**STATISTICS AND BASIC ECONOMETRICS**

**MPBA G505**

**STATISTICS REPORT**



**Submitted to:**

**Dr. Udayan Chanda**

**Submitted by:**

|  |  |
| --- | --- |
| **Name** | **BITS ID** |
| Ashish Karanje | 2021H1540857P |

**INDEX**

|  |  |  |
| --- | --- | --- |
| **Sr. No.** | **Title** | **Page No.** |
|  | Assignment 1a | 1 |
| 2. | Assignment 1b | 9 |
| 3. | Assignment 2 | 17 |
| 4. | Lab Assignment 2a | 25 |
| 5. | Lab Assignment 2b | 26 |
| 6. | Assignment 3 | 27 |
| 7. | Lab Assignment 3a | 29 |
| 8. | Lab Assignment 3b | 31 |
| 9. | Assignment 4 | 33 |
| 10. | Lab Assignment 4a | 35 |
| 11. | Lab Assignment 4b | 35 |
| 12. | Assignment 5 | 37 |
| 13. | Lab Assignment 5a | 41 |
| 14. | Assignment 6 | 43 |
| 15. | Lab Assignment 6a | 47 |
| 16. | Lab Assignment 6b | 47 |
| 17. | Assignment 7 | 49 |
| 18. | Lab Assignment 7a | 55 |
| 19. | Lab Assignment 7b | 55 |
| 20. | Assignment 8 | 57 |
| 21. | Lab Assignment 8a | 61 |
| 22. | Lab Assignment 8b | 64 |

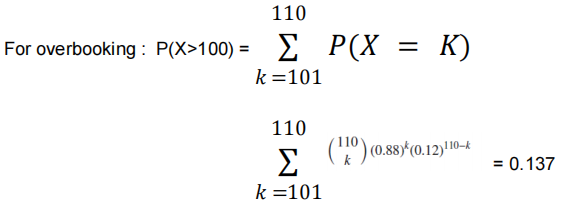
**1.Title of the Practical 2(a): Binomial Distribution**

**2.Problem/Questions:** According to Leder et al. (2002), many airlines consistently report that about 12% of all booked passengers do not show up to the gate due to cancellations and no-shows. If an airline sells 110 tickets for a flight that seats 100 passengers, what is the probability that the airline overbooked (sold more tickets than seats) ?

For a 100-seat flight, suppose the airline would like to sell the maximum number of tickets such that the chance of overbooking is less than 5%. The airlines call this a “5% bump threshold” overbooking strategy. How many tickets should the airline sell? Simulate the probability of overbooking when 108 tickets are sold.

**3.Solution : Formulas :**

Let us assume that the passenger arrival at the gate is independent, then X be the binomial distribution with n = 110 and p = 1- 0.12 = 0.88 .



**4.Code:**

# Setting the seed in order to produce the same results.

**1.**set.seed(1)

#P(X > 100) = 1 - P(X<= 100)

**2.**pb\_over\_bkd<- 1- pbinom(100,110,0.88) sprintf("The probability that the airline overbooked is :%f",pb\_over\_bkd)

#Finding the value of 'n' so that P(X > 100) <= 0.05 , where X is the binomial probability distribution function.

sprintf("The total cost of an accident for an year is:%d ",total\_cst)

#"The total cost of an accident for an year is:2225000 "

#Simulating for 1000 scenarios.

smlt\_list <- replicate(1000,sum(25000\*rpois(12,7.5))) avg\_cost\_y<- mean(smlt\_list)

sprintf("The average cost of accidents over a year's time is :%d",avg\_cost\_y)

**3.** for(x in 101:110){ if ((1-pbinom(100,x,0.88)) <

0.05) n <- x }

sprintf("The value of n so that P(X>100) <= 0.05 is: %s",n)

# "The value of n so that P(X>100) <= 0.05 is: #108".

#Simulating the probability of overbooking when #108 tickets are sold.

**4.**chk\_ob <- replicate(10000,if (rbinom(1,108,0.88) >100 ) 1 else 0)

**5.**mean\_chk\_ob<-mean(chk\_ob)

**6.**sprintf("The probability of overbooking when 108 tickets are sold:%f",mean\_chk\_ob)

# "The probability of overbooking when 108 tickets are sold:0.046600" .

# "The average cost of accidents over a year's time is: 2251900"

**5.Inference:**

Our goal is to find P(X>100) <= 0.05, Using trial and error we found n =108. In order to simulate the probability of overbooking when 108 tickets are sold, we approach the problem using the Binomial random variable with n =108 and p = 0.88. We simulate this condition on 10,000 trials.

**Title of the Practical 2(b): Poisson Distribution**

**Problem/Questions:** The number of accidents per month at a busy intersection has a Poisson distribution with parameter λ = 7.5. Conditions at the intersection have not changed much over time. Suppose each accident costs the local government about $25,000 for clean-up. How much do accidents cost, on average, over a year’s time?

**Solution : Formulas :**

The poisson distribution function is given by f(x) =



Here, λ represents the average number of accidents per month.

**Code:**

# Setting the seed in order to produce the same results.

**1.**set.seed(1)

# The number of accidents for 12 months forms a Poisson's distribution with lambda= 7.5. 2.accidents <- rpois(12,7.5)

3.total\_cst <- sum(25000\*accidents)

**4.Code :**

# Setting the seed in order to produce the same results.

**1.**set.seed(1)

#P(X > 100) = 1 - P(X<= 100)

**2.**pb\_over\_bkd<- 1- pbinom(100,110,0.88) sprintf("The probability that the airline overbooked is :%f",pb\_over\_bkd)

#Finding the value of 'n' so that P(X > 100) <= 0.05 , where X is the binomial probability distribution function.

sprintf("The total cost of an accident for an year is:%d ",total\_cst)

#"The total cost of an accident for an year is:2225000 "

#Simulating for 1000 scenarios.

smlt\_list <- replicate(1000,sum(25000\*rpois(12,7.5))) avg\_cost\_y<- mean(smlt\_list)

sprintf("The average cost of accidents over a year's time is :%d",avg\_cost\_y)

# "The average cost of accidents over a year's time is : 2251900"

**Inference:** ‘rpois’ function generates Poisson random variable values from the Poisson distribution and returns the results. Then, we used the values returned to calculate the total cost of the accidents. Similarly , we created 1000 scenarios and for each scenario calculated the cost of accidents over a year’s time. And from the list of the values we calculated the average cost of the accidents.

**ASSIGNMENT 3**

**Name :** Ashish Karanje

**BITS ID:** 2021H1540844P

**Date:** 6st September’ 2021

**Regression Analysis on ceosal1 dataset**

**Exercise:** Perform linear regression to obtain best model.

**Data Source:** ceosal1

**Theory:** Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable(Y). The variable you are using to predict the other variable's value is called the independent variable(X).

A linear regression line has an equation of the form

Y = B0+B1X , where X is the explanatory variable and Y is the dependent variable. The slope of the line is B1 and B0 is the intercept.

