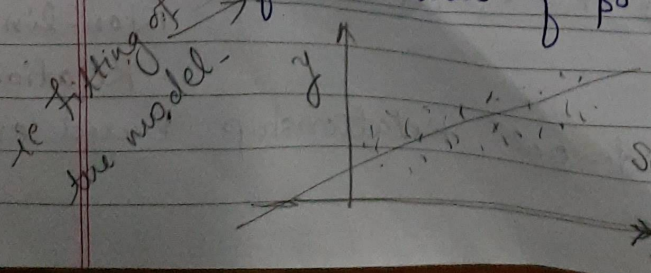
Once you have data, you plot the data & if there is linear Relationship then find values of β_0 & β_1 i.e fitting your model

Non-linear

eg $y = \beta_0 + \beta_1 e^x$

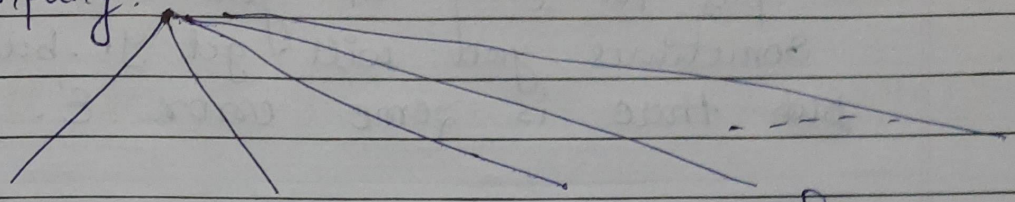
z

linear



Page No.
Date: / /

eg. Price of an end product in a manufacturing company.



Cost of ingredients (x_1) | carrying cost (x_2) | holding cost (x_3) | profit (x_4)

$$y = f(x_1, x_2, \dots x_n)$$

eg x ↑ fertilizer — y ↑ yield of crop.

$x_1 = 5\,kg$ $x_2 = 4\,kg$ ----- $x_n = 7\,kg$

$y_1 = 50kg$ $y_2 = 45kg$ $y_n = 60kg$

∴ If we take avg then we can say on an avg for 5kg fertilizer, you are getting 50 kg of crop.

$x_1 = 5$ — 50
$x_2 = 5$ — 45
$x_3 = 5$ — 55
$x_4 = 5$ → 60

$$\frac{50+45+55+60}{4}$$

$$\Rightarrow \frac{210}{4} = 52.5kg$$

on avg for 5kg

All the model might not fit

But for every x_i you dont always y_1 sometimes you will get y_1. but most time But there is some error 'ε'.

so if for $x_1 = 5kg$ you get $y_1 = 49$ kg means there is error of 1 kg $\Rightarrow \varepsilon_1 = 1kg$.

$E(y_1) \Rightarrow \beta_0 + \beta_1 x_1$

so once you fit $x=5kg$, on an avg how much yield you are getting that is expectn of y_1

Mathematic Model Vs Statistical Model

↓ Concrete

↓ Actually takes care what is happening

$y_1 = \beta_0 + \beta_1 x_1$

$y_2 = \beta_0 + \beta_1 x_2$

$\vdots$

$y_n = \beta_0 + \beta_1 x_n$

$y_1 = \beta_0 + \beta_1 x_1 + \varepsilon_1$

$y_2 = \beta_0 + \beta_1 x_2 + \varepsilon_2$

$\vdots$

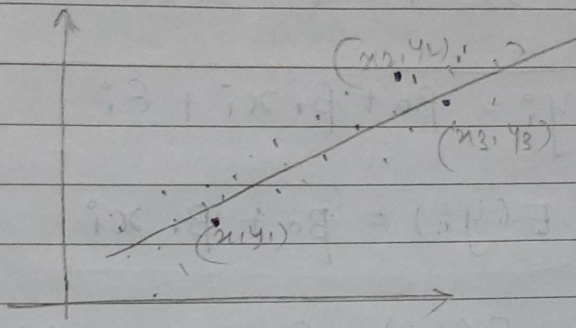
$y_n = \beta_0 + \beta_1 x_n + \varepsilon_n$ ↑ Unknown OR Random Error

ie If you fit $x_1 = 7kg$ then on an avg you will get 60kg of crop

↳ $E(y_1) = \beta_0 + \beta_1 x_1$, $E(\varepsilon_i) = 0$

Model ⇒ $y = \beta_0 + \beta_1 x$

dataset ⇒ $(x_1, y_1), (x_2, y_2) - - - - (x_n, y_n)$



If you assume that $y = \beta_0 + \beta_1 x$ is rel^n bet^n x & y then all points should lie on a line. So

The eq^n $y = \beta_0 + \beta_1 x$ should satisfy.

$$y_1 = \beta_0 + \beta_1 x_1$$
$$y_2 = \beta_0 + \beta_1 x_2$$
$$\vdots$$
$$y_n = \beta_0 + \beta_1 x_n$$

} Mathematical Model

But you don't actually get exactly response variable always rather there is some error.

$$y_1 = \beta_0 + \beta_1 x_1 + \varepsilon_1$$
$$y_2 = \beta_0 + \beta_1 x_2 + \varepsilon_2$$
$$\vdots$$
$$y_n = \beta_0 + \beta_1 x_n + \varepsilon_n$$

} Statistical Model.

Random as it is of RV

deterministic

Unknown and is Random Error (R.V)

$$E(y_1) = \beta_0 + \beta_1 x_1, \quad E(\varepsilon_1) = 0$$
$$\vdots$$
$$E(y_n) = \beta_0 + \beta_1 x_n, \quad E(\varepsilon_n) = 0.$$

Once we know parameters β_0 and β_1, will know the Relationship betⁿ x and y

So our model in general is

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

ε = error.

dataset should satisfy this model

$$E(y) = \beta_0 + \beta_1 x \text{ , i.e., } E(\varepsilon) = 0.$$

$$\therefore \quad y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad , \quad i = 1, 2, \ldots n$$

imp

so that $E(y_i) = \beta_0 + \beta_1 x_i$

where $E(\varepsilon_i) = 0$, $\forall i = 1, 2 \ldots n$

vvimp

Assumptions ⇒ Observations :-

① Random errors ε_i's are such that

$E(\varepsilon_i) = 0$ and $\text{var}(\varepsilon_i) = \sigma^2$ (variability is unknown & fixed)

② $\text{cov}(\varepsilon_i, \varepsilon_j) = 0$, $\forall i \neq j$

i.e. ε_i & ε_j are unrelated.

Error of i^{th} & j^{th} obs are not related

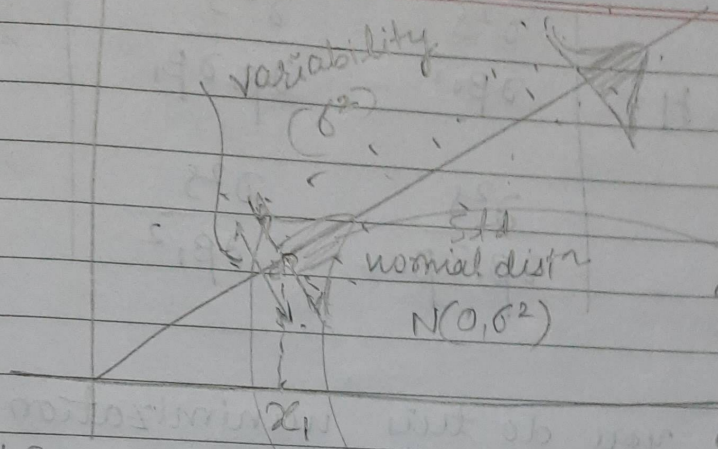
③ $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$

Errors are iid RV & has $N(0, \sigma^2)$ distribution

ie. on avg error is zero

$E(\varepsilon_i) = 0$ & $\text{var}(\varepsilon_i) = \sigma^2$

& follows normal distribution



higher prob of being on line & less probab of being off the line

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

to find β_0 & β_1 and on avg you can have ε error

* Least Square Estimation Method :- (Point Estimation)

find out β_0 and β_1 s.t. sum of the sq^r of error is Min

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} \varepsilon_i^2 \text{ is minimum}$$

(sum of squares of error is min)

theoretically all pts should lie on line

(i.e. most points should lie on the line & those not lying on line should be close to line)

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

fn of β_0, β_1 s.t. f^n should take min value

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

Should take minimum value.

