Data Analytics

Course Taught at IIFT

Day 8: Statistical Modelling with RStudio

Dr. Tanujit Chakraborty

Centre for Data Sciences

IIIT Bangalore

TEST OF HYPOTHESIS

TEST OF HYPOTHESIS

Introduction:

In many situations, it is required to accept or reject a statement or claim about some parameter

Example:

1. The average cycle time is less than 24 hours
2. The % rejection is only 1%

The statement is called the hypothesis

The procedure for decision making about the hypothesis is called hypothesis testing

Advantages

1.Handles uncertainty in decision making

2.Minimizes subjectivity in decision making

3.Helps to validate assumptions or verify conclusions

TEST OF HYPOTHESIS

Commonly used hypothesis tests on mean of normal distribution:

- Checking mean equal to a specified value ($mu = mu_0$)
- Two means are equal or not ($mu_1 = mu_2$)

Null Hypothesis:

A statement about the status quo

One of no difference or no effect

Denoted by H0

Alternative Hypothesis:

One in which some difference or effect is expected

Denoted by H1

TEST OF HYPOTHESIS

Methodology demo: To Test Mean = Specified Value (mu = mu_0)

Suppose we want to test whether mean of a process characteristic is 5 based on the following sample data from the process

4	4	5	5	6
5	4.5	6.5	6	5.5

Calculate the mean of the sample, xbar = 5.15

Compare xbar with specified value 5

or xbar - specified value = xbar - 5 with 0

If xbar - 5 is close to 0

then conclude mean = 5

else mean ≠ 5

TEST OF HYPOTHESIS

Methodology demo : To Test Mean = Specified Value (mu = mu_0)

Consider another set of sample data. Check whether mean of the process characteristic is 500

400	400	500	500	600
500	450	650	600	550

Mean of the sample, xbar = 515

xbar - 500 = 515 - 500 = 15

Can we conclude mean ≠ 500?

Conclusion:

Difficult to say mean = specified value by looking at xbar - specified value alone

TEST OF HYPOTHESIS

Methodology demo: To Test Mean = Specified Value ($mu = mu_0$)

Test statistic is calculated by dividing (xbar - specified value) by a function of standard deviation

To test Mean = Specified value

Test Statistic t_0 = (xbar - Specified value) / (SD / $\sqrt{n}$)

If test statistic is close to 0, conclude that Mean = Specified value

To check whether test statistic is close to 0, find out p value from the sampling distribution of test statistic

TEST OF HYPOTHESIS

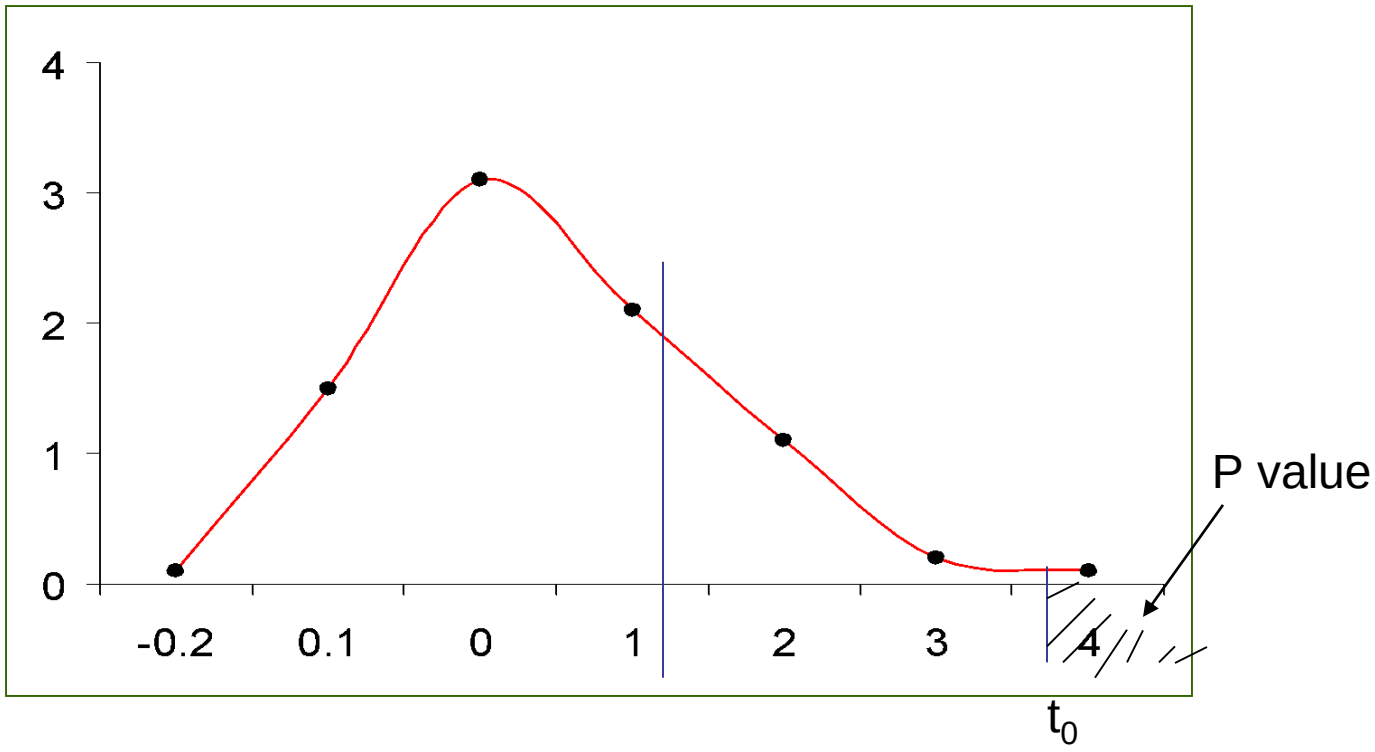
Methodology demo: To Test Mean = Specified Value

P value

The probability that such evidence or result will occur when H0 is true

Based on the reference distribution of test statistic

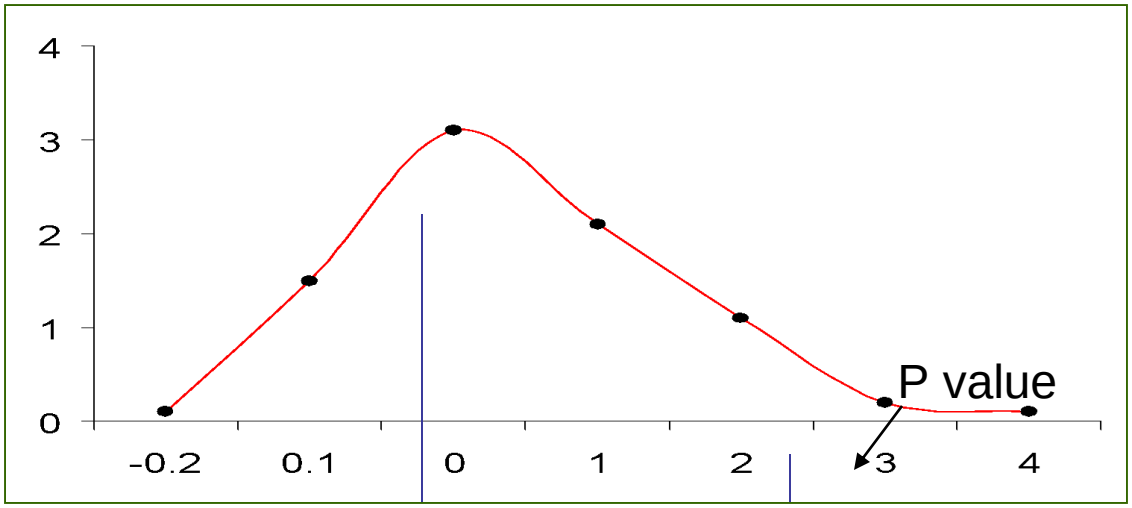
The tail area beyond the value of test statistic in reference distribution



TEST OF HYPOTHESIS

Methodology demo : To Test Mean = Specified Value

P value



If test statistic t_0 is close to 0 then p will be high

If test statistic t_0 is not close to 0 then p will be small

If p is small , $p < 0.05$ (with alpha = 0.05), conclude that $t \neq 0$, then

Mean ≠ Specified Value, H0 rejected

TEST OF HYPOTHESIS

To Test Mean = Specified Value (mu = mu_0)

Example: Suppose we want to test whether mean of the process characteristic is 5 based on the following sample data

