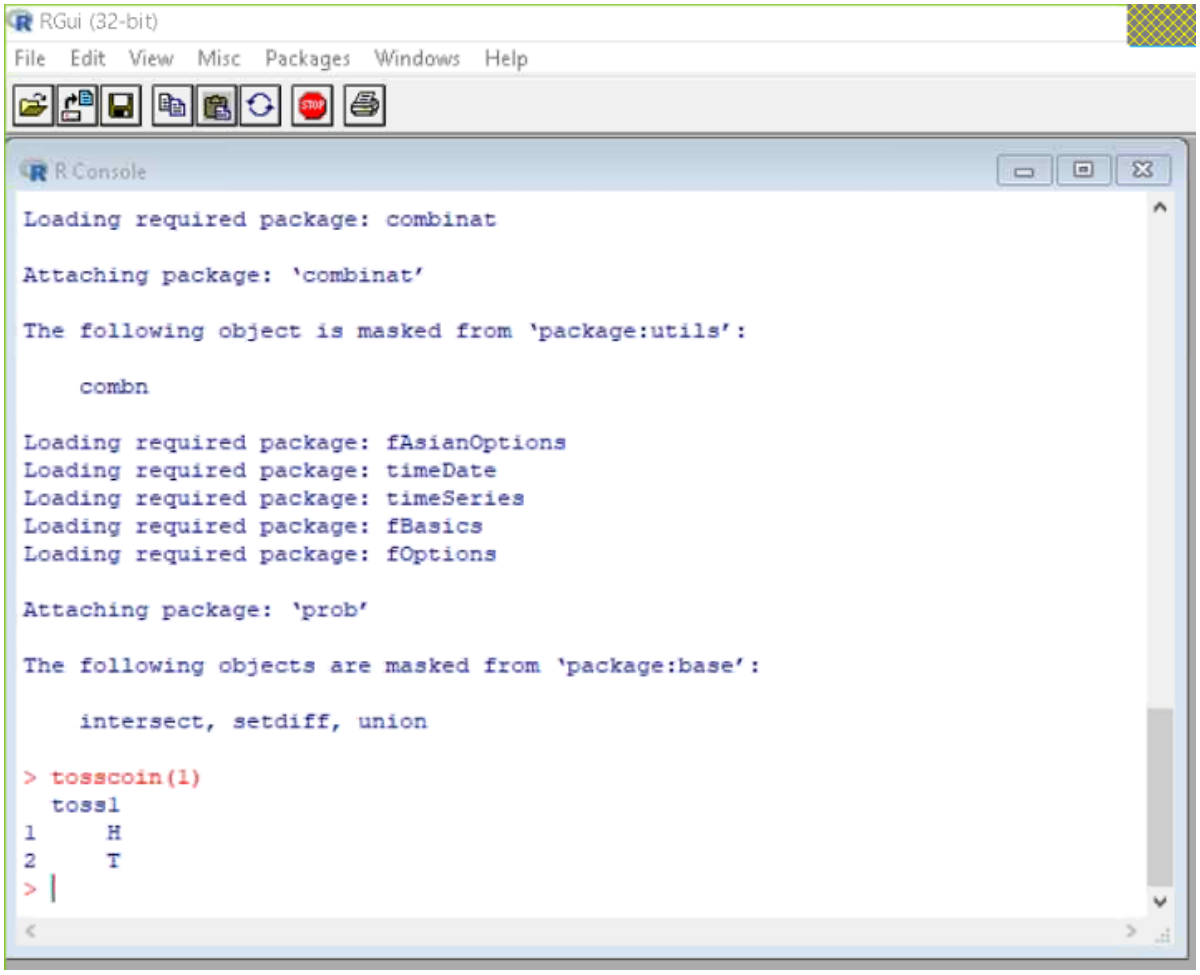
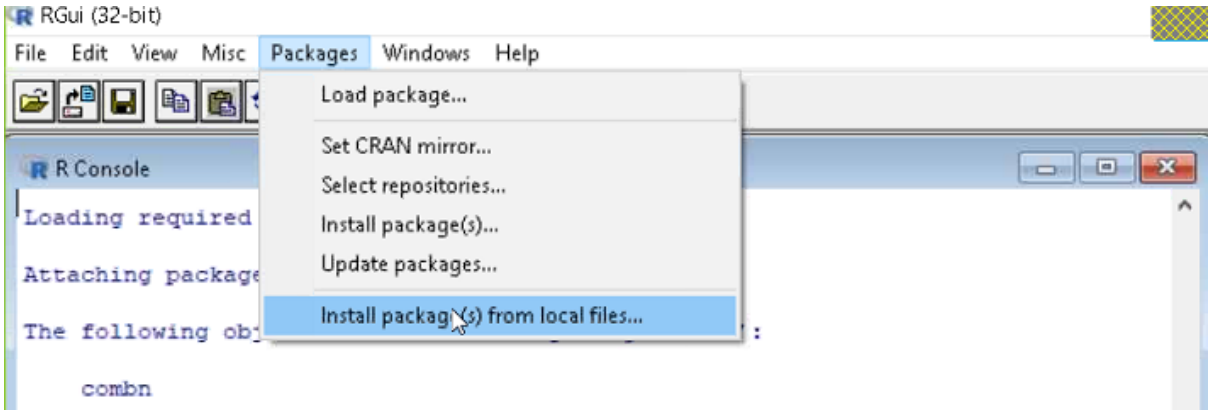
Lec 10 Language R Session – Implementation





2).Probability - Notepad

File Edit Format View Help

```
======================Sample Space===========

S <- data.frame(lands = c("down", "up", "side"))

S


install.packages("prob")   # installing prob package


library(prob)
tosscoin(1)

tosscoin(3)

rolldie(1)

rolldie(3, nsides = 4)      # 4 sided die 3 times

head(cards())

##Let our urn simply contain three balls, labeled 1, 2, and 3, respectively. We are
going to take a sample of size 2 from the urn.

1) Ordered,with Replacement

urnsamples(1:3, size = 2, replace = TRUE, ordered = TRUE)
```

```
Loading required package: timeSeries
Loading required package: fBasics
Loading required package: fOptions

Attaching package: 'prob'

The following objects are masked from 'package:base':

    intersect, setdiff, union

> tosscoin(1)
  toss1
1     H
2     T
> trees
   Girth Height Volume
1    8.3     70   10.3
2    8.6     65   10.3
3    8.8     63   10.2
4   10.5     72   16.4
5   10.7     81   18.8
6   10.8     83   19.7
7   11.0     66   15.6
8   11.0     75   18.2
9   11.1     80   22.6
```

```
> names(trees)
[1] "Girth"  "Height" "Volume"
>
```

1).Programming in R - Notepad

File Edit Format View Help

```
====================================Basics=====================

trees          # 2.to see the data set

names(trees)    # 3.names of the colums in the data set

trees[1:5,]    # 4.to see data sets from rows 1 to 5

trees[1:5,2]    # 5.data sets from rows  1 to 5 with colum 2

trees[1:5,"Volume"]   # 6.data sets from 1 to 5 wrt Volume

trees[,2]                    # 7.gives all the heights

trees[,3]                    #8. gives all Volume

str(trees)                  # 9.gives the structure of the data object rahter than stastical summary

summary(trees)        # 10,gives stasrical summary of data objects

attach(trees)           # 11. makes the contents of trees as a directory

====================================Basics=====================

mean(Height)         # 12. mean of Heights
mean(Girth)
mean(Volume)
mean(trees[,2])        #13.  gives mean of second column- height

apply(trees,2,mean)     # 14.  mean of each column(dimension of 2) of trees




================Explore Individual Variables============================


#Descriptive Statistics

par(mfrow=c(2,1))
```

```
8   11.0     75   18.2
9   11.1     80   22.6
10  11.2     75   19.9
11  11.3     79   24.2
12  11.4     76   21.0
13  11.4     76   21.4
14  11.7     69   21.3
15  12.0     75   19.1
16  12.9     74   22.2
17  12.9     85   33.8
18  13.3     86   27.4
19  13.7     71   25.7
20  13.8     64   24.9
21  14.0     78   34.5
22  14.2     80   31.7
23  14.5     74   36.3
24  16.0     72   38.3
25  16.3     77   42.6
26  17.3     81   55.4
27  17.5     82   55.7
28  17.9     80   58.3
29  18.0     80   51.5
30  18.0     80   51.0
31  20.6     87   77.0
>
```

Name columns – visualization – any relation between two columns – titles

```
> trees[1:5,]
  Girth Height Volume
1   8.3     70   10.3
2   8.6     65   10.3
3   8.8     63   10.2
4  10.5     72   16.4
5  10.7     81   18.8
>
```

```
> trees[1:5,2]
[1] 70 65 63 72 81
```

```
> summary(trees)
     Girth           Height       Volume
 Min.   : 8.30   Min.   :63   Min.   :10.20
 1st Qu.:11.05   1st Qu.:72   1st Qu.:19.40
 Median :12.90   Median :76   Median :24.20
 Mean   :13.25   Mean   :76   Mean   :30.17
 3rd Qu.:15.25   3rd Qu.:80   3rd Qu.:37.30
 Max.   :20.60   Max.   :87   Max.   :77.00
```

With respect with particular column – how these values are distributed – summary – instead of seeing whole set of data

```
> mean(Height)
Error in mean(Height) : object 'Height' not found
> attach(trees)
> mean(Height)
[1] 76
> mean(Volume)
[1] 30.17097
```

Set the directory & attach the file – then start analysis

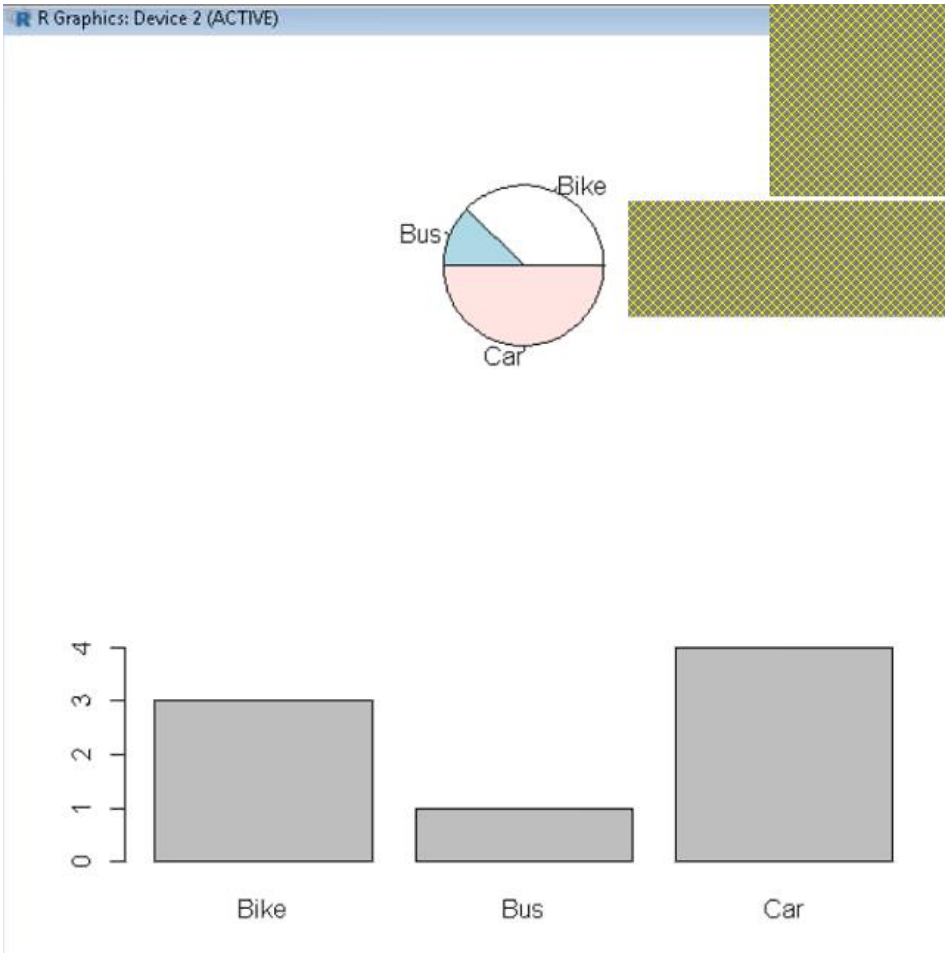
```
R version 3.5.1 (2018-07-02) -- "Feather Spray"
Copyright (C) 2018 The R Foundation for Statistical Computing
Platform: i386-w64-mingw32/i386 (32-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[Previously saved workspace restored]

> par(mfrow=c(2,1))
```

Window creation for visualization

```
> par(mfrow=c(2,1))
> Transportation=c("Car", "Bus", "Car", "Bike", "Car", "Bike", "Bike", "Car")
> Transportation
[1] "Car"  "Bus"  "Car"  "Bike" "Car"  "Bike" "Bike" "Car"
```

```
> table(Transportation)
Transportation
Bike  Bus  Car
   3    1    4
> pie(table(Transportation))
```

```
> pie(table(Transportation))
> barplot(table(Transportation))
```

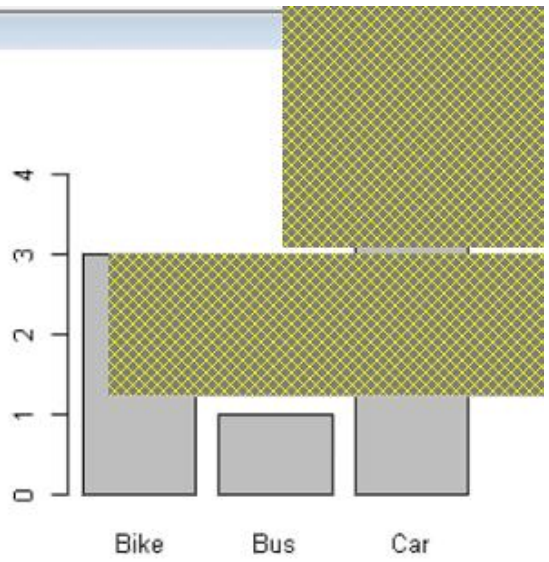


2 X 1 window created

```
> par(mfrow=c(2,2))
>  pie(table(Transportation))
> barplot(table(Transportation))
```

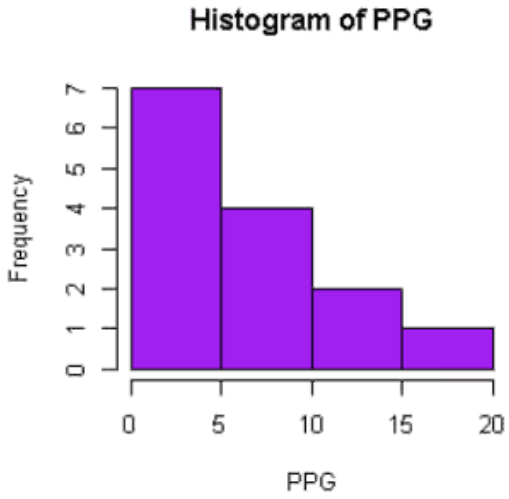
2 X 2 window created

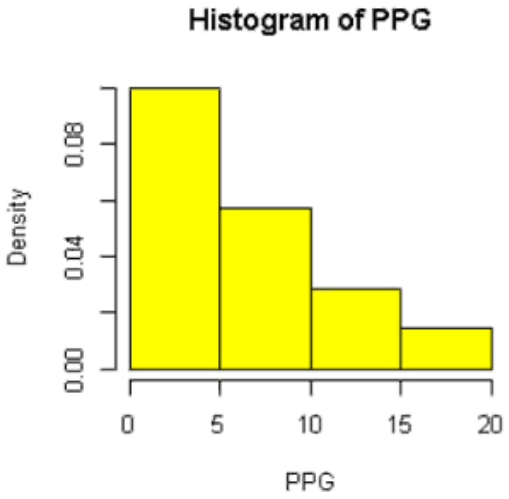
R Graphics: Device 2 (ACTIVE)




```
> PPG = c(15.6, 13.4, 11.5, 8.3, 7.5, 6.6, 5.2, 2.8, 2.4, 1.9, 1.1, 0.0, 0.0, 0$
> PPG
 [1] 15.6 13.4 11.5  8.3  7.5  6.6  5.2  2.8  2.4  1.9  1.1  0.0  0.0  0.0
> hist(PPG,col="purple")
```

```
> hist(PPG,col="yellow",prob=TRUE)
>
```





```
==========Explore Individual Variables==========


#Descriptive Statistics

par(mfrow=c(2,1))

Transportation=c("Car", "Bus", "Car", "Bike", "Car", "Bike", "Bike", "Car")
table(Transportation)
pie(table(Transportation))


Transportation=c("Car", "Bus", "Car", "Bike", "Car", "Bike", "Bike", "Car")
barplot(table(Transportation))

PPG = c(15.6, 13.4, 11.5, 8.3, 7.5, 6.6, 5.2, 2.8, 2.4, 1.9, 1.1, 0.0, 0.0, 0.0)
hist(PPG,col="purple")

hist(PPG,col="yellow",prob=TRUE)

mybreaks=seq(-0.2,15.8,4)
hist(PPG,breaks=mybreaks, xlab="Average point per game for U of A, Men's basketball team", main="points per game",col="green")




X <- c(56, 31,56,8,32)
X.bar  <- mean(X)
AD  <- abs(X - X.bar)
AAD  <- mean(AD)
print(AAD)
```

```
> hist(PPG,col="purple")
> hist(PPG,col="yellow",prob=TRUE)
> S <- data.frame(lands = c("down", "up", "side"))
> S
  lands
1  down
2    up
3  side
> install.packages("prob")
Installing package into 'C:/Users/YVK/Documents/R/win-library/3.5'
(as 'lib' is unspecified)
--- Please select a CRAN mirror for use in this session ---
trying URL 'https://cloud.r-project.org/bin/windows/contrib/3.5/prob_1.0-1.zip'
Content type 'application/zip' length 756052 bytes (738 KB)
downloaded 738 KB

package 'prob' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\YVK\AppData\Local\Temp\RtmpS6R9Js\downloaded_packages
```

2).Probability - Notepad

File Edit Format View Help

```
======================Sample Space===========

S <- data.frame(lands = c("down", "up", "side"))

S


install.packages("prob")   # installing prob package


library(prob)
tosscoin(1)

tosscoin(3)

rolldie(1)

rolldie(3, nsides = 4)      # 4 sided die 3 times

head(cards())

##Let our urn simply contain three balls, labeled 1, 2, and 3, respectively. We are
going to take a sample of size 2 from the urn.

1) Ordered,with Replacement

urnsamples(1:3, size = 2, replace = TRUE, ordered = TRUE)

2) Ordered without Replacement

urnsamples(1:3, size = 2, replace = FALSE, ordered = TRUE)

3) Unordered without Replacement

urnsamples(1:3, size = 2, replace = FALSE, ordered = FALSE)

4)Unordered, With Replacement

urnsamples(1:3, size = 2, replace = TRUE, ordered = FALSE)

======================Events===========
```

```
> tosscoin(1)
  toss1
1     H
2     T
> tosscoin(2)
  toss1 toss2
1     H     H
2     T     H
3     H     T
4     T     T
```

```
> library(prob)
Loading required package: combinat

Attaching package: 'combinat'

The following object is masked from 'package:utils':

    combn

Loading required package: fAsianOptions
Loading required package: timeDate
Loading required package: timeSeries
Loading required package: fBasics
Loading required package: fOptions

Attaching package: 'prob'

The following objects are masked from 'package:base':

    intersect, setdiff, union
```

```
> rolldie(1)
  X1
1  1
2  2
3  3
4  4
5  5
6  6
```

```
> rolldie(3, nsides = 4)
   X1 X2 X3
1   1  1  1
2   2  1  1
3   3  1  1
4   4  1  1
5   1  2  1
6   2  2  1
7   3  2  1
8   4  2  1
9   1  3  1
10  2  3  1
11  3  3  1
12  4  3  1
13  1  4  1
14  2  4  1
15  3  4  1
16  4  4  1
17  1  1  2
18  2  1  2
19  3  1  2
20  4  1  2
21  1  2  2
```

```
======================Events===========

S <- tosscoin(2)

S <- tosscoin(3)

S[1:3, ]

S <- cards()

rolldie(3)

subset(rolldie(3), X1 + X2 + X3 > 16)


################## Set Union,Intersection and Difference#############

S = cards()

A = subset(S, suit == "Heart")

B = subset(S, rank %in% 7:9)

union(A, B)

intersect(A, B)

setdiff(A, B)

setdiff(B, A)

setdiff(S,A)
```

```
> S <- tosscoin(2)
> S <- tosscoin(3)
> S[1:3, ]
  toss1 toss2 toss3
1     H     H     H
2     T     H     H
3     H     T     H
```

```
> S = cards()
> A = subset(S, suit == "Heart")
> A
   rank  suit
27    2 Heart
28    3 Heart
29    4 Heart
30    5 Heart
31    6 Heart
32    7 Heart
33    8 Heart
34    9 Heart
35   10 Heart
36    J Heart
37    Q Heart
38    K Heart
39    A Heart
```

```
> B = subset(S, rank %in% 7:9)
> B
   rank    suit
6     7    Club
7     8    Club
8     9    Club
19    7 Diamond
20    8 Diamond
21    9 Diamond
32    7   Heart
33    8   Heart
34    9   Heart
45    7   Spade
46    8   Spade
47    9   Spade
```

```
> intersect(A, B)
   rank  suit
32    7 Heart
33    8 Heart
34    9 Heart
```

```
> prob(A)
'prob' is deprecated; use 'Prob' instead.
'space' is missing a probs column
Error in Prob.default(x, ...) : see ?probspace
> Prob(A)
'space' is missing a probs column
Error in Prob.default(A) : see ?probspace
```

```
> union(A, B)
   rank    suit
6     7    Club
7     8    Club
8     9    Club
19    7 Diamond
20    8 Diamond
21    9 Diamond
27    2   Heart
28    3   Heart
29    4   Heart
30    5   Heart
31    6   Heart
32    7   Heart
33    8   Heart
34    9   Heart
35   10   Heart
36    J   Heart
37    Q   Heart
38    K   Heart
39    A   Heart
45    7   Spade
46    8   Spade
47    9   Spade
```

```
> S <- cards(makespace = TRUE)
> A <- subset(S, suit == "Heart")
> A
   rank  suit      probs
27    2 Heart 0.01923077
28    3 Heart 0.01923077
29    4 Heart 0.01923077
30    5 Heart 0.01923077
31    6 Heart 0.01923077
32    7 Heart 0.01923077
33    8 Heart 0.01923077
34    9 Heart 0.01923077
35   10 Heart 0.01923077
36    J Heart 0.01923077
37    Q Heart 0.01923077
38    K Heart 0.01923077
39    A Heart 0.01923077
```

```
> prob(A)
'prob' is deprecated; use 'Prob' instead.
[1] 0.25
```

```
> S
   rank     suit      probs
1     2     Club 0.01923077
2     3     Club 0.01923077
3     4     Club 0.01923077
4     5     Club 0.01923077
5     6     Club 0.01923077
6     7     Club 0.01923077
7     8     Club 0.01923077
8     9     Club 0.01923077
9    10     Club 0.01923077
10    J     Club 0.01923077
11    Q     Club 0.01923077
12    K     Club 0.01923077
13    A     Club 0.01923077
14    2  Diamond 0.01923077
15    3  Diamond 0.01923077
16    4  Diamond 0.01923077
17    5  Diamond 0.01923077
18    6  Diamond 0.01923077
```

```
> S <- rolldie(2, makespace = TRUE)
> S
   X1 X2      probs
1   1  1 0.02777778
2   2  1 0.02777778
3   3  1 0.02777778
4   4  1 0.02777778
5   5  1 0.02777778
6   6  1 0.02777778
7   1  2 0.02777778
8   2  2 0.02777778
9   3  2 0.02777778
10  4  2 0.02777778
11  5  2 0.02777778
```

```
> A <- subset(S, X1 == X2)
> A
   X1 X2      probs
1   1  1 0.02777778
8   2  2 0.02777778
15  3  3 0.02777778
22  4  4 0.02777778
29  5  5 0.02777778
36  6  6 0.02777778
```

```
> B <- subset(S, X1 + X2 >= 8)
> B
   X1 X2      probs
12  6  2 0.02777778
17  5  3 0.02777778
18  6  3 0.02777778
22  4  4 0.02777778
23  5  4 0.02777778
24  6  4 0.02777778
27  3  5 0.02777778
28  4  5 0.02777778
29  5  5 0.02777778
30  6  5 0.02777778
32  2  6 0.02777778
33  3  6 0.02777778
34  4  6 0.02777778
35  5  6 0.02777778
36  6  6 0.02777778
```

```
======================Probability ===========

 S <- cards(makespace = TRUE)
A <- subset(S, suit == "Heart")
prob(A)


======================Conditional Probability ===========

problem solving

library(prob)
S <- rolldie(2, makespace = TRUE)
A <- subset(S, X1 == X2)
B <- subset(S, X1 + X2 >= 8)

prob(A)
prob(B)
prob(A, given = B)
prob(B, given = A)
```

```
> prob(A)
'prob' is deprecated; use 'Prob' instead.
[1] 0.1666667
> prob(B)
'prob' is deprecated; use 'Prob' instead.
[1] 0.4166667
> prob(A, given = B)
'prob' is deprecated; use 'Prob' instead.
[1] 0.2
> prob(B, given = A)
'prob' is deprecated; use 'Prob' instead.
[1] 0.5
```

