Project 1

Yonghyeon Kweon

Logan Lee

Taeheon Park

Jay Choi

STAT 6021

Professor Taylor Brown

Section 001

July 19, 2019

**Introduction**

From the class, our project group has been given a diamond dataset. The dataset has one response variable (price) and four explanatory variables in terms of the diamond. There are three categorical variables (clarity, color, and cut) and one continuous variable (carat) in the explanatory variables. By using the dataset, our group tried to come up with the most interesting topic, and we ended up obtaining one topic which is figuring out the price of the diamond shown in a famous movie, *Titanic(1997)*. The name of the diamond is Heart of the Ocean, and it has 56-carat weight. Even though the diamond exists only in the movie story, Heart of the Ocean story in the movie was based on Hope diamond, a real piece of jewelry. For the better estimation of Heart of the Ocean, our group decided to estimate Hope diamond first because Hope diamond already has a price estimated by jewel appraisers. Hope diamond’s weight is 45.5-carat. Through the estimation of Hope diamond’s price, we can see how accurate our model is. Also, it will be possible to estimate the price of Heart of the Ocean by considering all other factors.

**Data summarization**

The given dataset includes one response variable (price) and four independent variables (carat, clarity, color, and cut). Price represents the price of each diamond. Carat represents the weight of each diamond. In this dataset, the range of carat is from 0.23 to 20.45. The tables below explain the details of each categorical variable.

**Clarity**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Flawless | Internally Flawless | Very Very Slightly Included | | Very Slightly Included | | Slightly Included | | Included | | |
| **Grade** | **FL** | **IF** | **VVS1** | **VVS2** | **VS1** | **VS2** | **SI1** | **SI2** | | |

**Color**

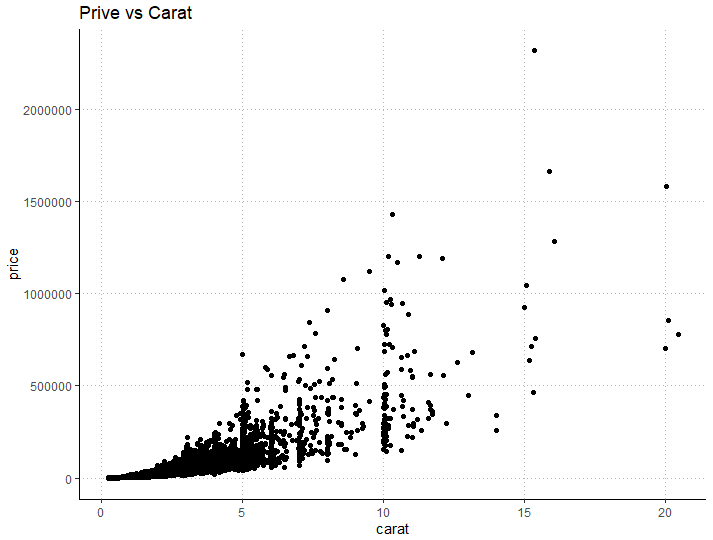
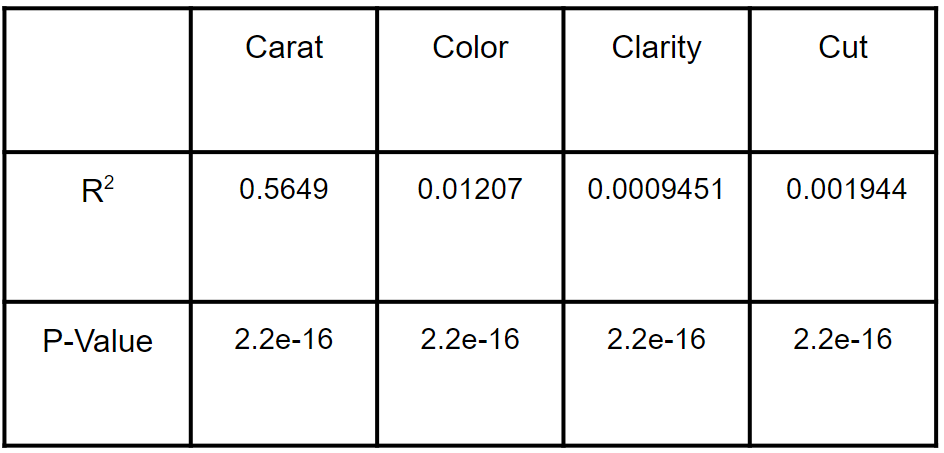
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| D | E | F | G | H | I | J |
| Colorless | | | Near Colorless | | | |