B1= covariance (X AND Y) and B0 = mean(Y) – B1\*mean(X) .

variance

**Code:**

var = var(ceosal1$lsalary) #calculating variance

cov = cov(ceosal1$lsalary, ceosal1$salary) #calculating covariance

B1 = cov/var

B0 = mean(ceosal1$salary)-B1\*mean(ceosal1$lsalary) #calculating intercept

ceosal = lm(ceosal1$salary~ceosal1$lsalary) #lm is used to fit linear model

plot(ceosal1$lsalary, ceosal1$salary, ylim = c(0,4000))

abline(ceosal , col ="blue")

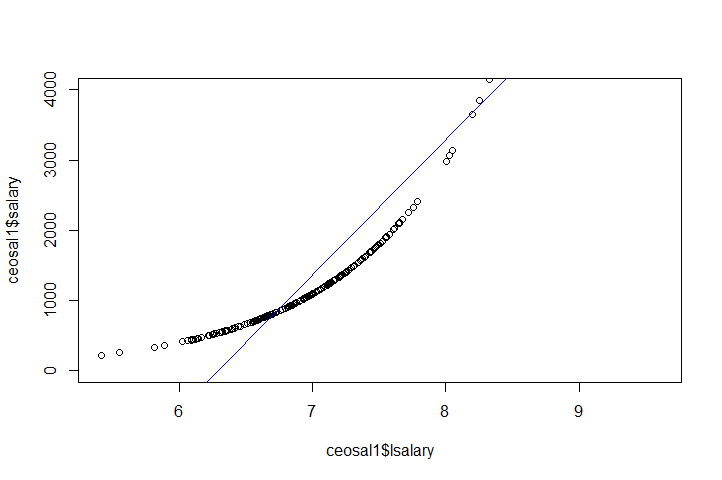
names(ceosal)

sal.hat = ceosal$fitted.values

u.hat = ceosal$residuals

cbind(ceosal1$lsalary,ceosal1$salary,sal.hat,u.hat) #binding dependent and independent variables

summary(ceosal) # gives summary of linear regression



Here, the dependent variable Y (salary) remains the same in every model, while the independent variable used in each model is provided in tabular format below.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model No. | X | Y | R-squared Value | Explained % |
| 1 | pcsalary | salary | 7.52E-05 | 0.007521 |
| 2 | sales | salary | 0.01437 | 1.437 |
| 3 | roe | salary | 0.01319 | 1.319 |
| 4 | pcroe | salary | 0.0008243 | 0.08243 |
| 5 | ros | salary | 0.001134 | 0.1134 |
| 6 | lsalary | salary | 0.6308 | 63.08 |
| 7 | lsales | salary | 0.03767 | 3.767 |

**Inference:**

R-squared value is very low for models except model 6 (lsalary)

the R-squared value of Isalary is 0.6308 , This implies that only 63.08% of variation in the salary is explained by Isalary.

**Assignment 3(a)- The Simple Regression Model**

**Problem/Questions:** Find the R-Square values of the linear regression model with

different sets of input features.

**Theory:** The linear regression model is:

y= β\_0 + β\_1 \* x + α where β\_0 and β\_1 are

the parameters of the model., and α is the

random variable also known as error.

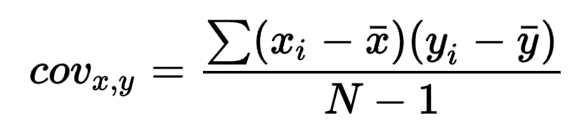
E[Y] = E[β\_0] + E[β\_1 \*x] + E(α)

E[Y] = β\_0 + β\_1 \* E[x] + 0

y\_hat = β\_0 + β\_1 \* x

β\_0\_hat = y\_hat - β\_1\_hat \*x\_bar

β\_1\_hat = cov(x,y)/var(x)



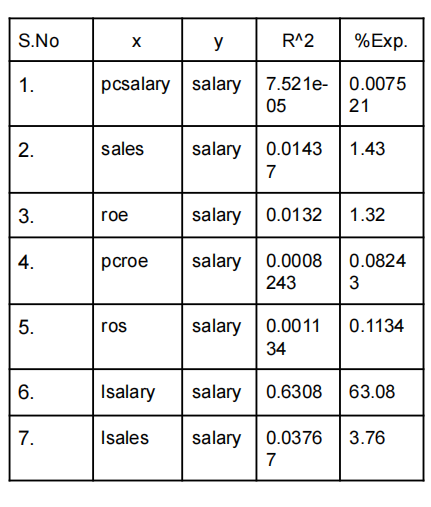
N = Sample Size

**Code:**

library(wooldridge)

ceo\_sal1 <- wooldridge::ceosal1

View(ceo\_sal1)



y <- ceo\_sal1$salary

co\_var <- cov(x,y)

y\_bar <- mean(y)

x\_bar <- mean(x)

beta\_1 <- co\_var/var\_a

beta\_o <- y\_bar - beta\_1\*x\_bar

#Salary\_hat = 963.1913 + 18.50119 \* roe#

lm\_model <- lm(ceo\_sal1$salary~x)

plot(x,y,ylim = c(0,4000))

abline(lm\_model)

## [1] "coefficients" "residuals" "effects"

"rank"

#Find the coefficients of the model.

lm\_model$coefficients

## (Intercept) x

## 963.19134 18.50119

#Fitted\_values = salary\_hat

sal.hat <- lm\_model$fitted.values

#Residuals values.

u.hat <- lm\_model$residuals

cbind(y,x,sal.hat,u.hat)

## y x sal.hat u.hat

## 1 1095 14.1 1224.0581 -129.0580710

## 2 1001 10.9 1164.8543 -163.8542606

## 3 1122 23.5 1397.9692 -275.9692156

## 4 578 5.9 1072.3483 -494.3483377

# 0 < R\_square < 1

# R\_square should be closer to 1 .

**summary(lm\_model)**

**##** **Residuals:**

## Min 1Q Median 3Q Max

## -1160.2 -526.0 -254.0 138.8 13499.9

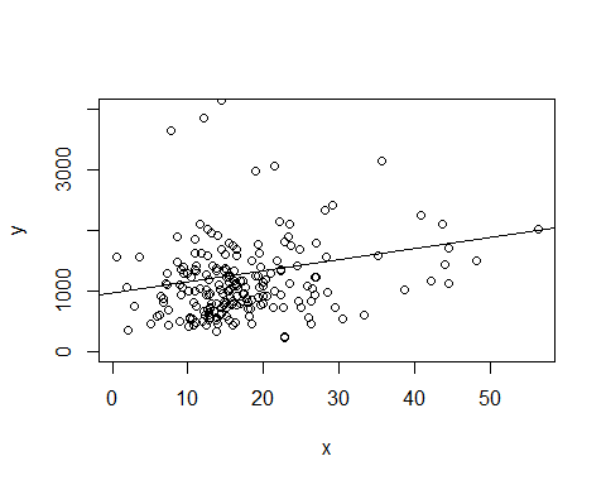
**##** **Coefficients:**

## Estimate Std. Error t value Pr(>|t|)

## (Intercept) 963.19 213.24 4.517 1.05e-05 \*\*\*

## x 18.50 11.12 1.663 0.0978 .

**## Multiple R-squared: 0.01319 .**



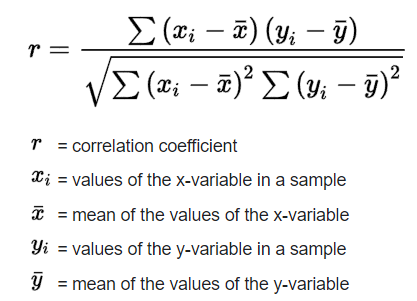
**Inference:** Isalary has the R^2 as 0.6308. This implies that 63.08 % variation in salary is

explained by Isalary.

