Rachelle Soh, Kendall Laflin, Yoojin Jeong, Derek Gingerich

Dr. Sadeghkhani

Statistics 3302

25 April 2022

Price of Diamonds Predicted by Regression

**Introduction of the Dataset**

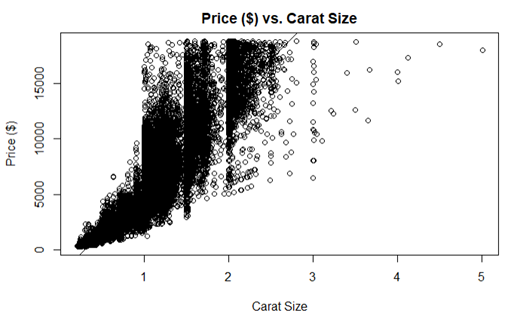
The dataset that our team chose is the “Diamonds” from Kaggle. Dataset contains price, carat, cut, color, and other attributes analyzing 54,000 diamonds. The price of the diamond is stored in price attribute, weight of the diamond is stored in carat attribute. The cut attribute describes the cut quality of the diamond categorized into Fair, Good, Very Good, Premium, Ideal. The color of diamond is stored in the color attribute, J (worst) to D (best). Clarity of diamond describes how clear the diamond is; (I1 (worst), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (best)). Attributes x, y, and z each describe length, width, and depth of diamonds and these are used to calculate the depth attribute, which is a total depth percentage. Lastly, the table attribute describes the width of top of diamond relative to widest point.

**Why We Chose this Dataset**

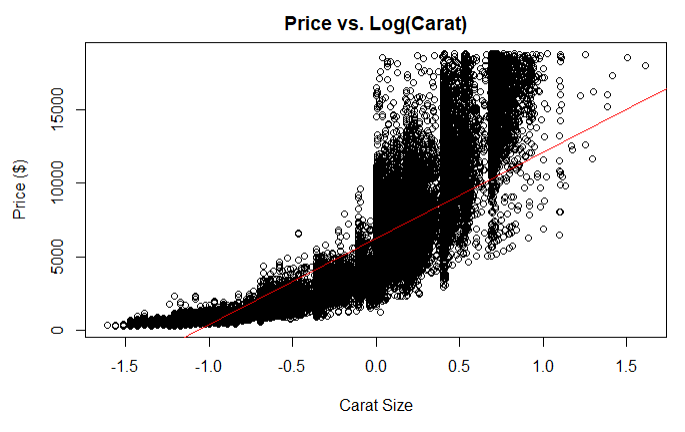
Our team selected the “Diamonds” dataset from Kaggle because we had an interest in learning how various attributes such as clarity, cut and price affected the price of the diamond. Most importantly, we felt that this was a dataset where various forms of regressions and statistical analysis can be applied to. With variables such as carat size and price, there are several variables where the user can logically predict and implement a regression model to fit these variables. There was not much domain-specific knowledge needed to understand the dataset, as compared to, for example, various types of radiation used on various patients. It is rather imperative that a larger carat size of diamond would correspond to higher price, so our team chose to focus on that as the beginning of our analysis, and we analyzed different aspects of the dataset from that starting point. When selecting a dataset, we wanted to push ourselves and not perform a statistical analysis on a similar dataset like the ones that we have learned about in class.