**Cut**

|  |  |  |  |
| --- | --- | --- | --- |
| Ideal | Astor Ideal | Very Good | Good |

**Exploratory Data Analysis**

Prior to conducting analysis, linear models were fitted for each predictor to decide which predictor is the most important. By using R, the values of R2 were calculated and recorded in the table. The table below shows that the carat predictor explains the total variability in price the most, so was chosen as the first model. The scatter plot on the right shows the relationship between the price and the carat. These two variables have a positive relationship, and the price variance clearly increases as the carat increases.

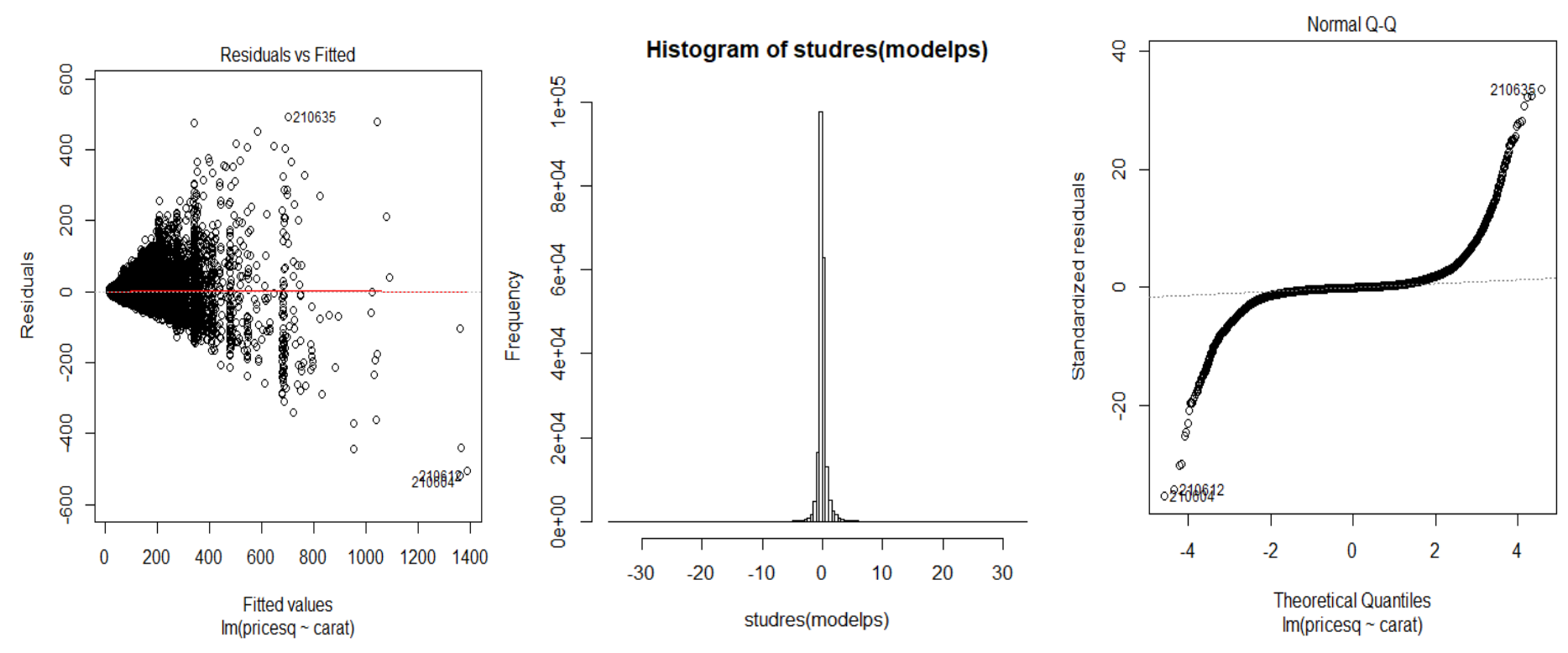


**Data Analysis**

Our goal is to find the most adequate model to predict the price of diamonds based on the regressors in the dataset. In order to fit the straight-line model, a response variable and predictors are transformed. In addition, linearity, homoskedasticity, and normality assumptions have been tested to see if the model have constant variances and linear relationships between the response variable and the predictors.

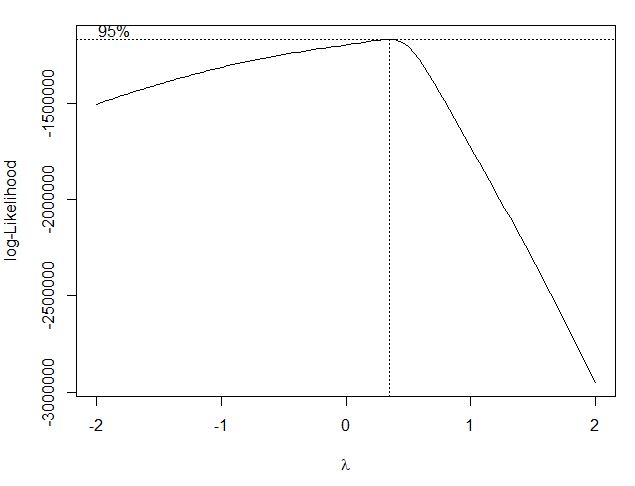
*Transformation of response variable (price)*

Since price versus carat scatter plot is showing a quadratic curve, price is transformed into square root of price to fit . The new model, , has shown an increase of R2 from 0.5649 to 0.9158 which is a huge improvement. Assumptions of homoskedasticity and normality should be checked by using a residual versus predicted response plot, histogram of residual distribution, and Q-Q plot; an external studentized residual is used for these three assumptions. If the points in the residual plot have the pattern, it’s a violation of the homoskedasticity assumption. If the points in the residual plot don’t follow the horizontal direction, it’s a violation of the linearity assumption. If the residuals are not normally distributed in the Q-Q plot or the histogram, it’s also a violation of the normality assumption.