This becomes problem of calculus now

$\frac{\partial S}{\partial \beta_0} = 0$ and $\frac{\partial S}{\partial \beta_1} = 0$, find critical points

and then check Hessian matrix

$$H = \begin{vmatrix} \frac{\partial^2 S}{\partial \beta_0^2} & \frac{\partial^2 S}{\partial \beta_0 \partial \beta_1} \\ \frac{\partial^2 S}{\partial \beta_1 \partial \beta_0} & \frac{\partial^2 S}{\partial \beta_1^2} \end{vmatrix}$$

If you do this minimization problem then:

2nd derivatives should come -ve &
H matrix should come +ve then you can say that pt is Local min and then check at Boundary pts & write conclusion Global minimum

After Least Sq° Estimate.

No. of Passengers (x)	Cost ($1000) (y)

→ Reln ⇒ $y = \beta_0 + \beta_1 x$

fitted Model ⇒ $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$

④ first calculate $\hat{\beta}_1$

① $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$

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

$$= \sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)$$

Date: / /

① $S_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2$

② $S_{xy} = \sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)$

③ $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$

④ $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

⑤ $\hat{\sigma}^2 = MS_{Res} = \frac{SS_{Res}}{n-2} = \frac{S_{yy} - \hat{\beta}_1 S_{xy}}{n-2}$

⑥ $SS_R = \hat{\beta}_1 S_{xy}$

⑦ $SS_T = SS_{Res} + SS_R$
$\|$
S_{yy}

⑧ $R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_{Res}}{SS_T}$

* Least Square Estimates for $\hat{\beta}_0$ and $\hat{\beta}_1$ (Imp page)

① $\hat{\beta}_{1\,LSE} = \dfrac{S_{xy}}{S_{xx}}$

② $\hat{\beta}_{0\,LSE} = \bar{y} - \hat{\beta}_1 \bar{x}$

③ $S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2$ → Sum of sq's of x_i's

④ $S_{xy} = \sum_{i=1}^{n} y_i (x_i - \bar{x}) = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$

$\therefore S_{xy} \Rightarrow \sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)$

replace y by x

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

$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad , \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

the fitted Regression model is given by.

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

fitted Regression line using least square Estimat" Method.

check example.

(point estimation)

To check How good you have fitted the model $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ check properties for LSE for that

Properties of LSE :-

(whether it is unbiased Estimator & MSE=0) $\hat{\beta}_0$ & $\hat{\beta}_1$ are least square estimators.

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

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

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

Taking Σ inside

$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$

$\sum x_i = n\bar{x}$

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

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

$$\hat{\beta}_1 = \sum_{i=1}^{n} c_i \cdot y_i \quad \text{, where } c_i = \frac{x_i - \bar{x}}{S_{xx}}$$

So this estimator $\hat{\beta}_1$ is linear in terms of observation.

c_i satisfies following properties ① $\hat{\beta}_1$ is linear Estimator

* ① $\sum_{i=1}^{n} c_i = 0$ ← ② sum of coefficient terms is zero.

* ② $\sum c_i \cdot x_i = 1$ ← ③ multiplying by independent variable & summing up will give value 1

Page No.
Date: / /

LSE is point estimation & pt estimation doesn't provide confidence or certainty.

* Point estimation ⇒ No guarantee.

For Least Sq^r check → ① Unbiasedness
② MSE

$$\hat{\beta}_1 = \sum_{i=1}^{n} c_i y_i$$

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

$$= \frac{1}{n}\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} c_i y_i \bar{x}$$

$$\hat{\beta}_0 = \sum_{i=1}^{n} d_i y_i \quad \text{where } d_i = \frac{1}{n} - c_i \bar{x}$$

so $\hat{\beta}_0$ is also linear estimator of observation.

$$\hat{\beta}_1 = \sum_{i=1}^{n} c_i y_i \quad \text{, where, } c_i = \frac{x_i - \bar{x}}{S_{xx}}$$

$$\hat{\beta}_0 = \sum_{i=1}^{n} d_i y_i \quad \text{, where } d_i = \frac{1}{n} - c_i \bar{x}$$

Both $\hat{\beta}_1$ and $\hat{\beta}_0$ are linear estimators.

$y = \beta_0 + \beta_1 x + \varepsilon$ ___ starting model with error

↓ put the data & use least square estimation to get the fitted model

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

$E(\hat{\beta}_0) = \beta_0$ and $E(\hat{\beta}_1) = \beta_1$

$\therefore$ $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased estimators of β_0 & β_1

3) $var(\hat{\beta}_1) = \frac{\sigma^2}{Sxx}$ How much $\hat{\beta}_1$ can vary

$var(\hat{\beta}_0) = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{Sxx}\right]$ How much $\hat{\beta}_0$ can vary

Gauss-Markov Thm :-

For the regression model $y = \beta_0 + \beta_1 x + \varepsilon$, with $E(\varepsilon) = 0$ and $var(\varepsilon) = \sigma^2$ and uncorrelated error $[Cov(\varepsilon_i, \varepsilon_j) = 0]$, the least squared estimators are unbiased as well as minimum variance.

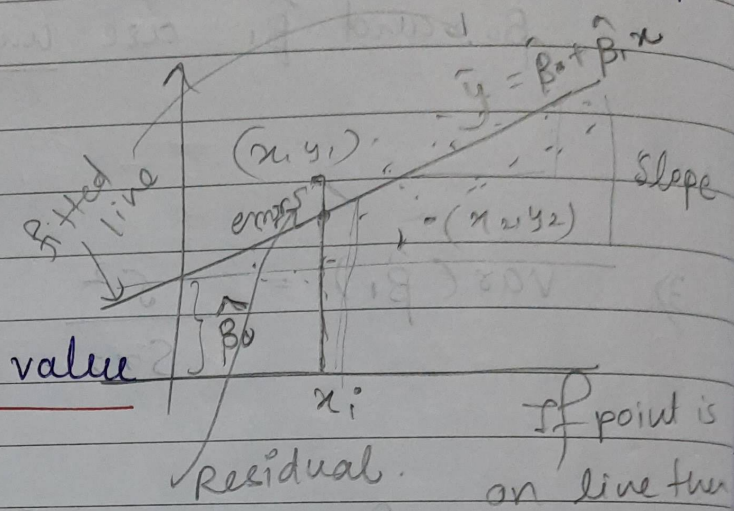
Among all unbiased estimators, least sqr estimator method has min variance .ie., we get good line so we can say its a Good Estimator

But how good you have fitted the line?

4) The sum of the residuals in any regression model that contains intercept term, is zero.

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

$\hat{\beta}_0$ ↓ intercept, $\hat{\beta}_1$ ↓ slope



Residual = obsv value − fitted value

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

If point is on line then Residual = 0

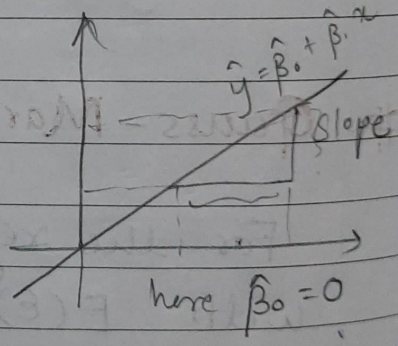
Then sum of the residuals is zero i.e.

$$\sum_{i=1}^{n} e_i = 0$$

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

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

$$\Rightarrow 0$$



$\frac{\partial S}{\partial \beta_0} = 0$ check.

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

$\frac{\partial S}{\partial \beta_0} = 0 \Leftarrow y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$ Normal eqn

5) Sum of the observed value is same as sum of fitted value.

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

$$\boxed{\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{y}_i}$$

6) The least sq^r regression line passes through centroid $(\bar{x}, \bar{y})$

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ this eq^n of line is always passing through points $(\bar{x}, \bar{y})$ which is centre of the data.

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

* All these properties are used to see how good our Estimation is. i.e. least sq^r estimator is one of the good estimator to estimate para of Regression model.

