QANT 530 Homework #1

Phan Nguyen

07/12/2022

For this homework, we are going to do a simple regression analysis on UtilityCosts and MaintenanceCosts from the practice dataset. UtilityCosts will be explanatory variable, and MaintenanceCosts will be the response variable.

```
##   UtilityCosts MaintenceCosts
## 1     29.89512      17.818890
## 2     35.77786      14.547290
## 3     36.56817      18.462665
## 4     40.46540       8.226194
## 5     40.63734       7.087728
## 6     42.57490      18.452979
```

Table 1: Dataset example

1.Exploratory Analysis on the response variable

Descriptive numerical summary table

```
##   vars  n  mean   sd median trimmed  mad   min  max range  skew kurtosis  se
## 1    1 60 10.43 5.39  10.83   10.61 5.22 -1.71 20.8 22.51 -0.25     -0.6 0.7
##   Q0.25 Q0.75
## 1  7.58 14.26
```

Table 2: Descriptive Statistics table for MaintenanceCosts

MaintenanceCosts has a minimum of -1.711, a negative number. This is quite strange, but since we don't know the context of this dataset, we won't know if this number makes sense. Mean and median for MaintenanceCosts are approximately equal, thus meaning the distribution for this variable is symmetrical.
The interquartile range for this variable is between 7.58 and 14.25, meaning 50% of values for this variable are in this range.
The standard deviation for MaintenanceCosts is 5.39. Using Empirical Rule, we can expect that 95% of values are between -0.35 and 21.21 (2 stanadard deviation).

Skewness MaintenanceCosts are very close to 0, suggesting that the distribution for the variable is not quite skewed to either left or right. The measures of center also suggested the same. The excess kurtosis is around -0.6. Excess kurtosis is equal to normal kurtosis - 3. In this case, the kurtosis shows that the distribution is very close to a normal distribution. MaintenanceCosts's distribution has heavier tails and lower bell, due to its kurtosis to be lower than 0.

