Predicted Model of Diamonds Price

# Introduction

The price of a diamond depends on numerous factors, among which carat, color, clarity, and certification attract more attention from the public. While carat, the weight of a diamond stone, largely accounts for the price of a diamond, customers also care about the color purity and internally flawlessness. Moreover, different certification can have various judgment standard of a diamond properties resulting different cost. Therefore, a predicted model is needed to predict a diamond price.

The paper is to explore the relationship between Diamonds Price and Carat, Color, Clarity, Certification, and to set up a model with strong predictive power. In order to figure out the relationship, I utilize the multilinear analysis methods with R programming. After a series of analysis, I get the final mode (ln\_price = 3.24 + 1.35Caratc - 0.94Caratcsq + 0.20D + 0.16E + 0.13F + 0.09G + 0.14IF + 0.10VVS1), which can well explain the relationship between each factor and the price of diamonds.

# Analysis and Results

## Preparation for the Regression

Before conducting the regression, I draw the histograms of variable “Price” and “Carat” to see if there are any sever outliers and how the value is spreading. Here, outliers mean the values which are extremely abnormal or derived from the majority. Figure 1.1 and Figure 1.2 show that the data are quite normal and there are nor severe outliers.

Meanwhile, it’s necessary to plot between each predictor variables and response variable. Based on the result of Figure 1.3, Figure 1.4, Figure 1.5 and Figure 1.6, we can notice that there is something wrong with the Figure 1.6 *Scatterplot of Price vs Carat*. Although there is a clear positive relationship between Price and Carat, the trend of the plot appears to fan out, indicating higher carat will have higher price. Generally, a one-carat diamond will cost more than two half – carat diamonds, assuming all other qualities are equal (The 4Cs, 2018).

Therefore, we need to transform the variable “Price”, making it suitable for linear regression. In the following part, I take the logarithm of prices instead of price.

## The First Predicted Model

Based on Figure 2.1 *Scatterplot of ln(Price) vs Carat*, the relationship between Carat and the logarithms of Price appears more linear, compared to the former scatter plot. Therefore, it would be more judicious to employ ln(Price) in developing a linear regression model.

Before doing the multilinear regression, I define color “I” as the baseline, and compared it to the other five colors; I select clarity “VS2” as the baseline and compared it to the other four clarity grades (Chu Singfat, 2001)

From Figure 2.4, we can get the first predicted model with strong relationship between each variable and “ln\_price”.

**First Model: *ln\_price = 2.39 + 1.25Carat + 0.21D + 0.17E + 0.15F + 0.09G + 0.05H + 0.13IF + 0.05VS1 + 0.12VVS1 + 0.09VVS2 + 0.003GIA - 0.08IGI***

We need to do residuals analysis to test whether the model perform well to the data. However, from Figure 2.2 the points do not scatter randomly; also, the Q-Q plot does not form as a linear line. These show that the residuals are not normally distributed, and the predicted model is not appropriate enough. The residual plot indicates that the regression model underestimates prices at both ends of the price range and overestimates the midrange prices (Chu Singfat, 2001).

## Revise the model

Based on the statistical analysis and residuals analysis (Figure 3.1, 3.2, 3.3, 3.4), we can notice that there is a strong and positive relationship between each variable and ln\_price (exclude the certifications). And the residuals plots look much better. However, as I have added a transformed variable “caratsq” which is highly correlated with the untransformed variable carat, the VIF value of carat and “caratsq” is larger than 10 (the VIF carat = 42, the VIF caratsq = 37). (An ideal model should have VIF < 4, which is used to test whether there is correlation between each variable.)

For the quadratic model, replacing carat with carat - mean(carat) will remove the dependence while preserving the model. Based on the statistical analysis and residuals analysis output above, we can notice that there is a strong and positive relationship between each variable and ln\_price (exclude the certifications). And the residuals plots are nearly normally distributed. Most importantly, the VIF values have decreased to 2 and 1, which are smaller than 4. Here, we can get the second predicted model:

**Second Model: *ln\_price = 3.24 + 1.34Caratc - 0.92Caratcsq + 0.20D + 0.16E + 0.13F + 0.09G + 0.04H + 0.14IF + 0.03VS1 + 0.10VVS1 + 0.07VVS2 + 0.0004GIA - 0.01IGI***

## Consider about the intersections

From Figure 4.1, I employ several possible intersections into the second predicted model. I utilize the forward subset selection method to select a model ("Adjusted R^2" as criterion). This is an automatic model search strategy used to select the predictors in a regression model. Based on the output of "Adjusted R^2", we need to select the model with highest adjusted R^2 value 0.99. We can see that the intersections are not include in the selected model, so I will not add any intersections into the final model.

## Confirm the Final Model

Again, I use the forward subset selection method to select a model ("Adjusted R^2" as criterion), along with the "Mallow cp" method. They are both automatic model search strategies used to select the predictors in a regression model. This time the regression analysis is based on the second predicted model. Based on the output of two methods (Figure 5.1, 5.2), we can get the final predicted model, which includes the variable caratc, caratcsq, D, E, F, G, IF, VVS1.

**Final Model: *ln\_price = 3.24 + 1.35Caratc - 0.94Caratcsq + 0.20D + 0.16E + 0.13F + 0.09G + 0.14IF + 0.10VVS1***

# Conclusions

After a series of analysis, we get the final model, and from Figure 5.3, 5.4, 5.5, there is a strong relationship between chosen variables and ln(Price). The VIF values are all less than 4, and residuals are normally distributed. These showing our model has strong predicted power, and the coefficients are severely sensible.

**Final Model: *ln\_price = 3.24 + 1.35Caratc - 0.94Caratcsq + 0.20D + 0.16E + 0.13F + 0.09G + 0.14IF + 0.10VVS1***

While all other things being equal, based on the final model, ln\_price increases 0.20 according to D, while increases 0.16 according to E. The price will increase 100.20 and 100.16 respectively. Thus, the average price difference between a grade D diamond and one graded E is (100.20 - 100.16) US $.

**Second Model: *ln\_price = 3.24 + 1.34Caratc - 0.92Caratcsq + 0.20D + 0.16E + 0.13F + 0.09G + 0.04H + 0.14IF + 0.03VS1 + 0.10VVS1 + 0.07VVS2 + 0.0004GIA - 0.01IGI***

From the second model, we can see there are price differences among the three certification bodies, while the coefficient of GIA, IGI are 0.0004 and -0.01 compared to HRD.

