BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI
WORK INTEGRATED LEARNING PROGRAMMES
M .Tech (Data Science & Engineering)
I Semester, 2019-20
Course Handout

Course Title	Introduction to Statistical Methods
Course No(s)	

Course Description

This course will cover the statistical techniques which are very important in Data Science. It covers the models related to descriptive statistics, inferential statistics, predictive analytics and applied multivariate analytics.

Course Objectives

CO1	Understanding the data representation and analysis which is very important in Data Science
CO2	Understanding the predictive & inferential statistical models used in Data Science

Text Books

No	Author(s), Title, Edition, Publishing House
T1	Probability and Statistics for Engineering and Sciences,8th Edition, Jay L Devore, Cengage Learning
T2	Applied Logistic Regression, Hosmer and Lemeshow,3rd Edition, Wiley
T3	Introduction to Time Series and Forecasting, Second Edition, Peter J Brockwell, Richard A Davis, Springer.

Reference Books

No	Author(s), Title, Edition, Publishing House
R1	Miller and Freund's Probability and statistics for Engineers, 8th Edition, PHI
R2	Statistics for Business and Economics by Anderson, Sweeney and Wiliams, CENAGE learning

BITS Pilani
Pilani | Dubai | Goa | Hyderabad

Modular Content Structure

1. Descriptive Statistics
    - 1.1. Data Visualisation
    - 1.2. Measures of Central Tendency
    - 1.3. Measures of Variability
2. Probability
    - 2.1 Probability – Introduction and Basics
    - 2.2 Conditional probability
    - 2.3 Bayes' theorem
3. Probability Distributions
    - 3.1. Random variables – Discrete & Continuous
    - 3.2. Probability Distributions
        - 3.2.1. Binomial Distribution
        - 3.2.2. Poisson Distribution
        - 3.2.3. Normal Distribution
4. Testing of Hypothesis
    - 4.1. Sampling & Estimation
    - 4.2. Type I, Type II errors
    - 4.3. Testing of Hypothesis – Mean – one and two mean
    - 4.4. Testing of hypothesis – Proportions – one and several Proportions
    - 4.5. ANOVA
5. Regression
    - 5.1. Covariance
    - 5.2. Correlation
    - 5.3. Sum of Least Squares
    - 5.4. Simple linear regression
    - 5.5. Ridge Models &Lasso Model
    - 5.6. Assumptions of linear regression
    - 5.7. Model validation
    - 5.8. Multiple linear regression
    - 5.9. Nonlinear regression
    - 5.10. Logistic regression
6. Forecasting Model
    - 6.1. Principles of Forecasting
    - 6.2. Time series Analysis
        - 6.2.1. Smoothing & decomposition methods
        - 6.2.2. ARIMA Model

6.2.3 Moving Averages

6.2.4 Exponential smoothing

7. Applied Multivariate Analytics

6.1 Introduction

6.2 Joint distributions – Discrete & Continuous

6.3 Multivariate Normal Distribution

6.4 Principal Component Analysis

Learning Outcomes:

No	Learning Outcomes
LO1	Clear understanding of the various statistical models to model the data
LO2	Drawing conclusions from the models selected to understand the data

Part B: Course Handout

Academic Term	I semester, 2019 – 20
Course Title	**Introduction to Statistical Methods**
Course No	

Course Contents

Contact Session 1: Module 1(Descriptive Statistics)

Contact Session	List of Topic Title	Reference
CS - 1	Descriptive Statistics: Data Visualisation, Measures of Central Tendency, Measures of Variability	T1:Chapter 1
HW	Problems on Descriptive Statistics	T1:Chapter 1
Lab		

Contact Session 2: Module 2 - Probability

Contact Session	List of Topic Title	Reference
CS - 2	Probability - Introduction and Basics, Conditional probability, Bayes' theorem	T1:Chapter 2
HW	Problems on probability	T1:Chapter 2
Lab		

Contact Session 3: Module 3 – Probability Distributions

Contact Session	List of Topic Title	Reference
CS - 3	Random Variables – Discrete & Continuous	T1:Chapter 3 & 4
HW	Problems on Random Variables	T1:Chapter 3 & 4
Lab		

Contact Session 4: Module 3 – Probability Distributions

Contact Session	List of Topic Title	Reference
CS - 4	Probability Distributions – Binomial, Poisson and Normal Distributions	T1:Chapter 3 & 4
HW	Problems on probability distributions	T1:Chapter 3 & 4
Lab		

Contact Session 5: Module 4 – Testing of Hypothesis

Contact Session	List of Topic Title	Reference
CS - 5	Sampling & Estimation	R1
HW	Problems on Interval Estimation	R1
Lab		

Contact Session 6: Module 4 – Testing of Hypothesis

Contact Session	List of Topic Title	Reference

CS - 6	Testing of Hypothesis - Type I & II errors, Mean and Proportions models (one mean, Two mean, One proportions and Several proportions with small and big samples wherever applicable)	T1:Chapter 7 ,8,9 & 10
HW	Problems on Testing of Hypothesis	T1:Chapters 7 to 10
Lab		

Contact Session 7: Module 4 – Testing of Hypothesis

Contact Session	List of Topic Title	Reference
CS - 7	Testing of Hypothesis - Problems discussion	T1:Chapter 7 ,8,9 & 10
HW	Problems on Testing of Hypothesis	T1:Chapter 7 ,8,9 & 10
Lab		

Contact Session 8:

Contact Session	List of Topic Title	Reference
CS - 8	REVISION OF THE TOPICS COVERED	
HW		
Lab		

MID SEMESTER EXAMINATION

Contact Session 9: Module 5 – Regression

Contact Session	List of Topic Title	Reference
CS - 9	Covariance , correlation, Sum of least squares	T1:Chapter 12 & 13
HW	Problems on correlation and co variance	T1:Chapter 12 & 13
Lab		

Contact Session 10: Module 5 – Regression

Contact Session	List of Topic Title	Reference
CS - 10	Simple Linear regression model, Assumption of the model, interpretation of the model	T1:Chapter 12 & 13
HW	Problems on Linear regression	T1:Chapter 12 & 13
Lab		

Contact Session 11: Module 5 – Regression

Contact Session	List of Topic Title	Reference
CS - 11	Multiple linear regression model, non – linear regression & Logistic regression	T1:Chapter 12 & 13 and T2
HW	Problems on Linear regression	T1:Chapter 12 & 13
Lab		

Contact Session 12: Module 6 – Forecasting Models

Contact Session	List of Topic Title	Reference
CS - 12	Principles of Forecasting, Time series models _ smoothing and decomposition methods, AR,MA,ARIMA Models	T3
HW	Problems Time series models	
Lab		

Contact Session 13: Module 6 – Forecasting Models

Contact Session	List of Topic Title	Reference
CS - 13	Moving Averages and Exponential smoothing models	T3
HW	Problems Time series models	
Lab		

Contact Session 14: Module 7 – Applied Multivariate Analytics

Contact Session	List of Topic Title	Reference

CS - 14	Introduction – Joint Distributions	T1:Chapter 5
HW	Problems on Joint Distributions	
Lab		

Contact Session 15: Module 7 – Applied Multivariate Analytics

Contact Session	List of Topic Title	Reference
CS - 15	Principal component Analysis , Multivariate Normal Distribution	
HW	Problems on PCA	
Lab		

Contact Session 16:

Contact Session	List of Topic Title	Reference
CS - 16	**REVISION OF THE SYLLABUS**	
HW		
Lab		

L- 7: Inferential statistics & Predictive Analytics

Agenda

- Central limit theorem
- Type I, Type II Errors
- Testing of Hypothesis – continuation from previous session
- Covariance
- Correlation
- Introduction to regression

Central Limit Theorem

If $\bar{X}$ is the mean of a sample of size n taken from a population having the mean μ and variance σ^2, then $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ is a random variable whose distribution function approaches that of the standard normal distribution as $n \to \infty$.

Central Limit theorem

→ It does not matter what the distribution of X_i's is

→ In many real applications, the random variable is a sum of independent random variables. In all such cases, CLT helps to use normal distribution.

Examples

→ random noise in Comm. Systems

→ Errors in Lab measurements

→ Errors in regression analysis etc

Errors

	H_0 is true	H_0 is False
Decision	Accept H_0	reject H_0

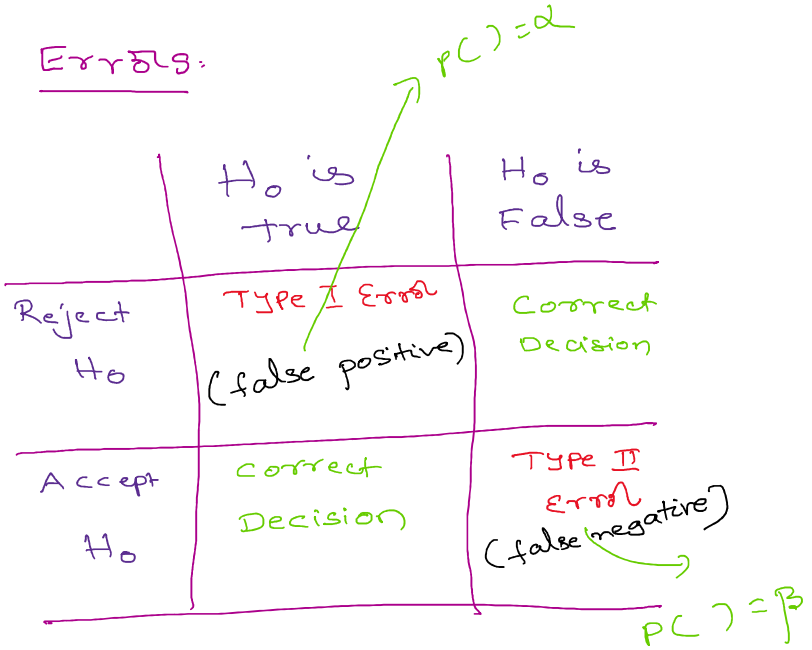
(1) reject H_0, when it is false ✓

(2) reject H_0, when it is TRUE

(3) Accept H_0, when it is false

(4) Accept H_0, when it is TRUE ✓

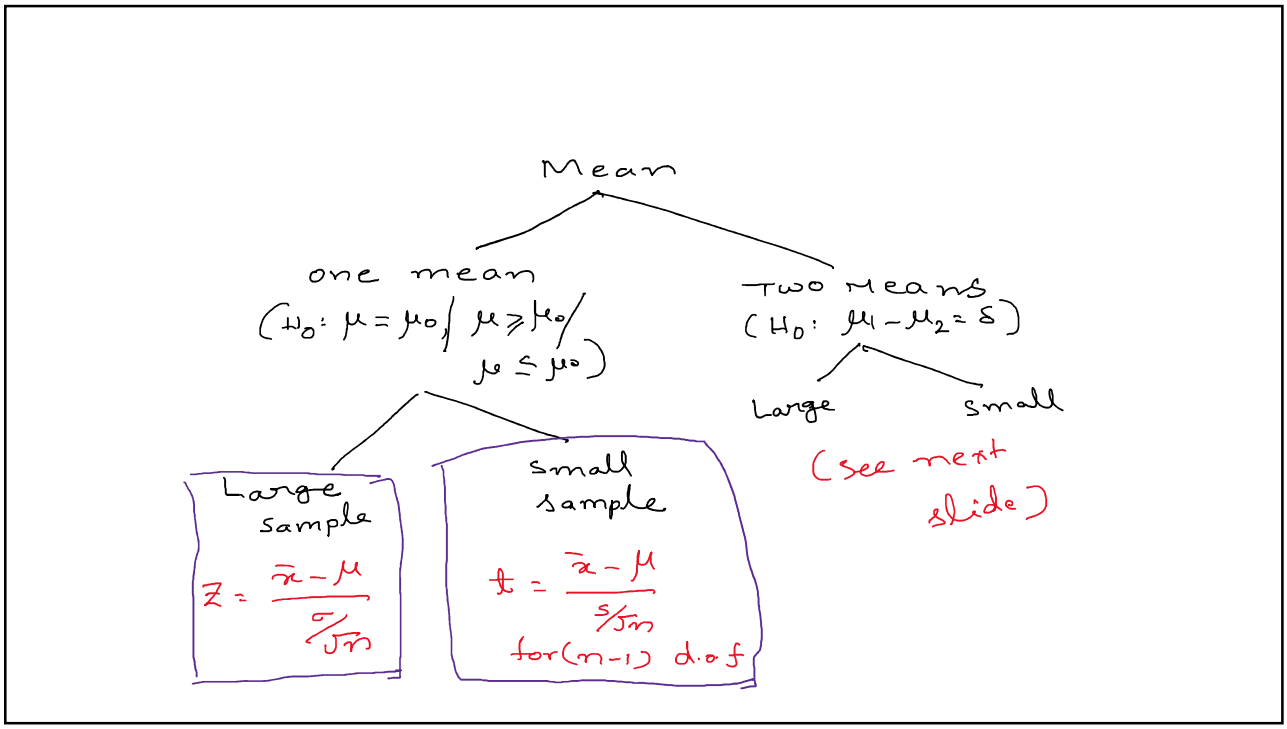
Errors:

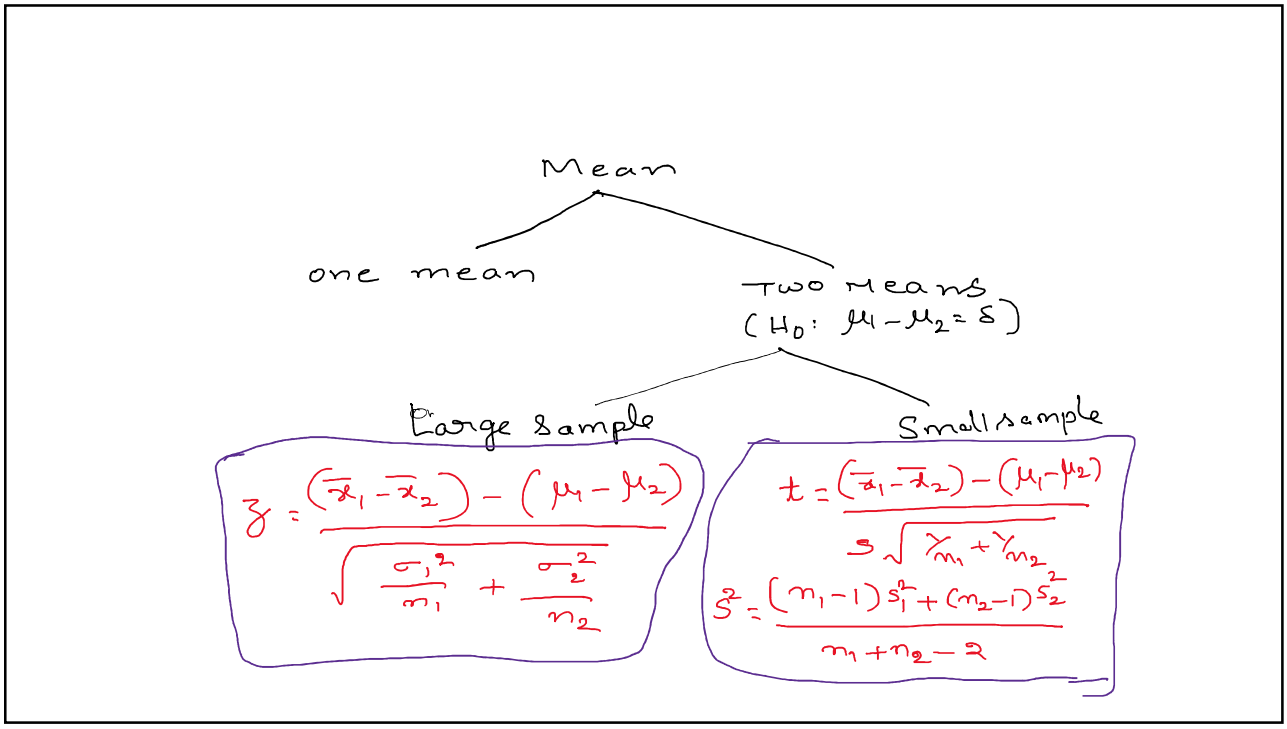


Testing of Hypothesis

or

Hypothesis testing


Mean
one mean
($H_0: \mu = \mu_0 / \mu \geq \mu_0 / \mu \leq \mu_0$)
Two means
($H_0: \mu_1 - \mu_2 = \delta$)
Large
Small
(see next slide)
Large sample
$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$
Small sample
$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$
for $(n-1)$ d.o.f


Mean
one mean
Two means
($H_0: \mu_1 - \mu_2 = \delta$)
Large sample
Small sample
$Z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$
$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$
$s^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2}$

Testing of Hypothesis

Example - 1

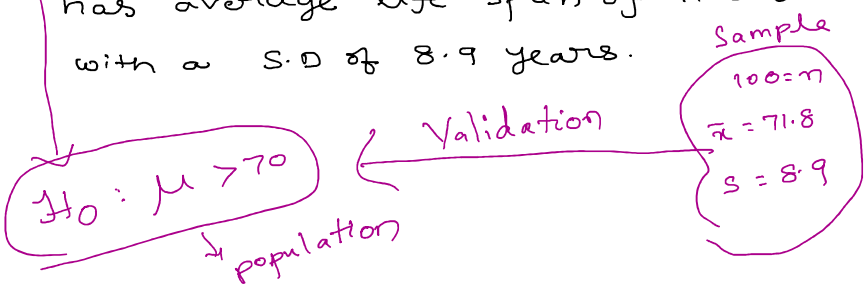
Example:- (1)

Can it be concluded that the average life span of Indians is more than 70yrs. If a random sample of 100 indians has average life span of 71.8 years with a S.D of 8.9 years.

Example:- (contd)

Can it be concluded that the average life span of Indians is more than 70yrs.

If a random sample of 100 indians has average life span of 71.8 years with a S.D of 8.9 years.



Example:- (contd)

Can it be concluded that the average life span of Indians is more than 70yrs.

If a random sample of 100 indians has average life span of 71.8 years with a S.D of 8.9 years.

→ one mean problem

→ $n = 100$: Large sample, so Z-test

∴ $Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$ or $\frac{\bar{x} - \mu}{s/\sqrt{n}}$

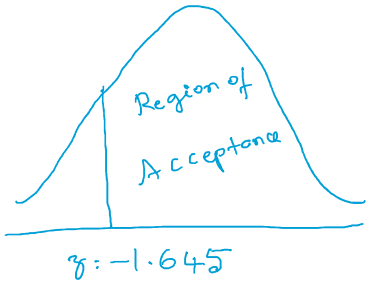
$H_0: \mu > 70$

$H_1: \mu \leq 70$ → Left tailed test

$\alpha = 5\%$ (Let)

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{71.8 - 70}{8.9/\sqrt{100}}$$

$$= 2.022$$



Lies in the region of acceptance

∴ H_0 is accepted

i.e Avg life is more than 70 years

Testing of Hypothesis

Example – 2.

Example - 2

A machine which produces Mica insulating washers for use in electronic devices said to have a thickness of 10mm. A sample of 10 washers has an average thickness of 9.52 mm with a S.D of 0.6mm. Whether the sample is drawn from the given population (use 5% Level of significance)

Example - 2

Small sample

Mean

A machine which produces Mica insulating washers for use in electronic devices said to have a thickness of 10mm. A sample of 10 washers has an average thickness of 9.52 mm with a S.D of 0.6mm. Whether the sample is drawn from the given population (use 5% Level of significance)

$\bar{x}$ S

$H_0 : \mu = 10$

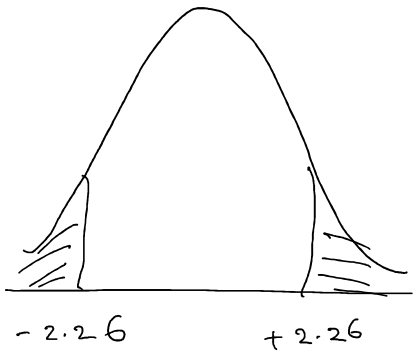
$H_1 : \mu \neq 10$

$\alpha = 0.05$

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

$$= \frac{9.52 - 10}{0.6/\sqrt{10}}$$

$$= -2.52 \longrightarrow \text{Reject } H_0$$



Testing of Hypothesis

Example – 3

Example 3 — modified problem

A random sample of 40 items produced by a company A have a mean life time of 647 hours with S.D 27 hours. While a sample of 40 items by company B has a mean life time of 638 hours with S.D of 31 hours.

Does this substantiate the claim of the company A that their items are ~~superior to those~~ produced by company B in

same as those

terms of mean life

Example 3 — Solution.

A random sample of 40 items produced by a company A have a mean life time of 647 hours with S.D 27 hours. While a sample of 40 items by company B has a mean life time of 638 hours with S.D of 31 hours.

