c.w
20/12/24

Paper Source
Subject.................................
Date :......................Time:..................

Inf. Stat → CI, Test of hypothesis

chi square → non parametric

z, t → Parametric

Mean test, μ : দুইভাবে করা যায়

(Parametric / Non-parametric)

(Sample came from Natural Population. Sample আসে ~~Not~~ Normal distribution থেকে

Mean → ২+ mean (anova)
Mean → ~~sig~~ single mean z/t
Mean → two mean

anova → analysis of variance
→ F-test (fisher)

সবগুলোই parametric
আবার non-parameter version ও আছে

$\hat{s} = s$

$\hat{\mu} = \bar{x}$

z depend α

t depend α, df

z always horizontal way তে চলে

এটা এভাবে টান দিতে হবে, মাঝ থেকে টান দেয়া যাবে না,

Conclusion: We have sufficient evidence that statistically and significantly the true mean no. of children in BD Homes is not equal 3 at 5% level of significance.

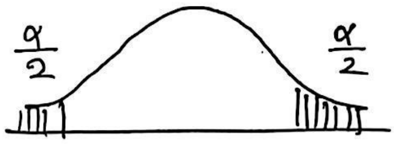
~~Et depends~~.

Example: Two-tail Test

$H_0 : \mu = 168$

$H_1 : \mu \neq 168$

কাজ sample নিয়ে hyp করি population দিয়ে

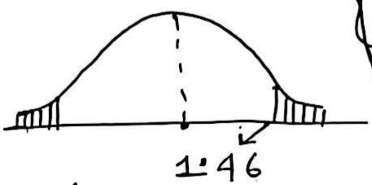


here, α is not given, we have to perform t test.

$$t = \frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}} = \frac{172.50-168}{\frac{15.40}{\sqrt{25}}} = 1.46 = t_{cal}$$

t_{tab} এ negative value নাই

$t_{tab} = 2.063$



Here, t_{cal} falls into the acceptance region on H_0, So we may not reject H_0. We have sufficient evidence that statisticals and significantly. The avg cost of blood test in a hospital in TK. 168. at 5% confidence level.

Two tail test: t test → independent

Paired / Matched

কিছু ক্ষেত্রে Match করতে হবে

Example: Matched কিভাবে হলো?

difference দুইটাই করা যাবে, (b-A)/(A-b)

μ_x = before

μ_y = after

$$\bar{d} = \frac{\sum d_i}{n} = -4.2$$

$$S_d = \sqrt{\frac{\sum(d_i-\bar{d})^2}{n-1}} = 5.67$$

$\mu_1 = \mu_2$ difference নাই

$\mu_1 \neq \mu_2$ diff আছে

$H_0 : \mu_x - \mu_y = 0$

$H_1 : \mu_x - \mu_y \neq 0$

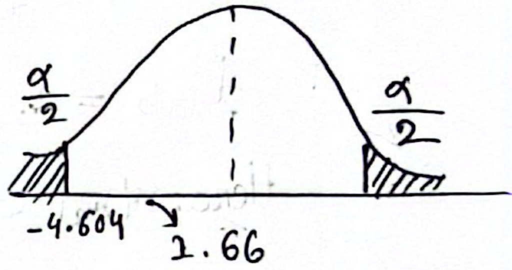
$\alpha = 0.01 \quad \bar{d} = -4.2$

here, we are performing two means test and samples came from the matched population. So, we have to perform matched/paired t-test.

$$t = \frac{\bar{d} - D_0}{S_d/\sqrt{n}} = \frac{-4.2-0}{5.67/\sqrt{5}} = -1.66 = t_{cal}$$