4	4	5	5	6
5	4.5	6.5	6	5.5

H0: Mean = 5

H1: Mean ≠ 5

Calculate xbar = 5.15

SD = 0.8515

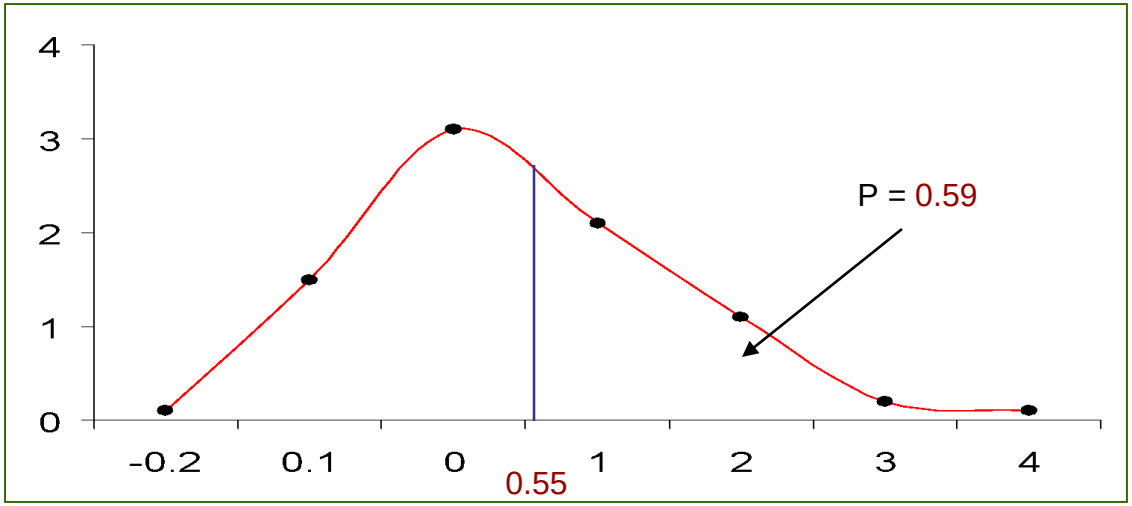
n = 10

Test statistic t_0 = (xbar - 5)/(SD / √n) = (5.15 - 5) / (0.8515 / √10) = 0.5571

TEST OF HYPOTHESIS

Example: To Test Mean = Specified Value (mu = mu_0)

$t_0 = 0.5571$



P ≥ 0.05, hence Mean = Specified value = 5.

H0: Mean = 5 is not rejected

TEST OF HYPOTHESIS

Hypothesis Testing: Steps

1.Formulate the null hypothesis H0 and the alternative hypothesis H1

2.Select an appropriate statistical test and the corresponding test statistic

3.Choose level of significance alpha (generally taken as 0.05)

4.Collect data and calculate the value of test statistic

5.Determine the probability associated with the test statistic under the null hypothesis using sampling distribution of the test statistic

6.Compare the probability associated with the test statistic with level of significance specified

TEST OF HYPOTHESIS

Install the necessary packages

```
> install.packages("car")
> library(car)
> install.packages("gplots")
> library(gplots)
> install.packages("ggplot2")
> library(ggplot2)
> install.packages("qqplotr")
> library(qqplotr)
> install.packages("boot")
> library(boot)
```

TEST OF HYPOTHESIS

One sample t test

Exercise 1 : A company claims that on an average it takes only 40 hours or less to process any purchase order. Based on the data given below, can you validate the claim? The data is given in PO_Processing.csv

Reading data to mydata

```
> mydata = read.csv('PO_Processing.csv',header = T,sep = ",")
> PT = mydata$Processing_Time
```

Performing one sample t test

```
> t.test(PT, alternative = 'greater', mu = 40)
```

Statistics	Value
t	3.7031
df	99
P value	0.0001753

NORMALITY TEST

NORMALITY TEST

Normality test

A methodology to check whether the characteristic under study is normally distributed or not

Two Methods :

Normality test - Quantile – Quantile (Q- Q) plot

Plots the ranked samples from the given distribution against a similar number of ranked quantiles taken from a normal distribution

If the sample is normally distributed then the line will be straight in the plot

Normality test – Shapiro – Wilk test

H0: Deviation from bell shape (normality) = 0

H1 : Deviation from bell shape ≠ 0

If p value ≥ 0.05 (5%), then H0 is not rejected, distribution is normal

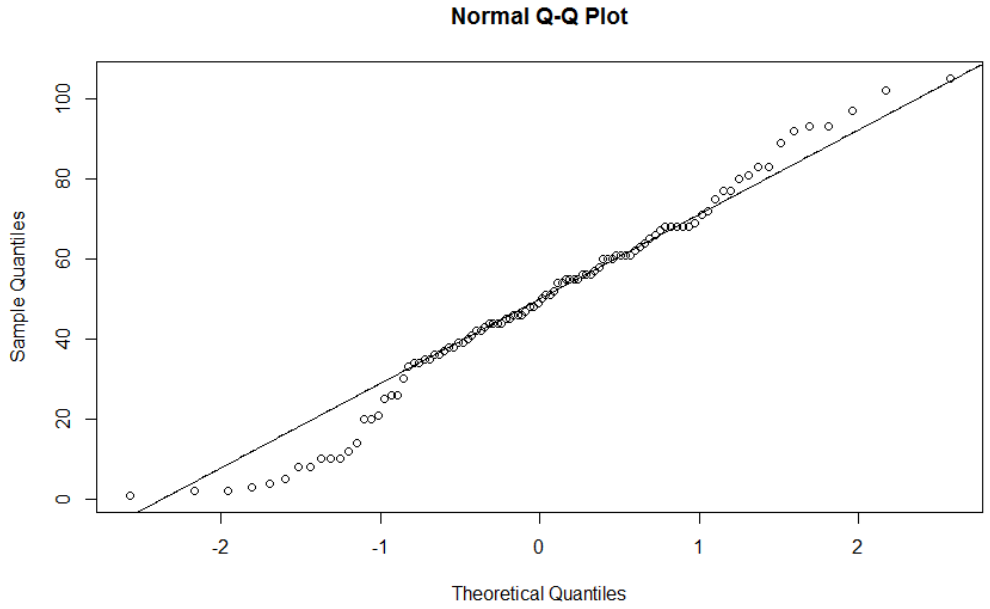
NORMALITY TEST

Normality test

Exercise 1 : The processing times of purchase orders is given in PO_Processing.csv. Is the distribution of processing time is normally distributed?

Normality Check using Normal Q – Q plot

```
> qqnorm(PT)
> qqline(PT)
```



NORMALITY TEST

Normality test

Exercise 1 : The processing times of purchase orders is given in PO_Processing.csv. Is the distribution of processing time is normally distributed?

Normality Check using Shapiro – Wilk test

```
> shapiro.test(PT)
```

Statistics	Value
W	0.9804
p value	0.1418

Conclusion: The data is Normal if p-value is above 0.05

ANALYSIS OF VARIANCE

ANALYSIS OF VARIANCE

ANOVA

Analysis of Variance is a test of means for two or more populations

Partitions the total variability in the variable under study to different components

H0 = $Mean_1 = Mean_2 = - - - = Mean_k$

Reject H0 if p – value < 0.05

Example:

To study location of shelf on sales revenue

ANALYSIS OF VARIANCE

One Way ANOVA : Example

An electronics and home appliance chain suspect the location of shelves where television sets are kept will influence the sales revenue. The data on sales revenue in lakhs from the television sets when they are kept at different locations inside the store are given in sales revenue data file. The location is denoted as 1:front, 2: middle & 3: rear. Verify the doubt? The data is given in Sales_Revenue_Anova.csv.

Factor: Location(A)

Levels : front, middle, rear

Response: Sales revenue

ANALYSIS OF VARIANCE

One Way Anova : Example

Step 1: Calculate the sum, average and number of response values for each level of the factor (location).

Level 1 Sum(A_1):

Sum of all response values when location is at level 1 (front)

= 1.55 + 2.36 + 1.84 + 1.72

= 7.47

nA_1: Number of response values with location is at level 1 (front)

= 4

ANALYSIS OF VARIANCE

One Way Anova : Example

Step 1: Calculate the sum, average and number of response values for each level of the factor (location).

