Session 8
Predictive Analytics - Correlation

Agenda

- Problems on Chi Square distribution
- Correlation
- Revision of the topics covered till session 7

Example:-
Consider the following data

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

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

???

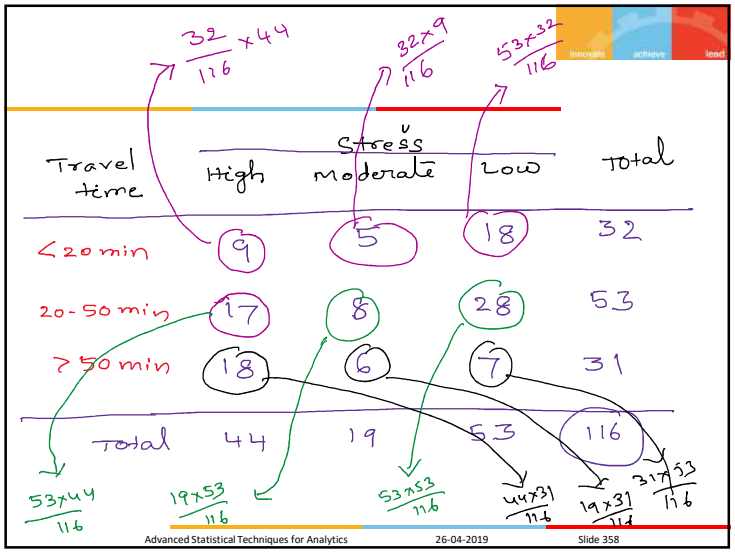
Chi-Square (χ^2) distribution

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

O: observed frequencies

E: Expected frequencies

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



$$\chi^2 = \sum \frac{(O-E)^2}{E} = \sum \frac{(9-12.14)^2}{12.14} \cdots = 9.836$$

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

Calculated $\chi^2 = 9.836$

Let $\alpha = 0.01$

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

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

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

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

H_0 accepted

Example

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

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

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

observed frequency

$E = \frac{180}{300} \times 100 = 60$

$E = \frac{180}{300} \times 200 = 120$

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

$\frac{120}{300} \times 100 = 40$

$\frac{120}{300} \times 200 = 80$

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

$= 14.063$

From χ^2-table at 0.05 LOS

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

$(2-1) \times (2-1)$

$= 1$

$= 3.841$

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

$\chi^2 = 14.063 > 3.841$

Reject H_0 $\rightarrow$

Correlation

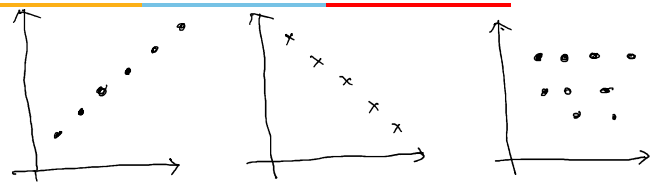
Correlation

→ Sales of a company and Expenditure on advertisement

→ Price and Demand of a product

→ Inflation and Gold Price

→ IQ and performance in Entrance.



positive correlation

Negative correlation

No correlation

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 "

Coefficient of correlation:

$$r = \frac{cov(x,y)}{\sigma_x \sigma_y} = \frac{\Sigma dx\, dy}{\sqrt{\Sigma dx^2 \; \Sigma dy^2}}$$

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

$dy = y - \bar{y}$

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

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

Example - 1

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

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

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

r

x	dx $x-5$	dx^2	y	dy $y-14$	dy^2	$dxdy$
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

$$r = \frac{\Sigma dx\,dy}{\sqrt{dx^2 \Sigma dy^2}} = \frac{59}{\sqrt{60 \times 60}} = 0.9833$$

Coefficient of Determination

r is coeff. of correlation

r^2 is coeff of determination

↓

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

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

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

remaining 19% is due to some other factors

Revision

Probability

1) Probability $P(A) = \frac{m}{n}$

2) $0 \leq P(A) \leq 1$

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

4) A and B are mutually exclusive events, then
$P(A \cup B) = P(A) + P(B) - P(A \cap B)$

5) If A and B are independent events
$\Leftrightarrow P(A \cap B) = P(A) \cdot P(B)$

6) Set theory
↓
happening of both $A \cap B$
atleast one ... $A \cup B$
not happening --- $\bar{A}$ etc

Conditional probability

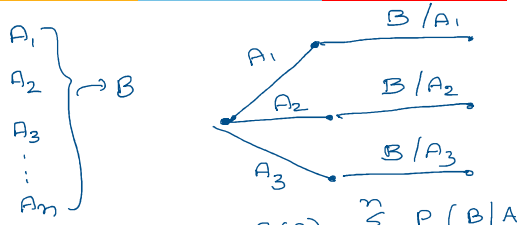
7) $P(A/B) = \frac{P(A \cap B)}{P(B)}$

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

8) $P(A \cap B) = P(A/B)\,P(B)$
$= P(B/A)\,P(A)$

9) If A, B are indept. events,
$P(A/B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A) \cdot P(B)}{P(B)} = P(A)$
Similarly $P(B/A) = P(A)$

Baye's theorem



$\therefore$ Total prob: $P(B) = \sum_{i=1}^{n} P(B/A_i)\,P(A_i)$

$P(A_1/B) = \frac{P(B/A_1)\,P(A_1)}{\sum_{i=1}^{n} P(B/A_i)\,P(A_i)}$

Random Variables

- Discrete: $X = 1, 2, 3, 4$ → $P(x)$ prob. dist. fn $p(x)$
  - (i) $0 \le p(x) \le 1$
  - (ii) $\sum p(x) = 1$
- Continuous: $x \in (1, 3)$ → $f(x)$ prob. density function $f(x)$
  - (i) $0 \le f(x) \le 1$
  - (ii) $\int f(x)\,dx = 1$

Prob. distributions

1) Binomial distribution.

$$P(x) = {}^{n}C_{x}\, p^{x} q^{n-x}$$

$n = 0, 1, 2, \ldots x$

2) Poisson distribution

$$P(x) = \frac{e^{-\lambda} \lambda^{x}}{x!}$$

$x = 0, 1, 2, \ldots$

3. Normal distribution

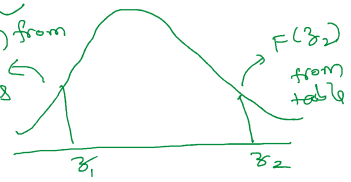
$$P(x_1 \le x \le x_2) \qquad z = \frac{x - \mu}{\sigma}$$

$$P(z_1 \le z \le z_2) = F(z_2) - F(z_1)$$

$F(z_1)$ from tables

$F(z_2)$ from table

$F(-z) = 1 - F(z)$



Testing of Hypothesis

- Mean
  - one mean
    - Large → Z
    - small → t
  - Two means
    - Large → Z
    - small → t
- proportion
  - one → Z
  - Two → Z
  - several → χ^2 - dist.

L- 9: Predictive Analytics

Agenda

- Introduction to regression
- Method of least squares
- Simple linear regression
- Multiple linear regression

Regression

Regression :-

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

when x = 7 : y = ?

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

when x = 7, y = ?

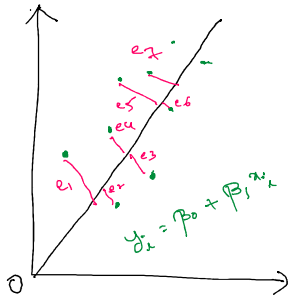
Method of Least squares

y : Dependent Variable

x : Independent Variable

predictor variable

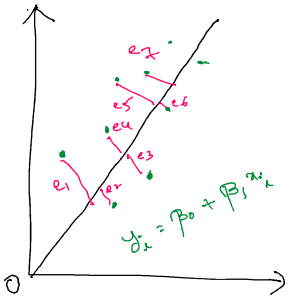
response Variable



Method of Least squares

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

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



Method of Least squares

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

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

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

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

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

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

Linear regression

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

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

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

Normal equations.

Matrix Approach:

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

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

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

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

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

Find β to minimize

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

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

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

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

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

Linear regression
(Multiple regression)

Example:-

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

Example:

Consider the following data

x	1	2	4	0
y	0.5	1	2	0

Fit a linear regression line

Estimate y when $x = 5$.

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

$y = \beta_0 + \beta_1 x$

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

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

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

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

on solving these

$\beta_0 = 0$

$\beta_1 = 0.5$

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

when $x = 5$, $y = (0.5)5 = 0.25$

BITS Pilani

Session 10: Predictive Analytics(Continued)

Agenda

- Linear Regression(Revise)
- Model validation
- Ridge and lasso models
- Assumptions of Linear regression
- Logistic regression

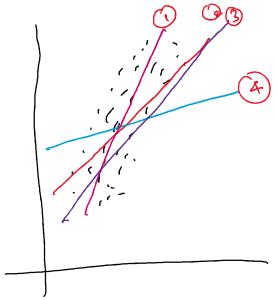
Classical Linear Regression (OLS)

- 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:

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

Method of Least squares Errors

which one is the best ?



$y = \beta_0 + \beta_1 x$

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

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

$\beta_0 = ?$, $\beta_1 = ?$

then $y = (\beta_0) + (\beta_1) x$.