# References

Caspi, A. (1997, August). MODERN DIAMOND CUTTING AND POLISHING. GEMS & GEMOLOGY.

doi:<https://www.gia.edu/doc/Modern-Diamond-Cutting-and-Polishing.pdf>

Chu Singfat (1996) Diamond Ring Pricing Using Linear Regression, Journal of Statistics Education, 4:3, DOI:

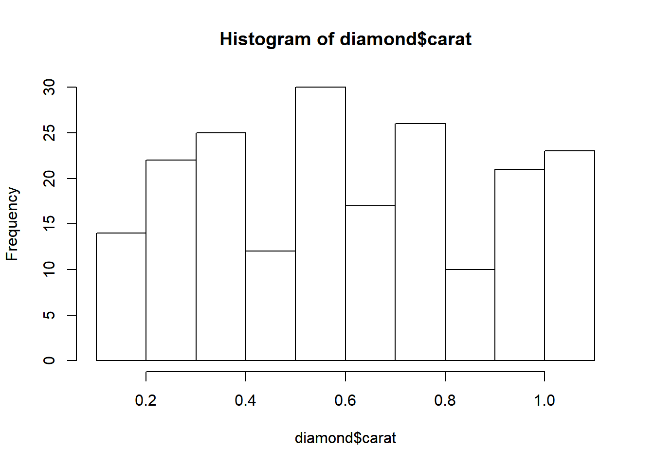
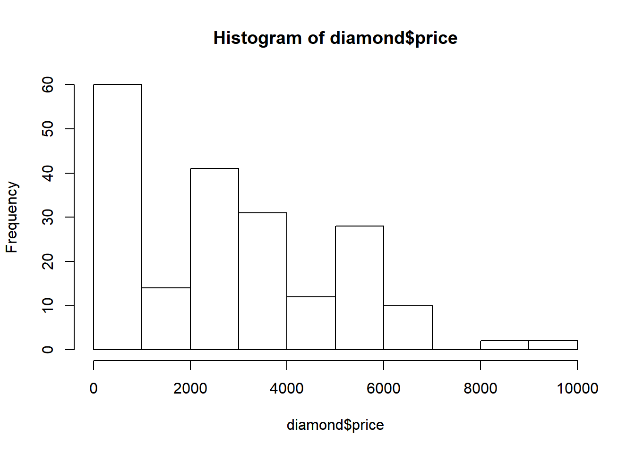
10.1080/10691898.1996.11910519

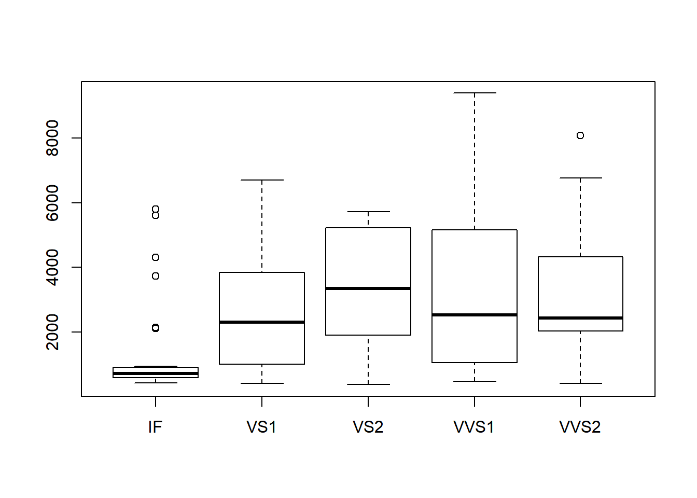
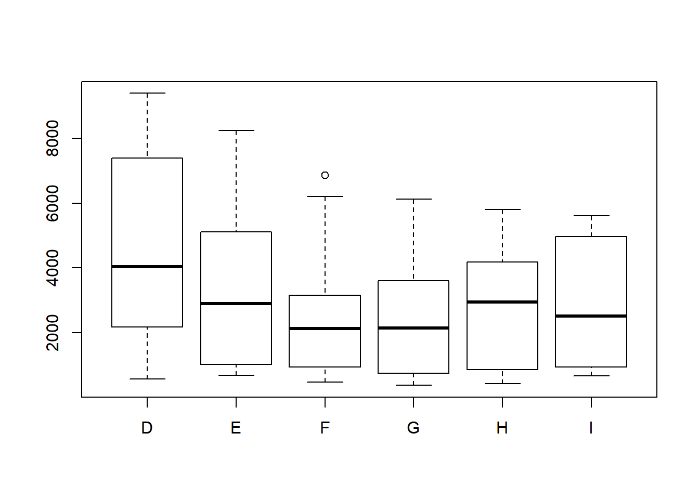
Chu Singfat (2001) Pricing the C's of Diamond Stones, Journal of Statistics Education, 9:2, DOI:

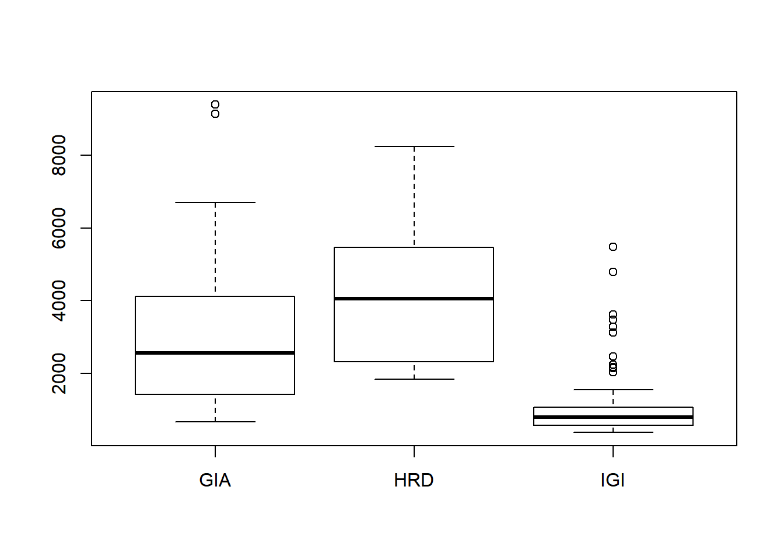
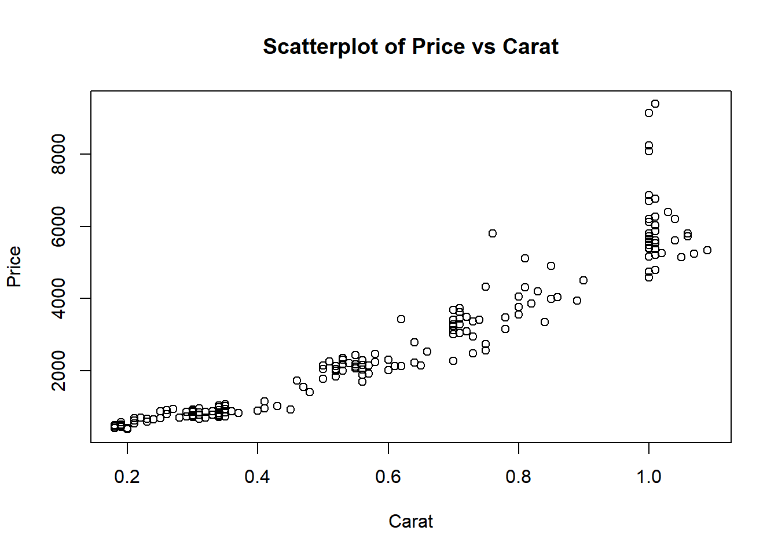
10.1080/10691898.2001.11910659

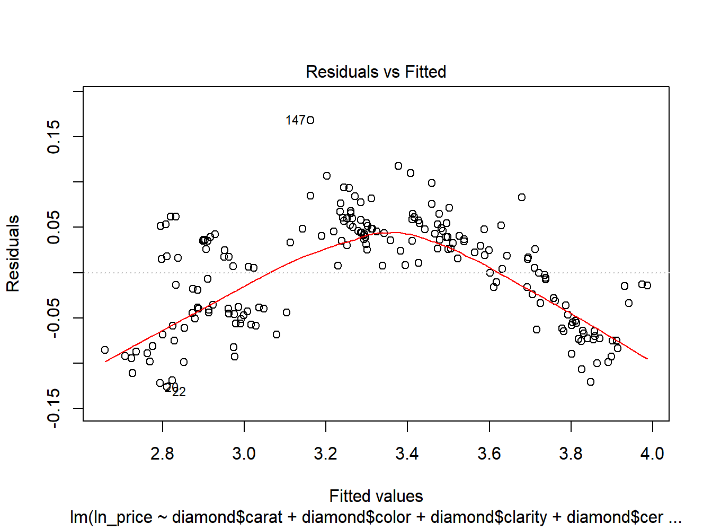
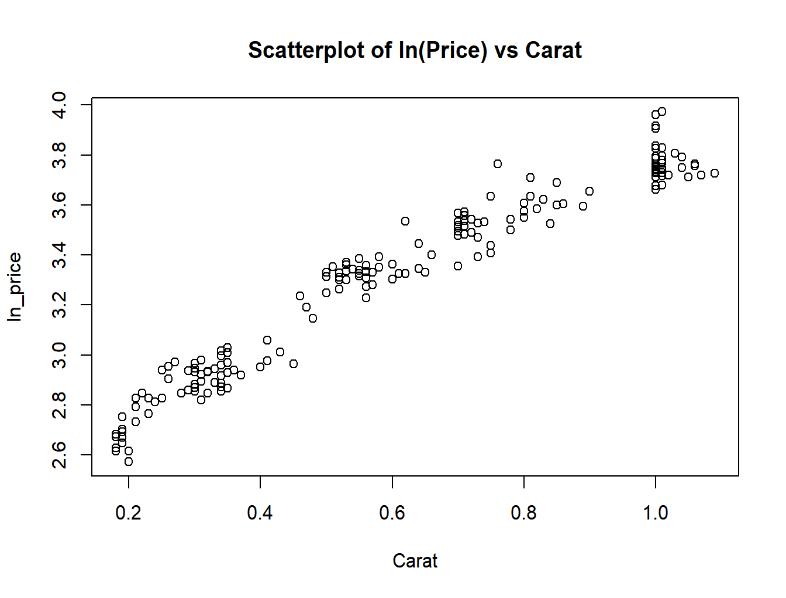
THE 4Cs. (n.d.). Retrieved from <https://www.debeersgroup.com/a-diamonds-journey/the-4cs>

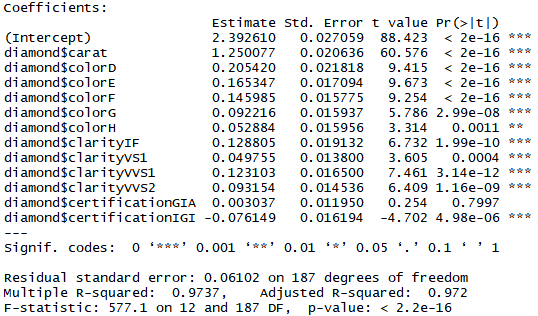
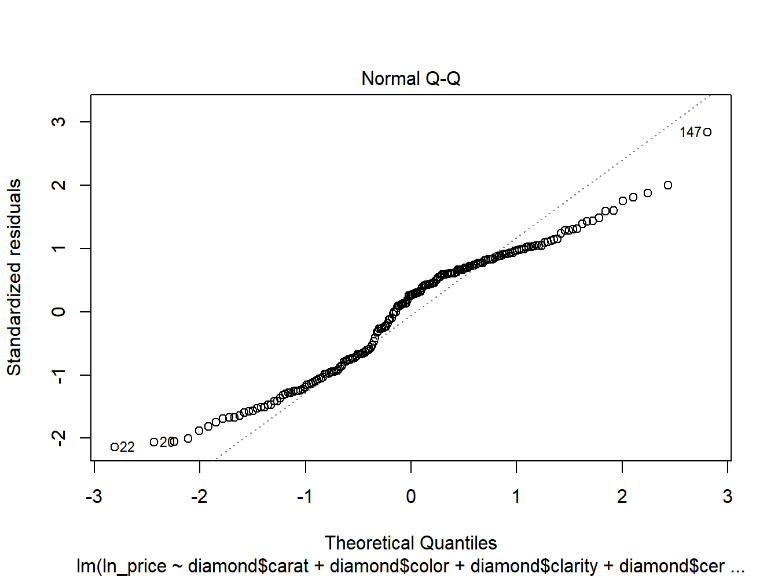
# Appendices