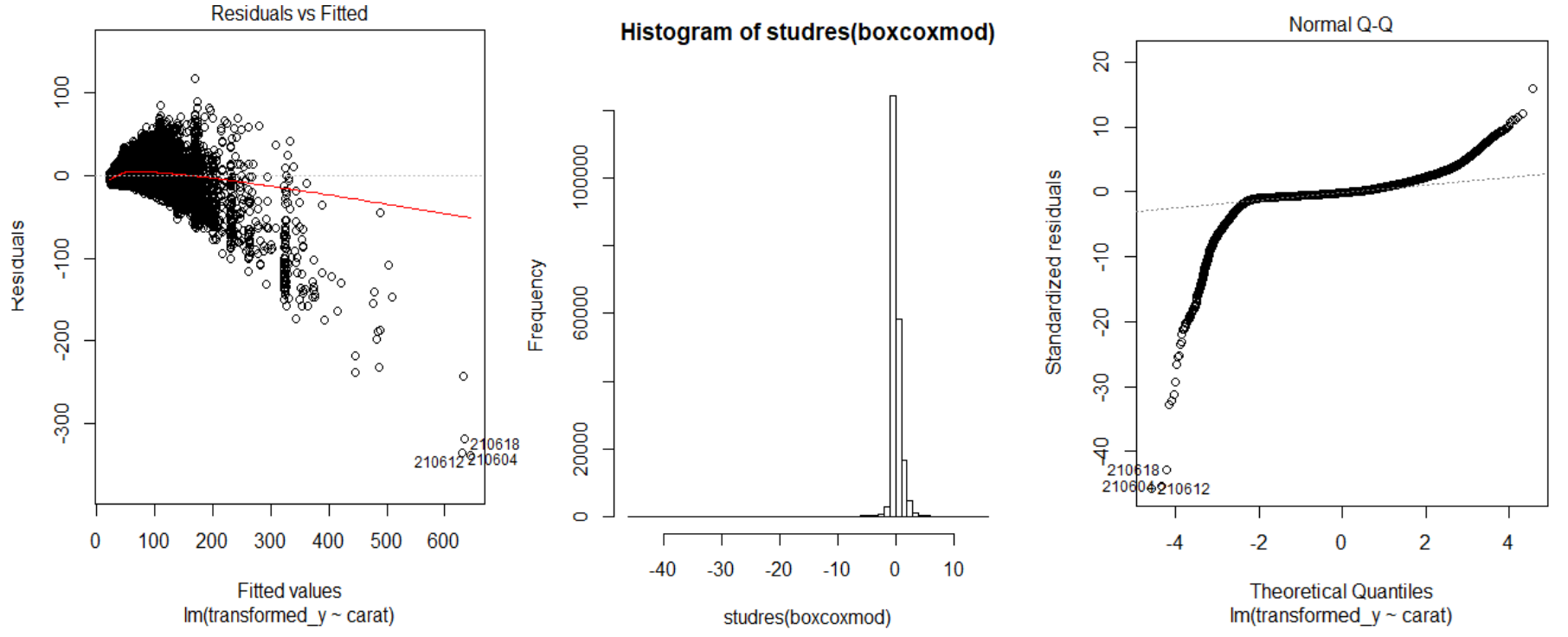


The residual versus predicted response plot shows the residuals have fanning out pattern horizontally which means the linearity assumption is satisfied while the homoskedasticity assumption is violated. Q-Q plot shows the residual has a fat-tailed distribution, so the normality assumption is violated.

*BoxCox procedure*



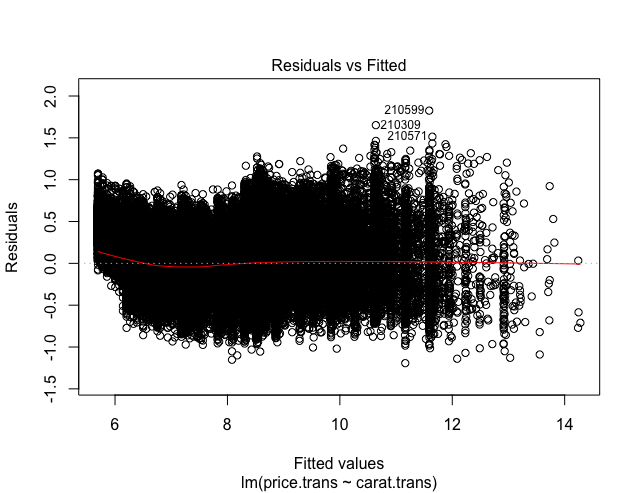
Transformation by square-rooting the price was not the best choice, so BoxCox procedure was used to transform the price. This procedure is useful when a variance is not stabilized. The lambda was calculated by