**Title of the Practical 3 (b): Regression Analysis using R.**

**Problem/Questions:** Find the coefficients and various parameters of a simple regression model using the dataset library wooldridge.

**Correlation:**  It is used to find the degree of strength between two variables



The equation of MLR model is :

Y = β\_0 + β\_1\*X + U\_i (error term)

**Assumption :**

1. U\_i is a random variable(scattered)

2. U\_i follows normality with mean 0 and standard deviation σ

E[Y] = E[β\_0] + E[β\_1\*x] + E(U\_i)

E[Y] = β\_0+ β\_1\* E[X] + 0

Y\_hat = β\_0+ β\_1\* X

We want to minimize : Σ(Y-Y\_hat)^2.

To do that we will use the ordinary least square method.

β\_0 \_hat = Y\_hat - β\_1\_hat\* x\_bar.

Β\_1\_hat = cov(x,y)/var(x)

**Data\_Set:** wooldridge::vote1

**Code :**

library(wooldridge)

#Vote Data set

vote1 <- wooldridge::vote1

#Get the shareA of the dataset

x\_vote\_a <- vote1$shareA

#Get the shareB of the dataset

y\_vote\_a <- vote1$voteA

#Find the co\_variance between x and y co\_var\_vote\_a <- cov(x\_vote\_a,y\_vote\_a)

**[1] 520.0194**

#Find the variance of x

var\_a\_vote\_a <- var(x\_vote\_a)

**[2] 1121.15**

y\_bar\_vote\_a <- mean(y\_vote\_a)

x\_bar\_vote\_a <- mean(x\_vote\_a)

beta\_o\_vote\_a <- y\_bar\_vote\_a - beta\_1\*x\_bar\_vote\_a

beta\_1\_vote\_a <- co\_var\_vote\_a/var\_a\_vote\_a

beta\_1\_vote\_a

**[3] 0.4638**

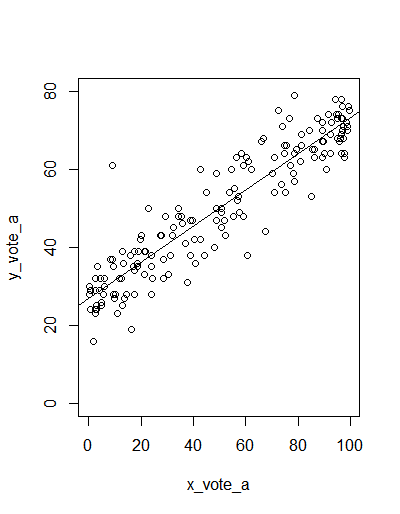
beta\_o\_vote\_a

**[4] 26.81221**

**lm\_model\_vote <- lm(vote1$voteA~vote1$shareA)**

**plot(x\_vote\_a,y\_vote\_a,ylim = c(0,80))**

**abline(lm\_model\_vote)**



**lm\_model\_vote$coefficients**

(Intercept) vote1$shareA

26.8122141 0.4638269

**Inference:** There is a covariance between x and y and the linear regression model doesn’t fit the data completely. There are residuals between the actual value and the predicted values.

ASSIGNMENT 4

Name : Ashish Karanje

**BITS ID:** 2021H1540857P

**Exercise** – Perform ANOVA and analyze significant difference using Tukey test.

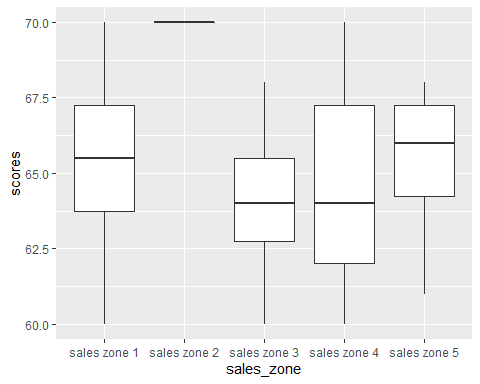
**Theory** - Analysis of variance (ANOVA) is a collection of statistical models and their associated estimation procedures (such as the "variation" among and between groups) used to analyze the differences among means.

The one-way ANOVA can help you know whether or not there are significant differences between the means of your independent variables (such as the first example: age, sex, income).

One-way ANOVA between groups: used when you want to test two groups to see if there’s a difference between them.

Two-way ANOVA without replication: used when you have one group and you’re double-testing that same group.

Tukey's range test, also known as Tukey's test, is a single-step multiple comparison procedure and statistical test. It can be used to find means that are significantly different from each other.

library(ggplot2)  
library(dbplyr)  
  
sales =read.csv("E:/MBA BA/G505 - Statistics and Econometrics/Sales.csv")  
library(ggplot2)  
ggplot(sales,aes(x=sales\_zone, y=scores))+geom\_boxplot()

sales.lm = lm(sales$scores~sales$sales\_zone,data=sales)  
sales.lm

##   
## Call:

lm(formula = sales$scores ~ sales$sales\_zone, data = sales)

Coefficients:

(Intercept) sales$sales\_zonesales zone 2 sales$sales\_zonesales zone 3

6.537e+01 4.625e+00 -1.250e+00

sales$sales\_zonesales zone 4 sales$sales\_zonesales zone 5

-7.500e-01 9.104e-15

sales.aov<-aov(sales.lm)  
summary(sales.aov)

## Df Sum Sq Mean Sq F value Pr(>F)

sales$sales\_zone 4 177.1 44.27 6.186 0.000712 \*\*\*

Residuals 35 250.5 7.16

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

sales.test=TukeyHSD(sales.aov)  
sales.test

##Tukey multiple comparisons of means

## 95% family-wise confidence level

##Fit: aov(formula = sales.lm)

$`sales$sales\_zone`

diff lwr upr p adj

sales zone 2-sales zone 1 4.625000e+00 0.7791989 8.4708011 0.0118332

sales zone 3-sales zone 1 -1.250000e+00 -5.0958011 2.5958011 0.8814981

sales zone 4-sales zone 1 -7.500000e-01 -4.5958011 3.0958011 0.9798420

sales zone 5-sales zone 1 1.421085e-14 -3.8458011 3.8458011 1.0000000

sales zone 3-sales zone 2 -5.875000e+00 -9.7208011 -2.0291989 0.0008885

sales zone 4-sales zone 2 -5.375000e+00 -9.2208011 -1.5291989 0.0025798

sales zone 5-sales zone 2 -4.625000e+00 -8.4708011 -0.7791989 0.0118332

sales zone 4-sales zone 3 5.000000e-01 -3.3458011 4.3458011 0.9956496

sales zone 5-sales zone 3 1.250000e+00 -2.5958011 5.0958011 0.8814981

sales zone 5-sales zone 4 7.500000e-01 -3.0958011 4.5958011 0.9798420

Inference:

H0 : Mean of all factors are the same.