Event wise both are same but probability wise both are different

3).Bayes Theorem - Notepad

File Edit Format View Help

```
======================Baye's Theorem ===========


install.packages("prob")   # installing prob package
library(prob)

## Example 1##
prior <- c(0.8,0.2)
like <- c(0.7,0.2)
post <- prior * like
post/sum(post)

## Example 2###
prior <- c(0.2, 0.6, 0.15,0.05)
like <- c(0.05, 0.10, 0.1,0.05)
post <- prior * like
post/sum(post)

======================Random Variables ===========
S <- rolldie(3, nsides = 4, makespace = TRUE)
S <- addrv(S, U = X1 - X2 + X3)
prob(S, U > 6)

### V = max(X1; X2; X3) and W = X1 + X2 + X3.
S <- addrv(S, FUN = max, invars = c("X1", "X2", "X3"), name = "V")
S <- addrv(S, FUN = sum, invars = c("X1", "X2", "X3"), name = "W")


====================== Probability Distributions ===========
x <- c(0,1,2,3)
f <- c(1/8, 3/8, 3/8, 1/8)
mu <- sum(x * f)
sigma2 <- sum((x-mu)^2 * f)
sigma2
sigma <- sqrt(sigma2)

library(distrEx)
> X <- DiscreteDistribution(supp = 0:3, prob = c(1,3,3,1)/8)
> E(X); var(X); sd(X)

#### Expectation of a r.v
```

```
> prior <- c(0.8,0.2)
> like <- c(0.7,0.2)
> post <- prior * like
> post/sum(post)
[1] 0.93333333 0.06666667
```

```
> S <- rolldie(3, nsides = 4, makespace = TRUE)
>
```

```
55  3  2  4 0.015625
56  4  2  4 0.015625
57  1  3  4 0.015625
58  2  3  4 0.015625
59  3  3  4 0.015625
60  4  3  4 0.015625
61  1  4  4 0.015625
62  2  4  4 0.015625
63  3  4  4 0.015625
64  4  4  4 0.015625
```

```
> S <- addrv(S, U = X1 - X2 + X3)
```

```
54  2  2  4  4 0.015625
55  3  2  4  5 0.015625
56  4  2  4  6 0.015625
57  1  3  4  2 0.015625
58  2  3  4  3 0.015625
59  3  3  4  4 0.015625
60  4  3  4  5 0.015625
61  1  4  4  1 0.015625
62  2  4  4  2 0.015625
63  3  4  4  3 0.015625
64  4  4  4  4 0.015625
```

```
> prob(S, U > 6)
'prob' is deprecated; use 'Prob' instead.
[1] 0.015625
>
```

```
====================== Probability Distributions ===========
###### Binomial Distribution ##########

pbinom(9, size=12, prob=1/6) - pbinom(6, size=12, prob=1/6)

diff(pbinom(c(6,9), size = 12, prob = 1/6))

dbinom(17, size = 31, prob = 0.447)
pbinom(13, size = 31, prob = 0.447)

sum(dbinom(16:19, size = 31, prob = 0.447))
diff(pbinom(c(19, 15), size = 31, prob = 0.447, lower.tail = FALSE))

library(distrEx)
X = Binom(size = 31, prob = 0.447)
E(X)
var(X)
```

pbinom(c(6,9)) → P(6) < x < P(9) → Cumulative

```
> pbinom(9, size=12, prob=1/6) - pbinom(6, size=12, prob=1/6)
[1] 0.001291758
>  pbinom(9, size=12, prob=1/6)
[1] 0.9999992
```

```
### Poisson Distribution########

ppois(0,lambda = 5)

diff(ppois(c(47, 50), lambda = 50))



## crvn####

# method 1

f <- function(x) 3 * x^2
integrate(f, lower = 0.14, upper = 0.71)
```

```
> ppois(0,lambda = 5)
[1] 0.006737947
> diff(ppois(c(47, 50), lambda = 50))
[1] 0.1678485
>
```

Binomial Distribution - n is very large & P tends to 0 → mean == variance → Binomial distribution - towards Poisson distribution

6) linearregression - Notepad

File Edit Format View Help

```
......data exploration.........
data(anscombe)

attach(anscombe)

cor(x1,y1)

cor(x2,y1)

par(mfrow = c(2,2))

plot(x1,y1,main = "plot 1")
plot(x2,y2,main = "plot 2")
plot(x3,y3,main = "plot 3")
plot(x4,y4,main = "plot 4")
```

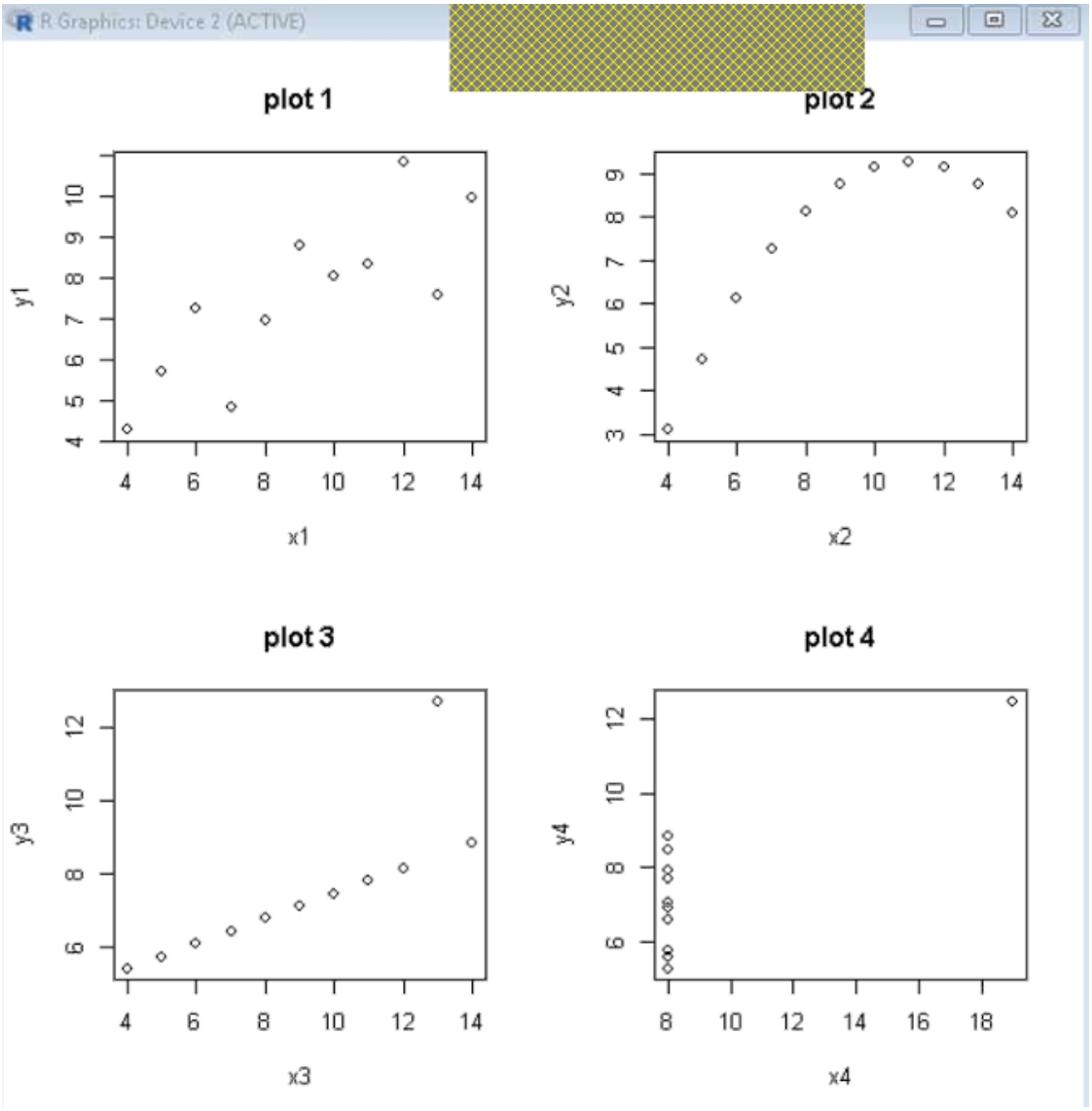
```
> data(anscombe)
> anscombe
   x1 x2 x3 x4    y1   y2    y3    y4
1  10 10 10  8  8.04 9.14  7.46  6.58
2   8  8  8  8  6.95 8.14  6.77  5.76
3  13 13 13  8  7.58 8.74 12.74  7.71
4   9  9  9  8  8.81 8.77  7.11  8.84
5  11 11 11  8  8.33 9.26  7.81  8.47
6  14 14 14  8  9.96 8.10  8.84  7.04
7   6  6  6  8  7.24 6.13  6.08  5.25
8   4  4  4 19  4.26 3.10  5.39 12.50
9  12 12 12  8 10.84 9.13  8.15  5.56
10  7  7  7  8  4.82 7.26  6.42  7.91
11  5  5  5  8  5.68 4.74  5.73  6.89
```

```
> attach(anscombe)
> cor(x1,y1)
[1] 0.8164205
> cor(x2,y1)
[1] 0.8164205
```

Find correlation coefficient – to understand pattern – what are the variables correlated – distributed – based on that which kind of regression

```
> par(mfrow = c(2,2))
> plot(x1,y1,main = "plot 1")
Error in plot(x1, y1, main = "plot 1") : object 'x1' not found
> data(anscombe)
> attach(anscombe)
> plot(x1,y1,main = "plot 1")
> plot(x2,y2,main = "plot 2")
> plot(x3,y3,main = "plot 3")
> plot(x4,y4,main = "plot 4")
```

```
.........Linear regression.............

install.packages("alr3")
library(alr3)
data(snake)
dim(snake)
head(snake)

names(snake) <-c("content","yield")
attach(snake)
head(snake)

plot(content , yield)
plot(content , yield,xlab = "water content of the snow", ylab ="water yield")

yield.fit <- lm(yield ~ content)
summary(yield.fit)
abline(yield.fit,lwd=3,col="red")

par(mfrow = c(2,2)
plot(yield.fit)
```

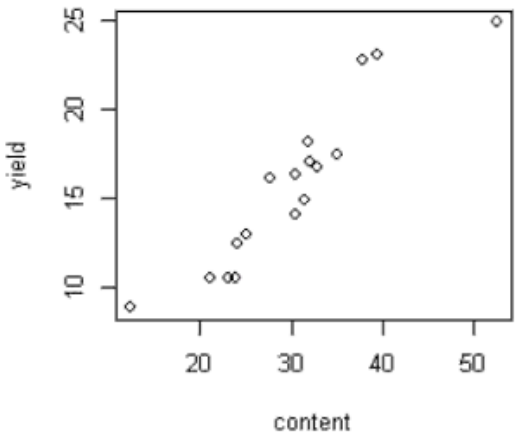


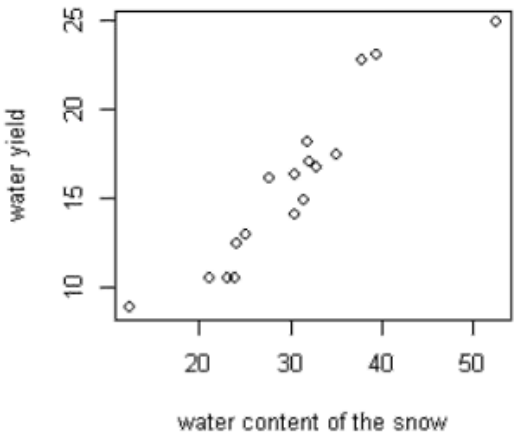
Plot 1 – Linear regression, Plot 2 – Curvilinear, Plot 3 – Linear, Plot 4 – no useful pattern – remove out layers – this is where arts & science meets – out layer is very important which affects pattern – removing out layer will not give exact picture

Plot 2 – Non Linear pattern – curve kind of pattern – compulsory to fit curve – freedom of application – difficulty when work with no linear data – complexity increases – like medical application – can't take risk of considering single error

```
> install.packages("alr3")
Installing package into 'C:/Users/YVK/Documents/R/win-library/3.5'
(as 'lib' is unspecified)
--- Please select a CRAN mirror for use in this session ---
trying URL 'https://cloud.r-project.org/bin/windows/contrib/3.5/alr3_2.0.8.zip'
Content type 'application/zip' length 616181 bytes (601 KB)
downloaded 601 KB

package 'alr3' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\YVK\AppData\Local\Temp\RtmpgVsxST\downloaded_packages
> library(alr3)
Loading required package: car
Loading required package: carData

Attaching package: 'alr3'

The following object is masked _by_ '.GlobalEnv':

    snake
```

```
> data(snake)
> dim(snake)
[1] 17  2
> head(snake)
     X    Y
1 23.1 10.5
2 32.8 16.7
3 31.8 18.2
4 32.0 17.0
5 30.4 16.3
6 24.0 10.5
> names(snake) <-c("content","yield")
> attach(snake)
> head(snake)
  content yield
1    23.1  10.5
2    32.8  16.7
3    31.8  18.2
4    32.0  17.0
5    30.4  16.3
6    24.0  10.5
> plot(content , yield)
> plot(content , yield,xlab = "water content of the snow", ylab ="water yield")
>
```

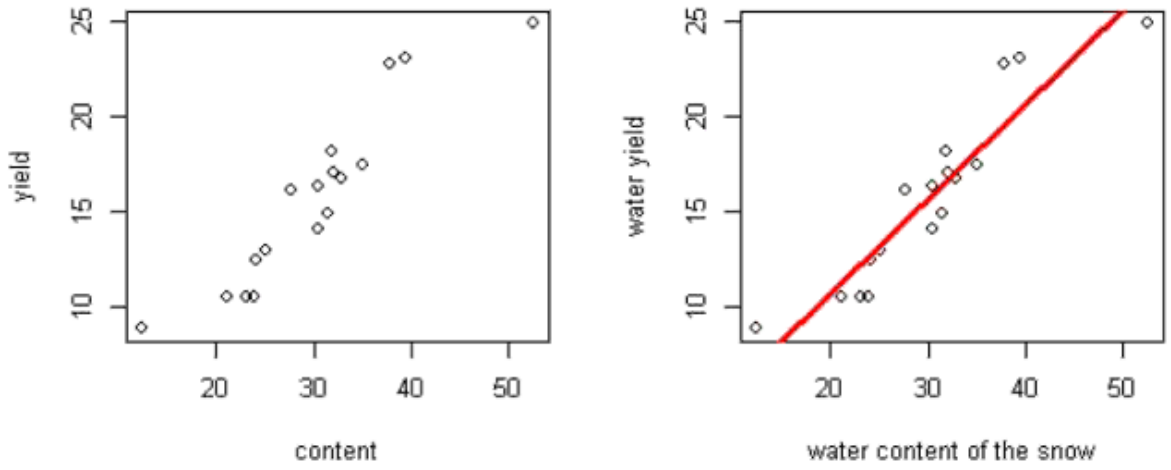




```
> yield.fit <- lm(yield ~ content)
> summary(yield.fit)

Call:
lm(formula = yield ~ content)

Residuals:
    Min      1Q  Median      3Q     Max
-2.1793 -1.5149 -0.3624  1.6276  3.1973

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.72538    1.54882   0.468    0.646
content      0.49808    0.04952  10.058 4.63e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.743 on 15 degrees of freedom
Multiple R-squared:  0.8709,    Adjusted R-squared:  0.8623
F-statistic: 101.2 on 1 and 15 DF,  p-value: 4.632e-08
```

$y = \beta_0$ (Intercept) + β_1 * x (content) → as if sum of errors is minimum

```
abline(yield.fit,lwd=3,col="red")
```



$y = \beta_0 + \beta_1 * x$ → such that most of the points are nearer to line

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

Agenda

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

How to validate the model – which parameters to consider, Assumptions relations to linear regression, Classification of regression – Logistic

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

Method of least square errors – linear regression or straight line – Beta parameters

Multiple regression

Numeric Response variable (y)

p Numeric predictor variables

Model:

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

- 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)

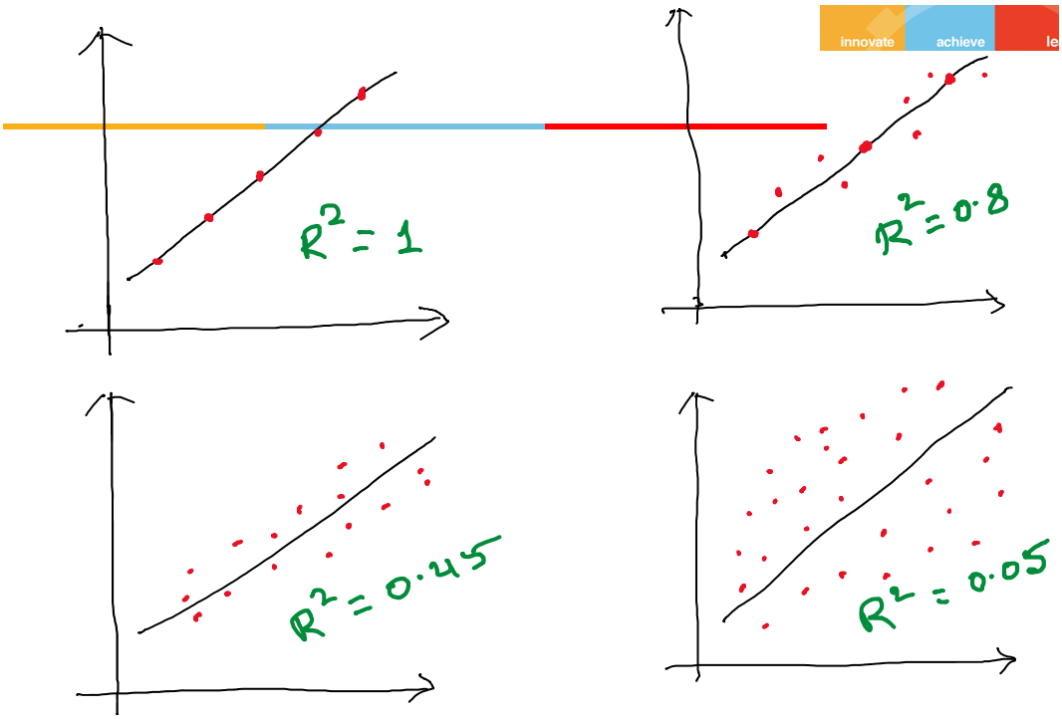
Two steps – set of data points – build a model → validate model with rest of the data points

RSS → Residual sum of squares

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

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

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



R^2 → more – model we built is more appropriate – data points lying on line of regression

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

R^2 using that we can validate the model

Multiple regression –data with many independent variables **adjusted R^2** is helpful – not sure of affect or relation between that independent variable & predicting variable (dependent)

Y → X1, X2, X3, X4, X5 → multiple regression model – Beta coefficient value - R^2 & adjusted R^2 - R^2 value → whenever add variable – simply R^2 value increases – not checking whether this variable is related to model or not - income of person – height of a person (not related)

Implement a model – look into this point – conclusion

Test data – overfitting a model → solution Regularization – constraints on coefficient – to avoid overfitting data

LASSO – we can drop some coefficients, RIDGE – can become 0 – we can't drop them

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

innovate achieve

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

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

Lambda – penalty coefficient – Lambda = 0 – simple OLS estimation – control with help of LASSO or RIDGE estimation – OLS – ordinary least squares – minimize SSE for best fit

Lambda – choose to control over fitting points → OLS – we don't have any control over confidents – LASSO & RIDGE – control the model by varying values of Lambda

Assumptions in Regression Analysis

Assumptions

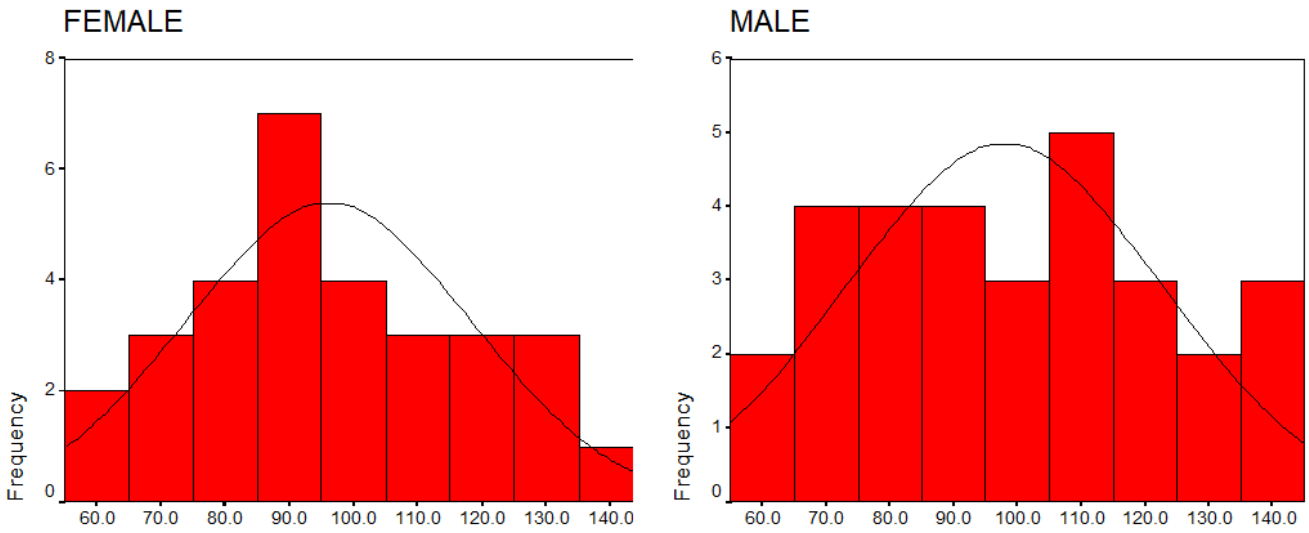
innovate achieve le

- 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

Distribution is normal - Residuals & plot it is – almost like normal distribution

- 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



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

Negative Skew – square – remove skewness so we can see normal distributed residuals

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

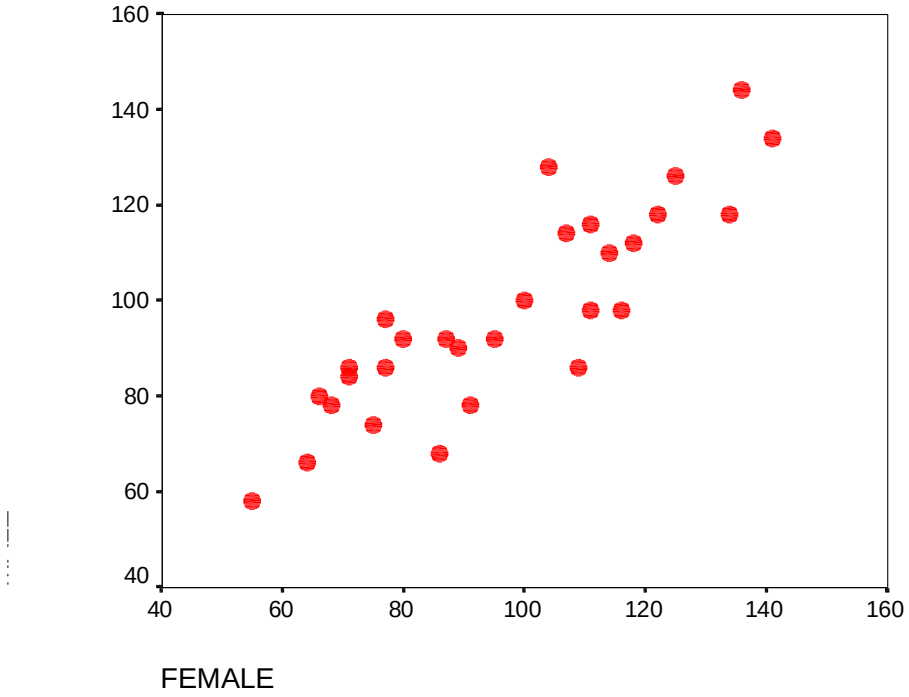
- Hetero=different
- Scedastic = scattered

We don't want heteroscedasticity

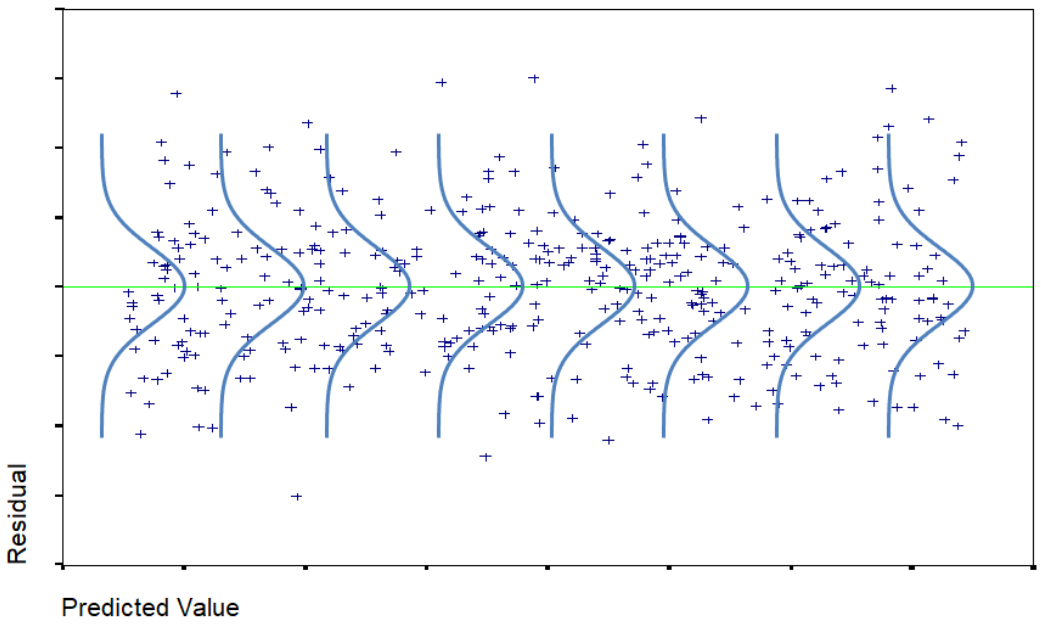
- we want our data to be homoscedastic

Draw a scatterplot to investigate

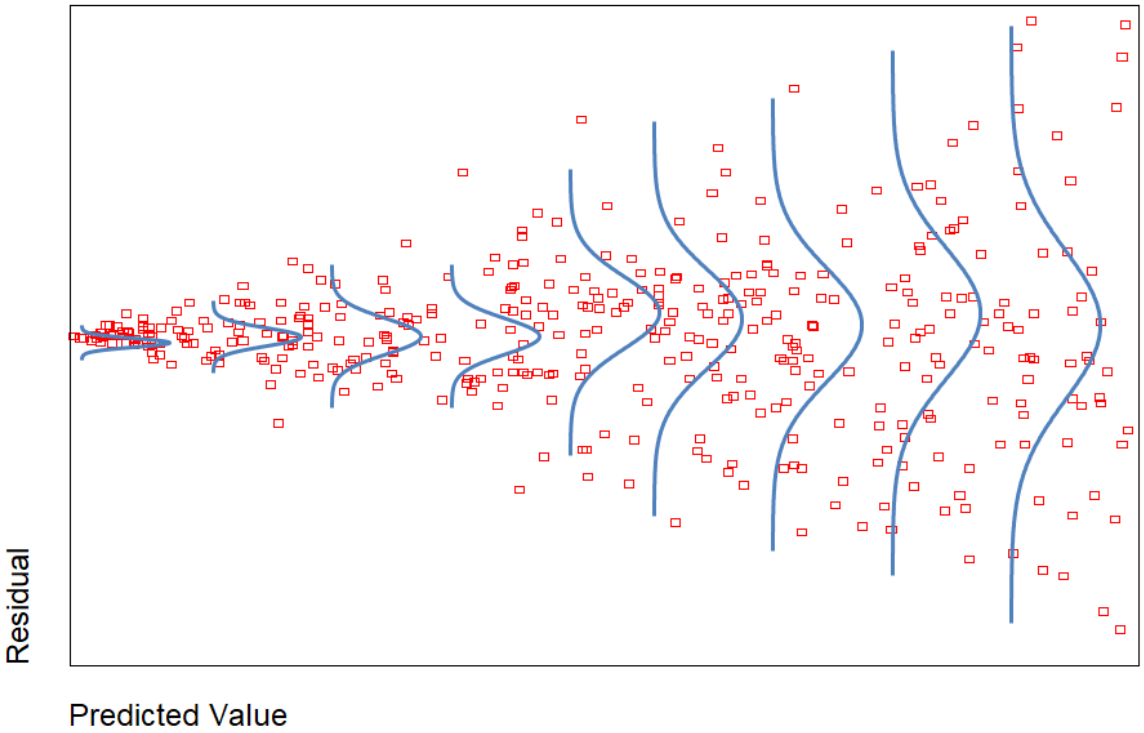
Heteroscedasticity or Homoscedasticity – using help of scattered plot