$\boxed{\alpha,\ df}$ → $\frac{\alpha}{2} = 0.005$, → 4

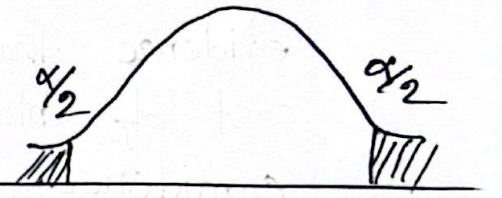
here t_{cal} falls into the ~~expectance~~ accepted region of H_0, So, we may not reject the H_0. We have ......, the training has not made a difference in the number at 0.01 (1%) significance level



Draw the hypothesis:

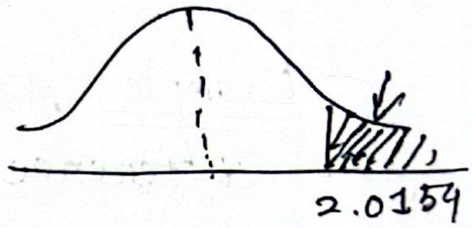
$H_0: \mu_D - \mu_S = 0$

$H_1: \mu_D - \mu_S \neq 0$



So, we have two means test and sample came from the independent population. So we have to perform independent t test.

$$t = \frac{(\bar{x} - \bar{x}_i) - \overbrace{(\mu_1 - \mu_2)}^{D_0}}{\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = 2.040 = t_{cal}$$

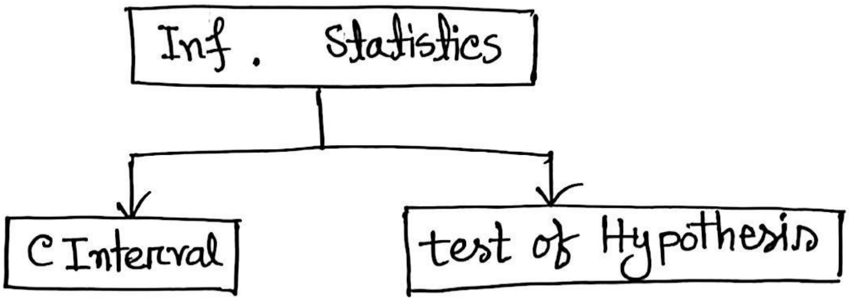


$t_{tab} = \frac{\alpha}{2}$, $df\ (n_1 + n_2 - 2) = 2.0154$

here, t_{cal} falls into the rejection region of H_0. So we may the H_0. We have ......, There is difference in ava number of dying at 5% significance level.

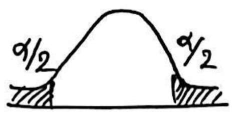
C.W 18/12/2024

Hypothesis Testing 1



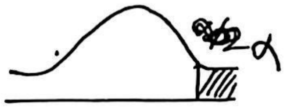
(Null) $H_0 : \mu_B = \mu_G$

(Alternative) $H_1 / H_a : \mu_B \neq \mu_G$



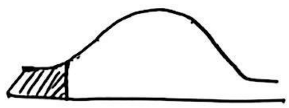
(two tail)

$\mu_B > \mu_G$



(one tail/ right/upper tail)

$\mu_B < \mu_G$



(left tail/ one tail/lower tail)

Null may be rejected ✓	
Null may not be rejected ✓	Null accepted ✗

- There are two types of hypotheses
  Null (H_0)
  Alternative (H_a)
- Hypothesis must be based on the parameter (μ).
- Only tail depends on H_1 / H_a. Everything else

depends on H_0.

- Decision should be based on H_0.
  - (i) H_0 may be rejected
  - (ii) H_0 may not be rejected.

α = level of significance

statistically and significantly

type 1 error	type 2 error
(b) Innocent → Guilt	(a) Guilt → Innocent

The probability of type-1 error α.
" " " tyepe-2 " β.