Level 1 Average:

Sum of all response values when location is at level 1 / number of response values with location is at level 1

= A_1 / nA_1 = 7.47 / 4 = 1.87

ANALYSIS OF VARIANCE

One Way Anova : Example

Step 1: Calculate the sum, average and number of response values for each level of the factor (location).

	Level 1 (front)	Level 2 (middle)	Level 3 (rear)
Sum	A_1: 7.47	A_2: 30.31	A_3: 15.55
Number	nA_1: 4	nA_2: 8	nA_3: 6
Average	1.87	3.79	2.59

ANALYSIS OF VARIANCE

One Way Anova : Example

Step 2: Calculate the grand total (T)

T = Sum of all the response values

= 1.55 + 2.36 + - - - + 2.72 + 2.07 = 53.33

Step 3: Calculate the total number of response values (N)

N = 18

Step 4: Calculate the Correction Factor (CF)

CF = (Grand Total)^2 / Number of Response values

= T^2 / N = $(537.33)^2$ / 18 = 158.0049

ANALYSIS OF VARIANCE

One Way Anova : Example

Step 5: Calculate the Total Sum of Squares (TSS)

TSS = Sum of square of all the response values - CF

$= 1.55^2 + 2.36^2 + - - - + 2.72^2 + 2.07^2 – 158.0049$

$= 15.2182$

Step 6: Calculate the between (factor) sum of square

$SS_A = A_1^2 / nA_1 + A_2^2 / nA_2 + A_3^2 / nA_3 - CF$

$= 7.47^2 / 4 + 30.31^2 / 8 + 15.55^2 / 4 – 158.0049$

$= 11.0827$

Step 7: Calculate the within (error) sum of square

SS_e = Total sum of square – between sum of square

$= TSS - SS_A = 15.2182 – 11.0827 = 4.1354$

ANALYSIS OF VARIANCE

One Way Anova : Example

Step 8: Calculate degrees of freedom (df)

Total df = Total Number of response values - 1

= 18 - 1 = 17

Between df

= Number of levels of the factor - 1

= 3 - 1 = 2

Within df = Total df – Between df

= 17 - 2 = 15

ANALYSIS OF VARIANCE

One Way Anova : R Code

Reading data and variables to R

```
> mydata = read.csv('Sales_Revenue_Anova.csv',header = T,sep = ",")
> location = mydata$Location
> revenue = mydata$Sales.Revenue
```

Converting location to factor

```
> location = factor(location)
```

Computing ANOVA table

```
> fit = aov(revenue  ~  location)
> summary(fit)
```

ANALYSIS OF VARIANCE

One Way Anova : Example

Anova Table:

Source	df	SS	MS	F	F Crit	P value
location	2	11.08272	5.541358	20.09949	3.68	0.0000
Residuals	15	4.135446	0.275696			
Total	17	15.21816				

MS = SS / df

$F = MS_{Between} / MS_{Within}$

F Crit =finv (probability, between df, within df) , probability = 0.05

P value = fdist (F, between df, within df)

ANALYSIS OF VARIANCE

One Way Anova : Decision Rule

If p value < 0.05, then

The factor has significant effect on the process output or response.

Meaning:

When the factor is changed from 1 level to another level,

there will be significant change in the response.

ANALYSIS OF VARIANCE

One Way Anova : Example Result

For factor Location, $p = 0.000 < 0.05$

Conclusion:

Location has significant effect on sales revenue

Meaning:

The sales revenue is not same for different locations like front, middle & rear

ANALYSIS OF VARIANCE

One Way Anova : Example Result

The expected sales revenue for different location under study is equal to level averages.

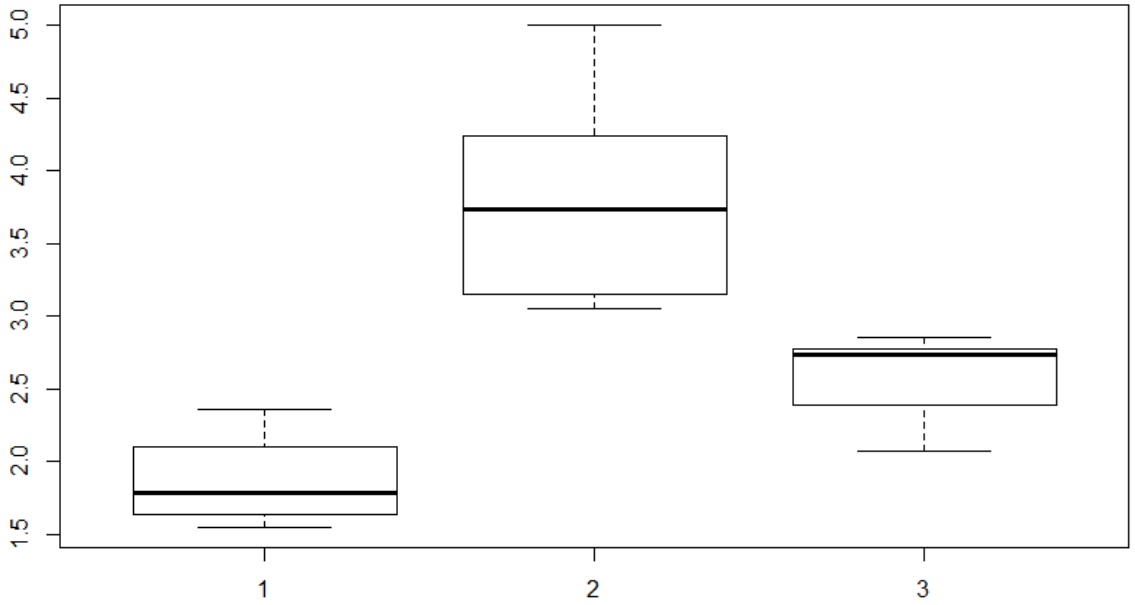
Location	Expected Sales Revenue
Front	1.8675
Middle	3.78875
rear	2.591667

```
> aggregate(revenue ~ location, FUN = mean)
```

ANALYSIS OF VARIANCE

One Way Anova : Example Result

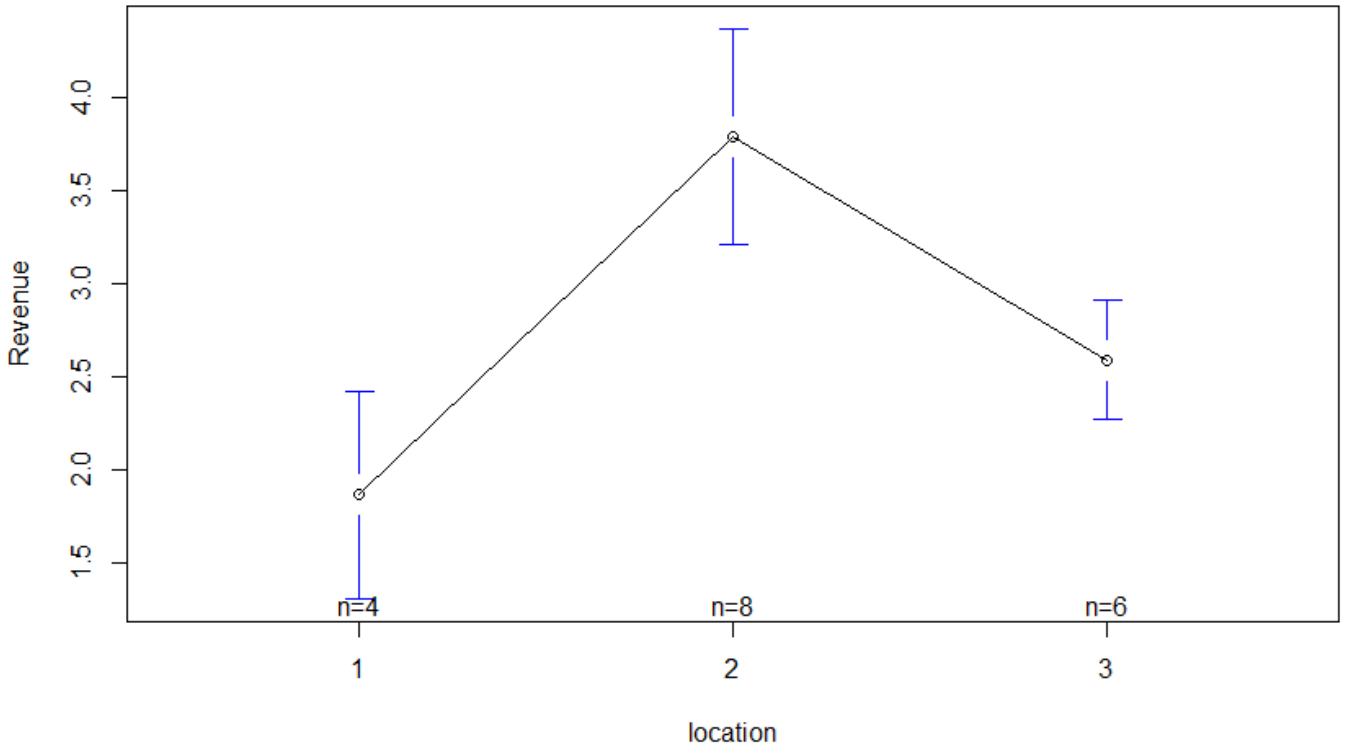
> boxplot(revenue ~ location)