using R and it came out to be around 0.3434 (the plot above shows the best lambda). The third linear model fitted was



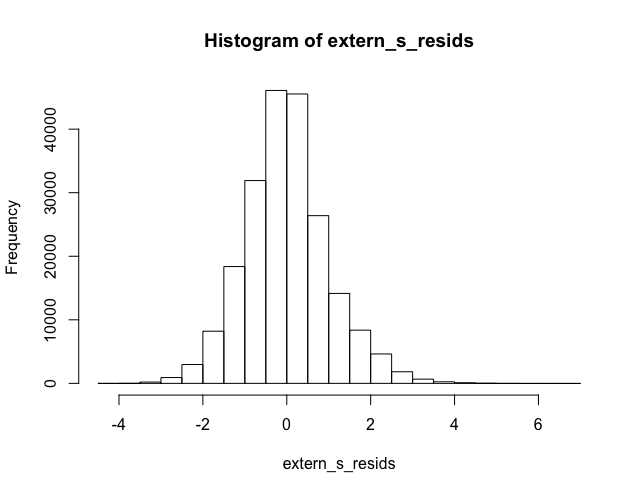
The result became worse than the square-root transformation. The R2 decreased from 0.9158 to 0.897. Also, all three of the assumptions were violated because the residuals had the downward fanning-out pattern in the first plot, and the residual distribution is heavily skewed to the left in the second and third plot. The assumptions were violated even after the price transformation. It indicated both the price and the predictor transformations were needed.