$1-\alpha = CI$ (সত্যকে সত্য বলা)

$1-\beta$ = Power of test (মিথ্যাকে মিথ্যা বলা)

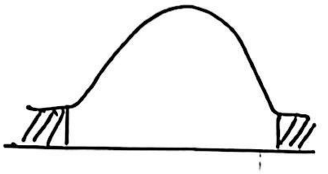
one mean test:

- σ দেওয়া থাকলে z test. Else t-test

- Null hypothesis কখনোই $\neq$ নিতে পারবে না,
- H_0 only $=, \geq, \leq$
- $H_0 : \mu = 3$
  $H_1 : \mu \neq 3$

(I) Step 1: draw the hypothesis. $H_0 : \mu = 3$
$H_1 : \mu \neq 3$

(II) Step 2: draw the picture and fix the tail



$H_1 : \mu \neq 3$

if α is not given, $\alpha = 5\%$

(III) Step-3: fix the test

Here σ is given and it is a single mean test. So we have to perform z test.

$\mu_0 = 3$

$$\therefore z_{cal} = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} = -2$$

Step-4:

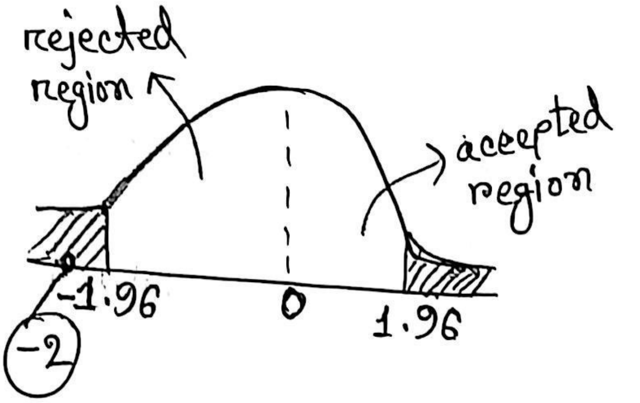
(IV) Find the z_{tab}

z depends on α.

~~$\alpha = 5\% = 0.05$~~

(two tail) $\alpha/2 = 2.5\% = 0.0250$

Probability $= -1.96$



(V) Decision: Here z_{cal} falls into the rejection region of Ho. So, we may reject the null Hypothesis

(VI) Conclusion: True mean number of children in BD home is not equal to 3

C.W 29/12/2021

Paper Source
Subject: CSE303
Date: 23/12/24 Time: 9:25

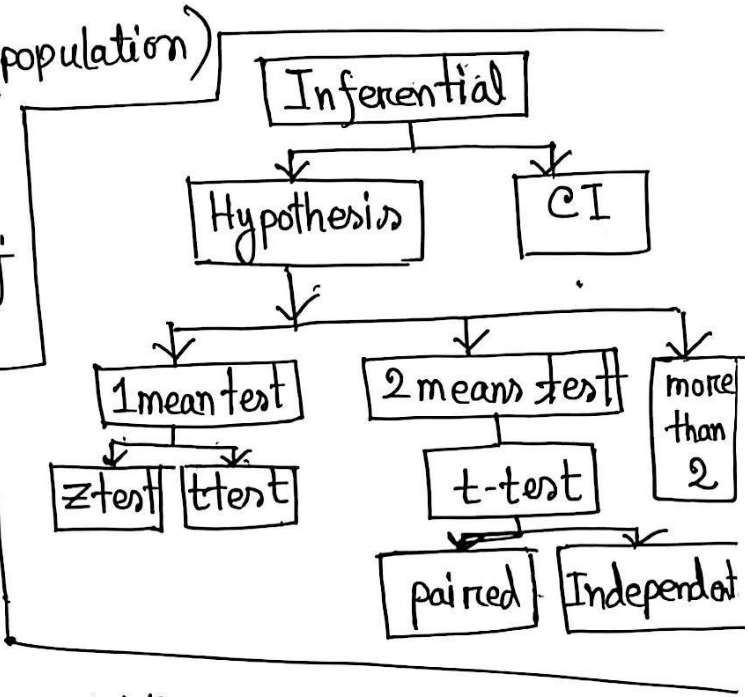
ANOVA (Analysis of variance):

- ANOVA - Analysis of variance
- mean test
- পূর্বশর্ত : Normal Distribution (population)
- $H_0 : \mu_1 = \mu_2 = \mu_3$