ANALYSIS OF VARIANCE

One Way Anova : Example Result

```
> library(gplots)
> plotmeans(revenue ~ location)
```



ANALYSIS OF VARIANCE

One Way Anova : Tukey's Honestly Significant Difference (HSD) Test

Used to do pair wise comparison between the levels of factors

R code

```
>TukeyHSD(fit)
```

Comparison	Mean difference	Lower	Upper	p value
2 - 1	1.92125	1.086067	2.756433	0.0000
3 - 1	0.724167	-0.15619	1.604527	0.1158
3 - 2	-1.19708	-1.93365	-0.46052	0.0020

ANALYSIS OF VARIANCE

Anova logic:

Two Types of Variations:

1. Variation within the level of a factor
2. Variation between the levels of factor

ANALYSIS OF VARIANCE

Anova logic :

Variation between the level of a factor:

The effect of Factor.

Variation within the levels of a factor:

The inherent variation in the process or Process Error.

	Location		
	Front	Middle	rear
Sales Revenue	1.34 1.89 1.35 2.07 2.41 3.06	3.20 2.81 4.52 4.40 4.75 5.19 3.42 9.80	2.30 1.91 1.40 1.48

ANALYSIS OF VARIANCE

Anova logic :

If the variation between the levels of a factor is significantly higher than the inherent variation

then the factor has significant effect on response

To check whether a factor is significant:

Compare variation between levels with variation within levels

ANALYSIS OF VARIANCE

Anova logic :

Measure of variation between levels: MS of the factor ($MS_{between}$)

Measure of variation within levels: MS Error (MS_{within})

To check whether a factor is significant:

Compare MS of between with MS within

i.e. Calculate $F = MS_{between} / MS_{within}$

If F is very high, then the factor is significant.

ANALYSIS OF VARIANCE

Variation Within levels:

Ideally variation within all the levels should be same

To check whether variation within the levels are same or not

Do Bartlett's test

If p value ≥ 0.05, then variation within the levels are equal, otherwise not

R Code for Bartlett's test

```
> bartlett.test(revenue, location, data = mydata)
```

ANALYSIS OF VARIANCE

Variation Within levels:

Bartlett's Test result for sales revenue (location of TV sets) example

Bartlett's K^2 Statistic	df	p value
3.8325	2	0.1472

Since p value = 0.1472 > 0.05, the variance within the levels are equal

REGRESSION ANALYSIS

REGRESSION ANALYSIS

Regression

Correlation helps

To check whether two variables are related

If related

Identify the type & degree of relationship

Regression helps

- To identify the exact form of the relationship
- To model output in terms of input or process variables

REGRESSION ANALYSIS

Exercise 1: The data from the pulp drying process is given in the file DC_Simple_Reg.csv. The file contains data on the dry content achieved at different dryer temperature. Develop a prediction model for dry content in terms of dryer temperature.

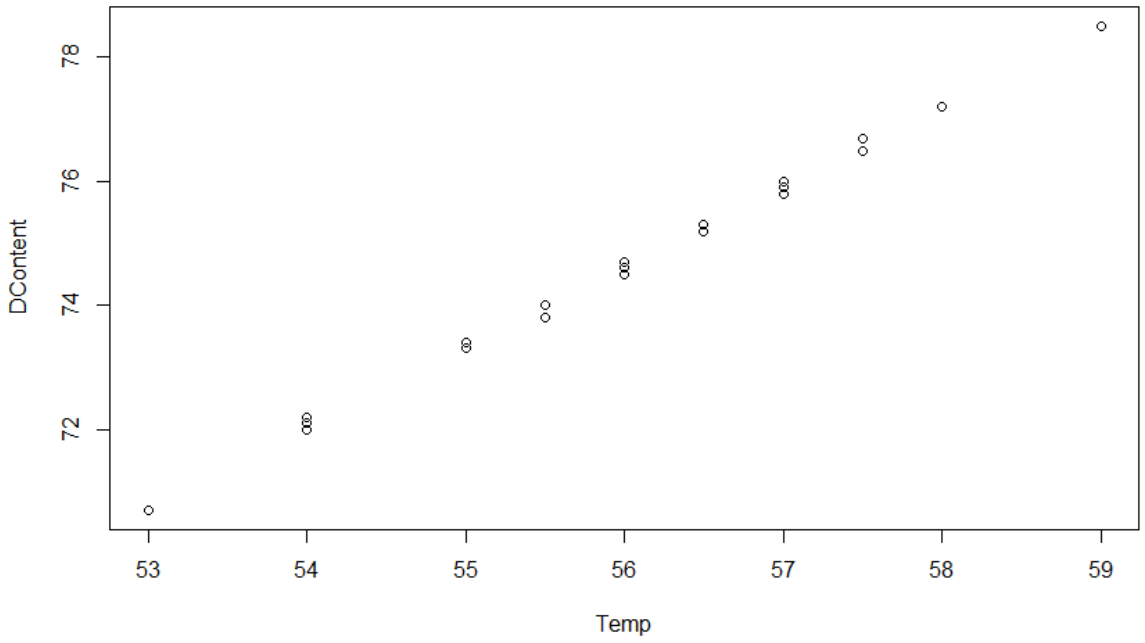
1. Reading the data and variables

```
> mydata = read.csv('DC_Simple_Reg.csv',header = T,sep = ",")
> Temp = mydata$Dryer.Temperature
> DContent = mydata$Dry.Content
```

REGRESSION ANALYSIS

2. Constructing Scatter Plot

```
> plot(Temp, DContent)
```



REGRESSION ANALYSIS

3. Computing Correlation Matrix

> cor(Temp, DContent)

Attribute	Dry Content
Temperature	0.9992

Remark:

Correlation between y & x need to be high (preferably 0.8 to 1 to -0.8 to -1.0)

REGRESSION ANALYSIS

4: Performing Regression

```
> model = lm(DContent ~ Temp)
> summary(model)
```

Statistic	Value	Criteria
Residual standard error	0.07059	
Multiple R-squared	0.9984	> 0.6
Adjusted R-squared	0.9983	> 0.6

Model	df	F	p value
Regression	1	24497	0.000
Residual	40		
Total	41		

Criteria:
P value < 0.05

REGRESSION ANALYSIS

4: Performing Regression

Attribute	Coefficient	Std. Error	t Statistic	p value
Intercept	2.183813	0.463589	4.711	0.00
Temperature	1.293432	0.008264	156.518	0.00

Interpretation

The p value for independent variable need to be < significance level α (generally α = 0.05)

Model: Dry Content = 2.183813 + 1.293432 x Temperature

REGRESSION ANALYSIS

5: Regression Anova

> anova(model)

ANOVA					
Source	SS	df	MS	F	p value
Temp	122.057	1	122.057	24497	0.000
Residual	0.199	40	0.005		
Total	122.256	41			

Criteria: P value < 0.05

REGRESSION ANALYSIS

5: Residual Analysis

```
> pred = fitted(model)
> Res = residuals(model)
> write.csv(pred,"D:/ISI/DataSets/Pred.csv")
> write.csv(Res,"D:/ISI/DataSets/Res.csv")
```

