**QMB-6304 Analytical Methods for Business**

**Module 4 Assignment**

Write a simple R script to execute the following data preprocessing and statistical analysis. Where required show analytical output and interpretations.

**Preprocessing**

1. Load into an object the data from the "Craigs List Cars" worksheet in the "6304 Module 4 Assignment Data.xlsx" spreadsheet file. This file contains information on 46,484 vehicles in the United States listed for sale on Craig’s List. This is the master data set.
2. Using the numerical portion of the U number as a random number seed, select a random sample of n=200 autos from the master data set. Hold these observations in a separate data frame.

**Analysis**

1. Conduct a simple regression analysis using your sample data with the dependent variable being "price" and the independent variable being "odometer".
2. Give verbal interpretations of all beta coefficients in your regression model. Make certain the language you use is understandable to a reasonably competent lay person shopping for a car on Craig's List.
3. Evaluate and interpret both the p value and the confidence interval on the "odometer" coefficient in your regression model.
4. Run appropriate diagnostics on your regression model to determine if it is in conformity with the LINE assumptions of regression.
5. Ms. Trayla Parks is considering offering her Toyota for sale on Craig's List. The Toyota currently has an odometer reading of 78,521 miles. Use your regression model to predict the price of the vehicle on Craig's List. Determine and verbally interpret the appropriate confidence interval on this prediction. If Trayla kept her vehicle one more year until the odometer showed 98,000 by how much would your model predict the price of her car would change?

Your deliverable will be a single MS-Word file showing 1) the R script which executes the above instructions and 2) the results of those instructions. The first line of your script file should be a “#” comment line showing your name as it appears in Canvas. Results should be presented in the order in which they are listed here. Deliverable due time will be announced in class and on Canvas. **This is an individual assignment to be completed before you leave the classroom. No collaboration of any sort is allowed on this assignment.**

**Preprocessing:**

#Varun Teja Kolluru

#Assignment 4

#PREPROCESSING

#remove all the variables in Environment window

rm(list=ls())

#Import the required libraries

library(rio)

#Load the data into R

my\_data=import('6304 Module 4 Assignment Data.xlsx')

colnames(my\_data)=tolower(make.names(colnames(my\_data)))

#create sample data from the loaded data

set.seed(97)

my\_sample=my\_data[sample(1:nrow(my\_data),200,

replace=FALSE),]

attach(my\_sample)

Using rm statement, we can clear all the variables in the environment window. All the useful libraries are imported and using the import statement the data is loaded into R. As per the given steps data is split and a seed is set with my last 2 digits of U number and a random sample of 200 observations are taken for the analysis part. Proceeding to the analysis part.

**Analysis:**

1. Conduct a simple regression analysis using your sample data with the dependent variable being "price" and the independent variable being "odometer".

Rcode:

#1

plot(odometer,price,pch=19,

main="Odometer & Price Raw Data Plot")

output=lm(price~odometer,data=my\_sample)

Chart, scatter chart

Description automatically generated

From the sample data, I have plotted the basic scatter plot for the initial analysis and from the graph we can observe that there is a lot of distribution, and this might be a linear relationship but with a lot of outliers and we can assume that the slope might be negative for the given dataset.

Here I have conducted a simple linear model using ‘lm’ statement in R studios, with price as DV and odometer as IV with tilda operator by giving the sample data.

1. Give verbal interpretations of all beta coefficients in your regression model. Make certain the language you use is understandable to a reasonably competent lay person shopping for a car on Craig's List.

Rcode:

output=lm(price~odometer,data=my\_sample)

#2

summary(output)

Result in console window:

> #1

> output=lm(price~odometer,data=my\_sample)

> summary(output)

Call:

lm(formula = price ~ odometer, data = my\_sample)

Residuals:

Min 1Q Median 3Q Max

-16211 -5501 -2009 3717 31936

Coefficients:

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

(Intercept) 1.851e+04 1.163e+03 15.923 < 2e-16 \*\*\*

odometer -5.723e-02 8.884e-03 -6.443 8.74e-10 \*\*\*

---

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

Residual standard error: 8042 on 198 degrees of freedom

Multiple R-squared: 0.1733, Adjusted R-squared: 0.1691

F-statistic: 41.51 on 1 and 198 DF, p-value: 8.736e-10

Here, the output of the linear regression can be viewed using ‘summary’ statement in R studios.

Explaining call part:

So from the start, call is the statement tells that we have used linear regression model with the dependent and independent variables.

Explaining residuals part:

Here, we have the min residual value as -16211 and the maximum residual value is 31936 and the median is -2009. 1Q is the 25% quartile and it is around -5501 and 3Q indicates 75% quartile and it is around 3717.

Explaining Coefficients part:

The Beta coefficients are B0(Intercept) is 1.851e+04 and B1(Slope) is -5.723e-02. We can derive the equation as below with these coefficients.

Estimated Y = (1.851e+04) – (5.723e-02)X.

And the p values(significance) for both the coefficients is less than 5 percent and we reject the null hypothesis and accept the alternate saying that the beta coefficients are not ‘zero’.

Residual standard error is the standard deviation, and the SD is around 8042 with 198 degree of freedom.

Multiple R-Squared value tells us how well the linear regression model is fit. Any number close to 0 means not identifiable linear relationship. If the value is equals to 1 then it is a best fit and if the value is 0 then it is a worst fit. From the above data we got R-Squared value is 0.1733, which is not so good.