  $H_1 : \mu_i \neq \mu_j ;\ i \neq j$

  Any μ is different

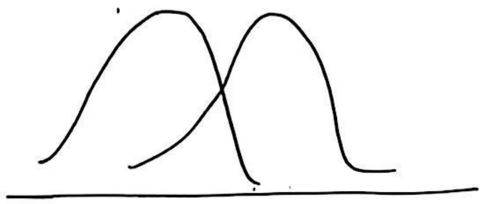
$x \sim N(\mu, \sigma^2)$



Hypothesis of one-way ANOVA:-

~~$H_1 : \mu_1 \neq \mu_2$~~ $H_0 : \mu_1 = \mu_2 = \mu_3 = \dots = \mu_K$

H_1 : Not all μ_i are same



Variability:-

Partition ~~the~~ Variation: $SST = SSW + SSG$

SST ↓ Total sum of square

SSW ↓ within group

SSG ↓ group to group

$\therefore SST = \sum_{i=1}^{k}\sum_{j=1}^{n}(X_{ij} - \bar{X})^2$ ~~observati~~ overall sample mean

$\therefore SSW = \sum_{i=1}^{k}\sum_{j=1}^{n}(X_{ij} - \bar{X}_i)^2$

R.A. Fisher

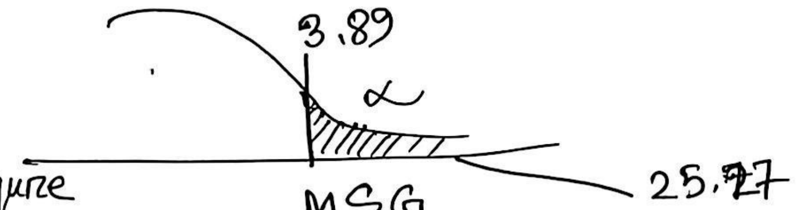
$\therefore$ F test , $MSW = \frac{SSW}{n-K}$ degress of freedom

- One-way ANOVA Table

$z^2 \sim \lambda^2_{(1)}$
$z_1^2 + z_2^2 \sim \lambda^2_{(2)}$

- F_{cal}

F_{tab}



Ftest = 2টি Chisquare এর ratio

$Ftest = \frac{MSG}{MSW}$

☐ Example

Drug 1 mean , $\mu_1 = 249.2$

Drug 2 mean , $\mu_2 = 226.0$

Drug 3 mean , $\mu_3 = 205.8$

$SSW = 1119.6$
$SSG = 4716.4$
$\therefore F_{cal} = \frac{4716.4}{1119.6} \doteq 25.28$

$H_0 : \cancel{\mu_1 \neq \mu_2 \neq \mu_3}\ \mu_1 = \mu_2 = \mu_3$

$H_1 : \mu_i \neq \mu_j \; ; i \neq j$

any μ is different

F_{tab} : 3.89

$F_{cal} > F_{tab}$

Paper Source
Subject...
Date :..............................Time:...................

Here, ~~We~~

we have sufficient evidence.......

So, we may reject H_0. (There is a difference . . . .

(write from question)

C.W 06/01/2025

Paper Source
Subject.............
Date :............Time:...........

Inferential

Correlation
↓
Simple linear Regression
↓
Multiple linear Regression
↓
nonlinear logistic Regn } dep cat.

} ~~Inferential~~ dep1

Classical stat → PCA } dep 1+
↑
classification

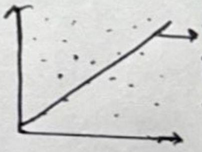
ROC

ML
↓
overall
↓
SVM (Classification test)

- Prediction is the key task of statistics:

Correlation : x এর change এর জন্য y কী change হচ্ছে?

Regression: কতটুকু change হচ্ছে?