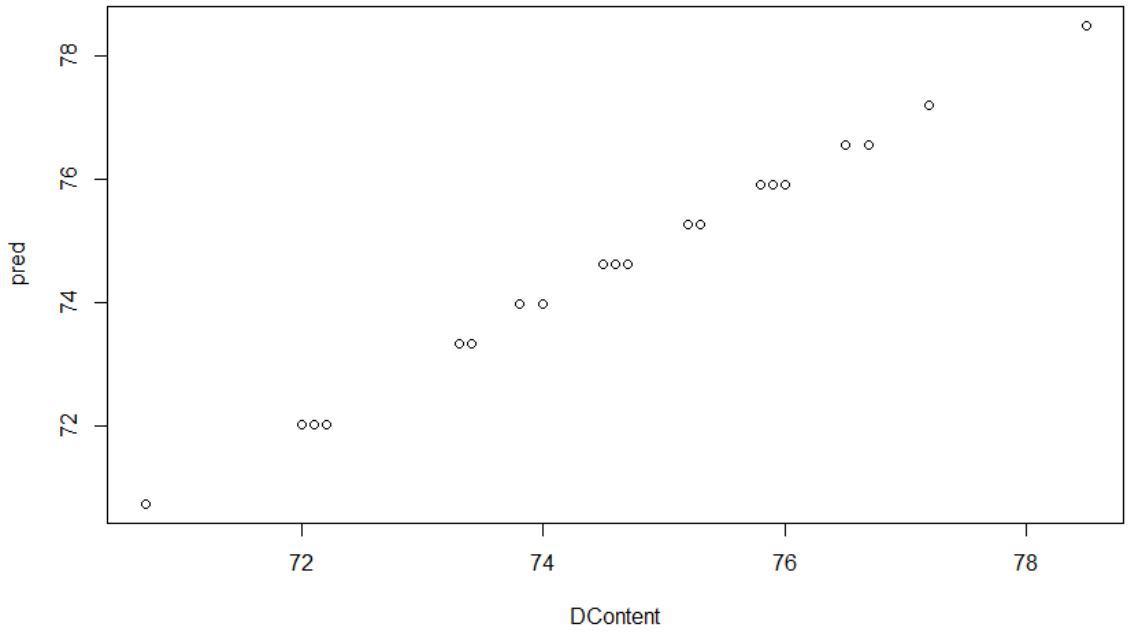
SL No.	Fitted	Residuals	SL No.	Fitted	Residuals
1	73.32259	-0.02259	22	74.61602	-0.01602
2	74.61602	-0.01602	23	75.26274	-0.06274
3	73.96931	0.030693	24	73.96931	0.030693
4	78.49632	0.00368	25	75.90946	-0.00946
5	74.61602	-0.01602	26	75.26274	0.03726
6	73.96931	0.030693	27	73.96931	0.030693
7	75.26274	-0.06274	28	78.49632	0.00368
8	77.20289	-0.00289	29	76.55617	-0.05617
9	75.90946	-0.00946	30	74.61602	-0.11602
10	74.61602	-0.01602	31	75.90946	0.090544
11	73.32259	-0.02259	32	76.55617	-0.05617
12	75.90946	-0.00946	33	76.55617	0.143828
13	75.90946	0.090544	34	75.90946	0.090544
14	74.61602	-0.01602	35	75.90946	-0.10946
15	74.61602	0.083977	36	73.96931	-0.16931
16	74.61602	-0.11602	37	73.32259	-0.02259
17	70.73573	-0.03573	38	74.61602	-0.01602
18	72.02916	-0.02916	39	73.32259	0.077409
19	72.02916	0.070841	40	75.90946	0.090544
20	72.02916	0.170841	41	73.96931	0.030693
21	70.73573	-0.03573	42	75.26274	-0.06274

REGRESSION ANALYSIS

5: Residual Analysis

Scatter Plot: Actual Vs Predicted (fit)

> plot(DContent, pred)



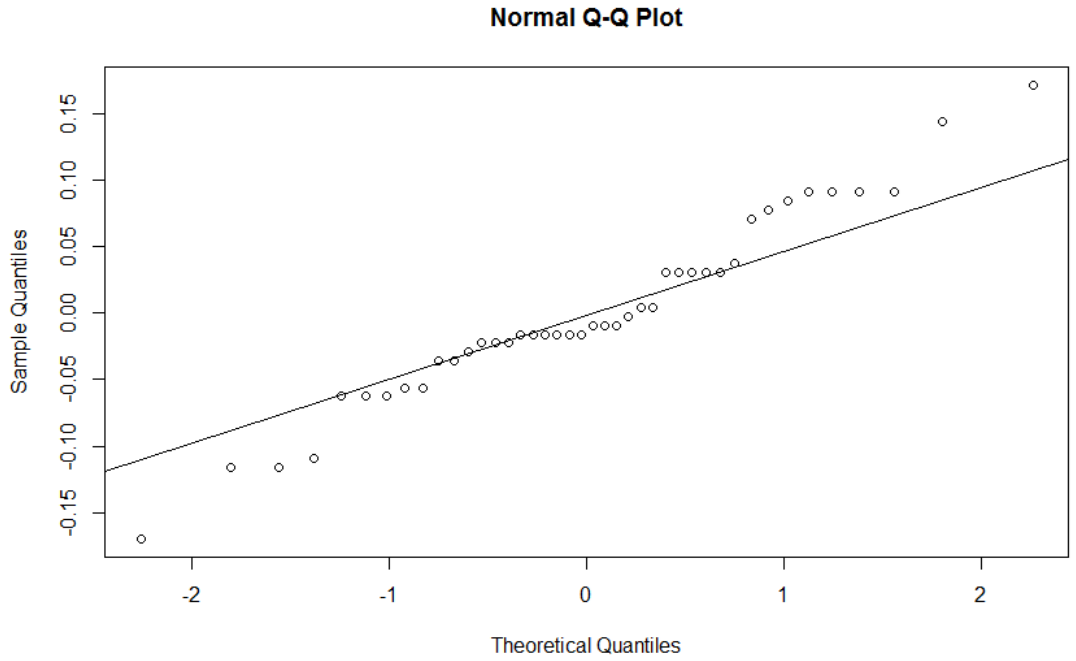
REGRESSION ANALYSIS

5: Residual Analysis

Normality Check on residuals

> qqnorm(Res)

> qqline(Res)



Residuals should be normally distributed or bell shaped

REGRESSION ANALYSIS

5: Residual Analysis

Normality Check on residuals

```
> shapiro.test(Res)
```

Shapiro-Wilk normality Test:	
W	p value
0.9693	0.3132

Residuals should be normally distributed or bell shaped

REGRESSION ANALYSIS

5: Residual Analysis

> plot(pred, Res)

> plot(Temp, Res)

Residuals should be independent and stable

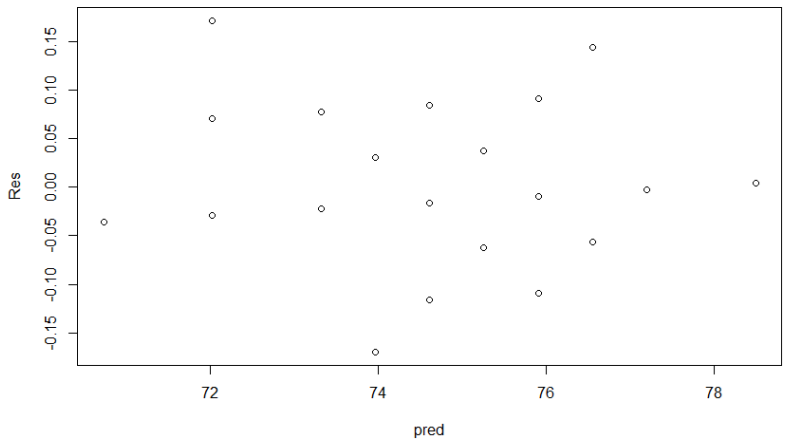
Plot the residuals against fitted value. The points in the graph should be scattered randomly and should not show any trend or pattern. The residuals should not depend in anyway on the fitted value.

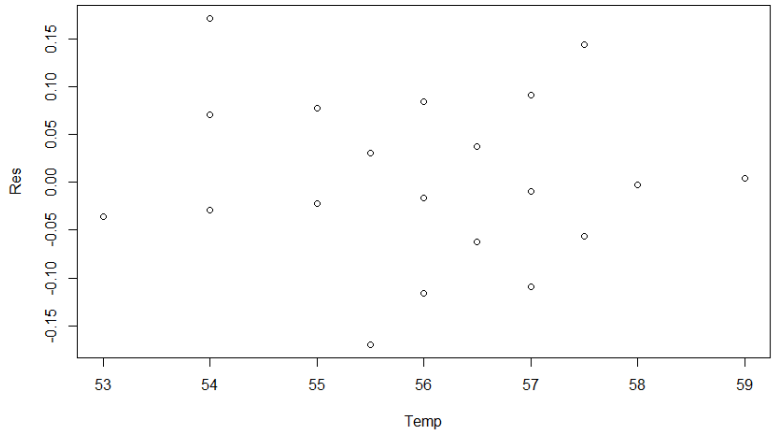
If there is a pattern then a transformation such as log y or $\sqrt{y}$ to be used

Similarly the residuals shall not depend on x. This can be checked by plotting residuals vs x. A pattern in this plot is an indication that the residuals are not independent of x. Instead of x, develop the model with a function of as predictor (Eg: x^2, 1/x, $\sqrt{x}$, log(x), etc.)