Descriptive chart summaries

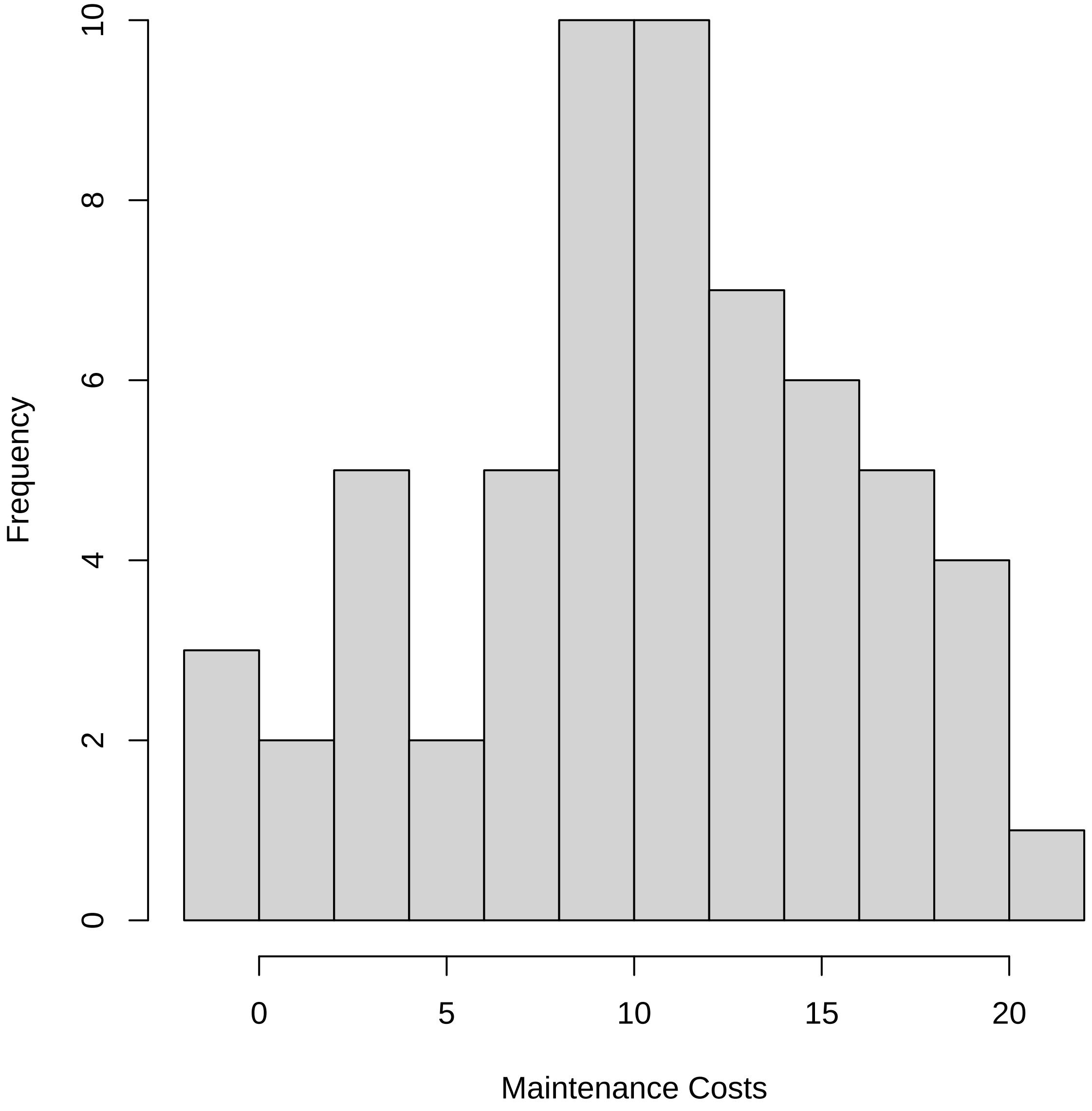

Figure 1: Histogram of MaintenanceCosts

The histogram shows the frequency of data values of MaintenanceCosts variable. We can cross check it with the numerical summary we've done on the top. The mean is approximately 10. The distribution is close to a normal distribution, however, the bins on the left side of the mean have less frequency that those on the right of the mean.
The variable is not skewed hardly to one side or another.