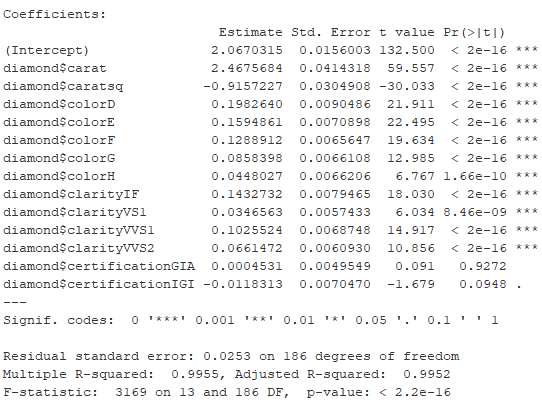
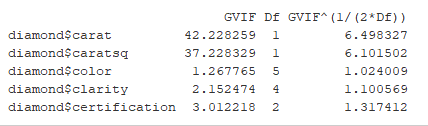
 Figure 1.1 Histogram of Price Figure 2.2 Histogram of Carat

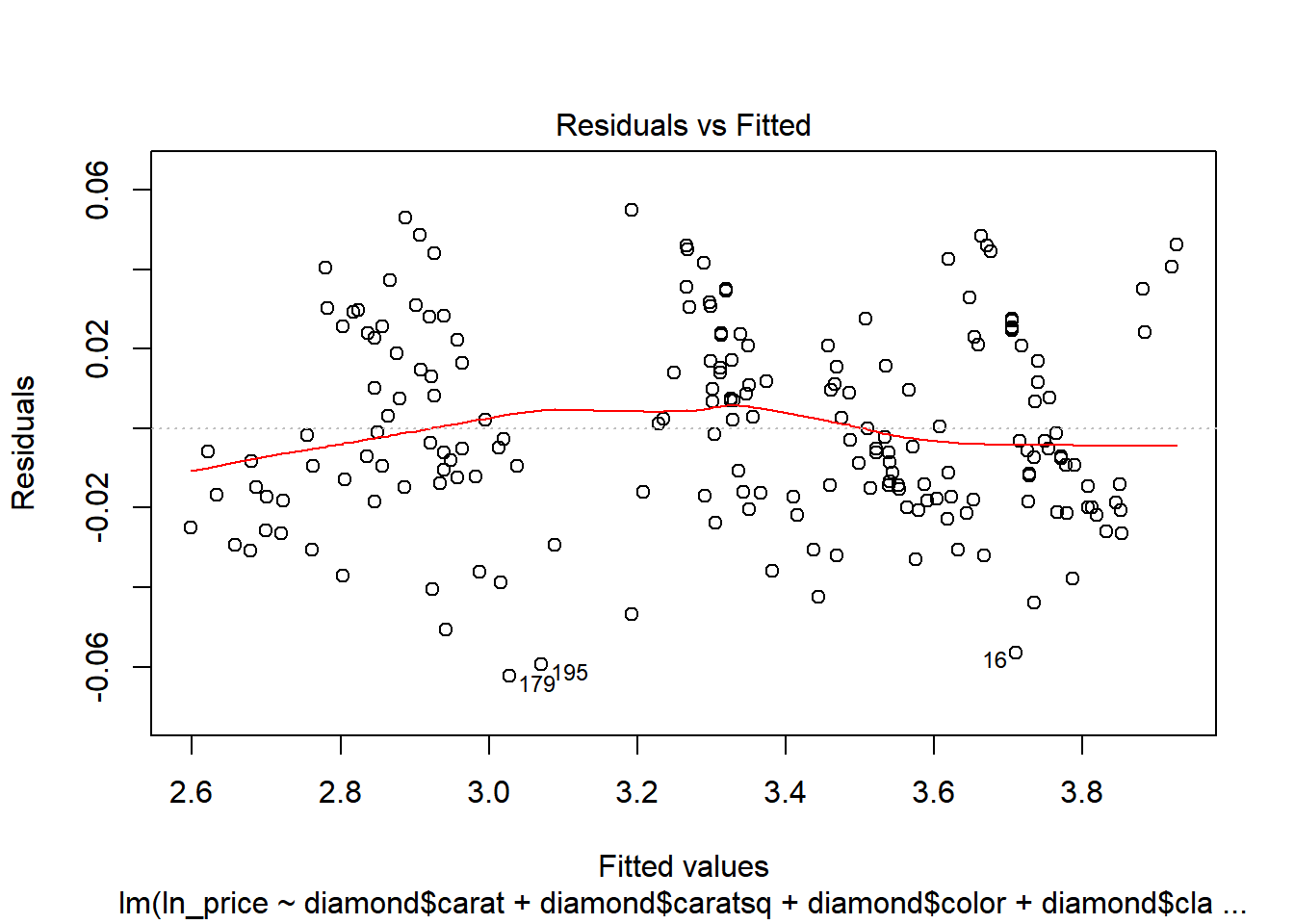
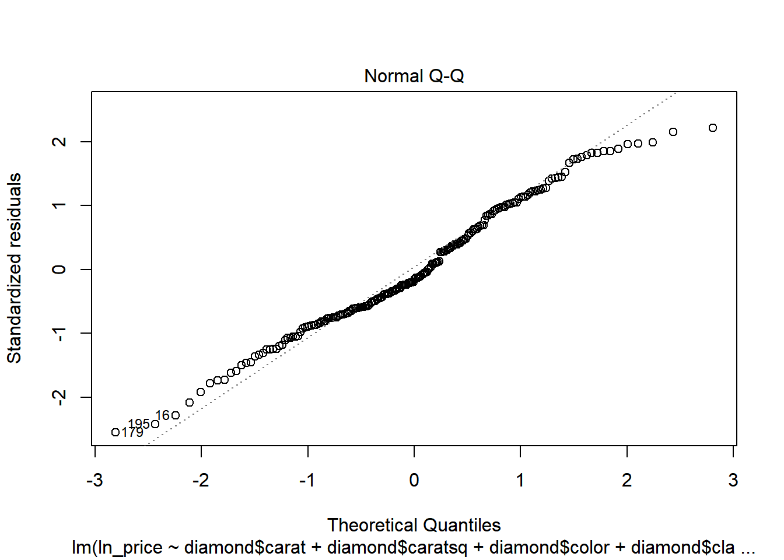
 Figure 1.3 Plot of Price vs Color Figure 1.4 Plot of Price vs Clarity