(1) Unbiasness ✓
(2) Min Variance
(3) Sum of fitted value = Sum of observed value
(4) passes through origin & data

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

classmate Date Page

12th_evening

* Estimation procedure for unknown variance σ^2

Assumptions ⇒

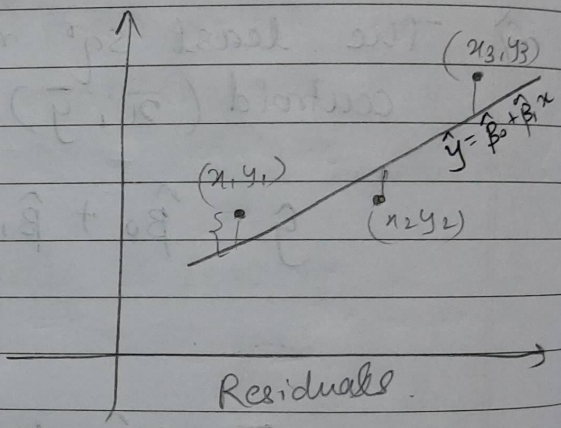
① $E(\varepsilon_i) = 0$ & $var(\varepsilon_i) = \sigma^2$ — fixed but unknown

② $cov(\varepsilon_i, \varepsilon_j) = 0$

③ $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$

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

Estimation of σ^2 :-



The sum of the square of the residual

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

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

$e_1 = y_1 - \hat{y}_1$

$\rightarrow e_2 = y_2 - \hat{y}_2$

$$\boxed{SS_{Res} = S_{yy} - \hat{\beta}_1 S_{xy}}$$

// Random Variable

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

$$S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

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

$$S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

classmate
Date
Page

Result ⇒

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

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

corollary ⇒

$$X \sim \chi^2_n$$

then $E(X) = n$.

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

$$X \sim \chi^2_{(n-2)}$$

$\therefore E(X) = n-2.$

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

$E(aX) = a\,E(X)$

$$\Rightarrow \frac{1}{n-2} E(SS_{Res}) = \sigma^2$$

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

MS_{Res}.

Mean Sq^r Residual.

$E(MS_{Res}) = \sigma^2$ where $MS_{Res} = \frac{SS_{Res}}{n-2}$

estimate for σ^2

$\therefore$ Estimate for σ^2 $\boxed{\hat{\sigma}^2 = MS_{Res}}$

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

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

How good your model is

Evaluate Model ⇒ Hypothesis testing of the Slope & intercept

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

① $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$

② $H_0 : \beta_1 = 3$ vs $H_1 : \beta_1 \neq 3$

for intercept ③ $H_0 : \beta_0 = \beta_0^*$ vs $H_1 : \beta_0 \neq \beta_0^*$

If you accept your null or reject null then what will be conclusion?

Hypothesis testing problem ⇒ 13th_morning.

* original model ⇒ $y = \beta_0 + \beta_1 x + \varepsilon$
* fitted model ⇒ $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

and assumptions are

① $E(\varepsilon_i) = 0$, $var(\varepsilon_i) = \sigma^2$

② $cov(\varepsilon_i, \varepsilon_j) = 0$

③ $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$

You want to test your belief, also cheking how good your estimator is

Ten

eg ~~$H_0 = \beta_1 = \beta_1^*$~~

Testing for Slope (β_1)

Testing problems for slope (β_1) parameter

$H_0 : \beta_1 = \beta_1^*$	$H_0 : \beta_1 = \beta_1^*$	$H_0 : \beta_1 = \beta_1^*$
vs	vs	vs
$H_1 : \beta_1 \neq \beta_1^*$	$H_1 : \beta_1 > \beta_1^*$	$H_1 : \beta_1 < \beta_1^*$

Testing Problems for intercept also same as above.

* Result ⇒

(i) $$Z_1 = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\sigma^2}{S_{xx}}}} \sim N(0,1) \qquad \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$

(ii) $$T_1 = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{MSRes}{S_{xx}}}} \sim t_{(n-2)}$$

* Testing Problems ⇒

$H_0 : \beta_1 = \beta_1^*$ vs $H_1 : \beta_1 \neq \beta_1^*$

① Testing slope (β_1) parameter :

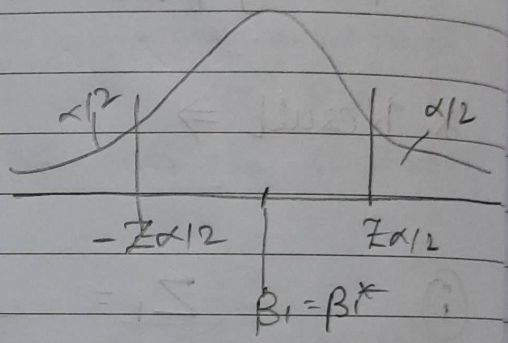
Case I ⇒ σ^2 is known

① $H_0: \beta_1 = \beta_1^*$ vs $H_1: \beta_1 \neq \beta_1^*$

case I : σ^2 is known

$$Z_1 = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\sigma^2}{S_{xx}}}} \sim N(0,1)$$

level-α-test :



Reject H_0 if

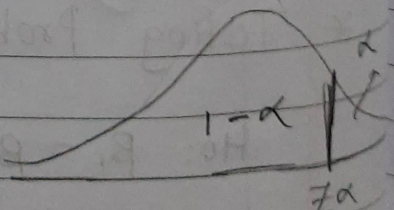
$$|Z_1| = \left| \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\sigma^2}{S_{xx}}}} \right| > Z_{\alpha/2}.$$

↳ either $Z_1 > Z_{\alpha/2}$ or

or $Z_1 < -Z_{\alpha/2}$.

P-value ⇒ $2P(Z_1 > |Z_{obs}|)$

② $H_0: \beta_1 = \beta_1^*$ vs $H_1: \beta_1 > \beta_1^*$

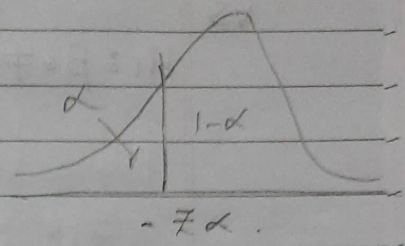


Reject H_0 if $Z_1 = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\sigma^2}{S_{xx}}}} > Z_\alpha$

P-value = $P(Z_1 > Z_{obs})$

③ $H_0 = \beta_1 = \beta_1^*$ vs $H_1 : \beta_1 < \beta_1^*$

Reject H_0 if, $Z_1 = \dfrac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\sigma^2}{S_{xx}}}} < -Z_\alpha$



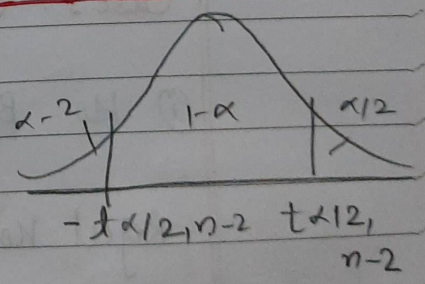
P-value = $P(Z_1 < Z_{obs})$.

* If σ^2 is unknown, we have to use estimate for σ^2 it $\hat{\sigma}^2 = MSRes = \dfrac{SSRes}{n-2} = \dfrac{S_{yy} - \hat{\beta}_1^2 S_{xx}}{n-2}$

Case II : σ^2 is unknown

① $H_0 : \beta_1 = \beta_1^*$ vs $H_1 : \beta_1 \neq \beta_1^*$

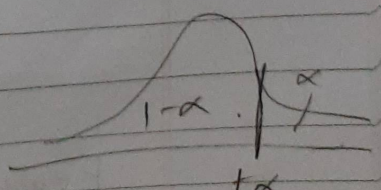
Reject H_0 if $|T_1| = \dfrac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{MSRes}{S_{xx}}}} > t_{\alpha/2, n-2}$



p-value = $2\,P(T_1 > |T_{obs}|)$

② $H_0 : \beta_1 = \beta_1^*$ vs $H_1 : \beta_1 > \beta_1^*$

Reject H_0 if $T_1 = \dfrac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{MSRes}{S_{xx}}}} > t_{\alpha, n-2}$



③ ——"—— $< t_{\alpha, n-1}$

② Testing intercept (β_0) parameter.