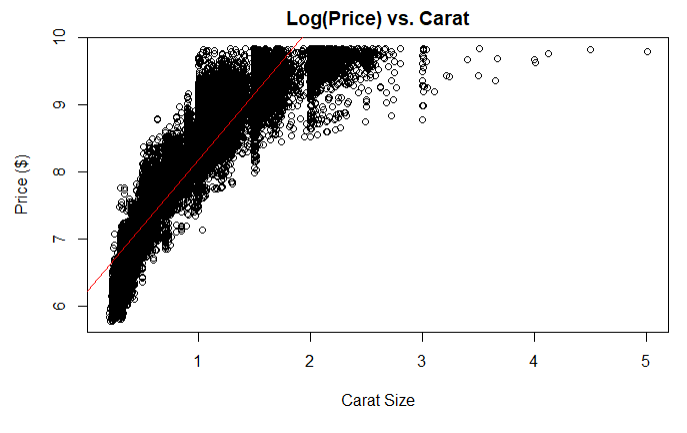
**Linear Regression**

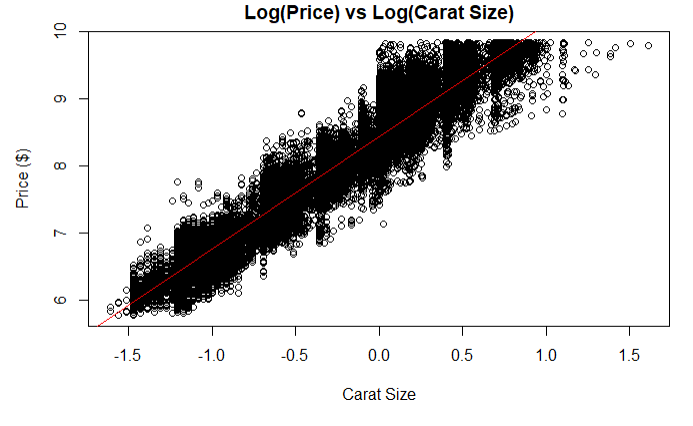
To begin our analysis, our team chose to fit the two variables, carat size and price, into a linear regression model to give ourselves a frame of reference for the dataset selected, with carat [size] being the predictor variable and price being the response variable. We utilized the function “lm” to craft the linear regression model, lm( price ~ carat, data = diamonds). From there, we witnessed that B0 was -2256.36 and B1 was 7756.43. The value -2256.36 does not have a natural interpretation, as the response variable is price, and a natural interpretation of price would mean that price cannot be negative. However, the B1 is a point estimate that gives insight to the increase in price as the carat size increases. As the carat size increases by 1, the estimated price of the diamond is expected to increase by $7756.43. In addition, when running the summary function of the linear model, the reported “Adjusted R-squared” value is 0.88493, which means that taking the square root of this value, r is equal to 0.9216.

This means there is a strong positive correlation between price and carat size. However, as shown in the graph, there seems to be a non-linear pattern between the Carat Size and the Price in dollars when we fit these variables to a linear regression model. As the report continues, we found better ways to establish a data relationship between the data provided in the dataset.

**Log Predictor and Response**

The first logical step our team took after fitting carat size and price into a linear regression model, we decided to experiment with taking the log of predictor and the response variable. First, taking the log of the predictor variable carat size was opted, and the corresponding graph appears to look like an exponential distribution rather than a linear spread of the data.

Next, we took the log of the response variable of price and the corresponding graph indicated to us that taking the log of the response variable would not be appropriate, as the data did not look linear, and the fitted linear regression line was vertical.

Lastly, when we took the log of both the predictor and response variable, we saw that the data appeared to be linear, and the data was fitted around the linear regression line. We concluded that taking both log of the predictor and response variable was needed to fit the linear regression model for both the carat size and price variable. Lastly, looking at the summary of the log predictor and response linear regression model (called by using the function summary on the model “lm(log(price) ~ log(carat), data = diamonds). The B0 was 8.449 and the B1 was 1.676. The B0 does not have a natural interpretation because the B0 takes into account the intercept or starting value. A diamond cannot have a carat size of zero. But in regard to the B1 value, it means for every additional increase in log carat size of the diamond, the price of the diamond is expected to increase by e1.676 = 5.34 thousand dollars. One thing of note is the R-value, calculated by taking the square root of the adjusted R-squared value of 0.933, is 0.966. Compared to the linear regression model’s R-value of 0.9216, there is an even stronger, positive relationship between log price and carat size.

**Regression with “Cut” as Factor Variable**

Even though we found a model to fit the data well using log of predictor and response in the last section, this regression does not fully demonstrate our understanding of the different types of models from this course’s content. The next direction we went with the analysis—in order to demonstrate this understanding and utilize more of the variables available to us in the dataset—was to add the diamonds’ cut quality to the regression as a factor variable. The different categories of the cut include “Fair” (the reference category), “Good”, “Very Good”, “Premium”, and “Ideal”. Using the glm() function in R, we summarized the model with log of carat size and cut quality to predict the log-price of a diamond. The first thing that stands out about the results of this model is that the p-values for each predictor read “<2e-16”. These values are approximately 0, telling us that they are each very significant variables in predicting the price of diamonds.