Good – no heteroscedasticity



Bad – heteroscedasticity



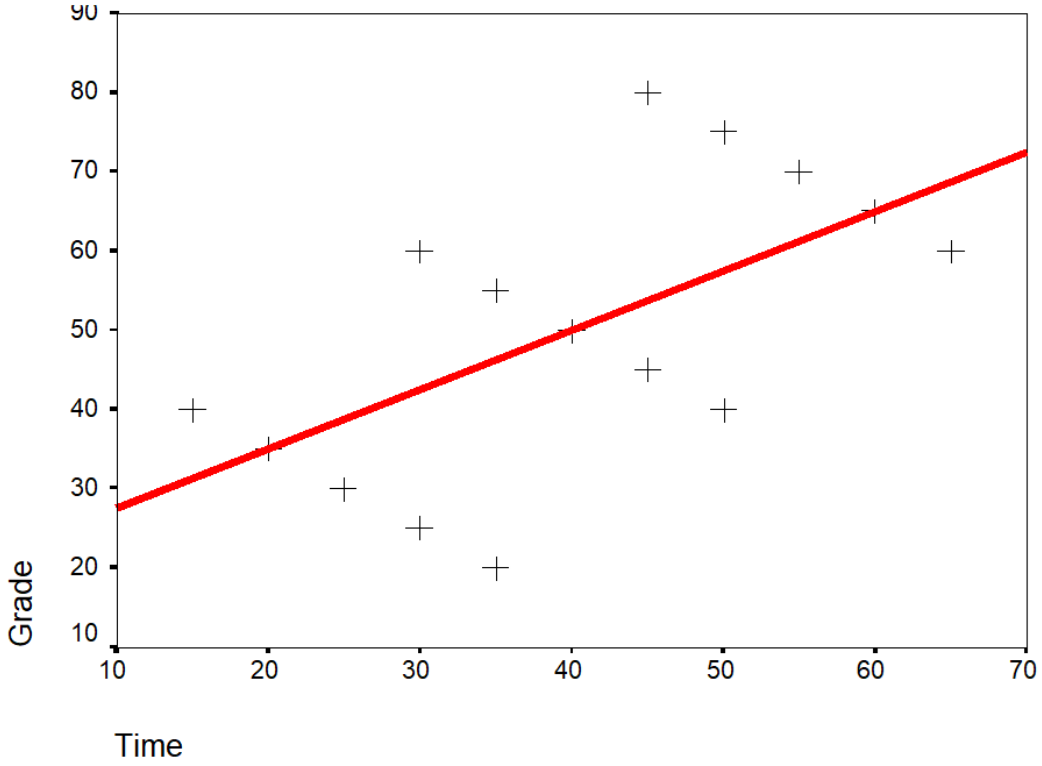
Assumption 3:
The Error Term is Additive

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

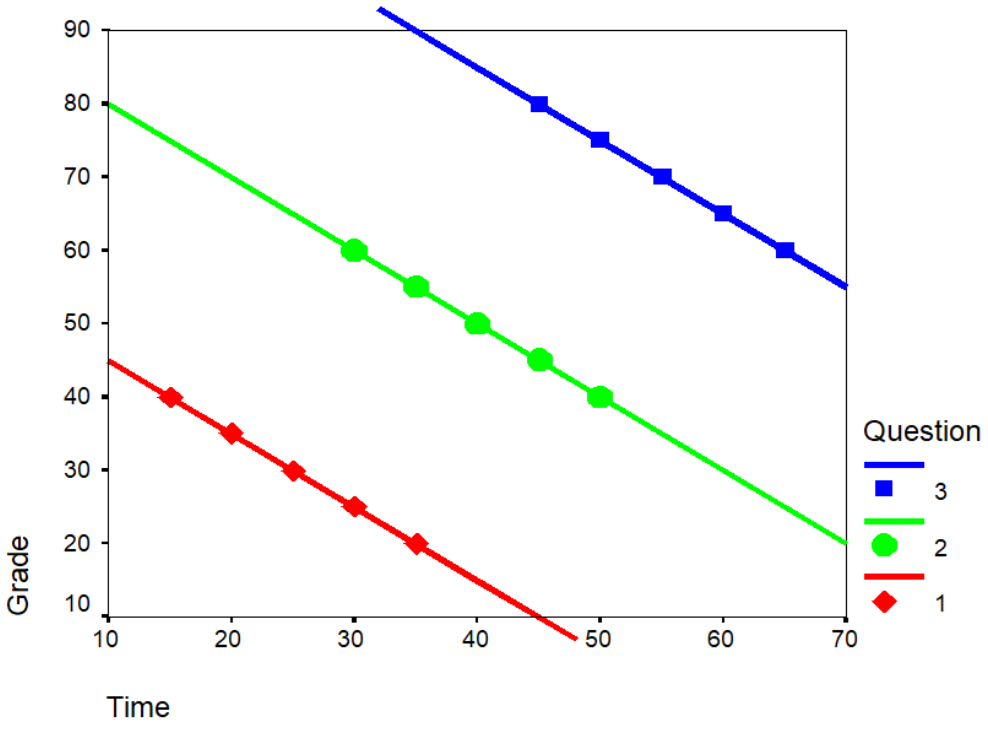
Link with normal distribution – every dependent variable it follows standard normal distribution → mean of the standard normal distribution is 0

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

- Result, with line of best fit



Now somewhat different



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

y – Dependent variable, x_1, x_2, x_3 – independent variables – find which variables are more collinear – if some of the variable is not collinear with other variable – then we can remove

x_1, x_3 → more positive correlation (high correlation), x_1, x_2 → more negative correlation – helps us to understand the model – build a model & eliminate variable which is not related

LASSO & RIDGE – regular regression – OLS – model can be overfitting model – overfitting can be controlled – control over coefficient – Lambda parameter – LASSO & RIDGE - control over model & avoid overfitting

Independent variable – correlation coefficient = 0 → change in one variable – doesn't affect

Build a model – hike in salary – performance of company, performance of individuals, login time → hike & login time – is there any relation? – Performance of the team – sometimes multicollinearity which also affects

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?

innovate achieve lead

- 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)

Regression – build a model & estimate value of some unknown

Classify data – positive or negative, Email – spam or not spam, Doctor Visit – Diagnosis results – Diabetic or not – reports – parameters → conclusion – expected – we don't expect y value from x_1, x_2, x_3 – we want conclusion – whether – Diabetic or not – decision

Given the customer & bank scenario –going to be defaulter? Approve a loan? Binomial – Y/N

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

CIBIL Score, Credit score – whether he will become defaulter on – classify

After certain time – performance prediction → machine will fail or not etc.

Categorical Response Variables

Examples:

Whether or not a person smokes — Binary Response

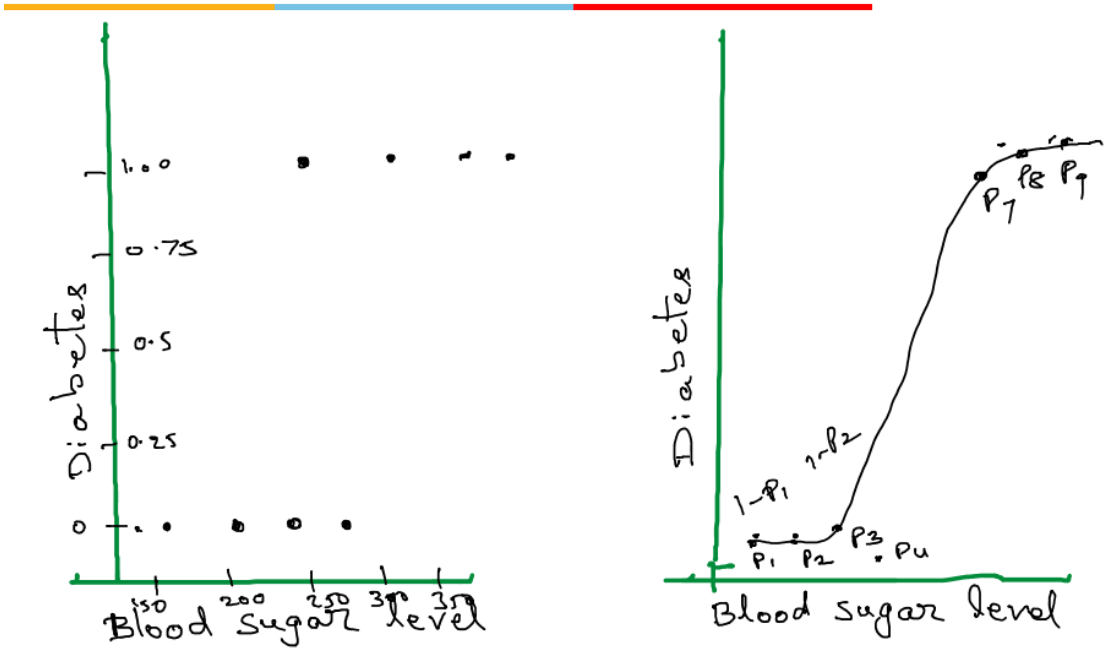
$$Y = \begin{cases} \text{Non} - \text{smoker} \\ \text{Smoker} \end{cases}$$

Success of a medical treatment — Binary Response

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

Opinion poll responses — Ordinal Response

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



Probability of odds – happening → p, not happening → 1 – p

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

$$\text{Likelihood} = (1-p_1)(1-p_2)(1-p_3)(1-p_4)\, p_5 (1-p_6)\, p_7\, p_8\, p_9\, p_{10}$$

i.e. $[(1-p_i)(1-p_j)\ldots$ for all non diabetics$]$ · $[p_i \cdot p_i \ldots$ for all diabetes$]$

Logistic regression – relation between dependent & independent variable is linear

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

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

Sigmoid function we use to have logistic regression

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

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

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

$$Odds\ ratio = \frac{0.639}{0.282} = 2.27$$

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

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

Logistic regression is a classification model – relation between dependent & independent variable – which is linear – also not only binary – it can be extended – most of the applications – many independent variables

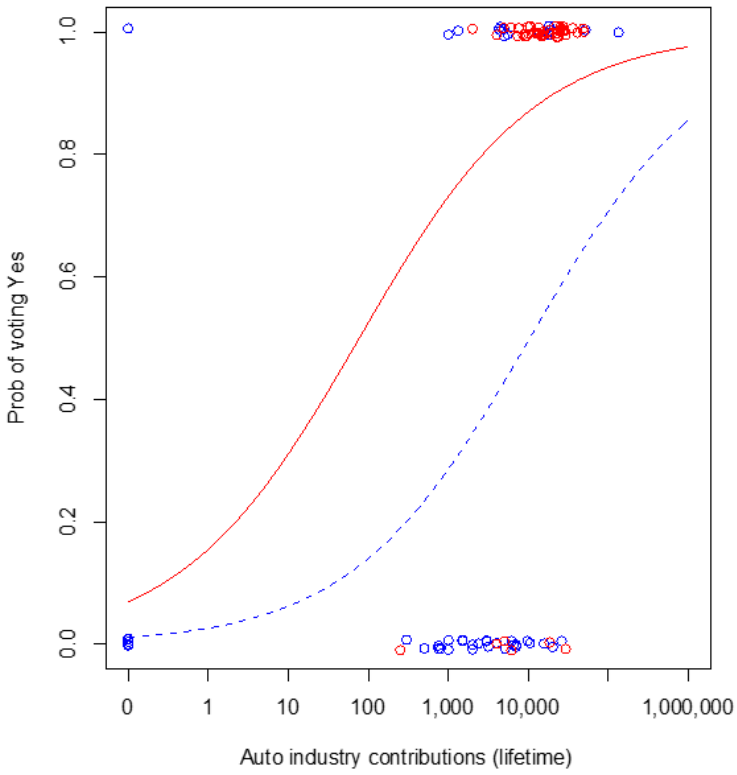
No restriction on number of independent variables here