$H_0: \beta_0 = \beta_0^*$	$H_0: \beta_0 = \beta_0^*$	$H_0: \beta_0 = \beta_0^*$
vs	vs	vs
$H_1: \beta_0 \neq \beta_0^*$	$H_1: \beta_0 > \beta_0^*$	$H_1: \beta_0 < \beta_0^*$

Results ⇒

(i) $$Z_0 = \frac{\hat{\beta}_0 - \beta_0}{\sqrt{\sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}} \sim N(0,1)$$

σ^2 known →

(ii) $$T_0 = \frac{\hat{\beta}_0 - \beta_0}{\sqrt{MSRes\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}} \sim t_{(n-2)}$$

σ^2 unknown →

Case I: σ^2 known (same

① $H_0: \beta_0 = \beta_0^*$ vs $H_1: \beta_0 \neq \beta_0^*$

Reject H_0 if

$$|Z_0| = \left|\frac{\hat{\beta}_0 - \beta_0^*}{\sqrt{\sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}}\right| > Z_{\alpha/2}$$

Same for other cases

$$P\text{-value} = 2P(Z_0 > |Z_{obs}|)$$

Reject H_0 if

② $H_0: \beta_0 = \beta_0^*$ vs $H_1: \beta_0 > \beta_0^*$

$$Z_0 = \frac{\hat{\beta}_0 - \beta_0^*}{\sqrt{\sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}} > Z_\alpha$$

③ $H_0: \beta_0 = \beta_0^*$ vs $H_1: \beta_0 < \beta_0^*$

$$Z_0 = \frac{\hat{\beta}_0 - \beta_0^*}{\sqrt{\sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}} < -Z_\alpha$$

Case II: σ^2 is unknown

① $H_0: \beta_0 = \beta_0^*$ vs $H_1: \beta_0 \neq \beta_0^*$

Reject H_0 if

$$|T_0| = \frac{\hat{\beta}_0 - \beta_0}{\sqrt{MSRes\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}} > t_{\alpha/2,\, n-2}$$

p-value: $2P(T_0 > |T_{obs}|)$

② $H_0: \beta_0 = \beta_0^*$ vs $H_1: \beta_0 > \beta_0^*$

$$T_0 = \frac{\hat{\beta}_0 - \beta_0}{\sqrt{MSRes\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}} > t_{\alpha,\, n-2}$$

③ $H_0: \beta_0 = \beta_0^*$ vs $H_1: \beta_0 < \beta_0^*$

$$T_0 = \frac{\hat{\beta}_0 - \beta_0}{\sqrt{MSRes\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}} < -t_{\alpha,\, n-2}$$

$\beta_1 * = 0$

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

Based on data if you accept Null (H_0) means slope of ~~the~~ para is zero.

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

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

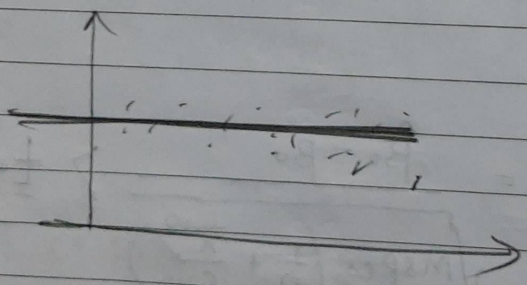
if we accept null then $\hat{\beta}_1 = 0$ slope zero.

$$\hat{\beta}_0 = \bar{y}$$

$$\therefore \boxed{\hat{y} = \bar{y}}$$ for all the points x fitted line will be some constant

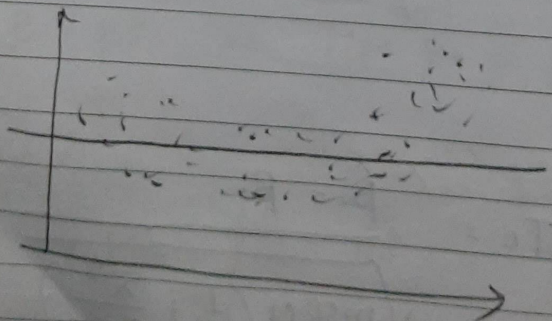
When you Accept null

C1)



C1: Regression line ||el to x-axis

C2)



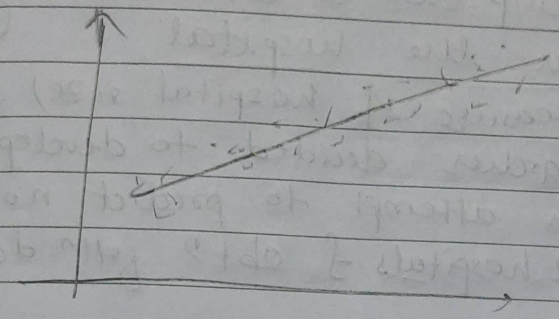
C2) data has non-linear behaviours if you are trying to fit straight line

classmate
Date
Page

13th evening

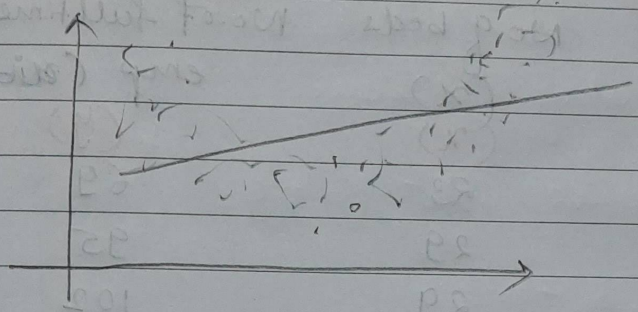
when you reject Null means slope para $\neq 0$ (β_1)

C1:



line fits the model.

C2:



non-linear data

when you accept alternative either you fitted well the model OR the model doesn't fit well but if try to fit polynomial model it might fit well

$$y = \hat{\beta}_0 + \hat{\beta}_1 x^2 + \varepsilon$$

13th evening

Ex.1) A specialist in hospital admin stated that no of full time emp can be estimated by counting no. of beds in the hospital (common measure of hospital size) A healthcare business researcher decided to develop a regression model in an attempt to predict no. of beds. He surveyed 12 hospitals & obtd foll" data.

Hospital	No. of beds (x)	No. of full time emp (y)
1	23	69
2	29	95
3	29	102
4	35	118
5	42	126
6	46	138
7	50	178
8	54	156
9	64	184
10	66	176
11	78	225
12	76	125

↑ dependent variable.

(1)

⇒ Once there is linear relationship then we fits the straight line

Least sqr estimat" ⇒

Fitted Regression line ⇒ $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \quad \text{——} ①$$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \quad \text{——} ②$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad \text{——} ③$$

putting ③ in ①

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

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

① $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$

$\Sigma x_i = 592, \quad \Sigma y_i = 1692.$

$\bar{x} = \frac{1}{n} \Sigma x_i \qquad \bar{y} = \frac{1}{n} \Sigma y_i$

$\bar{x} = \frac{1}{12} \times 592 \qquad \bar{y} = \frac{1}{12} \times 1692$

$\boxed{\bar{x} = 49.33} \qquad \boxed{\bar{y} = 141}$

② $S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = 3838.7$

$S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = 8566.$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{8566}{3838.7} = 2.232$$

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

$$\hat{y} = 141 - 2.232 (x - 49.33)$$ fitted eq^n of line.

If we have testing problem.

* Generally σ^2 are not known.

Case I $\Rightarrow$ $H_0: \beta_1 = 0$ Vs $H_1: \beta_1 \neq 0$

σ^2 unknown

$$T_1 = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{MSRes \left(\frac{1}{n} + \right)}}$$

Reject $$|T_1| = \left| \frac{\hat{\beta}_1 - \beta_1^*}{\sqrt{\frac{MSRes}{S_{xx}}}} \right| > t_{\alpha/2, n-2}$$

$$T_1 = \frac{2.232 - 0}{\sqrt{\frac{244.886}{3838.7}}} = 8.84$$

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

$$MSRes = \frac{SSRes}{n-2} = \frac{S_{yy} - \hat{\beta}_1 S_{xy}}{n-2}$$

$$= \frac{2448.86 - 2.232 \times 8566}{11}$$

$$MSRes = \frac{SSRes}{n-2} = \frac{S_{yy} - \hat{\beta}_1 S_{xy}}{n-2}$$

$\therefore MSRes = 244.886.$

~~if~~

(2) $t_{\alpha/2, n-2} \Rightarrow$

P

(3) P-value $= 2P(T_1 > |Tobs|)$

$= 2P(T_1 > 8.84)$

P-value $\equiv$ ~~2~~ close to zero

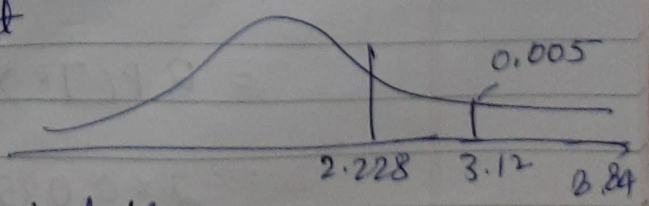
dof $= n-2$

$= 10$

If P-value $\leq \alpha$ then reject

for any α p-value is close to zero so we reject H_0.

when $\alpha = 0.025$



α will be close to zero

the slope means β_1 is non-zero quantity.