REGRESSION ANALYSIS

Residual Analysis





There is no trend or pattern on residuals vs fitted value ,residuals vs observation order or residuals vs x plot. Hence the assumptions of independence and stability of residuals are satisfied.

REGRESSION ANALYSIS

6: Outlier test

Observations with Bonferonni p – value < 0.05 are potential outliers

```
> library(car)
> outlierTest(model)
```

Observation	Studentized Residual	Bonferonni p value
20	2.723093	0.40417

REGRESSION ANALYSIS

7: Leave One Out Cross Validation (LOOCV)

- Split the data into two parts : training data and test data
- Test data consists of only one observation (x_1, y_1)
- Training data consists of the remaining n – 1 observations namely (x_2, y_2),(x_3, y_3) , - - -, (x_n, y_n)
- Develop the model using n – 1 training data observations and predict the response y_1 of the test data observation
- Compute the residuals and mean square error $MSE_1 = (y_{1actual} - y_{1pred})^2$
- Repeat the process by taking (x_1, y_1) as test data and the remaining n – 1 observations as training data
- Compute MSE_2
- Repeating the procedure n times produces n squared errors MSE_1, MSE_2, - - -, MSE_n
- LOOCV estimate of the test MSE is the average of these n test error estimates

$$CV_{(n)} = \frac{1}{n}\sum_{i=1}^{n} MSE_i$$

REGRESSION ANALYSIS

7: Leave One Out Cross Validation (LOOCV)

```
> library(boot)
> attach(mydata)
> mymodel = glm(Dry.Content ~ Dryer.Temperature)
> valid = cv.glm(mydata, mymodel)
> valid$delta[1]
```

Statistic	Value
Delta	0.005201004

REGRESSION ANALYSIS

Multiple Linear Regression

To model output variable y in terms of two or more variables.

General Form:

$$y = a + b_1x_1 + b_2x_2 + - - - + b_kx_k + \varepsilon$$

Two variable case:

$$y = a + b_1x_1 + b_2x_2 + \varepsilon$$

Where

a: intercept (the predicted value of y when all x's are zero)

b_j: slope (the amount change in y for unit change in x_j keeping all other x's constant, j = 1,2,---,k)

REGRESSION ANALYSIS

Exercise : The effect of temperature and reaction time affects the X.yield. The data collected in given in the Mult-Reg_Yield file. Develop a model for X.yield in terms of temperature and time?

Step 1: Reading the data and variables

```
> mydata = read.csv('Mult_Reg_Yield.csv',header = T,sep = ",")
> mydata[,-1]  # Removing 1st  column
> attach(mydata)
```

REGRESSION ANALYSIS

Exercise : The effect of temperature and reaction time affects the X.yield. The data collected in given in the Mult-Reg_Yield file. Develop a model for X.yield in terms of temperature and time?

Step 2: Correlation Matrix

> cor(mydata)

	Time	Temperature	X.Yield
Time	1	-0.00756	0.89671
Temperature	-0.00756	1	-0.05457
X.Yield	0.89671	-0.05457	1

Correlation between xs & y should be high

Correlation between xs should be low

REGRESSION ANALYSIS

Step 3: Regression Output – Identify the model

```
> model = lm(X.Yield ~ Temperature + Time)
> summary(model)
```

Attribute	Coefficient	Std. Error	t Statistic	p value
Time	0.9061	0.12337	7.344	0.0000
Temperature	-0.0642	0.16391	-0.392	0.702
Intercept	-67.8844	40.58652	-1.67	0.118

Interpretation: Only time is related to % yield as p value < 0.05

REGRESSION ANALYSIS

Step 4: Regression Output

Statistic	Value	Criteria
Adjusted R Square	0.7766	≥ 0.6

ANOVA

> anova(model)

Source	SS	df	MS	F	p value
Time	6777.8	1	6777.8	53.9872	0.000
Temp	19.3	1	19.3	0.1534	0.702
Residual	1632.1	13	125.5		

Criteria: P value < 0.05

REGRESSION ANALYSIS

Step 5: Modified Regression Output – Identify the model

```
> model_m = lm(X.Yield ~ Time)
> summary(model_m)
```

Attribute	Coefficient	Std. Error	t Statistic	p value
Time	0.9065	0.1196	7.580	0.0000
Intercept	-81.6205	19.7906	-4.124	0.00103

Model % Yield= 0.9065 x Time - 81.621

REGRESSION ANALYSIS

Step 6: Regression Output

Statistic	Value	Criteria
Adjusted R Square	0.7901	≥ 0.6

ANOVA

> anova(model_m)

Source	SS	df	MS	F	p value
Time	6777.8	1	6777.8	57.462	0.000
Residual	1651.3	14	118.0		

Criteria: P value < 0.05

REGRESSION ANALYSIS

Step 7: Residual Analysis

```
>pred = fitted(model_m)
>Res = residuals(model_m)
>write.csv(pred,"C:/Users/Downloads/data_and_code/Data and Code/Pred_m.csv")
>write.csv(Res,"C:/Users/Downloads/data_and_code/Data and Code/Res_m.csv")
```

Step 8: Standardizing Residuals using Scale function

```
>Std_Res = scale(Res, center = TRUE, scale = TRUE)
>write.csv(Std_Res,"C:/Users/Downloads/data_and_code/Data and Code/Std_Res_m.csv")
```

The "center" parameter (when set to TRUE) is responsible for subtracting the mean on the numeric object from each observation.
The "scale" parameter (when set to TRUE) is responsible for dividing the resulting difference by the standard deviation of the numeric object.

REGRESSION ANALYSIS

Residual Analysis

SL No.	Temperature	% Yield	Predicted	Time
1	190	35.0	36.22	130
2	176	81.7	76.10	174
3	205	42.5	39.84	134
4	210	98.3	91.51	191
5	230	52.7	67.94	165
6	192	82.0	94.23	194
7	220	34.5	48.00	143
8	235	95.4	86.98	186
9	240	56.7	44.38	139
10	230	84.4	88.79	188
11	200	94.3	77.01	175
12	218	44.3	59.79	156
13	220	83.3	90.61	190
14	210	91.4	79.73	178
15	208	43.5	38.03	132
16	225	51.7	52.53	148

REGRESSION ANALYSIS

6: Normality test using Shapiro Wilk Normality Test

```
> shapiro.test(Res)
```

Shapiro-Wilk normality Test:	
W	p value
0.9693	0.3132

6: Outlier test

Observations with Bonferonni p – value < 0.05 are potential outliers

```
> library(car)
> outlierTest(model_m)
```

Observation	Studentized Residual	Bonferonni p value
11	1.781515	NA

REGRESSION ANALYSIS

7: Leave One Out Cross Validation (LOOCV)

```
> library(boot)
> attach(mydata)
> mymodel = glm(X.Yield ~ Time)
> myvalidation = cv.glm(mydata, mymodel)
> myvalidation$delta[1]
```

Statistic	Value
Delta	128.8541

REGRESSION ANALYSIS

Exercise : The effect of temperature, time and kappa number of pulp affects the % conversion of UB pulp to Cl_2 pulp. inspection. The data collected in given in the Mult_Reg_Conversion file. Develop a model for % conversion in terms of exploratory variables?

Step 1: Reading the data and variables

```
> data = read.csv('Mult_Reg_Conversion.csv',header = T,sep = ",")
> mydata[,-1]  # Removing 1st  column
> attach(mydata)
```

REGRESSION ANALYSIS

Step 1: Correlation Analysis

> cor(mydata)

	Temperature	Time	Kappa #	X..Conversion
Temperature	1.00	-0.96	0.22	0.95
Time	-0.96	1.00	-0.24	-0.91
Kappa #	0.22	-0.24	1.00	0.37
X..Conversion	0.95	-0.91	0.37	1.00

Interpretation

High Correlation between X..Conversion and Temperature & Time

High Correlation between Temperature & Time - Multicollinearity

REGRESSION ANALYSIS

Measure for Multicollinearity

Variance Inflation Factor (VIF)

Measures the correlation (linear association) between each x variable with other x's

$VIF_i = 1/(1- R_i^2)$

Where R_i is the coefficient for regressing x_i on other x's

Criteria: VIF > 5 indicates multicollinearity.