Interactions in logistic regression

innovate achieve

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

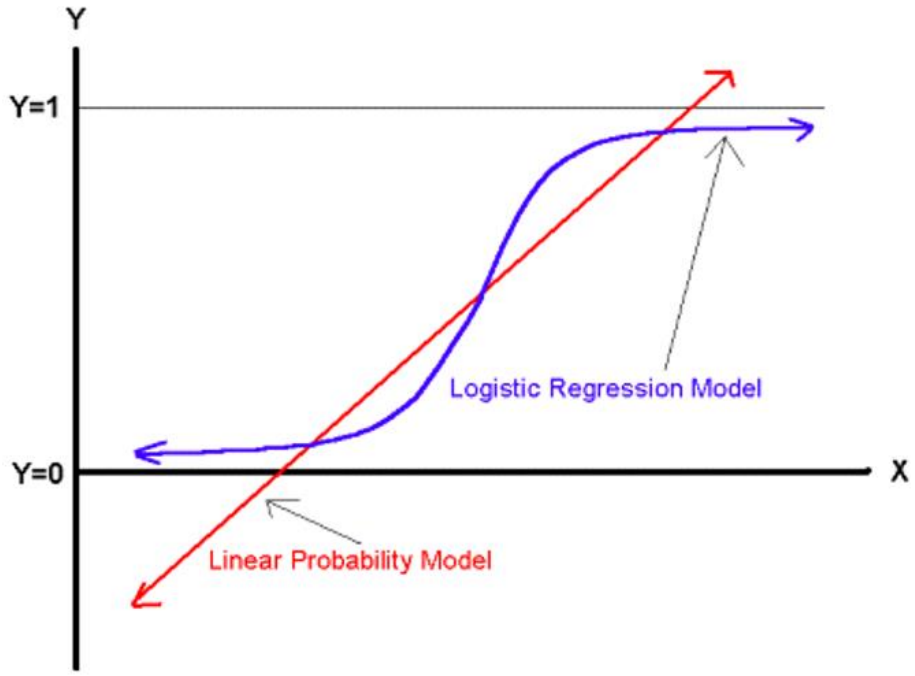
```
> intermodel=glm(Survive~BP*Sex, family=binomial, data=ICU)
> summary(intermodel)

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

Comparing the LP and Logit Models



First line (red) – y = mx + c, second curve – sigmoid function

Log(probabilities) → we can do logistic regression

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, ?

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

a – fixed pattern, b – arithmetic progression

c → sales of government outlet – month wise data – regression technique to find relation – how much inventory need to maintain for coming Diwali season – previous year data (Jan-Oct) – use data to predict sales of next season – right approach? → **No –** not complete data helps – Diwali sales does matter - previous year sales - predict for next Diwali – same season

Dealer of cement industry – how much stock needed for next November → previous data – July to October – not good sale because of rainy season → we can't use this data directly – time series analysis → Stock market analysis – season term – TSA

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?

If we are not having proper analysis of past data we may end up in extreme situations – high inventory – less sale (loss of business) or high sale – low inventory

Importance of Forecasting

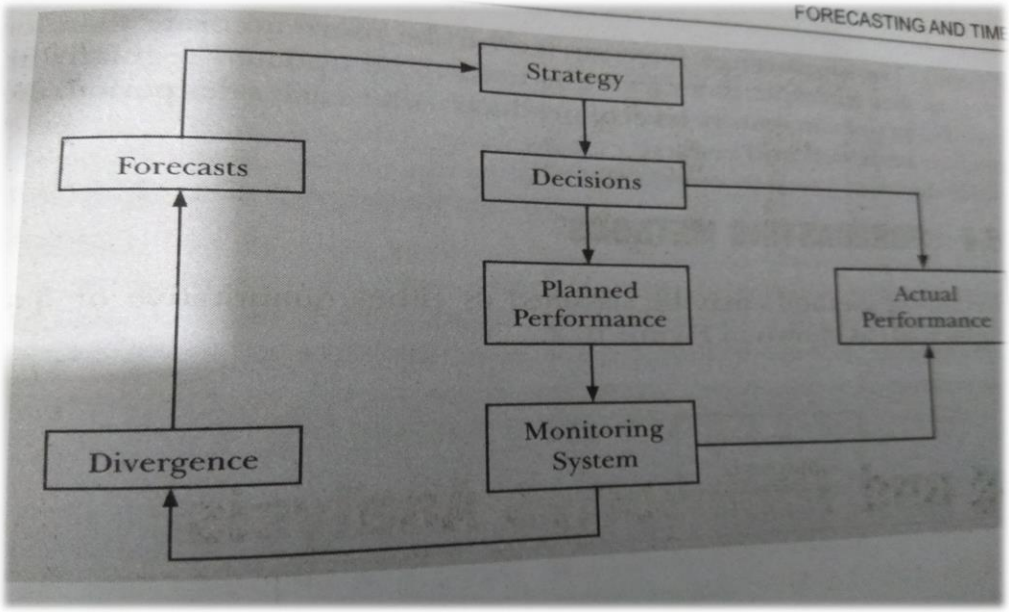
Departments throughout the organization depend on forecasts to formulate and execute their plans.

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

Technological forecasts → Demand for internet data, manufacturing industry – automobile – particular technology – latest car – we use tools & technology available in sector to design & manufacture

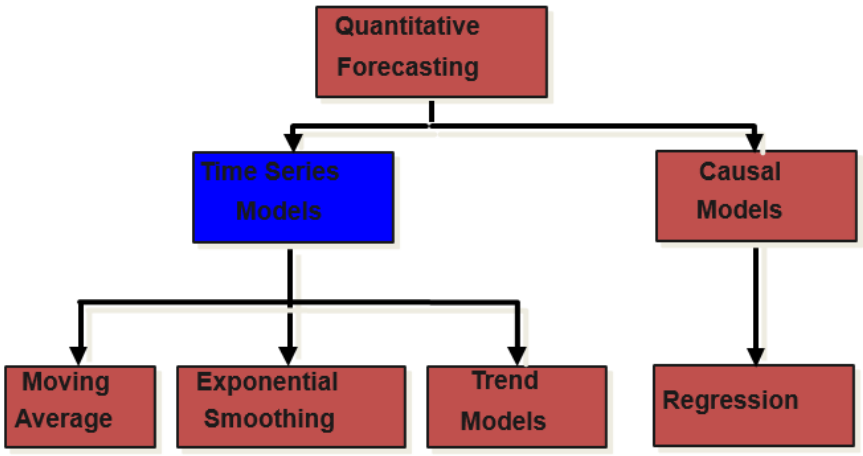
If industry is moving – next version – we need to forecast – our model will take care of those changes – or some language – new oops related language – whether this model can accommodate the change – we need to forecast

Timing of Forecasts

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

Short term – Quarter forecast in terms of financial or human resources, Medium – yearly forecast, Long term – 5 – 10 years etc.

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

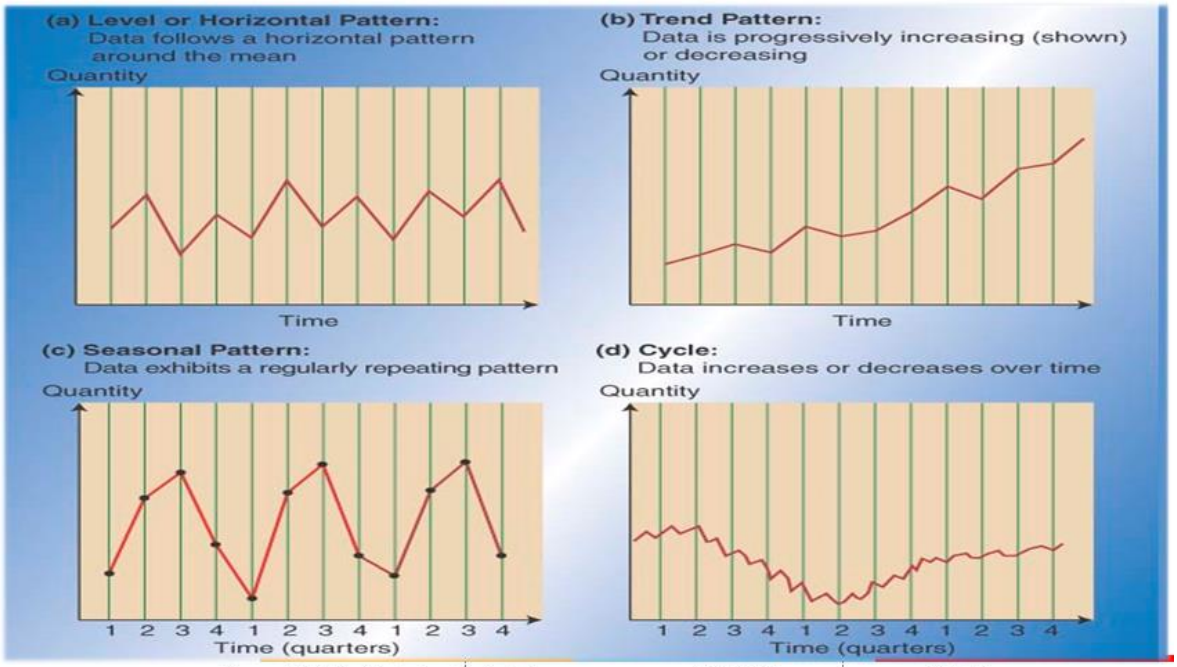
- Forecaster looks for data patterns as

  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

Sales pattern – not constant – year wise getting changed – we try to observe pattern which is associated with some kind of noise or variation → if it's possible for us to eliminate random variation – so we can find pattern based on leftover points

Time Series Patterns



Seasonal – every summer or winter – some sales → we have complete data – use seasonal data, Trend – trend in data – increasing or decreasing

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

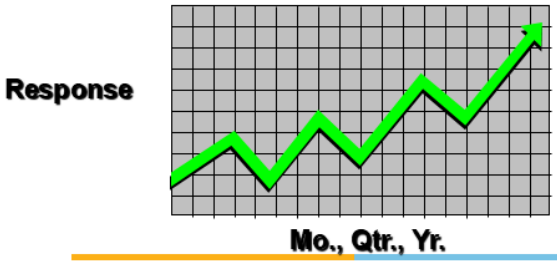
$$X_t = T_t S_t C_t I_t$$

When we have data we have to divide into these components

Additive – addition of all components, Multiplicative – product of all these components

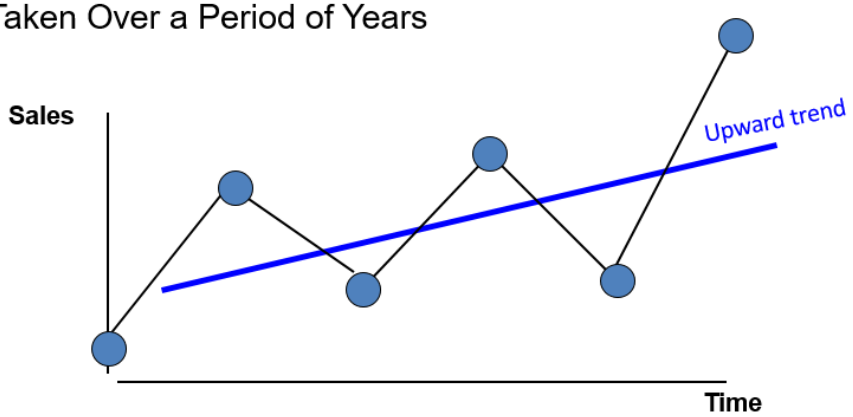
Trend Component

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



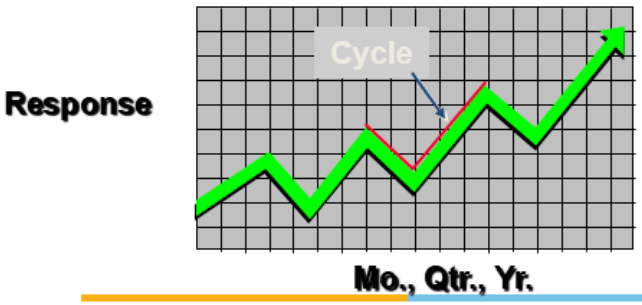
Trend Component

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



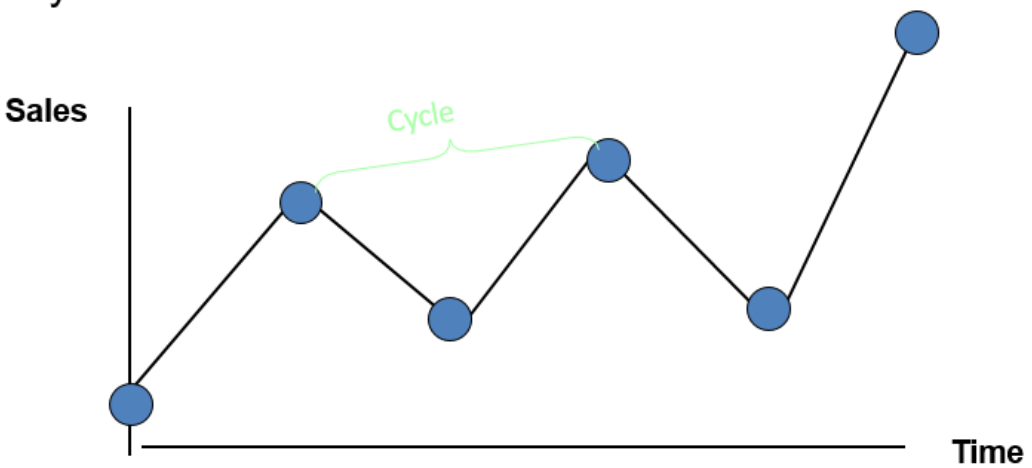
Cyclical Component

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



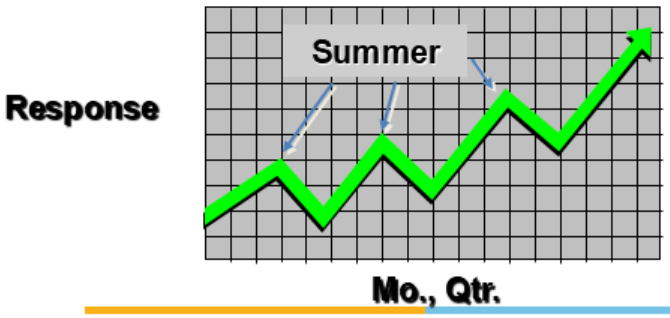
Seasonal Component

innovate achieve

Regular pattern of up & down fluctuations

Due to weather, customs etc.

Occurs within one year



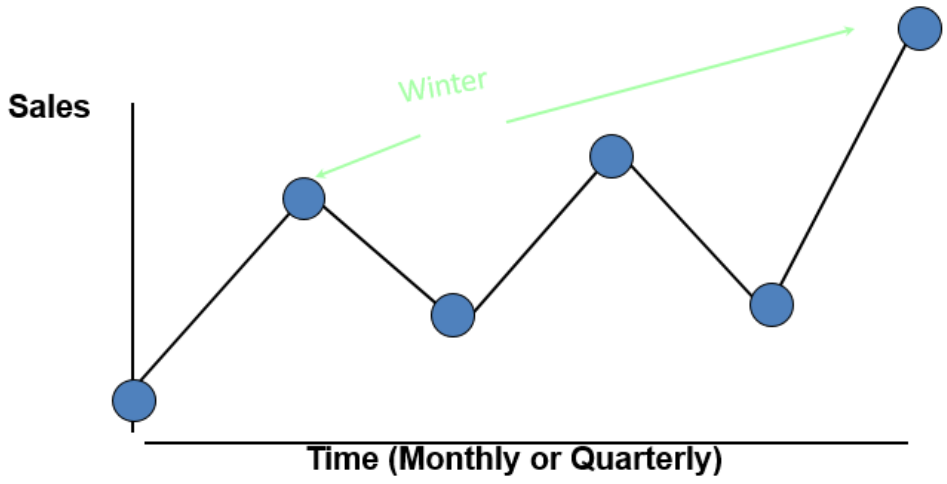
Seasonal Component

innovate achieve

Upward or Downward Swings

Regular Patterns

Observed Within One Year



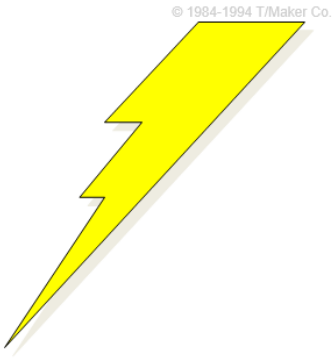
Irregular Component

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

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

Moving average forecast – fix a timeframe and take average value in timeframe → i.e. Jan to Sep data available – predict for October → 2 months frame → Jan + Feb sales – average

Weighted moving average – time is given some weights – last year values are given more weight than the previous years – in case of sales data

Also in organization – more experience – more weight – makes sense

Employees – experience not more than 4 years → more than 4 years – weight is 0

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

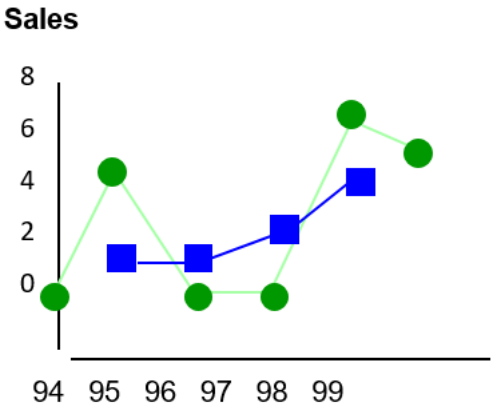
Moving Average

[Solution]

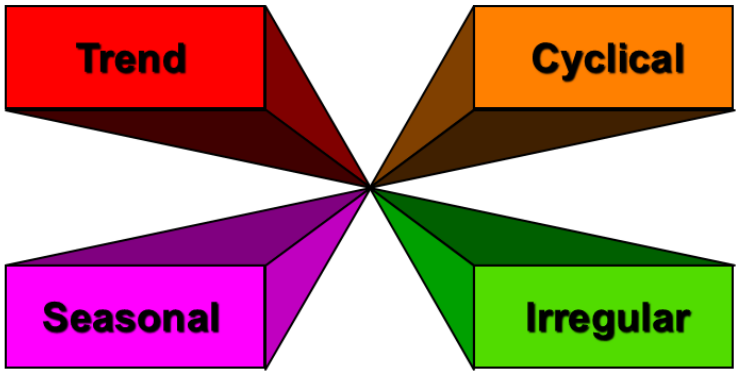
Year	Sales	MA(3) in 1,000
1995	20,000	NA
1996	24,000	(20+24+22)/3 = 22
1997	22,000	(24+22+26)/3 = 24
1998	26,000	(22+26+25)/3 = 24
1999	25,000	NA

Moving Average

Year	Response ●	Moving Ave ■
1994	2	NA
1995	5	3
1996	2	3
1997	2	3.67
1998	7	5
1999	6	NA



Time Series Components



Lec 12 Predictive Analytics _ Time Series Analysis

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 – time component – yearly, quarterly, yearly data → forecast sales for coming years – difference between this model & earlier models (i.e. regression – x value this then what is y value) – here time is the constraint – data we collected time is the imp param

Complete Sales data – year wise & month wise – stock vs sales – prediction, some industry – garment – seasonal data helps (complete data may not)

Given period – sales increased or decreased – randomness

Cyclic – periodic data – time intervals – predict for upcoming interval – every 5 years – when we are dealing with period of time – lengthy period

Trends – voting trends, IRCTC – seasonality & trends – ticket booking scenario

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

Remove components one by one – try to find suitable model – based on selected model – predict – also refinement of model

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

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

It’s not easy to predict sales for next cycle – decompose into various components – easy to find trend, seasonality in data – cyclic behaviour – periodicity – every five years change in price of commodity products – periodicity is for longer time – not days or years

Additive model – linearly connected – change in one component affects linearly, Multiplicative model – change in one model – multiplication – one by one we can take out components & see how time series model behaves

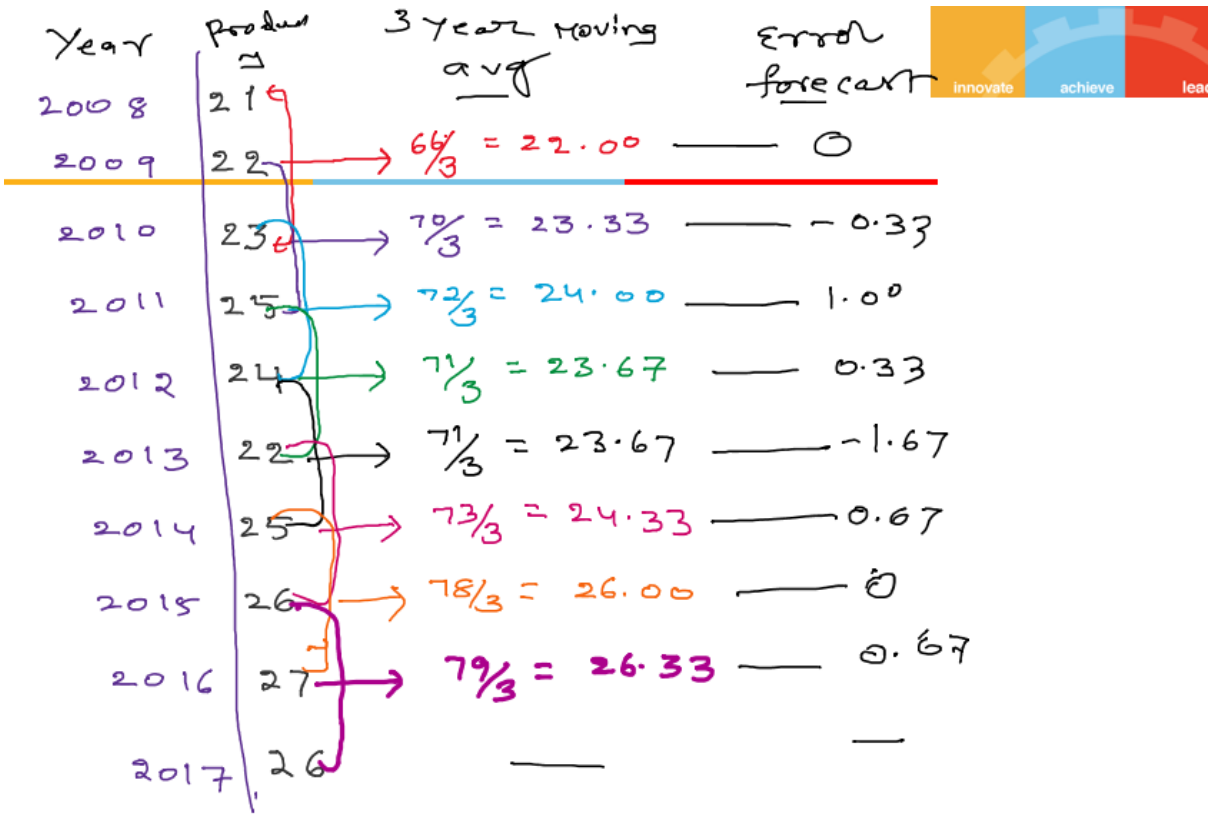
Moving average – takes certain period – cluster or timeframe – i.e. every 3 months – what is the average – trying to build a model – based on error we can determine – how effective model is – depending upon window period we can have different averages