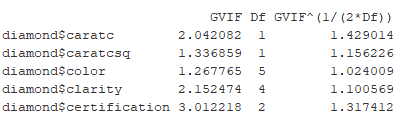
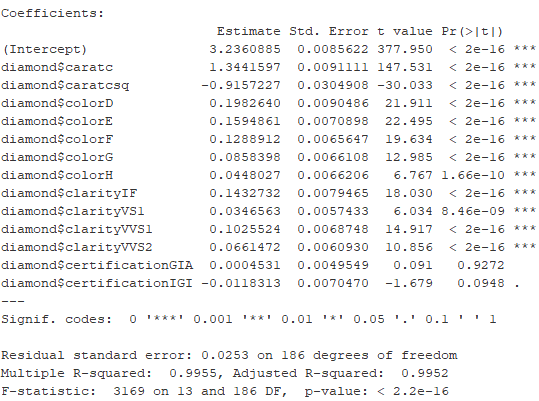
 Figure 1.5 Plot of Price vs Certification Figure 1.6 Scatterplot of Price vs Carat

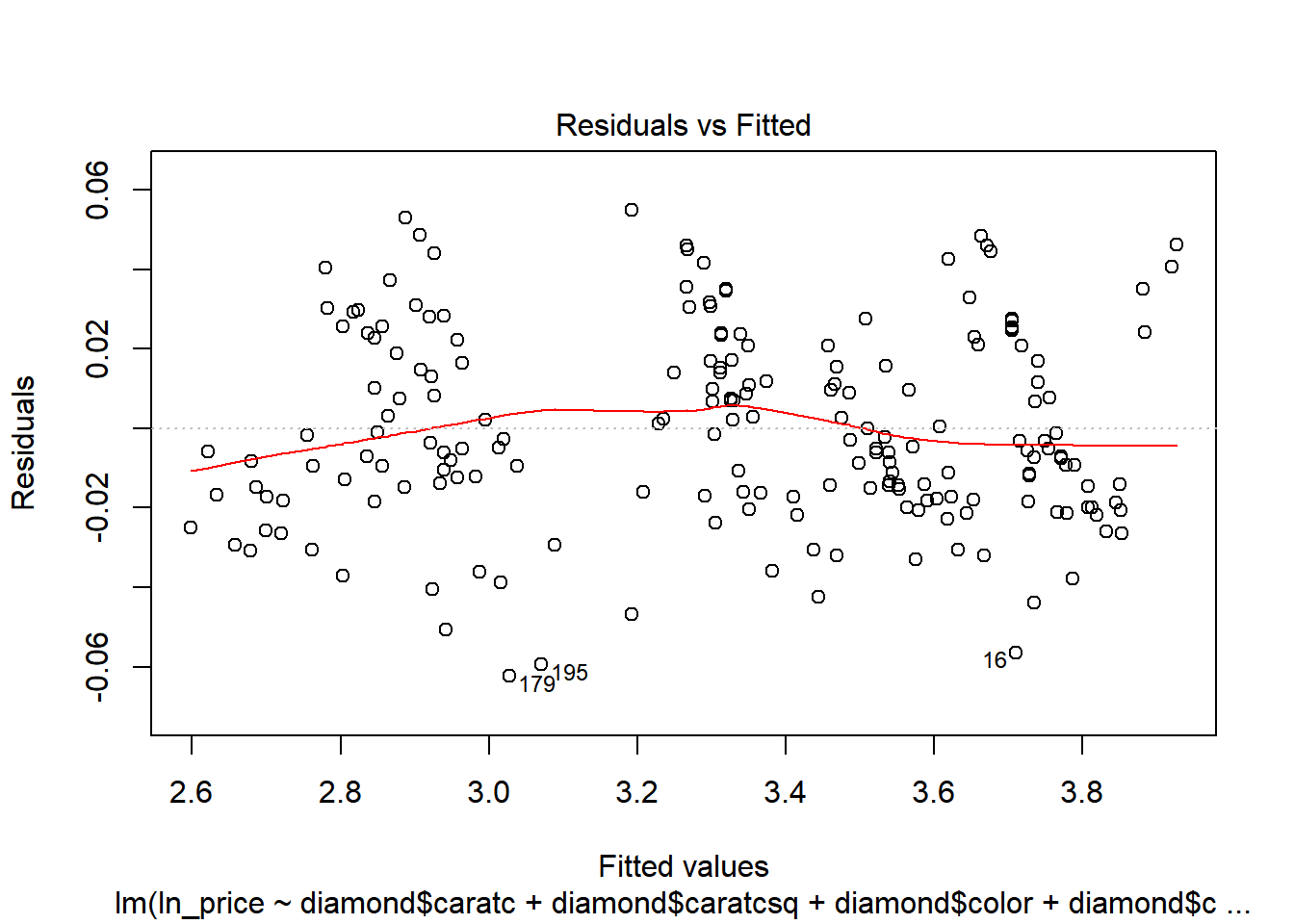
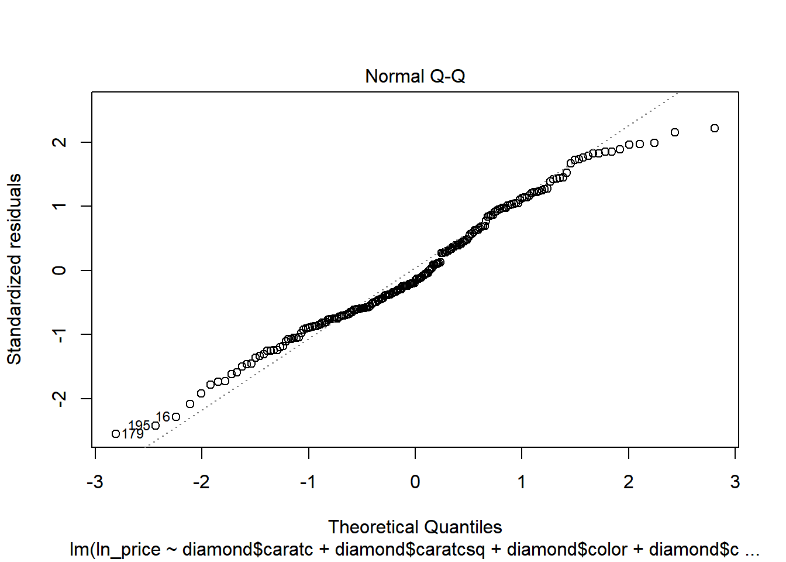
 Figure 2.1 Scatterplot of ln(Price) vs Carat Figure 2.2 Residuals vs Fitted