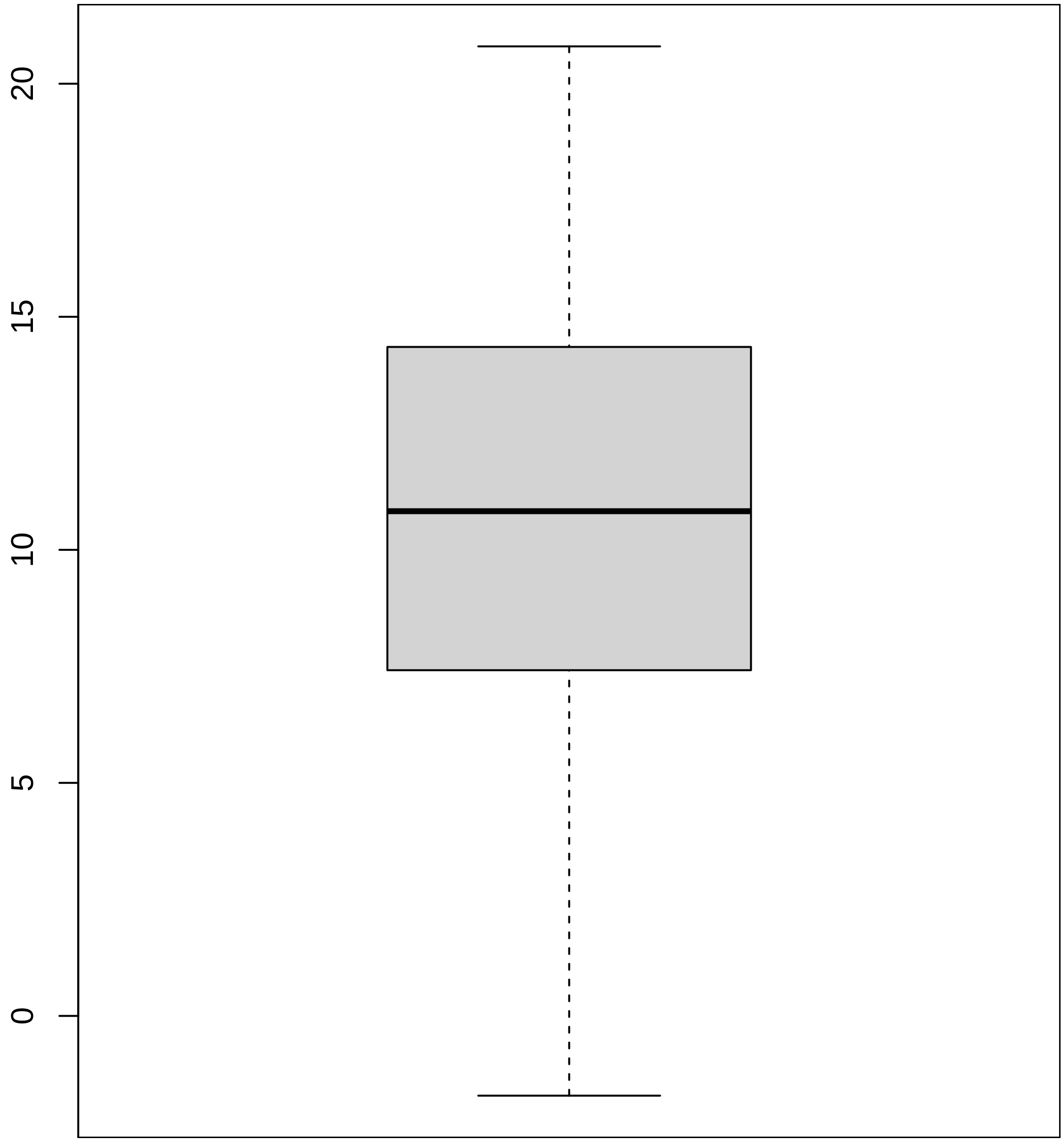

Figure 2: Boxplot of MaintenanceCosts

Visually, the boxplot of this variable looks normal. There are whiskers indicating the maximum and minimum values, corresponding to what we saw in the numerical descriptive statistcs. The gray box represents the interquartile range, where the lower bound is the first quartile and the upper bound is the third quartile. The black line in the middle indicates the median. We can see that the median is in the middle of this gray box, meaning it has approximately the same range to the first and third quartiles. From this boxplot, we can see that the variable's distribution is skewed to the left by a small amount.