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

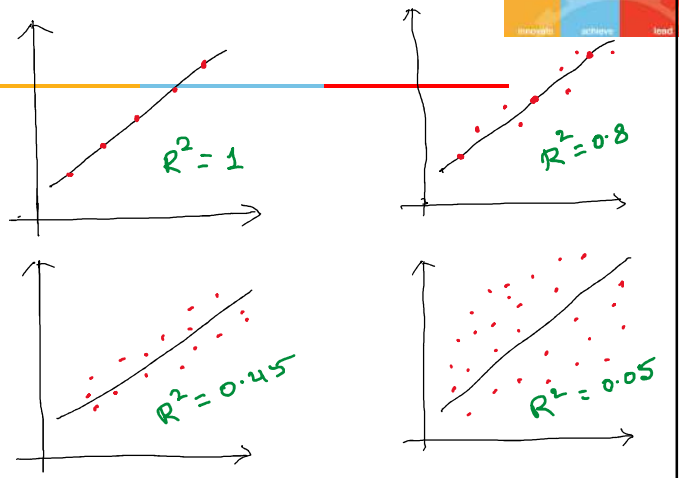
1) Coefficient of determination (R^2)
2) Residual Standard error (RSE)

RSS → Residual sum of squares

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

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

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



R – Squared vs Adjusted R - Squared

- In multiple regression, adjusted R – squared is better metric than R – squared asses the goodness of fit of the model
- R – squared always increases if additional variables are added into model , even if they are not related to the dependent variable

Regularization

- Over fitting can be solved with regularization
- Regularization can be done by putting constraints on the coefficients and variables.
- LASSO: Least Absolute Shrinkage and Selection Operator
  Some coefficients can be dropped(i.e become zero)
- RIDGE: The coefficients will approach zero, but never dropped

Lasso & Ridge

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

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

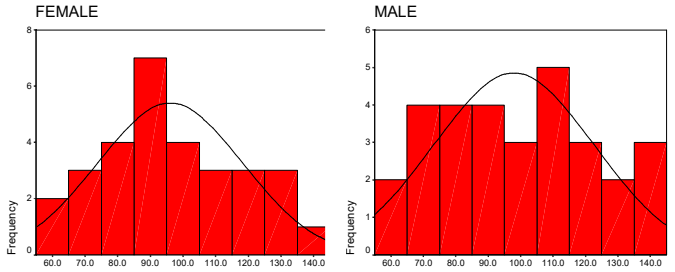
Assumptions in Regression Analysis

Assumptions

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

- 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

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

Heteroscedasticity

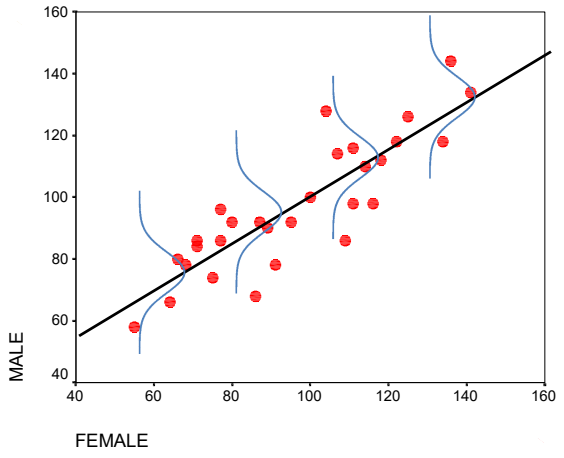
This assumption is a about heteroscedasticity of the residuals

- Hetero=different
- Scedastic = scattered

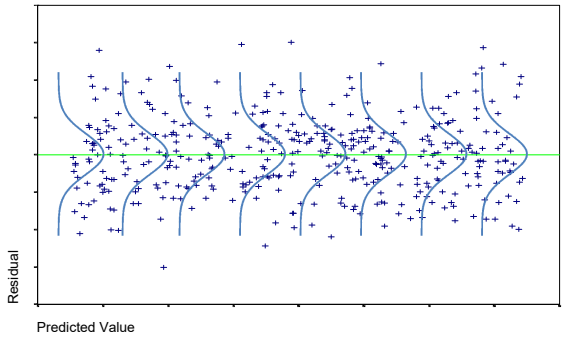
We don't want heteroscedasticity

- we want our data to be homoscedastic

Draw a scatterplot to investigate

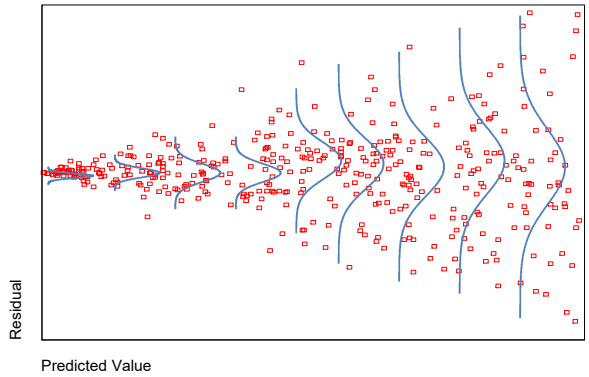


Good – no heteroscedasticity



417

Bad – heteroscedasticity



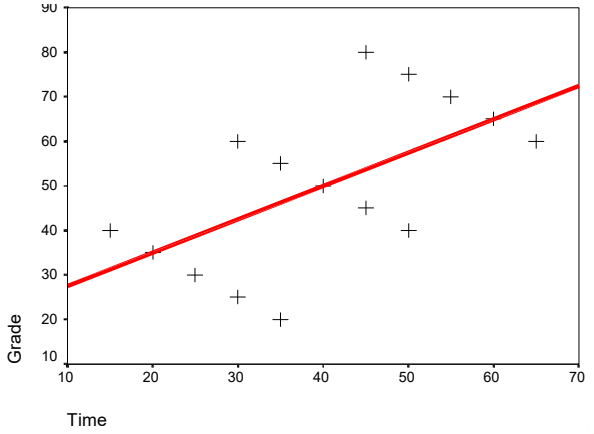
418

Assumption 3:
The Error Term is Additive

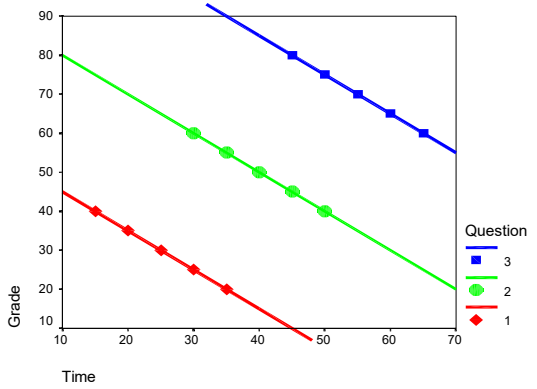
Assumption 4: At every value of the dependent variable the expected (mean) value of the residuals is zero

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

- Result, with line of best fit



Now somewhat different



423

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

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

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

Multicollinearity

Correlation Matrix

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

VIF(Variance Inflation Factor)

VIF(Variance Inflation Factor)

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

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

If VIF = 1 then Variables are not correlated

1< VIF < 5 then the variables are moderately correlated

VIF > 5 then highly correlated and need to be eliminated from the model

Logistic Regression

Consider the following data

Customer	Age	Income (Lakh)	Gender	response	
X	30	25	M	Yes	1
Y	45	40	F	No	0
Z	50	20	M	Yes	1
A	62	80	F	No	0
B	75	50	M	No	0
C	60	20	F	Yes	1

We want to model this, then

$$y = \beta_0 + \beta_1(\text{Age}) + \beta_2(\text{Income}) + \beta_3(\text{gender})$$

Issues

1. Errors/residuals are not normally distributed.
2. No guarantee that the target/output/estimation is between 0 & 1.

Let us consider a function

$$f(z) = \frac{1}{1+e^{-y}}$$

$$P(\text{response} = \text{yes} / \text{age, income, gender}) = \frac{1}{1+e^{-(\beta_0 + \beta_1 \cdot \text{age} + \beta_2 \cdot \text{income} + \beta_3 \cdot \text{gender})}}$$

Why use logistic regression?

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

Logistic Regression

Logistic regression is a supervised classification model.

This allows us to make predictions from labelled data ,if the target variable is categorical.

