AMOD 5240H – Data Analysis Project

Rachael Joan Dias, Student No.: 0651897

24/11/2018

1. Description of the data set and provenance

I have selected the **"diamond.dat.txt"**[1] dataset from the Journal of Statistics Education Data Archive.

Data Source: (http://jse.amstat.org/jse_data_archive.htm)

Data Set Description [2]:

- The data set consists of 48 observations in total and 2 variables.
- The 2 variables are the price of the diamond ring in Singapore dollars and the carat size measured in grams (1 carat=.2 grams) of the diamond stone.
- Each ring is made with gold of 20 carats purity and mounted with a single diamond stone.
- The observations only consists of 20K rings. The weight of the diamond stones varies from 0.12 to 0.35 gms while their prices vary between 223 and 1086 Singapore dollars.

Data Provenance [3]:

- The origin of the data is from a full page advertisment in the The Straits Times newspaper issue of February 29, 1992, by a Singapore-based retailer of diamond jewelry.
- The advertisment consisted of pictures of diamond rings and their corresponsing prices, diamond content and gold purity.
- For the purpose of this observational study only 20K rings i.e rings made with gold of 20 carat purity were considered.

Summary of Data: Column 'Carat_Size' represents the Carats (weight in gm) and column 'Price' represents the Price in Singapore dollars.

```
library(kableExtra)
diamond<-read.csv("diamond.csv",header=FALSE)
names(diamond)<-c("Carat_Size","Price")
kable(diamond[1:7,])
```

Carat_Size	Price
0.17	355
0.16	328
0.17	350
0.18	325
0.25	642
0.16	342
0.15	322

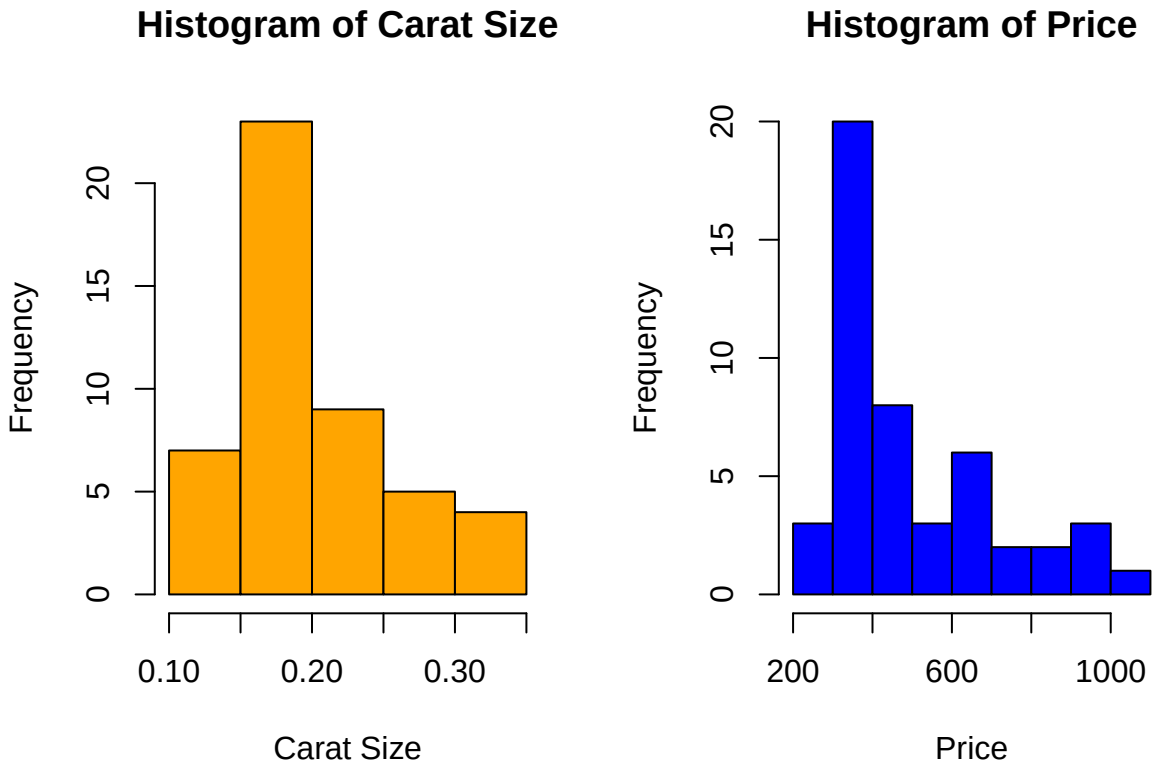
2. Appropriateness for statistical analysis

Conditions for statistical analysis:

Random Samples: The data is from a newspaper advertisment for a single Singapore-based retailer of diamond jewelry. There is no mention of the sampling method used to collect the data therefore, we cannot say that the observations are random samples from a population.