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

1 dep cont

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \longrightarrow \text{Simple}$$

1 indep cont/cat

$$y_i = b_0 + b_1 x_i + e_i$$

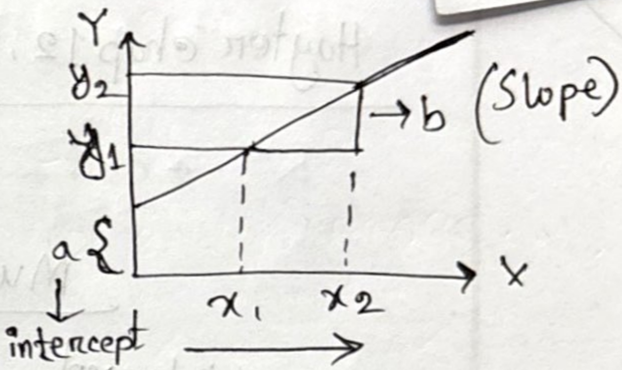
OLS - ordinary ~~test~~ least square (OLS)

$$E(\varepsilon_i) = 0$$

$$\hat{y}_i = b_0 + b_1 x_i =$$

$y_i = a + b x_i + e_i = 0$

$\hat{y}_i = a + b x_i$



□ **Regression to the mean:** let, $b = 2.32$. which means if we increase 1 unit of x, on average y will be increase by 2.32 units.

prediction
- Interpolation : $x=45$ হলে $y=?$
- Extrapolation : $x=85$ হলে $y=?$

forecasting

$x \rightarrow y$	
40	35
50	42
80	55

45, 85

Normal approximation in regression

Residuals:

$$\ln(y_i) = a + b(x_i) + e_i = 0$$

lin, log
log, log
log, log

(e based)

- First difference
- Leaverage and influential points.

c.w. 08/01/2024

Paper Source
Subject............
Date :............ Time:............

Hayter chap 12, 13

Multiple Regression

Y intercept — Population Slopes. — Random Error

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_K X_K + \varepsilon$$

R^2 = regn coffecient / coeff. determination

$-1 \le r \le 1$

$\downarrow$

0

$(-1)^n = 1$

$(1)^2 = 1$

$0^2 = 0$

$R^2 = 0.64$

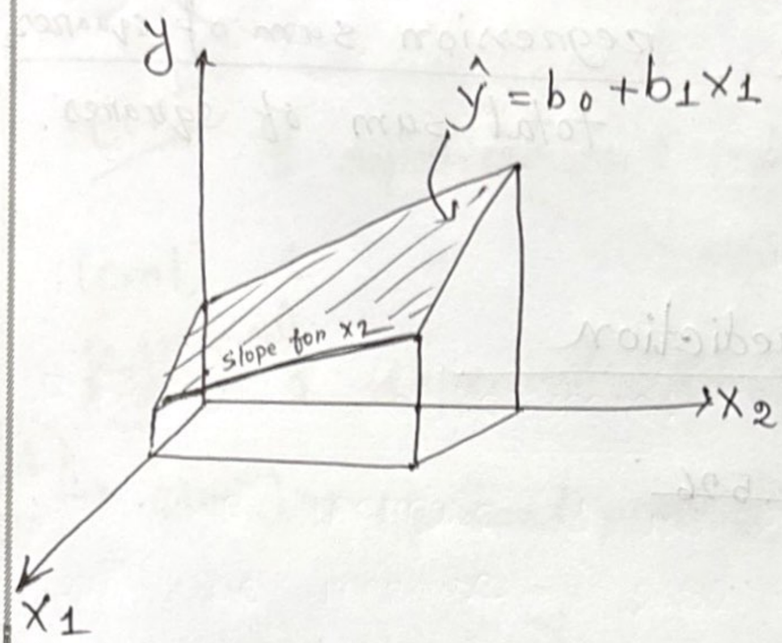
64% ~~regn~~ variation of y is explained by x

x always increas

Paper Source
Subject.
Date : Time:

Two variable model

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2$$



$\therefore$ Sales $= 306.526 - 24.975$ (Price) $+ 74.131$ (Advertis

$b_1 = -24.975$

b_1 If we increase 1 dollar price of pie, on average pie sells will be decrease by about 25unit keeping other things constant

$b_2 = 74.131$

b_2 if we increase 100 dollar of advertising cost, on average pie sells will be increase by 74 pieces. keeping other things constand.