H1 : Mean of any one of the factor is different from the other factor.

Running regression model, gives us a p-value of 0.000712 so we reject the Null Hypothesis. It implies that there are differences in the means of the factor.

But we can’t comment on where the differences lie.

So, we perform the Tukey Test. Tukey test tells us that the difference between

1. sales zone-3 & sales zone-2 (diff = -5.875, p -value = 0.000885)

1. sales zone 4 & sales zone 2(diff = -5.375, p- value = 0.0025)
2. sales zone 5 & sales zone 2 (diff= -4.625, p-value = 0.0118)

are statistically significant.

**Assignment 4(a) : One way ANOVA**

**Problem/Questions: Perform Analysis of Variance.**

**Theory: ANOVA :** Analysis of variance (ANOVA) is an analysis tool used in statistics that splits an observed aggregate variability found inside a data set into two parts: systematic factors and random factors. The systematic factors have a statistical influence on the given data set, while the random factors do not. Analysts use the ANOVA test to determine the influence that independent variables have on the dependent variable in a regression study.If no true variance exists between the groups, the ANOVA's F-ratio should equal close to 1.

**Code :**

**city** <- c("D","D","D","D","D","D","M","M","M","M","M","M","K","K","K","K","K","K","C","C","C","C","C","C")

**price** <- c(22,22.5,21.5,22,22.5,21.5,19,19.5,19,20,19,21,18,17,18.5,17,18.5,17,21,20,21.5,20,21,20)

**aov.mod <- aov(price~city)**

**summary(aov.mod)**



**Inference :** Since the F\_stat >>1 and the corresponding p\_value is <<0.05 , we will reject the null hypothesis i.e. there exists sufficient statistical evidence that true variance exists between price and city variables.

**Assignment 4(b):** One - Way ANOVA (F-Test Statistics)

**1.Problem/Questions:** A company is in the process of launching a new product. Before

launching, the company wants to ascertain the status of its product as a second alternative. For doing so, the company prepared a questionnaire consisting of 20 questions on a five-point rating scale with 1 being “strongly disagree” and 5 being “strongly agree.” The company administered this questionnaire to 8 randomly selected respondents from five potential sales zones. The scores obtained from the respondents are given in the table. Analyze the Significant difference in the scores.

**2.Theory:**

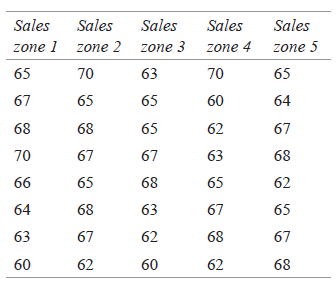
ANOVA is a technique of testing the hypothesis about the significant difference in several population means.

Suppose if there is “z” samples being analyzed, then the null and alternative hypothesis can be set:

H0 : mu1 = m2= …= (mu)n

H1: Not all mu are equal.

**3.Data:**

****

**4.Code :**

**#Linear Regression Model.**

model\_reg <- lm(sales\_csv$Value~sales\_csv$Sales\_Zone)

**#Summary of the linear model.** summary(model\_reg)

**## Multiple R-squared: 0.09223,**

**##Adjusted R-squared: -0.01152**

**## F-statistic: 0.889 on 4 and 35 DF, p-value: 0.4807**

**#One-way Anova model.**

**data.aov <- anova(model\_reg)** **TukeyHSD(data\_tukey)**

**Diff**  **lwr**  **upr**  **p\_adj**

1.125 -2.898067 5.148067 0.9276433

-1.250 -5.273067 2.773067 0.8974526

**Inference :** F\_Stat:0889 with the corresponding p-value : 0.487, this implies that we cannot reject Null Hypothesis. Also there is sufficient evidence that the population mean of all the levels are equal.

**T-test:** The t-values of the independent variable is less than 1, this implies that sales zones don’t effect the value of scores.

**Assignment 5(a):** Analysis of Covariance(ANCOVA)

**1.Problem/Questions:** Perform the ANCOVA on the FETUSA data set and explain the findings.

**2.Data :** FETUSA

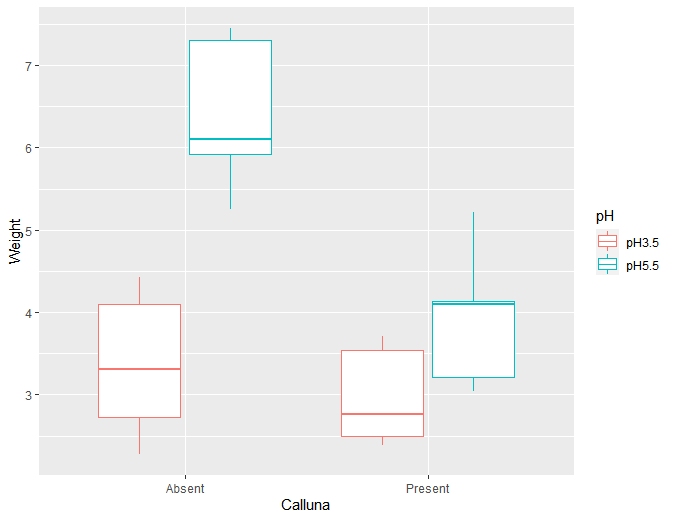
**3.Code:**

**library(dplyr)**

**library(ggplot2)**

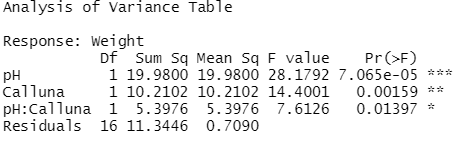
**festuca <- read.csv("FESTUCA.CSV")**

**ggplot(data= festuca,aes(x=Calluna, y = Weight, color = pH)) + geom\_boxplot()**

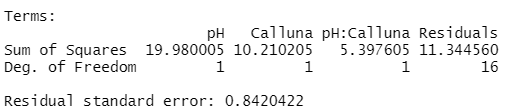
****

**festuca\_model** <- **lm(Weight ~ pH + Calluna + pH : Calluna , data = festuca)**

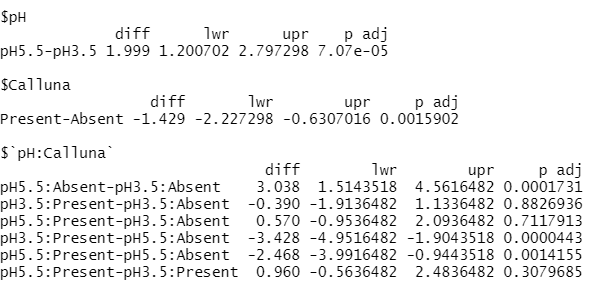
**festuca\_model.anov** <- **anova(festuca\_model)**



**festuca\_model.aov** <- **aov(festuca\_model)**

****

**TukeyHSD(festuca\_model.aov**



**Inference:** From the box plot we can come to the conclusion that the median value of weight in case of Calluna : Absent and ph value : 5.5 is higher compared to the rest of the cases.