Normality: Both the numerical variables Price and Carat Size do not follow a normal distribution.

```
par(mfrow =c(1,2))
hist(diamond$Carat_Size,xlab="Carat Size", main="Histogram of Carat Size", col="orange")
hist(diamond$Price, xlab="Price", main="Histogram of Price", col="blue")
```



Independance: The sample size consists of only 48 observations which is less than 10% of the population. If we sample without replacement, the sample size is not more than 10% of the population. The individual observations can be considered as independant since removing a single observation doesn't significantly affect the population.

3. Two Questions

Question 1: Is there a relationship between the price of diamond rings and carat size?

Question 2: How does the price of a diamond ring vary with the carat size?

Using a linear regression model we can identify the relationship between the two numerical variables carat size and price of diamond rings. I will use the linear regression technique that we learnt this semester to find an answer to Question 1 and determine the type of relationship between price of diamond rings and carat size.

4. Data summary using R

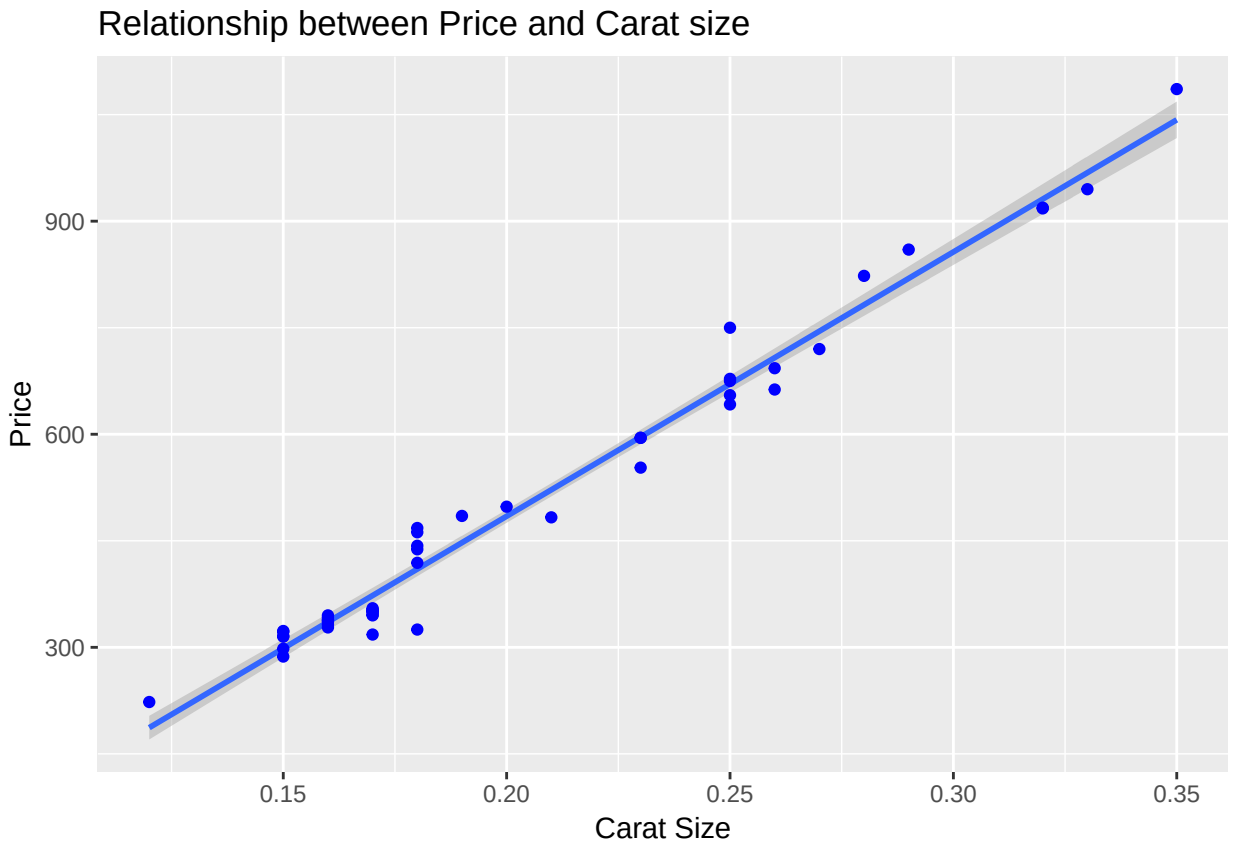
Scatterplot: Since our data set only consists of price of the rings and carat size we need not subset it. We start by visualizing our data to see if there exist a relationship between the 2 variables. A scatterplot is the primary tool we use to examine a linear relationship between 2 numerical variables.