Case II: $H_0: \beta_0 = 0$ vs $H_1: \beta_0 \neq 0$ σ^2 unknown (natural scenario)

$$T_0 = \frac{\hat{\beta}_0 - \beta_0^*}{\sqrt{MSRes\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}}$$

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

$$= 141 - 2.232 \times 49.33$$

$$\hat{\beta}_0 = 30.895$$

$$\beta_0^* = 0$$

$$MSRes = 244.886$$

$$S_{xx} = 3838.7$$

$\therefore |T_0| = 2.333 \;?\; t_{\alpha/2, n-2}$? check

P-value $= 2P(T_0 > t_{obsv})$

$= 2P(T_0 > 2.33)$ at 10 dof

$= 2 \times 0.035$ $\alpha = 0.035$

$= 0.07$

If $\alpha = 0.05$

$t_{\alpha/2, n-2} = t_{0.025, 10} = 2.228.$

$\therefore T_0 = 2.333 > t_{\alpha/2, n-2} = 2.228$

So we reject Null. means intercept term $\neq 0$.

If $\alpha = 0.01$

$t_{\alpha/2, n-2} = t_{0.005, 10} = 3.169.$

$T_0 = 2.333 \not> t_{\alpha/2, n-2} = 3.169$

So we accept Null (H_0) i.e.

$$MSRes = \frac{SSRes}{n-2}$$

15th_M

* Interval Estimation in Simple linear Regression Analysis ⇒

$$y = \beta_0 + \beta_1 x + \varepsilon \quad \text{---- our model.}$$

Results ⇒ σ^2 unknown

(i) $$T_1 = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{MSRes}{Sxx}}} \sim t(n-2)$$

(ii) $$T_0 = \frac{\hat{\beta}_0 - \beta_0}{\sqrt{MSRes\left(\frac{1}{n} + \frac{\bar{x}^2}{Sxx}\right)}} \sim t(n-2)$$

Let $$MSRes = \frac{SSRes}{n-2}$$

Estimated Standard Error

$$Se(\hat{\beta}_1) = \sqrt{\frac{MSRes}{Sxx}} = \frac{Syy - \hat{\beta}_1 Sxy}{n-2}$$

$$Se(\hat{\beta}_0) = \sqrt{MSRes\left(\frac{1}{n} + \frac{\bar{x}^2}{Sxx}\right)}$$

$$\therefore T_1 = \frac{\hat{\beta}_1 - \beta_1}{Se(\hat{\beta}_1)} \quad \text{and} \quad T_0 = \frac{\hat{\beta}_0 - \beta_0}{Se(\hat{\beta}_0)}$$

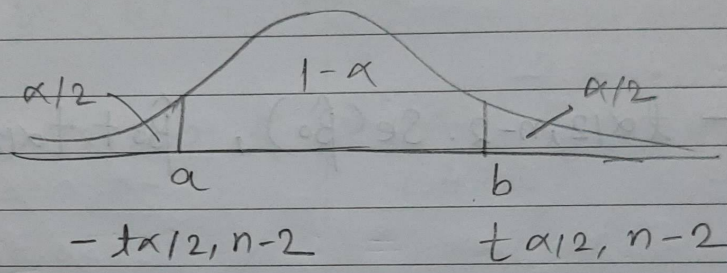
P(

inte

These T_1 and T_0 will be used as pivot quantity to estimate parameteres $\hat{\beta}_0$ & $\hat{\beta}_1$

* Interval estimation for $\hat{\beta}_1$ $\Rightarrow$ σ^2 unknown.

$100(1-\alpha)\%$ CI for parameter β_1 based on pivot quantity T_1 is given by,



$\therefore P(a < T_1 < b) = 1-\alpha$

$$P\left(-t_{\alpha/2, n-2} < \frac{\hat{\beta}_1 - \beta_1}{Se(\hat{\beta}_1)} < t_{\alpha/2, n-2}\right) = 1-\alpha$$

$$P\left(-t_{\alpha/2, n-2} \cdot Se(\hat{\beta}_1) < \hat{\beta}_1 - \beta_1 < t_{\alpha/2, n-2} \cdot Se(\hat{\beta}_1)\right) = 1-\alpha.$$

$$P\left(\hat{\beta}_1 - t_{\alpha/2, n-2} \cdot Se(\hat{\beta}_1) < \beta_1 < \hat{\beta}_1 + t_{\alpha/2, n-2} \cdot Se(\hat{\beta}_1)\right)$$

interval $\Rightarrow$ $\left(\hat{\beta}_1 - t_{\alpha/2, n-2} \cdot Se(\hat{\beta}_1),\ \hat{\beta}_1 + t_{\alpha/2, n-2} \cdot Se(\hat{\beta}_1)\right)$

* Interval Estimation for $\beta_0 \Rightarrow$

$$T_0 = \frac{\hat{\beta}_0 - \beta_0}{\sqrt{MSRes\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}} \sim t(n-2)$$

$100(1-\alpha)\%$ CI for β_0 based on T_0 is given by

$$\left(\hat{\beta}_0 - t_{\alpha/2, n-2} \cdot se(\hat{\beta}_0),\ \hat{\beta}_0 + t_{\alpha/2, n-2} \cdot se(\hat{\beta}_0)\right)$$

where, $se(\hat{\beta}_0) = \sqrt{MSRes\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}$

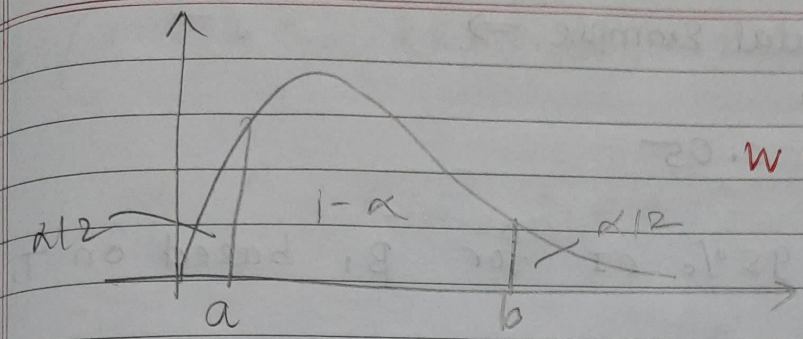
* Estimation for model parameter σ^2

$var(\varepsilon_i) = \sigma^2$ variance quantity of error

Result $\Rightarrow$

$$W = \frac{SSRes}{\sigma^2} \sim \chi^2(n-2)$$

Interval Estimation of $\sigma^2 \Rightarrow$



$$W = \frac{SSRes}{\sigma^2} \sim \chi^2_{n-2}$$

$\chi^2_{1-\alpha/2, n-2}$ $\chi^2_{\alpha/2, n-2}$

$$P(a < W < b) = 1-\alpha.$$

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

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

$100(1-\alpha)\%$ CI for σ^2 based on pivot quantity W is

$$\left(\frac{SSRes}{\chi^2_{\alpha/2, n-2}}, \frac{SSRes}{\chi^2_{1-\alpha/2, n-2}}\right)$$

where, $SSRes = S_{yy} - \hat{\beta}_1 S_{xy}$.