From the ANOVA table, we can come to the following conclusion: (i) In case of pH , the F\_stat is 28.17 and the corresponding p-value is 7.065e-05 , hence we will be rejecting the null hypotheses. This means that there is statistical evidence that the means of the different unrelated groups of pH are not equal.

(ii) In case of Calluna , the F\_stat is 14.041 and the corresponding p-value is 0.00159 , this indicates that there is statistical evidence that the means of the different unrelated groups of Calluna are not equal.

In case of pH: Calluna , the F\_stat : 7.62 and the p-value : 0.0139 , this implies that there is not much variation that can be explained by the interaction between pH and Calluna.

Since the overall F\_stat is 16.73 and the corresponding p\_value is 3.447e-05 , this signifies the overall significance of the model and hence we can proceed for the Tukey Test.

The Tukey test signifies that the mean difference between pH5.5:Absent and pH3.5:Absent,

pH3.5:Present-pH5.5:Absent and pH5.5:Present-pH5.5:Absent is statistically significant.

**Assignment 6(a): The Simple regression model & VIF**

**Regression Model & VIF**

**Problem/Questions:** Find the R-Square

values of the linear regression model with

different sets of input features and also interpret the value of VIF.

**Code :**

**library(wooldridge)**

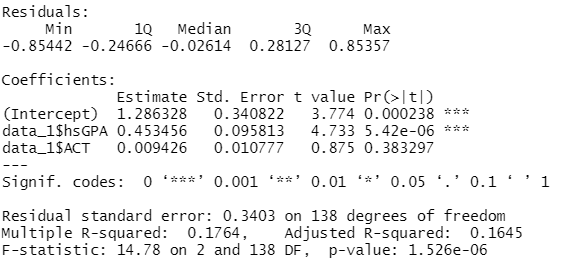
**library(dplyr)**

**data\_1** <- wooldridge::gpa1

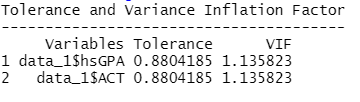
glimpse(**data\_1**)

**GPAres**<- **lm**(data\_1$colGPA ~ data\_1$hsGPA + data\_1$ACT,data = data\_1)

summary(**GPAres**)



**ols\_coll\_diag**(GPAres)



**Inference : Step 1 :**

Residual Analysis : The min and max value indicate that there is no skewness present in the distribution of the error term. Overall the residuals look to have a fairly normal distribution.

**F-test :** F\_statistics 14.78 and its corresponding **p-value** is 1.52e^-06. This implies that the overall model is significant and there is at least one coefficient in the model that is not zero. This leads us to reject the null hypothesis and conclude that there is strong evidence that a relationship does exist between the independent variable and dependent variable.

**t-test :** **t-stat** reveals that the MLR model has only one significant independent variable . This also means that there is enough statistical evidence that the standard error of these variables is small

**R\_square :** The Multiple R-Squared value : 0.1764 indicates that only 17% variation in the dependent variable is explained by the independent variables. This also means that the model is not fitting the data very well.

**VIF :** This value indicates that there is no presence of multicollinearity i.e. the independent variable is not correlated with other variables.

**Title of the Practical 6 (b) : The Simple Regression Model & VIF**

**Problem/Questions:** Find the R-Square

values of the linear regression model with

different sets of input features.

**Code :**

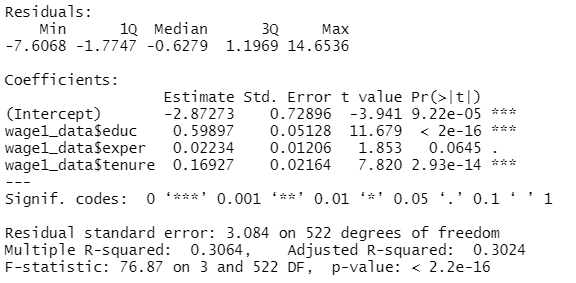
library(olsrr)

**wage1\_data <- wooldridge::wage1**

glimpse(wage1\_data)

wage\_mode<**lm**(wage1\_data$wage~ wage1\_data$educ+wage1\_data$exper+wage1\_data$tenure)

**summary**(lm(wage1\_data$wage~ wage1\_data$educ+wage1\_data$exper+wage1\_data$tenure))



**ols\_coll\_diag**(wage\_model)

Second Data

**Problem/Questions:** Find the R-Square values of the linear regression model with different sets of input features.

**Code :**

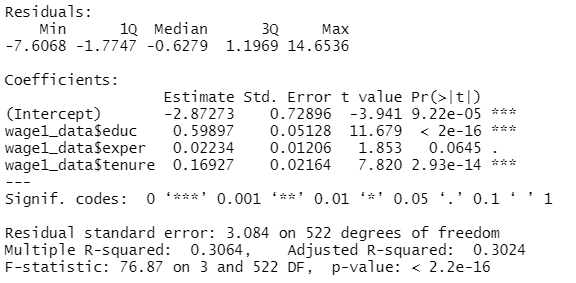
library(olsrr)

**wage1\_data <- wooldridge::wage1**

glimpse(wage1\_data)

wage\_model <- **lm**(wage1\_data$wage ~ wage1\_data$educ+wage1\_data$exper+wage1\_data$tenure)

**summary**(lm(wage1\_data$wage ~ wage1\_data$educ+wage1\_data$exper+wage1\_data$tenure))



**ols\_coll\_diag**(wage\_model)

**Inference : Step 1 :**

Residual Analysis : The min and max value indicate that there is no skewness present in the distribution of the error term. Overall the residuals look to have a fairly normal distribution.

**F-test :** F\_statistics : 76.87 and its corresponding **p-value** is 2.26e^-16. This implies that the overall model is significant and there is at least one coefficient in the model that is not zero. This leads us to reject the null hypothesis and conclude that there is strong evidence that a relationship does exist between the independent variable and dependent variable.

**t-test:** **t-stat** reveals that the MLR model has three significant independent variables. This also means that there is enough statistical evidence that the standard error of these variables is small in comparison to coefficients.

**R\_square:** The Multiple R-Squared value: 0.3064 indicates that only 30% variation in the dependent variable is explained by the independent variables. This also means that the model is not fitting the data very well.

**VIF:** This value indicates that there is no presence of multicollinearity i.e. the independent variable is not correlated with other variables.

**Title of the Practical 7(a) : Heteroskedasticity & Autocorrelation**

**Problem/Questions:** Check for heteroskedasticity and autocorrelation in the dataset.

**Theory: Heteroskedasticity:** refers to the error variance, or dependence of scattering, within a minimum of one independent variable within a particular sample.

In statistics, heteroskedasticity happens when the standard errors of a variable, monitored over a specific amount of time, are non-constant.

**Autocorrelation:** It represents the degree of similarity between a given time series and a lagged version of itself over successive time intervals.