2. Scatterplot analysis

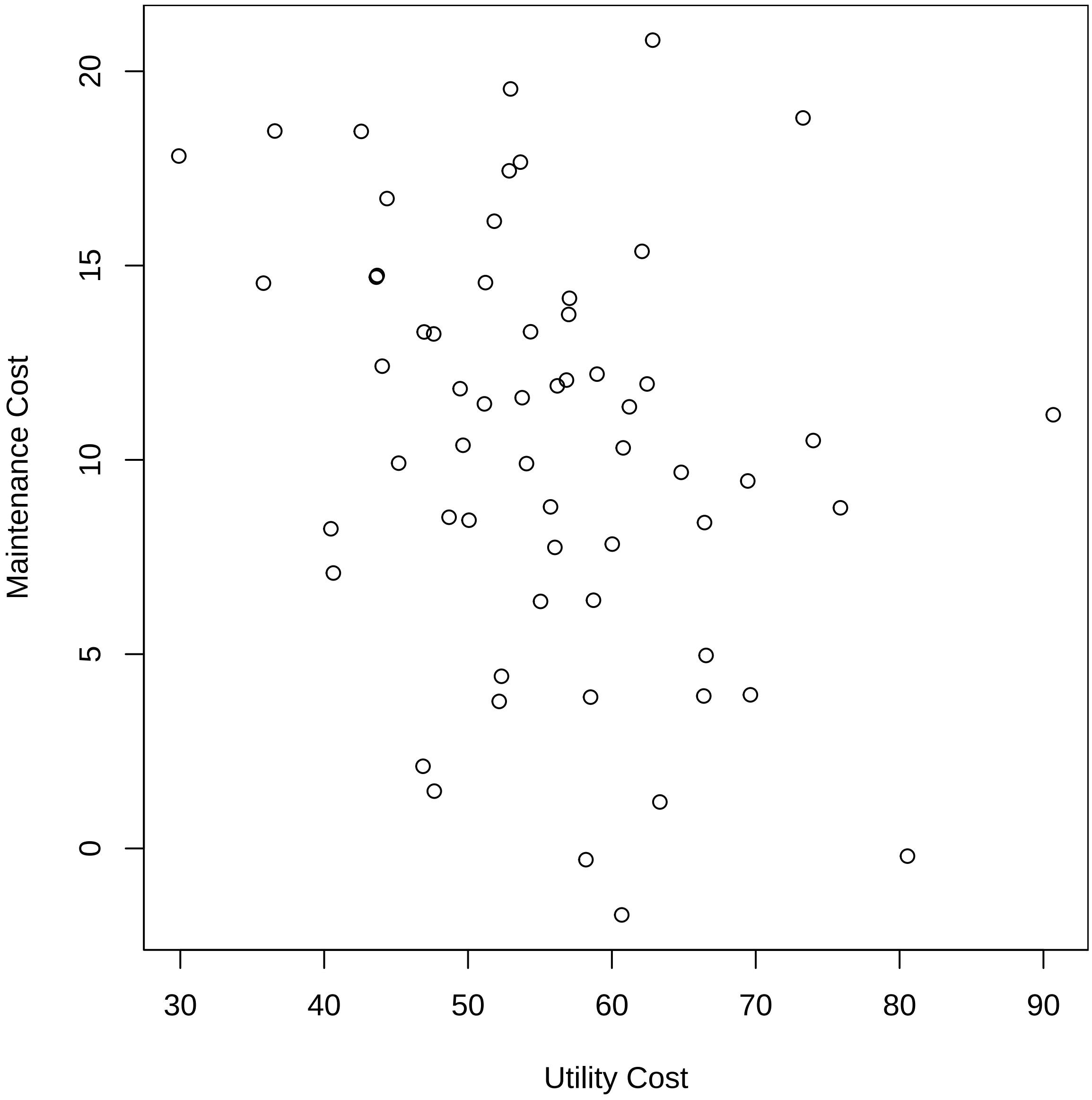

Figure 3: Scatterplot for UtilityCosts and MaintenanceCosts

There seems to be low correlation between the two variables, when we first look at the scatterplot. The data points seems to be everywhere. However, if we analyze closely, there are some types of downward trend for the points. We can expect that the Pearson correlation is small, and negative.

And the correlation between UtilityCosts and MaintenanceCosts is

```
## [1] -0.2933223
```

, which is in line with our visual analysis above.