Some of the application latest data is given more weight or vice a versa – based on it we build a model & forecast it – different k – 2,3,4 – what is the error in forecast model – less error select this model

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



Which one is optimal value of k – take different k values – what is the error – trends of k values – we can decide k with minimal error

Test the model based on data we have – 3 year window – forecast for 2018 year

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

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

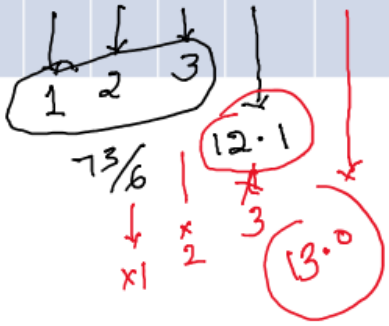
Example(Weighted moving Averages)

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

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



Cost of an apartment or vehicle – last year model vs previous year's models – last year model's price is more – weight we carry is more

(16*3) + (13*2) + (12*1) / 6 → 14.33

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 wts

a)

$$\frac{114 + 121 + 130 + 128 + 137}{5}$$

$$= 126$$

b)

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

$$= 133 \text{ units}$$

Center moving average – first five → middle element – 3rd average

For different k values – forecast value – find k for which we get optimal error

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, ,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 values may place too much weight on last period's random variation

A_t - F_t – Forecasting error – it is controlled by smoothing parameter

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 ie 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 whereas the actual demand is 450 units

a) Forecast demand for the next week ie March 8

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

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

$$= 495$$

Week	(A_t) Demand	(F_t) (old) Forecast	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	$500 + 0.1(510-500) = 501$

Different alpha values → conclude particular alpha is optimal – for different alpha values – how alpha behaves

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**

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

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

Observation – Error is minimal – negligible throughout – optimal alpha value

Measuring Forecasting Accuracy

Mean Absolute Deviation (MAD)

- measures the total error in a forecast without regard to sign

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

Cumulative Forecast Error (CFE)

- Measures any bias in the forecast

$$CFE = \sum (actual - forecast)$$

Mean Square Error (MSE)

- Penalizes larger errors

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

Tracking Signal

- Measures if your model is working

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

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

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

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

x_t – random variable – auto covariance between every pair

Auto Correlation Function

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

Lec 13 Time Series Analysis (cont..)

L- 13:Time Series Analysis(cont..) ☺ ?

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

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

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

$$= 495$$

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

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

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

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

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 $\longrightarrow$ AR(p)
- MA Model $\longrightarrow$ MA(q)
- ARMA Model $\longrightarrow$ 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

y_t = f(ϵ_t (today's announcement), ϵ_{t-1}, ϵ_{t-2} (Yesterday's), ⋯)

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

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

$\theta_q \neq 0$

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

Epsilon – follow **normal distribution** – mean = 0, variance = sigma^2 – for all values of 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)

Different values – different models – moving average or weighted moving average model

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

For which value of k – we get optimal model, Length of horizon – smaller period or longer

Time series data – components – decompose – how particular component behaves

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

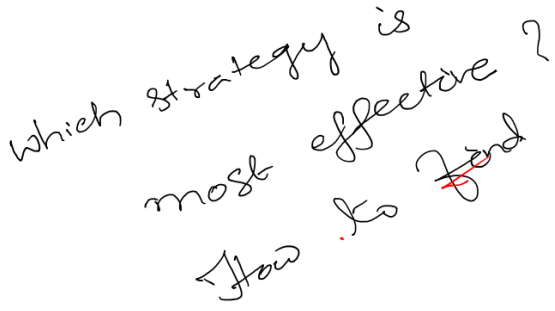
Group 3 eats junk food &does cardio exercise every other day

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…….

We have various components & we need to find various observations between given data – impact of nutrition on various groups – same or different



Case 1 – one group at a time, Case 2 – compare two groups, Case 3 – multiple groups

One group at a time – Hypothesis (one mean problem) – z or t distribution based on group size, Two groups at a time – means of two samples – Mu1 = Mu2

Generally we want to know Impact of certain condition on more number of groups – medicine – effects on different set of people before moving to production (i.e. age wise)

ANOVA-analysis of variance

* Significance of difference between two sample means

$H_0 = \mu_1 = \mu_2 = \cdots = \mu_K$

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

Null Hypothesis

Alternate hypothesis

Whether impact of certain condition on different groups – are similar or not – mean of the sample & population

Sales – One - location wise performance – Another - also how different groups are in marketing (teams)

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 σ_i^2

Each sample is drawn randomly and independent of other samples

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

Impact of various parameters on different groups – ANOVA Model – i.e. impact of different seeds on yield of the crop

F Distribution we have to use – S1 – variance of first sample – degrees of freedom – two samples – two degrees of freedom – n1, n2

ANOVA summary

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

Short cut method

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

Cal. Fact: $CF = \frac{T^2}{m}$, $m = m_1 + m_2 + \cdots + m_n$

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

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

$$SSE = SST - SSTR$$

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

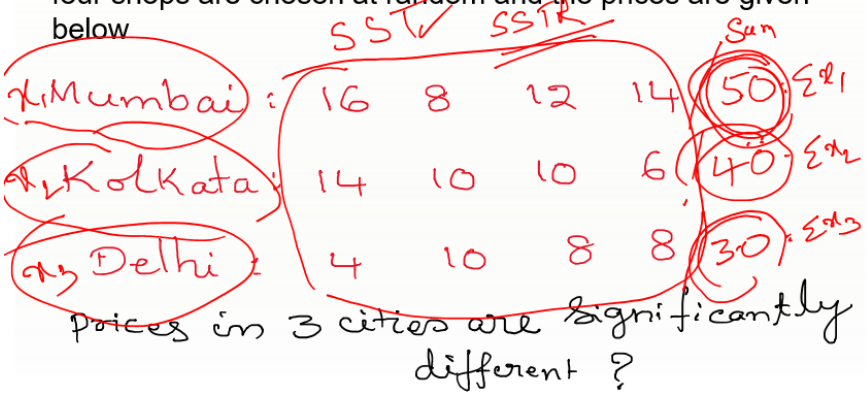
Kolkata : 14 10 10 6

Delhi : 4 10 8 8

Prices in 3 cities are significantly different ?

Two cities – Mu1 = Mu2 – Null hypothesis, Alternate Mu1 – Mu2 > 0 → t distribution

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



Short cut method

