Rupesh Gudipudi

Jan Hannig

STOR 455

21 November 2019

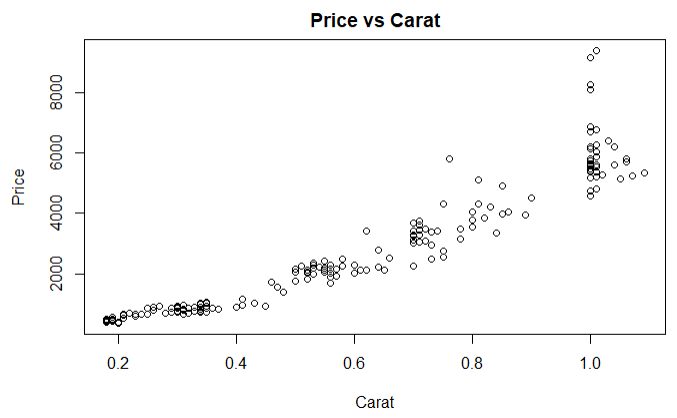
Predicting the Price of Diamonds

**Introduction**

Diamonds are highly valued stones that are known for their use in jewelry, because of their durability and luster. They are highly valued because they are pretty rare to find, so therefore, they can be pretty expensive. But the price of diamonds tends to depend on the characteristics of them. Some characteristics that are taken into consideration are carats, color, clarity, and certification. There are different types of color – they are graded on a scale that ranges from D (colorless) to Z. Clarity refers to the amount of inclusions in a diamond and are also on a scale, ranging from F (flawless) to I (included). Lastly, certification means where the diamond was certified.

**Analysis and Results**

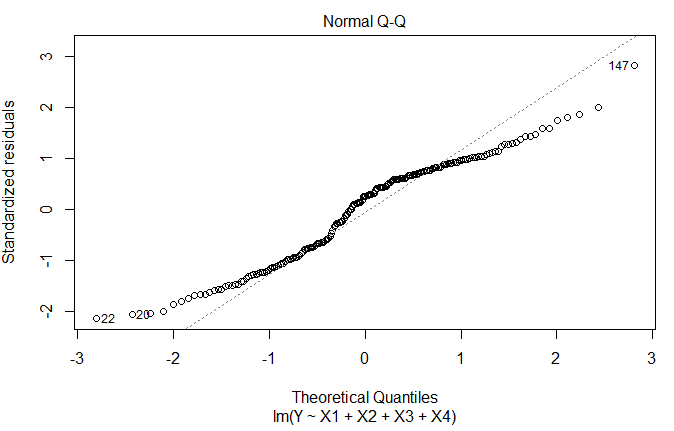
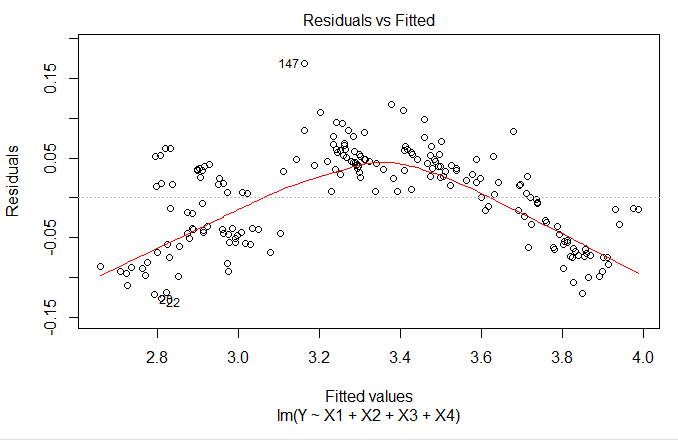
I am currently a consultant for a firm that sells diamonds and my job is to create a model that predicts the price of a diamond based on these characteristics. I first simply created plots that showed the correlation between price and each of the characteristics. The plot for Price vs Carat was a normal scatterplot while the others were box and whisker plots because the others tend to be categorical variables.