Looking more specifically at some of the coefficient estimates in this summary, the estimated value for “log-carat” equals 1.696. Because we are using the log of the variables, we take e raised to each coefficient in order to interpret them. For each additional carat added to a diamond, the price increases by $5.45, if all other variables are held constant. We can also do this for each cut quality coefficient. A diamond with a "Good" cut will cost 18% (e^0.1632 = 1.1773) more than a diamond of the same carat size with a "Fair" cut. A diamond with a "Very Good" cut will cost 27% (e^0.2408 = 1.272) more than a diamond of the same carat size with a "Fair" cut. A diamond with a "Premium" cut will cost 27% (e^0.2382 = 1.269) more than a diamond of the same carat size with a "Fair" cut. A diamond with an "Ideal" cut will cost 37% (e^0.3172 = 1.373) more than a diamond of the same carat size with a "Fair" cut. A diamond with a "Premium" cut will be worth 27% (e^0.2382 = 1.269) more than a diamond of the same carat size with a "Fair" cut.

**Multivariate multiple regression**

After fitting linear regression with response variable as price and predictor variable as carat size, we wanted to try adding one more attribute to the predictor variable. For model 1, we added depth variable to the prior linear regression to see how the coefficients change. With price being the response variable and carat size, depth being the predictor variable, we used lm() function to fit this model. lm(price ~ carat + depth, data = diamonds). From the summary of this lm() function, B0 was 4045.333, B1 was 7765.141, and B2 was –102.165. B1 is a point estimate implying that the price increases by $7765.141 as the carat size increases by 1. B2 is a point estimate showing that the price decreases as the depth increases, shown through the negative value of –102.165. In other words, as the depth increases by 1, the estimated price of the diamond is expected to decrease by $102.165.

For our second multivariate multiple regression model, we added the variable table to see if having one more variable demonstrates a better fit. From the summary of lm(price ~ carat + depth + table, data = diamonds), point estimate B3 implies that price decreases as the table increases since it is the negative value. As the table increases by 1, the estimated price of the diamond is expected to decrease by $104.473. Comparison between these two models was done by using ‘anova()’ function. From the output, p-value was 2.2e-16, which is very close to 0. This small p-value implies that model 2, adding table variable, improved fit over model 1.

**Proportional odds logistic regression**

A probabilistic approach can be used in addition to a more conventional method. In order to do this, the possible outcomes must be grouped into a manageable number of categories. The categories chosen are $0-4999, $5000-9999, $10000-14999, and $15000+. With four categories, three intercept values are generated. These reflect the probability of reaching $5000, $10000, and $50000, respectively, for a hypothetical 0-carat diamond. The coefficient shows the impact of the variable on the probability of reaching each mark. A chi-square test found a p-value rounded to zero.

With intercept values equal to 8.5409, 12.3064, and 15.8770, respectively, the probabilities for a diamond with zero carats are 2.0\*10-4, 4.5\*10-6, and 1.3\*10-7. The coefficient on the carat variable is 8.012. This indicates an increase of one carat multiplies the odds of reaching each benchmark by e8.012=3018.

The probability of a diamond with the median carat value costing at least $5000 is about 5.1%. The probability of a diamond with the tenth percentile carat value costing at least $5000 is about .2%. The probability of a diamond with the ninetieth percentile costing at least $5000 is about 97.2%. At $10000, the probabilities are .1% for the fiftieth percentile and 44.8% for the ninetieth percentile. A diamond with the ninetieth percentile carat value has a 2.2% chance of costing $15000. The probabilities of a tenth percentile diamond costing $10000 or a fiftieth percentile diamond costing $15000 are miniscule.

**Conclusion**

To wrap things up, we conclude that linear regression is not a good fit for variables carat size and price. Taking both the log predictor and response was appropriate to fit a linear regression model, seen through the resulting scatterplot and R summary output. Adding the factor variable, “Cut”, allowed us to incorporate this semester’s content, as well as get a better analysis of our dataset. Adding more variables to the linear regression and fitting the multivariate multiple regression, we found out that model with three predictor variables (carat, depth, table) improved fit over the model with two predictor variables (carat, depth). A proportional-odds logistic regression is less versatile in its utility but is a viable method for evaluating probabilities.