**Problem 4.**

#Given Dataset saved in csv format

#dataset imported

>dataset4 <- read.csv(file.choose(),header=T)

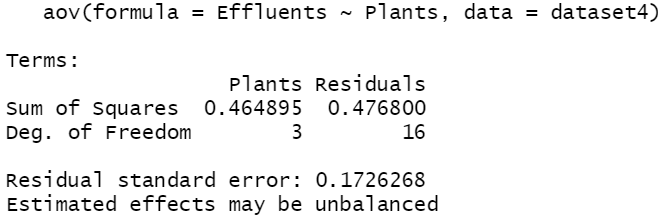
#Dataset attached

>attach(dataset4)

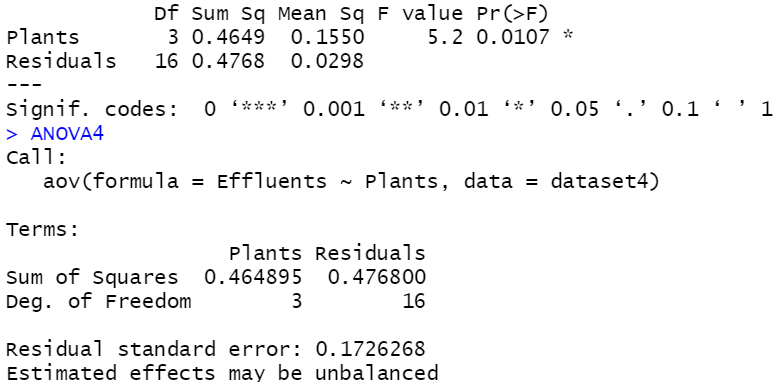
#Conducting One-Way ANOVA

>ANOVA4 <- aov(Effluents~Plants, data=dataset4)

>ANOVA4



>summary(ANOVA4)



#Sum of squares between groups is 0.464895.

**From the ANOVA summary results, it can be seen that the p-value of F-test is 0.0107<0.05. So, at 0.05 significance level, Null Hypothesis gets rejected, and it can be safely concluded that atleast one mean is significantly different from 0.**

**Thus, all the means are not simultaneously 0, as per this test.**

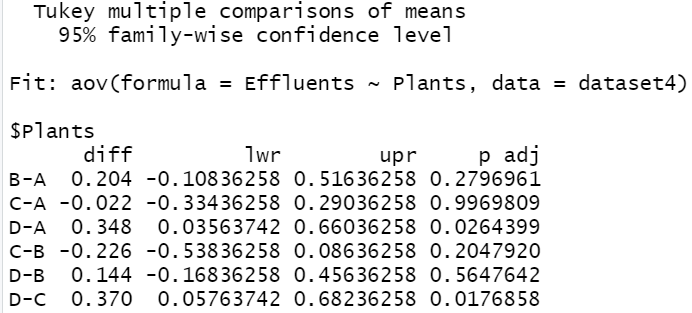
**a.)**

#agricolae is the package required for Turkey test

>library(agricolae)

>tukey.test <- TukeyHSD(ANOVA4)

>tukey.test



**Conclusion (based on obtained p -value) about the truth or falsity of null hypothesis of equality of population means for populations of polluting effluents in five different locations:**

Looking at the p-values, it can be seen that the p-values for D-C and D-A, are lesser than 0.05. Thus, null hypothesis (means of D and C are same) gets rejected, and it can be safely concluded that the mean differences D-C and D-A, are significantly different from 0.

This implies that there is some (non-zero) difference between the means of the pair, exists.

**Results of Tukey’s comparisons (R output)**:

Here we are running multiple comparisons to see which means may differ from the other.

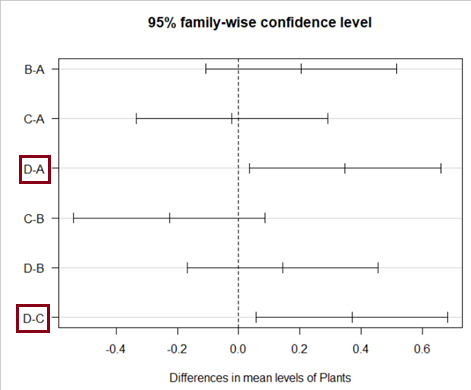
Since, 0 lies doesn’t lie in the Confidence interval of D-A & D-C, this implies that the difference of means for these pairs, are significantly different from 0 (at 0.05 significance level).