From the scatter plot there appears to be a clear linear relationship between price and carat size. Examing some of the feature of the scatterplot we can conclude that

- There is an uphill trend or positive correlation between price and carat size.

- The strength of the relationship is strongs since there is very little amount of vertical variation.
- The shape is linear which means fitting a linear model is reasonable.

```
library(ggplot2)
ggplot(diamond, aes(x=Carat_Size, y=Price))+geom_smooth(method=lm)+
geom_point(color="blue")+ylab("Price")+
xlab("Carat Size")+
ggtitle("Relationship between Price and Carat size")
```



Correlation Coefficient 'r': We also check the correlation coefficient to estimate the strength of the relationship between the price and carat size variables. Since we get a value very close to '1', there exist a strong linear relationship between the variables and the are positively correlated.

```
cor(diamond$Carat_Size,diamond$Price)
```

```
## [1] 0.9890707
```

5. Data analysis using R

Linear Model: We assume all assumptions for statistical analysis are valid. Since price and carat size have a strong linear relationship, we fit a linear regression model to the data using the lm() function.

```
linear_model<-lm(diamond$Price~diamond$Carat_Size)
summary(linear_model)
```

```
##
## Call:
## lm(formula = diamond$Price ~ diamond$Carat_Size)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -85.159 -21.448  -0.869  18.972  79.370
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -259.63      17.32  -14.99   <2e-16 ***
## diamond$Carat_Size  3721.02      81.79   45.50   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 31.84 on 46 degrees of freedom
## Multiple R-squared:  0.9783, Adjusted R-squared:  0.9778
## F-statistic:  2070 on 1 and 46 DF,  p-value: < 2.2e-16
```

Testing for the Slope:

Hypothesis:

$H_0 : \beta_1 = 0$ The Null Hypothesis states that the slope/coefficient parameter is 0 i.e there is no linear relationship between price of rings and carat size.

$H_0 : \beta_1 \neq 0$ The Alternative Hypothesis states that the slope/coefficient parameter is not 0 i.e there is a linear relationship between price of rings and carat size.

Conclusion: Since we get a P-value of **2e-16** for the slope, which is less than our significance level of $\alpha = 0.05$, we reject H_0, we can conclude that a linear relationship exist between the price of the rings and carat size.