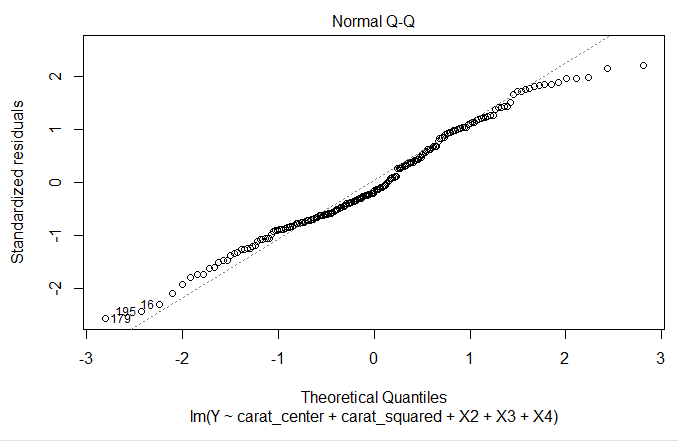
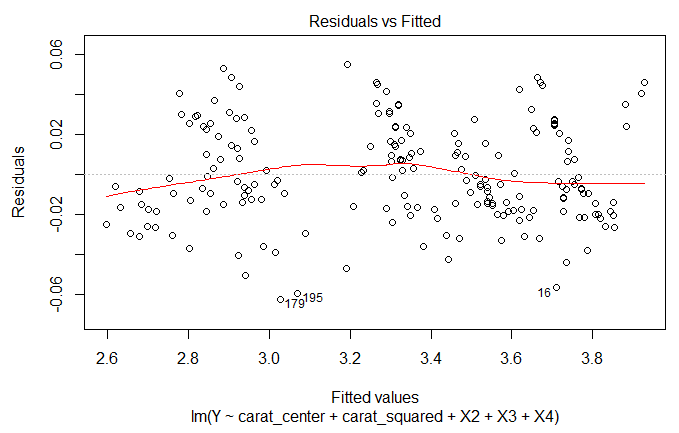
As you can see the from the scatterplot of Price vs Carat, there is a clear positive relationship between these two. This means the more carats in a diamond, the more expensive it will be.

Next, I created a linear model that uses all the other characteristics to predict price. The variable Y is used for price and is variable we’re measuring, while the X’s explain the other characteristics and are used for the independent variables. I made sure to use a log scale for the predictor because the intercept tends to be the expected value of the response when predictor is 1, and the slopes measure the expected changes in the responses when the predictor increases by a certain amount. It’s also easier to analyze the data with smaller coefficient values. I used log base 10 to compute common logarithms because this tends to be beneficial for large data sets, such as the one that we’re using (Biomembers 2018). My linear model uses I color grade, IF clarity, and GIA certification as the base cases, which means these coefficients get absorbed into the intercept and are not present in the equation. The final result of my model was price = 2.52 + 1.25Carat + .21D + .17E + .15F + .09G + .05H - .08VS1 - .13VS2 - .01VVS1 - .04VVS2 - .003HRD - .08IGI. This model seems pretty reasonable considering the coefficient for carat is the highest, since that is the one that will have the highest impact on price. All of the clarities are in the negatives because IF is being used for the intercept, which means that IF probably has a huge impact on price. The same goes for the certifications; GIA probably has the biggest impact on price out of the three certifications.

However, when forming the residual plot and doing the residual analysis, it seems that the ab-line is not straight. In addition, the residuals are not normally distributed and the QQ plot is not linear. Therefore, this model is not sufficient enough.



I then decided to create another model. For this one, I centered the carat variable to make the regression constants more meaningful. I basically took the difference between the carat value and the mean and then squared it and incorporated both of these into the linear regression model. Once I have done that, my final linear model turned out to be price = 3.38 + 1.34carat\_center - .92carat\_squared + .20D + .16E + .13F + .09G + .05H - .11VS1 - .14VS2 - .04VVS1 - .08VVS2 - .0005HRD - .01IGI. This also seems reasonable, considering the statistical analysis is similar to the one mentioned earlier. However, after analyzing the residual plots, this model seems to be fitting. The ab-line is much straighter and the residuals are more normally distributed. The QQ plot is also pretty linear.



**Conclusion**

Based on both the statistical and residual analyses, this model seems very sensible. There is a very strong relationship between these characteristics and price – carats seem to have the highest impact on it. The average price difference between a grade D and grade E diamond is 10^(.16) – 10^(.2) which is -0.13945. There is a price difference between the HRD and IGI certification bodies and that is equal to 10^(.01) – 10^(.0005) which is equal to 0.02444. However, you can’t incorporate the third certification body (GIA) into the price difference because this is the base case and is being taken into account within the intercept.

References

Biomembers. “Log Base 2 or e or 10?” *BioTuring's Blog*, 14 Nov. 2018, blog.bioturing.com/2018/04/26/log-base-2-or-e-or-10/.