Adjusted R-Squares is also called as coefficient of determination and the value we got is 0.1691, which is also not good.

Estimated Y = (1.851e+04) – (5.723e-02)X.

From the linear regression model equation, we can say that if the odometer reading is increased by 1 mile, then there is a significant decrease in price with 5.723e-02 value.

1. Evaluate and interpret both the p value and the confidence interval on the "odometer" coefficient in your regression model.

Rcode:

confint(output)

Result in console window:

> confint(output)

2.5 % 97.5 %

(Intercept) 16221.8912986 20808.0492893

odometer -0.0747512 -0.0397142

As explained above from the summary of the linear regression model, the p values for both the coefficients is less than 5 percent and we reject the null hypothesis and accept the alternate saying that the beta coefficients are not ‘zero’.

The 95% confidence intervals for the intercept are, lower interval is 16221.89 and higher interval is 20808.04.

The 95% confidence intervals for the slope are, lower interval is -0.074 and higher interval is -0.039.

We are 95% confident that our intercept and slope values are in between these lower and higher intervals.

1. Run appropriate diagnostics on your regression model to determine if it is in conformity with the LINE assumptions of regression.

Rcode:

#4

# Linearity

plot(my\_sample$price,output$fitted.values,

pch=19,main="Odometer&Price Actual v. Fitted Values")

abline(0,1,col="red",lwd=3)

plot(my\_sample$price,output$fitted.values,

pch=19,

xlim=c(0,60000),ylim=c(0,30000),

main="Odometer&Price Actual v. Fitted Values")

abline(0,1,col="red",lwd=3)

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

We check the sample price with the fitted values to check the linearity and from the above graph, we can say that the data doesn’t follow linearity, because below the red line there are a lot of errors or residuals far from the line, which doesn’t look good. To satisfy the linearity the sum of errors should be equals to zero or the summed squared errors value should be minimum.

Rcode:

# Normality

qqnorm(output$residuals,pch=19,main="O&P Normality Plot")

qqline(output$residuals,col="red",lwd=3)

Chart, line chart

Description automatically generated

QQ plot is one of way to check the Normality. We give the residuals as the input for the qq plot and check the data is following normality. From the graph we can say that, the data is follows normality( its good ), because the residuals follow the standard line( red line ).

Rcode:

# Equality of Variances

plot(output$fitted.values,scale(output$residuals),

pch=19,main="O&P Standardized Residuals",

xlab=c("Fitted Values"),ylab=c("Residuals"))

abline(0,0,col="red",lwd=3)

Chart, scatter chart

Description automatically generated

We check the Equality of variance plotting fitted values and residuals. We will be looking for patterns in those dots relative to the horizontal line. If we find a pattern, then the residuals are not

equally distributed.

From the above graph, we don’t see any patterns but there are some outlines above far from the horizontal line which are fine, and we can conclude that the data follows Equality of variance.

1. Ms. Trayla Parks is considering offering her Toyota for sale on Craig's List. The Toyota currently has an odometer reading of 78,521 miles. Use your regression model to predict the price of the vehicle on Craig's List. Determine and verbally interpret the appropriate confidence interval on this prediction. If Trayla kept her vehicle one more year until the odometer showed 98,000 by how much would your model predict the price of her car would change?

Rcode:

#5

# predicting what the price of car would be if odometer reading was 78,521 miles

newdata=data.frame(odometer=78521)

# predict is a point in time

predict(output,newdata,interval="predict")

# confidence is a long term average

predict(output,newdata,interval="confidence")

Result in console window:

> #5

> # predicting what the price of car would be if odometer reading was 78,521 miles

> newdata=data.frame(odometer=78521)

> # predict is a point in time

> predict(output,newdata,interval="predict")

fit lwr upr

1 14021 -1889.866 29931.87

> # confidence is a long term average

> predict(output,newdata,interval="confidence")

fit lwr upr

1 14021 12737.39 15304.61

Note: If we use the ‘predict’ in interval, that mean we are checking the price at that specific point of time. And if we use ‘confidence’ in interval, the confidence interval we get is for long term average time. And we get a tighter confidence interval for ‘confidence’ and a narrow confidence interval for ‘predict’ statement when compared to ‘confidence’ statement.

As per the linear regression model equation, if the odometer reading are 78,521 the price would be 14021 dollars. In interval predict, the lower limit is -1889.866 and the upper limit is 29931.87. In interval confidence, the lower limit is 12737.39 and the upper limit is 29931.87

Rcode:

# predicting what the price of car would be if odometer reading was 98,000 miles

newdata=data.frame(odometer=98000)

# predict is a point in time

predict(output,newdata,interval="predict")

# confidence is a long term average

predict(output,newdata,interval="confidence")

Result in console window:

> # predicting what the price of car would be if odometer reading was 78,521 miles

> newdata=data.frame(odometer=98000)

> # predict is a point in time

> predict(output,newdata,interval="predict")

fit lwr upr

1 12906.17 -2994.963 28807.29

> # confidence is a long term average

> predict(output,newdata,interval="confidence")

fit lwr upr

1 12906.17 11749.52 14062.81

As per the linear regression model equation, if the odometer reading are 98,000 the price would be 12906.17 dollars. In interval predict, the lower limit is -2994.963 and the upper limit is 28807.29. In interval confidence, the lower limit is 11749.52 and the upper limit is 14062.81