Binary classification

Examples

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

Categorical Response Variables

Examples:

Whether or not a person smokes — Binary Response

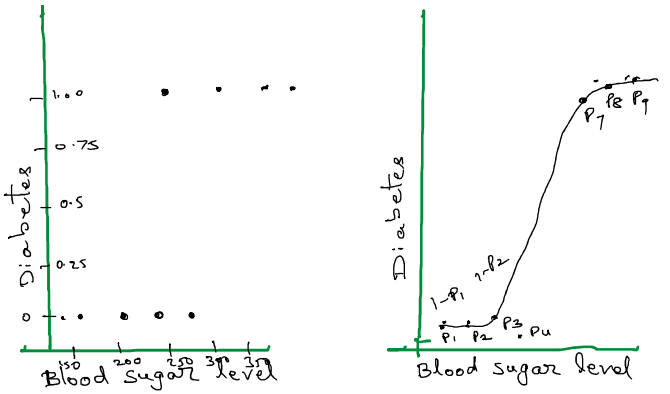
$$Y = \begin{cases} \text{Non} - \text{smoker} \\ \text{Smoker} \end{cases}$$

Success of a medical treatment

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

Opinion poll responses — Ordinal Response

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



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

Likelihood = $(1-P_1)(1-P_2)(1-P_3)(1-P_4)$ $P_5(1-P_6)P_7 P_8 P_9 P_{10}$

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

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

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

Binary Logistic Regression Model

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

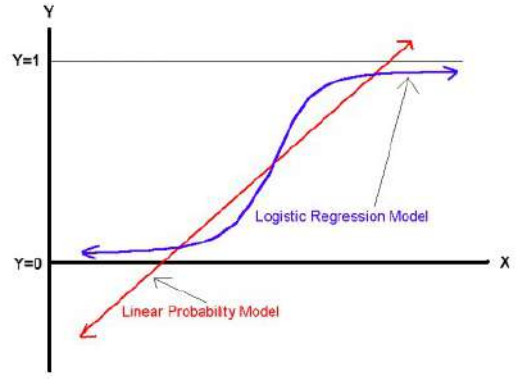
π = proportion of 1's at any $x_1, x_2, \ldots, x_k$

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

Comparing the LP and Logit Models



Session 11: Regression & ANOVA

Agenda

- Linear Regression(Revise)
- Model validation
- Ridge and lasso models
- Assumptions of Linear regression
- Logistic regression

Example:

consider the following data

y	370	386	443	499	528	616
x	22	25	28	31	33	38

Fit a regression line $y = \beta_0 + \beta_1 x$

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

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

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

x	y	x^2	xy
22	370	484	8140
25	386	625	9650
28	443	784	12404
31	499	961	15469
33	528	1089	17424
38	616	1444	23408
$\Sigma x = 177$	2842	5387	86495

$y = \beta_0 + \beta_1 x$

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

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

ie $2842 = 6(\beta_0) + \beta_1 (177)$

$86495 = 177(\beta_0) + \beta_1 (5387)$

on solving $\beta_0 = 0.195$

$\beta_1 = 16.05$

$\therefore y = (0.195) + (16.05)x$

Another form

Regression equation of y on x

$(y - \bar{y}) = r_{xy}(x - \bar{x})$

$\bar{y}$: mean of y

$\bar{x}$: mean of x

r_{xy}: Regression coefficient of y on x

$r_{xy} = \frac{\sigma_{xy}}{\sigma_x^2}$

BITS Pilani

Analysis of Variance (ANOVA)

- For example: A salesman in car sales wants to find the difference between two types of cars in terms of mileage:
  - Mid-Size Vehicles

  - Sports Utility Vehicles

- For example: A salesman in car sales wants to find the difference between three types of cars in terms of mileage:
  - Mid-Size Vehicles

  - Sports Utility Vehicles

PICKUP TRUCK

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…….

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

ANOVA

- ✓ In real life things do not typically result in two groups being compared
- ✓ Used with 3 or more groups to test for MEAN DIFFS.
- ✓ We have at least 3 means to test, e.g., H_0: $\mu_1 = \mu_2 = \mu_3$.
- ✓ Could take them 2 at a time, but really want to test all 3 (or more) at once.

Between/Within Groups

Variance can be separated into two major components

- ➢ **Within groups** – variability or differences in particular groups (individual differences)
- ➢ **Between groups** - differences depending what group one is in or what treatment is received

Null hypothesis:

no difference in means

Alternative hypothesis:

difference in means

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

ANOVA-analysis of variance

* Significance of difference between two sample means

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

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

Alternative Hypothesis

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

observations	population A	B	C
	26	18	23
2	25	16	19
3	28	17	26
4	12	8	30

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

Short cut method

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

Corr. Fact : $CF = \frac{T^2}{n}$; $n = n_1 + n_2 + \cdots + n_n$

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

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

$$SSE = SST - SST$$

Example

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

Mumbai :	16	8	12	14
Kolkata :	14	10	10	6
Delhi :	4	10	8	8

Prices in 3 cities are significantly different ?

Example

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
6	5	6
5	5	7
4	4	6
5	4	5
6	5	6
4	4	6

$c = 3, n = 18$

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

$= 30 + 27 + 36 = 93$

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

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

$= 154 + 123 + 218 = 14.50$

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

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

$= 7$

$SSE = SST - SSTR$

$= 14.50 - 7 = 7.50$

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

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

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

F-distribution tables

calculated value: 7

F table value: 3.68

$3.68 < 7 \Rightarrow$ Rejected

Two way ANOVA

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

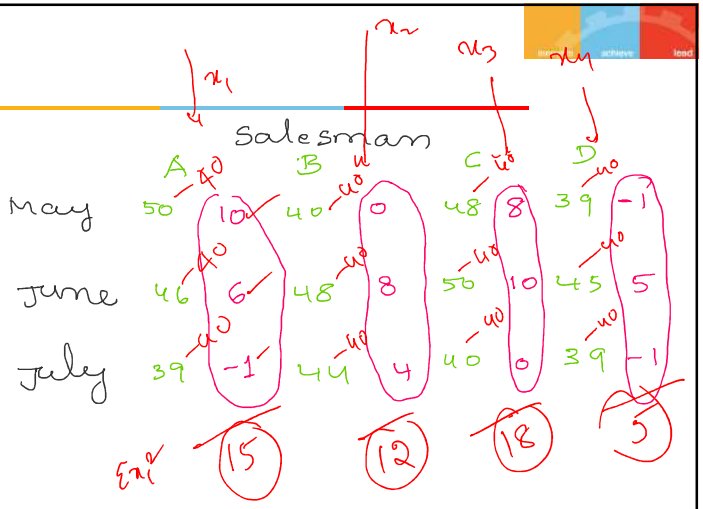
Example

Sales

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

→ Is there any significant difference in the sales made by A, B, C, D

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



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

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

SSTR = Sum of squares (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)

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

$= 91.5$

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

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

$= 216$

$SSE = SST - (SSTR + SSR)$

$= 216 - (42 + 91.5)$

$= 82$

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

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

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

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

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

MSTR < MSE → $F = \frac{MSE}{MSTR}$ ← MSTR > MSE → $F = \frac{MSTR}{MSE}$

Salesman → columns

	Sum of squares	D.o.f	Mean squares	Variance ratio
Between Salesmen	SSTR 42.0	c − 1 = 3	MSTR = 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) $F_{treatment} = 1.018 <$ (columns) $df_1 = 3, df_2 = 6, \alpha = 0.05$ accept

b) $F_{block} = 3.327 <$ (?) → accept

2, 6 → row

difference in the Sales by salesman

difference in Sales made during months.

L- 12:Appled Multivariate Analytics

Agenda

- → Bi Variate distribution
- → multivariate normal distribution
- → Eigen values and eigen vectors
- → Principal Component Analysis

Joint Discrete Distribution Function _Discrete

$P(x_i, y_j)$ → joint probability distribution function

↓

$P_X(x_i, y_j)$ → Marginal prob. dist. fun of X

$P_Y(x_i, y_j)$ → " " " of Y.

Joint Discrete Distribution Function _Continuous

$f(x,y)$ → joint p.d.f

↓

$f_X(x,y)$ → Marginal Pdf of X

$f_Y(x,y)$ → Marginal p.d.f of Y

Example:-

$$f(x,y) = \begin{cases} Kxy; & 0<x<4,\ 1<y<5 \\ 0: & \text{otherwise} \end{cases}$$

1) Find 'K' value ?
2) $P(1<x<2,\ 2<Y<3)$
3) $P(x \geq 3,\ Y \leq 2)$
4) $f_x(x,y)$
5) $f_Y(x,Y)$

1) $\int_{x=0}^{4}\int_{y=1}^{5} f(x,y)\,dy\,dx = 1$

$\Rightarrow K\int\int xy\,dy\,dx = 1$

$\Rightarrow K\left[\frac{x^2}{2}\right]_{x=0}^{4}\left[\frac{y^2}{2}\right]_{y=1}^{5} = 1$

$\Rightarrow K\left[\frac{16}{2}-0\right]\left[\frac{25}{2}-\frac{1}{2}\right] = 1$

$\Rightarrow K = 1/96$

2) $P(x \geq 3,\ y \leq 2)$

$= \int_{x=3}^{4}\int_{y=1}^{2} Kxy\,dy\,dx$

$\int x^n dx = \frac{x^{n+1}}{n+1}$

$= \frac{1}{96}\left[\frac{x^2}{2}\right]_{x=3}^{4}\left[\frac{y^2}{2}\right]_{y=1}^{2}$

$= \frac{1}{96\times2\times2}\left[4^2-3^2\right]\left[2^2-1^2\right]$

$= \frac{7}{128}$

3) $P(1<x<2,\ 2<Y<3)$

$= \int_{x=1}^{2}\int_{y=2}^{3} Kxy\,dx\,dy$

$= \frac{1}{96}\left[\frac{x^2}{2}\right]_{x=1}^{2}\left[\frac{y^2}{2}\right]_{y=2}^{3}$

$= \frac{5}{128}$

4) Marginal of X

$$f_X(x,y) = \int_{y=1}^{5} f(x,y)\,dy$$

$$= \int_{y=1}^{5} kxy\,dy = \frac{1}{96}x\left[\frac{y^2}{2}\right]_{y=1}^{5}$$

$$= \frac{x}{192}(25-1) = \frac{x}{8}$$

i.e. $f_X(x,y) = x/8$, $0 < x < 4$

5) Marginal of Y.