- Setarous Peribous.

Paper Source
Subject..........
Date :.......... Time:..........

Coefficient of Determination, R^2

$$R^2 = \frac{SSR}{SST} = \frac{\text{regression sum of squares}}{\text{total sum of squares}}.$$

Prediction

~~Sales = 306.526~~

Dummy Variables

Categorical Data

- There will be on average 15 pieces higher average pie sales in holidays compare to non ~~holdi~~ holiday keeping all the other things constant.

Or,

On average in holidays, there will be higher
avg sales of pie by 15 pieces compare to non-holidays keeping other things constant.

C.W 13/01/24

Paper Source
Subject.
Date : Time:

Logistic Regression

Cont

1 dependent 1 indep $\longrightarrow$ Single Regression

(cont) 1 " 1+ " $\longrightarrow$ Multiple "

(categorical) ~~(Cont)~~ 1 dependent 1/1+" $\longrightarrow$ logistic "

(0-1) → Binary → We want prove 1

$$Pr(Y=1|X) = \pi(x)$$

(0 to1) Continuous given

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

$$\hat{y} = b_0 + b_1 x$$

$$\pi(x) \rightarrow \beta_0 + \beta_1(x) \longrightarrow -\infty \text{ to } \infty$$

$$Pr(y=1|X) = \beta_0 + \beta_1 X$$

$$\downarrow$$

$$\log \frac{\pi(x)}{1-\pi(x)}$$

$$\therefore \pi(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$

20/01/2025

Paper Source
Subject.........
Date :.........Time:.........

Ordinary list square (OLS)

$$y' = B_0 + B_1 X_i + \varepsilon_i$$

$$\hat{y} = b_0 + b_1 x_i ;$$

$$y = \hat{y} = \varepsilon_i$$

$$E(\varepsilon_i) = 0.$$

$$\cancel{(y' - \hat{y}) = y - b_0 - b_1 x_i = \varepsilon_i}$$

$$(y' - \hat{y})^2 = \varepsilon_i^2 = (y - b_0 - b_1 x_i)^2$$

/01/ 2025

Hayter- Chap 9, 10

Paper Source
Subject.................
Date :.................Time:.................

Logistic Regression

$y = 0, 1$
↓ ↓
CHD- CHD+

$x = age$

OR's Ratio

OR = 1.23

OR = −0.85

$= 1 - 0.85 = 0.15 = 15\%$

age 1 বছর বাড়লে Heart deaseases হওয়ার প্রবণতা 1.23 time বেড়ে যাবে

If we increase 1 unit of x the probability of CHD is 1.23 times higher than having non. CHD keeping other things constant.

(−) হয়েছে কারণ

* If we increase 1 unit of x age the probability of CHD is 15% lower than that of negative ~~keep~~ CHD keeping other things constant.

~~ORs Ratio যদি কখনো~~

~~if we increas~~

Machine learning (ML)

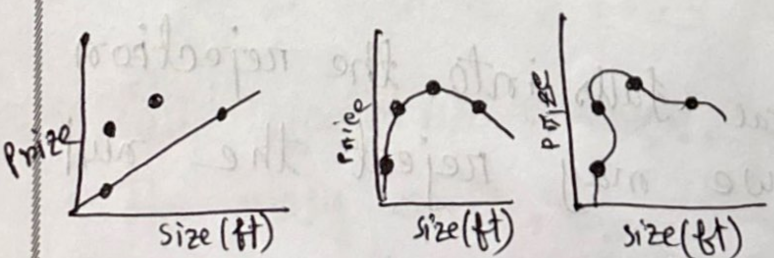
Theoretical ~~প্রশ্ন আসবে~~

linear eqn
$y = \beta_0 + \beta_1 x_i^2 + \varepsilon_i$

non linear eqn
$y = \beta_0 + \beta_1 x_i^2 + \varepsilon_i$

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

$y - \hat{y} = \text{bias}$



Under fit (bias) exact fit over fit

Paper Source
Subject
Date : Time:

Bias ঠিক করার জন্য data বাড়ালে হবে না,
data তখনই বেশি হলে ভালো যখন variance বেশি থাকে।

Variance কমানোর জন্য data add করতে হবে।
Bias কমানোর জন্য model verification.

* Chart

In the chart, linear Regression Algorithm is better because low variance.

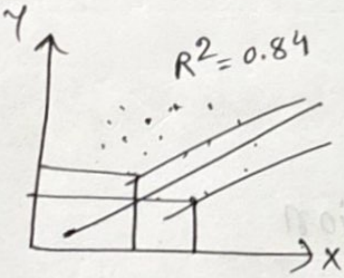
C.W 22/01/2025

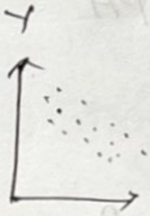
Paper Source
Subject.........
Date :.............Time:.........

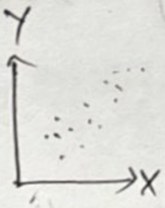
Simple Linear Regression

x এর কতটুকু change এর জন্য y change হচ্ছে

R^2 line কে ফিট করে কিনা সেটা check করে







4 Test
- Parameter test 1-11
- Hypothesis test
- Anova

Simple Linear Regression
Logistic Reg

Decision Tree
Random forest

$$\hat{y}_i = b_0 + b_1 x_i$$

estimate y

x এর এক unit increase এর জন্য y এর কত unit average change (increase/decrese) হচ্ছে তাকে b বলে।

Multiple Regression Equation

$$b_1 = r \frac{S_y}{S_x} \qquad r = \frac{(x_i - \bar{x})(y_i - \bar{y})}{n-1\,(S_x S_y)}$$

$$a = b_0 = \bar{y} - b_1 \bar{x}$$

Paper Source
Subject.........
Date :......... Time:.........

20

Library work

L.W:	5	6	4	~~5~~	9	15
CGPA	3.1	3.5	2.7	~~6~~	3.9	3.7

→ X

दुईटोर Cause → x → L.W
continuous Effect → y → CGPA
हुन्छन्

Logistic Regression

$$P(y|x) = \frac{e^{a+fx}}{1+e^{a+fx}}$$

linear $y \to 0 \to \infty$
Logiste $y \to 0-1$

-0.77

$1 - 0.77 = 0.23$

~~X~~ if we increase 1 hour/unit of x, the chance / probability of having over height from the normal weight will be decrease by 23%.
lower

60.5.

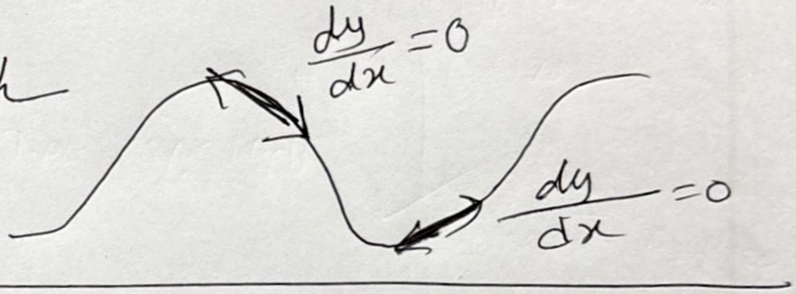
Paper Source
Subject...
Date :..Time:..........................

OLS এর প্রাথমিক ধারণা

OLS

$$\{y_i - (a+bx)\}^2 = \sum e_i^2 \longrightarrow$$

OLS এর math লাগবে না,



Multiple regn. → only ব্যবহার