REGRESSION ANALYSIS

Regression Output

➢model = lm(X..Conversion ~ Kappa.number + Temperature + Time)

> summary(model)

	Coeff	Std. Error	t	p value
Constant	-121.27	55.43571	-2.19	0.0492
Temperature	0.12685	0.04218	3.007	0.0109
Time	-19.0217	107.92824	-0.18	0.863
Kappa #	0.34816	0.17702	1.967	0.0728

Variance-inflation factors (VIF)

> vif(mymodel)

x	VIF
Temperature	12.23
Time	12.33
Kappa #	1.062

REGRESSION ANALYSIS

Regression Output

Statistic	Value	Criteria
Adjusted R Square	0.899	> 0.6

Regression ANOVA

> anova(model)

Model	SS	df	MS	F	p value
Kappa.number	290.79	1	290.79	20.4915	0.000694
Temperature	1662.19	1	1662.19	117.1310	0
Time	0.44	1	0.44	0.0311	0.8630417
Residual	170.290	12	14.191		
Total	2123.709	15			

REGRESSION ANALYSIS

Tackling Multicollinearity:

1. Remove one or more of highly correlated independent variable
2. Principal Component Regression
3. Partial Least Square Regression
4. Ridge Regression
5. Collecting Additional Data

REGRESSION ANALYSIS

Tackling Multicollinearity:

Method 1: Removing highly correlated variable – Stepwise Regression

Approach

- A null model is developed without any predictor variable x. In null model, the predicted value will be the overall mean of y
- Then predictor variables x's are added to the model sequentially
- After adding each new variable, the method also remove any variable that no longer provide an improvement in the model fit
- Finally the best model is identified as the one which minimizes Akaike information criterion *(*AIC)

$$AIC = \frac{1}{n\hat{\sigma}^2}(RSS + 2d\hat{\sigma}^2)$$

REGRESSION ANALYSIS

Tackling Multicollinearity:

Method 1: Removing highly correlated variable – Stepwise Regression

Akaike information criterion (AIC)

$$AIC = \frac{1}{n\hat{\sigma}^2}(RSS + 2d\hat{\sigma}^2)$$

n: number of observations

$\hat{\sigma}^2$: estimate of error or residual variance

d: number of x variables included in the model

RSS: Residual sum of squares

REGRESSION ANALYSIS

Tackling Multicollinearity:

Method 1: Removing highly correlated variable – Stepwise Regression

R code

```
> library(MASS)
> mymodel = lm(X..Conversion ~ Temperature + Time + Kappa.number)
> step =stepAIC(mymodel, direction = "both")
```

Step	x's in the model	AIC
1	Temperature, Time & Kappa Number	45.8
2	Temperature & Kappa Number	43.9

REGRESSION ANALYSIS

Tackling Multi collinearity:

Method 1: Stepwise Regression

> summary(step)

Attribute	Coefficient	Std. Error	t Statistic	p value
Temperature	0.13396	0.01191	11.250	0.0000
Kappa #	0.35106	0.16955	2.071	0.0589
Intercept	-130.68986	14.14571	-9.239	0.0000

X..Conversion = 0.13396 * Temperature + 0.35106 * Kappa # - 130.68986

Variance-inflation factors (VIF)

> vif(step)

x	VIF
Temperature	1.0526
Kappa #	1.0526

REGRESSION ANALYSIS

Tackling Multi collinearity:

Method 1: Stepwise Regression

```
> pred = predict(step)
> res = residuals(step)
 > cbind(X..Conversion, pred, res)
> mse = mean(res^2)
> rmse = sqrt(mse)
```

Statistic	Value
Mean Square Error (MSE)	10.7
Root Mean Square Error (RMSE)	3.27

REGRESSION ANALYSIS

k fold Cross Validation

Steps

1.Divide the data set into k equal subsets
2.Keep one subset (sample) for model validation
3.Develop the model using all the other k – 1 subsets data put together
4.Predict the responses for the test data and compute residuals
5.Return the test sample back to the original data set and take another subset for model validation
6.Go to step 3 and continue until all the subsets are tested with different models
7.Compute the overall Root Mean Square Residuals. RMSE of validation should not be high compared to the original model developed with all the data points together.

Note: when k = n, then k fold cross validation is same as leave one out cross validation

REGRESSION ANALYSIS

k fold Cross Validation

R code

```
> library(DAAG)
> cv.lm(mymodel, m = 16)
> cv.lm(mymodel, df = mydata, m = 16)
```

m: number of validations required. M = 16 = n, hence equal to leave one out cross validation

Model	MSE	RMSE
Original	10.7	3.27
Cross Validation	19.6	4.43

REGRESSION ANALYSIS

Tackling Multi collinearity:

Method 2: Principal Component Regression

1.Perform principal component analysis on x variables
2.Use the principal components as x variables and develop the model

R Code : Principal Component Regression

```
>install.packages("pls")
> library(pls)
> mymodel = pcr(X..Conversion ~ ., data = mydata, scale = TRUE)
> summary(mymodel)
```

- mymodel$loadings
- mymodel$scores

REGRESSION ANALYSIS

Tackling Multi collinearity:

Method 2: Principal Component Regression

1.Perform principal component analysis on x variables
2.Use the principal components as x variables and develop the model

Cum % Variance	PC1	PC2	PC3
x	68.66	98.61	100
Conversion (y)	90.48	90.62	91.98

Component 1 or 1 & 2 may be sufficient to include in the model

REGRESSION ANALYSIS

Tackling Multi collinearity:

Method 2: Principal Component Regression

1.Perform principal component analysis on x variables
2.Use the principal components as x variables and develop the model

Loadings	PC1	PC2	PC3
Temperature	-0.674	0.218	0.705
Time	0.677	-0.2	0.709
Kappa.number	-0.296	-0.955	0

Component 1 is taking care of information in temperature and Time and
Component 2 is mostly representing kappa number

REGRESSION ANALYSIS

Tackling Multi collinearity:

Method 2: Principal Component Regression

Principal Component Scores

SL No.	Comp 1	Comp 2	Comp 3
1	-1.079	1.2498	0.1202
2	-1.158	0.9967	0.1236
3	-1.273	0.6625	0.117
4	-1.371	0.2313	0.1563
5	-1.543	-0.362	0.1756
6	-1.889	-1.365	0.1558
7	0.4709	1.1733	-0.133
8	0.3133	0.8148	-0.173
9	0.0021	0.2622	-0.299
10	-0.257	-0.122	-0.428
11	-0.268	-0.763	-0.24
12	-0.432	-1.819	-0.07
13	2.2484	0.6246	-0.022
14	2.4329	0.165	0.2963
15	2.1218	-0.388	0.1699
16	1.6801	-1.362	0.0493

REGRESSION ANALYSIS

Tackling Multi collinearity:

Method 2: Principal Component Regression

Identifying the required number of components in the model

```
> pred = predict(mymodel, type = "response", ncomp = 1)
> res = X..Conversion - pred
> mse = mean(res^2)
> prednew = predict(mymodel, type = "response", ncomp = 2)
> resnew = X..Conversion - prednew
> msenew = mean(resnew^2)
```

Statistics	Regression with	
	PC1	PC1 & PC2
MSE	12.64226	12.45593

Since there is not much reduction in MSE by including the second principal component , only PC1 is required for modelling

REGRESSION ANALYSIS

Tackling Multi collinearity:

Method 3: Partial Least Square Regression

Principal component regression involves the identification of a linear combinations of predictors that best represents the x variables

The response y is not used to help the determination of principal components

The response does not supervise the identification of principal components

Identifies the best linear combinations which best explains the predictor variables x but may not the ones best for predicting the response y

Partial least square regression is a supervised alternative to principal component regression

Partial least square method identifies the components or directions (linear combinations of x variables) using the response variable y.