Interval Estimations for $\beta_1, \beta_0, \sigma^2 \Rightarrow$

Same hospital example $\Rightarrow$

① for β_1

Let $\alpha = 0.05$

$100(1-\alpha) = 95\%$ CI for β_1 based on T_1 is given by

$$\left(\hat{\beta}_1 - t_{\alpha/2, n-2}\, Se(\hat{\beta}_1),\ \hat{\beta}_1 + t_{\alpha/2, n-2}\, Se(\hat{\beta}_1)\right)$$

$\hat{\beta}_1 = 2.232$

$$Se(\hat{\beta}_1) = \sqrt{\frac{MS\,Res}{Sxx}} = \sqrt{\frac{244.886}{3838.7}} =$$

$$= \sqrt{\frac{SS\,Res}{(n-2)\,Sxx}}$$

$$= \sqrt{\frac{2448.86}{10 \times 3838.7}} \qquad MS\,Res = 244.8$$

Se = 0.08

$$\boxed{Se(\hat{\beta}_1) = 0.25}$$

$t_{\alpha/2, n-2} = t_{0.025, 10} = 2.228$

$\therefore \left(2.232 - 2.228 \times 0.25,\ 2.232 + 2.228 \times 0.25\right)$

$\Rightarrow (1.675, 2.789)$

So in pt estimation $\hat{\beta}_1 = 2.232$.

In interval estimation we can say β_1 (True para.) lies in interval $(1.675, 2.789)$ with 95% confidence.

② for β_0.

$\alpha = 0.05$, $\beta_0 = 30.895$.

$\left(\hat{\beta}_0 - t_{\alpha/2, n-2}\, Se(\hat{\beta}_0),\ \hat{\beta}_0 + t_{\alpha/2, n-2}\, Se(\hat{\beta}_0)\right)$.

$$Se(\hat{\beta}_0) = \sqrt{MSRes\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}$$

$Se(\hat{\beta}_0) = 13.25$

$(1.368,\ 33.944)$ interval for β_0

③ for σ^2 — 95% CI for σ^2.

$$\left(\frac{SSRes}{\chi^2_{\alpha/2,\, n-2}} , \frac{SSRes}{\chi^2_{1-\alpha/2,\, n-2}} \right)$$

$\alpha = 0.05$.

$SSRes = 2448.6$.

$\chi^2_{\alpha/2,\, n-2} = \chi^2_{0.025,\, 10} = 20.483$

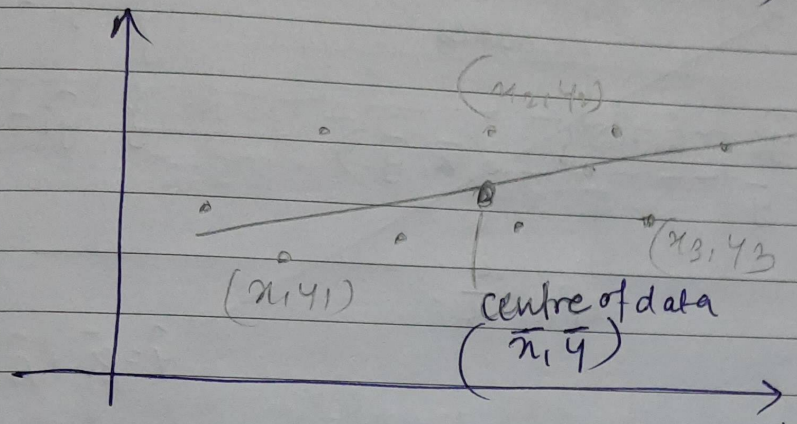
$\chi^2_{1-\alpha/2,\, n-2} = \chi^2_{0.975,\, 10} = 3.247$

interval for σ^2 (119.58 , 753.49)

15th afternoon

* Analysis of Variance (ANOVA) ⇒

(x_1, y_1) (x_2, y_2) ——— (x_n, y_n)



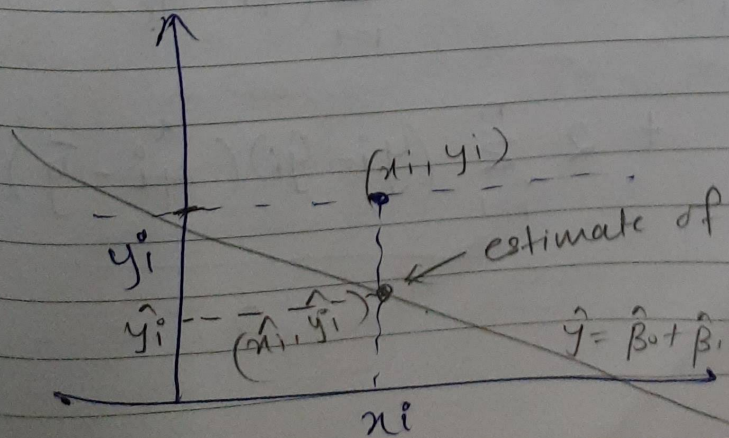
Avg Variability in data.

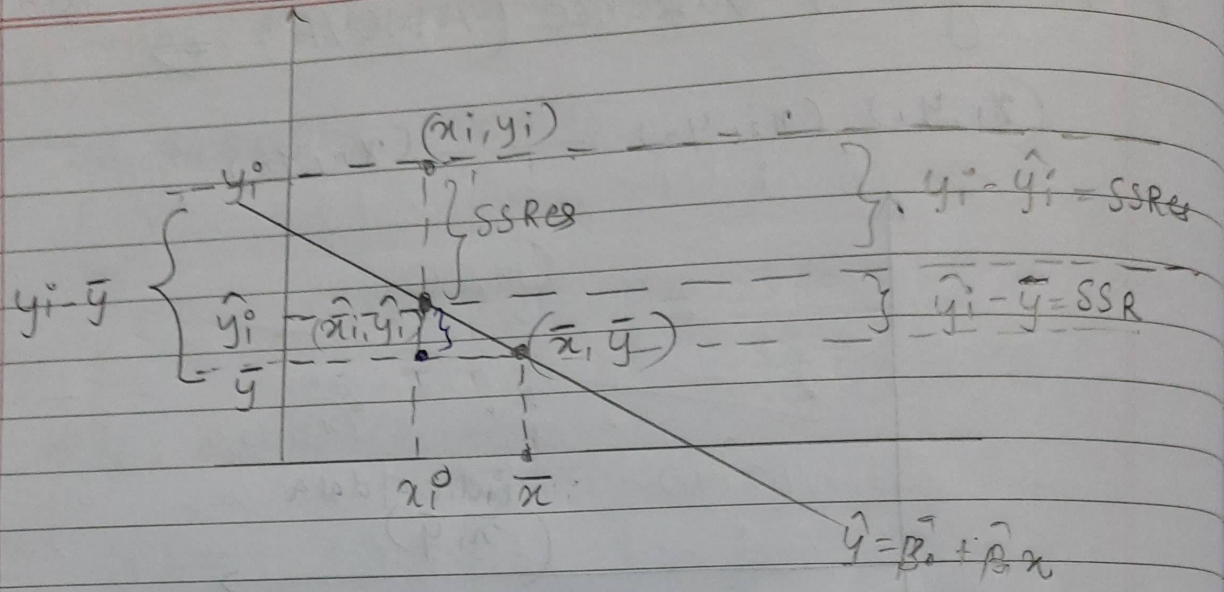
$$\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})^2$$

how far are my data pts from some centre of the data

Any Regression line always passes through centre of the data
centroid

$$\text{Total variability} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$





so $(y_i - \bar{y})$ is total variability in ith obsv.

deviation, $y_i - \bar{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$ variation of data when x is not considered.

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

Total variability of whole data

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

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

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

$$= 0$$

* Sum of the squares of Total Variation

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

$(SS_T \equiv S_{yy})$

$\Downarrow$ SSRes — Sum of the squares of residual.

$\Downarrow$ SSR — Sum of the square due to regression

$$\therefore \boxed{SS_T = SS_{Res} + SSR}$$

① $$\boxed{SS_{Res} = S_{yy} - \hat{\beta}_1 S_{xy}}$$

$$SS_{Res} = SS_T - \hat{\beta}_1 S_{xy}$$

$$= SS_T -$$

$$SSR = SS_T - SS_{Res}$$

$$= SS_T - [SS_T - \hat{\beta}_1 S_{xy}]$$

② $$\boxed{SS_R = \hat{\beta}_1 S_{xy}}$$

$$\boxed{SS_{Res} = SS_T - \hat{\beta}_1 S_{xy}}$$ ← same

* Coefficient of determination ⇒

The quantity

coeff of determinat

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSRes}{SST}$$

how good your model fits.

observations :-

① $R^2 = 1$ iff $SSRes = 0$ (when sum of sq^r of residuals is zero

when all pts lie on fitted straight line. (when obs is exactly on the line) $\sum_{i=1}^{n} (y_i - \hat{y}_i) = 0$

then Residual term = 0

So you fitted the model very well.