Does this substantiate the claim of the company A that their items are ~~superior to those~~ produced by company B in

same as those

terms of mean life

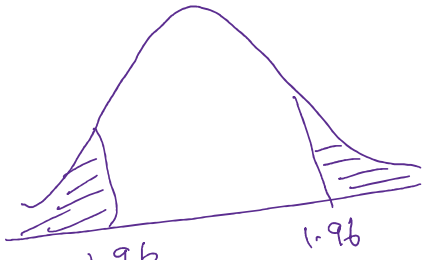
i.e. $\mu_1 = \mu_2$

Solution:-

$H_0: \mu_1 - \mu_2 = 0$

$H_1: \mu_1 - \mu_2 \neq 0$

$\alpha = 0.05$



$$Z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{(647 - 638) - 0}{\sqrt{\frac{(27)^2}{40} + \frac{(31)^2}{40}}}$$

$= 3.73$ → Reject H_0

Example 3

Earlier problem on it is

A random sample of 40 items produced by a company A have a mean life time of 647 hours with S.D 27 hours. While a sample of 40 items by company B has a mean life time of 638 hours with S.D of 31 hours.

Does this substantiate the claim of the company A that their items are superior to those produced by company B.

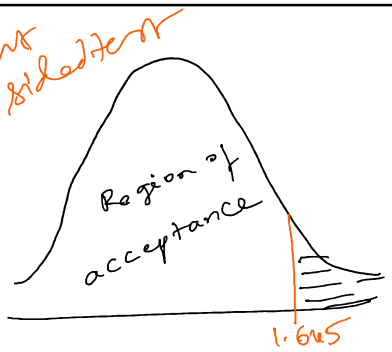
i.e. $\mu_1 > \mu_2$

Solution:-

$H_0: \mu_1 - \mu_2 = 0$

$H_1: \mu_1 - \mu_2 > 0$ → right sided test

$\alpha = 0.05$



$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{(647 - 638) - 0}{\sqrt{\frac{(27)^2}{40} + \frac{(31)^2}{40}}}$$

$= 3.73$ → Reject H_0

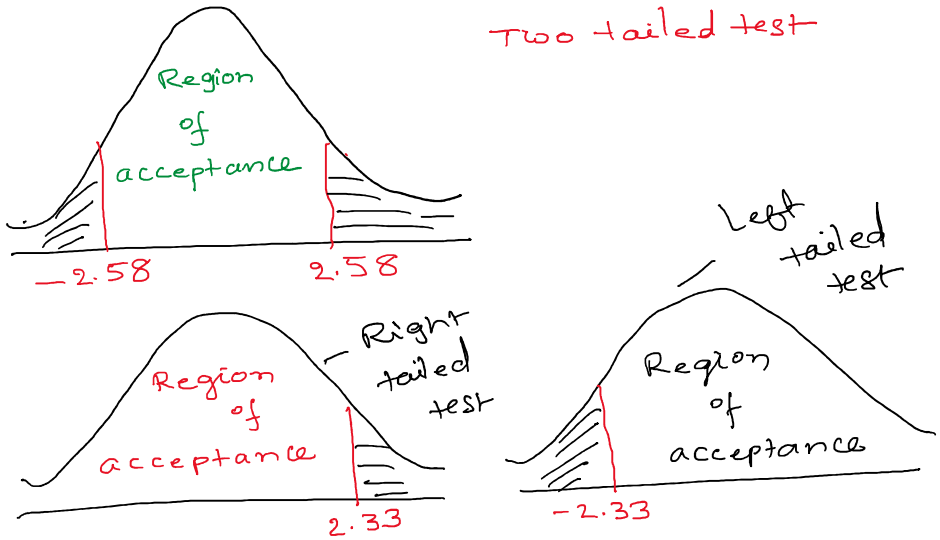
i.e. Accept alternate Hypothesis

Critical region for some 'α's (Z-distribution)

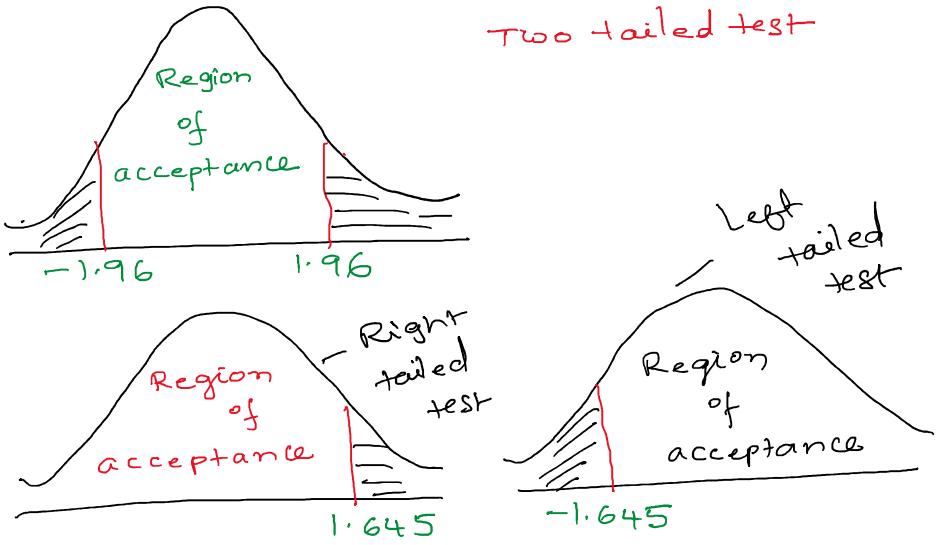
useful table values (Z- distribution)

	Level of significance		
	0.01	0.05	0.1
Two - tailed test	± 2.58	± 1.96	± 1.645
Right tailed test	2.33	1.645	1.28
Left tailed test	-2.33	-1.645	-1.28

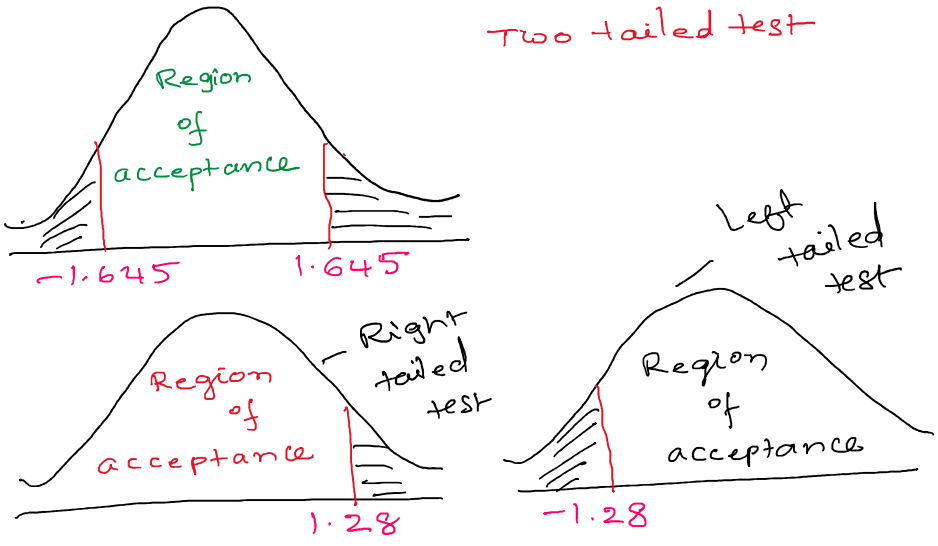
$\alpha = 1\%$ (or 0.01)



$\alpha = 5\%$ (or 0.05)



$\alpha = 10\%$ (or 0.1)



Note :-

In case of 't' distribution, the critical region depends on the degrees of freedom ($\nu = n-1$, where n is the size of the sample)

Testing of Hypothesis

Example – 4

Example- 4 :-

A Company believes that the advertisement A is more effective than advt. B. To test this sampling is done.

In a random sample of 60 customers who saw advertisement A, 18 tried the product. In a random sample of 100 customers, who saw advt B, 22 tried the product.

Does this indicate that advt A is more effective than advt B.

Example- 4 :-

A Company believes that the advertisement A is more effective than advt. B. To test this sampling is done.

In a random sample of 60 customers who saw advertisement A, 18 tried the product. In a random sample of 100 customers, who saw advt B, 22 tried the product.

Does this indicate that advt A is more effective than advt B.

Sample A : 18 out of 60
Sample B : 22 out of 100
} Advt (A) > Advt (B) ???

Sample A : 18 out of 60

Sample B : 22 out of 100

$\bar{P}_1 = \frac{18}{60} = 0.3$, $\bar{P}_2 = \frac{22}{100} = 0.22$ Let α : 5%

$H_0 : P_1 - P_2 = 0$

$H_1 : P_1 > P_2$

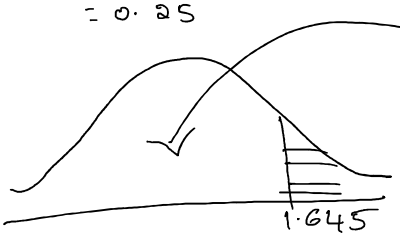
$$\bar{P} = \frac{n_1 \bar{P}_1 + n_2 \bar{P}_2}{n_1 + n_2}$$

$= 0.25$

$$z = \frac{(\bar{P}_1 - \bar{P}_2) - \delta}{\sqrt{\bar{P}(1-\bar{P})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$= \frac{(0.3 - 0.22) - 0}{\sqrt{(0.25)(0.75)\left(\frac{1}{60} + \frac{1}{100}\right)}}$$

$= 1.131$



Accept H_0

∴ no significant diff in the effectiveness of advertisements

Testing of Hypothesis

Example – 5

Example:

consider the following data

Travel time	Stress: High	Moderate	Low	Total
< 20 min	9	5	18	32
20-50 min	17	8	28	53
> 50 min	18	6	7	31
Total	44	19	53	116

Based on this data, Can we conclude that stress levels depends on travel time

???

$\frac{x}{n} = \bar{p} \longrightarrow p_0$

One proportion: (Large sample)

$H_0: p = p_0 \mid p \geq p_0 \mid p \leq p_0$

$$Z = \frac{\bar{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

Two proportions: (large sample)

$H_0: p_1 - p_2 = \delta \mid p_1 - p_2 \geq \delta \mid p_1 - p_2 \leq \delta$

$$Z = \frac{(\bar{p}_1 - \bar{p}_2) - \delta}{\sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \quad \text{where } \bar{p} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$$

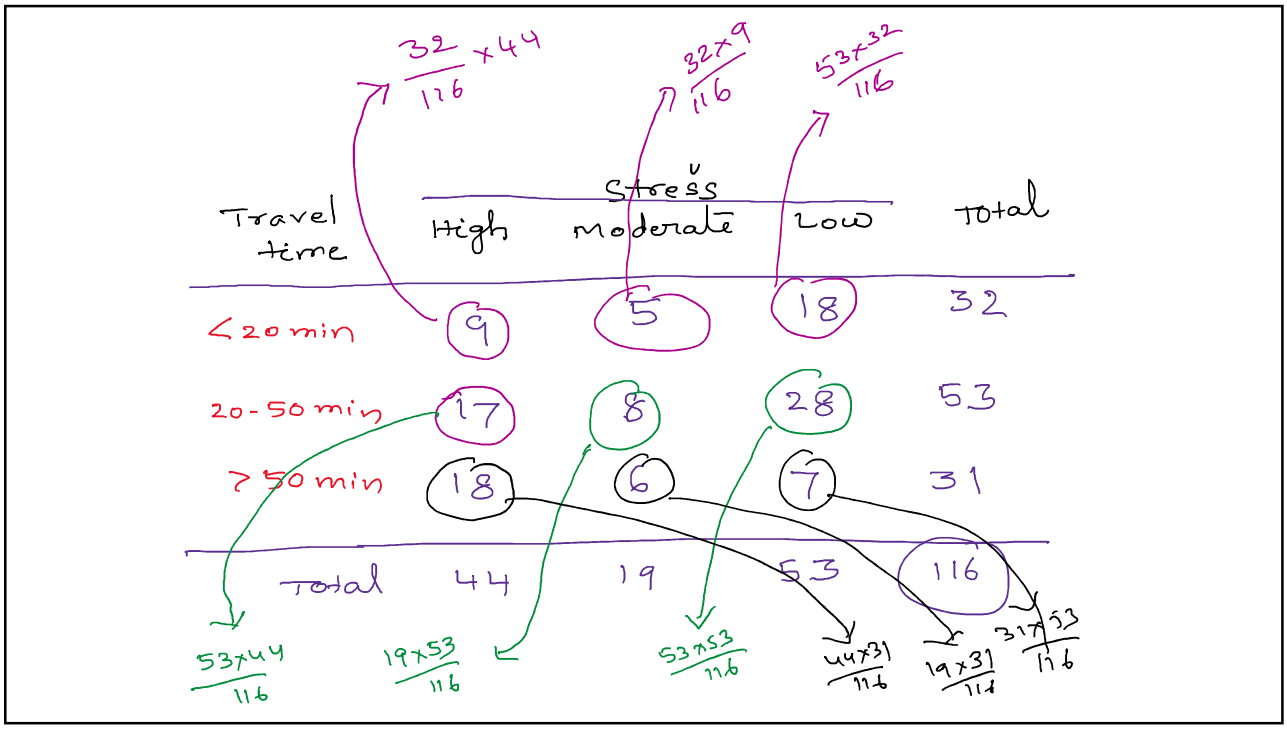
Chi-Square (χ^2) distribution

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

O: observed frequencies

E: Expected frequencies

for $(r-1) * (c-1)$ degrees of freedom



$$\chi^2 = \sum \frac{(O-E)^2}{E} = \sum \frac{(9-12.14)^2}{12.14} + \cdots$$

$$= 9.836$$

Travel time	Stress: High	moderate	Low	Total
< 20 min	9 / 12.14	5 / 5.24	18 / 14.62	32
20-50 min	17 / 20.10	8 / 8.68	28 / 24.22	53
> 50 min	18 / 11.75	6 / 5.08	7 / 14.17	31
Total	44	19	53	116

Calculated $\chi^2 = 9.836$

Let $\alpha = 0.01$

d.o.f : $(r-1) \times (c-1)$

$= (3-1) \times (3-1) = 4$

$\chi^2_{0.01,4} = 13.30$

$\chi^2_{cal} = 9.836 < 13.30$

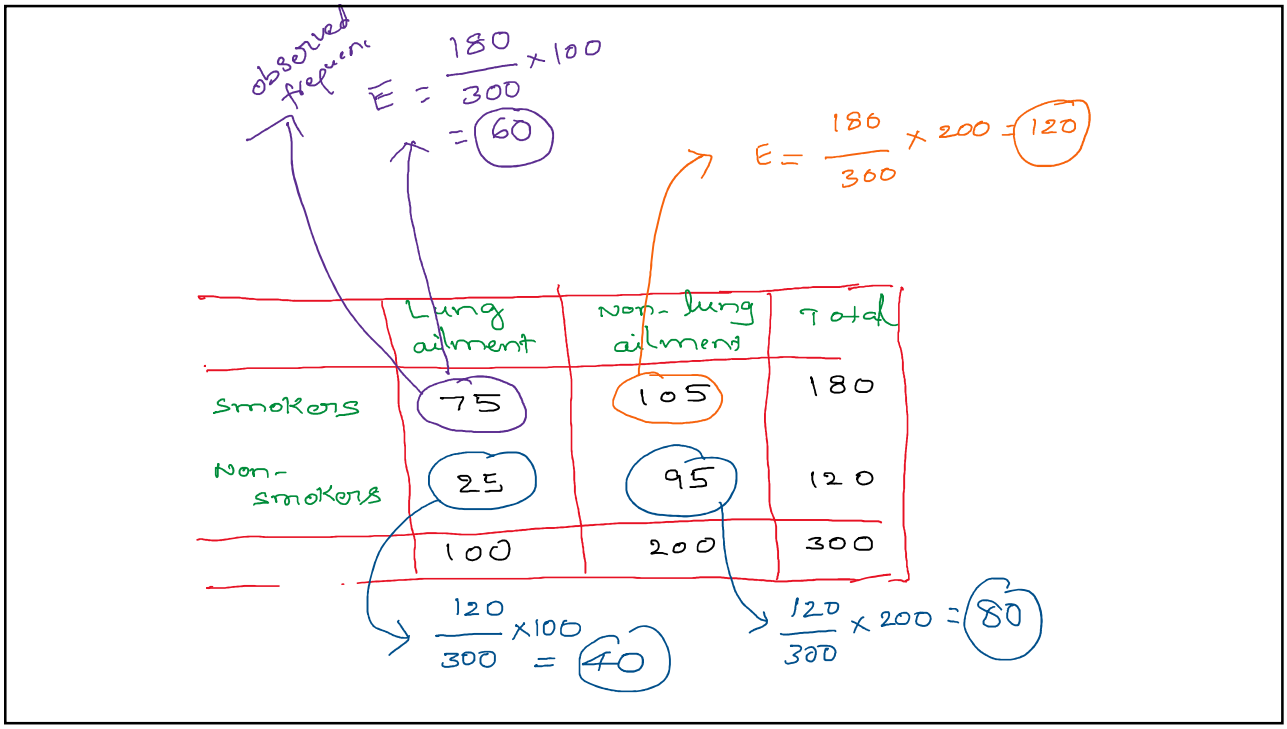
H_0 accepted

Example

A tobacco company claims that there is no relationship between smoking and lung ailments. ↓ H_0

	Lung ailment	Non-lung ailment	Total
Smokers	75	105	180
Non-smokers	25	95	120
	100	200	300

Based on this data, can we accept / reject the claim



$$\chi^2 = \frac{(75-60)^2}{60} + \frac{(105-120)^2}{120} + \cdots$$

$$= 14.063$$

From χ^2-tables at 0.05 LOS

d.o.f $= (r-1) \times (c-1)$
$= (2-1) \times (2-1)$
$= 1$

$= 3.841$

	Lung ailment	Non-lung ailments	Total
Smokers	75 ~~60~~	105 ~~120~~	180
Non-smokers	25 ~~40~~	95 ~~80~~	120
	100	200	300

$\chi^2 = 14.063 > 3.841$

Reject H_0 $\rightarrow$

L- 8: Predictive Analytics

Agenda

- Covariance
- Correlation
- Introduction to regression
- Method of least squares
- Simple linear regression

Covariance of X and Y

$$Cov(X,Y) = E\left[(X-\mu_X)(Y-\mu_Y)\right]$$

$$= \sum\sum (x-\mu_X)(y-\mu_Y)\, P(x,y) \quad \text{if discrete}$$

$$= \int\int (x-\mu_X)(y-\mu_Y)\, f(x,y)\, dx\, dy \quad \text{if continuous}$$

consider the following

⇒ whether spending on advertising of a company is related to overall sales of the company.

→ If it is related, how it is related

⇒ Forecasting the sales, given the budget for advertising

And also

⇒ Farmer has an impression that if he uses more fertilizers, then the crop yield increases.

We need to validate this?

How → ?

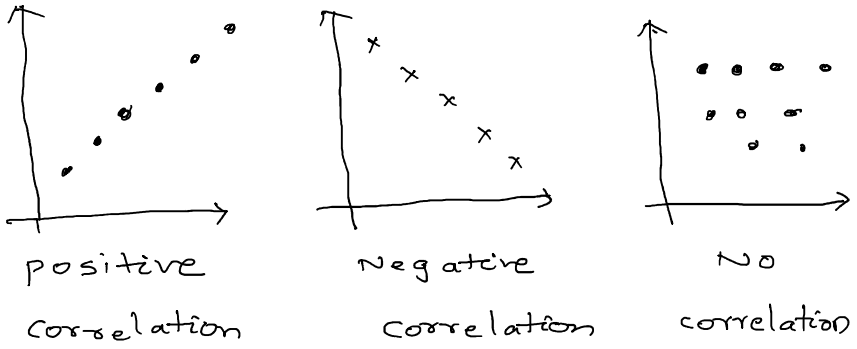
Correlation

→ Sales of a company and Expenditure on advertisement

→ Price and Demand of a product

→ Inflation and Gold Price

→ IQ and performance in Entrance.


positive
correlation
Negative
correlation
No
correlation


Perfo-
rmance
Stress
results
efforts

Coefficient of correlation:

$$r = \frac{cov(x,y)}{\sigma_x \sigma_y} = \frac{\Sigma XY}{\sqrt{\Sigma X^2 \cdot \Sigma Y^2}}$$

where $X = x - \bar{x}$

$Y = y - \bar{y}$

$X^2 = (x - \bar{x})^2$

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

Coefficient of Correlation

$r = 1 \Rightarrow$ Perfect and positive relation

$r = -1 \Rightarrow$ " " negative relation

$r = 0 \Rightarrow$ No relation

$0 < r < 1 \Rightarrow$ Partial positive relation

$-1 < r < 0 \Rightarrow$ " negative "

Example - 1

x	1	2	3	4	5	6	7	8	9
y	10	11	12	14	13	15	16	17	18

$$\bar{x} = \frac{\Sigma x}{n} = \frac{45}{9} = 5$$

$$\bar{y} = \frac{\Sigma y}{n} = \frac{126}{9} = 14$$

x	$X = x-5$	X^2	y	$Y = y-14$	Y^2	XY
1	-4	16	10	-4	16	16
2	-3	9	11	-3	9	9
3	-2	4	12	-2	4	4
4	-1	1	14	0	0	0
5	0	0	13	-1	1	0
6	1	1	15	1	1	1
7	2	4	16	2	4	4
8	3	9	17	3	9	9
9	4	16	18	4	16	16
		60			60	59

$$r = \frac{\Sigma XY}{\sqrt{\Sigma X^2 \Sigma Y^2}}$$

$$= \frac{59}{\sqrt{60 \times 60}}$$

$$= 0.9833$$

Coefficient of Determination

r is coeff. of correlation

r^2 is coeff of determination

↓

Indicates the extent to which variation in one variable is explained by the variation in the other.