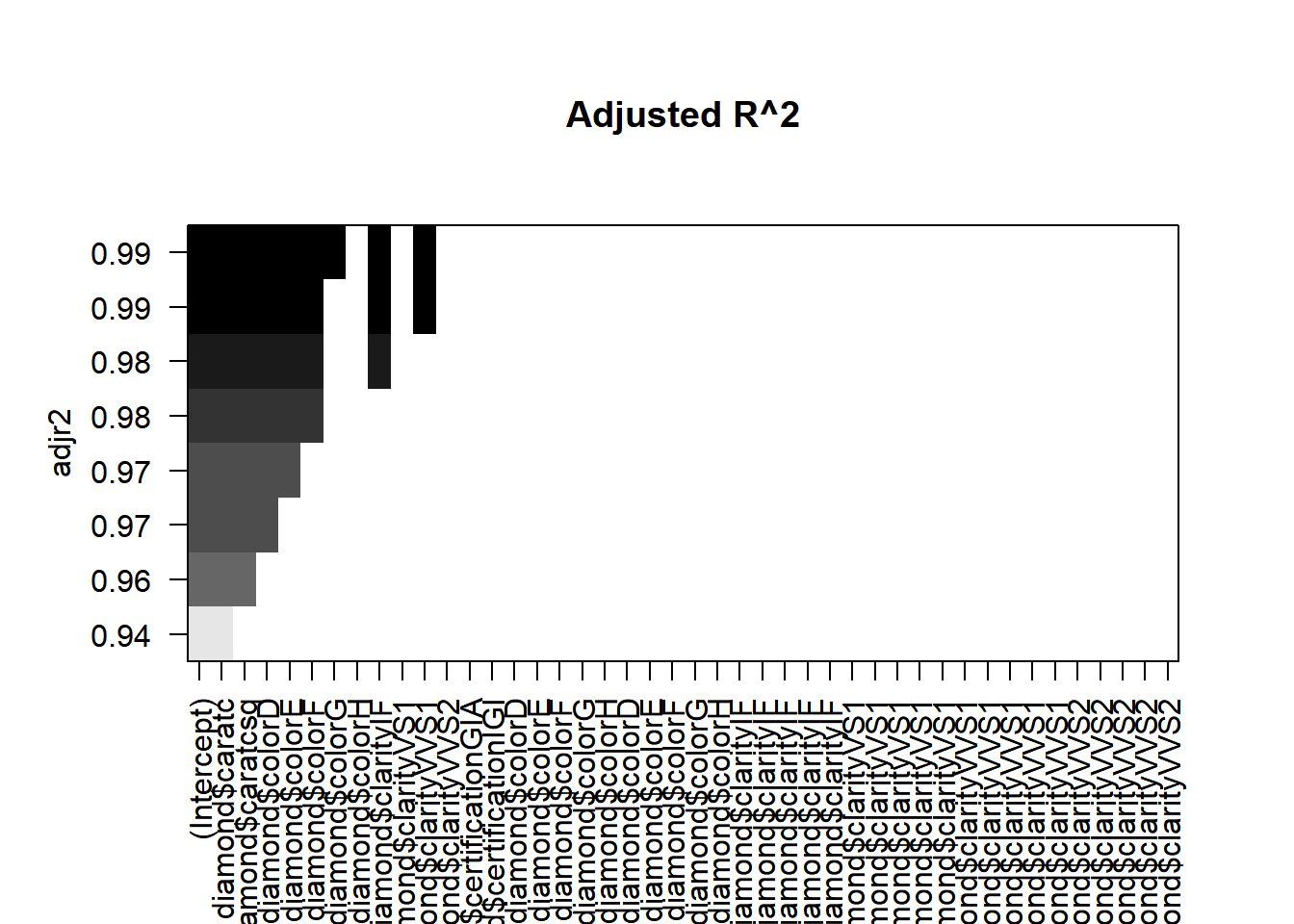
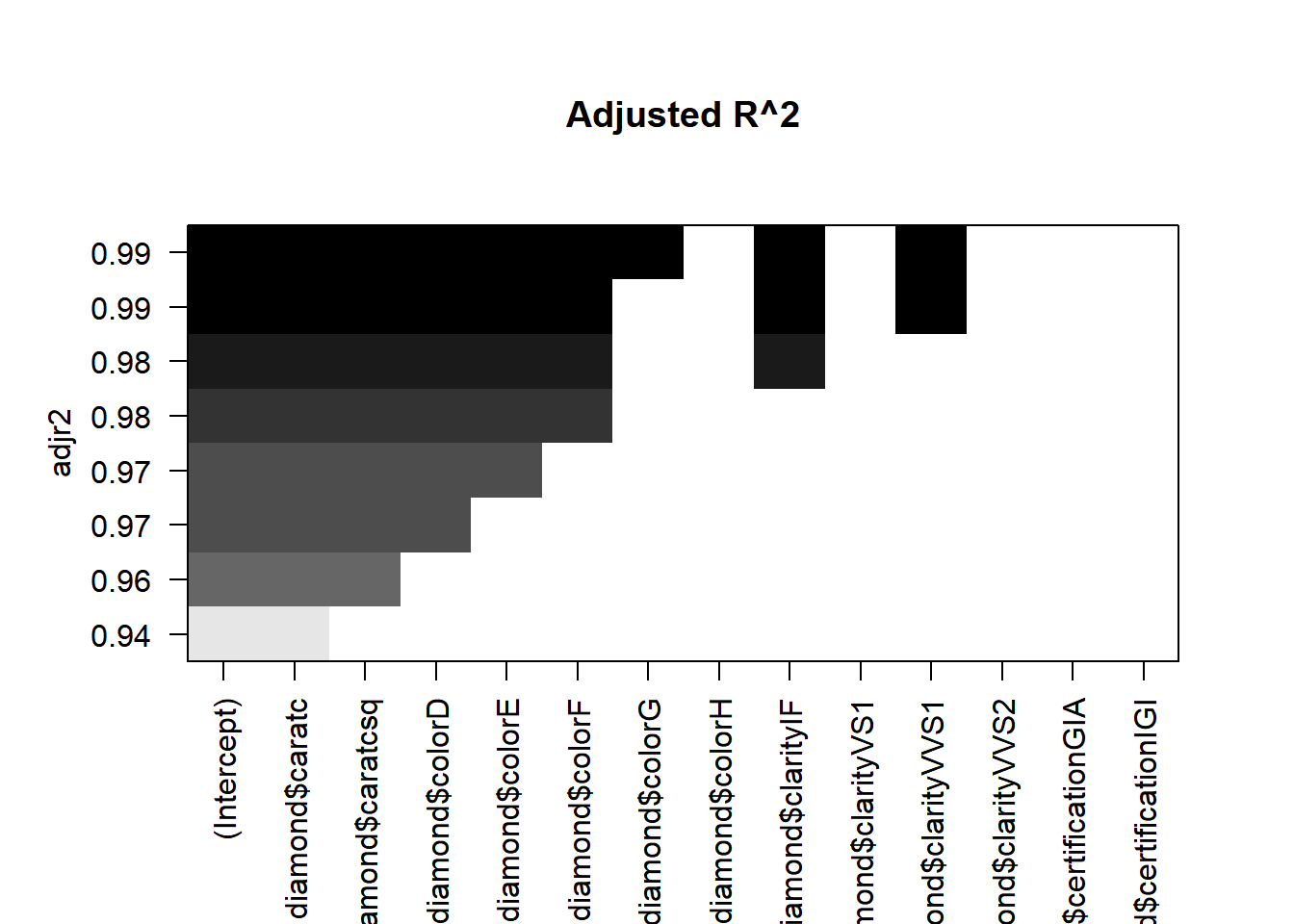
 Figure 2.3 Normal Q-Q Plot Figure 2.4 First Predicted Model