Analysing the R Output of our Linear Regression Model

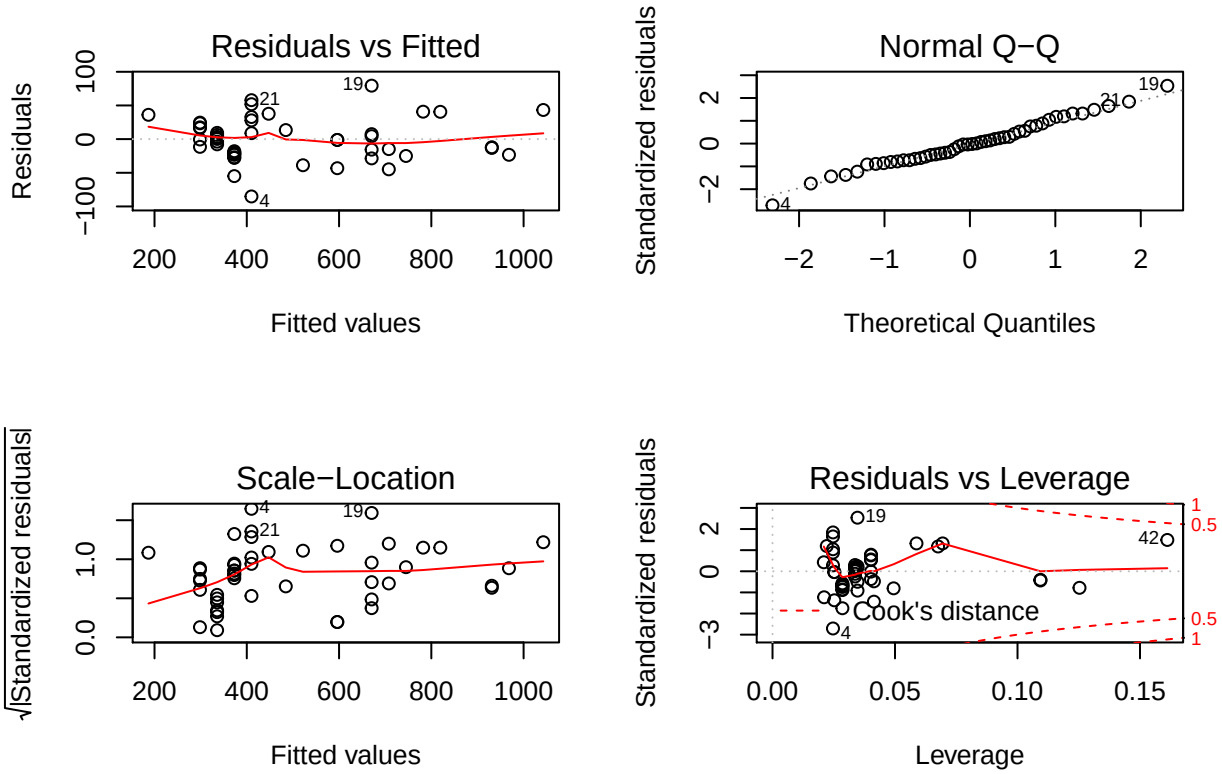
Adjusted R-squared: The model is able to explain 97.78% variablity in the data i.e 97.78% variation in the price of diamond rings can be explained by carat size.

Degrees of Freedom: 46 degrees of freedom shows that there were 48 observations.

P-value: P-value of the slope and overall P-value are the same, there is a significant linear relationship. The model is useful for predictions.

Model Plots:

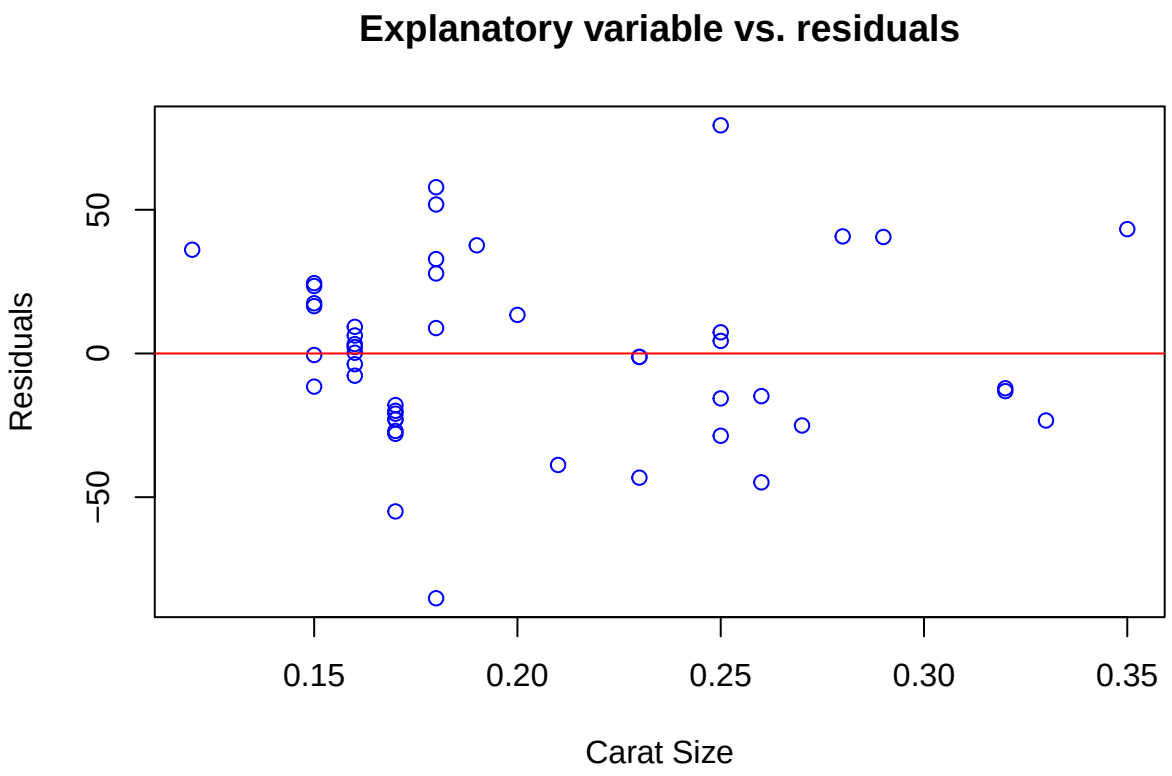
```
par(mfrow =c(2,2))
plot(linear_model)
```