For other pairs, the difference of mean is not significantly different from 0.

#Plot of the results of Tukey test

#Visual display which helps us decide which mean differ from each other.

>plot(tukey.test,las=1)



At 0.05 significance level, the mean differences of D-C and DA, appear to be deviating from the zero (0) line.

**Problem 7.**

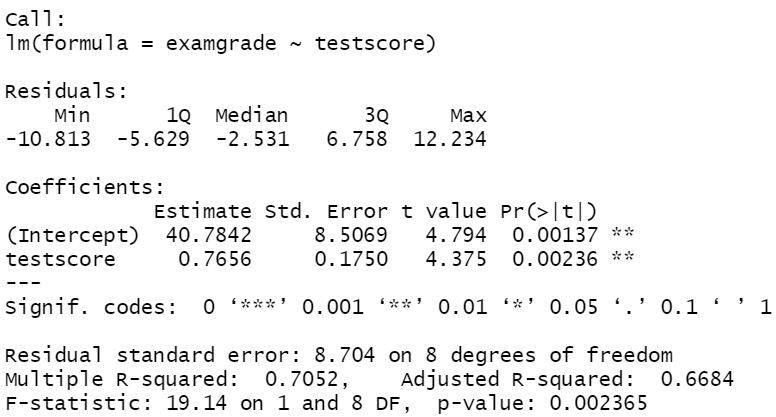
#a

> testscore <- c(39,43,21,64,57,47,28,75,34,52)

> examgrade <- c(65,78,52,82,92,89,73,98,56,75)

> gradeversusscore <- lm(examgrade~testscore)

>summary(gradeversusscore)



# Equation of estimated regression line

It can be seen from the summary results that both 'Intercept' and 'testscore' are statistically

significant at 0.01 level of significance.

Also, the multiple R-squared value of 0.7052, which indicates that 70.52% of the total

variability in the response variable (examgrade) is explained by the predictor 'testscore'.

However, the model can still be made better by adding more relevant predictors or introducing interaction

variables.

Adjusted R-squared(0.6684) can also be improved.

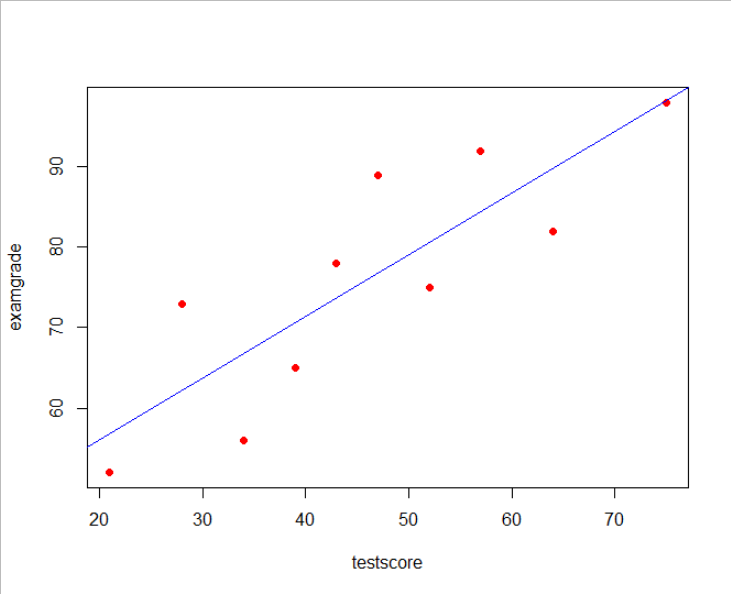
**b.)**

#Scatterplot

>plot(testscore, examgrade,pch=19 ,col='red')

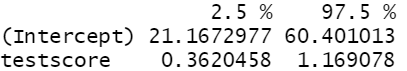
#Line of best fit

>abline(lm(examgrade ~ testscore), col='blue') # plot the regression line



#To find Confidence interval of Slope for Calculus grade data

>confint(gradeversusscore)



95% Confidence Interval for Slope(testscore):

**(0.3620458, 1.169078)**

**Problem 8.**

#Given Dataset saved in csv format

#dataset imported

>dataset <- read.csv(file.choose(),header=T)