$$T = \sum x_1 + \sum x_2 + \dots + \sum x_n = 120$$

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

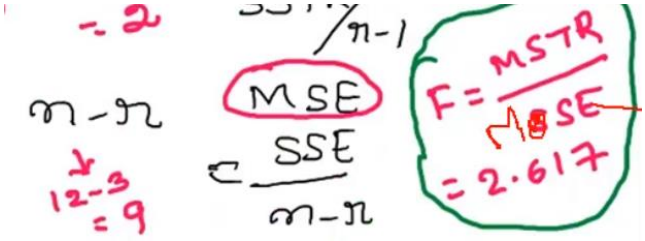
$$SST = \left[\sum (x_1^2) + \sum x_2^2 + \dots \sum x_n^2\right] - CF \rightarrow 136$$

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

$$SSE = SST - SSTR = 86$$

ANOVA summary

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



F = MSTR/MSE

Calculated F = 2.617

From tables, for $\gamma_1 = 2$, $\gamma_2 = 9$

$\gamma_1 \downarrow n-1$ $\quad$ $\gamma_2 \downarrow n-r$

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

innovate achieve

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

$k = 3, n = 18$

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

$= 30 + 27 + 36 = 93$

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

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

$= 154 + 123 + 218 = 14.50$

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

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

$= 7$

$SSE = SST - SSTR$

$= 14.50 - 7 = 7.50$

$MSTR = \frac{SSTR}{df_1} = \frac{7}{2} = 3.5$ → $k-1 = 3-1 = 2$

$MSE = \frac{SSE}{df_2} = \frac{7.5}{df_2} = 0.5$ → $n-k = 18-3 = 15$

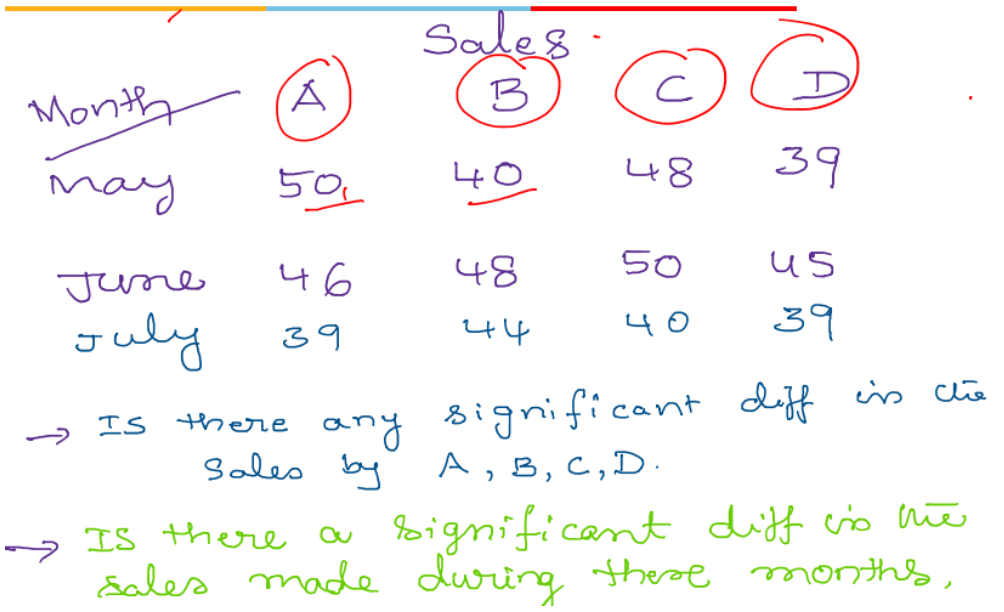
$F = \frac{MSTR}{MSE} = \frac{3.5}{0.5} = 7$

calculated value: 7

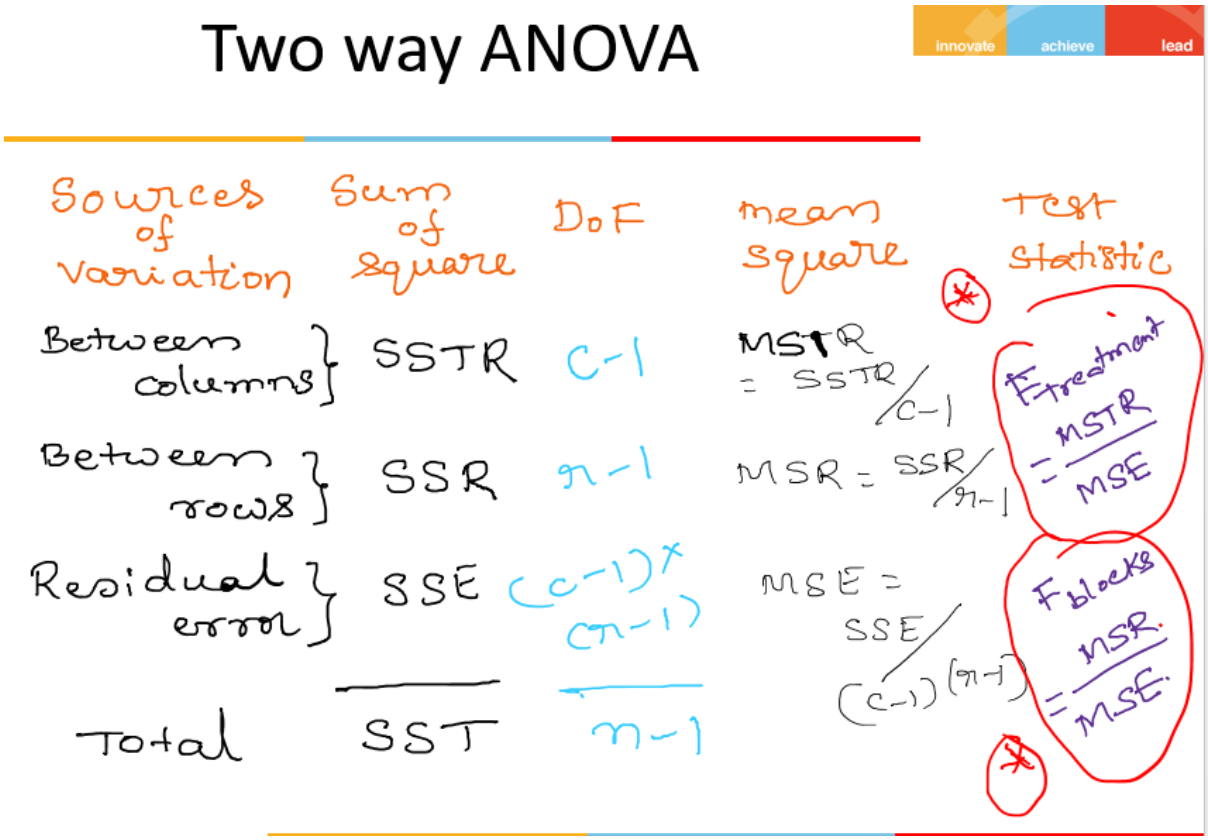
Table value: 3.68 (at 5%)

$7 > 3.68 \Rightarrow$ Rejected.

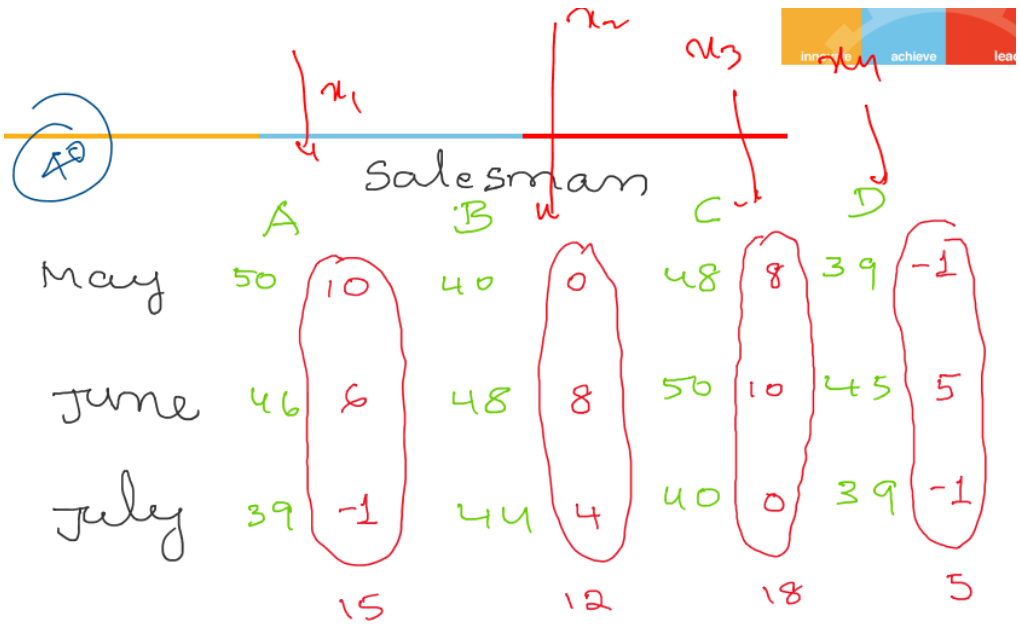
Example



Two way ANOVA



If we have larger values – we can subtract some constant from all the entries – for easier calculation



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

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

SSTR = Sum of squares (of salesman columns)

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

$$= 42$$

SSR = Sum of squares between months (rows)

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

$$= 91.5$$

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

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

$$= 216$$

$$SSE = SST - (SSTR + SSR)$$
$$= 216 - (42 + 91.5)$$
$$= 82$$

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

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

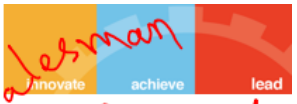
$$MSTR = \frac{SSTR}{c-1} = \frac{42}{3} = 14$$

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

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

$MSTR < MSE$ ← $MSTR > MSE$

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



columns

	Sum of squares	D.o.f	Mean squares	Variance ratio
* Between Salesmen	SSTR 42.0	$c-1$ 3	MSTR $= \frac{42.0}{3} = 14.$	$F_{treatment} = \frac{14}{13.75} = 1.018$
* Between months	SSR 91.5	$r-1$ 2	MSR = 45.75	$F_{block} = \frac{45.75}{13.75} = 3.327$
* residual error	SSE 82.5	$(c-1)(r-1)$ 6	MSE 13.75	
total	216	11		

dof

rows → Months

a) $F_{treatment} = 1.018 <$

$df_1 : 3, df_2 = 6$

$\alpha = 0.05$

accept

b) $F_{block} : 3.327 < X$

2, b

row

accept

difference in the Sales ~~by salesman~~

Difference in Sales made during months.

Multi Variate analysis

Introduction

Multivariate normal distribution

Principal component Analysis

Factor Analysis

Discriminant Analysis

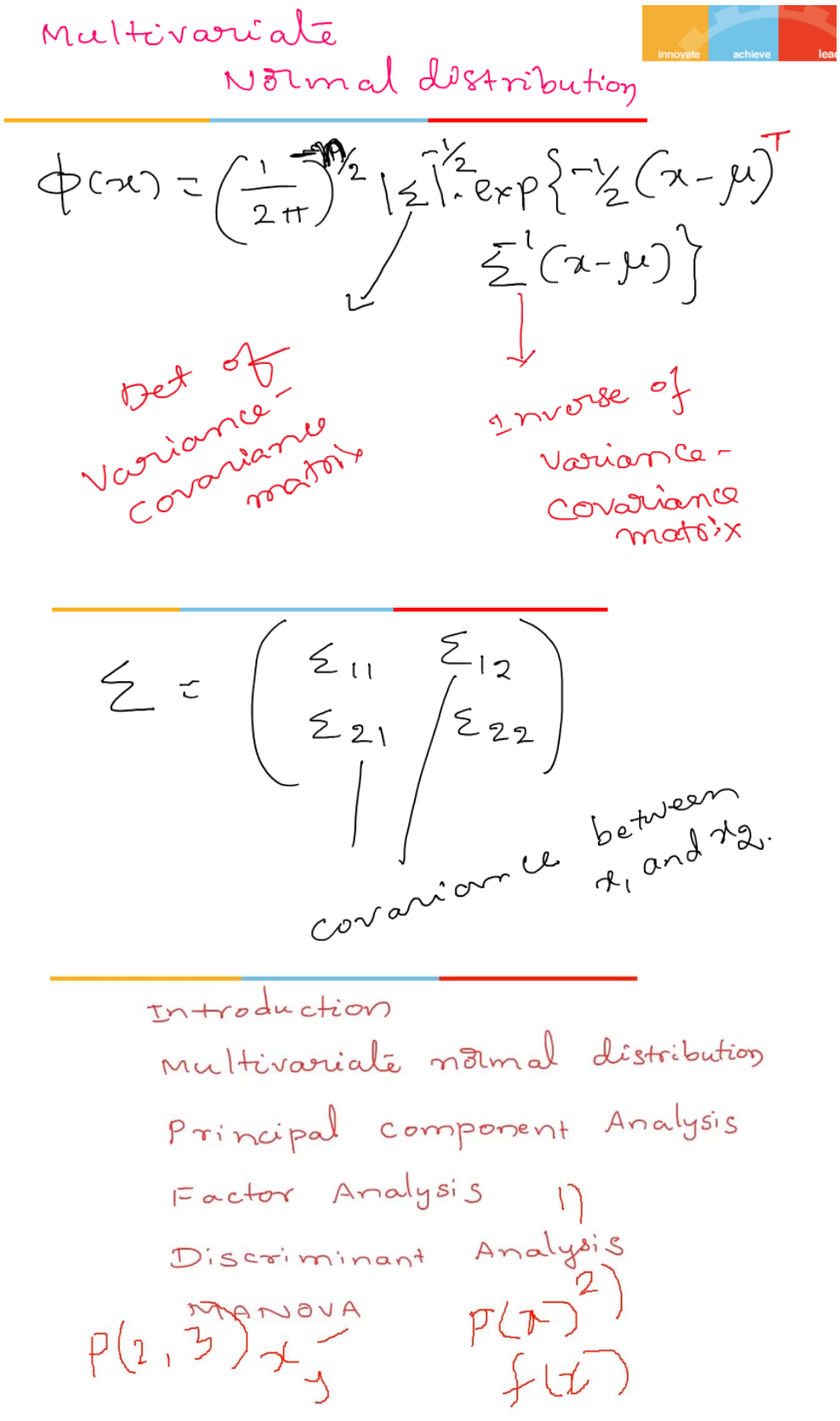
MANOVA

Bi-Variate Normal dist

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

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

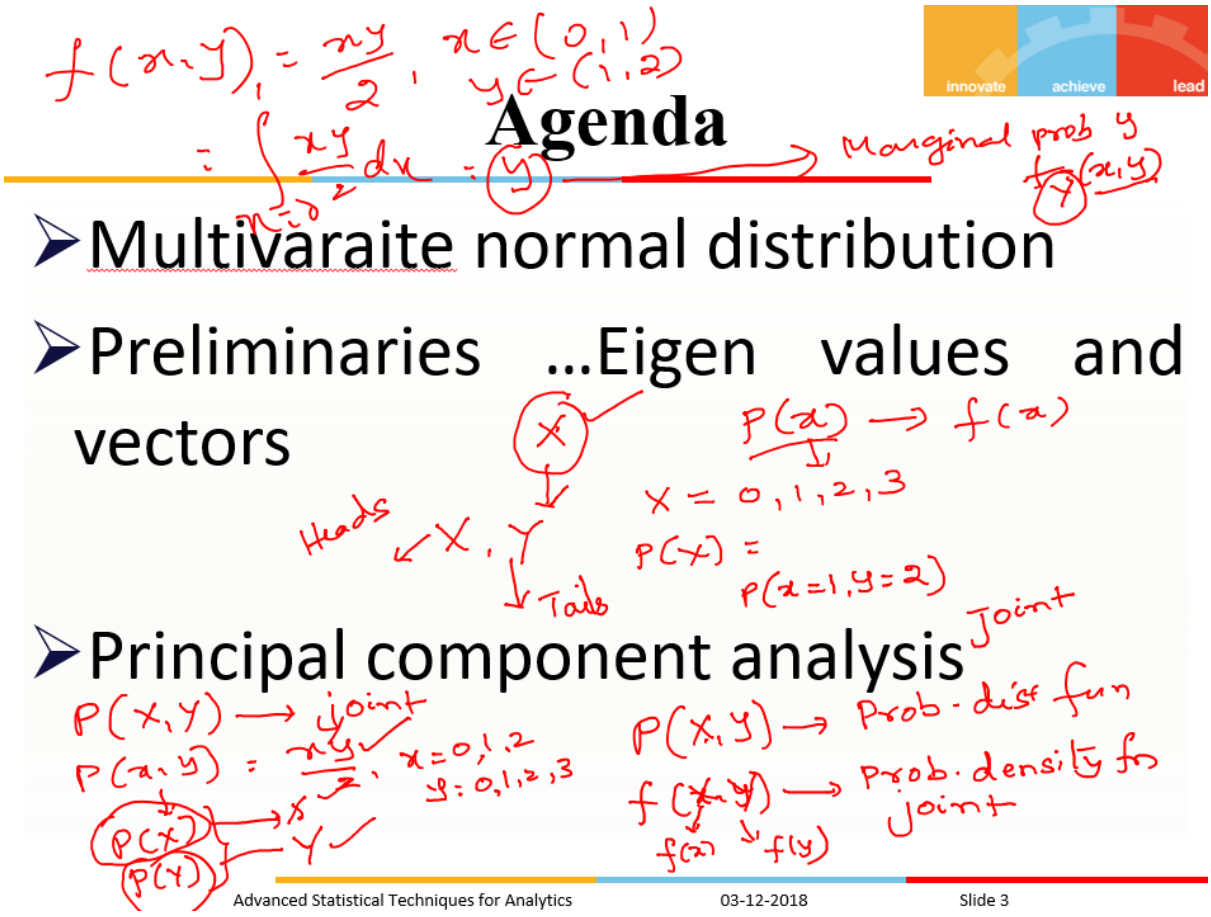
$$\rho = cor(x_1, x_2) = \frac{cov(x_1, x_2)}{\sigma_1 \sigma_2}$$



Probability distribution of many variable – joint distribution – P(x, y)

Lec 14 Applied Multivariate Analytics

L- 14:Applied Multivariate Analytics



Can we continue with multiple number of variables – two random variables here – one w.r.t. f(x) and one is w.r.t. f(y) → P(x, y) – Probability distribution function of x & y

f(x, y) → Joint probability density function

We want w.r.t. one variable – from joint to separate from → Marginal probability of x or y

Bi-Variate Normal dist

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

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

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

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

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

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

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

More random variables – P(x |y) → what is the probability of head given two tails

Multivariate Normal distribution

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

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

innovate achieve

$\sigma = 5$

$\sigma^2 = 25$

$\frac{\Sigma(x-\mu)^2}{n}$

- 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

$cov(x,y) = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\sigma_x \sigma_y}$

$cov(x,x) =$ Variance 9.490233116

Covariance Matrix

innovate achieve

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

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

x, y, z

Symmetric matrix

$A^T = A$

- 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

Preliminaries – Helps us to understand how data is spread w.r.t. mean of the data?

Covariance – how much each of the dimension varies from another – w.r.t. x how y is varying and vice a versa

Covariance – how variable differs it self taking mean as base

Covariance matrix is a **symmetric** 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:

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

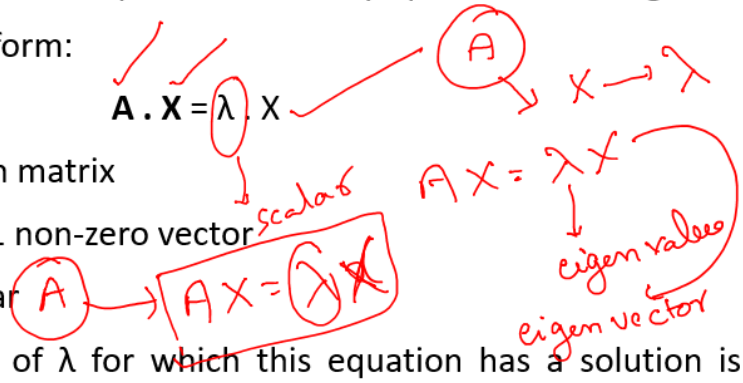
Square transformation matrix transforms (3,2) from its original location. Now if we were to take a multiple of (3,2)

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

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

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.



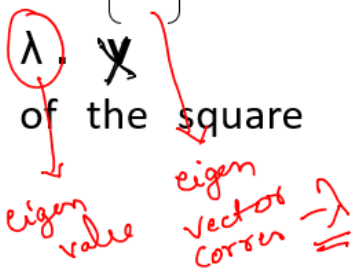
Depending on Lambda X varies, X is a eigen vector corresponding to eigen value Lambda

eigenvalue problem

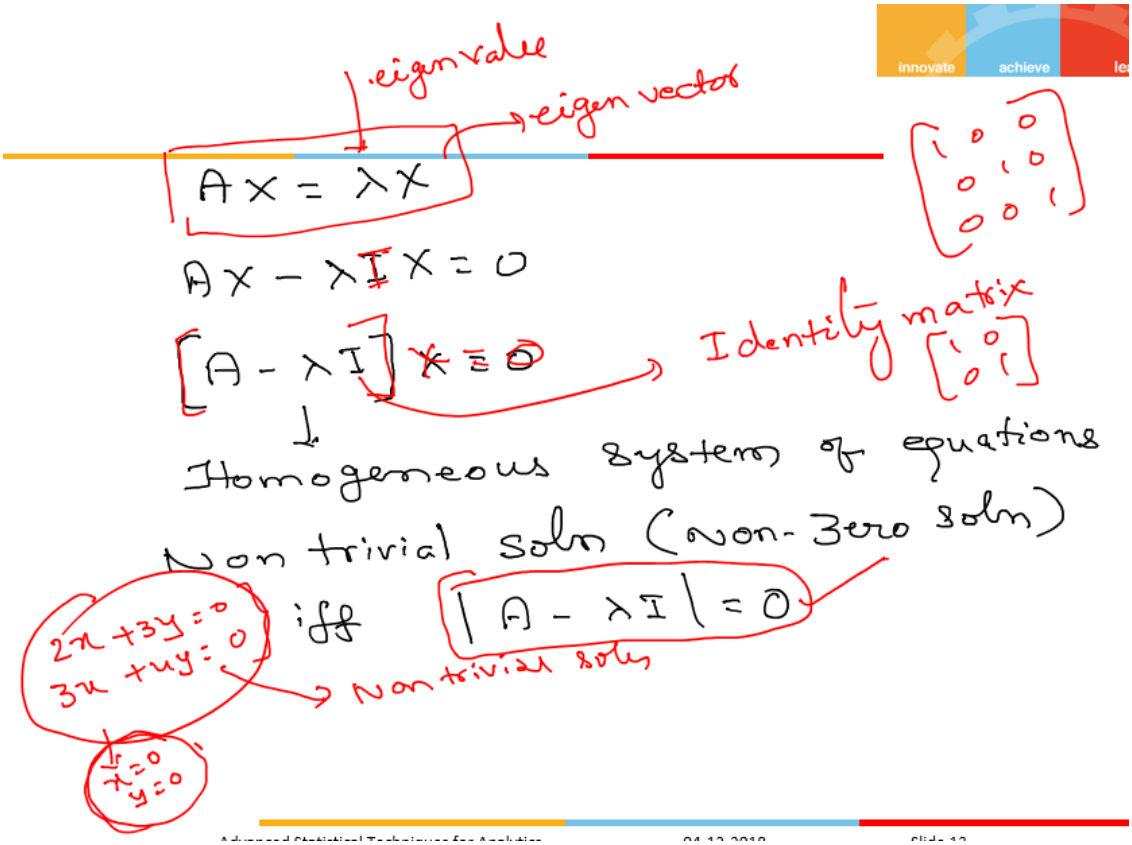
$$\begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix} \times \begin{bmatrix} 3 \\ 2 \end{bmatrix} = \begin{bmatrix} 12 \\ 8 \end{bmatrix} = 4 \times \begin{bmatrix} 3 \\ 2 \end{bmatrix}$$

A . X = λ . X

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



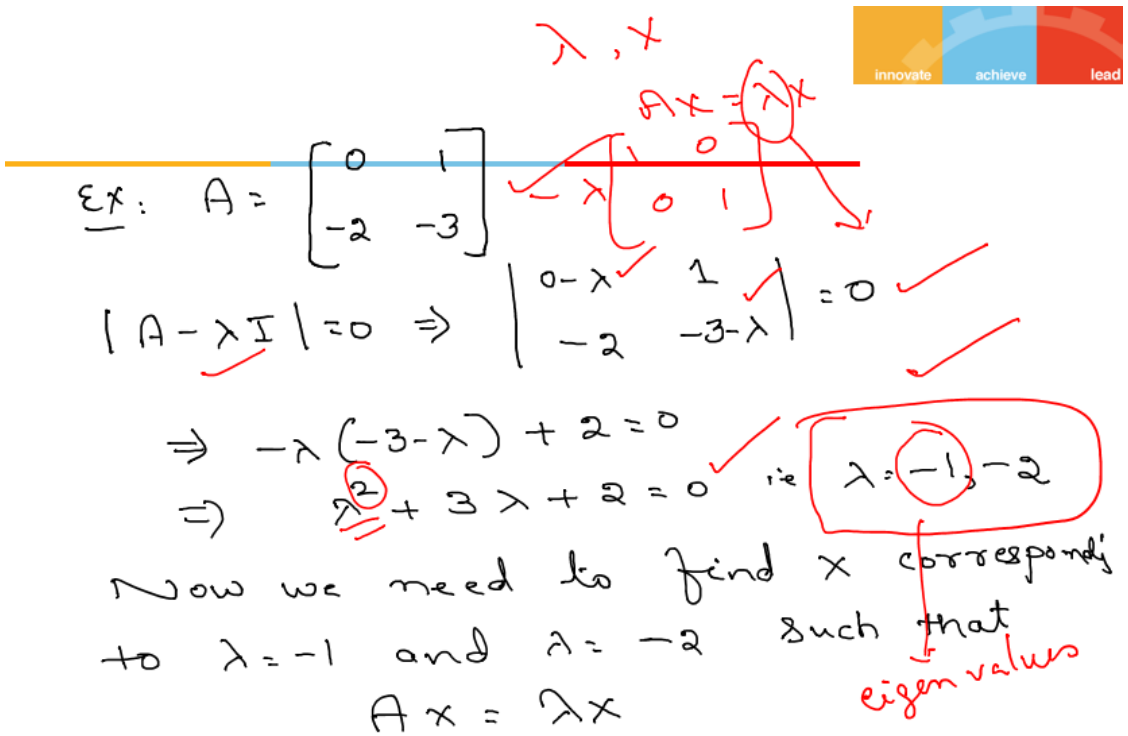
Given matrix A, how can we calculate the eigenvector and eigenvalues for A?

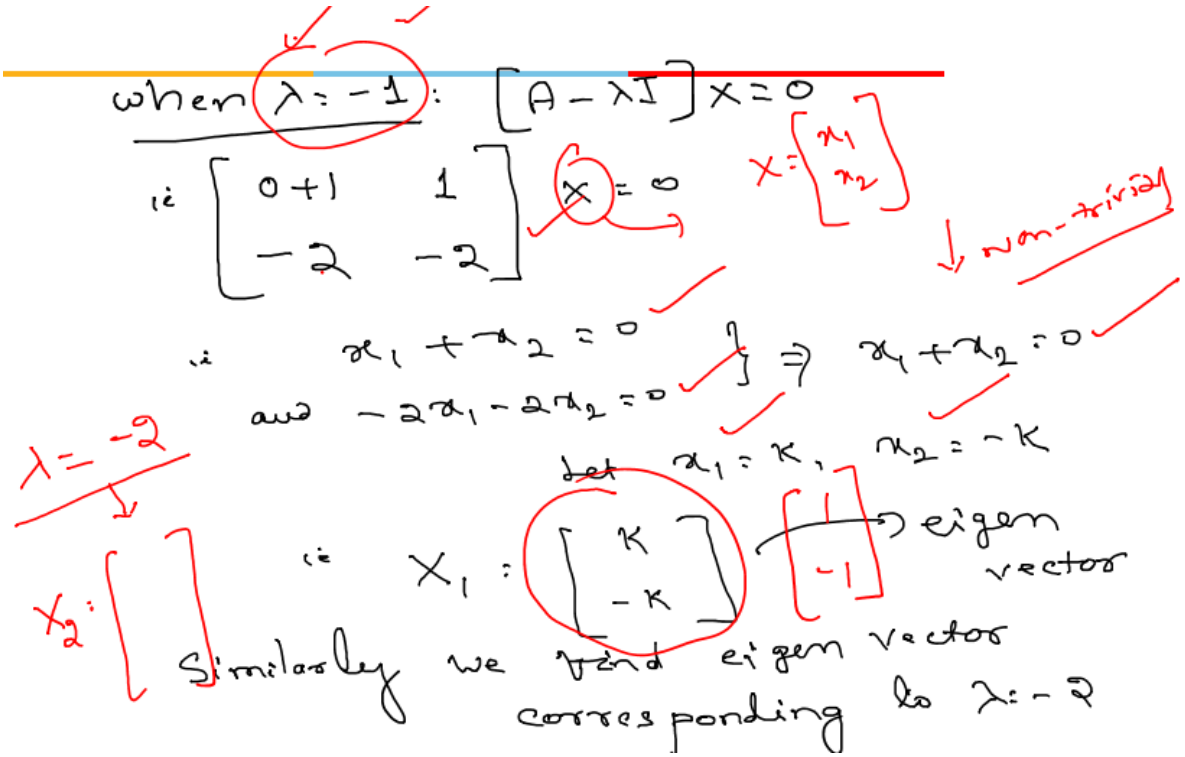


Homogeneous system → ax+b = 0, ax+by = 0, while ax + by = c is not homogenoous

One possibility → x = 0 and y = 0 → we have to check for any non trivial solution

Det(A – Lambda*I) = 0 → extend in terms of Lambda so we can have Lambda value





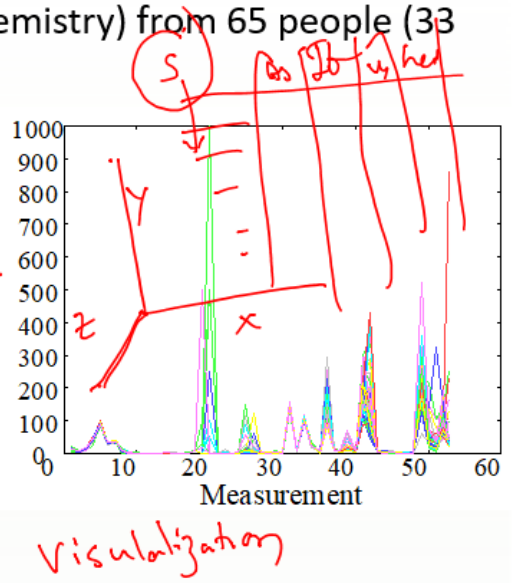
k is an arbitary constant – for homogenous system – infinite number of solutions

Data Presentation

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