$r = 0.9 \Rightarrow r^2 = 0.81$

i.e. 81% of the variation in y due to variation in 'x'

remaining 19% is due to some other factors.

Regression

Regression :-

x	1	2	3	4	5
y	1	4	9	16	25

when $x = 7$: $y = ?$

x	1	2	3	4	5
y	1	6	2	5	4

when $x = 7$, $y = ?$

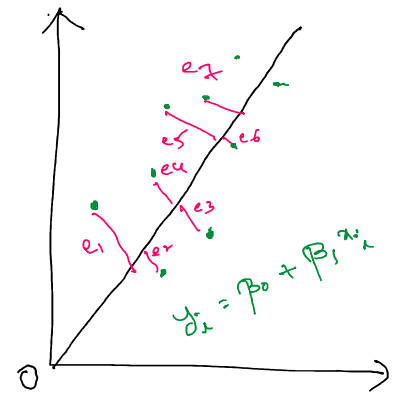
Method of Least squares

y : Dependent Variable

x : Independent Variable

predictor variable

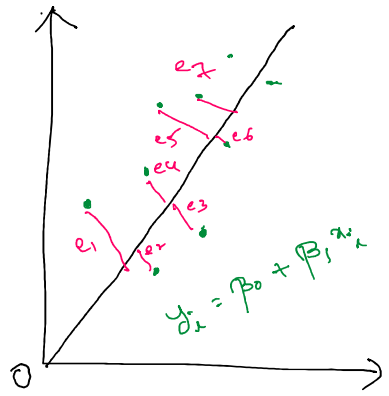
response Variable



Method of Least squares

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

β_0 and β_1 are chosen to minimize the sum of squared errors.



Method of Least squares

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

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

$$\Rightarrow \sum_{i=1}^{n} y_i = n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i \checkmark$$

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

$$\Rightarrow \sum_{i=1}^{n} x_i y_i = \beta_0 \sum x_i + \beta_1 \sum x_i^2 \checkmark$$

on solving these, we get β_0 & β_1 which minimizes error.

Linear regression

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

$$\Sigma y = \beta_0 n + \beta_1 \Sigma x$$

$$\Sigma xy = \beta_0 \Sigma x + \beta_1 \Sigma x^2$$

Normal equations.

Matrix Approach:

Let $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_{p-1} x_{p-1}$

observations $y_i = 1, 2, \ldots n$ → by a vector Y

unknowns $\beta_0, \beta_1 \cdots \beta_{p-1}$ → " " β

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p-1} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p-1} \\ \vdots & & & & \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np-1} \end{bmatrix}$$

$$\underset{n\times 1}{\hat{Y}} = \underset{n\times p}{X}\ \underset{p\times 1}{\beta}$$

Find β to minimize

$$S(\beta) = \sum_{i=1}^{m} \left(y_i - \beta_0 - \beta_1 x_1 - \beta_2 x_2 - \cdots \right)^2$$

$$= \| Y - X\beta \|^2 = \| Y - \hat{Y} \|^2$$

Diff S wrt to each β we get Linear eqns

$$X^T X \hat{\beta} = X^T Y \longrightarrow \text{normal eqns}$$

If $X^T X$ is non-singular, the soln is

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

Computationally, it is sometimes unwise even to form the normal equations because the multiplications involved in forming $X^T X$ can introduce undesirable round-off errors.

Linear regression
(Multiple regression)

Example:-

x_0	Size	No of rooms	No of floors	Age of home	Price Lakhs
1	2000	5	2	45	4000
1	1400	3	1	40	2000
1	1600	3	2	30	3000
1	800	2	1	35	2000
	x_1	x_2	x_3	x_4	y

Linear regression
(Multiple regression)

Example:-

$$X = \begin{bmatrix} 1 & 2000 & 5 & 2 & 45 \\ 1 & 1400 & 3 & 1 & 40 \\ 1 & 1600 & 3 & 2 & 30 \\ 1 & 800 & 2 & 1 & 35 \end{bmatrix} \quad y = \begin{bmatrix} 4000 \\ 2000 \\ 3000 \\ 2000 \end{bmatrix}$$

$$\beta = (X^T X)^{-1} X^T y$$

Example:

Consider the following data

x	1	2	4	0
y	0.5	1	2	0

Fit a linear regression line

Estimate y when $x = 5$.

x	y	xy	x^2
1	0.5	0.5	1
2	1	2	4
4	2	8	16
0	0	0	0
$\Sigma = 7$	$\Sigma 3.5$	$\Sigma 10.5$	$\Sigma 21$

$y = \beta_0 + \beta_1 x$

$\Sigma y = n\beta_0 + \beta_1 \Sigma x$

$\Sigma xy = \beta_0 \Sigma x_1 + \beta_1 \Sigma x^2$

$3.5 = 4\beta_0 + \beta_1 (7)$

$10.5 = 7\beta_0 + \beta_1 (21)$

On solving these

$\beta_0 = 0$

$\beta_1 = 0.5$

i.e $y = 0 + (0.5)x$

When $x = 5$, $y = (0.5)5$

$= 0.25$

L- 9: Predictive Analytics & Revision

Agenda

- Review of last session
- Introduction to regression
- Method of least squares
- Simple linear regression

Covariance of X and Y

$$Cov(X,Y) = E[(X-\mu_X)(Y-\mu_Y)]$$

$E(x) = \sum x\,p(x)$ or $\int x f(x)\,dx$

$$= \sum_x \sum_y (x-\mu_x)(y-\mu_y)\,P(x,y)$$

if discrete ($P(x,y)$: joint p.d.f)

$$= \int\int (x-\mu_x)(y-\mu_y)\,f(x,y)\,dx\,dy$$

if continuous ($f(x,y)$: joint prob. density fn)

$$Cov(X,Y) = \frac{\sum_{i=1}^{n} (x-\mu_x)(y-\mu_y)}{n-1}$$

X → income, Y → expenditure

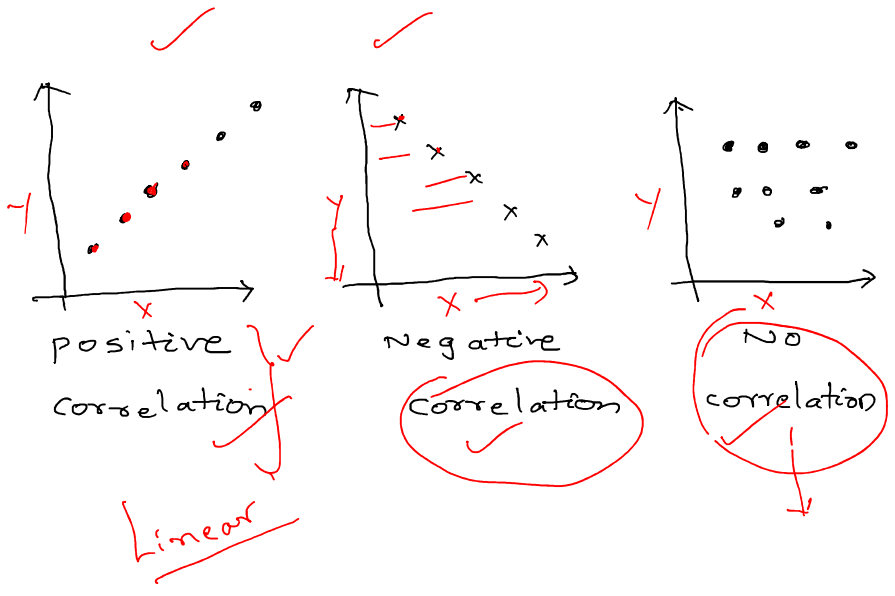
Correlation → $(n-1)$, n

And also

⇒ Farmer has an impression that if he uses more fertilizers, then the crop yield increases.

We need to validate this?

How → ?



non-linear

Coefficient of correlation:

$$r = \frac{cov(x,y)}{\sigma_x \sigma_y} = \frac{\Sigma XY}{\sqrt{\Sigma X^2 \cdot \Sigma Y^2}}$$

where $X = x - \bar{x}$

$Y = y - \bar{y}$

$X^2 = (x - \bar{x})^2$

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

Coefficient of Correlation

$r = 1 \Rightarrow$ Perfect and positive relation

$r = -1 \Rightarrow$ " " negative relation

$r = 0 \Rightarrow$ No relation

$0 < r < 1 \Rightarrow$ Partial positive relation

$-1 < r < 0 \Rightarrow$ " negative "

$-1 \leq r \leq 1$

Example - 1

x	1	2	3	4	5	6	7	8	9
y	10	11	12	14	13	15	16	17	18

$$\bar{x} = \frac{\Sigma x}{n} = \frac{45}{9} = 5$$

$$\bar{y} = \frac{\Sigma y}{n} = \frac{126}{9} = 14$$

x	$X = x-5$	X^2	y	$Y = y-14$	Y^2	XY
1	−4	16	10	−4	16	16
2	−3	9	11	−3	9	9
3	−2	4	12	−2	4	4
4	−1	1	14	0	0	0
5	0	0	13	−1	1	0
6	1	1	15	1	1	1
7	2	4	16	2	4	4
8	3	9	17	3	9	9
9	4	16	18	4	16	16
		60			60	59

$$r = \frac{\Sigma XY}{\sqrt{\Sigma X^2 \Sigma Y^2}}$$

$$= \frac{59}{\sqrt{60 \times 60}}$$

$$= 0.9833$$

$r = 0.9833$

→ Corrected

→ +ve correlated

x	$X = x-5$	X^2	y	$Y = y-14$	Y^2	XY
1	−4	16	10	−4	16	16
2	−3	9	11	−3	9	9
3	−2	4	12	−2	4	4
4	−1	1	14	0	0	0
5	0	0	13	−1	1	0
6	1	1	15	1	1	1
7	2	4	16	2	4	4
8	3	9	17	3	9	9
9	4	16	18	4	16	16
		60			60	59

$$\text{cov}(x,y) = \frac{\Sigma XY}{n-1}$$

$$= \frac{59}{8}$$

$$= 7.375$$

Coefficient of Determination

r is coeff. of correlation

r^2 is coeff of determination

↓

Indicates the extent to which variation in one variable is explained by the variation in the other.

$r = 0.9 \Rightarrow r^2 = 0.81$

i.e. 81% of the variation in y due to variation in 'x'

remaining 19% is due to some other factors.

---

r (x, y)

$r = 0.9833$ → Coeff correlation

$Cov(x, y) = 7.375$ → Covariance Interpretation

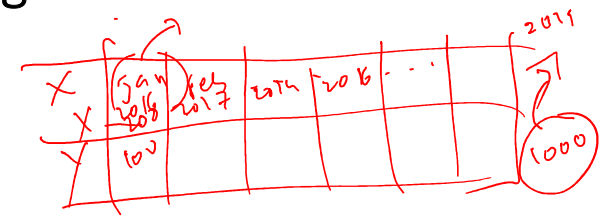
$r^2 = 0.81$ → coeff of determinate

$-1 \le r \le 1$

9490233116

farmer:
x: fertilizers
y: crop yield
Correlation (r)

Regression



Regression :-

Correlation (r)

x	1	2	3	4	5
y	1	4	9	16	25

$y = x^2$ when $x = 7$, $y = 49$

when $x = 7$: $y = ?$

5.4, 5.6, 5.8

x	1	2	3	4	5
y	1	6	2	5	4

$y = 2 \ldots$

when $x = 7$, $y = ?$

Correlation	Regression
→ Measuring strength or degree of the relationship between two variables	→ Having an algebraic equation between two variables
→ No estimation	→ estimation
→ both variables are independent	→ one is dept variable and other indepr variables

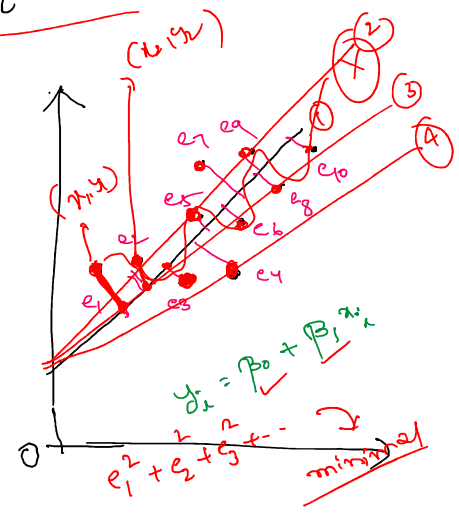
Method of Least squares

y : Dependent Variable

x : Independent Variable

→ predictor variable

→ response Variable

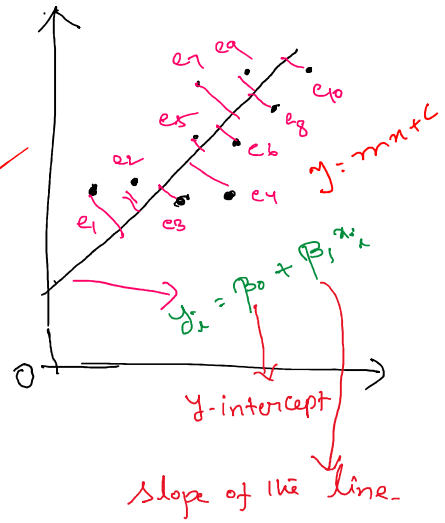


Method of Least squares

"minimizing the error"

i.e minimize

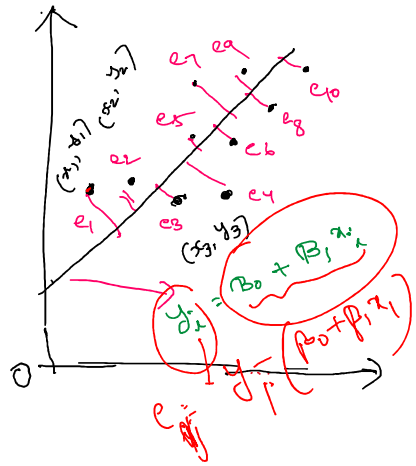
$e_1^2 + e_2^2 + e_3^2 + \cdots e_{10}^2$



Method of Least squares

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

we need to choose β_0 and β_1 which minimizes the error.



Method of Least squares

Max / Min → first derivative = 0

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

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

$$\Rightarrow \sum_{i=1}^{n} y_i = n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i$$

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

$$\Rightarrow \sum_{i=1}^{n} x_i y_i = \beta_0 \sum x_i + \beta_1 \sum x_i^2$$

on solving these, we get β_0 & β_1 which minimizes error.

Linear regression

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

$$\Sigma y = \beta_0 n + \beta_1 \Sigma x$$

$$\Sigma xy = \beta_0 \Sigma x + \beta_1 \Sigma x^2$$

Normal equations. β_0, β_1

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

Regression Coefficients

$y = a + bx$ — regression line of y on x

b_{yx} : Regression coeff of y on x

$x = c + dy$ — regression line of x on y

b_{xy} : regression coeff of ~~y on x~~ x on y

Correlation coefficient

$$r = \sqrt{b_{yx} \times b_{xy}}$$

Positive

Geometric mean

Negative

$r = \pm 0.9$

r^2 = Determination

$= 0.81$

Example :-

company	Advt Expt	Sales Revenue
A	1	1
B	3	2
C	4	2
D	6	4
E	8	6
F	9	8
G	11	8
H	14	9

$y = a + bx$

$\Sigma y = an + b\Sigma x$

$\Sigma xy = a\Sigma x + b\Sigma x^2$

Example :-

Sales Revenue y	Advt expt. x	x^2	xy
1	1	1	1
2	3	9	6
2	4	16	8
4	6	36	24
6	8	64	48
8	9	81	72
8	11	121	88
9	14	196	126
$\Sigma 40$	$\Sigma 56$	$\Sigma 524$	$\Sigma 373$

$y = \beta_0 + \beta_1 x$

$\Sigma y = n\beta_0 + \beta_1 \Sigma x$

$\Sigma xy = \beta_0 \Sigma x + \beta_1 \Sigma x^2$

$\Rightarrow 40 = 8\beta_0 + 56\beta_1$

$373 = 56\beta_0 + 524\beta_1$

on solving

$\beta_0 = 0.072$

$\beta_1 = 0.704$

$\therefore y = (0.072) + (0.704)x$

i.e. $y = (0.072) + (0.704)x$

when $x = 0.075$, then

$y = (0.072) + (0.704)(0.075)$

$= 0.1248 \simeq 12.48\%$

Example:

Consider the following data

x	1	2	4	0
y	0.5	1	2	0

Fit a linear regression line

Estimate y when $x = 5$.

x	y	xy	x^2
1	0.5	0.5	1
2	1	2	4
4	2	8	16
0	0	0	0
$\Sigma = 7$	$\Sigma 3.5$	$\Sigma 10.5$	$\Sigma 21$

$y = \beta_0 + \beta_1 x$

$\Sigma y = n\beta_0 + \beta_1 \Sigma x$

$\Sigma xy = \beta_0 \Sigma x_1 + \beta_1 \Sigma x^2$

$3.5 = 4\beta_0 + \beta_1 (7)$

$10.5 = 7\beta_0 + \beta_1 (21)$

on solving these

$\beta_0 = 0$

$\beta_1 = 0.5$

i.e $y = 0 + (0.5)x$

When $x = 5$, $y = (0.5)^5$

$= 0.25$

Linear regression

(Multiple regression)

Example:-

No ↓	Size	No of rooms	No of floors	Age of home	Price Lakhs
1	2000	5	2	45	4000
1	1400	3	1	40	2000
1	1600	3	2	30	3000
1	800	2	1	35	2000
	x_1	x_2	x_3	x_4	y

House → price

$y \rightarrow x$ (my part)

y = increment

$(y) =$

20 years
1 floor
2 rooms
1200 sqft

Multiple Linear regression

$y = \beta_0 + \beta_1 x$ → $\Sigma y = \beta_0 n + \beta_1 \Sigma x$, $\Sigma xy = \beta_0 \Sigma x + \beta_1 \Sigma x^2$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$$

$$\Sigma y = \beta_0 n + \beta_1 \Sigma x_1 + \beta_2 \Sigma x_2 + \beta_3 \Sigma x_3 + \beta_4 \Sigma x_4$$

$$\Sigma x_1 y = \beta_0 \Sigma x_1 + \beta_1 \Sigma x_1^2 + \beta_2 \Sigma x_1 x_2 + \beta_3 \Sigma x_1 x_3 + \beta_4 \Sigma x_1 x_4$$

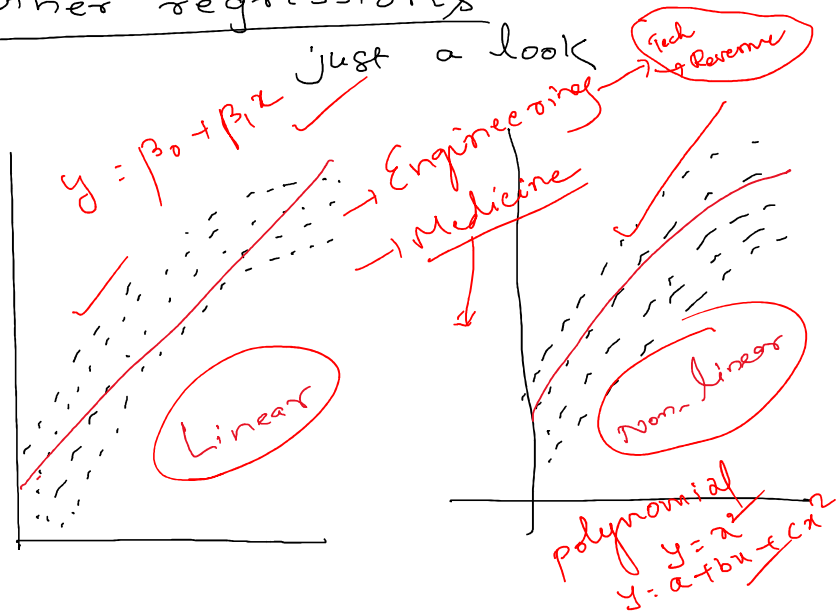
$$\Sigma x_2 y = \beta_0 \Sigma x_2 + \beta_1 \Sigma x_1 x_2 + \beta_2 \Sigma x_2^2 + \beta_3 \Sigma x_2 x_3 + \beta_4 \Sigma x_2 x_4$$

$$\Sigma x_3 y = \beta_0 \Sigma x_3 + \beta_1 \Sigma x_1 x_3 + \beta_2 \Sigma x_2 x_3 + \beta_3 \Sigma x_3^2 + \beta_4 \Sigma x_3 x_4$$

$$\Sigma x_4 y = \beta_0 \Sigma x_4 + \beta_1 \Sigma x_1 x_4 + \beta_2 \Sigma x_2 x_4 + \beta_3 \Sigma x_3 x_4 + \beta_4 \Sigma x_4^2$$

solve $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4$

other regressions just a look



Suppose $y = ae^{bx}$ → exponential curve → non-linear

$\log y = \log a + b \log x$

$Y \quad A \quad X$

i.e. $Y = A + bX$ → linear eqn

$\Sigma Y = An + b \Sigma X \rightarrow (1)$

$\Sigma XY = A \Sigma X + b \Sigma X^2 \rightarrow (2)$

$A = ?$ ⇒ a

b

Hence, we get $y = ae^{bx}$

Suppose $y = ax^b$ → non Linear → Power Curve

$\log y = \log a + b \log x$

$Y \quad A \quad X$

i.e. $Y = A + bX$

$\Sigma Y = An + b \Sigma X$

$\Sigma XY = A \Sigma X + b \Sigma X^2$

$A \rightarrow a$

$b \rightarrow b$

$y = ax^b$

Matrix Approach:

Let $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_{p-1} x_{p-1}$

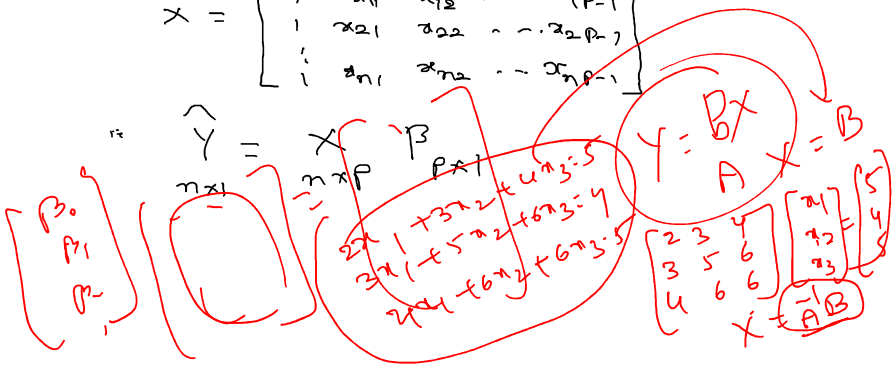
observations $y_i = 1, 2, \ldots n$ → by a vector Y

unknowns $\beta_0, \beta_1 \cdots \beta_{p-1}$ → " " β

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1(p-1)} \\ 1 & x_{21} & x_{22} & \cdots & x_{2(p-1)} \\ \vdots & & & & \\ 1 & x_{n1} & x_{n2} & \cdots & x_{n(p-1)} \end{bmatrix}$$

$$\hat{Y} = X \beta$$

$n \times 1 \quad n \times p \quad p \times 1$



Find β to minimize

$$S(\beta) = \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_1 - \beta_2 x_2 \cdots \right)^2$$

$$= \| Y - X\beta \|^2 = \| Y - \hat{Y} \|^2$$

Diff S wrt to each β we get Linear eqns

$$X^T X \hat{\beta} = X^T Y \rightarrow \text{normal eqns}$$

If $X^T X$ is non singular, the soln is

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

$AX = B$

$X = A^{-1}B$

computationally, it is sometimes unwise even to form the normal equations because multiplications involved in forming X^TX can introduce undesirable round-off error.

→ If X^TX is non-invertible ... ?

✓ Redundant features

✓ too many features → scaling

Revision

→ probability → $P(A \cup B)$

$P(A \cap B)$

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$

→ Conditional probability:

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

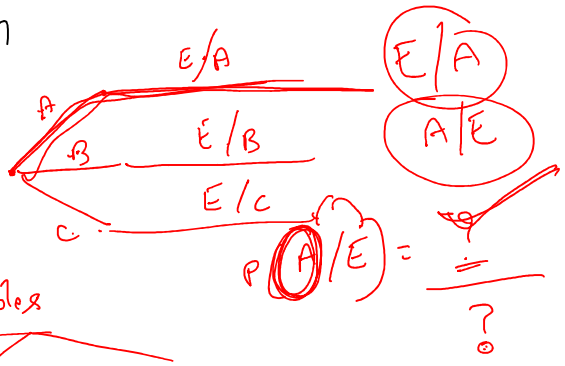
$P(B/A) = \frac{P(A \cap B)}{P(A)}$

$P(A \cap B) = P(A|B)P(B)$

$= P(B|A)P(A)$

Revision

→ Bayes theorem :-



→ Random variables

Discrete $P(x)$

i) $0 \le P(x) \le 1$

ii) $\sum P(x) = 1$

Mean $= E(X) = \sum x P(x)$

Continuous $f(x)$

i) $0 \le f(x) \le 1$

ii) $\int f(x)\,dx = 1$

$= \int x f(x)\,dx$

Variance : $E(x-\mu)^2$

$= E(X^2) - \mu^2$

$= E(X^2) - [E(X)]^2$

Revision

→ Binomial dist $P(x) = {}^nC_x\, p^x q^{n-x}$, $x = 0,1,2 \cdots n$

poisson dist $P(x) = \dfrac{e^{-\lambda}\lambda^x}{x!}$, $x = 0,1,2 \cdots \infty$

→ Normal distribution :- $P(30 \le x \le 50)$

$z = \dfrac{x-\mu}{\sigma}$

$P(z_1 \le z \le z_2)$

$= F(z_2) - F(z_1)$

→ Testing of Hypothesis → Mean → one → z, t

two → z, t

→ Proportion → χ^2 - distribution

L- 11: Predictive Analytics(Continued) & Forecasting Models

Agenda

- Model validation
- Ridge and lasso models
- Assumptions of Linear regression
- Logistic regression

369/54

Classical Linear Regression (OLS)

- Explanatory and Response Variables are Numeric
- Relationship between the mean of the response variable and the level of the explanatory variable assumed to be approximately linear (straight line)
- Model:

$$Y = \beta_0 + \beta_1 x + \varepsilon \qquad \varepsilon \sim N(0, \sigma)$$

- $beta_1 > 0 \Rightarrow$ Positive Association
- $beta_1 < 0 \Rightarrow$ Negative Association
- $beta_1 = 0 \Rightarrow$ No Association

370/54

Multiple regression

Numeric Response variable (y)

p Numeric predictor variables

Model:

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon$$

371/54

• Population Model for mean response:

$$E(Y \mid x_1, \ldots x_p) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

• Least Squares Fitted (predicted) equation, minimizing *SSE*:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p \qquad SSE = \sum \left(Y - \hat{Y} \right)^2$$

Accuracy of a model

- By Using the following the strength of the linear model can be tested

1) Coefficient of determination (R^2)

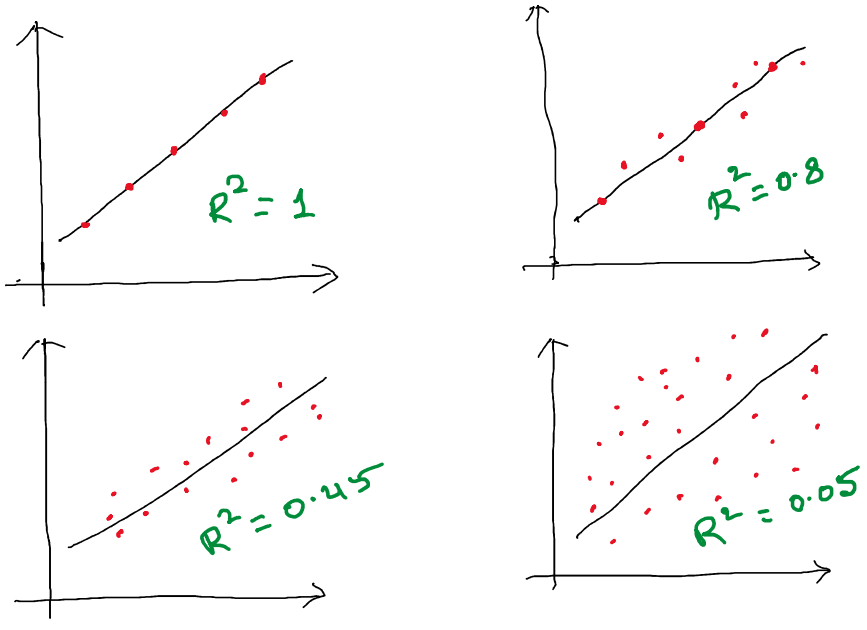
2) Residual Standard error (RSE)

RSS → Residual sum of squares

$$= \sum_{i=1}^{n} \left(y_i - (\alpha + \beta x_i) \right)^2$$

TSS → $\sum_{i=1}^{n} \left(y_i - \bar{y} \right)^2$ → mean of respective variables

$$R^2 = 1 - \frac{RSS}{TSS}$$



R – Squared vs Adjusted R - Squared

- In multiple regression, adjusted R – squared is better metric than R – squared asses the goodness of fit of the model

- R – squared always increases if additional variables are added into model , even if they are not related to the dependent variable

Regularization

- Over fitting can be solved with regularization

- Regularization can be done by putting constraints on the coefficients and variables.

- LASSO: Least Absolute Shrinkage and Selection Operator
  - Some coefficients can be dropped(i.e become zero)

- RIDGE: The coefficients will approach zero, but never dropped

Lasso & Ridge

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p$$

- OLS estimation: $\min SSE = \sum \left(Y - \hat{Y} \right)^2$
- LASSO estimation: $\min SSE = \sum_{i=1}^{n} \left(Y - \hat{Y} \right)^2 + \lambda \sum_{j=1}^{p} \left| \beta_j \right|$
- Ridge regression estimation: $\min SSE = \sum_{i=1}^{n} \left(Y - \hat{Y} \right)^2 + \lambda \sum_{j=1}^{p} \left| \beta_j^{\,2} \right|$

Assumptions in Regression Analysis

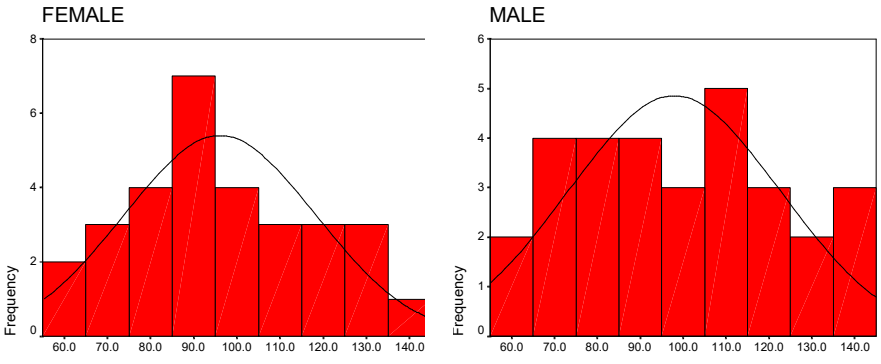
Assumptions

- The distribution of residuals is normal (at each value of the dependent variable).
- The variance of the residuals for every set of values for the independent variable is equal.
  - ✓ violation is called heteroscedasticity.
- The error term is additive
  - ✓ no interactions.
- At every value of the dependent variable the expected (mean) value of the residuals is zero
  - ✓ No non-linear relationships

379

- The expected correlation between residuals, for any two cases, is 0.
  - The independence assumption (lack of autocorrelation)
- ✓ All independent variables are uncorrelated with the error term.
- ✓ No independent variables are a perfect linear function of other independent variables (no perfect multicollinearity)
- ✓ The mean of the error term is zero.

Assumption 1: The Distribution of Residuals is Normal at Every Value of the Dependent Variable



382

Non-Normality

- **Skew and Kurtosis**
  - Skew – much easier to deal with
  - Kurtosis – less serious anyway
- **Transform data**
  - removes skew
  - positive skew – log transform
  - negative skew - square

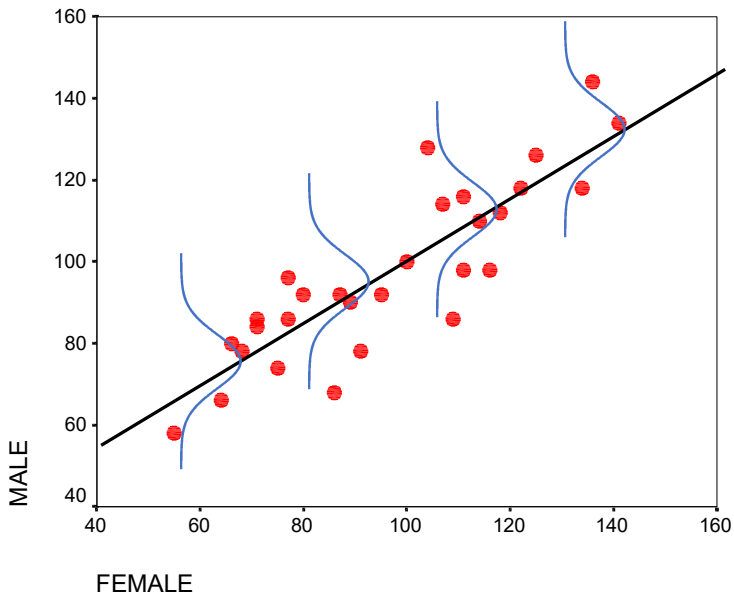
383

Assumption 2: The variance of the residuals for every set of values for the independent variable is equal.

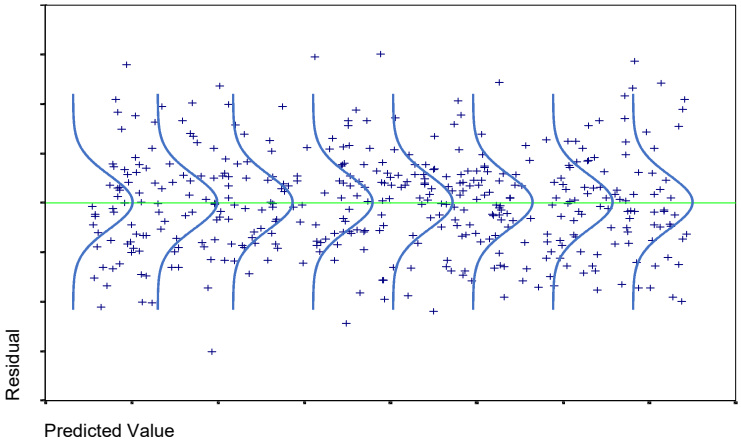
Heteroscedasticity

- This assumption is a about heteroscedasticity of the residuals
  - Hetero=different
  - Scedastic = scattered
- We don't want heteroscedasticity
  - we want our data to be homoscedastic
- Draw a scatterplot to investigate

385

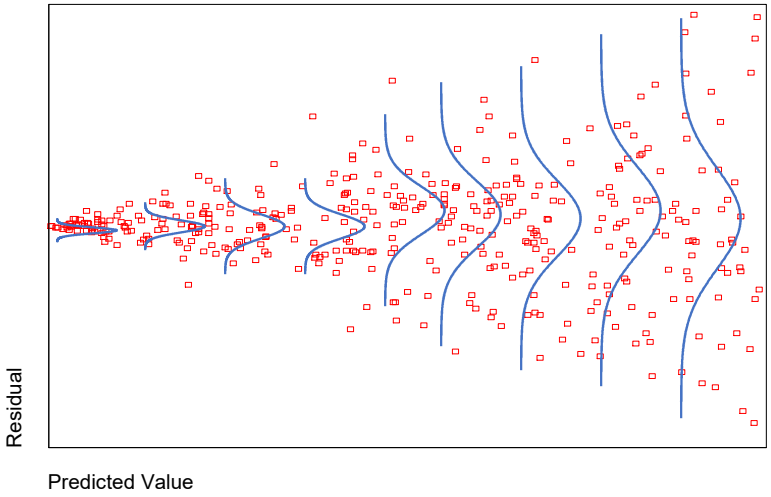


Good – no heteroscedasticity



387

Bad – heteroscedasticity



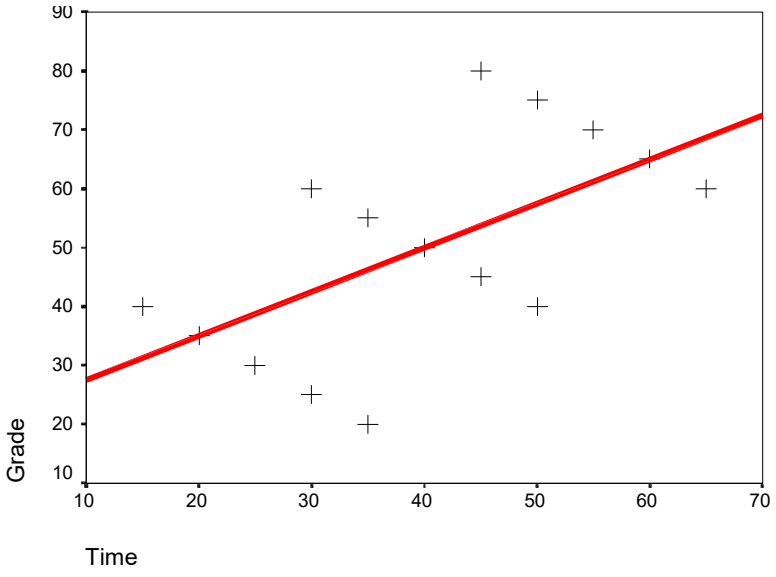
388

Assumption 3:
The Error Term is Additive

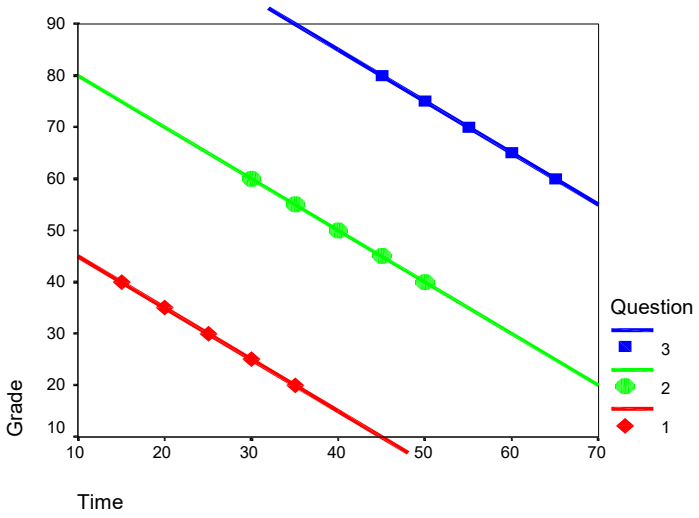
Assumption 4: At every value of the dependent variable the expected (mean) value of the residuals is zero

Assumption 5: The expected correlation between residuals, for any two cases, is 0.

- Result, with line of best fit



- Now somewhat different



393

Assumption 6: All independent variables are uncorrelated with the error term.

Assumption 7: No independent variables are a perfect linear function of other independent variables

Assumption 8: The mean of the error term is zero.

Multicollinearity

- Correlation Matrix

	x_1	x_2	x_3	x_4
x_1	1	-0.80	0.98	0.061
x_2	-0.80	1	-0.184	0.103
x_3	0.98	-0.184	1	0.119
x_4	0.061	0.103	0.119	1

VIF(Variance Inflation Factor)

- **VIF(Variance Inflation Factor)**

- The better way to assess multi collinearity is to compute the VIF

$$VIF = \frac{1}{1 - R^2}$$

- If VIF = 1 then Variables are not correlated
- 1< VIF < 5 then the variables are moderately correlated
- VIF > 5 then highly correlated and need to be eliminated from the model

Logistic Regression

Why use logistic regression?

- There are many important research topics for which the dependent variable is "limited."
- For example: voting, morbidity or mortality, and participation data is not continuous or distributed normally.
- Logistic regression is a type of regression analysis where the dependent variable is a dummy variable: coded 0 (did not vote) or 1(did vote)

Logistic Regression

- Logistic regression is a supervised classification model.
- This allows us to make predictions from labelled data ,if the target variable is categorical.
- Binary classification
- Examples

1. A customer will default on a loan or not
2. A particular machine will break down in the next month or not
3. Predicting whether an incoming email is spam or not

Categorical Response Variables

Examples:

Whether or not a person smokes — Binary Response

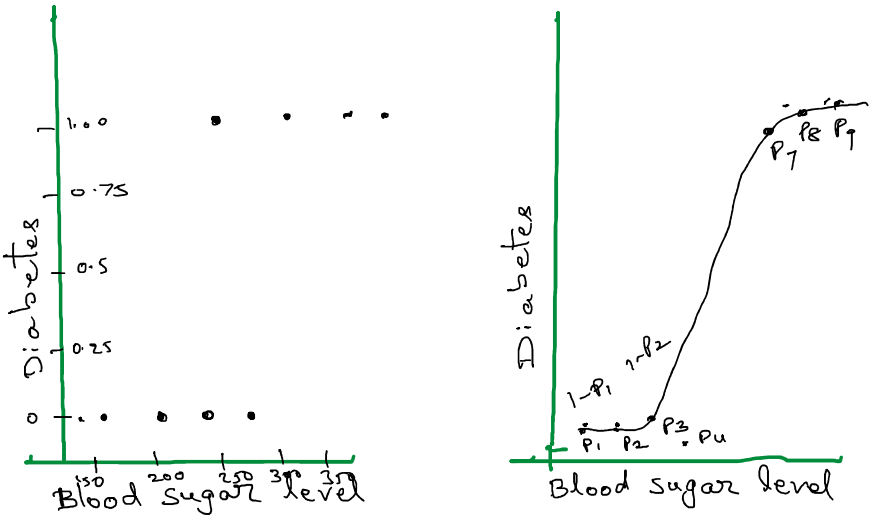
$$Y = \begin{cases} \text{Non} - \text{smoker} \\ \text{Smoker} \end{cases}$$

Success of a medical treatment

$$Y = \begin{cases} \text{Survives} \\ \text{Dies} \end{cases}$$

Opinion poll responses — Ordinal Response

$$Y = \begin{cases} \text{Agree} \\ \text{Neutral} \\ \text{Disagree} \end{cases}$$



$$P(\text{diabetes}) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

Likelihood $= (1-P_1)(1-P_2)(1-P_3)(1-P_4)$

$P_5 (1-P_6) P_7 P_8 P_9 P_{10}$

i.e. $[(1-P_i)(1-P_j)\ldots$ for all non diabetics$]$.