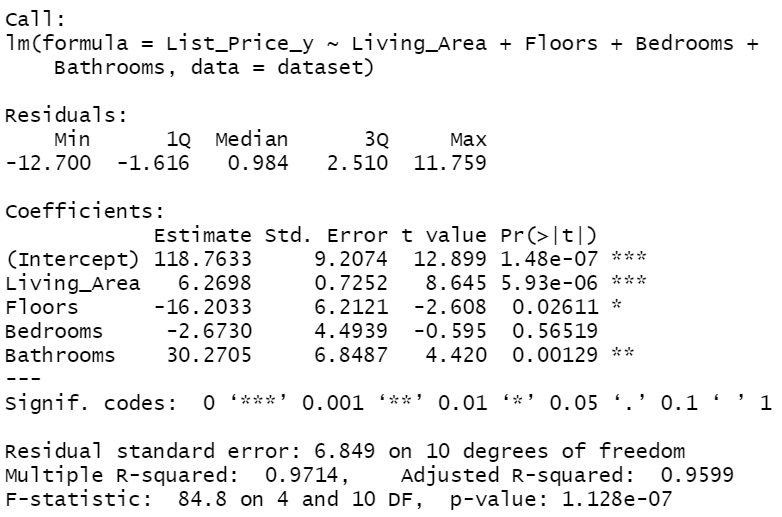
#Dataset attached

>attach(dataset)

#Multiple Linear regression model created

>model8 <- lm(List\_Price\_y~Living\_Area+Floors+Bedrooms+Bathrooms, data=dataset)

>summary(model8)

****

**a.)**

Yes,

Since, Multiple R-squared value is 0.9714. This implies that 97.14% of the variability in the model is explained by the independent variables.

Also, the Adjusted R-Squared is 95.99%, which signifies principal of parsimony of the model.

Model is not only accurate, but also uses minimum number of variables to give high accuracy.

From the model summary results, it can be seen most of the coefficients of the variables, except Bedrooms are statistically significant at 5% significance level.

Most of the variables like Intercept (p-value:1.48e-07), Living\_Area(p-value: 5.93e-06 ), Floors(p-value: 0.02611) and Bathrooms (p-value: 0.00129), are statistically significant.

**b.)**

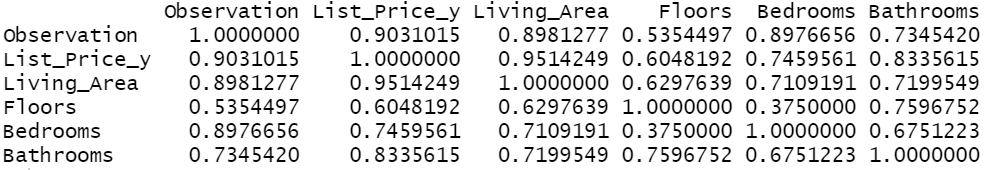
The model summary indicates that the p-value of List\_Price and predictor Living\_Area is 5.93e-06, which is less than 0.001. So, at 0.001% level of significance this predictor is significant in impacting the response variable.

Also, at 5% significance level, this predictor is significantly different from 0.

Since, the model is a multiple-linear regression one, this variable is indeed linearly correlated with the response variable ‘List\_price’.

#Correlation Matrix of the dataset

> cor(dataset)



Correlation matrix also says that the correlation coefficient (Pearson) between List\_Price\_y and Living\_Area is 0.9514249, which is very close to 1.

This indicates that these two variables have very high linear correlation between them.

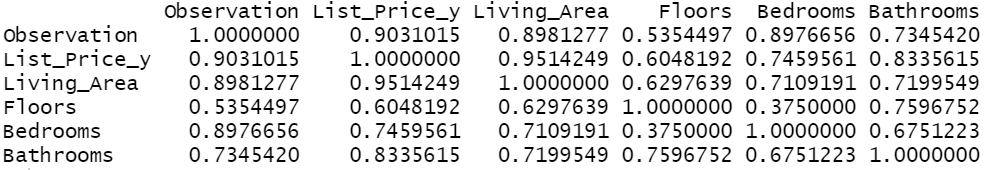
**c.)**

The model summary indicates that the p-value of List\_Price and predictor Bedrooms (Number of Bedrooms) is 0.56519, which is more than 0.05. So, at 0.05% level of significance this predictor is not significant in impacting the response variable.

Also, at 5% significance level, this predictor is not significantly different from 0.