$$f_Y(x,y) = \int_{x=0}^{4} kxy\,dx$$

$$= \frac{y}{96}\left[\frac{x^2}{2}\right]_{x=0}^{4}$$

$$= \frac{y}{12}$$

$f_Y(x,y) = y/12$, $1 < y < 5$

Example:

Y \ X	0	1	2
0	1/6	1/3	1/12
1	2/9	1/6	0
2	1/36	0	0

Example:

Y \ X	0	1	2	Marg of Y
0	1/6	1/3	1/12	7/12
1	2/9	1/6	0	7/18
2	1/36	0	0	1/36
X	15/36	1/2	1/12	

X	0	1	2
P(X)	15/36	1/2	1/12

Y	0	1	2
P(Y)	7/12	7/18	1/36

Conditional Probabilities

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

$$P(X/Y) = \frac{P(X,Y)}{P_Y(X,Y)}$$ → joint; → Marginal of Y

$$P(Y/X) = \frac{P(X,Y)}{P_X(X,Y)}$$ → joint; → Marginal of X

$$f(X/Y) = \frac{f(X,Y)}{f_Y(X,Y)}$$ → joint; → Marginal of Y

$$f(Y/X) = \frac{f(X,Y)}{f_X(X,Y)}$$ → joint; → Marginal of X

Independent r.Vs.

$$P(x,y) = P_X(x,Y) \cdot P_Y(x,Y)$$

$$f(x,y) = f_x(x,y) \cdot f_y(x,y)$$

(joint; Marginal x; Marginal y)

Bi-Variate Normal dist

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

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

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

Multivariate Normal distribution

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

x_1, x_2, x_3

Det of Variance-covariance matrix

Inverse of Variance-covariance matrix

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

covariance between x_1 and x_2.

Preliminaries

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

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

$\text{cov}(x,x)$ = Variance

Covariance Matrix

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

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

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

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

Transformation matrices

Consider:

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

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

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

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

eigenvalue problem

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

eigenvalue problem

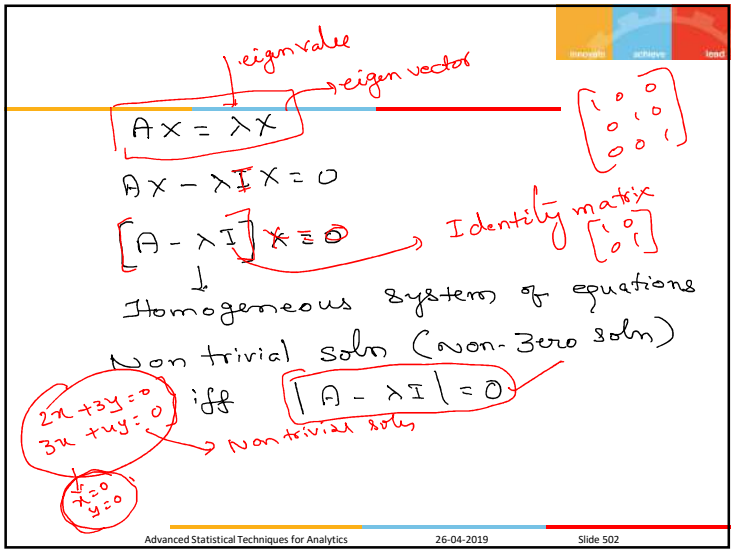
$$\begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix} \times \begin{bmatrix} 3 \\ 2 \end{bmatrix} = \begin{bmatrix} 12 \\ 8 \end{bmatrix} = 4 \times \begin{bmatrix} 3 \\ 2 \end{bmatrix}$$

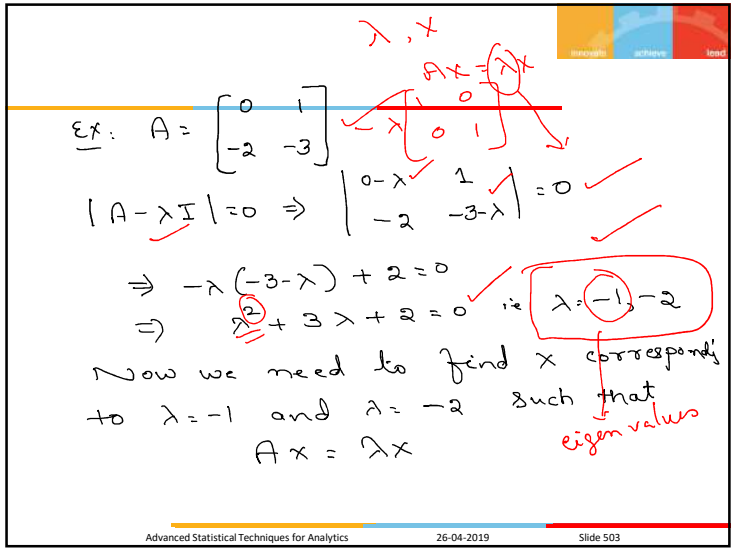
$$\mathbf{A} \cdot \mathbf{x} = \lambda \cdot \mathbf{x}$$

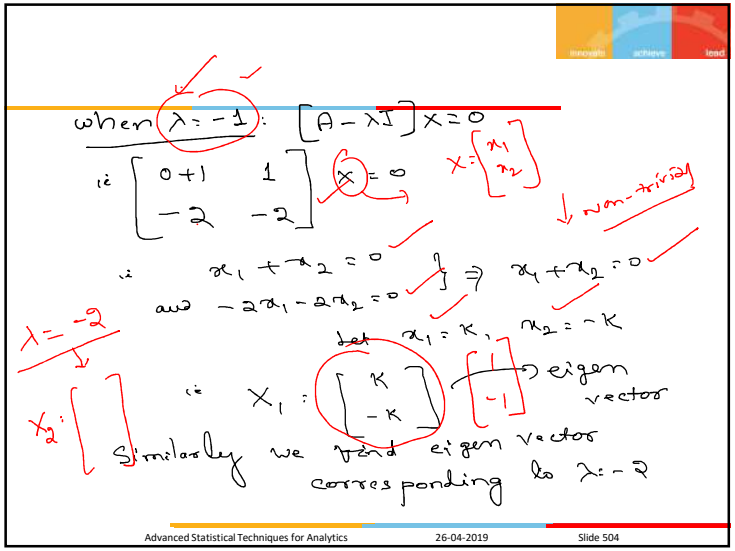
Therefore, (3,2) is an eigenvector of the square matrix **A** and 4 is an eigenvalue of **A**

eigen value, eigen vector corresponding

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





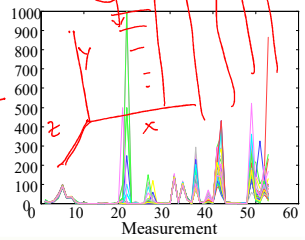


Data Presentation

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

Matrix Format

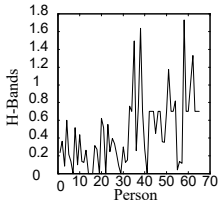
	H-WBC	H-RBC	H-Hgb	H-Hct	H-MCV	H-MCH	H-MCHC
A1	8.0000	4.8200	14.1000	41.0000	85.0000	29.0000	34.0000
A2	7.3000	5.0200	14.7000	43.0000	86.0000	29.0000	34.0000
A3	4.3000	4.4800	14.1000	41.0000	91.0000	32.0000	35.0000
A4	7.5000	4.4700	14.9000	45.0000	101.0000	33.0000	33.0000
A5	7.3000	5.5200	15.4000	46.0000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

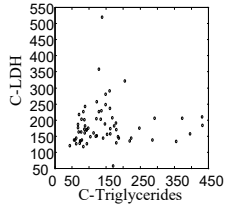


Visualization

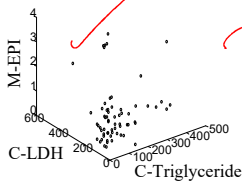
Univariate

Bivariate





Trivariate



Principal Components

Applications

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

Principal Component Analysis

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

As the dimensions of data increases, the difficulty to visualize it and perform computations on it also increases. So, how to reduce the dimensions of a data-

* Remove the redundant dimensions
* Only keep the most important dimensions

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

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

what does Principal Component Analysis (PCA) do?

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

How does PCA work

- Calculate the covariance matrix *X* of data points.
- Calculate eigen vectors and corresponding eigen values.
- Sort the eigen vectors according to their eigen values in decreasing order.
- Choose first k eigen vectors and that will be the new k dimensions.
- Transform the original n dimensional data points into k dimensions.

Example:

consider the following data

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

$r = 0.926$

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

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

$X = \begin{bmatrix} x & y \end{bmatrix}_{10 \times 2}$

Step 2 :-

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

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

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

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

$AX = \lambda x$

Step3.