* $[P_i \cdot P_i \ldots$ for all diabetes$]$

Bina Y = Binary response X = Quantitative predictor

p = proportion of 1's (yes, success) at any X

Equivalent forms of the logistic regression model:

Logit form

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X$$

Probability form

$$p = \frac{e^{\beta_0+\beta_1 X_1}}{1+e^{\beta_0+\beta_1 X_1}} = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1)}}$$

Binary Logistic Regression via R

```
> logitmodel=glm(Gender~Hgt,family=binomial, data=Pulse)
> summary(logitmodel)


Call:
glm(formula = Gender ~ Hgt, family = binomial)

Deviance Residuals:
     Min       1Q   Median       3Q      Max
-2.77443  -0.34870  -0.05375   0.32973   2.37928

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  64.1416     8.3694   7.664 1.81e-14 ***
Hgt          -0.9424     0.1227  -7.680 1.60e-14***
---
```

```
Call:
glm(formula = Gender ~ Hgt, family = binomial, data = Pulse)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  64.1416     8.3694   7.664 1.81e-14 ***
Hgt          -0.9424     0.1227  -7.680 1.60e-14***
---
```

$$p = \frac{e^{64.14 - 0.9424Ht}}{1 + e^{64.14 - .9424Ht}}$$

proportion of females at that Hgt

Example: TMS for Migraines

Transcranial Magnetic Stimulation vs. Placebo

Pain Free?	TMS	Placebo
YES	39	22
NO	61	78
Total	100	100

$P_{TMS} = 0.39$ $odds_{TMS} = \frac{39/100}{61/100} = \frac{39}{61} = 0.639$ $P = \frac{0.639}{1 + 0.639} = 0.39$

$P_{Placebo} = 0.22$ $odds_{Placebo} = \frac{22}{78} = 0.282$

$Odds\ ratio = \frac{0.639}{0.282} = 2.27$ Odds are 2.27 times higher of getting relief using TMS than placebo

Logistic Regression for TMS data

```
> lmod=glm(cbind(Yes,No)~Group,family=binomial,data=TMS)
> summary(lmod)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.2657     0.2414  -5.243 1.58e-07 ***
GroupTMS      0.8184     0.3167   2.584  0.00977 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6.8854  on 1  degrees of freedom
Residual deviance: 0.0000  on 0  degrees of freedom
AIC: 13.701
```

Note: $e^{0.8184} = 2.27$ = odds ratio

Binary Logistic Regression Model

Y = Binary $X_1, X_2, \ldots, X_k$ = Multiple

π = proportion of 1's at any $x_1, x_2, \ldots, x_k$

Equivalent forms of the logistic regression model:

Logit form $$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

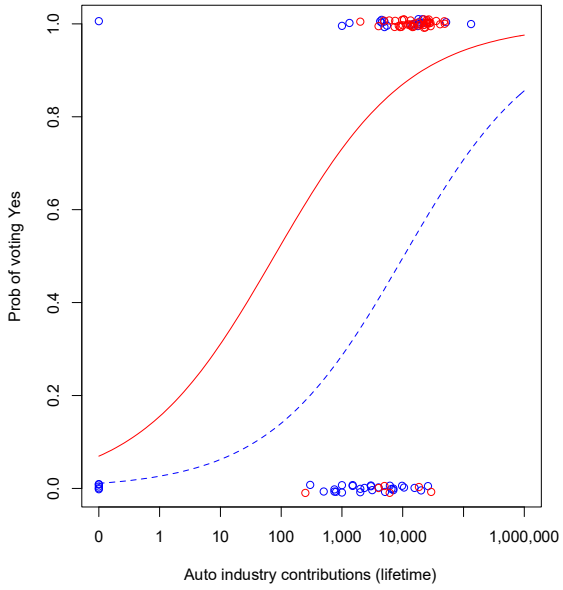
Probability form $$p = \frac{e^{\beta_o + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k}}{1 + e^{\beta_o + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k}}$$

$$= \frac{1}{1 + e^{-(\beta_o + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k)}}$$

Interactions in logistic regression

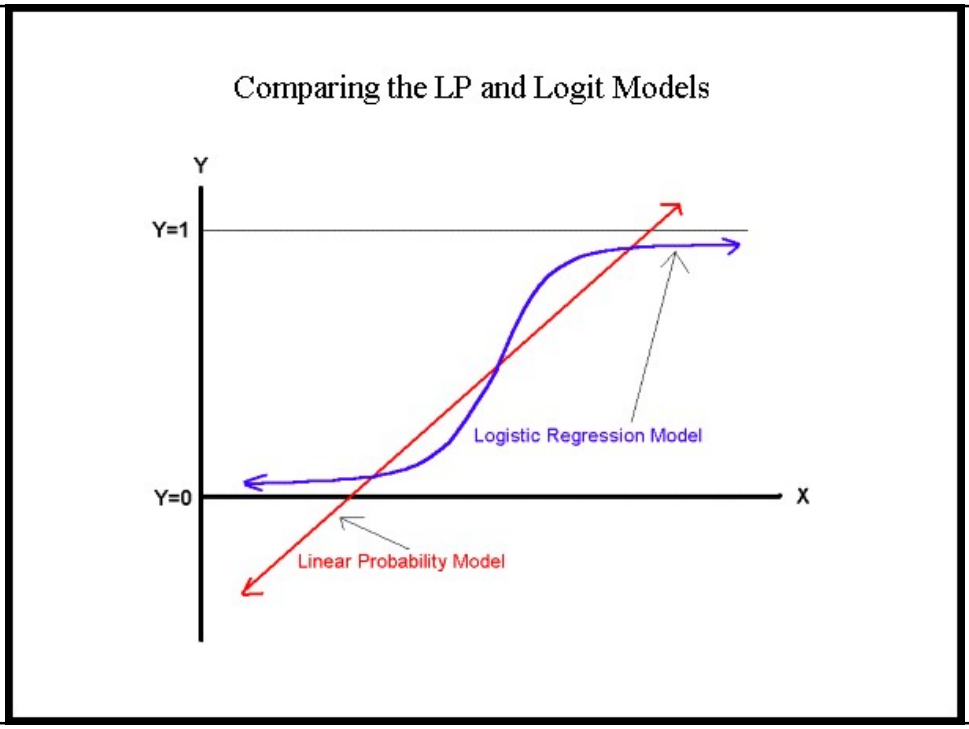
Consider Survival in an ICU as a function of SysBP -- BP for short – and Sex

```
> intermodel=glm(Survive~BP*Sex, family=binomial, data=ICU)
> summary(intermodel)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.439304   1.021042  -1.410  0.15865
BP           0.022994   0.008325   2.762  0.00575 **
Sex          1.455166   1.525558   0.954  0.34016
BP:Sex      -0.013020   0.011965  -1.088  0.27653

    Null deviance: 200.16  on 199  degrees of freedom
Residual deviance: 189.99  on 196  degrees of freedom
```



Rep = red,
Dem = blue

Lines are very close to parallel; not a significant interaction



Forecasting models

- Principles of forecasting
- Time series analysis
- Smoothing and decomposition methods
- ARIMA
- GARCH
- Holt – winter model
- Casual methods
- Moving averages
- Exponential smoothing

Forecasting

- **Predict the next number in the pattern:**

a) 3.7, 3.7, 3.7, 3.7, 3.7, ?

b) 2.5, 4.5, 6.5, 8.5, 10.5, ?

c) 5.0, 7.5, 6.0, 4.5, 7.0, 9.5, 8.0, 6.5, ?

Forecasting

- **Predict the next number in the pattern:**

a) 3.7, 3.7, 3.7, 3.7, 3.7, 3.7

b) 2.5, 4.5, 6.5, 8.5, 10.5, 12.5

c) 5.0, 7.5, 6.0, 4.5, 7.0, 9.5, 8.0, 6.5, 9.0

What Is Forecasting?

- Process of predicting a future event Underlying basis of all business decisions
  - Production
  - Inventory
  - Personnel
  - Facilities

Why do we need to forecast?

Importance of Forecasting

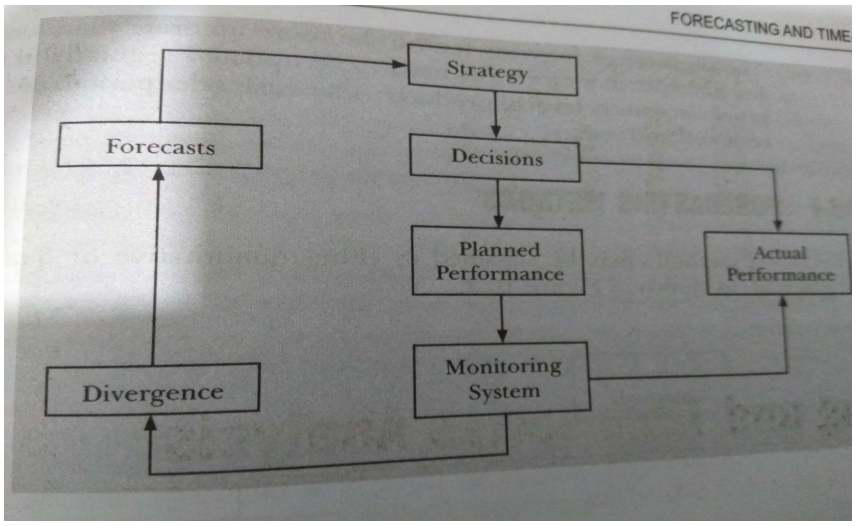
Departments throughout the organization depend on forecasts to formulate and execute their plans.

Finance needs forecasts to project cash flows and capital requirements.

Human resources need forecasts to anticipate hiring needs.

Production needs forecasts to plan production levels, workforce, material requirements, inventories, etc.

- ✓ Demand is not the only variable of interest to forecasters.
- ✓ Manufacturers also forecast worker absenteeism, machine availability, material costs, transportation and production lead times, etc.
- ✓ Besides demand, service providers are also interested in forecasts of population, of other demographic variables, of weather, etc.



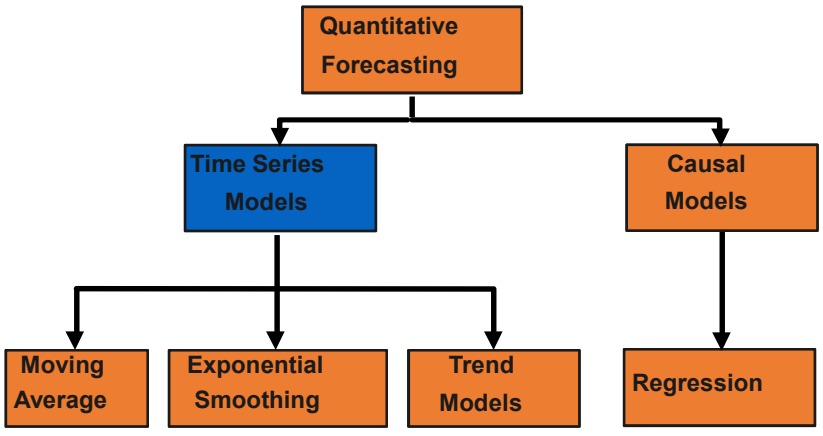
Types of forecasts

- Demand Forecasts
- Environmental Forecasts
- Technological Forecasts

Timing of Forecasts

- ✓ Short-range Forecast
- ✓ Medium – range Forecast
- ✓ Long – range Forecast

Quantitative Forecasting Methods



What is a Time Series?

- Set of evenly spaced numerical data
  - ➢Obtained by observing response variable at regular time periods
- Forecast based only on past values
  - ➢Assumes that factors influencing past, present, & future will continue
- Example

Year:	1995	1996	1997	1998	1999
Sales:	78.7	63.5	89.7	93.2	92.1

Time Series Models

➢Forecaster looks for data patterns as

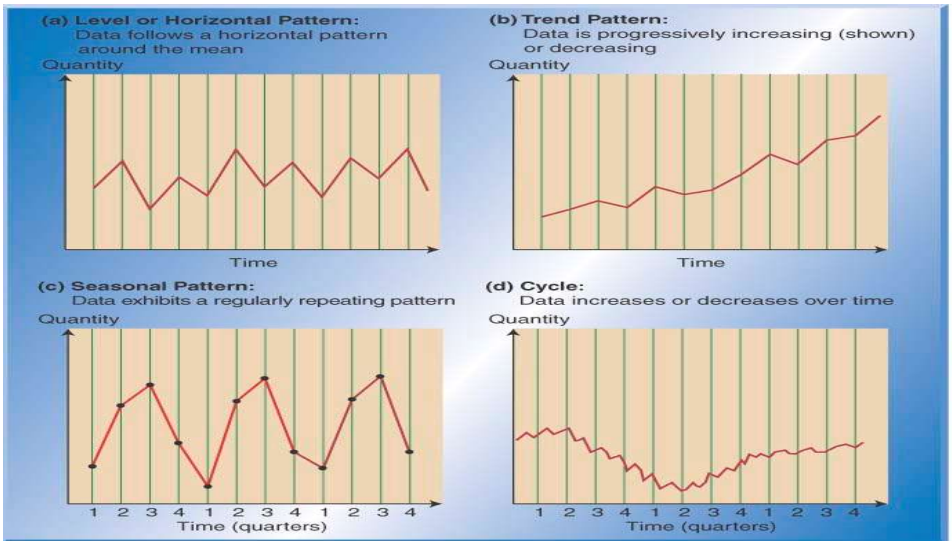
Data = historic pattern + random variation

➢Historic pattern to be forecasted:

- ➢Level (long-term average) – data fluctuates around a constant mean
- ➢Trend – data exhibits an increasing or decreasing pattern
- ➢Seasonality – any pattern that regularly repeats itself and is of a constant length
- ➢Cycle – patterns created by economic fluctuations

➢Random Variation cannot be predicted

Time Series Patterns



Time Series Components

A time series can be described by models based on the following components

$\mathbf{T_t}$	**Trend Component**
$\mathbf{S_t}$	**Seasonal Component**
$\mathbf{C_t}$	**Cyclical Component**
$\mathbf{I_t}$	**Irregular Component**

Using these components we can define a time series as the sum of its components or an **additive model**

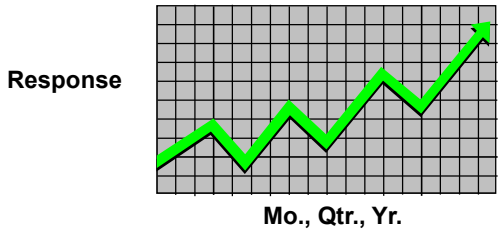
$$X_t = T_t + S_t + C_t + I_t$$

Alternatively, in other circumstances we might define a time series as the product of its components or a **multiplicative model** – often represented as a logarithmic model

$$X_t = T_t S_t C_t I_t$$

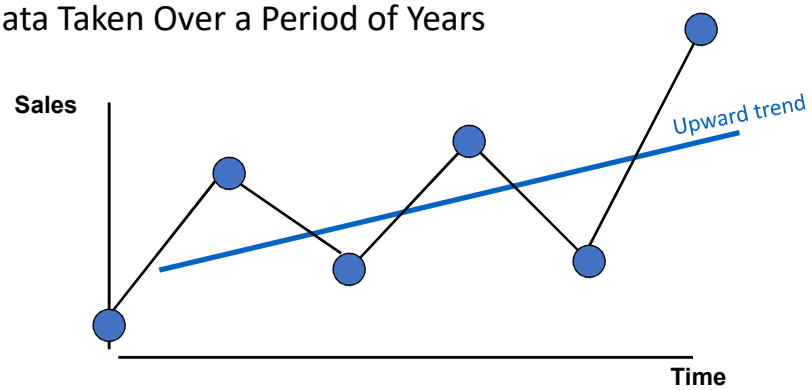
Trend Component

- Persistent, overall upward or downward pattern
- Due to population, technology etc.
- Several years duration



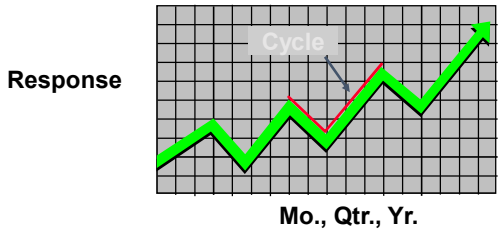
Trend Component

- Overall Upward or Downward Movement
- Data Taken Over a Period of Years



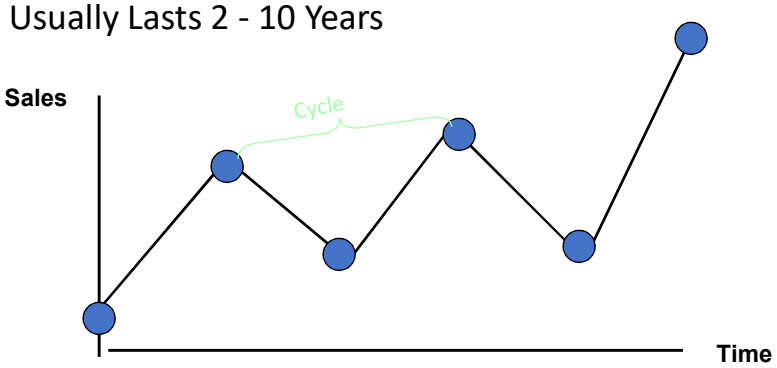
Cyclical Component

- Repeating up & down movements
- Due to interactions of factors influencing economy
- Usually 2-10 years duration



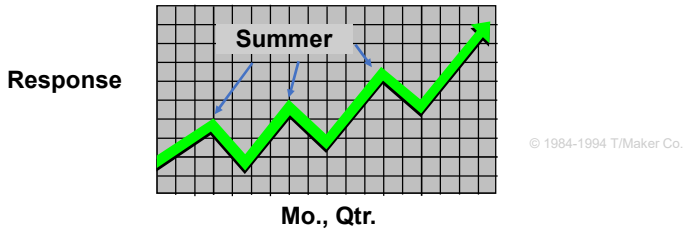
Cyclical Component

- Upward or Downward Swings
- May Vary in Length
- Usually Lasts 2 - 10 Years



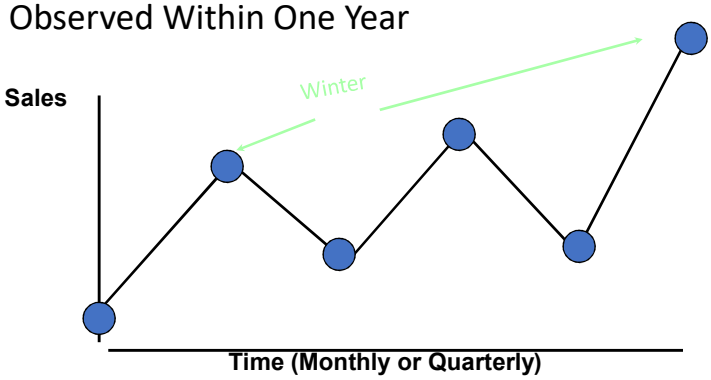
Seasonal Component

- Regular pattern of up & down fluctuations
- Due to weather, customs etc.
- Occurs within one year



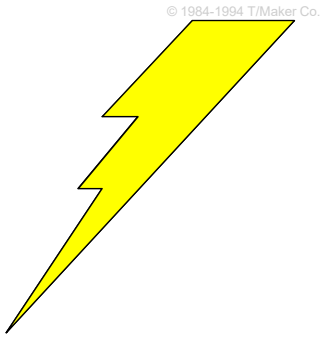
Seasonal Component

- Upward or Downward Swings
- Regular Patterns
- Observed Within One Year



Irregular Component

- Erratic, unsystematic, 'residual' fluctuations
- Due to random variation or unforeseen events
  - Union strike
  - War
- Short duration & nonrepeating



Moving Average Models

- Simple Moving Average Forecast

$$F_t = E(Y_t) = \frac{\sum_{i=t-k}^{t-1} Y_i}{k}$$

Weighted Moving Average Forecast

$$F_t = E(Y_t) = \frac{\sum_{i=t-k}^{t-1} w_i Y_i}{k}$$

Selecting the Right Forecasting Model

1. The amount & type of available data
   - Some methods require more data than others
2. Degree of accuracy required
   - Increasing accuracy means more data
3. Length of forecast horizon
   - Different models for 3 month vs. 10 years
4. Presence of data patterns
   - Lagging will occur when a forecasting model meant for a level pattern is applied with a trend

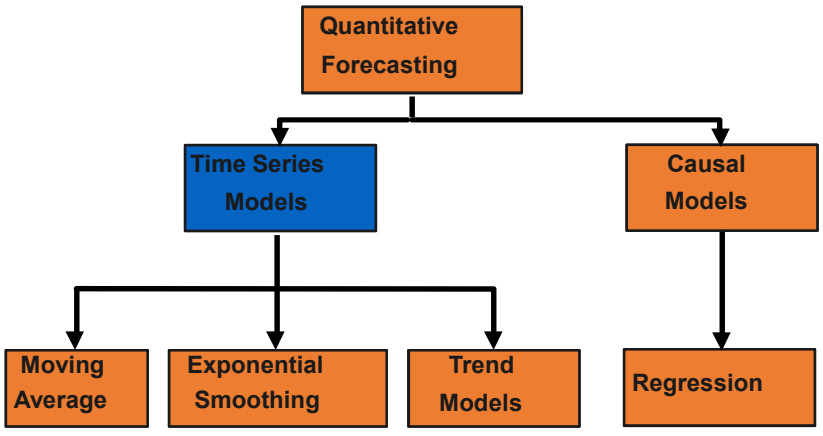
Moving Average
[Solution]

Year	Sales	MA(3) in 1,000
1995	20,000	NA
1996	24,000	(20+24+22)/3 = 22
1997	22,000	(24+22+26)/3 = 24
1998	26,000	(22+26+25)/3 = 24
1999	25,000	NA

Forecasting models

- Principles of forecasting
- Time series analysis
- Smoothing and decomposition methods
- Casual methods
- Moving averages
- Exponential smoothing
- AR,MA,ARMA & ARIMA Models

Quantitative Forecasting Methods



What is a Time Series?