Residual vs Fitted Plot: The residual vs. fitted plot does not have a wedge shape, the spread is randomly distributed, there is no pattern in the variance.

Normal Probability Plot: The residuals fall on a straight line hence they are normally distributed.

```
plot(diamond$Carat_Size, linear_model$residuals, main="Explanatory variable vs. residuals",
ylab="Residuals", xlab="Carat Size", col="blue")
abline(0,0,col="red")
```



Residual Plot vs Explanatory Variable:

- The points on the residual plot are random and have no trend, therefore the condition of linearity trent holds true
- It lacks a wedge shape, it does not become wider or narrower when viewed from left to right. This confirms the constant standard deviation of residuals.

Summary of Condition Checks:

1. **Linearity:** Since the points on the residual plot are randomly distributed and there is no trend, the condition for linearity is met.
2. **Constant Standard Deviation:** The residual plot lacks a wedge shape hence the constant standard deviation condition is satisfied.
3. **Normality:** The QQ plot of the residuals is a straight line, therefore the normality condition is satisfied.
4. **Independance:** The residuals are independant of each other, one observation does not tell us anything about another observation.

Equation of Linear Regression Line:

The equation of a regression line is given by $y = Intercept + slope * x$, where 'y' is the dependant variable and 'x' is the independant variable.

In this context it can be represented as $Price = -259.63 + 3721.02 * CaratSize$.

6. Conclusion

To conclude, from our linear regression analysis we were able to answer the first question there is definitely a strong linear relationship between the Price of diamond rings and Carat Size (measured in gms), the two variables are strongly positively correlated. We started by plotting a scatterplot of the 2 numerical variables Price and Carat Size, this helped us to see a clear linear pattern between the 2 variables. We also evaluated

the strength of the relationship by checking the correlation coefficient of the 2 variables. Our initial analysis led us to the conclusion that fitting a linear model is reasonable and so we built a linear regression model using the 2 variables.

The model gave us an overall P-value of **2e-16**, which tells us that it is useful for predicting the price of diamond rings for carat sizes between 0.12 and 0.35 gms, however we cannot extrapolate and make any predictions beyond this range. The model was also able to explain 97.78% variablity in the Price of diamond rings based on Carat size.

7. References

[1] Journal of Statistics Education (JSE) Home Page. [Online]. Available: http://jse.amstat.org/jse_data_archive.htm. [Accessed: 10-Dec-2018].

[2] "Diamond Ring Pricing Using Linear Regression," Journal of Statistics Education (JSE) Home Page. [Online]. Available: http://jse.amstat.org/v4n3/datasets.chu.html. [Accessed: 10-Dec-2018].

[3] Journal of Statistics Education (JSE) Home Page. [Online]. Available: http://jse.amstat.org/datasets/diamond.txt. [Accessed: 10-Dec-2018].