3. Residual plot analysis

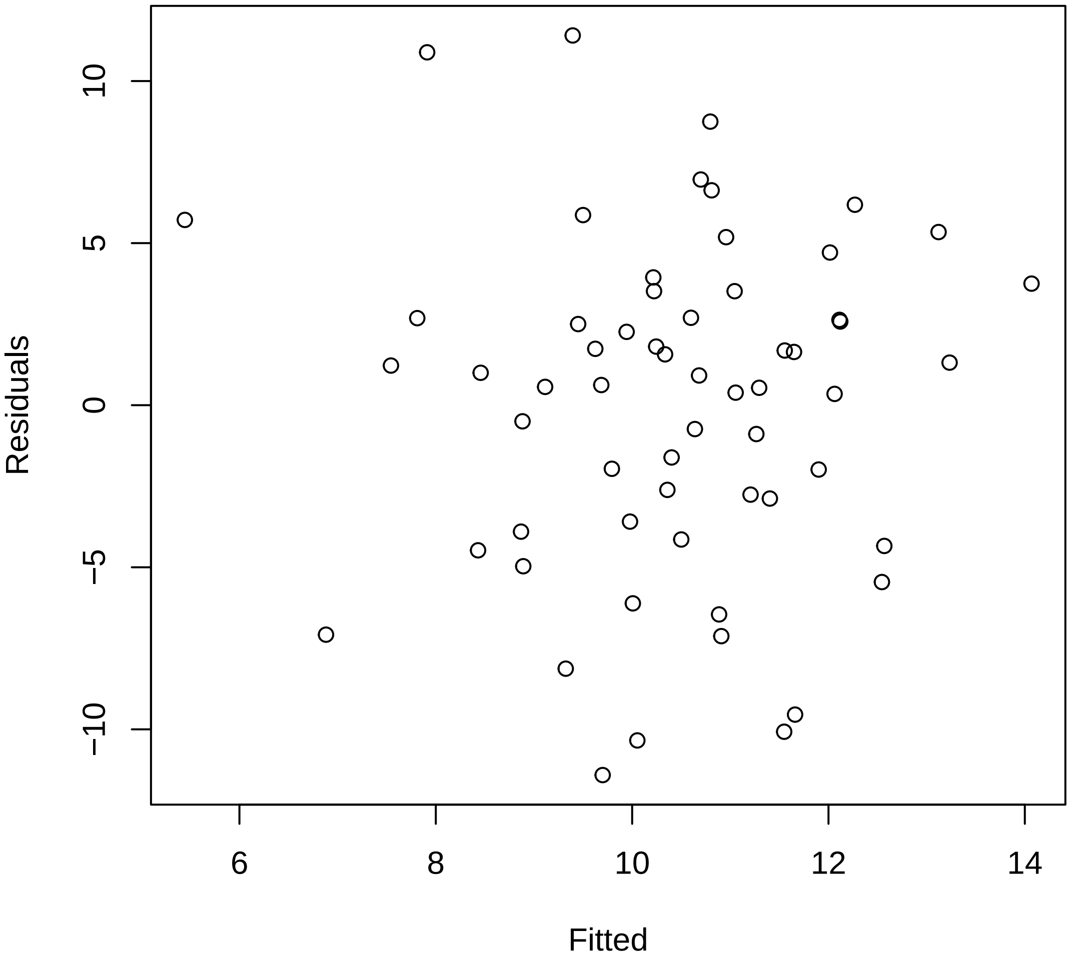

Figure 4: Residual Plot for UtilityCosts and MaintenanceCosts

The residual plot seems to be spread out, meaning that the variances between predicted and observed are distributed. There is no distinct pattern, thus we can say that the equal variance assumptions is met.