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

$\lambda_2 = 0.0490$

Eigen vectors corresponding

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

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

ie

Variable	Eigen vector 1 $\lambda_1 = 1.2840$	Eigen vector 2 $\lambda_2 = 0.0490$ → total 1.333
x_1	0.678	0.735
x_2	0.735	−0.678
% of total variance	$\frac{1.2840}{1.333} = 96.3\%$	$\frac{0.0490}{1.333} = 3.7\%$

Step 4: Select Variation matrix

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

or

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

Step5. Find new Data set $Y = XV$

ie $Y = XV$

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

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

$Y_2 = 0.735x_1 - 0.678x_2$

i.e $Y_1 = 0.678x_1 + 0.735x_2$

Step5. Find new Data set $Y = XV$

i.e. $Y = XV$

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

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

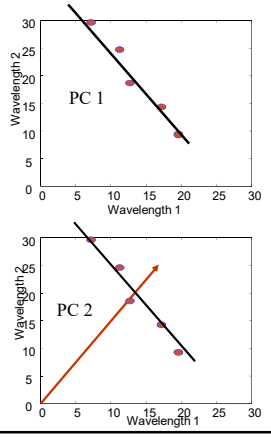
$\Rightarrow Y = 0.678\, x_1 + 0.735 x_2$

Summary :-

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 $Y = XV$

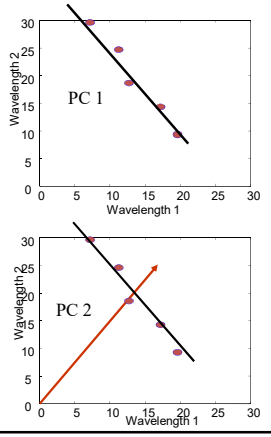
Principal Components

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



Principal Components

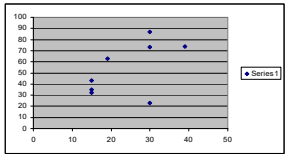
- All principal components (PCs) start at the origin of the ordinate axes.
- First PC is direction of maximum variance from origin
- Subsequent PCs are orthogonal to 1st PC and describe maximum residual variance

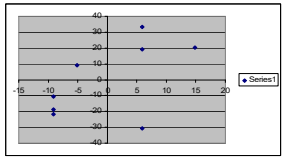


An Example

Mean1=24.1
Mean2=53.8

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





525

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.

L- 12:Appled Multivariate Analytics

Agenda

- Bi Variate distribution
- multivariate normal distribution
- Eigen values and eigen vectors
- Principal Component Analysis

Joint Discrete Distribution Function _Discrete

$P(x_i, y_j)$ → joint probability distribution function

↓

$P_X(x_i, y_j)$ → Marginal prob. dist. fun of X

$P_Y(x_i, y_j)$ → " " " of Y.

Joint Discrete Distribution Function _Continuous

$f(x,y)$ → joint p.d.f

↓

$f_X(x,y)$ → Marginal pdf of X

$f_Y(x,y)$ → Marginal p.d.f of Y

Example:-

$$f(x,y) = \begin{cases} Kxy; & 0<x<4,\ 1<y<5 \\ 0: & \text{otherwise} \end{cases}$$

1) Find 'K' value ?
2) $P(1<x<2,\ 2<Y<3)$
3) $P(x \geq 3,\ Y \leq 2)$
4) $f_X(x,y)$
5) $f_Y(x,y)$

1) $\int_{x=0}^{4}\int_{y=1}^{5} f(x,y)\,dy\,dx = 1$

$\Rightarrow K\iint xy\,dy\,dx = 1$

$\Rightarrow K\left[\frac{x^2}{2}\right]_{x=0}^{4}\left[\frac{y^2}{2}\right]_{y=1}^{5} = 1$

$\Rightarrow K\left[\frac{16}{2}-0\right]\left[\frac{25}{2}-\frac{1}{2}\right] = 1$

$\Rightarrow K = 1/96$

2) $P(x \geq 3, y \leq 2)$

$$= \int_{x=3}^{4} \int_{y=1}^{2} kxy \, dy \, dx$$

$$\int x^n dx = \frac{x^{n+1}}{n+1}$$

$$= \frac{1}{96} \left[\frac{x^2}{2} \right]_{x=3}^{4} \left[\frac{y^2}{2} \right]_{y=1}^{2}$$

$$= \frac{1}{96 \times 2 \times 2} \left[4^2 - 3^2 \right] \left[2^2 - 1^2 \right]$$

$$= \frac{7}{128}$$

3) $P(1 < x < 2, \; 2 < y < 3)$

$$= \int_{x=1}^{2} \int_{y=2}^{3} kxy \, dx \, dy$$

$$= \frac{1}{96} \left[\frac{x^2}{2} \right]_{x=1}^{2} \left[\frac{y^2}{2} \right]_{y=2}^{3}$$

$$= \frac{5}{128}$$

4) Marginal of X

$$f_X(x, y) = \int_{y=1}^{5} f(x, y) \, dy$$

$$= \int_{y=1}^{5} kxy \, dy = \frac{1}{96} x \left[y^2/2 \right]_{y=1}^{5}$$

$$= \frac{x}{192} (25 - 1) = \frac{x}{8}$$

i.e. $f_X(x, y) = x/8, \quad 0 < x < 4$

5) Marginal of Y.

$$f_Y(x, y) = \int_{x=0}^{4} kxy \, dx$$

$$= \frac{y}{96} \left[\frac{x^2}{2} \right]_{x=0}^{4}$$

$$= \frac{y}{12}$$

$$f_Y(x, y) = y/12, \quad 1 < y < 5$$

Example:

Y \ X	0	1	2
0	1/6	1/3	1/12
1	2/9	1/6	0
2	1/36	0	0

Example:

Y \ X	0	1	2	Marg of Y
0	1/6	1/3	1/12	7/12
1	2/9	1/6	0	7/18
2	1/36	0	0	1/36
X	15/36	1/2	1/12	

X	0	1	2
P(X)	15/36	1/2	1/12

Y	0	1	2
P(Y)	7/12	7/18	1/36

Conditional Probabilities

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

$$P(X/Y) = \frac{P(X,Y)}{P_Y(X,Y)}$$ → joint; → Marginal of Y

$$P(Y/X) = \frac{P(X,Y)}{P_X(X,Y)}$$ → joint; → Marginal of X

$$f(x/y) = \frac{f(x,y)}{f_y(x,y)}$$ → joint; → Marginal of y

$$f(y/x) = \frac{f(x,y)}{f_x(x,y)}$$ → joint; ↓ Marginal of X

Independent r.vs.

$$P(x,y) = P_x(x,y) \cdot P_y(x,y)$$

$P(x,y)$: joint; $P_x(x,y)$: Marginal x; $P_y(x,y)$: Marginal y

$$f(x,y) = f_x(x,y) \cdot f_y(x,y)$$

Bi-Variate Normal dist

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

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

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

Multivariate Normal distribution

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

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.

Covariance Matrix

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

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

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

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

Transformation matrices

Consider:

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

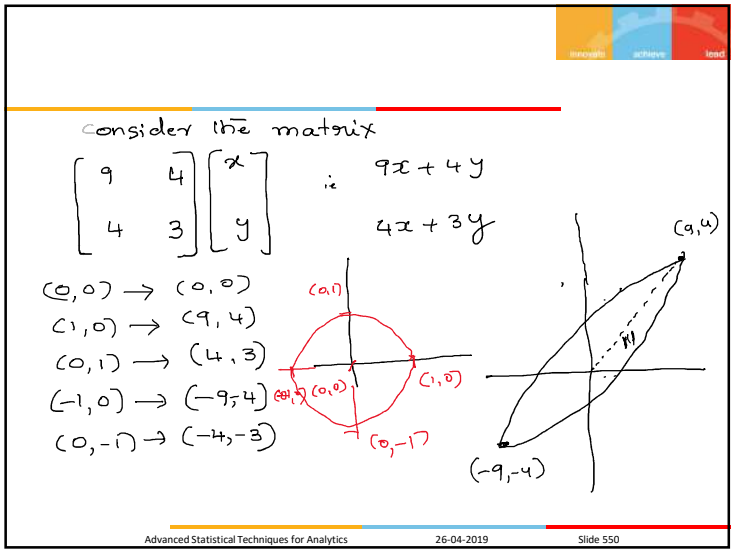
Square transformation matrix transforms (3,2) from its original location. Now if we were to take a multiple of (3,2)

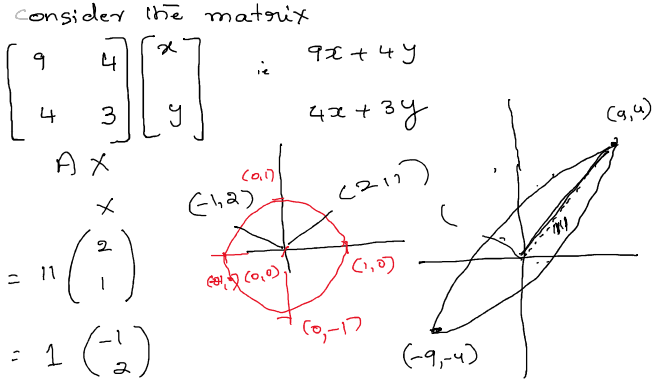
$$2 \times \begin{pmatrix} 3 \\ 2 \end{pmatrix} = \begin{pmatrix} 6 \\ 4 \end{pmatrix}$$