② $R^2 = \frac{SSR}{SST}$ ⇒ what proportion of ~~measure~~ variability is explained by the data.

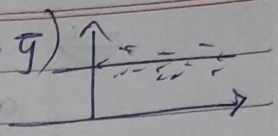
$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ variability of y by taking care of regression variables.

$SST = \sum_{i=1}^{n} (y_i - \hat{y}_i)$ variation of y's

③ $R^2 = 0$, $SSR = 0$ → (when $\hat{y} = \bar{y}$)

or

$SSRes = SST$. → 

④ $0 \leq R^2 \leq 1$ how good your model is fitted

if R^2 close to 1 – Good fit

R^2 → 0 – poor fit

$$SSR \leq SST$$

$$\frac{SSR}{SST} \leq 1.$$

Example continued. ⇒

$S_{xy} = 8566$

$SSRes = 2448.86.$

$\hat{\beta}_1 = 2.232$

$\hat{y} = 141 + 2.232 (x - 49.33)$

$SST = S_{yy} = 21,564$

$$R^2 = 1 - \frac{SSRes}{SST} \Rightarrow 1 - \frac{2448.86}{21,564} \Rightarrow 0.886$$

⇒ fitted line explains 88.6% variability of data & it cannot explain remaing 11.4% variability in data

(5) $R^2 = 1$ $\quad \hat{y} = \beta_0 + \beta_1 x + x^2$.

(6) $$E(R^2) = 1 - \frac{\sigma^2}{\hat{\beta}_1 S_{xx} + \sigma^2}$$

$S_{xx} \uparrow = E(R^2) \uparrow$

$S_{xx} \downarrow = E(R^2) \downarrow$

* Analysis of Variance (ANOVA) ⇒ F-test

① F-distribution ⇒

Let X and Y be two independent RV s.t.

$X \sim \chi^2(n)$ and $Y \sim \chi^2(m)$

$$F = \frac{X/n}{Y/m} \sim F_{n,m}.$$ (ratio of two chi-sq^r dist^{n}s)

② Non-central Chi-sq^r distribution ⇒

$$X_i \overset{iid}{\sim} N(0,1)$$

Central chi sq^r $W = \sum_{i=1}^{n} X_i^2 \sim \chi^2(n)$ ← $\mu_i = 0, \forall i$

Now if

$$X_i \overset{independent}{\sim} N(\mu_i, 1)$$

$\Rightarrow W = \sum_{i=1}^{n} X_i^2 \sim \chi^2(n)$ with non-central para. λ

Non-central chi-sq^r

$$\lambda = \sum_{i=1}^{n} \mu_i^2$$

Result ⇒

(i) $\frac{SSRes}{\sigma^2} \sim \chi^2(n-2)$ ← usual chi sq^r $\lambda = 0$

(ii) $\frac{SSR}{\sigma^2} \sim \chi^2(1)$ ← Non-central chi sq^r with non-central para $\lambda = \hat{\beta}_1^2 S_{xx}$

(iii) $F_0 = \frac{MSR}{MSRes} \sim F_{1, n-2}$

$MSR = \frac{SSR}{dof} = \frac{SSR}{1}$ mean sq^r due to regression

$MSRes = \frac{SSRes}{n-2}$ mean sq^r residual.

$MSR = SSR = \hat{\beta}_1 S_{xy}$.

original model $y = \beta_0 + \beta_1 x + \varepsilon$

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

* For Simple linear regression T-test & F-test will give same results.

* F-test ⇒

$H_0: \beta_1 = \beta_1^*$ vs $H_1:$

(1) $H_0: \beta_1 = 0$ vs $H_1: \beta_1 \neq 0$.

level-α-test ⇒

Reject H_0 if

$$F_0 = \frac{MSR}{MS\,Res} > F_{\alpha, 1, n-2}.$$

table F_{α, v_1, v_2}

* ANOVA Table ⇒

Source of Variation	Sum of squares	DOF	Mean Square	Fo
① Regression	$SS_R = \hat{\beta}_1 S_{xy}$	(1)	$MS_R = \frac{SS_R}{1}$	$F_0 = \frac{MS_R}{MS_{Res}}$
② Residual	$SS_{Res} = SS_T - \hat{\beta}_1 S_{xy}$ (where $SS_T = S_{yy}$)	$(n-2)$	$MS_{Res} = \frac{SS_{Res}}{n-2}$	
Total	$SS_T = SS_R + SS_{Res}$	$n-1$		

Ex.1) Hospital Example ⇒

ANOVA table ⇒

$$SS_R = \hat{\beta}_1 S_{xy} = 2.232 \times 3838.7 = 19119.3$$

$$SS_{Res} = S_{yy} - \hat{\beta}_1 S_{xy} = 2444.86$$

$$MS_R = SS_R = 19119.3$$

Source of variation	Sum of Squares	DOF	mean sq	Fo
1) Regression	$SSR = 19119.3$	1	$MSR = 19119.3$	$Fo = \frac{19119.3}{244.46} = 78.973$
2) Residual	$SS_{Res} = 2444.86$	10	$MS_{Res} = 244.446$	

MSR
MS Res

If $\alpha = 0.05$

$F_{\alpha,1,10} \Rightarrow F_{0.05,1,10} \Rightarrow 4.96.$

$\therefore Fo = 78.973 > F_{\alpha,1,10} = 4.96.$

$\therefore$ Reject H_0 i.e. $\beta_1 \neq 0$

Result $\Rightarrow$ Relation bet^n T-test and F-test $\Rightarrow$

$X \sim t(n)$ then $X^2 \sim F_{1,n}$ (dOF)

$$T_1 = \frac{\hat{\beta}_1 - \beta_1^*}{\sqrt{\frac{MS_{Res}}{S_{xx}}}} \sim t(n-2)$$

$T_1 = 8.84$

Imp → $T_1^2 = (8.84)^2 = 78.973 = F_0$

$T_{obs} > t_{\alpha/2, n-2}$

If $\alpha = 0.05$

$t_{0.025, 10} = 2.228$.

$T_1 = 8.84 > t_{\alpha/2, n-2} = 2.228$

∴ We reject H_0

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

OR $H_0: \beta_1 = 0$ vs $H_1: \beta_1 > 0$

} F-test not applicable.

F-test is applicable only for two sided hypothesis.

* Application of Simple linear Regression Analysis ⇒

corresponding to a regression (x) variable, how do you predict the mean response value.

our model ⇒ $y = \beta_0 + \beta_1 x + \varepsilon$

fitted model ⇒ $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

If $x = x_0$ (if $x = 40$ beds in hospital) then what is $E(y/x=x_0)$ (no of full time Emp on an avg when $x = 40$ beds)

$$\widehat{E(y/x=x_0)} \equiv \hat{\mu}_{y/x_0}.$$

① Point Estimation ⇒

$$y = \beta_0 + \beta_1 x_0 + \varepsilon$$

$$\hat{\mu}_{y/x_0} = \widehat{E(y/x=x_0)} = \beta_0 + \beta_1 x_0 \quad (\because E(\varepsilon)=0$$

$$\mu_{y/x_0} = E(y/x=x_0) = \beta_0 + \beta_1 x_0 \quad (\because E(\varepsilon)=0$$

$$\boxed{\hat{\mu}_{y/x_0} = \widehat{E(y/x=x_0)} = \hat{\beta}_0 + \hat{\beta}_1 x_0}$$

point of Estimation of mean response of y given $x = x_0$

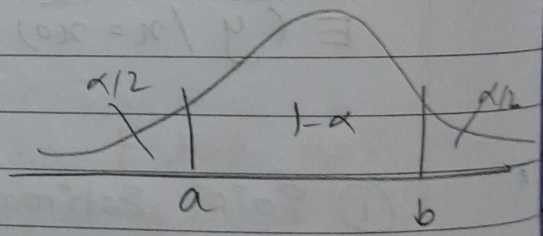
② Interval Estimation

If $x = 40$ Beds then what is 99% chance that or on an avg there will be how many no of employees.