- Set of evenly spaced numerical data
  - ➢Obtained by observing response variable at regular time periods
- Forecast based only on past values
  - ➢Assumes that factors influencing past, present, & future will continue
- Example

Year:	1995	1996	1997	1998	1999
Sales:	78.7	63.5	89.7	93.2	92.1

Applications

- ➢Retail sales
- ➢Spare parts planning
- ➢Stock trading

Time series _ components

- Trend
- Seasonality
- Cyclic
- Random

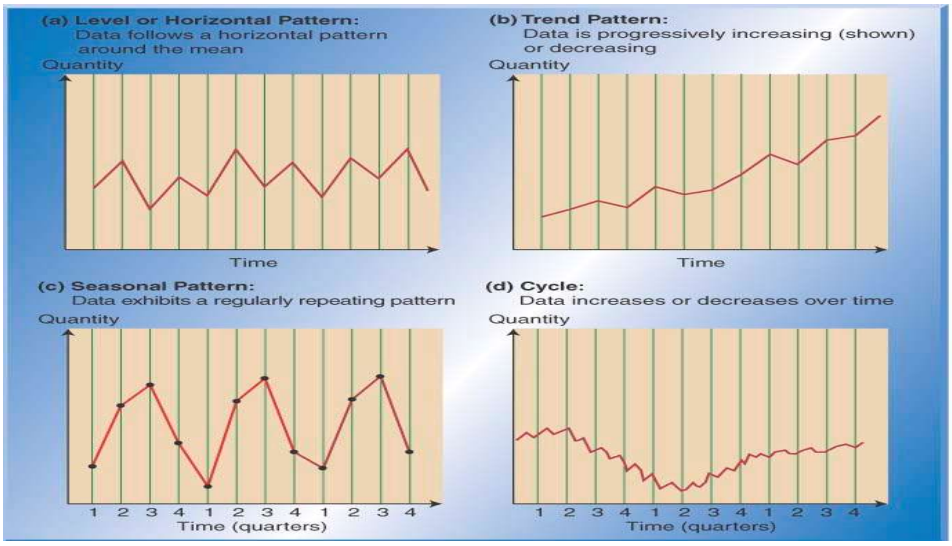
Time Series Models

- Forecaster looks for data patterns as

  Data = historic pattern + random variation

- Historic pattern to be forecasted:
  - Level (long-term average) – data fluctuates around a constant mean
  - Trend – data exhibits an increasing or decreasing pattern
  - Seasonality – any pattern that regularly repeats itself and is of a constant length
  - Cycle – patterns created by economic fluctuations
- Random Variation cannot be predicted

Time Series Patterns



Box – Jenkins Methodology

1. Condition data and select a model

- ❖ identify and account for any trends or seasonality in the time series
- ❖ examine the remaining time series and determine a suitable model

2. Estimate the model parameters

3. Assess the model and return to step 1,if necessary

Time Series Components

A time series can be described by models based on the following components

T_t	**Trend Component**
S_t	**Seasonal Component**
C_t	**Cyclical Component**
I_t	**Irregular Component**

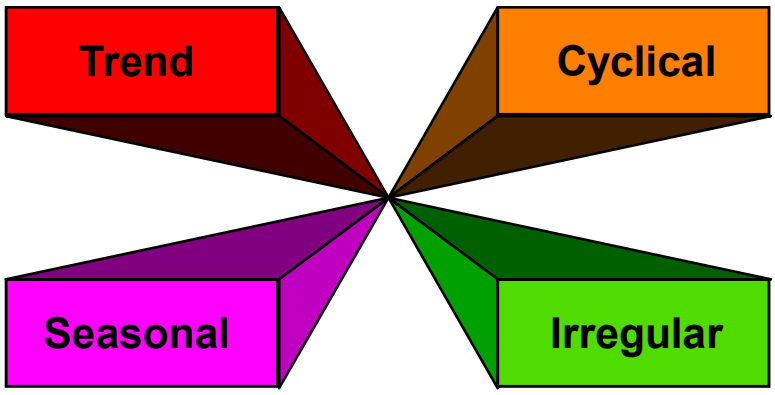
Using these components we can define a time series as the sum of its components or an **additive model**

$$X_t = T_t + S_t + C_t + I_t$$

Alternatively, in other circumstances we might define a time series as the product of its components or a **multiplicative model** – often represented as a logarithmic model

$$X_t = T_t S_t C_t I_t$$

Time Series Components



Smoothing Methods

Moving Average Models

- Simple Moving Average Forecast

$$F_t = E(Y_t) = \frac{\sum_{i=t-k}^{t-1} Y_i}{k}$$

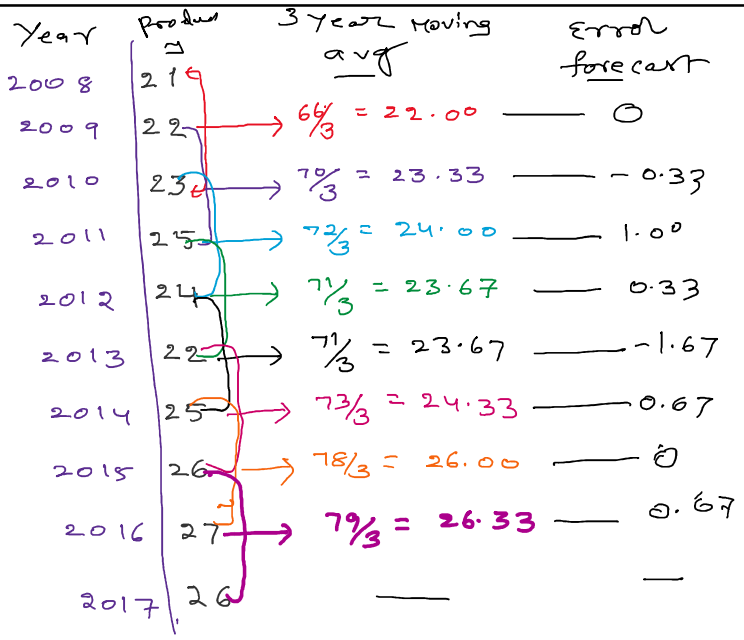
Weighted Moving Average Forecast

$$F_t = E(Y_t) = \frac{\sum_{i=t-k}^{t-1} w_i Y_i}{k}$$

Example(Moving averages)

- Use the following data to compute three year moving average for all available years. Find the trend and Forecast error

YEAR	Saleson (Lakhs)	YEAR	Saleson (Lakhs)
2008	21	2013	22
2009	22	2014	25
2010	23	2015	26
2011	25	2016	27
2012	24	2017	26



Time Series Models

- **Weighted Moving Average:**

- All weights must add to 100% or 1.00
  e.g. C_t **.5**, C_{t-1} **.3**, C_{t-2} **.2** (weights add to 1.0)

$$F_{t+1} = \sum C_t A_t$$

- Allows emphasizing one period over others; above indicates more weight on recent data (C_t=.5)

- Differs from the simple moving average that weighs all periods equally - more responsive to trends

Example(Weighted moving Averages)

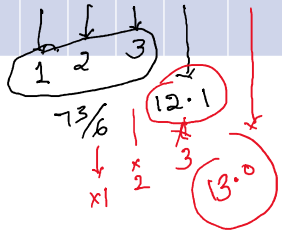
Weights	Month
3	Last month
2	Two months ago
1	Three months ago

Months	1	2	3	4	5	6	7	8	9	10	11	12
Sales	10	12	13	16	19	23	26	30	28	18	16	14

Example(Weighted moving Averages)

Weights	Month
3	Last month
2	Two months ago
1	Three months ago

Months	1	2	3	4	5	6	7	8	9	10	11	12
Sales	10	12	13	16	19	23	26	30	28	18	16	14



Example.

month	Demand
43	105
44	106
45	110
46	110
47	114
48	121
49	130
50	128
51	137

a) Forecast demand for month 52 using 5-month moving Avg

b) " weighted moving average with weights 3, 2, 1 — latest to descending

Example.

month	Demand	
43	105	–
44	106	–
45	110	109.50
46	110	112.2
47	114	117.0
48	121	120.6
49	130	126.0
50	128	–
51	137	–

5 wks

a) $114 + 121 + 130 + 128 + 137 / 5$

$= 126$

b) $\frac{3 \times 137 + 2 \times 128 + 1 \times 130}{6}$

$= 133$ units

Time Series Models

- **Exponential Smoothing:**

Most frequently used time series method because of ease of use and minimal amount of data needed

- Need just three pieces of data to start:
  - Last period's forecast (F_t)
  - Last periods actual value (A_t)
  - Select value of smoothing coefficient, α, between 0 and 1.0

$$F_{t+1} = \alpha A_t + (1-\alpha)F_t$$

$$= F_t + \alpha(A_t - F_t)$$

- If no last period forecast is available, average the last few periods or use naive method
- Higher α values may place too much weight on last period's random variation

Example :-

Forecast for the first week of March was 500 units where as the actual demand is 450 units

a) Forecast demand for the next week i.e March 8

b) Assume the actual demand during the March 8 is 505 units.

continue the forecasting, assuming that subsequent demands were actually 516, 488, 467, 554 and 510 units.

Example :-

Forecast for the first week of March was 500 units where as the actual demand is 450 units

a) Forecast demand for the next week i.e March 8

$$F_{t+1} = F_t + \alpha (A_t - F_t)$$

$$= 500 + 0.1 (450 - 500)$$

$$= 495$$

Week	Demand (A_t)	Forecast (F_t) (old)	New forecast
March 1	450	500	$500 + 0.1(450-500) = 495$
8	505	495	$495 + 0.1(505-495) = 496$
15	516	496	$496 + 0.1(516-496) = 498$
22	488	498	$498 + 0.1(488-498) = 497$
April 1	467	497	$497 + 0.1(467-497) = 494$
8	554	494	$494 + 0.1(554-494) = 500$
15	510	500	$500 + 0.1(510-500) = 501$

Forecasting Trend

- Basic forecasting models for trends compensate for the lagging that would otherwise occur
- One model, **trend-adjusted exponential smoothing** uses a three step process
  - **Step 1 - Smoothing the level of the series**

$$S_t = \alpha A_t + (1-\alpha)(S_{t-1} + T_{t-1})$$

  - **Step 2 – Smoothing the trend**

$$T_t = \beta(S_t - S_{t-1}) + (1-\beta)T_{t-1}$$

  - **Forecast including the trend**

$$FIT_{t+1} = S_t + T_t$$

Measuring Forecasting Accuracy

- **Mean Absolute Deviation (MAD)**
  - measures the total error in a forecast without regard to sign

$$\text{MAD} = \frac{\sum |\text{actual} - \text{forecast}|}{n}$$

- **Cumulative Forecast Error (CFE)**
  - Measures any bias in the forecast

$$\text{CFE} = \sum (\text{actual} - \text{forecast})$$

- **Mean Square Error (MSE)**
  - Penalizes larger errors

$$\text{MSE} = \frac{\sum (\text{actual} - \text{forecast})^2}{n}$$

- **Tracking Signal**
  - Measures if your model is working

$$\text{TS} = \frac{\text{CFE}}{\text{MAD}}$$

Stationarity

Stationary time series have no trend.

conditions

1. constant mean
2. Constant variance
3. An autocovariance that does not depend on time

Auto Correlation

$$\text{auto covariance}_h(X_t) = \text{cov}(X_t, X_{t-h})$$

$$\text{auto correlation}_h(X_t) = \frac{\text{Auto cov}_h(X_t)}{\text{Std}(X_t)\ \text{Std}(X_{t-h})}$$

Auto Correlation Function

$$\text{auto covariance}_h(X_t)$$

$$\gamma_x(h) = \text{cov}(X_t, X_{t-h})$$

$$\text{ACF} = \frac{\gamma_x(h)}{\gamma_x(0)} = \text{Cor}(X_t, X_{t-h})$$

Models

- AR Model → AR(p)
- MA Model → MA(q)
- ARMA Model → ARMA(p,q)
- ARIMA Model

AR Model(Auto regressive model)

AR(p)

$$y_t = \delta_t + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \epsilon_t$$

- δ_t → Constant
- y_{t-1} → value of time series at time t-1
- $\phi_p \neq 0$
- $\epsilon_t \sim N(0, \sigma^2)$ for all t.

Moving Average(MA) Model

today's announcement

Yesterday's

$$y_t = f(\epsilon_t, \epsilon_{t-1}, \epsilon_{t-2}, \dots)$$

$$MA(q) = \theta_0 + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \dots + \theta_q \epsilon_{t-q}$$

θ_k is constant for $k = 1, 2, \dots q$

$\theta_q \neq 0$

$\epsilon_t \sim N(0, \sigma^2)$ for all t.

ARMA model –ARMA(p,q)

$$y_t = \delta + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \dots + \phi_p y_{t-p} + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \dots + \theta_q \epsilon_{t-q}$$

If $p \neq 0$ and $q = 0$, then AR (P)

If $p = 0$ and $q \neq 0$, then MA (Q)

Example:-

Forecast for the first week of March was 500 units whereas the actual demand is 450 units

a) Forecast demand for the next week is March 8

$$F_{t+1} = F_t + \alpha (A_t - F_t)$$

$$= 500 + 0.1 (450 - 500)$$

$$= 495$$

Week	Demand (A_t)	Forecast (F_t) (old)	New forecast
March 1	450	500	$500 + 0.1 (450-500) = 495$
8	505	495	$495 + 0.1 (505-495) = 496$
15	516	496	$496 + 0.1 (516-496) = 498$
22	488	498	$498 + 0.1 (488-498) = 497$
April 1	467	497	$497 + 0.1 (467-497) = 494$
8	554	494	$494 + 0.1 (554-494) = 500$
15	510	500	$500 + 0.1 (510-500) = 501$

Measuring Forecasting Accuracy

- **Mean Absolute Deviation (MAD)**
  - measures the total error in a forecast without regard to sign

$$MAD = \frac{\sum |\text{actual} - \text{forecast}|}{n}$$

- **Cumulative Forecast Error (CFE)**
  - Measures any bias in the forecast

$$CFE = \sum (\text{actual} - \text{forecast})$$

- **Mean Square Error (MSE)**
  - Penalizes larger errors

$$MSE = \frac{\sum (\text{actual} - \text{forecast})^2}{n}$$

- **Tracking Signal**
  - Measures if your model is working

$$TS = \frac{CFE}{MAD}$$

iid noise

The time series in which there is no trend or seasonal component and the observations are simply independent and identically distributed (iid) random variables with zero mean.

Such sequence of random variables $x_1, x_2 \cdots x_n$ as iid noise

Auto Correlation

$$\text{auto covariance}_h(x_t) = \text{cov}(x_t, x_{t-h})$$

$$\text{auto correlation}_h(x_t) = \frac{\text{Auto cov}_h(x_t)}{\text{Std}(x_t)\,\text{Std}(x_{t-h})}$$

Auto Correlation Function

$$\text{auto covariance}_h(x_t)$$

$$\gamma_x(h) = \text{cov}(x_t, x_{t-h})$$

$$\text{ACF} = \frac{\gamma_x(h)}{\gamma_x(0)} = \text{Cor}(x_t, x_{t-h})$$

$$\downarrow$$

$$\rho_x(h)$$

Models

- AR Model → AR(p)
- MA Model → MA(q)
- ARMA Model → ARMA(p,q)

AR Model(Auto regressive model)

AR(p)

$$y_t = \delta_t + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \epsilon_t$$

δ_t → Constant

y_{t-1} → value of time series at time t-1

$\phi_p \neq 0$

$\epsilon_t \sim N(0, \sigma^2)$ for all t.

Moving Average(MA) Model

today's announcement

Yesterday's

$$y_t = f(\epsilon_t, \epsilon_{t-1}, \epsilon_{t-2}, \cdots)$$

$$MA(q) = \theta_0 + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots + \theta_q \epsilon_{t-q}$$

θ_k is constant for $k = 1, 2, \cdots q$

$\theta_q \neq 0$

$\epsilon_t \sim N(0, \sigma^2)$ for all t.

ARMA model –ARMA(p,q)

$$y_t = \delta + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots + \theta_q \epsilon_{t-q}$$

If $p \neq 0$ and $q = 0$, then AR (P)

If $p = 0$ and $q \neq 0$, then MA (Q)

Selecting the Right Forecasting Model

1. The amount & type of available data
   - Some methods require more data than others
2. Degree of accuracy required
   - Increasing accuracy means more data
3. Length of forecast horizon
   - Different models for 3 month vs. 10 years
4. Presence of data patterns
   - Lagging will occur when a forecasting model meant for a level pattern is applied with a trend

Case

- Testing the impact of nutrition and exercise on 60 candidates between age 18 and 50.They are grouped with different strategies.Now we need to find the most effective strategy
- Group 1 eats only junk food
- Group 2 eats only healthy food
- Group 3 eats junk food &does cardio exercise every other day
- Group 4 eats healthy food & does cardio ……
- Group 5 eats junk food& does both cardio & strength training every other day
- Group 6 eats healthy food…….

Which strategy is most effective ?

How to find ?

ANOVA-analysis of variance

- * Significance of difference between two sample means

$H_0 = \mu = \mu_2 = \cdots = \mu_K$

$H_1 = \mu_1 \neq \mu_2 \neq \cdots \neq \mu_K$

Null Hypothesis

Alternate hypothesis

ANOVA

- Effectiveness of different promotional activities
- Quality of a product produced by different manufacturers in terms of an attribute
- Yield of crop due to varieties of seeds , fertilisers and quality of soil

Assumptions

- Each population is normally distributed with mean μ_i With equal variances
- Each sample is drawn randomly and independent of other samples

$$F = \frac{S_1^2}{S_2^2} \quad \nu_1 = n_1 - 1, \quad \nu_2 = n_2 - 1$$

ANOVA summary

Source of Variation	Sum of squares	d.o.f	Mean squares	F-value
Between (Samples)	SSTR	$n-1$	$MSTR = \frac{SSTR}{n-1}$	$F = \frac{MSTR}{MSE}$
within Samples (Error)	SSE	$N-n$	$MSE = \frac{SSE}{N-n}$	
total	SST	$N-1$		

Short cut method

$$T = \Sigma x_1 + \Sigma x_2 + \cdots + \Sigma x_n$$

$$\text{Cal. Fact} = CF = \frac{T^2}{N}, \quad N = n_1 + n_2 + \cdots + n_n$$

$$SST = \left[\Sigma (x_1^2) + \Sigma x_2^2 + \cdots \Sigma x_n^2\right] - CF$$

$$SSTR = \frac{(\Sigma x_j)^2}{n_j} - CF$$

$$SSE = SST - SSTR$$

Example

- To test the significance of variation in the retail prices of a commodity in three metro cities,Mumbai,Kolkata and Delhi, four shops are chosen at random and the prices are given below

Mumbai : 16 8 12 14

Kolkata : 14 10 10 6

Delhi : 4 10 8 8

Prices in 3 cities are significantly different ?