*Transformation of both response variable and predictors*

As a new model, both the response variable (price) and the predictor (carat) have been transformed into log of price and log of carat. After the transformation, it represents greater R^2 from 0.5649 to 0.9501 which means the data is close enough to the fitted regression. In other words, around 95% of the variation is explained by the model, In order to see if the variance is stabilized in the model, the external studentized residual has been calculated and plotted.



As the plot shown above, it doesn’t have any pattern so it satisfies the linearity and the homoskedasticity assumptions. Therefore, there exist the equal variance along the regression line and the linear relationship between variables. Also, the test for the violation of normality assumption is accomplished through the inspection of residual from the regression model. The histogram below shows a normal distribution of residuals.



In order to check if other variables (cut, color, and clarity) contribute to the model to predict the price of a diamond, partial F test has been conducted.

Full model:

Reduced model:

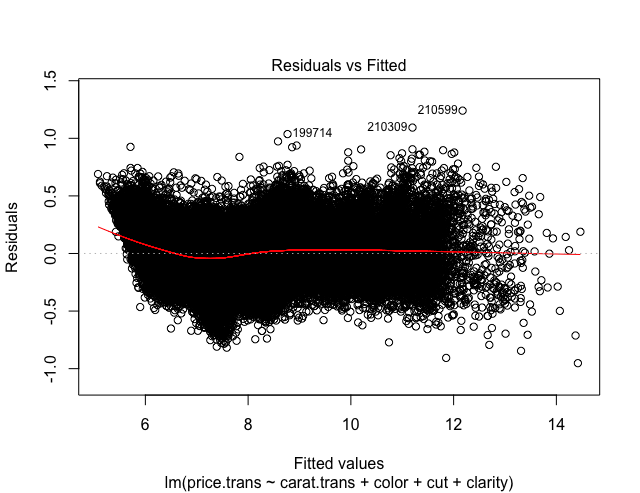
The null hypothesis and the alternative hypothesis of the test are:

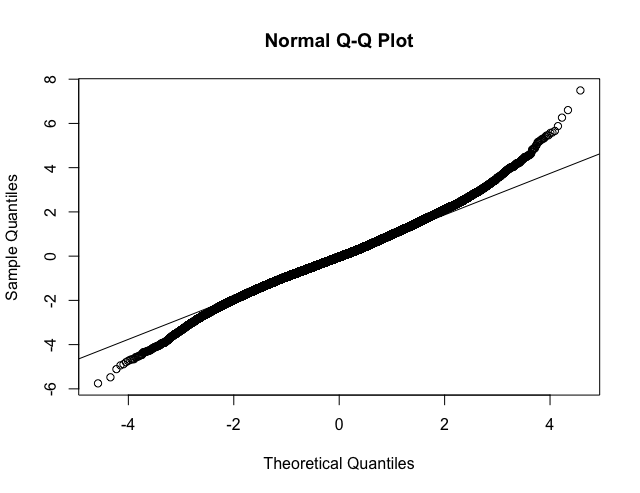
H0 : = = = 0

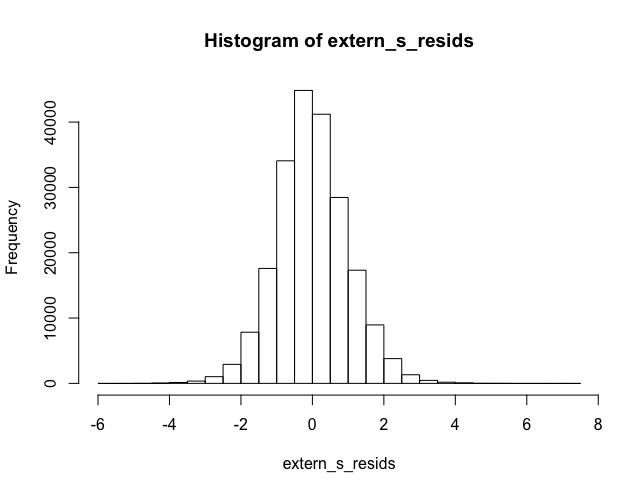
Ha : At least one of ≠ 0

As a result, the null hypothesis was rejected with p-value **<** 2.2e-16. Therefore, three variables (cut, color, and clarity) have been added to the final model. That is, the final model is:

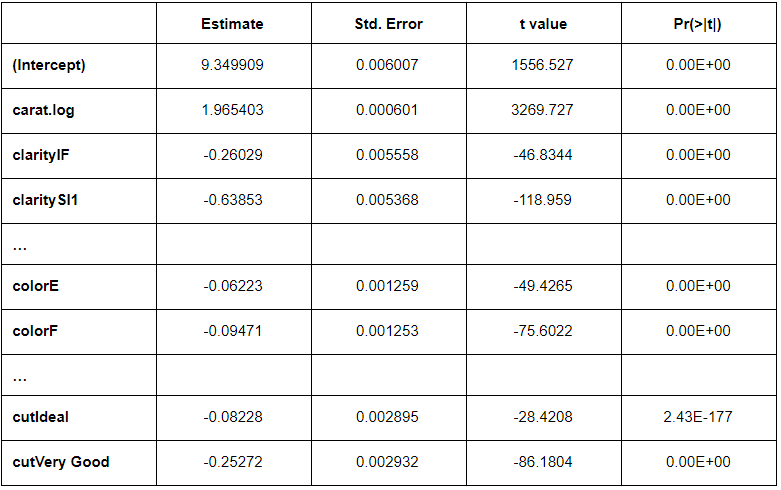
For the final model, the violations of linearity, homoskedasticity, and normality assumptions were tested again.

As three plots are shown below, all assumptions have been satisfied. 





**Conclusion**

****

*Model interpretation*

The result of our final linear regression model is given in the table above. All coefficients are statistically significant at the alpha level of 0.01. The coefficient of ‘log(carat)’ is positive, which means that the weight of a diamond is positively correlated to the price of the diamond. On the other hand, all other coefficients for categorical explanatory variables are negative. To understand what these numbers mean, one should keep in mind which value in each category was used as the default. In our model, ‘D’, ‘FL’, and ‘Astor Ideal’ were used as defaults for ‘color’, ‘clarity’, and ‘cut’ respectively. Thus, it can be said that the negativity of the coefficients stems from the fact that the defaults are the best values, in each category, that boost up the price of diamonds.

The beta of ‘log(carat)’, 1.96, means that when the weight (in carat) of a diamond increases by 1%, the price goes up by 1.96%. Such an intuitive explanation is possible thanks to the fact that the response variable (price) has also been log-transformed.

*Making predictions*

With the weight of 56 carats and the best ‘cut’, ‘clarity’, and ‘color’, the predicted price for Heart of the Ocean is $28,891,662. Plus, the model predicts the price of Hope diamond which weighs 45.5 carats to be $19,210,518.

Since Heart of the Ocean is fictional and Hope diamond is not for sale, the real market prices cannot be obtained. However, according to the Smithsonian Institute, Hope diamond was reportedly insured for $250 million as of 2009. Thus, our prediction of the price of Hope diamond seems to be far-fetched.

*Limitation*

The weights of both Heart of the Ocean and Hope diamond are far out of the range of your given dataset, 56 and 45.5 carats respectively. This exposes our prediction to a problem called ‘extrapolation’ which means that predictions made to far out of the range of given data are often inaccurate. This is because it is not guaranteed that the relationship witnessed in our analysis will stay the same for data far outside of the scope. However, it does not necessarily mean that our prediction is absurd.

One thing that can be said for sure is that the real market price of the Titanic diamond if there is such a thing, will be highly likely to be much higher than our predicted price $28.89 million thanks to its unique color, scarcity, and the story behind.