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

eigenvalue problem

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





eigenvalue problem

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

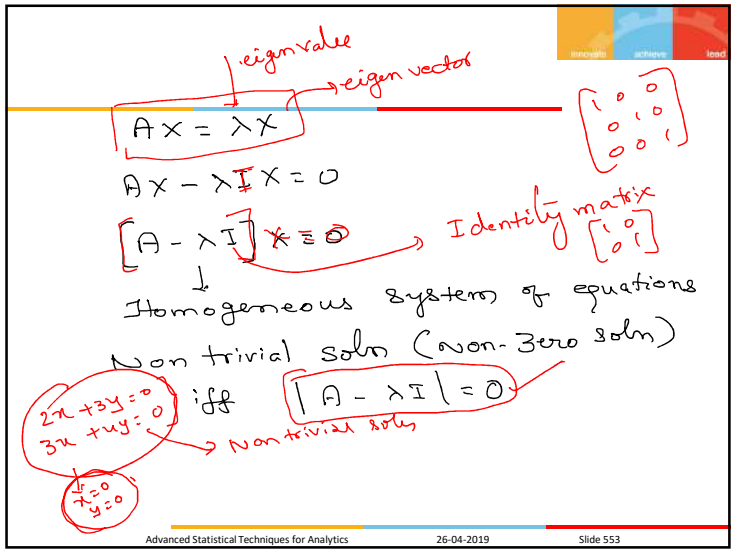
A . X = λ . X

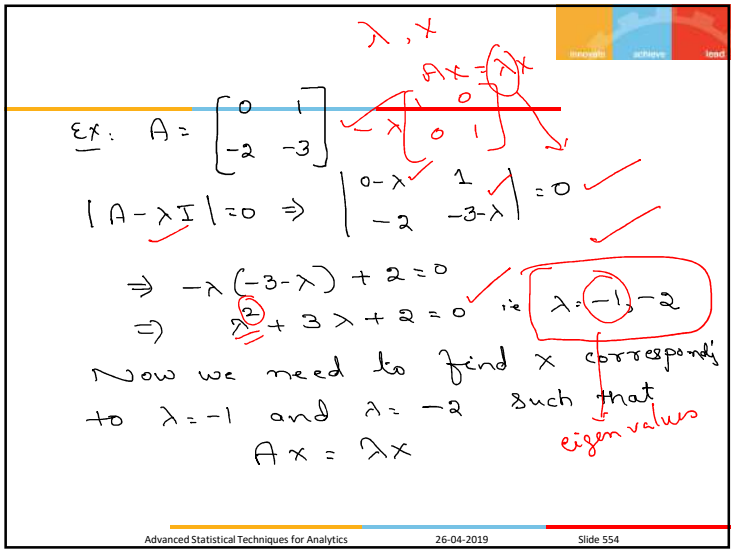
Therefore, (3,2) is an eigenvector of the square matrix **A** and 4 is an eigenvalue of **A**

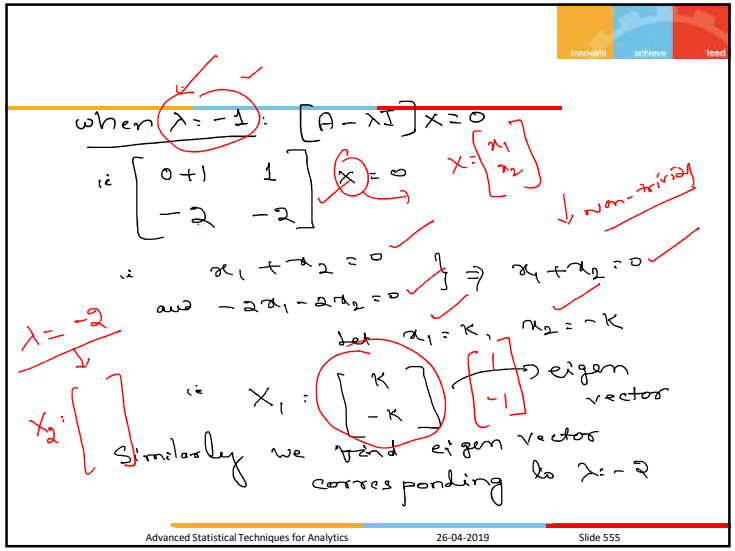
eigen value

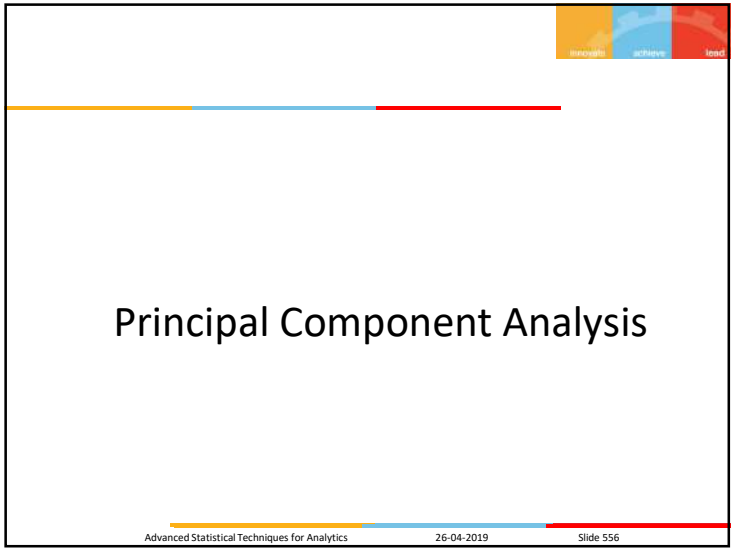
eigen vector corresponding

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


eigen value
eigen vector
$AX = \lambda X$
$AX - \lambda IX = 0$
$[A - \lambda I]X = 0$
Identity matrix
Homogeneous system of equations
Non trivial soln (non-zero soln)
iff
$|A - \lambda I| = 0$
$2x + 3y = 0$
$3x + 4y = 0$
Non trivial soln
$x = 0$
$y = 0$
Advanced Statistical Techniques for Analytics
26-04-2019
Slide 553


λ, X
$AX = \lambda X$
Ex: $A = \begin{bmatrix} 0 & 1 \\ -2 & -3 \end{bmatrix}$
$|A - \lambda I| = 0 \Rightarrow \begin{vmatrix} 0-\lambda & 1 \\ -2 & -3-\lambda \end{vmatrix} = 0$
$\Rightarrow -\lambda(-3-\lambda) + 2 = 0$
$\Rightarrow \lambda^2 + 3\lambda + 2 = 0$ i.e. $\lambda = -1, -2$
Now we need to find X corresponding to $\lambda = -1$ and $\lambda = -2$ such that
$AX = \lambda X$
eigen values
Advanced Statistical Techniques for Analytics
26-04-2019
Slide 554


when $\lambda = -1$: $[A - \lambda I]X = 0$
i.e. $\begin{bmatrix} 0+1 & 1 \\ -2 & -2 \end{bmatrix} X = 0$
$X = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$
non-trivial
$x_1 + x_2 = 0$
$-2x_1 - 2x_2 = 0$
$\Rightarrow x_1 + x_2 = 0$
$\lambda = -2$
X_2
let $x_1 = K$, $x_2 = -K$
i.e. $X_1 = \begin{bmatrix} K \\ -K \end{bmatrix}$
$\begin{bmatrix} 1 \\ -1 \end{bmatrix}$ eigen vector
Similarly we find eigen vector corresponding to $\lambda = -2$
Advanced Statistical Techniques for Analytics
26-04-2019
Slide 555


Principal Component Analysis
Advanced Statistical Techniques for Analytics
26-04-2019
Slide 556

→ Why PCA

↓

Dimensionality Reduction

Why PCA ?

→ With large number of variables, the matrix may be too large to study and interpret.

→ Visualization of the data is not helpful and complex also

→ If it is possible to reduce the number of variables to a few, interpretable linear combination, then it becomes simple.

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

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

what does Principal Component Analysis (PCA) do?

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

How does PCA work

- Calculate the covariance matrix *X* of data points.
- Calculate eigen vectors and corresponding eigen values.
- Sort the eigen vectors according to their eigen values in decreasing order.
- Choose first k eigen vectors and that will be the new k dimensions.
- Transform the original n dimensional data points into k dimensions.

Example:

consider the following data

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

$r = 0.926$

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

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

$X = \begin{bmatrix} x & y \end{bmatrix}_{10 \times 2}$

Step2 :-

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

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

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

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

$AX = \lambda X$

Step3.

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

$\lambda_2 = 0.0490$

Eigen vectors corresponding

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

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

ie

Variable	Eigen vector 1 $\lambda_1 = 1.2840$	Eigen vector 2 $\lambda_2 = 0.0490$ (total 1.333)
x_1	0.678	0.735
x_2	0.735	−0.678
% of total variance	$\frac{1.2840}{1.333} = 96.3\%$	$\frac{0.0490}{1.333} = 3.7\%$

Step 4 : Select Variation matrix

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

or

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

Step5. Find new Data set $Y = XV$

ie $Y = XV$

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

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

$Y_2 = 0.735x_1 - 0.678x_2$

$Y_1 = 0.678x_1 + 0.735x_2$

Step5. Find new Data set $Y = XV$

ie $Y = XV$

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

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

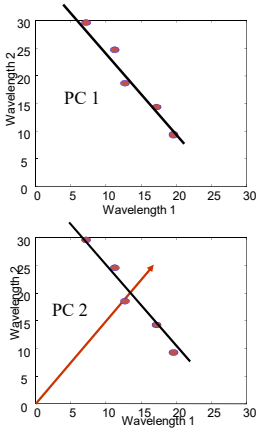
$Y = 0.678x_1 + 0.735x_2$

Summary :-

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

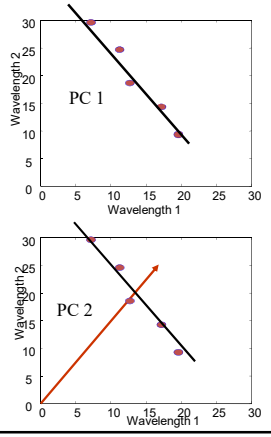
Principal Components

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



Principal Components

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



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.