Matrix Format

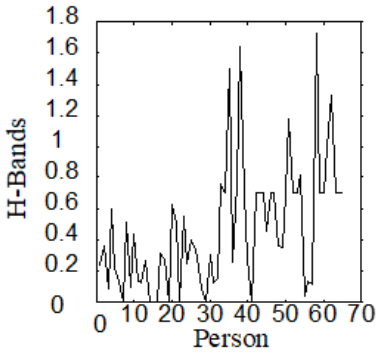
	H-WBC	H-RBC	H-Hgb	H-Hct	H-MCV	H-MCH	H-MCHC
A1	8.0000	4.8200	14.1000	41.0000	85.0000	29.0000	34.0000
A2	7.3000	5.0200	14.7000	43.0000	86.0000	29.0000	34.0000
A3	4.3000	4.4800	14.1000	41.0000	91.0000	32.0000	35.0000
A4	7.5000	4.4700	14.9000	45.0000	101.0000	33.0000	33.0000
A5	7.3000	5.5200	15.4000	46.0000	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



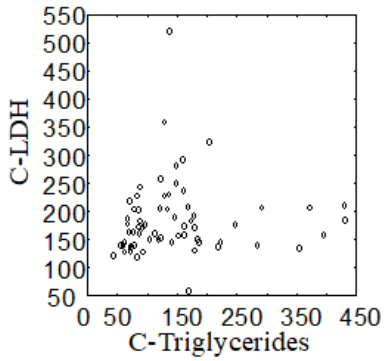
How data is related, how they are forming clusters? If we have more number of parameters then it is a problem

Which one is important or major parameter which helps us to determine performance? – major contribution for us to understand the pattern

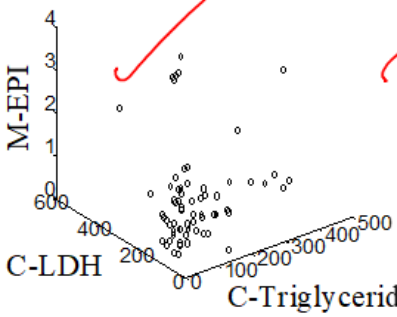
Univariate



Bivariate



Trivariate



Principal Components

Applications

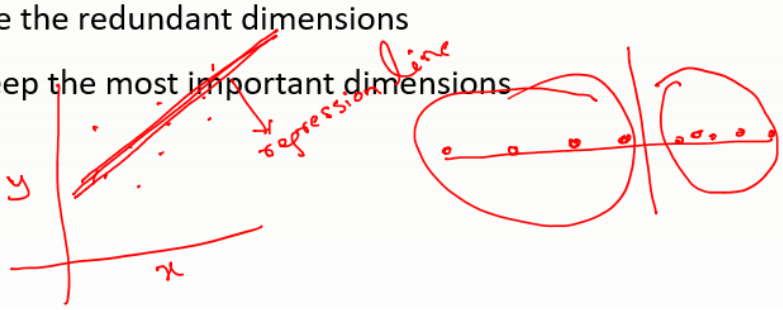
- Face Recognition
- Image Compression
- Gene Expression Analysis
- Data Reduction
- Data Classification
- Trend Analysis
- Factor Analysis
- Noise Reduction

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



Many dimensions – we try to reduce dimensions which are redundant

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.

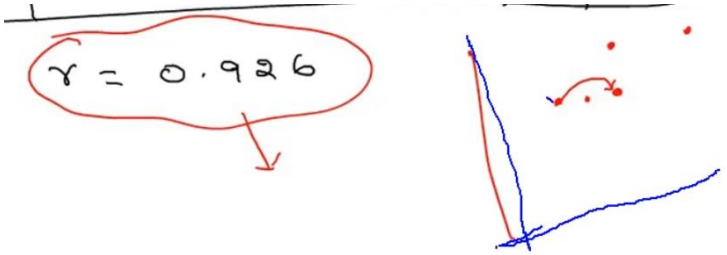
Transform data of n dimensions to k dimension data

Example:

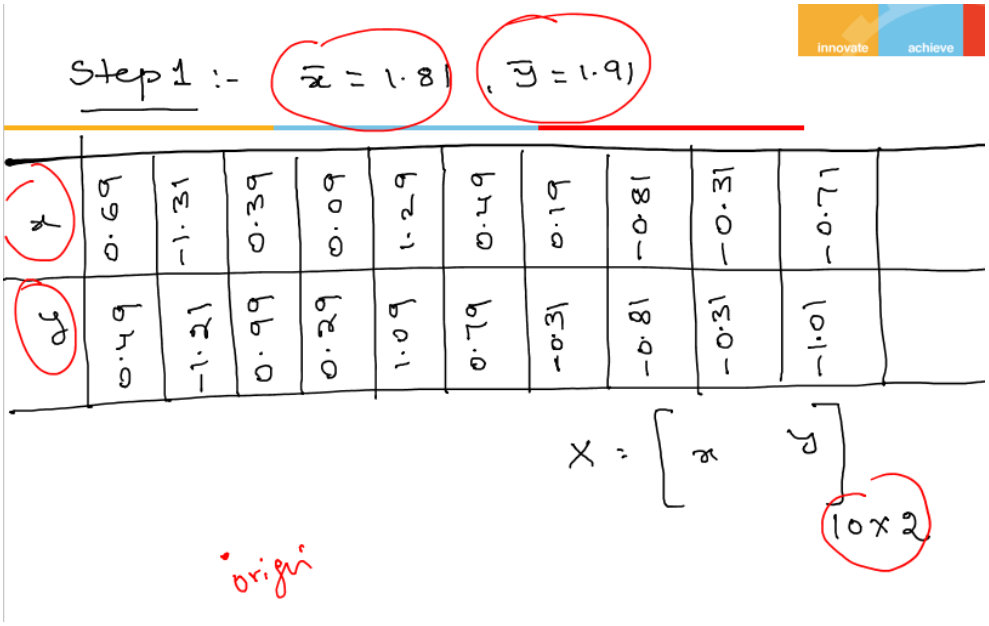
consider the following data

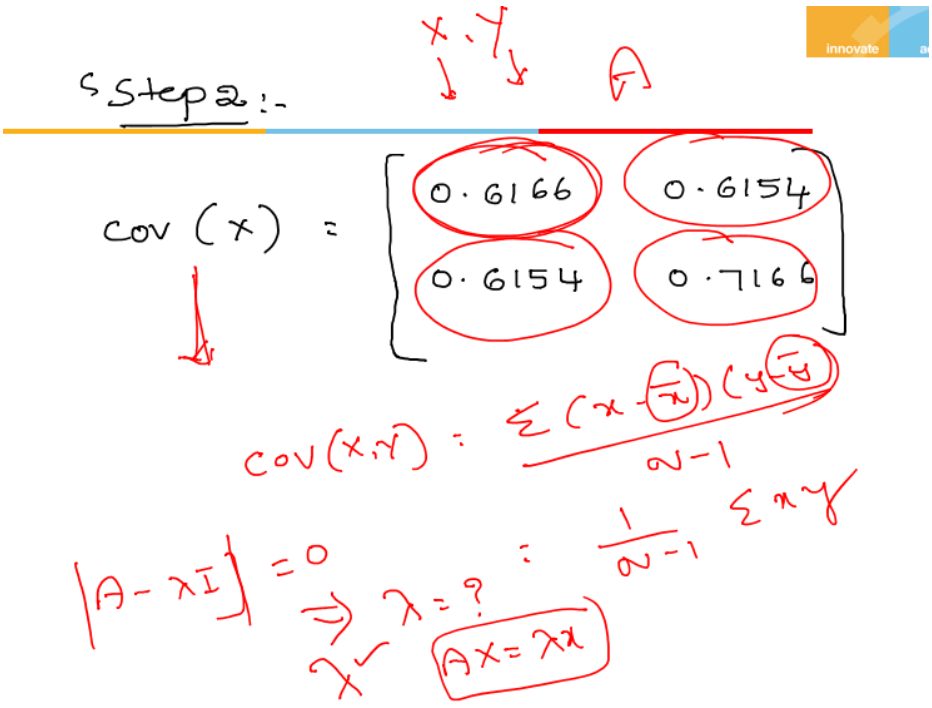
x	2.5	0.5	2.2	1.9	3.1	2.3	2	1	1.5	1.1
y	2.4	0.7	2.9	2.2	3.0	2.7	1.6	1.1	1.6	0.9

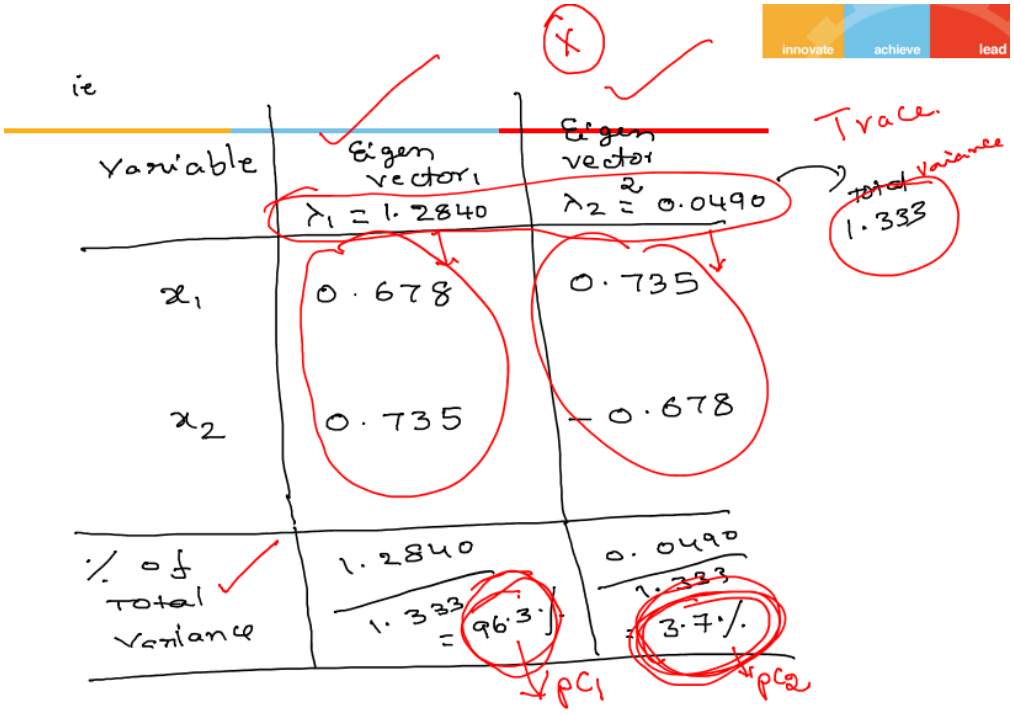
$r = 0.926$

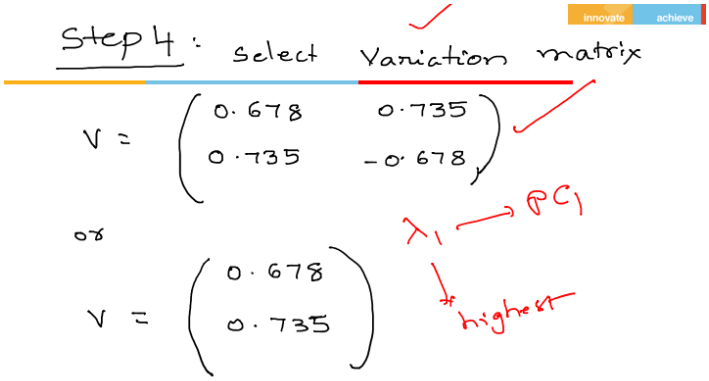


Transform w.r.t. to x & y – normailzation of points – shifting of origins









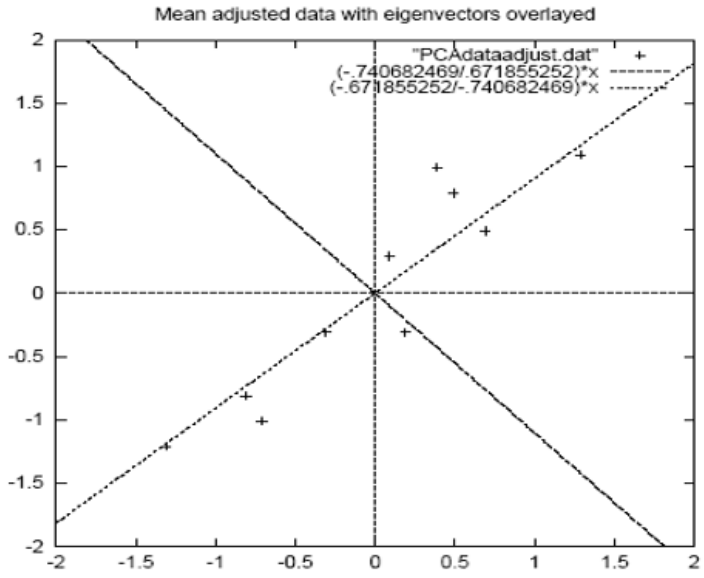


Figure 3.2: A plot of the normalised data (mean subtracted) with the eigenvectors of the covariance matrix overlayed on top.

innovate achieve

Step5. Find new Data set $Y = XV$

ie $Y = XV$

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

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

PC2 y_2 — PC1 y_1

$y_2 = 0.735 x_1 - 0.678 x_2$

$\therefore Y_1 = 0.678 x_1 + 0.735 x_2$

innovate achieve lead

Step5. Find new Data set $Y = XV$

ie $Y = XV$

$\lambda_1 =$

λ_1 ✓

λ_2 ✓

λ_2

$\lambda =$ 0.3

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

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

linear ~~transf~~ Combination

$\therefore Y = 0.678 x_1 + 0.735 x_2$

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

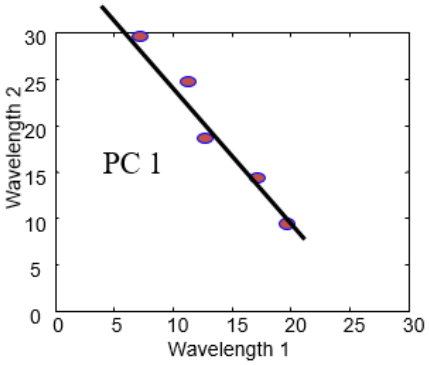
4) Find the transformation matrix V based on PC selection

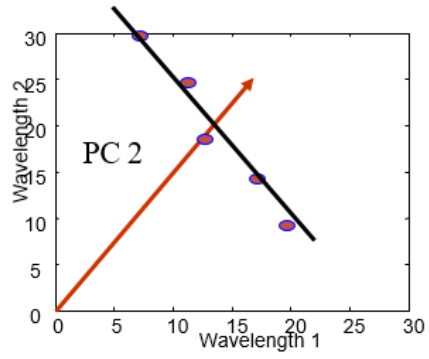
5) Derive the new data set by Y = XV

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

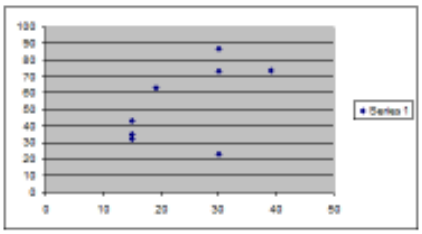


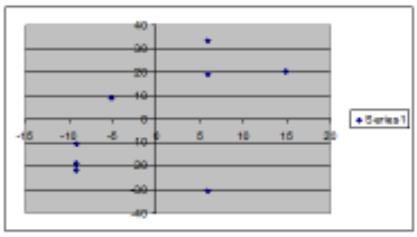


An Example

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





Covariance Matrix

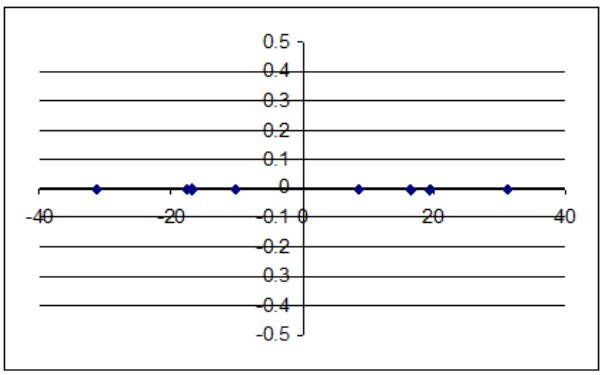
- C=

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

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

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.

PC1, PC2, PC3 all are perpendicular to eachother

Multi variables – its tedious task – takes time to solve a problem

Lec 15 Applied Multivariate Analytics & Revision

L- 15:Applied 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 $\boxed{P(A) + P(\bar{A}) = 1}$

Sample space – We can define different events → sample space is suppose n

Mutually exclusive events – A ∩ B = Φ

Conditional Probability

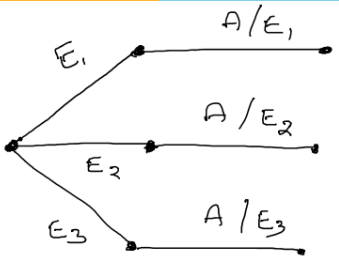
$$P(A/B) = \frac{P(A \cap B)}{P(B)}$$

or

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

Conditional probaibility – sequence of events – only two events else go for Baye's theorem

Baye's theorem



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

Denominator is same for all events – we can remove denominator (programming – saves time – division operation takes time)

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

Random Variables

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

Choosing appropriate distribution – discrete or continuous? – depends on x values – Poisson – rarest of the rare events – n is very large & p is very small → mean = Lambda

Normal – mean is 0 & variance is 1 → convert → standard Normal distribution

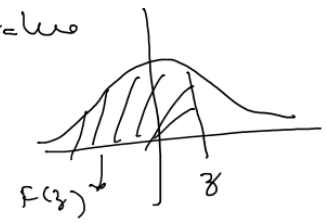
Normal distribution

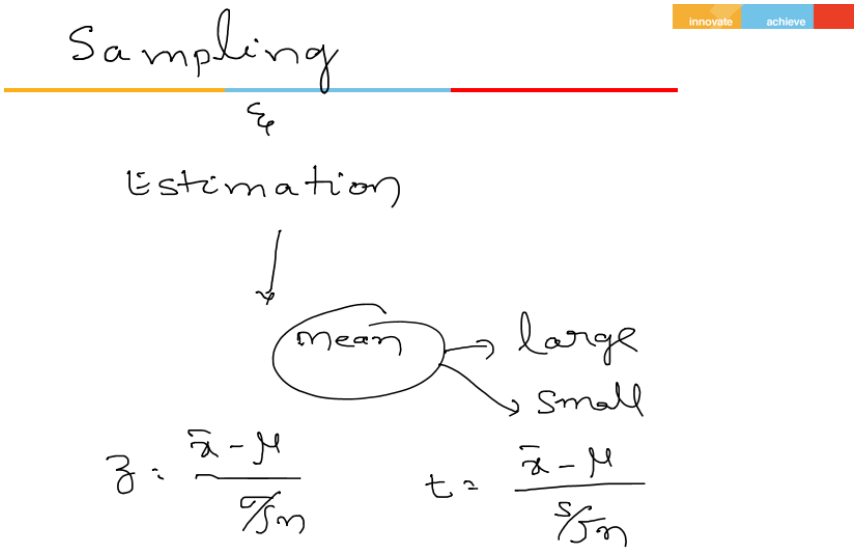
$$P(x_1 \le x \le x_2) \qquad z = \frac{x-\mu}{\sigma}$$

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

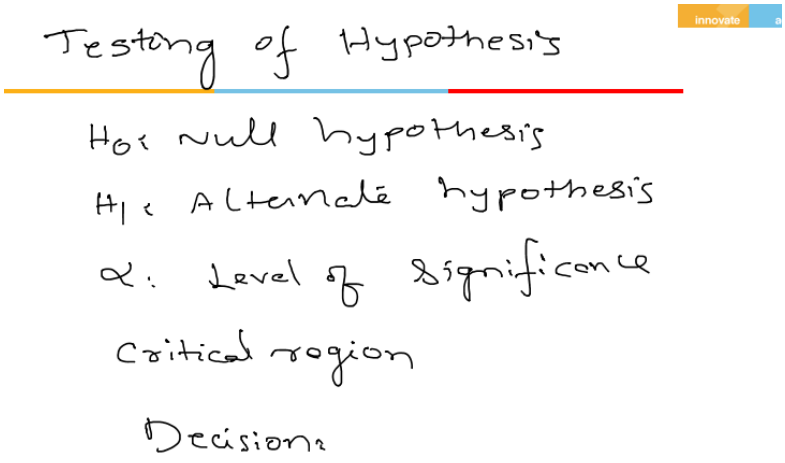
$$= F(z_2) - F(z_1)$$

$F(z)$ is tabulated value

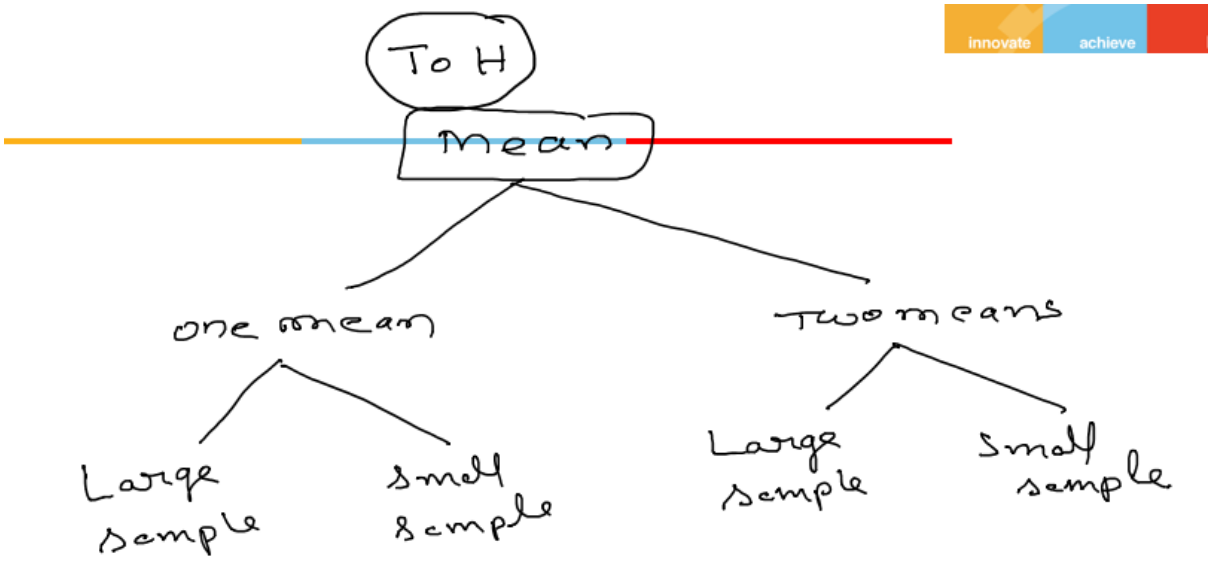


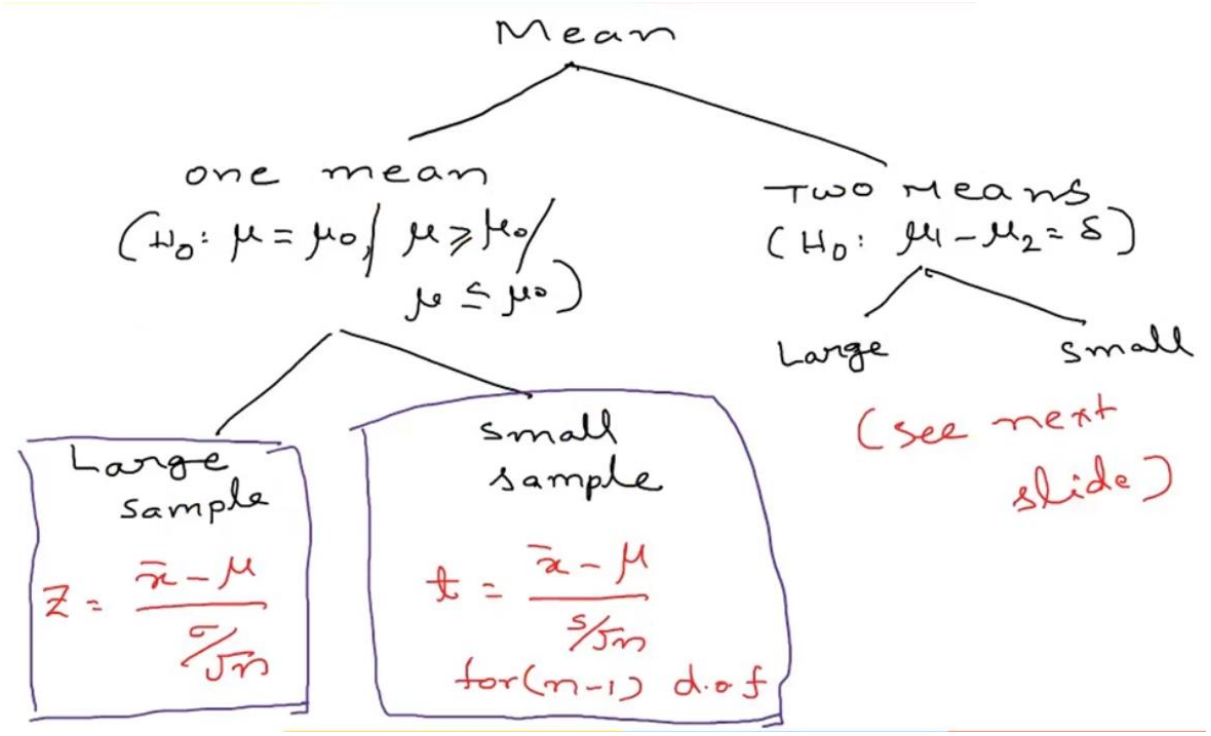


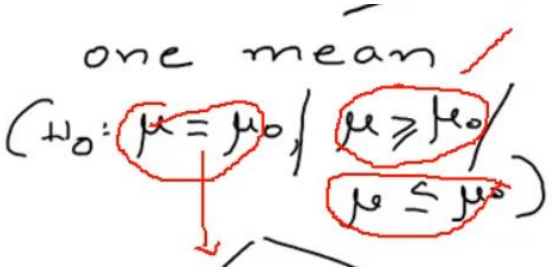
z & t distribution – degrees of freedom, n – size of sample



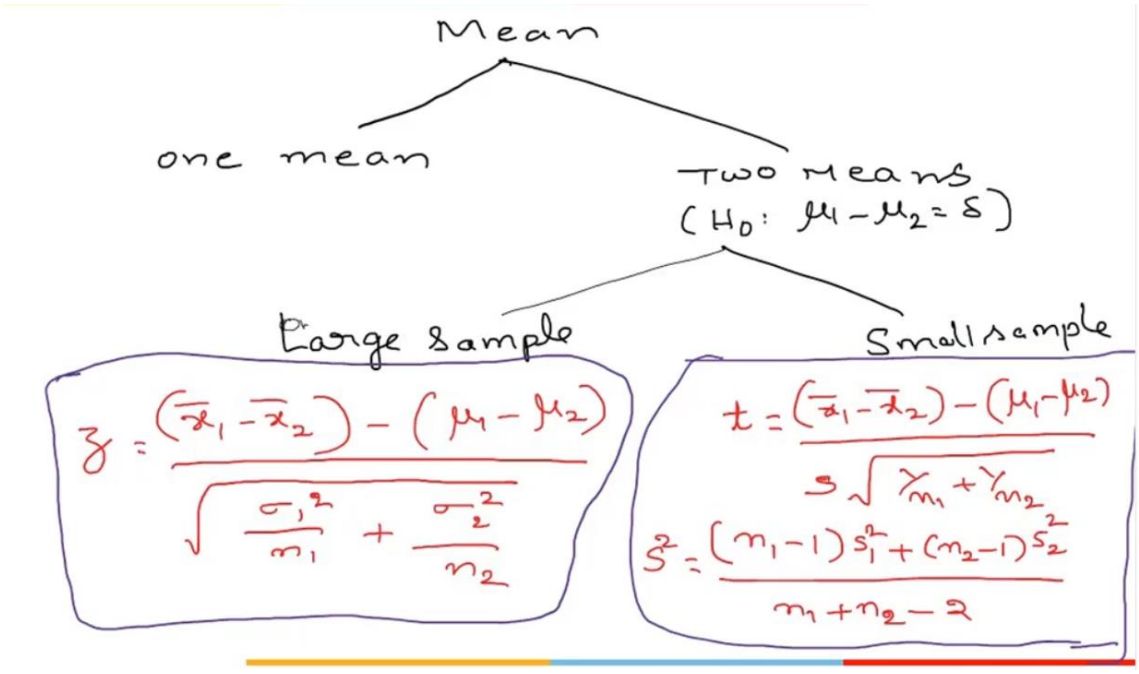
Critical region – depending on the boundaries – compare & conclude – accept or reject







Null & alternate hypothesis – both are compliment, Two tail test - > or <

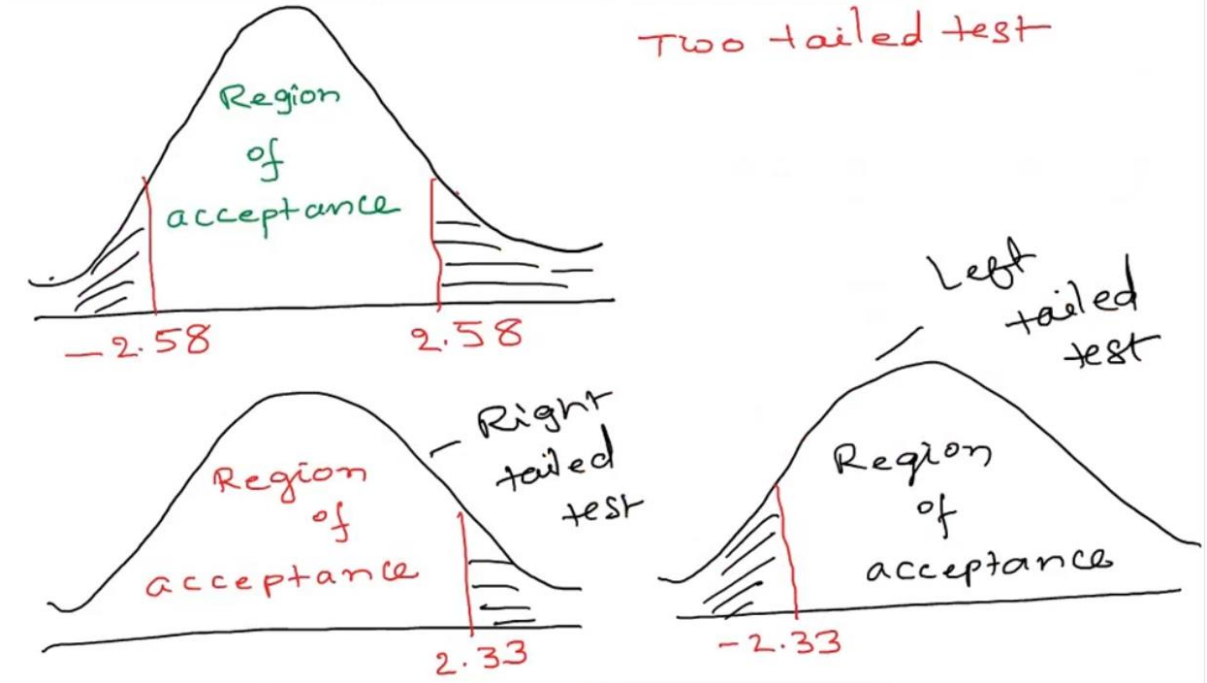


Two means problem – Mu1 – Mu2 = Delta

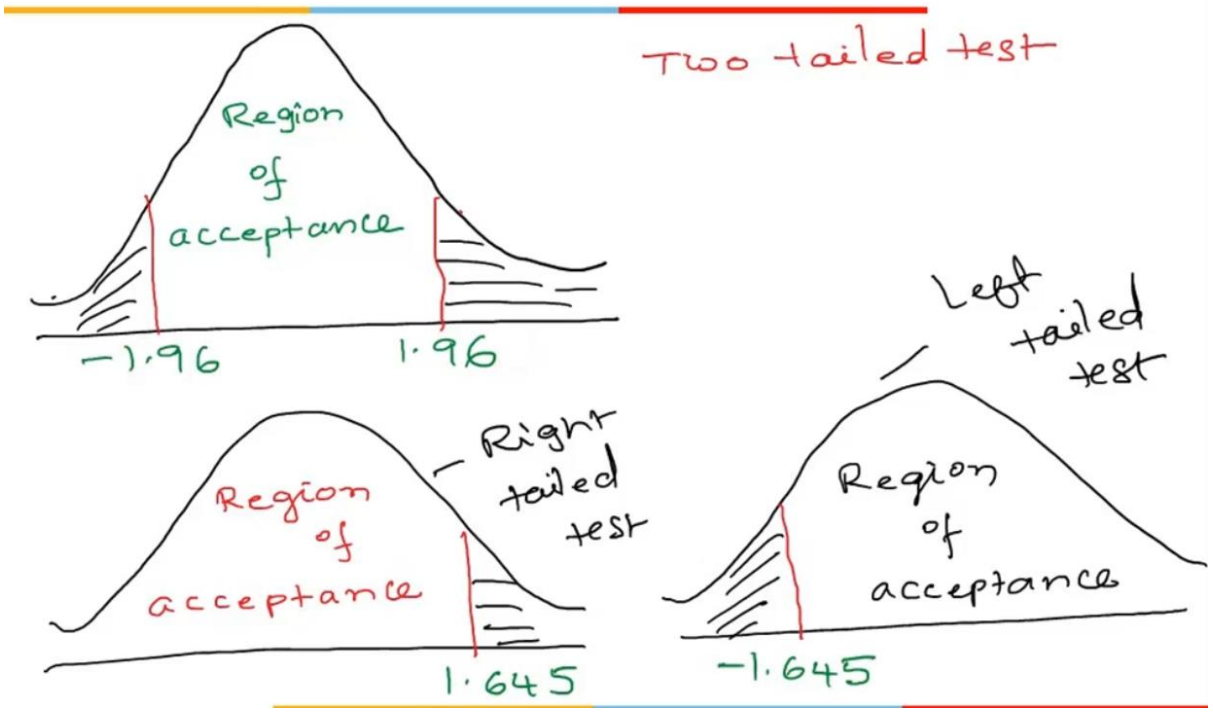
useful table values (Z- distribution)

	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

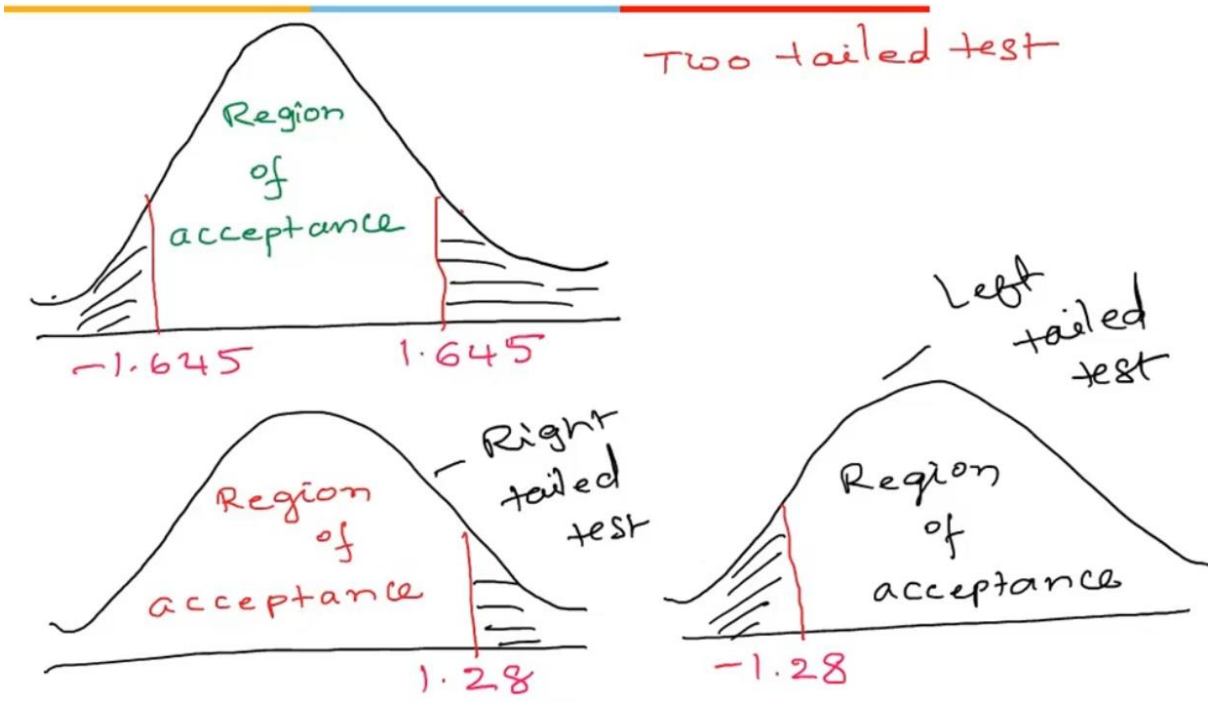
$\alpha = 1\%$ (or 0.01)



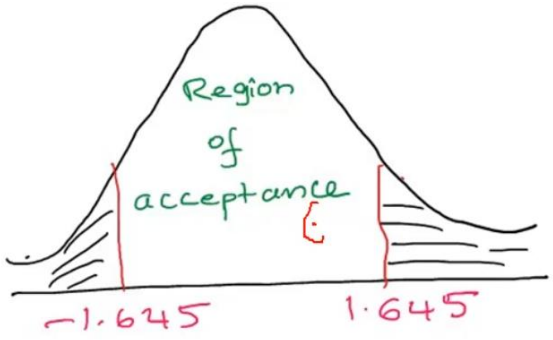
$\alpha = 5\%$ (or 0.05)



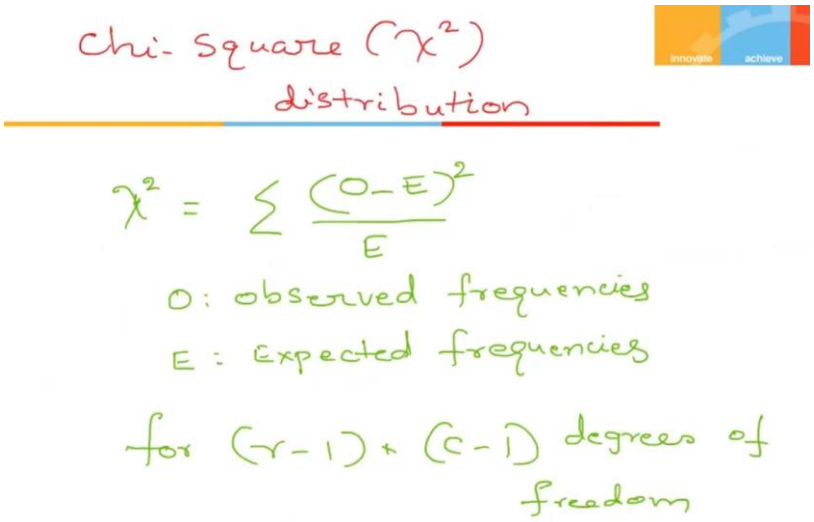
$\alpha = 10\%$ (or 0.1)



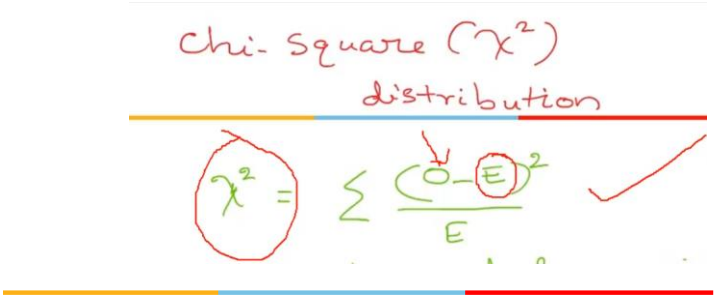
85% confidence given – how to calculate?



80% - 0.8 → leftover – 0.2 - We know the probabilities – we need to find z values – Alpha val



Something related to proportion or frequencies – smoker – how many are having cancer? – Is there relation between smoking & cancer?



1) correlation

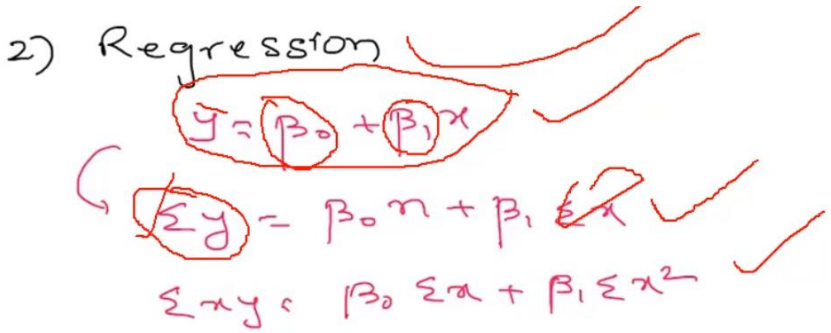
2) Regression

$y = \beta_0 + \beta_1 x$

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

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

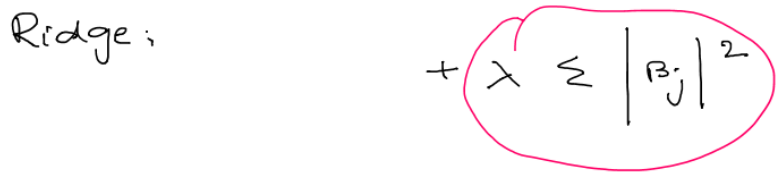
Correlation – relation between two variable – related or not, Regression – find or establish that relation, Regression – linear (most of the times) or multilinear regression



R^2 = coeff of determination

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

R^2 – exact relation between two variables → strength



Logistic regression

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

$y = \beta_0 + \beta_1 x$

or

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

When to go for linear – numerical value & estimate, x = 8 then what is y = ?

Logistic – all the data – set of parameters – predictive analysis – yes or no

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 = $\frac{S_1^2}{S_2^2}$ → $\gamma_1 - 1$, $\gamma_2 - 1$

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