Partial least square places highest weight on the variables that are most strongly related the response y

REGRESSION ANALYSIS

Tackling Multi collinearity:

Method 3: Partial Least Square Regression

R code

```
> mymodel = plsr(X..Conversion ~ ., data = mydata, scale = TRUE)
> summary(mymodel)
> mymodel$loading
```

REGRESSION ANALYSIS

Tackling Multi collinearity:

Method 3: Partial Least Square Regression

Cum % Variance	PLS1	PLS2	PLS3
x	68.65	96.92	100
Conversion (y)	90.63	90.86	91.98

Loadings	PLS1	PLS2	PLS3
Temperature	0.677	0.344	0.299
Time	-0.679	-0.207	0.607
Kappa.number	0.285	-1.391	0.736

REGRESSION ANALYSIS

Tackling Multi collinearity:

Method 3: Partial Least Square Regression

> ps = mymodel$scores

> score = ps[,1:2]

SL No	PLS1	PLS2
1	1.11324	0.89634
2	1.18502	0.73368
3	1.2913	0.51027
4	1.3792	0.25877
5	1.5361	-0.1142
6	1.85493	-0.7845
7	-0.4425	0.66627
8	-0.2949	0.40157
9	-0.0005	-0.0564
10	0.24599	-0.4059
11	0.24426	-0.6809
12	0.3833	-1.24
13	-2.2314	0.4067
14	-2.4222	0.35105
15	-2.1279	-0.1069
16	-1.7138	-0.8359

REGRESSION ANALYSIS

Tackling Multi collinearity:

Method 3: Partial least square regression

Identifying the required number of components in the model

```
> pred = predict(mymodel, data = mydata, scale = TRUE, ncomp = 1)
> res = X..Conversion - pred
> mse = mean(res^2)

> prednew = predict(mymodel, , data = mydata, scale = TRUE ,  ncomp = 2)
> resnew = X..Conversion - prednew
> msenew = mean(resnew^2)
```

Statistics	Regression with	
	PLS1	PLS11 & PLS2
MSE	12.44252	12.13185

Since there is not much reduction in MSE by including the second component , only PLS1 is required for modelling

REGRESSION ANALYSIS

Tackling Multi collinearity:

Method 4: Ridge regression

In least square regression, the coefficients β's of x variables are identified by minimizing the residual sum of squares

$$RSS = \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2$$

In ridge regression, the coefficients β's of x variables are identified by minimizing a slightly different quantity

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda\sum_{j=1}^{p}\beta_j^2 = RSS + \lambda\sum_{j=1}^{p}\beta_j^2$$

Where λ ≥ 0 is a turning parameter and $\lambda\sum_{j=1}^{p}\beta_j^2$ is the shrinkage penalty,

which will be small when $\beta_1, \beta_2, - - -, \beta_p$ are close to zero.

REGRESSION ANALYSIS

Tackling Multi collinearity:

Method 4: Ridge regression

Ridge regression seeks coefficient estimates that fit the data well by minimizing the RSS and the tuning parameter λ has the effect of shrinking the estimates β_j towards zero
The value of λ s identified through 10 fold cross validation

10 fold Cross Validation

- Divide the data set into 10 equal parts
- Develop the model using 9 parts and test it with the remaining one part
- Repeat the process 10 times to get an unbiased estimate of MSE

REGRESSION ANALYSIS

Tackling Multi collinearity:

Method 4: Ridge regression

R Code

```
> library(glmnet)
> set.seed(1)
> y = mydata[,4]
> x =mydata[,1:3]
> x = as.matrix(x)
```

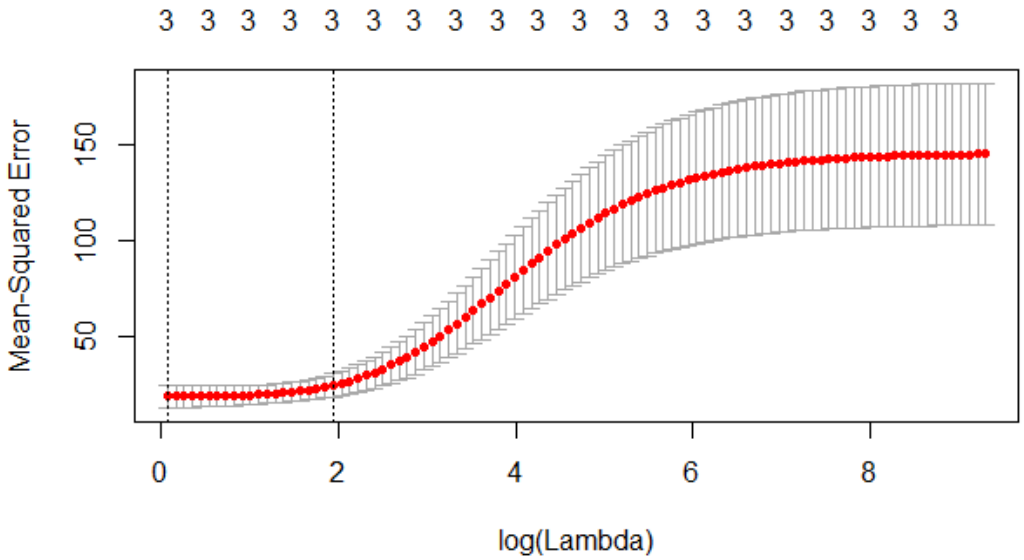
Cross Validation

```
> mymodel = cv.glmnet(x , y, alpha =0)
> plot(mymodel)
```

REGRESSION ANALYSIS

Tackling Multi collinearity:

Method 4: Ridge regression



Choose the λ which minimizes the mean square error

> bestlambda = mymodel$lambda.min

Best λ= 1.088771

REGRESSION ANALYSIS

Tackling Multi collinearity:

Method 4: Ridge regression

Develop the model with best λ and identify the coefficients

> mynewmodel = glmnet(x, y, alpha = 0)

> predict (mynewmodel, type = "coefficients", s = bestlambda)[1:4,]

Variable	Coefficients
(Intercept)	-63.0713
Temperature	0.0823
Time	-117.5048
Kappa.number	0.3268

CORRELATION & REGRESSION

Regression with dummy variables

When x's are not numeric but nominal

Each nominal or categorical variable is converted into dummy variables

Dummy variables takes values 0 or 1

Number of dummy variable for one x variable is equal to number of distinct values of that variable - 1

Example: A study was conducted to measure the effect of gender and income on attitude towards vocation. Data was collected from 30 respondents and is given in Travel_dummy_reg file. Attitude towards vocation is measured on a 9 point scale. Gender is coded as male = 1 and female = 2. Income is coded as low=1, medium = 2 and high = 3. Develop a model for attitude towards vocation in terms of gender and Income?

CORRELATION & REGRESSION

Regression with dummy variables

Variable		Dummy
Gender	Code	gender_Code
Male	1	0
Female	2	1

Variable		Dummy	
Income	Code	Income1	Income 2
Low	1	0	0
Medium	2	1	0
High	3	0	1

CORRELATION & REGRESSION

Regression with dummy variables

Read the fie and variables

- mydata = Travel_dummy_Reg
- attach(mydata)

> mydata = mydata[,2:4]

Converting categorical x's to factors

> gender = factor(Gender)

> income = factor(Income)

CORRELATION & REGRESSION

Regression with dummy variables – Output

> mymodel = lm(Attitude ~ gender + income)

> summary (mymodel)

Multiple R^2	0.8603
Adjusted R^2	0.8442
F Statistics	53.37
P value	0.00

	Estimate	Std. Error	t value	p value
(Intercept)	2.4	0.3359	7.145	0.00000
gender2	-1.6	0.3359	-4.763	0.00006
income2	2.8	0.4114	6.806	0.00000
income3	4.8	0.4114	11.668	0.00000

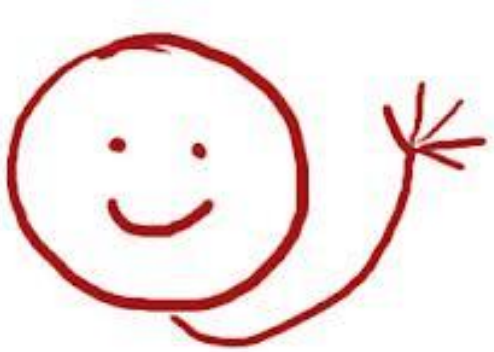
CORRELATION & REGRESSION

Regression with dummy variables – Output

> anova (mymodel)

	Df	Sum Sq	Mean Sq	F	p value
gender	1	19.2	19.2	22.691	0.0001
income	2	116.27	58.133	68.703	0.0000
Residuals	26	22	0.846		

For other queries mail me at
tanujitisi@gmail.com



New Visit: https://www.ctanujit.org/DA.html New

THANK YOU