4. Simple regression assumption check

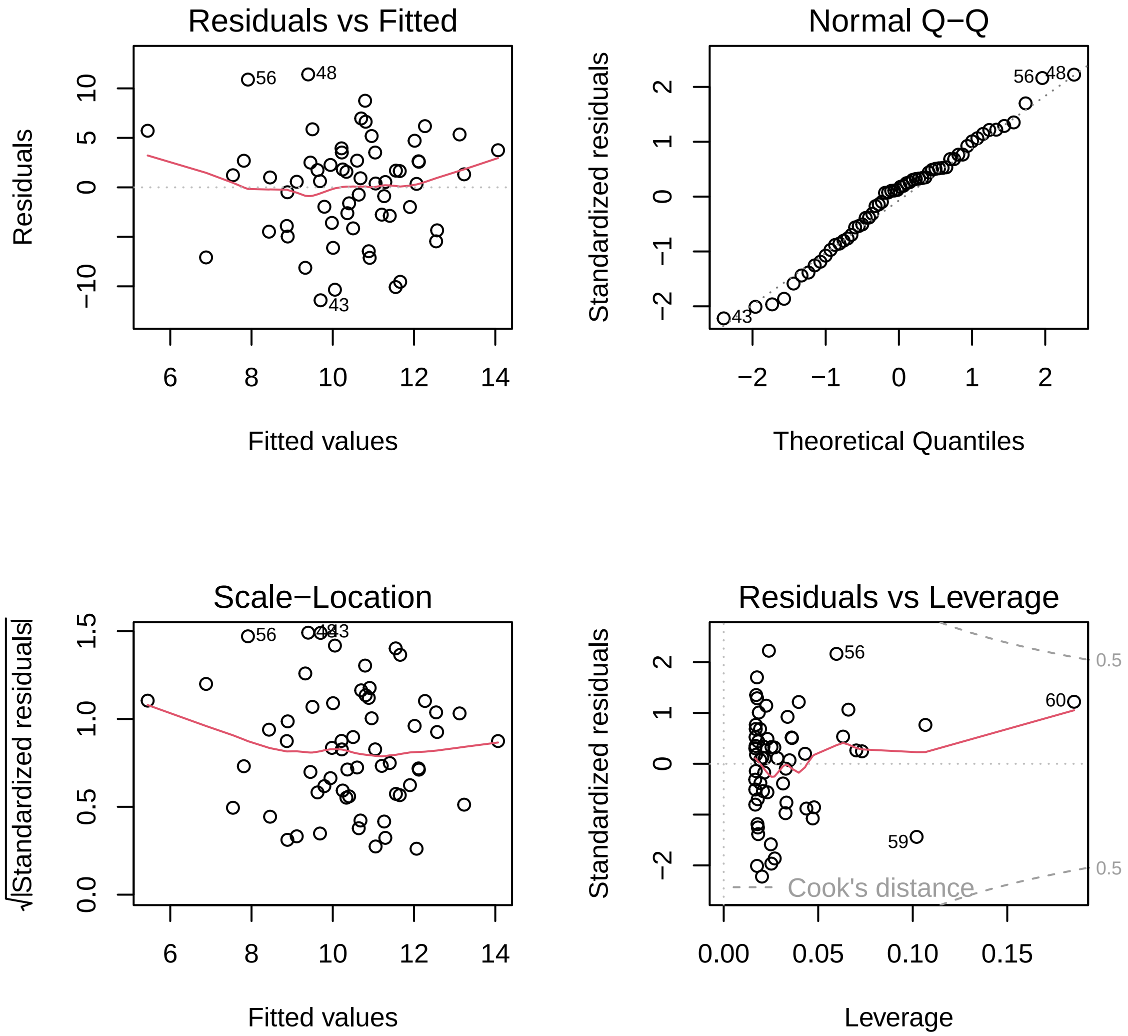

Figure 4: Simple Regression Plots for models

We can check the assumptions for the simple regression model using these plots. The residual plot on the left hand side, shows how the predicted versus observed values, similarly to Figure 4. Ideally, the red line will stay horizontally at 0. In this case, the red line is not fully horizontal, however it is still close to the zero line. The points are scattered and don't follow any pattern. We can conclude that the linearity and equal variances assumptions are met.

Independce of variance can be checked using the Scale-Location plot, on the bottom left hand. The red line supposed to be horizontally at zero for the assumption to be met. We can say that the assumption is met.

The normal Q-Q plots on the top right shows visually a straight line, meaning that the normality assumption is met.

5. Simple Linear Regression model

The assumptions are warranted for us to run a simple regression model.

```
##
## Call:
## lm(formula = MaintenceCosts ~ UtilityCosts, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.4106  -3.6688   0.7697   2.9020  11.4064
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  18.30949    3.43924   5.324 1.72e-06 ***
## UtilityCosts -0.14187    0.06072  -2.337   0.0229 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.193 on 58 degrees of freedom
## Multiple R-squared:  0.08604,    Adjusted R-squared:  0.07028
## F-statistic:  5.46 on 1 and 58 DF,  p-value: 0.02293
```

Table 2: Regression statistics

The slope of the regression is -0.14, meaning that for every one unit change in UtilityCosts, MaintenanceCosts changes averagely -0.14 units. Both the intercept and the slope are statistically significant with alpha = 0.05, due to p-value is close to 0. R squared is 0.08, meaning that the variation in UtilityCosts only explains 8% of variations in MaintenanceCosts.