Question :

How many principal Components should be considered ?

Any criteria - - - - ?

Component	Eigen value	Proportion
1	0.3775	0.7227 } 82%
2	0.0511	0.0977 }
3	0.0279	0.0535 — 87%
4	0.0230	0.0440 → 91
5	0.0168	0.0321 → 95
6	0.0120	0.0229 → 97
7	0.0085	0.0162 → 98
8	0.0039	0.0075 → 99
9	0.0018	0.0034 → 100
Total	0.5225	

1) First two eigen values together explains 82% of the variation.

First three eigen values explains 87% of the variation

$\lambda_1 \rightarrow PC_1$
$\lambda_2 \rightarrow PC_2$

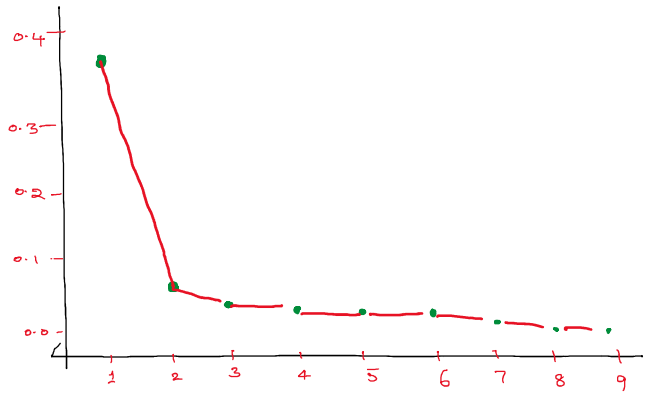
$\lambda_1 \rightarrow PC_1$
$\lambda_2 \rightarrow PC_2$
$\lambda_3 \rightarrow PC_3$

2) $\lambda_1 = 0.3775$
$\lambda_2 = 0.0511$ } → 0.326
$\lambda_3 = 0.0279$ → 0.0232
$\lambda_4 = 0.0230$ } → 0.0049

Like these, the differences are smaller and smaller.

This is another indicator of how many eigen values to consider

3) Scree plot:



L- 13:Appled Multivariate Analytics

Agenda_13th Session

- Principal Component Analysis
- ANOVA

Principal Component Analysis

Example:

consider the following data

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

$r = 0.926$

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

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

$X = \begin{bmatrix} x & y \end{bmatrix}_{10 \times 2}$

Step2 :-

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

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

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

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

$AX = \lambda x$

Step3.

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

$\lambda_2 = 0.0490$

Eigen vectors corresponding

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

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

ie

Variable	Eigen vector 1 $\lambda_1 = 1.2840$	Eigen vector 2 $\lambda_2 = 0.0490$ (total 1.333)
x_1	0.678	0.735
x_2	0.735	−0.678
% of total variance	$\frac{1.2840}{1.333} = 96.3\%$	$\frac{0.0490}{1.333} = 3.7\%$

Step 4 : Select Variation matrix

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

or

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

Step5. Find new Data set $Y = XV$

ie $Y = XV$

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

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

$Y_2 = 0.735 x_1 - 0.678 x_2$

i.e $Y_1 = 0.678 x_1 + 0.735 x_2$

Step5. Find new Data set $Y = XV$

ie $Y = XV$

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

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

i.e $Y = 0.678 x_1 + 0.735 x_2$

Summary:

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

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.

Question:

How many principal Component should be considered?

Any criteria - - - ?

Component	Eigen value	Proportion	
1	0.3775	0.7227	%
2	0.0511	0.0977	82%
3	0.0279	0.0535	87%
4	0.0230	0.0440	→ 91
5	0.0168	0.0321	→ 95
6	0.0120	0.0229	→ 97
7	0.0085	0.0162	→ 98
8	0.0039	0.0075	→ 99
9	0.0018	0.0034	→ 100
	Total 0.5225		

1) First two eigen values together explains 82% of the variation.

First three eigen values explains 87% of the variation

$\lambda_1 \rightarrow PC_1$
$\lambda_2 \rightarrow PC_2$

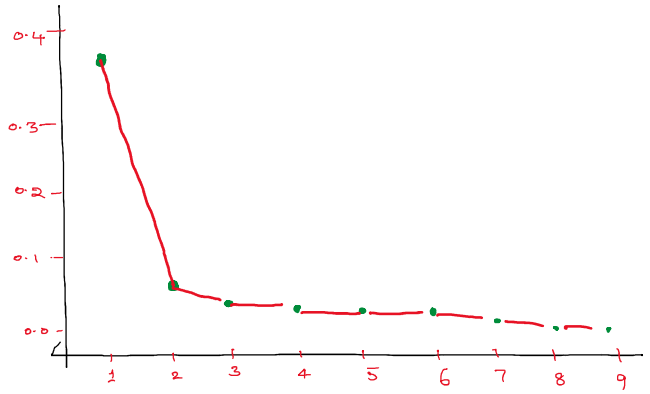
$\lambda_1 \rightarrow PC_1$
$\lambda_2 \rightarrow PC_2$
$\lambda_3 \rightarrow PC_3$

2) $\lambda_1 = 0.3775$ } $\rightarrow 0.326$
$\lambda_2 = 0.0511$ } $\rightarrow 0.0232$
$\lambda_3 = 0.0279$ } $\rightarrow 0.0049$
$\lambda_4 = 0.0230$

Like these, the differences are smaller and smaller.

This is another indicator of how many eigen values to consider

3) Scree plot:



Analysis of Variance (ANOVA)

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

Between/Within Groups

Variance can be separated into two major components

- **Within groups** – variability or differences in particular groups (individual differences)
- **Between groups** - differences depending what group one is in or what treatment is received

Null hypothesis:

no difference in means

Alternative hypothesis:

difference in means

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

ANOVA-analysis of variance

* Significance of difference between two sample means

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

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

Alternative Hypothesis

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

observations	population A	B	C
	26	18	23
2	25	16	19
3	28	17	26
4	12	8	30

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

Short cut method

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

Corr. Fact : $CF = \frac{T^2}{n}$; $n = n_1 + n_2 + \cdots + n_n$

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

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

$$SSE = SST - SST$$

Example

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

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

$n = 3, n = 18$

$$T = \Sigma x_1 + \Sigma x_2 + \Sigma x_3$$
$$= 30 + 27 + 36 = 93$$

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

$$SST = (\Sigma x_1^2 + \Sigma x_2^2 + \Sigma x_3^2) - CF$$
$$= 154 + 123 + 218 = 14.50$$

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

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

$$= 7$$

$$SSE = SST - SSTR$$

$$= 14.50 - 7 = 7.50$$

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

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

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

(df_1, df_2; F-distribution table)

calculated value: 7

F table value = 3.68

$3.68 < 7 \Rightarrow$ Rejected.

Two way ANOVA

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

Example

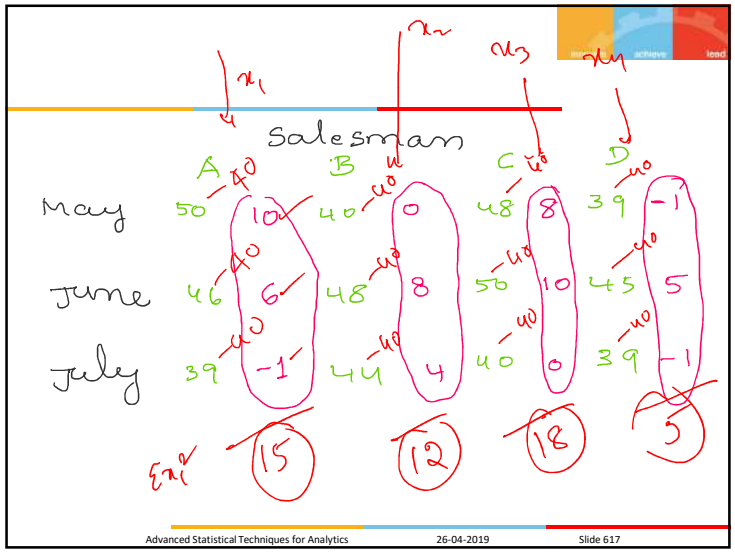
ANOVA

Sales

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

→ Is there any significant difference in the sales made, if any?

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



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

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

SSTR = Sum of squares (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)

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

$$= 91.5$$

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

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

$$= 216$$

$$SSE = SST - (SSTR + SSR)$$

$$= 216 - (42 + 91.5)$$

$$= 82$$

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

$df = (C-1)(n-1) = 3 \times 2 = 6$

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

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

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

MSTR < MSE ⇒ $F = \frac{MSE}{MSTR}$ ← MSTR > MSE ⇒ $F = \frac{MSTR}{MSE}$

Salesman → columns

	Sum of squares	D.o.f	Mean squares	Variance ratio
Between Salesmen	SSTR 42.0	c − 1 = 3	MSTR = 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		

rows → Months

a) $F_{treatment} = 1.018 <$ (columns) $df_1 = 3, df_2 = 6, \alpha = 0.05$ accept

b) $F_{block} = 3.327 <$ (?) accept

2, 6 → rows

difference in the Sales ~~by salesman~~

difference in Sales made during months.