The **Durbin-Watson** always produces a test number range from 0 to 4. Values closer to 0 indicate a greater degree of positive correlation, values closer to 4 indicate a greater degree of negative autocorrelation, while values closer to the middle suggest less autocorrelation.

**Bptest: Performs the Breusch-Pagan test against heteroskedasticity.**

**Data :**  wooldridge::hprice1

**Code:**

library(wooldridge)

library(lmtest)

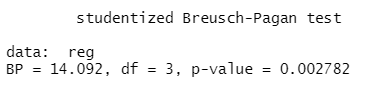
data\_h\_price <- wooldridge::hprice1

reg <- lm(price~lotsize+sqrft+bdrms,data = data\_h\_price)

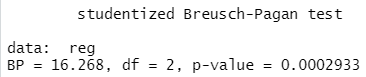
bptest(reg)

#Interaction between y\_hat and y\_hat\_square is also considered in the white test.

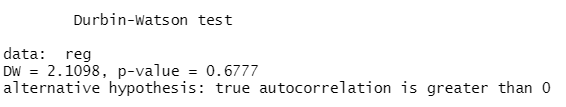
bptest(reg,~ fitted(reg)+I(fitted(reg)^2))

****

bptest(reg,~ fitted(reg)+I(fitted(reg)^2))



dwtest(reg)



**Inference :** We first performed the BPtest and the statistics are : BP = 16.268 with p-value = 0.002782, hence we will be rejecting the NULL hypothesis i.e there is enough evidence that there is heteroskedasticity in the model.

Then we performed the Durbin-Watson test , and the statistics are : DW = 2.1098 with p-value = 0.677 , hence we fail to reject the NULL hypothesis i.e. there is enough statistical evidence that there is no autocorrelation in the model.

**Title of the Practical 7(b) : Heteroskedasticity & Autocorrelation**

**Problem/Questions:** Check for heteroskedasticity and autocorrelation in the dataset.

**Data:** wooldridge::rental

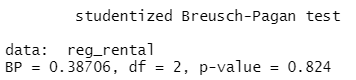
**Code :** data\_rental <- wooldridge::rental

**#rent is dependent,population is independent and #enrol is independent.**

**reg\_rental** <- **lm**(rent~pop+enroll,data = data\_rental)

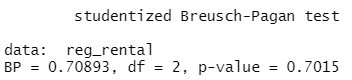
#homoskedastic

**bptest**(reg\_rental)



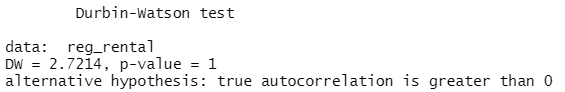
#Interaction between y\_hat and y\_hat\_square is also considered in the white test.

**bptest**(reg\_rental,~fitted(reg\_rental)+I(fitted(reg\_rental)^2))



#dwtest

dwtest(reg\_rental)



**Inference :** We first performed the BPtest and the statistics are : BP = 16.268 with p-value = 0.82, hence we will not be rejecting the NULL hypothesis i.e there is enough evidence that there is homoscedasticity in the model.

Then we performed the Durbin-Watson test , and the statistics are : DW = 2.1098 with p-value = 1 , hence we fail to reject the NULL hypothesis i.e. there is enough statistical evidence that there is no autocorrelation in the model.

**Practical 8(a): Multicollinearity**

**Problem/Questions: Analyse the dataset for multicollinearity.**

**Theory:**

**Data: Wage dataset.**

**Code :**

library(dplyr)

library(readxl)

library(GGally)

library(corpcor)

library(mctest)

library(ppcor)

library(car)

library(wooldridge)

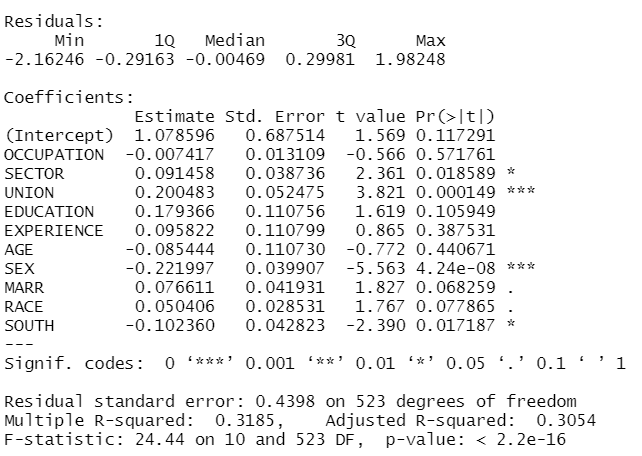
library(ggplot2)

**data\_wage <- read.csv('multicol.csv')**

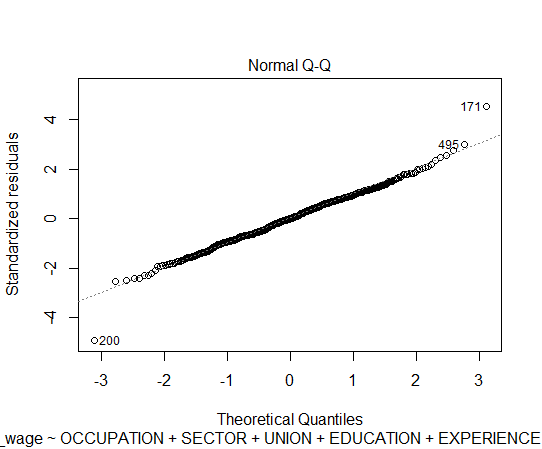
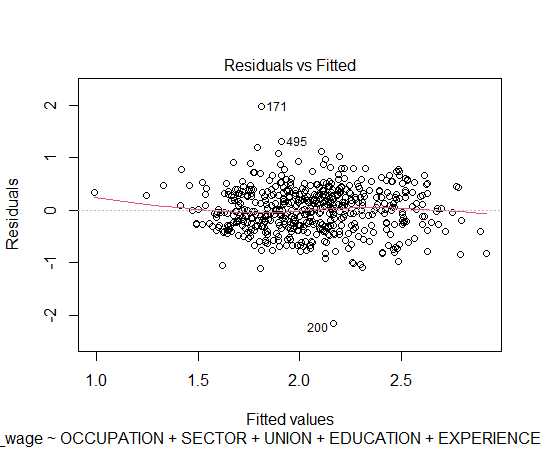
log\_wage <- log(data\_wage$WAGE)

**model\_lm <- lm(log\_wage ~ OCCUPATION+SECTOR+UNION+EDUCATION+EXPERIENCE+AGE+SEX+MARR+RACE+SOUTH,data = data\_wage)**

**summary(model\_lm)**

****

**#Since only four parameters are significant , then #the model is not significant.**



######**Checking for Multicollinearity**#########

x <- data\_wage[,2:11]

ggpairs(x)

**#We can plot the model diagnostics checking.**

par(mfrow = c(2,2))

plot(model\_lm)