Example

- To test the significance of variation in the retail prices of a commodity in three metro cities,Mumbai,Kolkata and Delhi, four shops are chosen at random and the prices are given below

SST SSTR

Mumbai : 16 8 12 14 50

Kolkata : 14 10 10 6 40

Delhi : 4 10 8 8 30

Prices in 3 cities are significantly different ?

Short cut method

$$T = \Sigma x_1 + \Sigma x_2 + \cdots + \Sigma x_r = 120$$

$$\text{Cal. Fact} = CF = \frac{T^2}{n} = \frac{(120)^2}{12} = 1200 \quad ; n = n_1 + n_2 + \cdots + n_r$$

$$SST = \left[\Sigma (x_1^2) + \Sigma x_2^2 + \cdots \Sigma x_r^2\right] - CF \rightarrow 136$$

$$SSTR = \frac{(\Sigma x_j)^2}{n_j} - CF \rightarrow 50$$

$$SSE = SST - SSTR = 86$$

ANOVA summary

Source of Variation	Sum of squares	d.o.f	Mean squares	F-value
Between (Samples)	SSTR → 50	$r-1$ → 3-1 = 2	MSTR = SSTR/(r-1) = 50/2 = 25	$F = \frac{MSTR}{SSE} = 2.617$
within Samples (Error)	SSE → 86	$n-r$ → 12-3 = 9	MSE = SSE/(n-r) = 86/9 = 9.55	
total	SST → 136	$n-1$ → 12-1 = 11		

Calculated $F = 2.617$

From tables, for $\nu_1 = 2$, $\nu_2 = 9$

($\nu_1 = n-1$, $\nu_2 = n-r$)

at 0.01 Level of significance

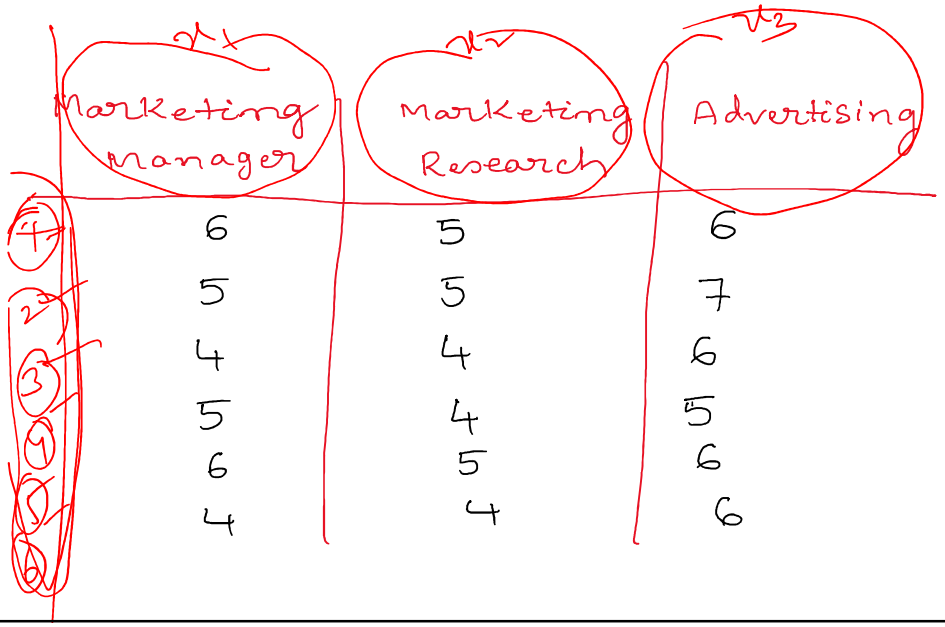
$F = 8.6$

$\therefore F_{cal} < F_{tab} \Rightarrow$ Accept H_0

(If $F_{cal} > F_{tab} \Rightarrow$ reject H_0)

Example

- A study was conducted to investigate the perception of corporate ethical values among individuals specialising in marketing. Using 0.05 level of significance and the data given below, test for significant differences in perception among three groups.(higher scores indicate higher ethical values)



	Marketing Manager	Marketing Research	Advertising
1	6	5	6
2	5	5	7
3	4	4	6
4	5	4	5
5	6	5	6
6	4	4	6

$r = 3, n = 18$

$T = \Sigma x_1 + \Sigma x_2 + \Sigma x_3$

$= 30 + 27 + 36 = 93$

$CF = \frac{T^2}{n} = \frac{(93)^2}{18} = 480.50$

$SST = (\Sigma x_1^2 + \Sigma x_2^2 + \Sigma x_3^2) - CF$

$= 154 + 123 + 218 = 14.50$

$$SSTR = \left(\frac{\Sigma x_1^2}{n_1} + \frac{\Sigma x_2^2}{n_2} + \frac{\Sigma x_3^2}{n_3} \right) - CF$$

$$= \frac{(30)^2}{6} + \frac{(27)^2}{6} + \frac{(36)^2}{6} - 480.50$$

$$= 7$$

$$SSE = SST - SSTR$$

$$= 14.50 - 7 = 7.50$$

$$MSTR = \frac{SSTR}{df_1} = \frac{7}{2} = 3.5$$

$df_1 \rightarrow n-1 = 3-1 = 2$

$$MSE = \frac{SSE}{df_2} = \frac{7.5}{df_2} = 0.5$$

$df_2 \rightarrow n-7 = 18-3 = 15$

$$F = \frac{MSTR}{MSE} = \frac{3.5}{0.5} = 7$$

calculated value: 7

Table value = 3.68 (at 5%)

$7 > 3.68 \Rightarrow$ Rejected.

Example

Month	Sales A	B	C	D
May	50	40	48	39
June	46	48	50	45
July	39	44	40	39

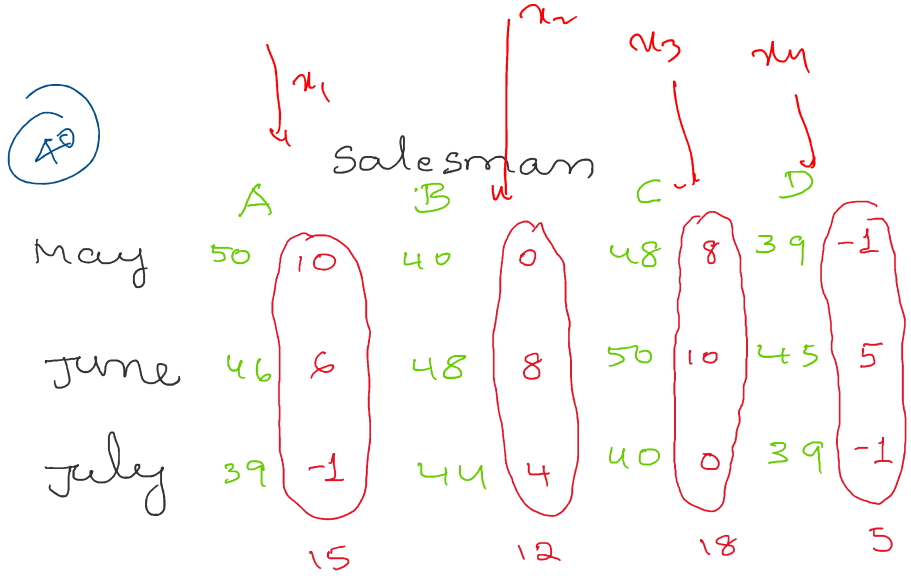
→ Is there any significant diff in the sales by A, B, C, D.

→ Is there a significant diff in the sales made during these months.

Two way ANOVA

Sources of Variation	Sum of Square	DoF	mean square	Test Statistic
Between columns	SSTR	$c-1$	$MSTR = SSTR/(c-1)$	$F_{treatment} = MSTR/MSE$
Between rows	SSR	$n-1$	$MSR = SSR/(n-1)$	$F_{blocks} = MSR/MSE$
Residual error	SSE	$(c-1)(n-1)$	$MSE = SSE/((c-1)(n-1))$	
Total	SST	$n-1$		

(A)



$$T = 15 + 12 + 18 + 3 = 48$$

$$CF = \frac{T^2}{n} = \frac{(48)^2}{12} = 192$$

SSTR = Sum of squares (Salesman / Column)

$$= \left(\frac{15^2}{3} + \frac{12^2}{3} + \frac{18^2}{3} + \frac{3^2}{3}\right) - 192$$

$$= 42$$

SSR = Sum of squares between months (rows)

$= \left(\frac{17^2}{4} + \frac{29^2}{4} + \frac{22^2}{4}\right) - 192$

$= 91.5$

$SST = \left(\Sigma x_1^2 + \Sigma x_2^2 + \Sigma x_3^2 + \Sigma x_4^2\right) - CF$

$= (137 + 80 + 164 + 27) - 192$

$= 216$

$SSE = SST - (SSTR + SSR)$

$= 216 - (42 + 91.5)$

$= 82$

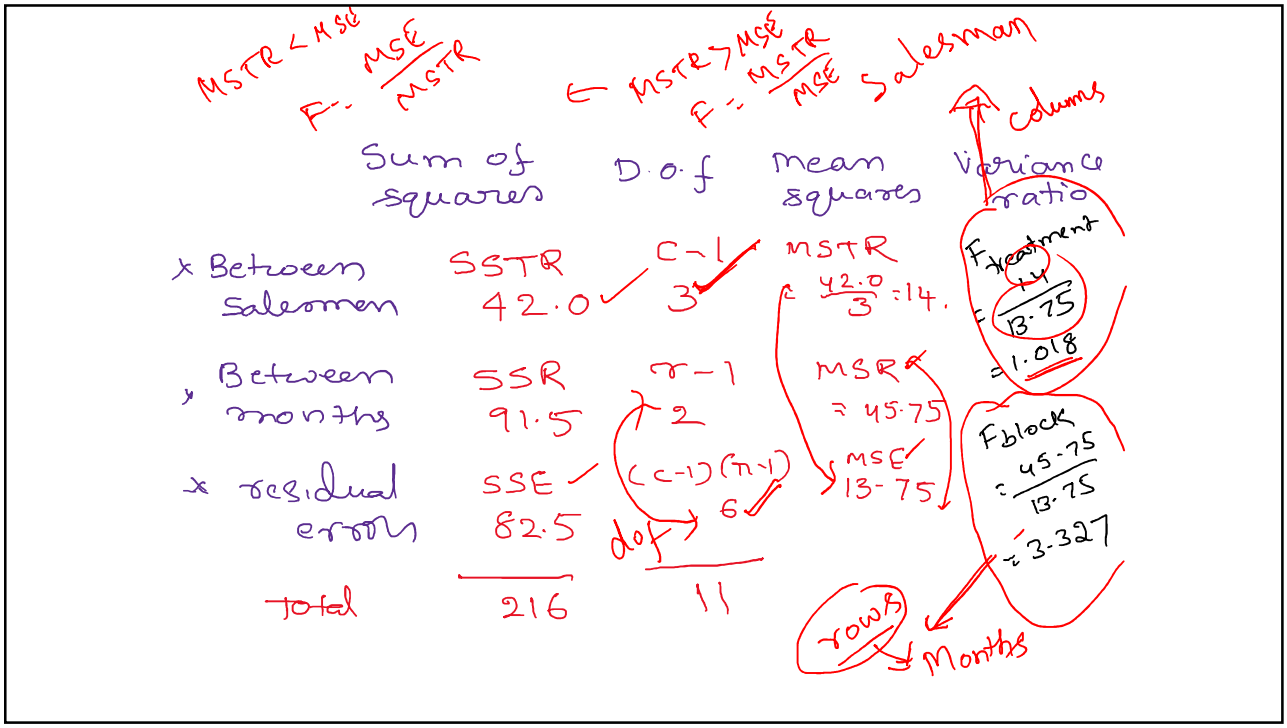
$df_c = 3, \quad df_r = r - 1 = 3 - 1 = 2$

$df = (c-1)(r-1) = 3 \times 2 = 6$

$$MSTR = \frac{SSTR}{c-1} = \frac{42}{3} = 14$$

$$MSR = \frac{SSR}{r-1} = \frac{91.5}{2} = 45.75$$

$$MSE = \frac{SSE}{(c-1)(r-1)} = \frac{82.5}{6} = 13.75$$



a) $F_{treatment} = 1.018 <$

$df_1 = 3, df_2 = 6$

$\alpha = 0.05$

accept

b) $F_{block} = 3.327 <$ ✗

2, 6

accept

difference in the sales ~~by salesman~~

difference in sales made during months.

Multivariate analysis

Introduction

Multivariate normal distribution

Principal component Analysis

Factor Analysis

Discriminant Analysis

MANOVA

Bi-Variate Normal dist

$$P(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left[\frac{-z}{2(1-\rho^2)}\right]$$

$$z = \frac{(x_1-\mu_1)^2}{\sigma_1^2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2} - 2\rho\frac{(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2}$$

$$\rho = \text{cor}(x_1, x_2) = \frac{\text{cov}(x_1, x_2)}{\sigma_1\sigma_2}$$

Multivariate
Normal distribution

$$\phi(x) = \left(\frac{1}{2\pi}\right)^{n/2} |\Sigma|^{-1/2} \exp\left\{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right\}$$

$|\Sigma|$: Det of Variance-covariance matrix

Σ^{-1} : Inverse of Variance-covariance matrix

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

Σ_{12} : covariance between x_1 and x_2.

L- 14:Appled Multivariate Analytics

Agenda

- Multivaraite normal distribution
- Preliminaries …Eigen values and vectors
- Principal component analysis

Bi-Variate Normal dist

$$\phi(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left[\frac{-z}{2(1-\rho^2)}\right]$$

$$z = \frac{(x_1-\mu_1)^2}{\sigma_1^2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2} - 2\rho\frac{(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2}$$

$$\rho = \text{cor}(x_1, x_2) = \frac{\text{cov}(x_1, x_2)}{\sigma_1\sigma_2}$$

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

$$P\left(\frac{X}{Y}\right) = \frac{P(X,Y) \rightarrow \text{joint}}{P_Y(x,y) \rightarrow \text{Marginal of } Y}$$

$P(x=1/y=2)$

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad x \in (-\sigma, \sigma)$$

Multivariate Normal distribution

$$\phi(x) = \left(\frac{1}{2\pi}\right)^{p/2} |\Sigma|^{-1/2} \exp\left\{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right\}$$

x_1, x_2, x_3

Det of Variance-covariance matrix

Inverse of Variance-covariance matrix

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

covariance between x_1 and x_2.

Preliminaries

- Standard Deviation is a measure of the spread of the data
- Variance – measure of the deviation from the mean for points in one dimension e.g. heights
- Covariance as a measure of how much each of the dimensions vary from the mean with respect to each other.
- Covariance is measured between 2 dimensions to see if there is a relationship between the 2 dimensions e.g. number of hours studied & marks obtained
- The covariance between one dimension and itself is the variance

Covariance Matrix

- Representing Covariance between dimensions as a matrix
- e.g.

$$C = \begin{pmatrix} cov(x,x) & cov(x,y) & cov(x,z) \\ cov(y,x) & cov(y,y) & cov(y,z) \\ cov(z,x) & cov(z,y) & cov(z,z) \end{pmatrix}$$

- Diagonal is the variances of x, y and z
- cov(x,y) = cov(y,x) hence matrix is symmetrical about the diagonal
- N-dimensional data will result in nxn covariance matrix

- A **positive value** of covariance indicates both dimensions increase or decrease together
- A **negative value** indicates while one increases the other decreases, or vice-versa
- If **covariance is zero**: the two dimensions are independent of each other .

Transformation matrices

- Consider:

$$\begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix} \times \begin{bmatrix} 3 \\ 2 \end{bmatrix} = \begin{bmatrix} 12 \\ 8 \end{bmatrix} = 4 \times \begin{bmatrix} 3 \\ 2 \end{bmatrix}$$

- Square transformation matrix transforms (3,2) from its original location. Now if we were to take a multiple of (3,2)

$$2 \times \begin{bmatrix} 3 \\ 2 \end{bmatrix} = \begin{bmatrix} 6 \\ 4 \end{bmatrix}$$

$$\begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix} \times \begin{bmatrix} 6 \\ 4 \end{bmatrix} = \begin{bmatrix} 24 \\ 16 \end{bmatrix} = 4 \times \begin{bmatrix} 6 \\ 4 \end{bmatrix}$$

eigenvalue problem

- The eigenvalue problem is any problem having the following form:
- **A . X** = λ . X
  - A: n x n matrix
  - X: n x 1 non-zero vector
  - λ: scalar
- Any value of λ for which this equation has a solution is called the eigenvalue of A and vector v which corresponds to this value is called the eigenvector of A.

eigenvalue problem

$$\begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix} \times \begin{bmatrix} 3 \\ 2 \end{bmatrix} = \begin{bmatrix} 12 \\ 8 \end{bmatrix} = 4 \times \begin{bmatrix} 3 \\ 2 \end{bmatrix}$$

$$A \cdot x = \lambda \cdot x$$

Therefore, (3,2) is an eigenvector of the square matrix **A** and 4 is an eigenvalue of **A**

eigen value

eigen vector corresponding to λ

Given matrix A, how can we calculate the eigenvector and eigenvalues for A?

eigenvalue

eigen vector

$AX = \lambda X$

$AX - \lambda IX = 0$

$[A - \lambda I]X = 0$ → Identity matrix $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$

↓

Homogeneous system of equations

Non trivial soln (non-zero soln)

iff $|A - \lambda I| = 0$

$2x + 3y = 0$

$3x + 4y = 0$

→ Non trivial soln

$x = 0$

$y = 0$

λ, X

$AX = \lambda X$

Ex: $A = \begin{bmatrix} 0 & 1 \\ -2 & -3 \end{bmatrix}$

$\lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$

$|A - \lambda I| = 0 \Rightarrow \begin{vmatrix} 0-\lambda & 1 \\ -2 & -3-\lambda \end{vmatrix} = 0$

$\Rightarrow -\lambda(-3-\lambda) + 2 = 0$

$\Rightarrow \lambda^2 + 3\lambda + 2 = 0$ i.e. $\lambda = -1, -2$

eigen values

Now we need to find X corresponding to $\lambda = -1$ and $\lambda = -2$ such that

$AX = \lambda X$

---

when $\lambda = -1$: $[A - \lambda I]X = 0$

i.e. $\begin{bmatrix} 0+1 & 1 \\ -2 & -2 \end{bmatrix} X = 0$ $\quad X = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$

i.e. $x_1 + x_2 = 0$

and $-2x_1 - 2x_2 = 0$ $\Rightarrow x_1 + x_2 = 0$

non-trivial

$\lambda = -2$

Let $x_1 = K$, $x_2 = -K$

$X_1 = \begin{bmatrix} K \\ -K \end{bmatrix} \rightarrow \begin{bmatrix} 1 \\ -1 \end{bmatrix}$ eigen vector

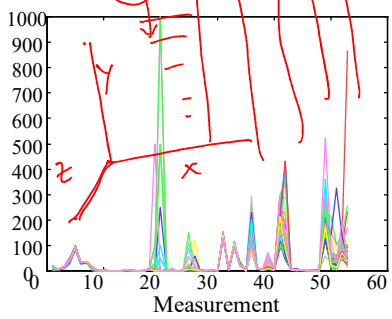
$X_2 = \begin{bmatrix} \quad \\ \quad \end{bmatrix}$

Similarly we find eigen vector corresponding to $\lambda = -2$

Data Presentation

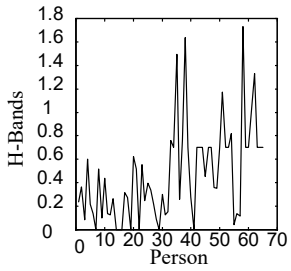
- Blood and urine measurements (wet chemistry) from 65 people (33 alcoholics, 32 non-alcoholics).
- Matrix Format

	H-WBC	H-RBC	H-Hgb	H-Hct	H-MCV	H-MCH	H-MCHC
A1	8.0000	4.8200	14.1000	41.0000	85.0000	29.0000	34.0000
A2	7.3000	5.0200	14.7000	43.0000	86.0000	29.0000	34.0000
A3	4.3000	4.4800	14.1000	41.0000	91.0000	32.0000	35.0000
A4	7.5000	4.4700	14.9000	45.0000	101.0000	33.0000	33.0000
A5	7.3000	5.5200	15.4000	46.0000	84.0000	28.0000	33.0000
A6	6.9000	4.8600	16.0000	47.0000	97.0000	33.0000	34.0000
A7	7.8000	4.6800	14.7000	43.0000	92.0000	31.0000	34.0000
A8	8.6000	4.8200	15.8000	42.0000	88.0000	33.0000	37.0000
A9	5.1000	4.7100	14.0000	43.0000	92.0000	30.0000	32.0000

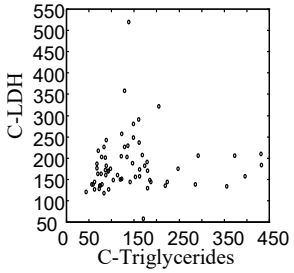


visualization

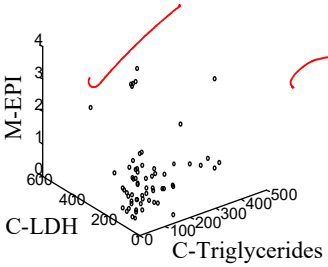
Univariate



Bivariate



Trivariate

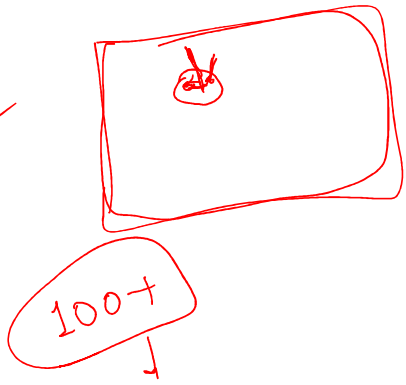


principal components

50

Applications

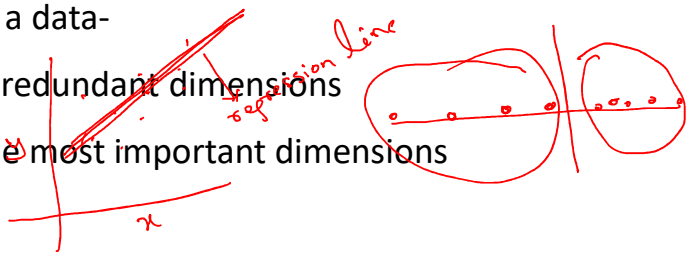
- Face Recognition
- Image Compression
- Gene Expression Analysis
- Data Reduction
- Data Classification
- Trend Analysis
- Factor Analysis
- Noise Reduction



Principal Component Analysis

- In real world data analysis tasks we analyze complex data i.e. multi dimensional data. We plot the data and find various patterns in it or use it to train some machine learning models. One way to think about dimensions is that suppose you have an data point **x** , if we consider this data point as a physical object then dimensions are merely a basis of view, like where is the data located when it is observed from horizontal axis or vertical axis.