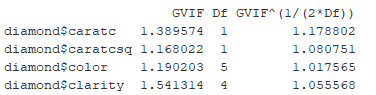
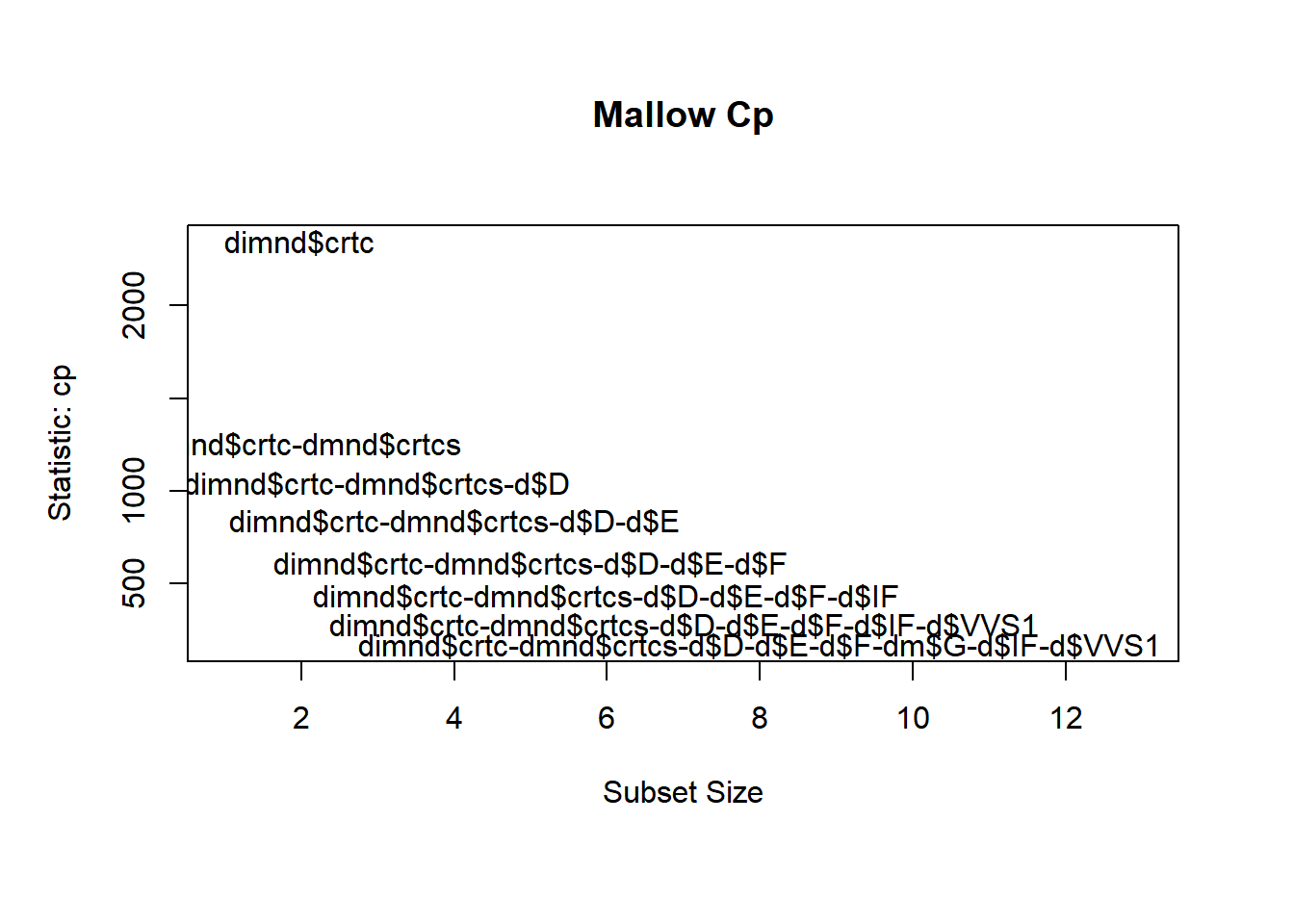
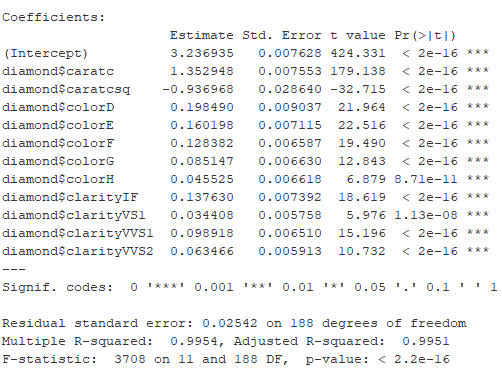
 Figure 3.1 Predicted Model Figure 3.2 VIF Value