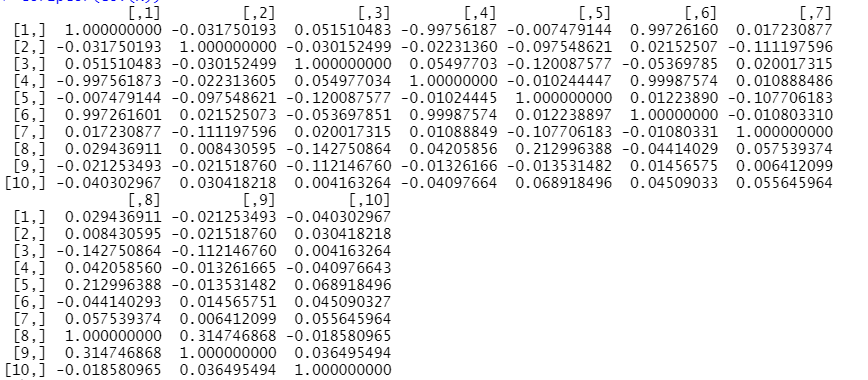
**Residual is not following any pattern.**

**#There is no heteroskedasticity . The data is homoscedastic.**

#The high correlation matrix shows that pair-wise correlation among all the explanatory variables are not very high , except age-experience.

#The high correlation between age and experience.

**cor2pcor(cov(x))**



**# H0 : p = 0**

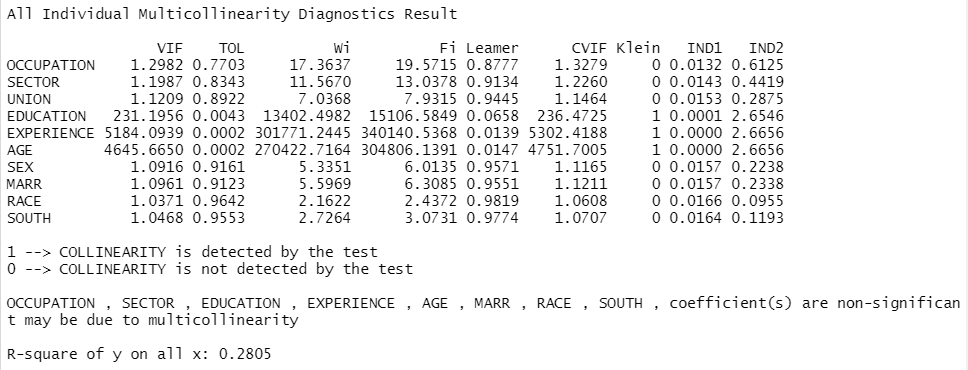
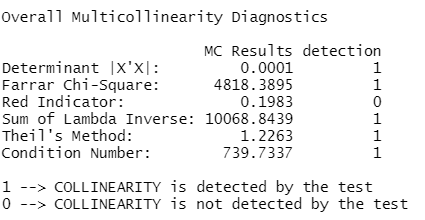
**# HA : p not equal to 0.**

**#Farrar-Glauber Test**

**#Whether there exists a multicollinearity or not .**

model\_w <- **lm(WAGE ~ OCCUPATION+SECTOR+UNION+EDUCATION+EXPERIENCE+AGE+SEX+MARR+RACE+SOUTH,data = data\_wage)**

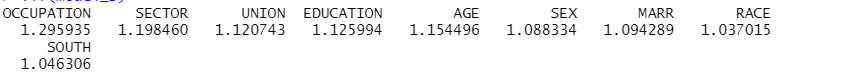
**omcdiag(model\_w)**

****

**pcor(x,method = 'pearson')**

**imcdiag(mod = model\_w)**

vif(model\_s)

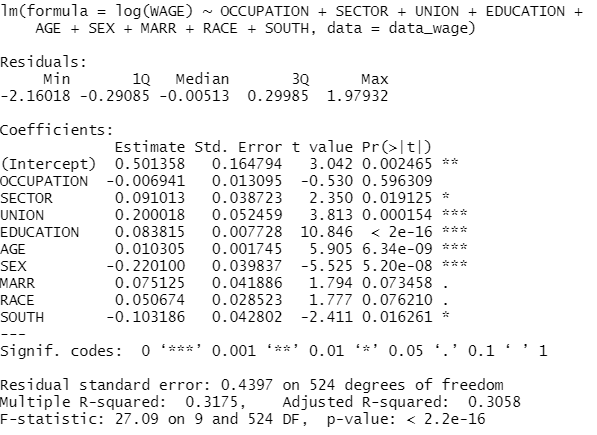


**#Remedial measures**

**model\_s<-**

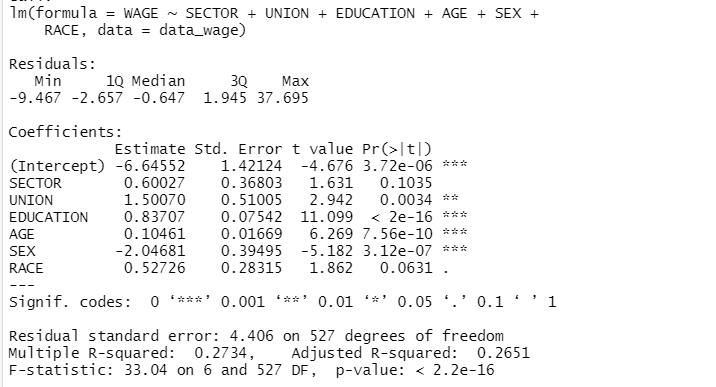
**lm(log(WAGE~ OCCUPATION+SECTOR+UNION+EDUCATION+AGE+SEX+MARR+RACE+SOUTH,data= data\_wage)**

**summary(model\_s)**

****

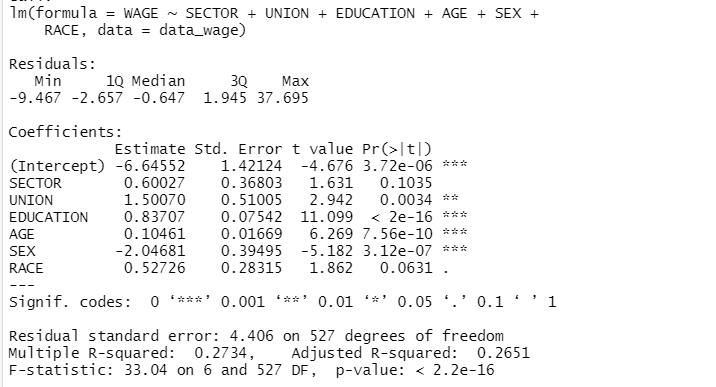
**model\_t** <- **lm(WAGE ~ SECTOR+UNION+EDUCATION+AGE+SEX+RACE,data = data\_wage)**

**summary(model\_t)**



**model\_t** <- **lm(WAGE ~ SECTOR+UNION+EDUCATION+AGE+SEX+RACE,data = data\_wage)**

**summary(model\_t)**



**Inference :**

**Step 1 :**

**F-test** For the model:

**lm**(log\_wage~OCCUPATION+SECTOR+UNION+EDUCATION+EXPERIENCE+AGE+SEX+MARR+RACE+SOUTH) , **F\_stat : 24.44 with p\_value : 2.2e-16**. This implies that there is enough evidence to indicate the overall significance of the model i.e. we will be rejecting the null hypotheses (H\_0 = β1=β2=..=0). This doesn’t mean all the independent variables are useful .