- As the dimensions of data increases, the difficulty to visualize it and perform computations on it also increases. So, how to reduce the dimensions of a data-
  * Remove the redundant dimensions
  * Only keep the most important dimensions

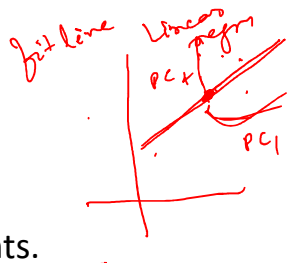


- Now lets think about the requirement of data analysis. Since we try to find the patterns among the data sets so we want the data to be spread out across each dimension. Also, we want the dimensions to be independent. Such that if data has high covariance when represented in some *n* number of dimensions then we replace those dimensions with *linear combination* of those n dimensions. Now that data will only be dependent on linear combination of those related n dimensions. *(related = have high covariance)*

- It is a linear transformation that chooses a new coordinate system for the data set such that
  - greatest variance by any projection of the data set comes to lie on the first axis (then called the first principal component),
  - the second greatest variance on the second axis, and so on.
- PCA can be used for reducing dimensionality by eliminating the later principal components.

- **what does Principal Component Analysis (PCA) do?**

- PCA finds a new set of dimensions (or a set of basis of views) such that all the dimensions are orthogonal (and hence linearly independent) and ranked according to the variance of data along them. It means more important principle axis occurs first. (more important = more variance/more spread out data)



- **How does PCA work**

➢Calculate the covariance matrix *X* of data points.

➢Calculate eigen vectors and corresponding eigen values.

➢Sort the eigen vectors according to their eigen values in decreasing order.

➢Choose first k eigen vectors and that will be the new k dimensions.

➢Transform the original n dimensional data points into k dimensions.

Example:

consider the following data

x	2.5	0.5	2.2	1.9	3.1	2.3	2	1	1.5	1.1
y	2.4	0.7	2.9	2.2	3.0	2.7	1.6	1.1	1.6	0.9

$r = 0.926$

Step 1 :- $\bar{x} = 1.81$, $\bar{y} = 1.91$

x	0.69	-1.31	0.39	0.09	1.29	0.49	0.19	-0.81	-0.31	-0.71
y	0.49	-1.21	0.99	0.29	1.09	0.79	-0.31	-0.81	-0.31	-1.01

$X = [x \quad y]$ — 10×2

origin

Step 2 :-

X, Y ↓ ↓ F1

$$\text{cov}(x) = \begin{bmatrix} 0.6166 & 0.6154 \\ 0.6154 & 0.7166 \end{bmatrix}$$

$$\text{cov}(x,y) = \frac{\sum (x - \bar{x})(y - \bar{y})}{n-1}$$

$$= \frac{1}{n-1} \sum xy$$

$|A - \lambda I| = 0 \Rightarrow \lambda = ?$

$AX = \lambda x$

Step3.

Eigen values (λ) = $\lambda_1 = 1.2840$

$\lambda_2 = 0.0490$

$Ax = \lambda x$

Eigen vectors corresponding

$\lambda_1 \longrightarrow \begin{bmatrix} 0.678 \\ 0.735 \end{bmatrix}$ → 1.2840 → PC_1

$\lambda_2 \longrightarrow \begin{bmatrix} 0.735 \\ -0.678 \end{bmatrix}$ → 0.0490 → PC_2

ie

Variable	Eigen vector 1 $\lambda_1 = 1.2840$	Eigen vector 2 $\lambda_2 = 0.0490$
x_1	0.678	0.735
x_2	0.735	−0.678
% of total variance	1.2840 / 1.333 = 96.3%	0.0490 / 1.333 = 3.7%
	PC_1	PC_2

X

Trace.

total variance

1.333

Step 4 : Select Variation matrix

$$V = \begin{pmatrix} 0.678 & 0.735 \\ 0.735 & -0.678 \end{pmatrix}$$

or

$$V = \begin{pmatrix} 0.678 \\ 0.735 \end{pmatrix}$$

$\lambda_1 \rightarrow PC_1$

λ_1 → highest

Step 5. Find new Data set $Y = XV$

i.e $Y = XV$

Case (i) :- $Y = \begin{bmatrix} 2.5 & 2.4 \\ 0.5 & 0.7 \\ \vdots & \\ 1.1 & 0.9 \end{bmatrix}_{10\times 2} \begin{bmatrix} 0.678 & 0.735 \\ 0.735 & -0.678 \end{bmatrix}_{2\times 2}$

$$= \begin{bmatrix} 3.459 & 0.211 \\ -0.854 & -0.107 \\ \vdots & \vdots \\ 1.407 & 0.199 \end{bmatrix}_{10\times 2}$$

PC_2 y_2

PC_1 y_1

$Y_2 = 0.735 x_1 - 0.678 x_2$

$Y_1 = 0.678 x_1 + 0.735 x_2$

Step5. Find new Data set $Y = XV$

i.e. $Y = XV$

Case (ii) :- $Y = \begin{bmatrix} 2.5 & 2.4 \\ 0.5 & 0.7 \\ \vdots & \\ 1.1 & 0.9 \end{bmatrix}_{10\times 2} \begin{bmatrix} 0.678 \\ 0.735 \end{bmatrix}$

$= \begin{bmatrix} 3.459 \\ -0.854 \\ \vdots \\ 1.407 \end{bmatrix}$

$\lambda_1 =$ λ_1 λ_2 λ_2 $\lambda_1 =$ 0.3

27

linear ~~transf~~ Combination

$Y = 0.678 x_1 + 0.735 x_2$

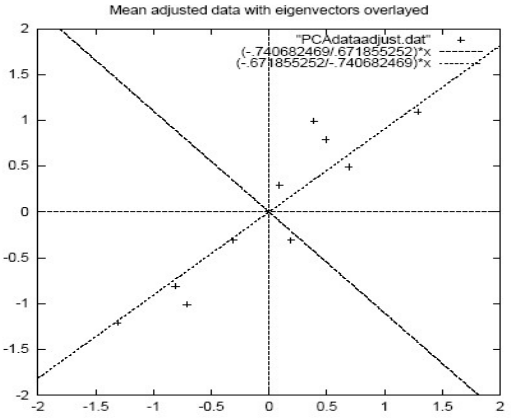


Figure 3.2: A plot of the normalised data (mean subtracted) with the eigenvectors of the covariance matrix overlayed on top.

Summary :- PCA

1) Re centre the original data set to the origin

2) Find covariance matrix x

3) Find eigen values and eigen vectors and also % of variability

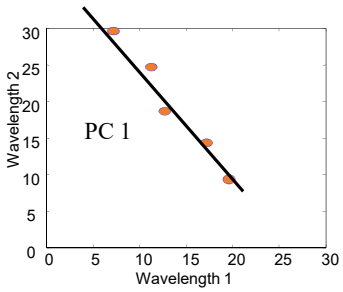
4) Find the transformation matrix V based on PC selection

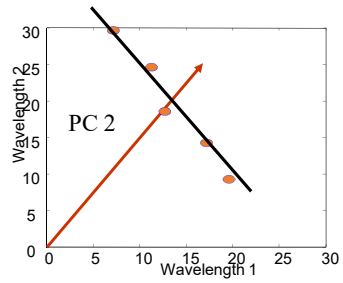
5) Derive the new data set by Y = XV

PC2 PC1 PC1 PC2 PC3

Principal Components

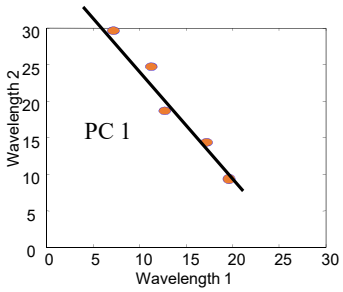
- All principal components (PCs) start at the origin of the ordinate axes.
- First PC is direction of maximum variance from origin
- Subsequent PCs are orthogonal to 1st PC and describe maximum residual variance

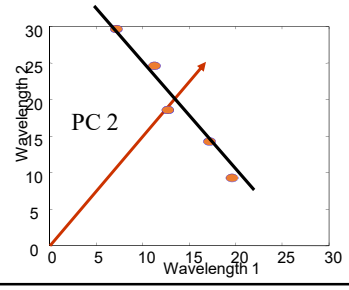




Principal Components

- All principal components (PCs) start at the origin of the ordinate axes.
- First PC is direction of maximum variance from origin
- Subsequent PCs are orthogonal to 1st PC and describe maximum residual variance

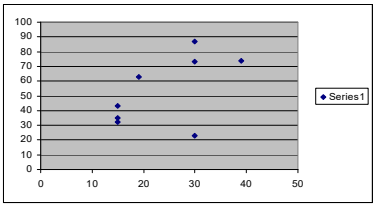


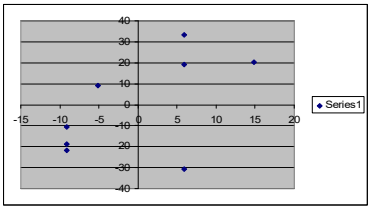


An Example

Mean1=24.1
Mean2=53.8

X1	X2	X1'	X2'
19	63	-5.1	9.25
39	74	14.9	20.25
30	87	5.9	33.25
30	23	5.9	-30.75
15	35	-9.1	-18.75
15	43	-9.1	-10.75
15	32	-9.1	-21.75





558

Covariance Matrix

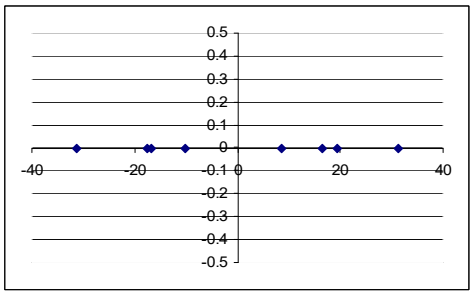
- C=

75	106
106	482

- Using MATLAB, we find out:
  - Eigenvectors:
  - e1=(-0.98,-0.21), λ1=51.8
  - e2=(0.21,-0.98), λ2=560.2
  - Thus the second eigenvector is more important!

If we only keep one dimension: e2

- We keep the dimension of e2=(0.21,-0.98)
- We can obtain the final data as



yi
-10.14
-16.72
-31.35
31.374
16.464
8.624
19.404
-17.63

$$y_i = \begin{pmatrix} 0.21 & -0.98 \end{pmatrix} \begin{pmatrix} x_{i1} \\ x_{i2} \end{pmatrix} = 0.21 * x_{i1} - 0.98 * x_{i2}$$

560

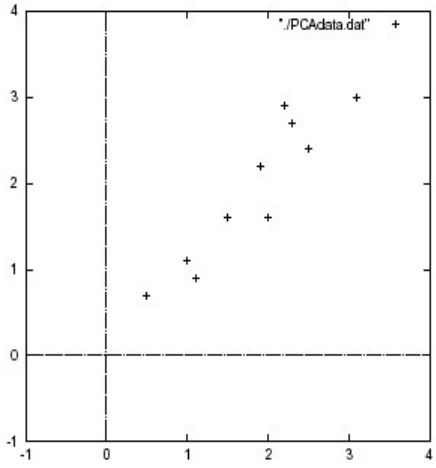
The first PC is the linear combination that captures the maximum variance in the data. The second PC is created by selecting another linear combination that max. variance with the constraint that its direction is perpendicular to the first component.

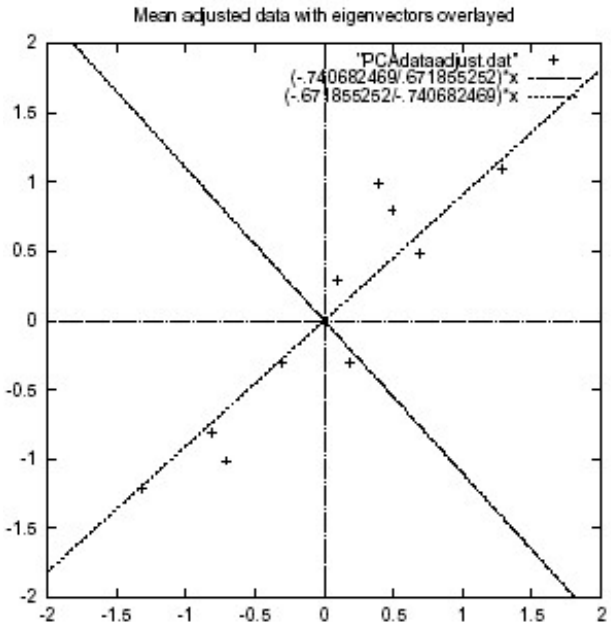
Data =

x	y
2.5	2.4
0.5	0.7
2.2	2.9
1.9	2.2
3.1	3.0
2.3	2.7
2	1.6
1	1.1
1.5	1.6
1.1	0.9

DataAdjust =

x	y
.69	.49
-1.31	-1.21
.39	.99
.09	.29
1.29	1.09
.49	.79
.19	-.31
-.81	-.81
-.31	-.31
-.71	-1.01



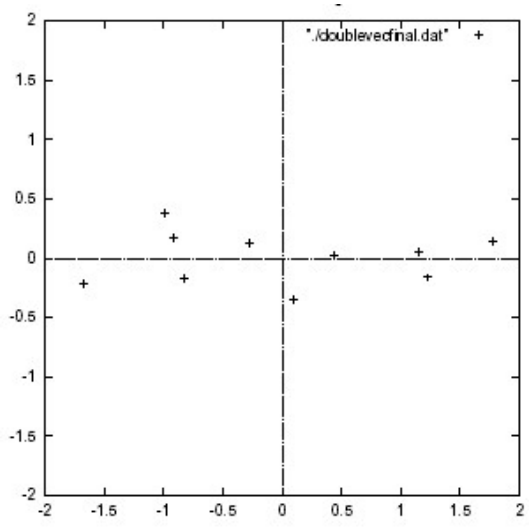


563

Transformed Data=

x	y
-.827970186	-.175115307
1.77758033	.142857227
-.992197494	.384374989
-.274210416	.130417207
-1.67580142	-.209498461
-.912949103	.175282444
.0991094375	-.349824698
1.14457216	.0464172582
.438046137	.0177646297
1.22382056	-.162675287

Data transformed with 2 eigenvectors



L- 15:Appled Multivariate Analytics & Revision

Agenda

Revision

Probability :-

Defn :- $P(A) = \frac{m}{n}$

ie $P(\bar{A}) = \frac{n-m}{n}$

$= 1 - \frac{m}{n}$

$= P(A)$

ie $\boxed{P(A) + P(\bar{A}) = 1}$

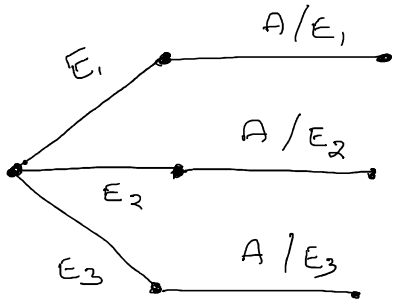
Conditional Probability

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

or

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

Baye's theorem



$$P(E_1/A) = \frac{P(A/E_i)\,P(E_i)}{\sum P(A/E_i)\,P(E_i)}$$

Random Variables

- Discrete — $P(x)$ → Distributions
  - Binomial: $P(x) = {}^nC_x\, p^x q^{n-x}$, $x = 0, 1, 2, \cdots n$
  - poisson: $P(x) = \frac{e^{-\lambda}\lambda^x}{x!}$, $x = 0, 1, 2 \cdots \infty$
- continuous — $f(x)$
  - Normal dist.: $f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{\frac{(x-\mu)^2}{2\sigma^2}}$, $x \in (-\infty, \infty)$

Normal distribution

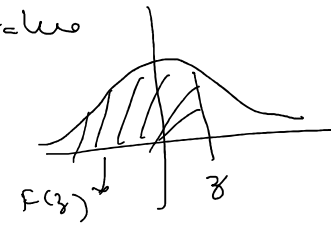
$P(x_1 \le x \le x_2)$ $\quad z = \frac{x-\mu}{\sigma}$

$\downarrow$

$P(z_1 \le z \le z_2)$

$= F(z_2) - F(z_1)$

$F(z)$ is tabulated value



Sampling

&

Estimation

$\downarrow$

mean → large

→ small

$z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$ $\quad t = \frac{\bar{x}-\mu}{s/\sqrt{n}}$

Testing of Hypothesis

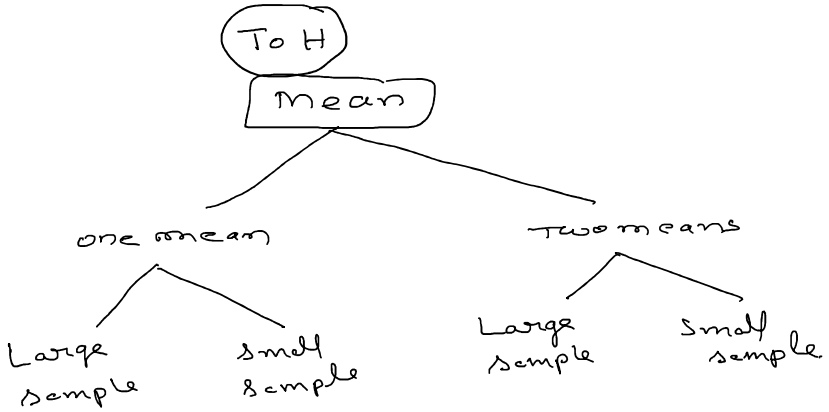
H_0: Null hypothesis

H_1: Alternate hypothesis

α: Level of significance

Critical region

Decision:



1) correlation

2) Regression

$y = \beta_0 + \beta_1 x$

$\Sigma y = \beta_0 n + \beta_1 \Sigma x$

$\Sigma xy = \beta_0 \Sigma x + \beta_1 \Sigma x^2$

R^2 = coeff of determination

$= 1 - \frac{RSS}{TSS}$

Lasso:

$$+ \lambda \sum |\beta_j|$$

Ridge:

$$+ \lambda \sum |\beta_j|^2$$

Logistic regression

$$\log\left(\frac{p}{1-p}\right) = y \quad \text{i.e.} \quad p = \frac{1}{1+e^{-y}}$$

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

or

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

Time series

↓

→ components

→ Methods

→ Moving averages

→ weighted moving avg

→ smoothing method - Expt

$$F_{t+1} = F_t + \alpha (A_t - F_t)$$

ANOVA.

↓

F - distribution : $\frac{S_1^2}{S_2^2}$ → $n_1 - 1$, $n_2 - 1$

$$F = \frac{MSTR}{MSE}$$

$$MSTR = \frac{SSTR}{n-1}$$

$$SSTR = \frac{(\Sigma x_j)^2}{n_j} - CF$$

$$SST = \left(\sum (x_i)^2 + \Sigma x_2^2 + \cdots \right) - CF$$

$$MSE = \frac{SSE}{n-1}$$

→ SSE : SST − SSTR

Principal comp. Analysis

→ when

→ why

→ How → wrt $\bar{x}$, $\bar{y}$

→ covariance matrix

→ eigen values & vector

→ V

→ $y = xV$.

Thanks