BITS Pilani

L- 14: Two way ANOVA & Time Series Analysis

Two way ANOVA

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

Example

Sales

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

→ Is there any significant difference in the sales made, Ex: D

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

Salesman

	A	B	C	D
May	50 (10)	40 (0)	48 (8)	39 (-1)
June	46 (6)	48 (8)	50 (10)	45 (5)
July	39 (-1)	44 (4)	40 (0)	39 (-1)
	15	12	18	5

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

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

SSTR = Sum of squares (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)

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

$= 91.5$

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

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

$= 216$

$SSE = SST - (SSTR + SSR)$

$= 216 - (42 + 91.5)$

$= 82.5$

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

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

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

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

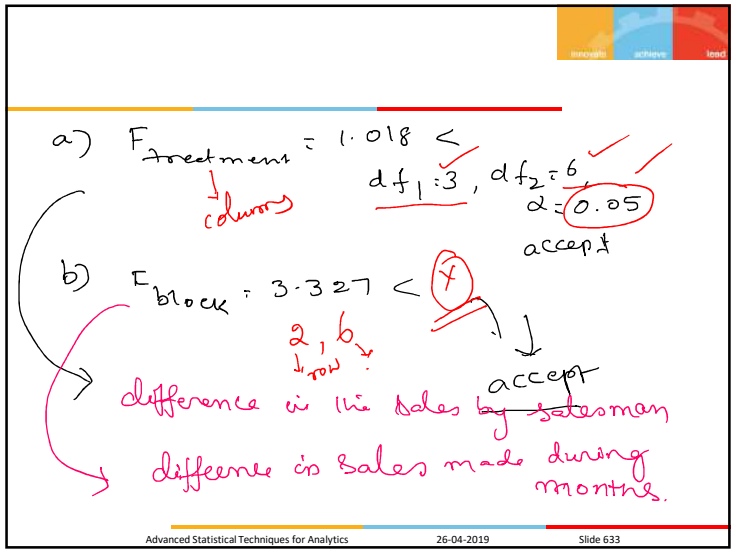
$MSE = \frac{SSE}{(c-1)(r-1)} = \frac{82.5}{6} = 13.75$

$MSTR < MSE \Rightarrow F = \frac{MSE}{MSTR}$ ← $MSTR > MSE \Rightarrow F = \frac{MSTR}{MSE}$ Salesman

	Sum of squares	D.o.f	Mean squares	Variance ratio
Between Salesmen	SSTR 42.0	c − 1 = 3 − 1 = 2	MSTR = 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		

columns

rows: Months



Time Series Analysis

- What is time series
- Why time series
- Components in time series
- Methods / models in time series

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

Why Time Series?

- To detect patterns of change over regular intervals of time.
- Predict for the future intervals of time

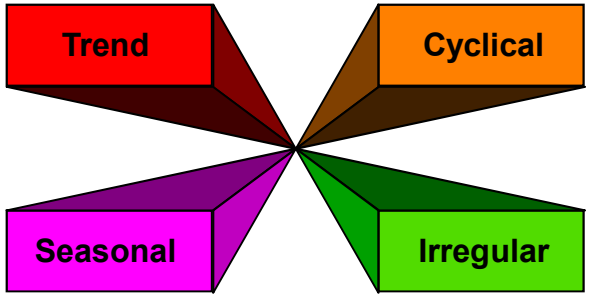
Applications

- Retail sales
- Spare parts planning
- Stock trading

Time series _ components

- Trend
- Cyclicality
- Seasonality
- Random / irregularity

Time Series Components

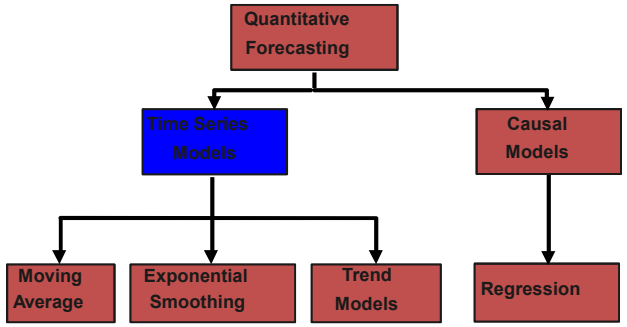


Time Series Models

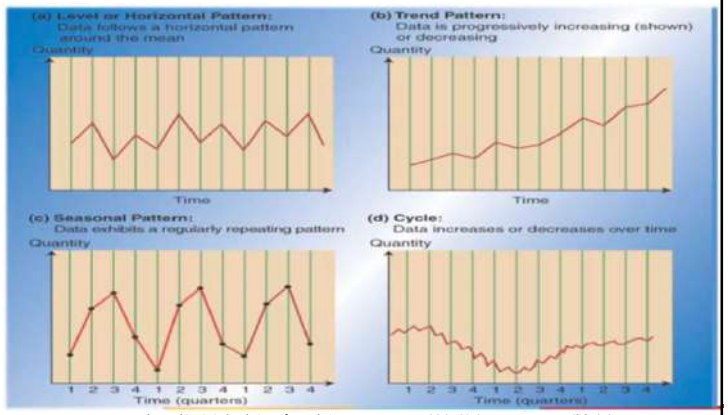
- Forecaster looks for data patterns as

  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

Quantitative Forecasting Methods



Time Series Patterns



Box – Jenkins Methodology

1. Condition data and select a model

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

2. Estimate the model parameters

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

Time Series Components

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

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

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

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

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

$$X_t = T_t S_t C_t I_t$$

Smoothing Methods

Moving Average Models

Simple Moving Average Forecast

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

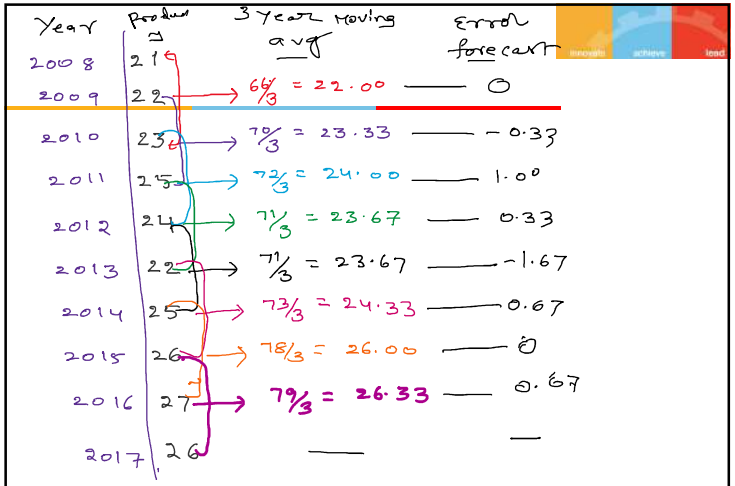
Weighted Moving Average Forecast

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

Example(Moving averages)

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

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



Time Series Models

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

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

- Allows emphasizing one period over others; above indicates more weight on recent data (C_t=.5)
- Differs from the simple moving average that weighs all periods equally - more responsive to trends

Example(Weighted moving Averages)

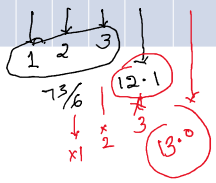
Weights	Month
3	Last month
2	Two months ago
1	Three months ago

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

Example(Weighted moving Averages)

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

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



Example.

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

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

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

Example.

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

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

$= 126$

b) $3 \times 137 + 2 \times 128 + 1 \times 130 / 6$

$= 133$ units

Time Series Models

- **Exponential Smoothing:**

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

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

Example:-

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

a) Forecast demand for the next week is March 8

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

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

Example:-

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

a) Forecast demand for the next week is March 8

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

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

$$= 495$$

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

Forecasting Trend

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

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

  - **Step 2 – Smoothing the trend**

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

  - **Forecast including the trend**

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

Measuring Forecasting Accuracy

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

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

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

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

Mean Square Error (MSE)
- Penalizes larger errors

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

Tracking Signal
- Measures if your model is working

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

Stationarity

Stationary time series have no trend.

conditions

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

Auto Correlation

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

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

Auto Correlation Function

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

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

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

Models

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

AR Model(Auto regressive model)

AR(p)

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

δ_t: Constant

y_{t-1}: value of time series at time $t-1$

$\phi_p \neq 0$

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

Moving Average(MA) Model

today's announcement

Yesterday's

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

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

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

$\theta_q \neq 0$

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

ARMA model –ARMA(p,q)

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

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

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

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

Example:-

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

a) Forecast demand for the next week 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

Forecasting Trend

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

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

  - **Step 2 – Smoothing the trend**

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

  - **Forecast including the trend**

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

Measuring Forecasting Accuracy

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

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

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

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

Mean Square Error (MSE)
- Penalizes larger errors

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

Tracking Signal
- Measures if your model is working

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

Stationarity

Stationary time series have no trend.

conditions

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

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 $x_1, x_2 \ldots x_n$ as iid noise

Auto Correlation

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

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

Auto Correlation Function

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

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

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

$$\downarrow$$

$$\rho_x(h)$$

Models

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

AR Model(Auto regressive model)

AR(p)

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

Constant

value of time series at time t-j

$\phi_p \neq 0$

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

Moving Average(MA) Model

today's announcement

Yesterday's

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

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

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

$\theta_q \neq 0$

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

ARMA model –ARMA(p,q)

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

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

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

Selecting the Right Forecasting Model

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