#Correlation Matrix of the dataset

> cor(dataset)



Correlation matrix says that the correlation coefficient (Pearson) between List\_Price\_y and Bedrooms is 0.7459561, which is close to 1.

This indicates that these two variables have very good linear correlation between them.

However, from the summary results this variable ‘Bedrooms’ doesn’t seem very significant in impacting the response variable. One suggestion could be introducing interaction variables using variable ‘Bedrooms’.

**Problem 9.**

#Given Dataset saved in csv format

#dataset imported

>dataset9 <- read.csv(file.choose(),header=T)

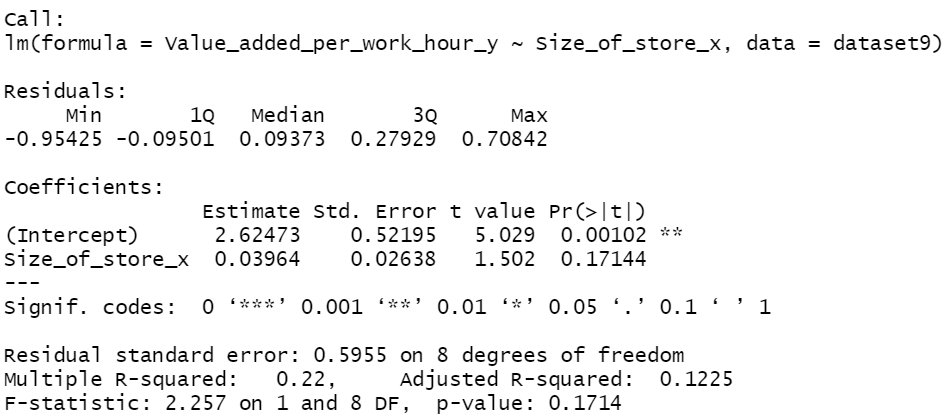
#attaching dataset

>attach(dataset9)

**a.)**

>model9\_linear <- lm(Value\_added\_per\_work\_hour\_y~Size\_of\_store\_x,data=dataset9)

>summary(model9\_linear)



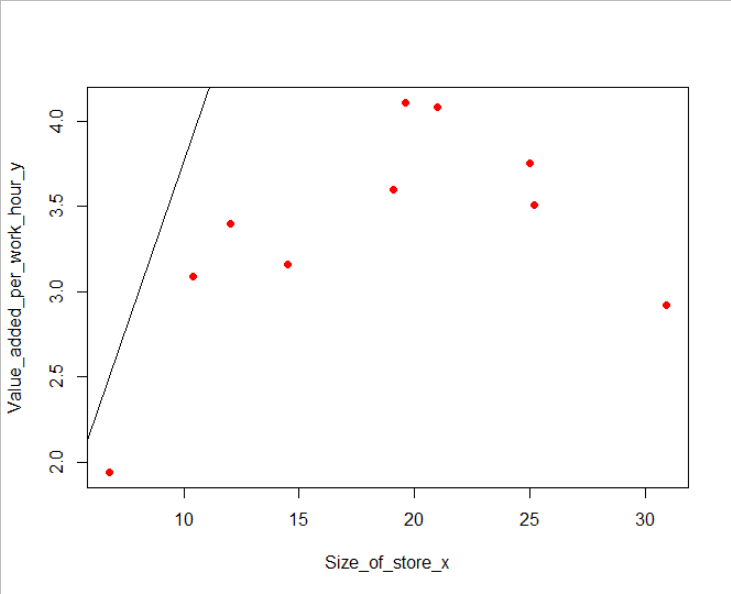
Linear Regression model suggests that the predictor ‘Size\_of\_store’ with a p-vaue of 0.17144, which is greater than even 0.1, is not statistically different from 0.

This predictor has negligible impact on the response variable Value\_added\_per\_work\_hour’.