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

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

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

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

→ SSE = SST − SSTR

Similar to testing of hypothesis – multiple means → deal with data – calculation between groups

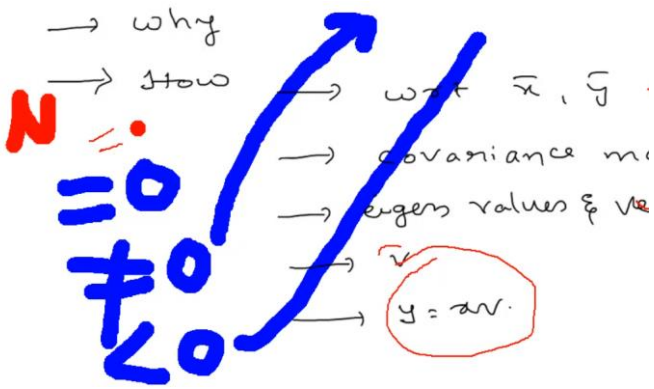
Principal comp. Analysis

- when
- why
- how
  - wrt $\bar{x}$, $\bar{y}$
  - covariance matrix
  - eigen values & vector
  - V
  - y = aV.

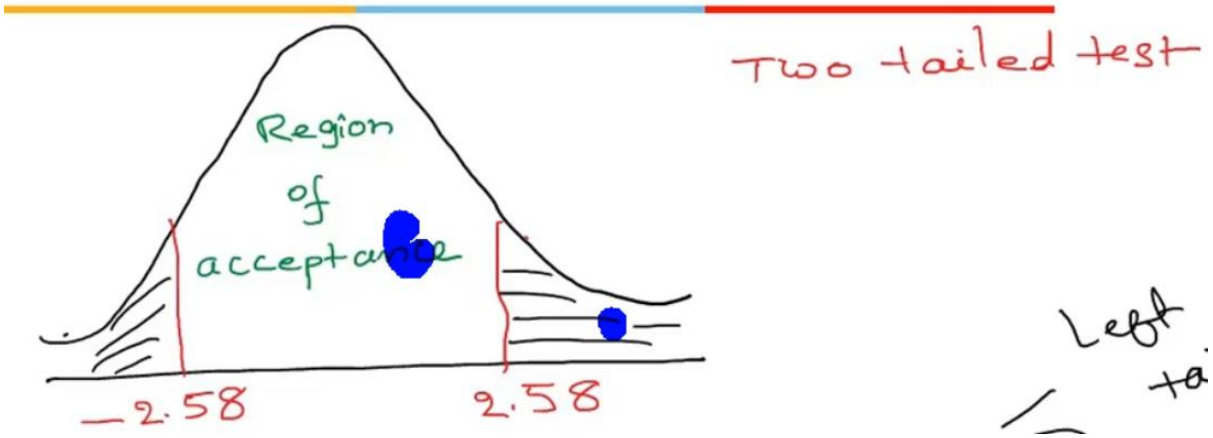
When we want to reduce dimensions – major dimension we want to find - PCA

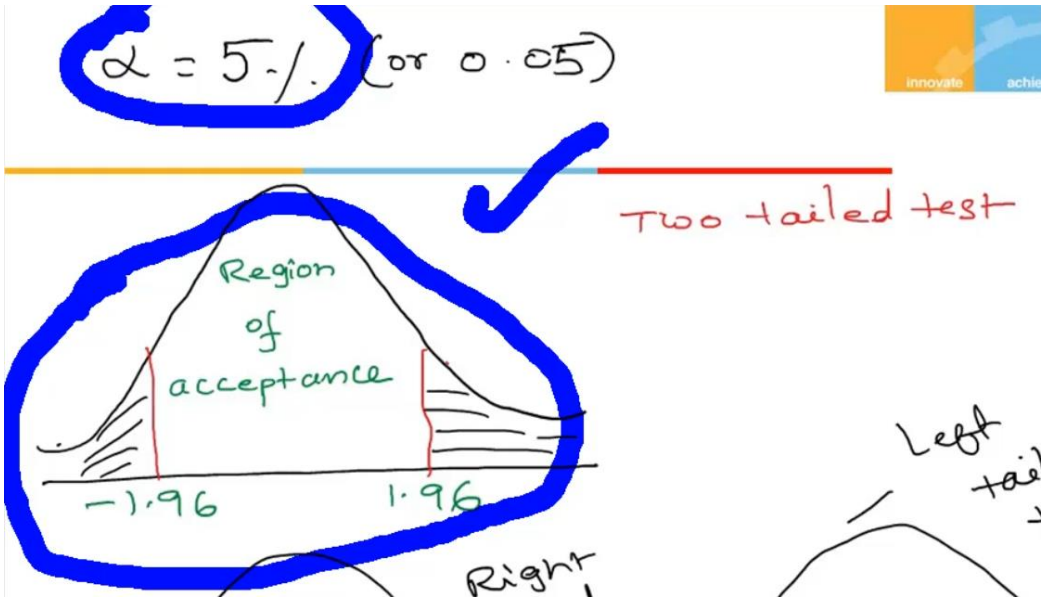
Eigen values – sort them in decreasing order – principal component

Choosing right tool is very important – given a scenario



Mu1 – Mu2 != 0 → two tailed test





Important topics – Regression, Correlation, TSA (MA & w/o MA), ANOVA, Testing of Hypothesis, Conclusions or which tool to use

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

innovate achieve lead

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

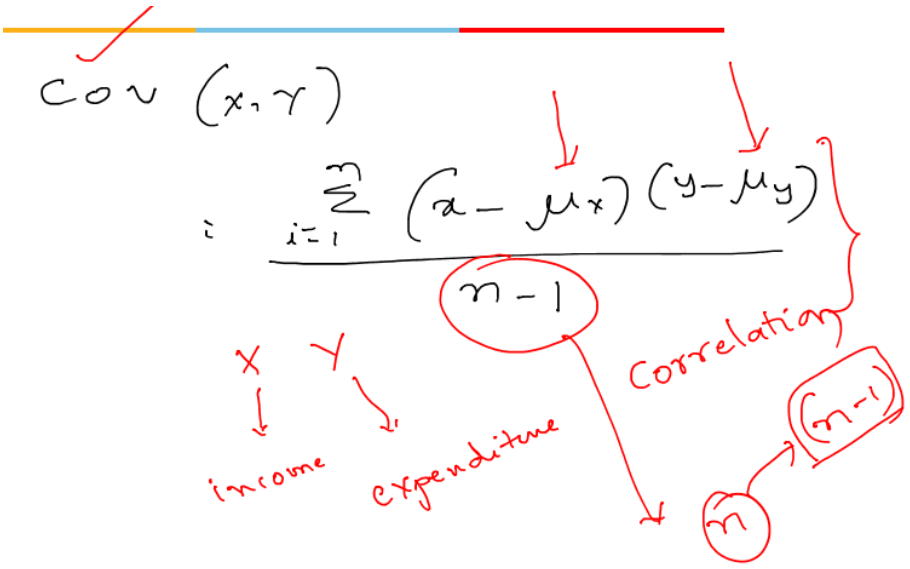
$$= \sum_x \sum_y (x-\mu_x)(y-\mu_y)\,P(x,y) \quad \text{if discrete}$$

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

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

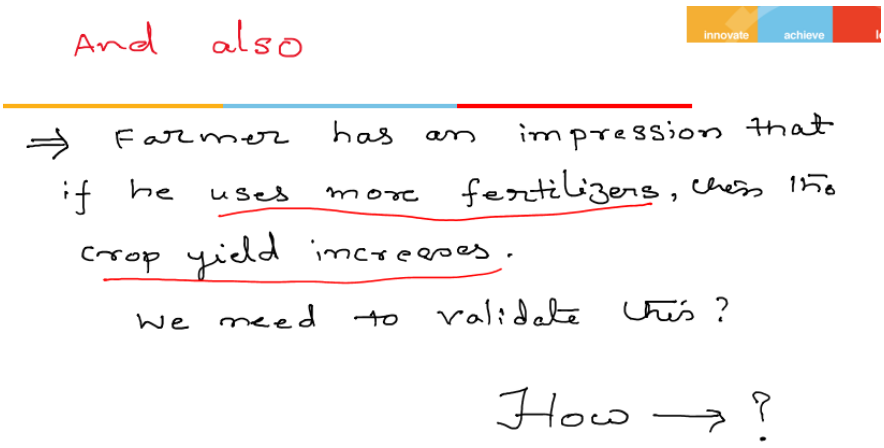
$P(x,y)$: joint p.d.f

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

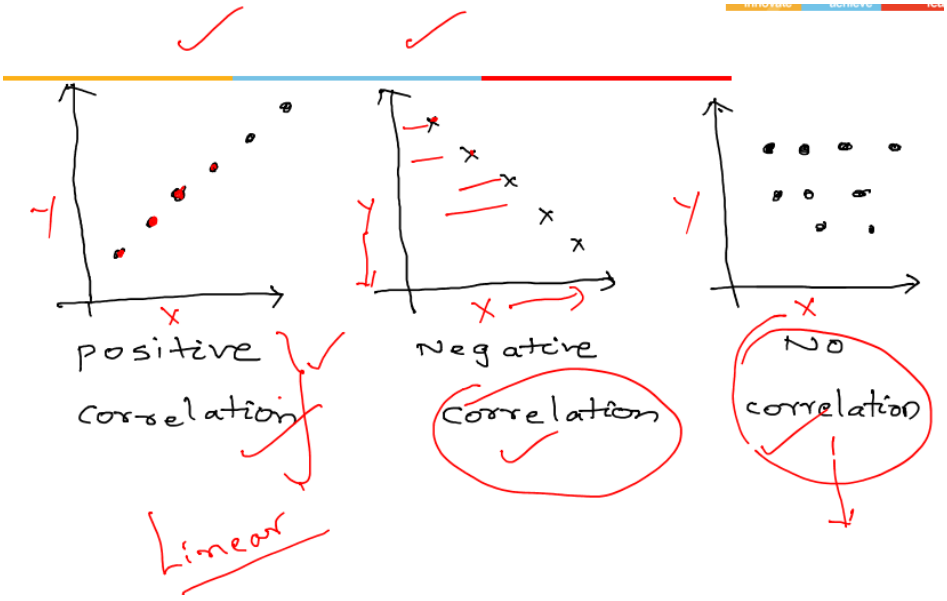


Covariance – w.r.t. mean how these two variables behaves, μ_x – mean of x

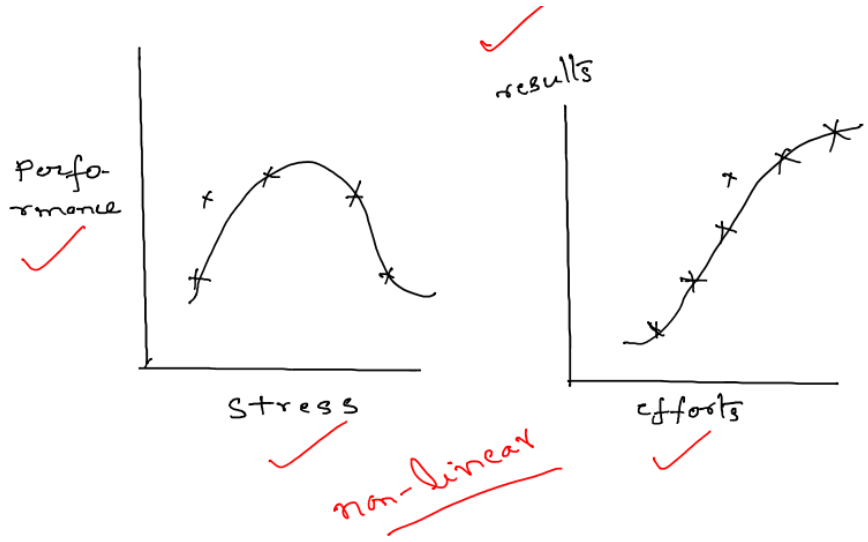
Relation between x & y – cov(x, y)



We want to validate whether crop yield depends on fertilizers? – Relation between x & y



Positive correlation can have linear or non-linear relation also – hyperbola or parabola



Coefficient of correlation:

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

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

$Y = y - \bar{y}$

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

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

Try to quantify relation between two variables

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

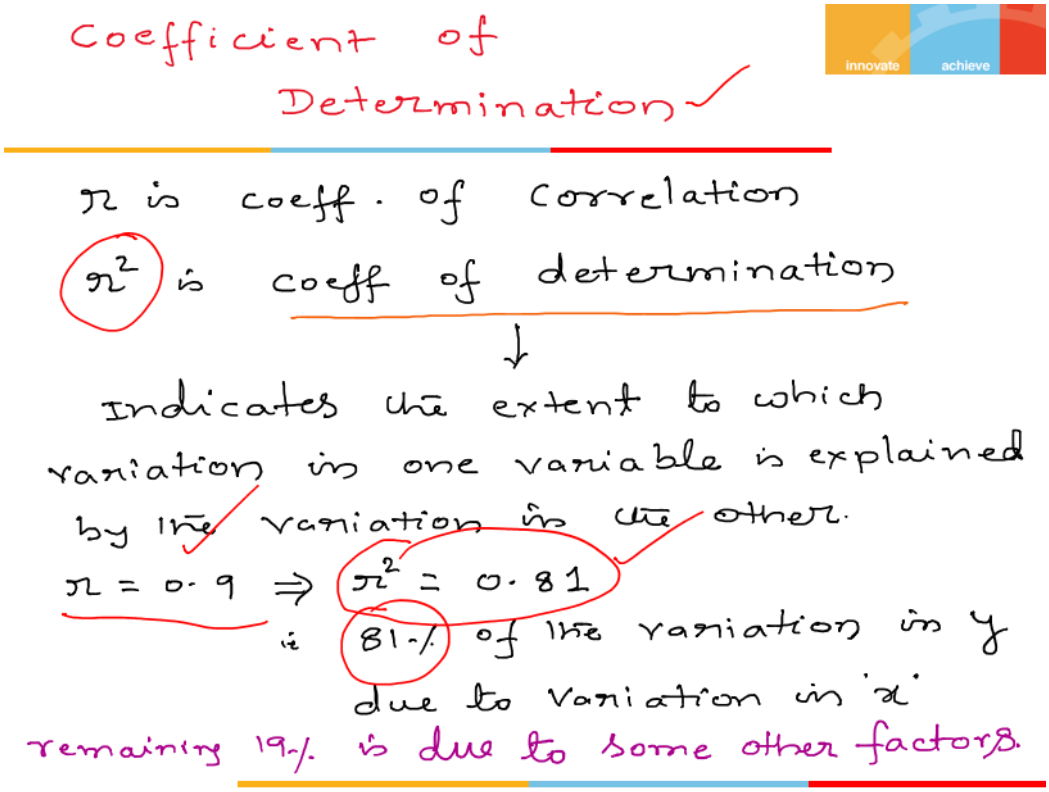
→ +ve correlated

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

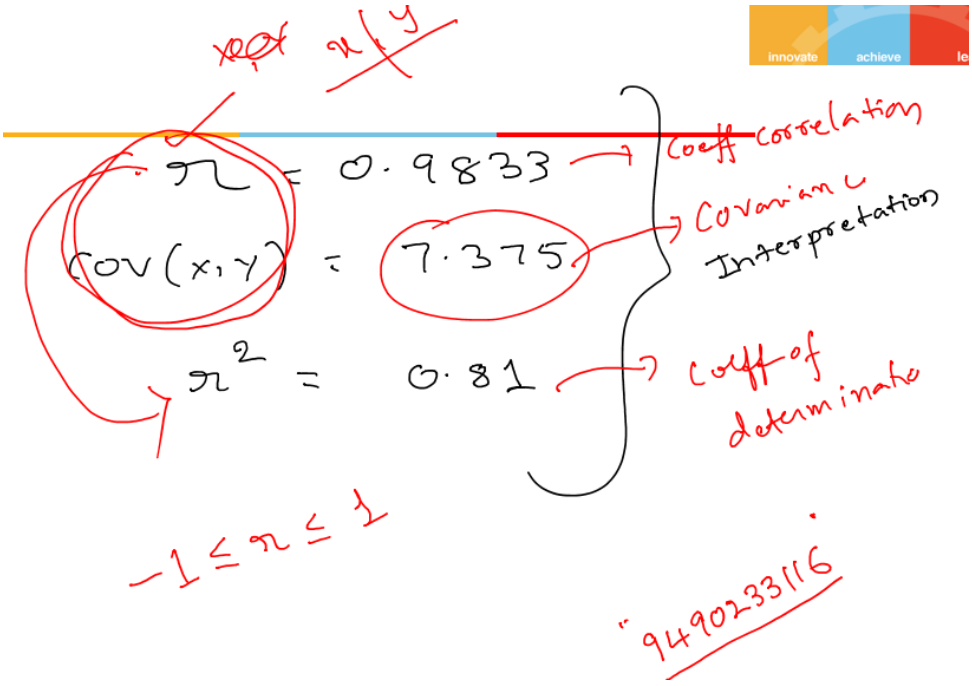
$$\text{Cov}(x,y) = \frac{\Sigma XY}{n-1}$$

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

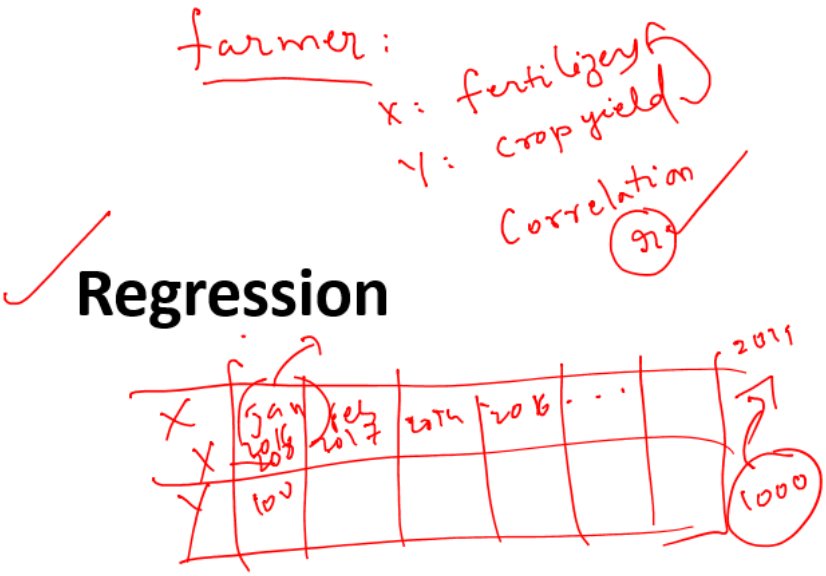
$$= 7.375$$



Whenever there's change in x – there's 81% chance that y changes – other 19% change in y due to some other factors



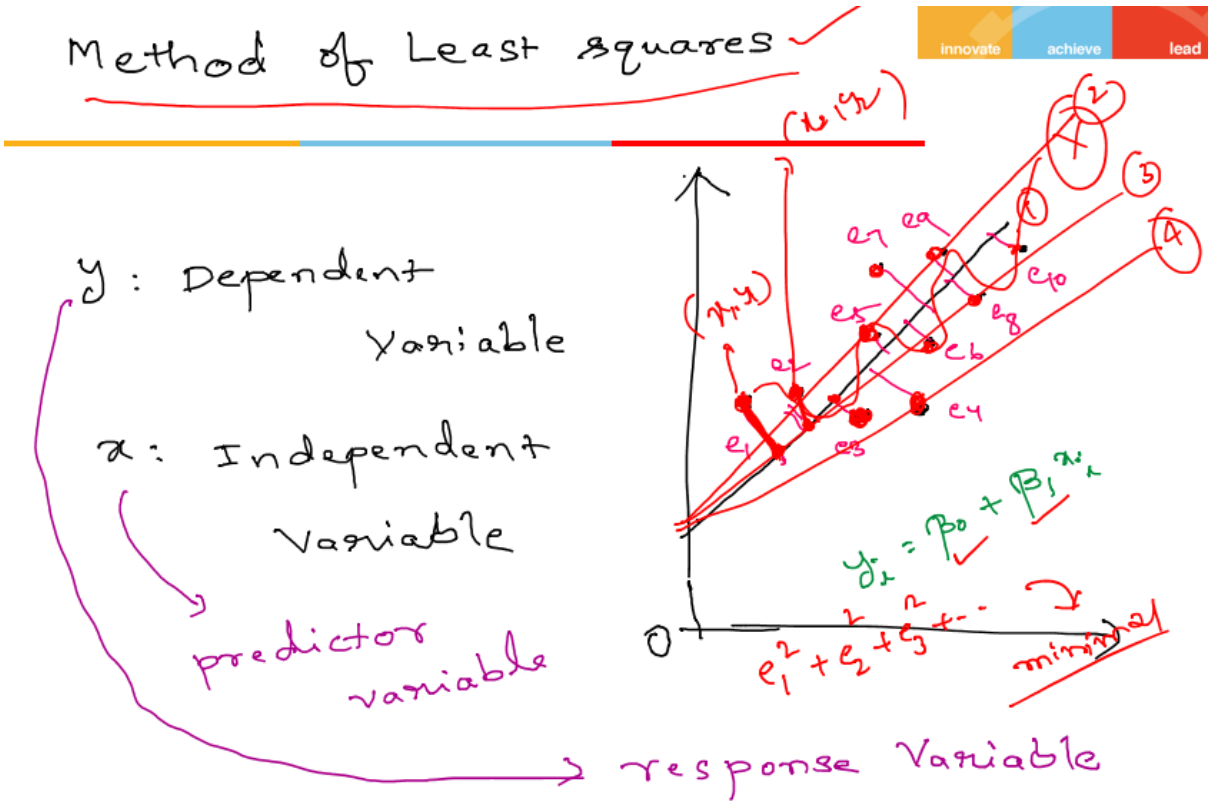
Covariance also helps us to understand relation between x & y – positive – two variables are related positively – how strongly related we can't say → Coefficient of determination – value - strength of the relation – similarly Coefficient of correlation → -1 <= r <= 1



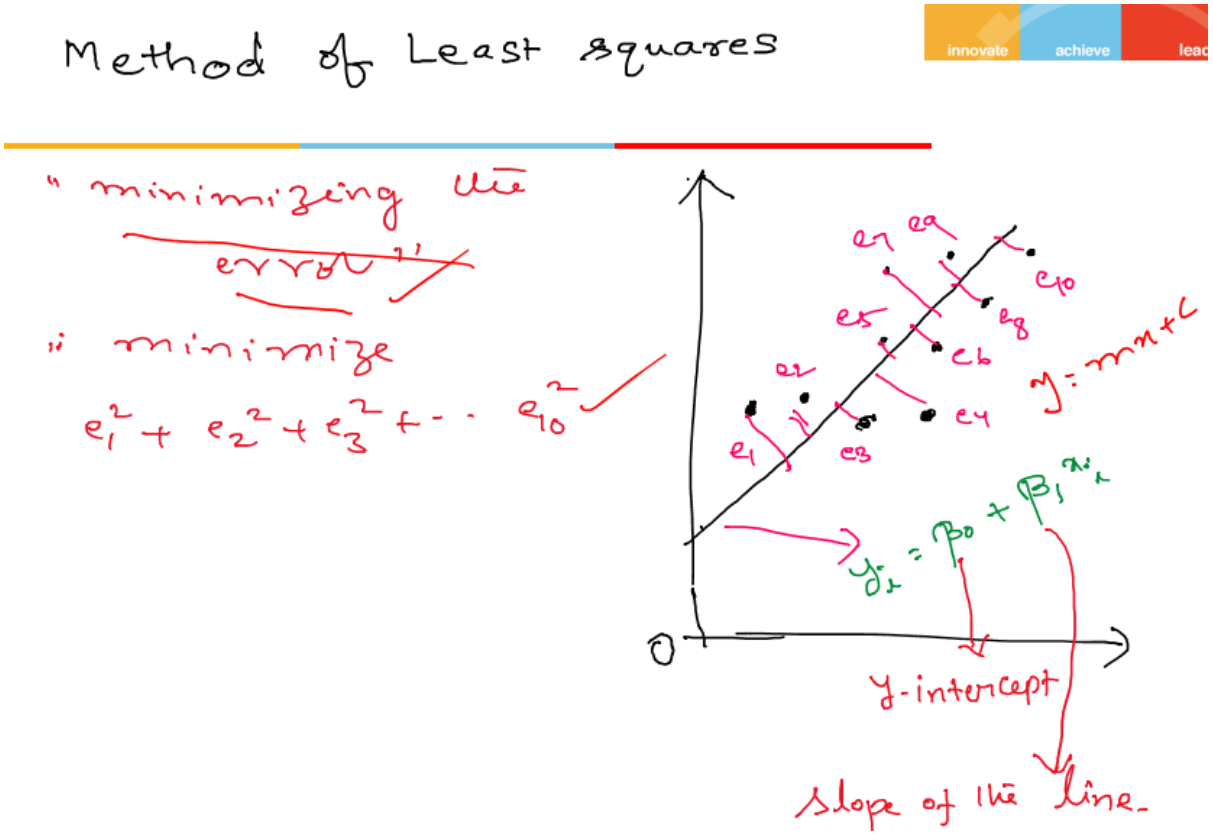
Next year – we want to have this much crop – how much fertilizer we should use?

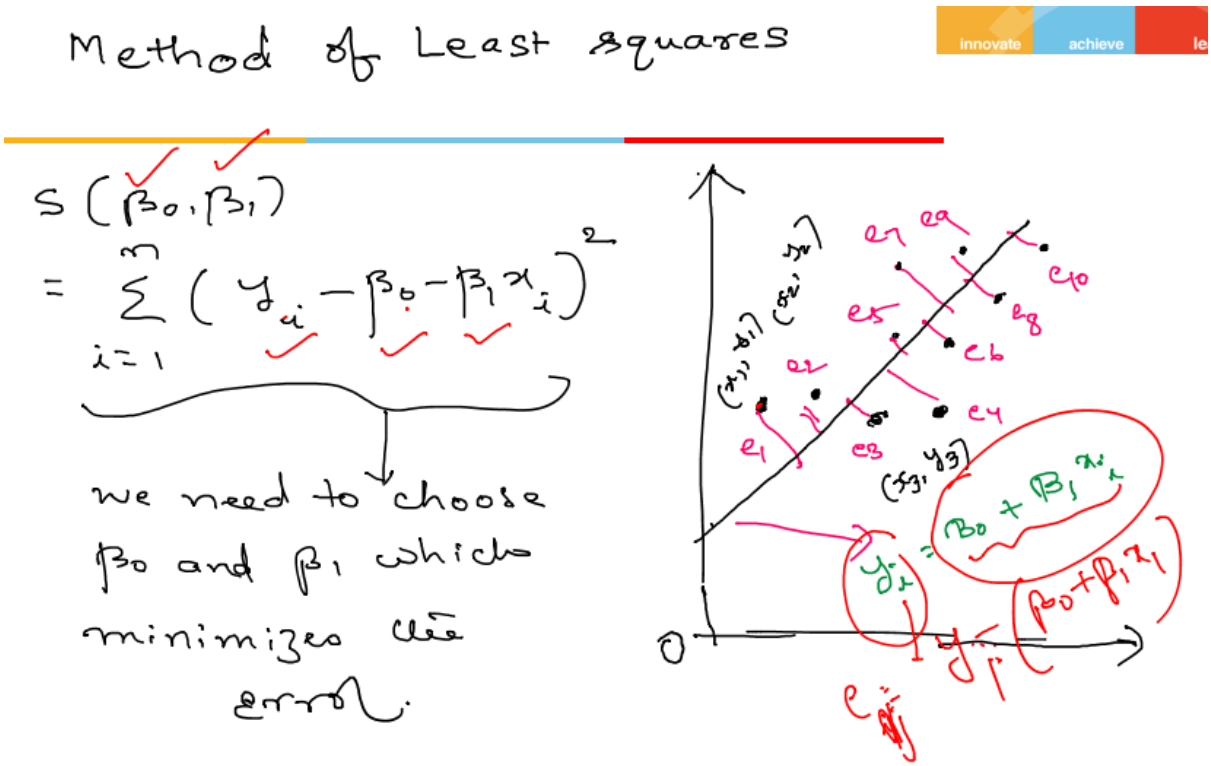
Correlation	Regression
→ Measuring strength or degree of the relationship between two variables	→ Having an algebraic equation between two variables
→ no estimation	→ estimation
→ both variables are independent	→ one is dept variable and other indept variables

Some relations are easy to find (direct visible) – others are not

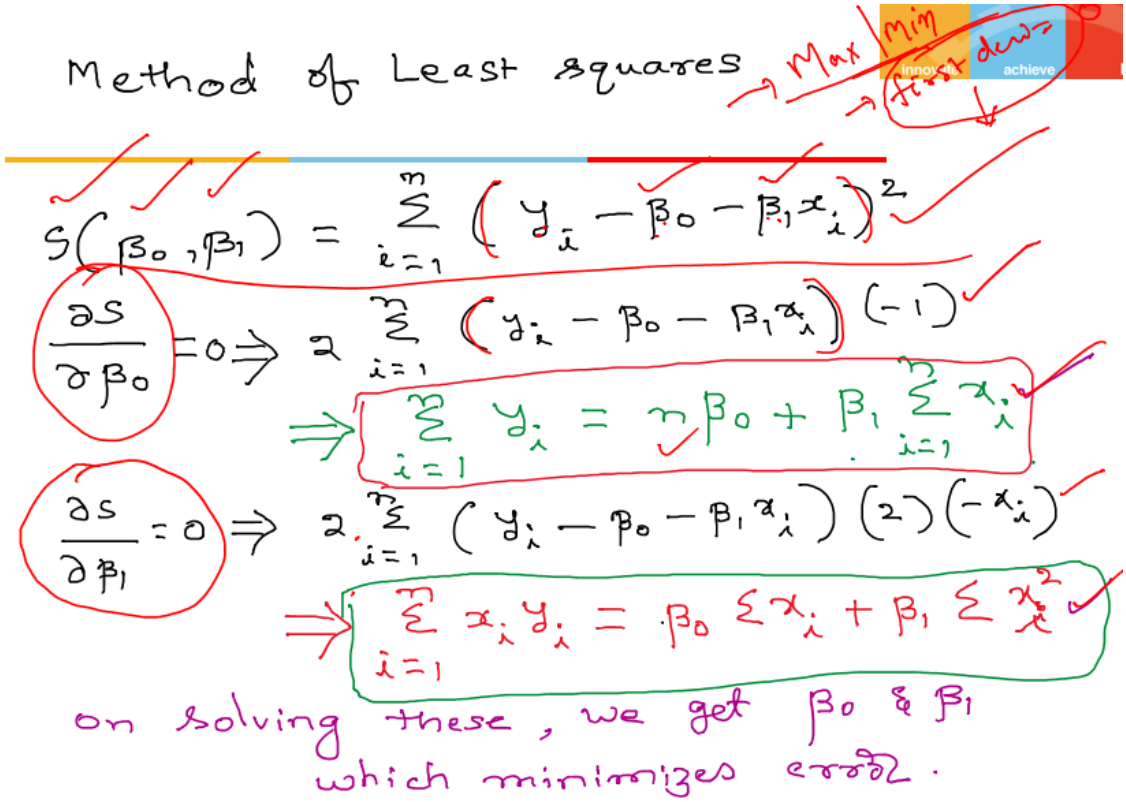


Many lines – which line is more nearer – optimal line – nearner to original points – line is best fit when error is minimized





Problem of optimization – minimize **S** w.r.t. β_0 & β_1 – choose appropriate values



Stationary points – multiple variables – take partial derivative w.r.t. independent variables = 0 – solve – function becomes minimum or maximum

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

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

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

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

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

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

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

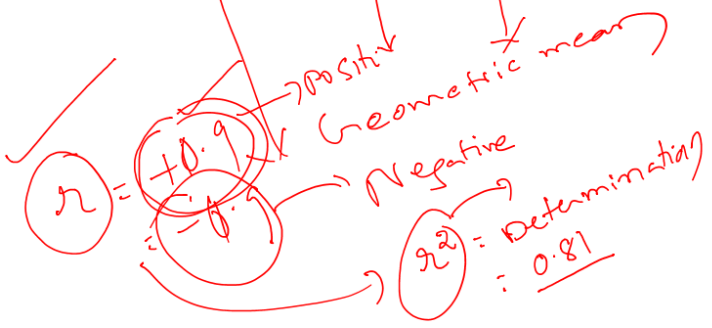
b_{yx} : Regression coeff of y on x

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

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

Correlation coefficient

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



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
Σ 40	Σ 56	Σ 524	Σ 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

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

$= 0.1248 \simeq 12.48\%$

innovate achieve

Example:

Consider the following data

x	1	2	4	0
y	0.5	1	2	0

Fit a linear regression line

Estimate y when $x = 5$.

innovate achieve lead

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

$\beta_0 = 0$

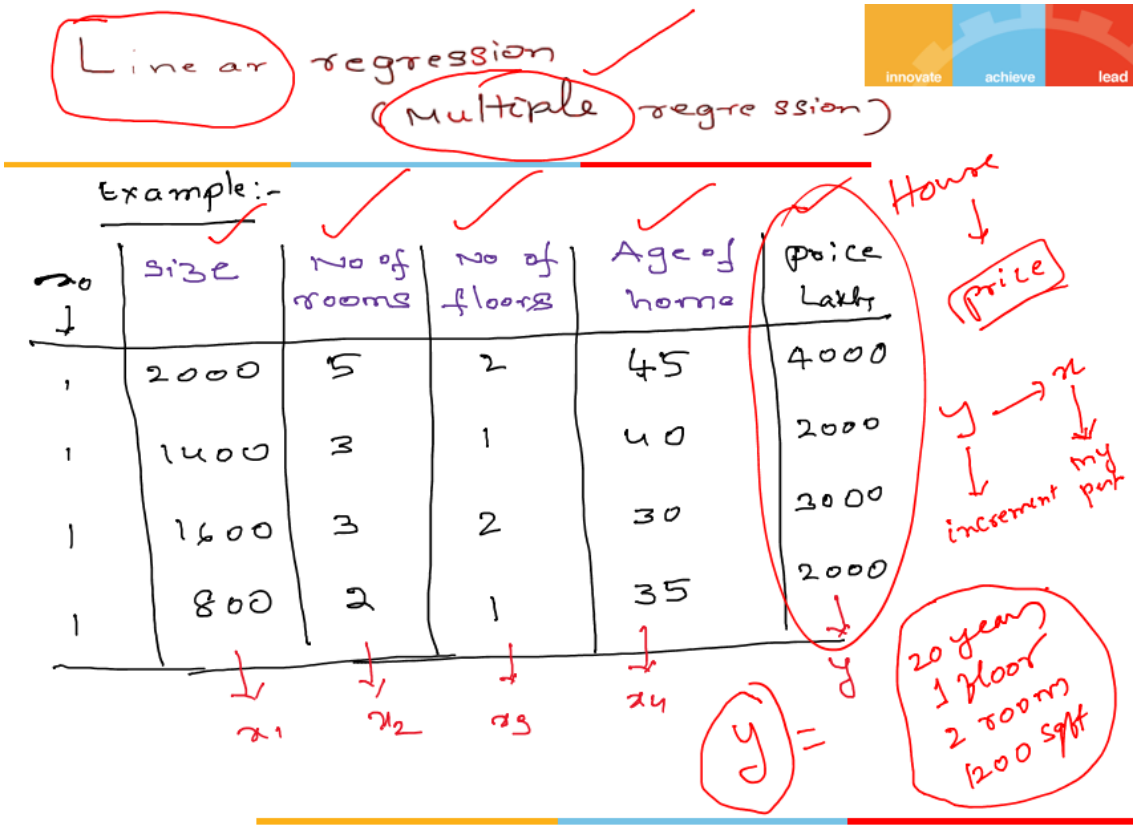
$\beta_1 = 0.5$

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

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

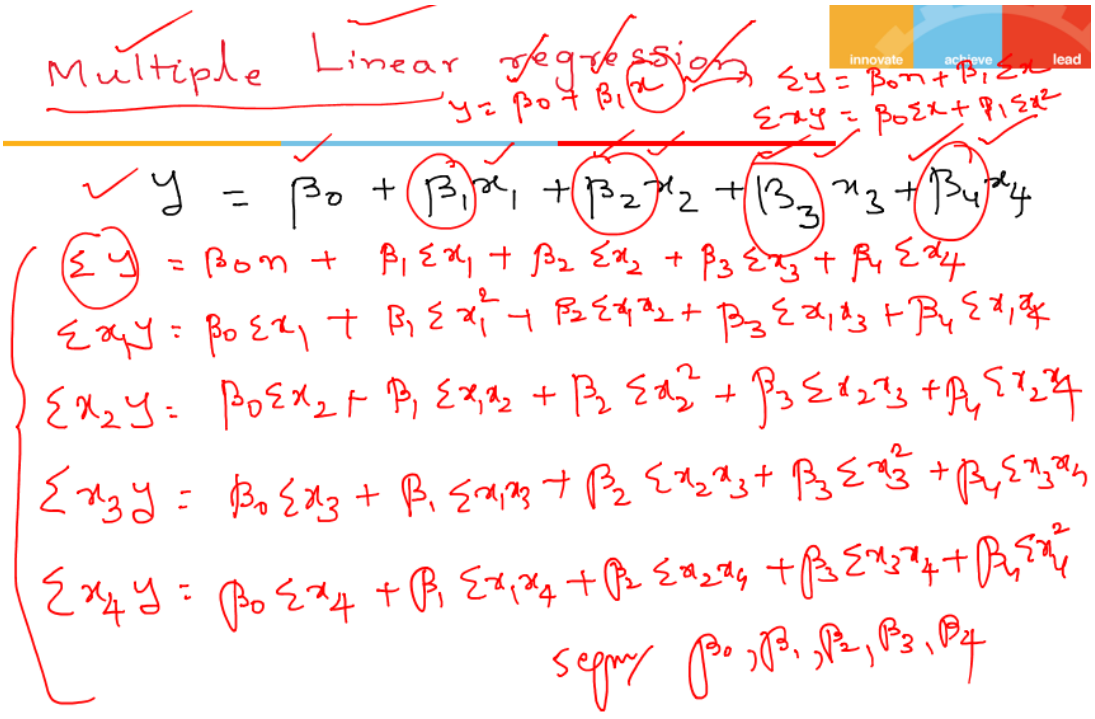
$= 0.25$

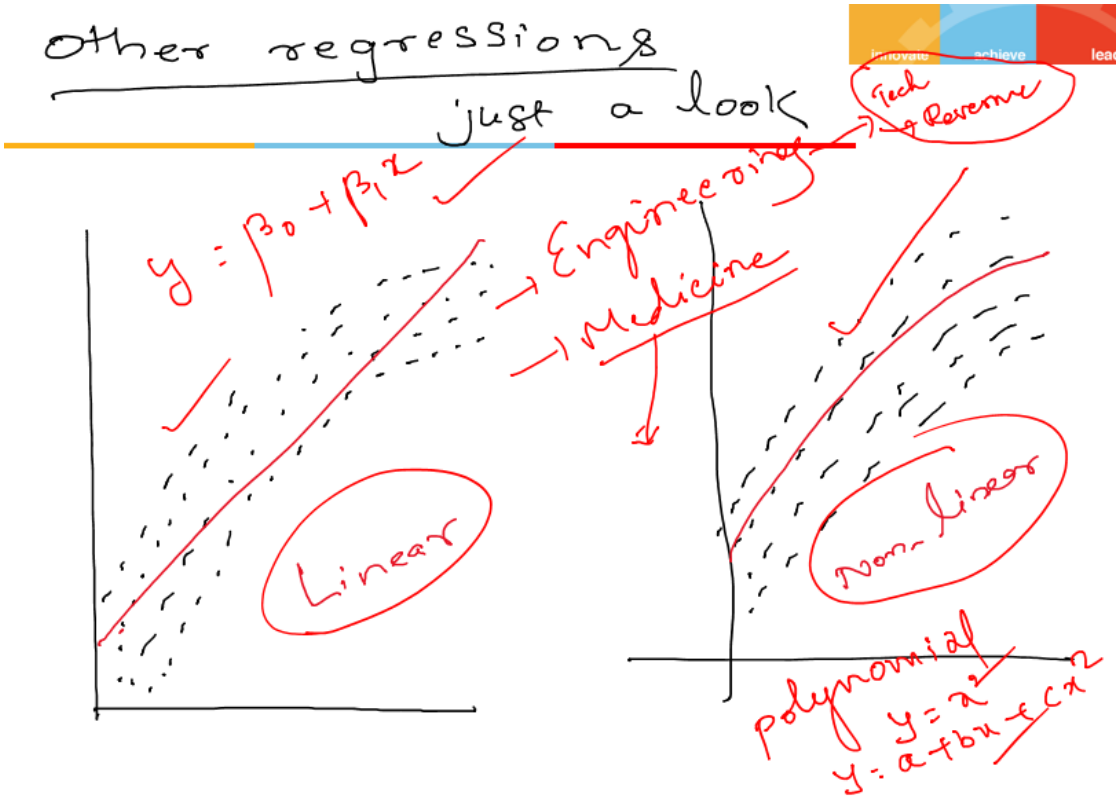
One dependent variable & one independent variable here



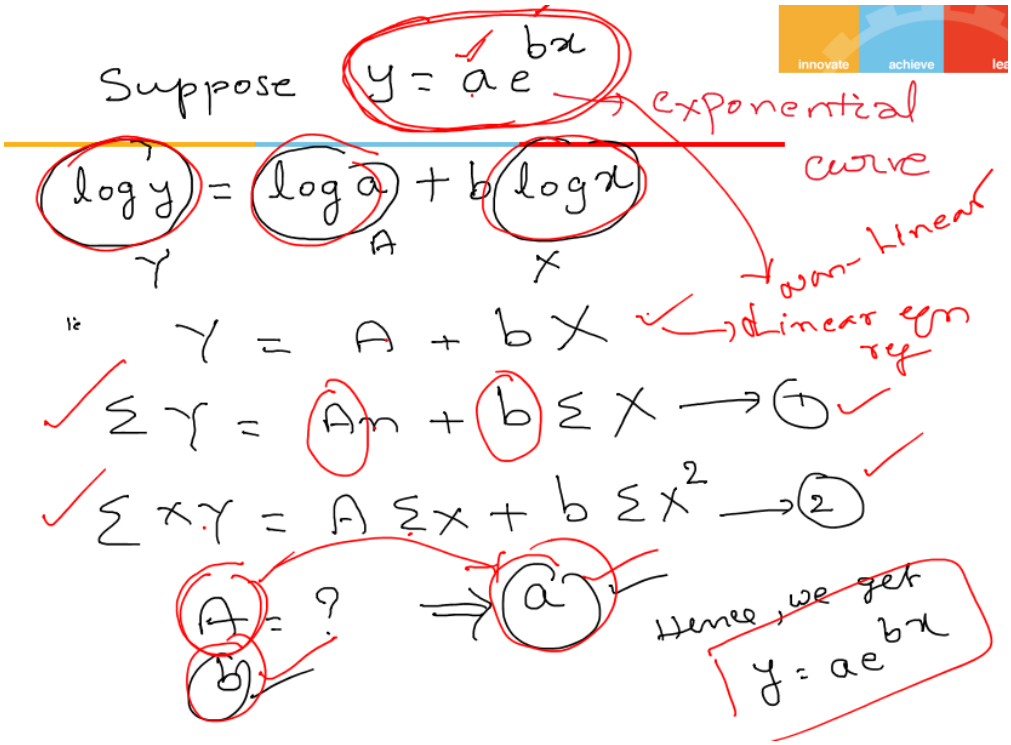
Real world scenario – that is not the case – many independent variables – increment → company profit, performance of the employee, project revenue etc.

y is a dependent variable which depends on many independent variables





Sometimes linear regression is not optimal – we have to go for nonlinear regression



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

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

$Y = A + bX$

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

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

$A \to a$, $b \to b$

$y = ax^b$

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

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

$\hat{Y} = X \beta$

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

$Y = BX$

$A X = B$

$X = A^{-1}B$

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

$$x^T x \hat{\beta} = x^T y \rightarrow \text{normal eqns}$$

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

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

β_i

$AX = B$

$X = A^{-1}B$

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

→ If $x^T x$ 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:

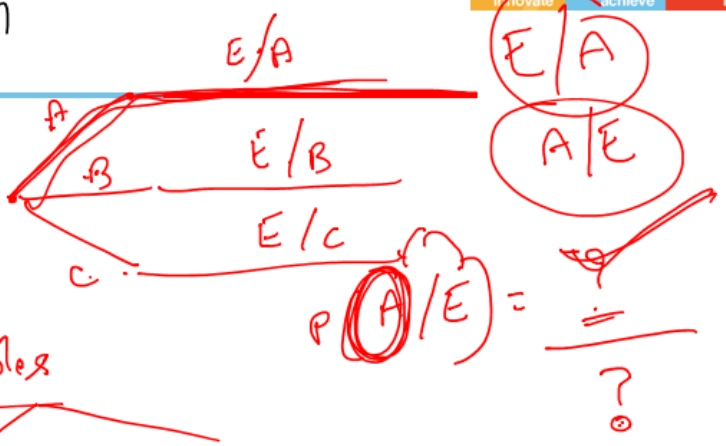
$$P(A/B) = \frac{P(A \cap B)}{P(B)}$$

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

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

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 $P(x) = {}^{n}C_x\, p^x q^{n-x}$, $x = 0,1,2 \cdots n$

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

→ Mean → two → z, t

→ Proportion → χ^2 - distribution