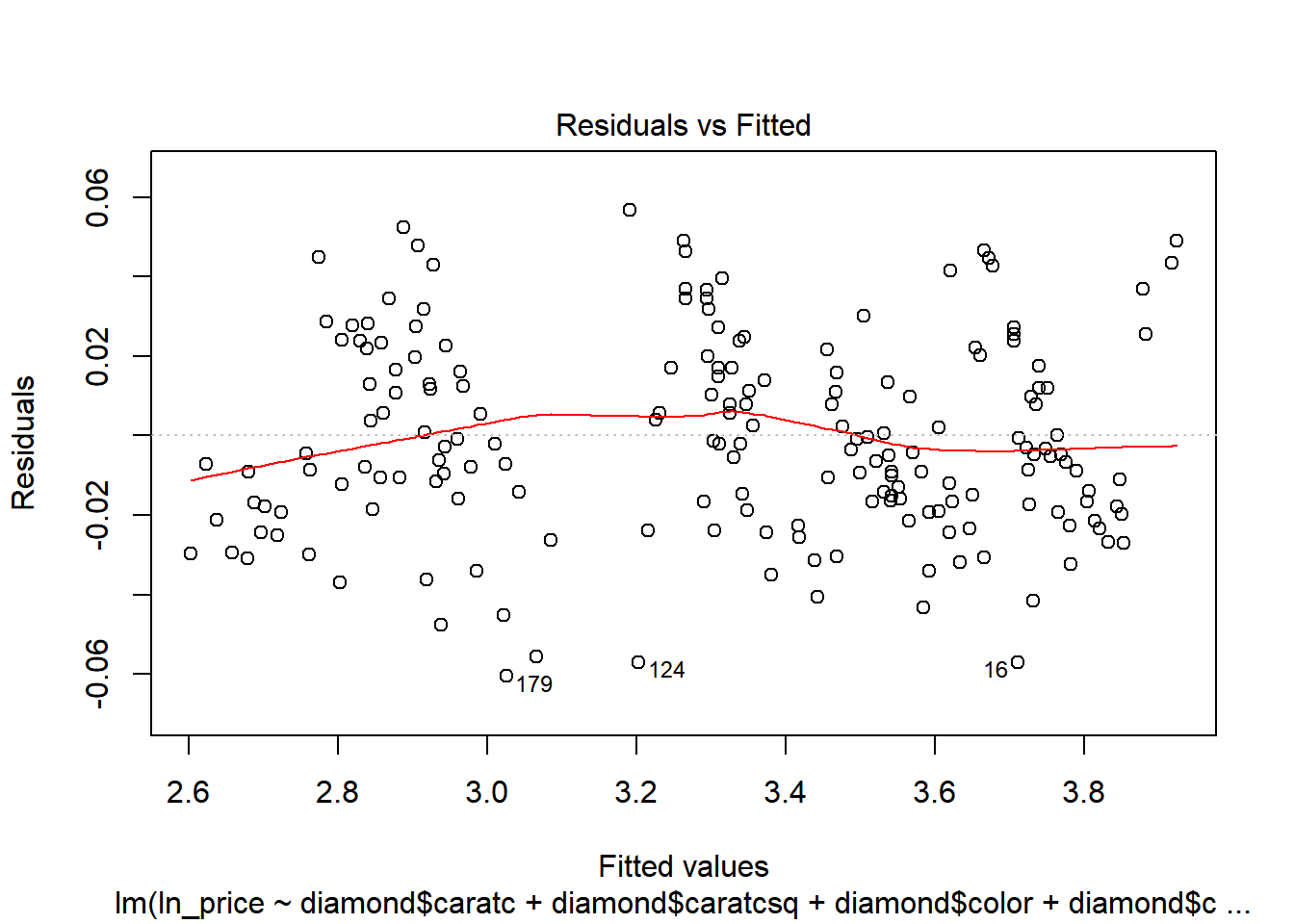
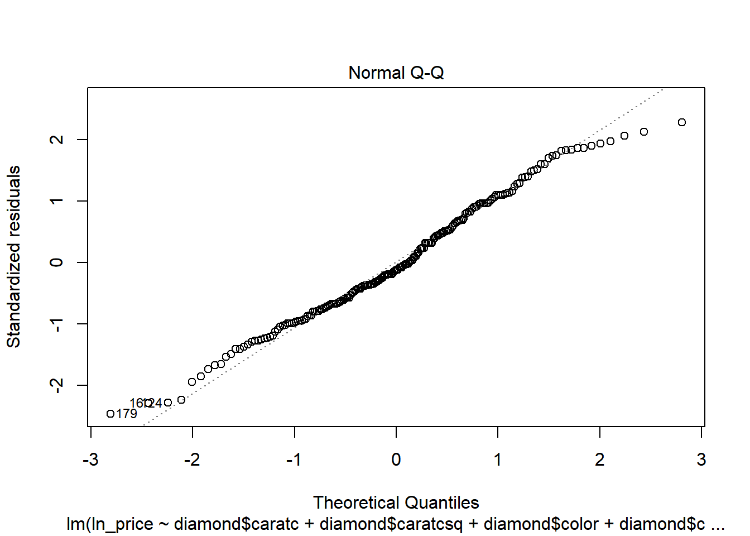
 Figure 3.3 Residuals vs Fitted Figure 3.4 Normal Q-Q

 Figure 3.5 Second Predicted Model Figure 3.6 VIF Value

 Figure 3.7 Residuals vs Fitted Figure 3.8 Normal Q-Q

 Figure 4.1 Adjusted R^2 Figure 5.1 Adjusted R^2

 Figure 5.2 Mallow Cp Figure 5.3 Final Predicted Model

 Figure 5.4 Residuals vs Fitted Figure 5.5 Normal Q-Q