#Scatterplot created

>plot(Size\_of\_store\_x,Value\_added\_per\_work\_hour\_y,pch=19,col='red')

#the plot doesn't seem to be linear

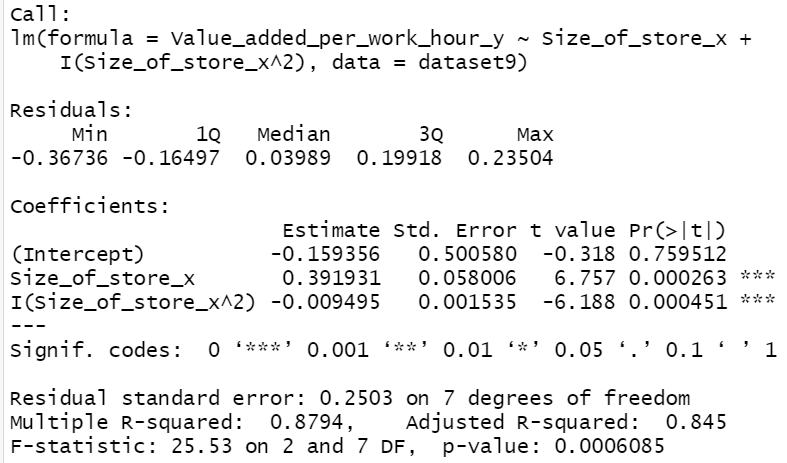


**b.)**

#Creating a Polynomial regression model of degree 2, or adding a Quadratic term into the model

>model9 <- lm(Value\_added\_per\_work\_hour\_y~Size\_of\_store\_x + I(Size\_of\_store\_x^2), data=dataset9)

>summary(model9)



Insertion of a quadratic term, makes the model a better fit.

The predictors Size\_of\_store\_x (p-value: 0.000263 less than 0.001) & Size\_of\_store\_x^2 (p-value: 0.000451 less than 0.001)

Both the p-values are less than 0.001.

Thus, the null-hypothesis that the coefficients of predictors are each 0, is rejected.

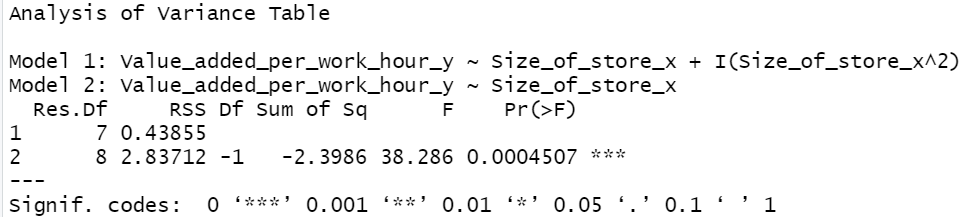
Hence, it can be safely concluded that the predictors are significantly away from 0, and make an impact on the response variable (Value\_added\_per\_work\_hour\_y).

Multiple R-squared (0.8794) indicates that 87.94% of the variability in the model is explained by the predictors.

Adjusted R-squared (0.845) is also close to 1.

This suggests that the Polynomial Regression model is a good fit.

>anova(model9,model9\_linear)

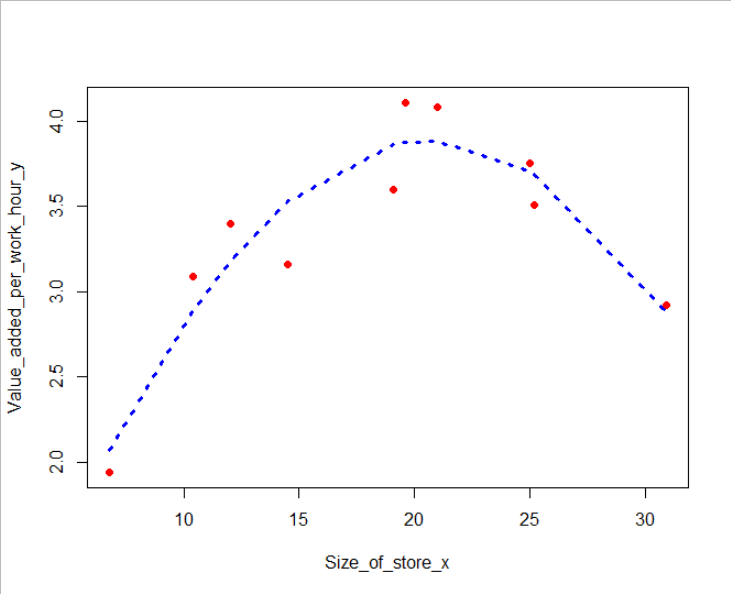


F-Ratio test suggests that model9(Polynomial Regression model) is significantly different from the basic linear model.

**c.)**

#Thick dashed blue line added to represent the Quadratic regression curve

>lines(smooth.spline(Size\_of\_store\_x ,predict(model9)),col='blue',lwd=3, lty=3)



From the plot, it is evident that the Quadratic Regression model is indeed a very good fit.