Result ⇒

$$\text{pivot, } T = \frac{\hat{\mu}_{y/x_0} - \mu_{y/x_0}}{\sqrt{MSRes\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}} \sim t(n-2)$$

$$P(a < T < b) = 1 - \alpha$$



$$P\left(-t_{\alpha/2,n-2} < \frac{\hat{\mu}_{y/x_0} - \mu_{y/x_0}}{\sqrt{MSRes\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}} < t_{\alpha/2,n-2}\right) = 1 - \alpha$$

$-t_{\alpha/2,n-2}$ $\quad$ $t_{\alpha/2,n-2}$

$$P\left(\hat{\mu}_{y/x_0} - t_{\alpha/2,n-2}\sqrt{MSRes\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)} < \mu_{y/x_0}\right.$$

$$\left. < \hat{\mu}_{y/x_0} + t_{\alpha/2,n-2}\sqrt{MSRes\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}\right) = 1 - \alpha$$

$100(1-\alpha)\%$ PCI for μ_{y/x_0} based on t is

$$\left(\hat{\mu}_{y/x_0} - t_{\alpha/2, n-2} \sqrt{MSRes\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}, \right.$$

$$\left. \hat{\mu}_{y/x_0} + t_{\alpha/2, n-2} \sqrt{MSRes\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)} \right)$$

Prediction of Mean Response corresponding to a given value $x = x_0$

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

Example continued.

Q. If there are 40 beds then how many full time employees ?

→ ① point estimate

$$\hat{\mu}_{y/x_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

$$= 30.895 + 2.332 \times 40$$

$$= 120.17 \approx 120 \text{ full time employees}$$

② Interval estimation ⇒

95% CI then $\alpha = 0.05$

$t_{\alpha/2, n-2} = t_{0.025, 10} = 2.228$

$n_0 = 40$, $\bar{x} = 49.33$, $\bar{y} = 141$, $S_{xx} = 3838.7$,

$\hat{\beta}_1 = 2.232$, $\hat{\beta}_0 = 30.895$.

$MS_{Res} = 244.886$.

$\hat{y}_{y/x_0} = 120.17$

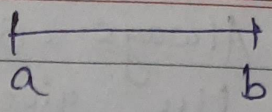
find values

(108.82, 131.52)

≈ (109, 132) no of full time employes

* Observations

The length of CI is

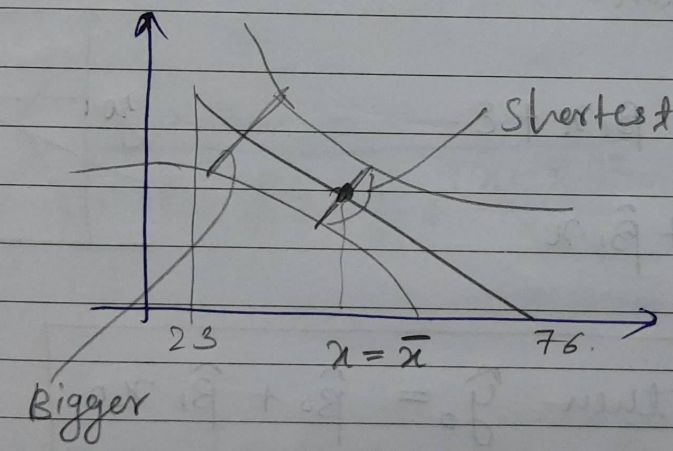


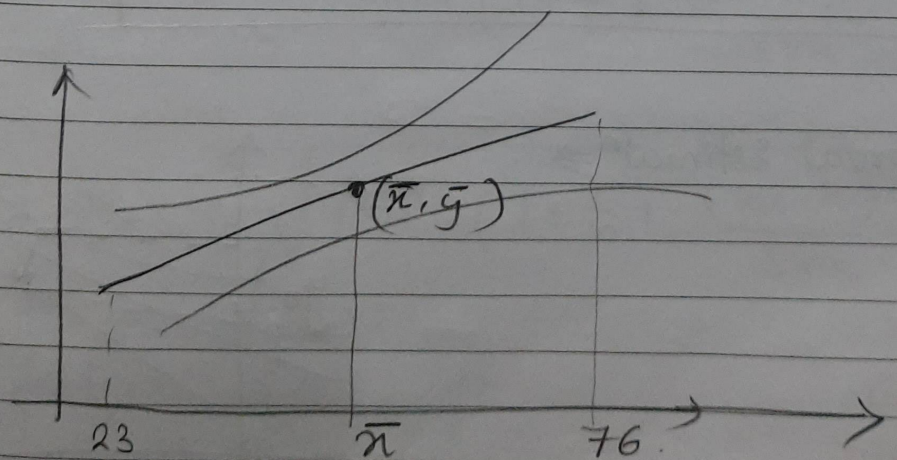
$$(b-a) = 2\, t_{\alpha/2,n-2} \sqrt{MSRes\left(\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}\right)}$$

$\therefore (b-a) =$

Imp * Then length of CI will be minimum when

$x_0 = \bar{x}$





for some point outside the range, there is chance of more error.

Always suggested to do prediction in the given value.

prediction of New Observation ⇒

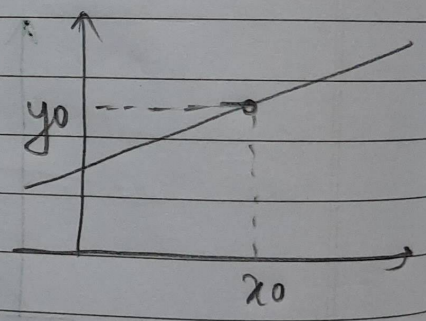
Predict future obs.

If $x = x_0$ then predict $y = y_0$? (not asking mean $E(y/x = x_0)$)

① point Estimation ⇒

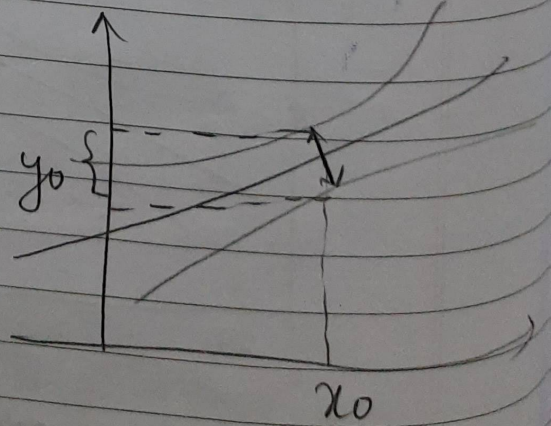
$$y = \beta_0 + \beta_1 x + \varepsilon.$$

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



$x = x_0$, then $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$

② Interval Estimatn ⇒

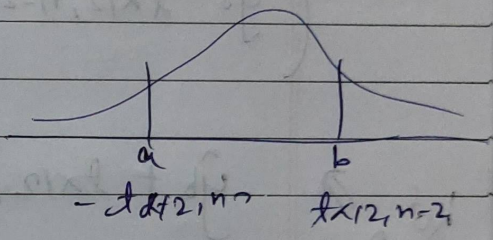


Result ⇒

$$T^* = \frac{\hat{y}_0 - y_0}{\sqrt{MSRes\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}} \sim t(n-2)$$

$100(1-\alpha)\%$ CI based on T^* is given as.

$$P(a < T^* < b) = 1 - \alpha.$$



$$P\left(-t_{\alpha/2, n-2} < \frac{\hat{y}_0 - y_0}{\sqrt{MSRes\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}} < t_{\alpha/2, n-2}\right)$$

$$P\left(\hat{y}_0 - t_{\alpha/2, n-2}\sqrt{MSRes\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}\right.,$$

$$\left.\hat{y}_0 + t_{\alpha/2, n-2}\sqrt{MSRes\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}\right)$$

interval for y_0

Called as prediction interval.

eg $x = 40$

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x = 120.17$$

predict interval for given $x = 40$ with 95% confidence based on T^* is

$$\left(\hat{y}_0 - t_{\alpha/2, n-2} \sqrt{MSRes \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}, \right.$$

$$\left. \hat{y}_0 + t_{\alpha/2, n-2} \sqrt{MSRes \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)} \right)$$

for y_0

$$\Rightarrow (83.5, 156.84)$$

CI for mean response $\Rightarrow$ (109, 132)

interval length = 23

prediction interval $\Rightarrow$ (156.84, 83.5)

interval leng = 73.34

predict int always wider than CI.

for $x = x_0$

or mean cannot vary much as compared to value corresp. to given prediction interval.