**Step 2: t-test :** For the same model , t\_test indicates that only 6 independent variables are significant. This can be due to the multicollinearity present between the independent variables.

**R\_square : 0.3185**. This means that approx. 32 % of the variation in wage is explained by the independent variables.

**Step 3 : Check for normality :** The Normal Q-Q plot indicates all other data points follow the straight line, hence the error term has normal distribution.

**Step 4: Check for Multicollinearity :**

Using the ggpair plot we get the following result:

**Age : Experience** **: 0.974223**

There is a presence of high multicollinearity between Age and Experience pair.Furthermore, we can use partial correlation coefficient matrix , it is clear that the partial correlation coefficient between Age and Experience are quite high which support the findings from the ggpairs plot.

We did a **Farrar Glauber Test** , which detected coefficients can be non- significant due to multicollinearity.

For remedial measures, we will first check the VIF of the MLR and the ones with more that 10 VIFvalue will be removed. Hence ,Experience is removed from the MLR model.

**Step 5 : Summary of the New model**

The number of significant coefficients of independent variables increased to 5.

The F-test indicates that the overall model is still significant.

The t-test indicates that the 5 coefficients of the independent variables are significant.

The **R\_square** has decreased a little : from 0.318 to 0.27. The **Adjusted\_R-squared** has increased a little : 0.7625 to 0.7719 .

**Now the model is free from multicollinearity. Log is used only when checking for #heteroskedasticity.**

**Title of the Practical 8(b): Autocorrelation**

**Problem/Questions: Analyse the dataset for Autocorrelation.**

**Theory:**

**Autocorrelation :** It represents the degree of similarity between a given time series and a lagged version of itself over successive time intervals.

The **Durbin-Watson** always produces a test number range from 0 to 4. Values closer to 0 indicate a greater degree of positive correlation, values closer to 4 indicate a greater degree of negative autocorrelation, while values closer to the middle suggest less autocorrelation.

**The ‘mctest’ package in R provides the Farrar-Glauber test and other relevant tests for multicollinearity.**

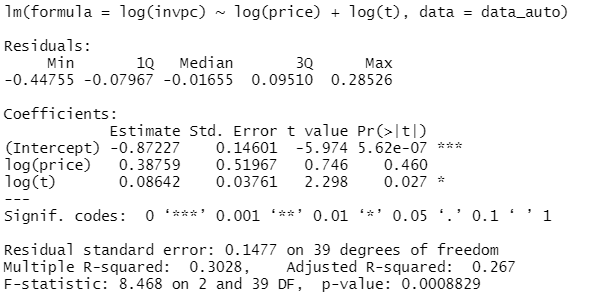
**There are two functions viz. ‘omcdiag’ and ‘imcdiag’ under ‘mctest’ package in R which will provide the overall and individual diagnostic checking for multicollinearity respectively.**

**Data:** wooldridge::hseinv

data\_auto <- wooldridge::hseinv

model\_l <- lm(log(invpc) ~ log(price)+log(t),data = data\_auto)

summary(model\_l)



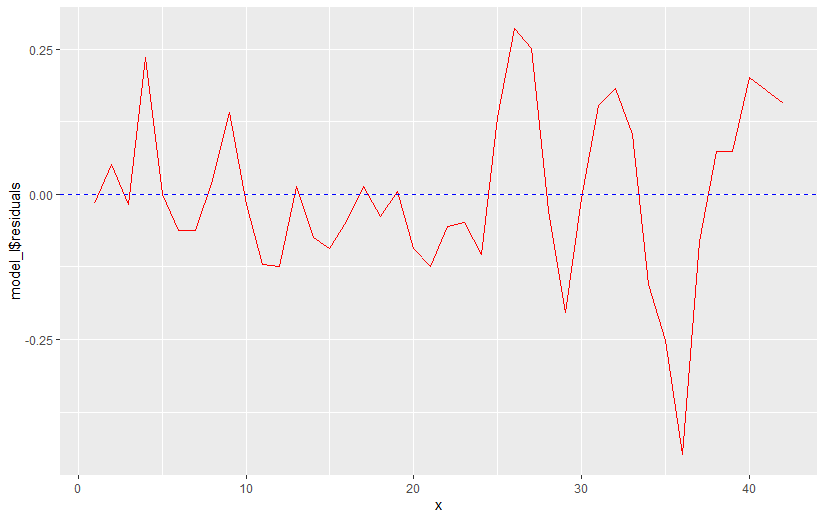
**par(mfrow = c(2,2))**

**plot(model\_l)**

**x <- seq(1,length(model\_l$residuals))**

**p <- ggplot()+geom\_line(data = hseinv,aes(x=x,y=model\_l$residuals),color = 'red')+ geom\_hline(yintercept = 0,linetype = "dashed",color = "blue")**

**print(p)**

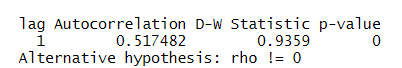


**Bptest:Performs the Breusch-Pagan test against heteroskedasticity.**

**#If there is no autocorrelation then the graph #will not have any pattern.**

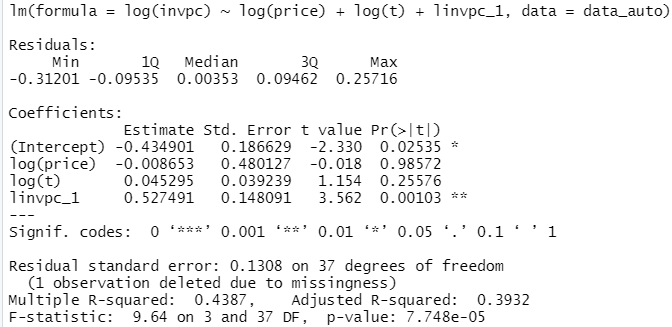
**#If the data were not autocorrelated the values #of the residuals would randomly jump as we #move from the left to right.**

**durbinWatsonTest(model\_l)**



**model\_2 <- lm(log(invpc) ~ log(price)+log(t)+linvpc\_1,data = data\_auto)**

**summary(model\_2)**

****

**Inferenc:**

**Step 1 : F- test** : For the model : lm(log(invpc) ~ log(price)+log(t),data = data\_auto)

, **F\_stat: 8.468 with p\_value :0.0008829**. This implies that there is enough evidence to indicate the overall significance of the model i.e. we will be rejecting the null hypotheses (H\_0 = β1=β2=..=0). This doesn’t mean all the independent variables are useful.

**Step 2: t-test:** For the same model , t\_test indicates that only 1 independent variable is significant.

**R\_square : 0.3028**. This means that approx. 30 % of the variation in log(invpc) is explained by the independent variables.

From the graph, we can conclude that there is no autocorrelation since there is no pattern in it.

Further this can be verified from the Durbin Watson test, since we cannot reject the NULL hypothesis i.e. there is sufficient evidence that there is homoscedasticity.