Predicting the Prices of Diamond using Regression Analysis

Annie Liang

STAT 6509- Theory of Application of Regression, Spring 2017

California University State, East Bay

Abstract

This paper investigates a brief history of diamond and build a regression model to validate the characteristics of diamond it has on the diamond prices. The summary statistics and model are examined using 308 observations collected from National University of Singapore. Estimates from prices, carat, color, clarity and certification bodies show that prices dispersion increases with color, carat, and certification bodies qualities.

1. Introduction

Diamond is a metastable form of carbons that is arranged in a way to form a very stable cubic crystal structure. Such structure makes diamond the hardest natural substance [1]. Diamonds have long been represented as the rite of marriage. People save money for months and months to purchase a diamond to what they would believe a gesture of unbreakable love between betrotheds. To understand the value of diamond, and ultimately the prices of diamond, it helps to first understand the origin of the diamond.

The sentiment of a diamond engagement ring began in 1215 when Pope Innocent III laid down new rules on wedding. He declared a waiting period between betrothal and marriage ceremony using rings to signify couple’s commitment in the interim [2]. Diamonds didn’t become a popular choice of gems until 1980s where miners discovered huge deposits of diamonds in South Africa [3]. These miners formed a group called De Beers Conslidatd Mines, Ltd to control the spread of diamonds and to keep the diamonds supplies scarcer and prices high. What is more, De Beers rolled out a brilliant marketing plan in the late 1930s to praise the value of diamond and penned the slogan “ A Diamond is Forever” that fostered the diamond value in the marriage process. At the end of the century, about 80% of engagement rings contained diamonds [4]. What is more, the total expenditures on diamond rings had raised to roughly $7 billion in the United State in 2012 [5].

Number of literature and statistical analyzes had done on the prices of diamonds using regression [6] and neural networks [7]. Also, independent organization like Rapaport Group (diamonds.net) publishes statistical reports on diamond prices annually that is available for access through paid subscription.

2. Materials and Methods

To explore the prices of diamond, a data set from a paper published at Journal of Statistics Education was used for analysis. The data set was collected by a group of Master of Business Administration (MBA) students from a class at National University of Singapore [6]. 308 observations on rounds diamonds were found in 2000 from an advertainment section in Singapore's Business Times edition. The data set contained values on prices, carat, color, clarity and certification bodies. The prices distribution on this diamond dataset ranged from $638 to $1,6008 in SGD (Singapore's dollar) with mean prices at $5019 SGD. The minimum diamond size was 0.18 carat and maximum size was 1.1 carat. Diamond prices and carat scatterplot was used to better judge if a regression analysis model would fit this dataset. From Figure 2, there seem to be a large stack of data points on the lower and higher ends of plot. A visible linear line between prices and caret was graphed that assumption of linearity was met for the first step of regression analysis.

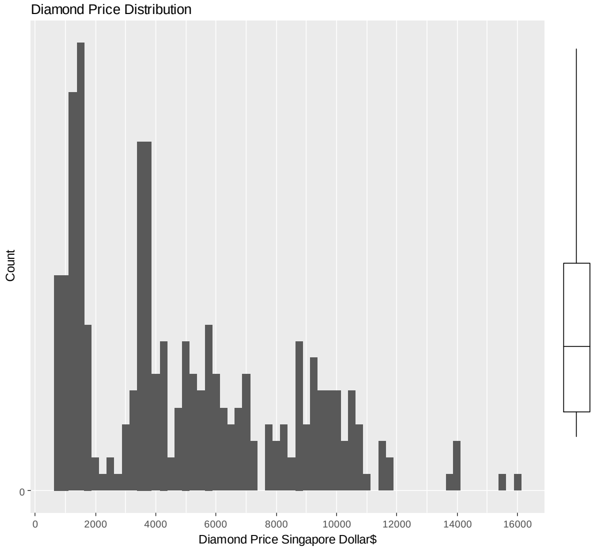

Figure 1. Diamond Prices Distribution

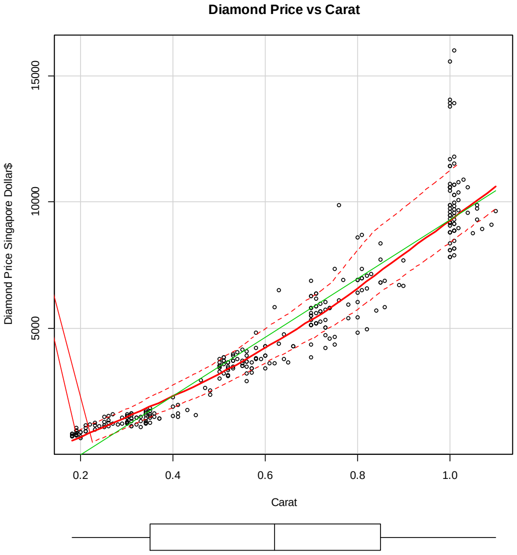

Figure 2. Diamond Prices Vs Carat

Diamond color has a grading scale of D to Z with D representing the highest grade of a diamond. Z is the lowest grade on the scale meaning a diamond is impure. There were 6 levels of color in this dataset, D to I (see figure 3). Grade D had the most wide spread of prices. Comparing the prices means of all color, D had the highest mean since it was the highest grading. The next variable in the data set was clarity. Clarity has a total of 11 point scale but only flawless (FL), IF (Internally Flawless, VVS1/2(Very,Very Slightly Included), and VS1/2 (Very Slightly Included) were available for analysis. Using boxplot, IF, VVS1 and VVS2 are skewed right whereas VS1 and VS2 were approximately normal distributed. The plausible exploration for these abnormal spread of IF, VVS1 and VVS was that the prices among these highest grades of clarity varied and there could be interaction relationship between other variables that affected the distribution of the prices. Three certification bodies were provided in this data set. They were Gemological Institute of America (GIA), Antwerp based International Gemology Institute (IGI) and Hoge Raad Voor Diamant (HRD).

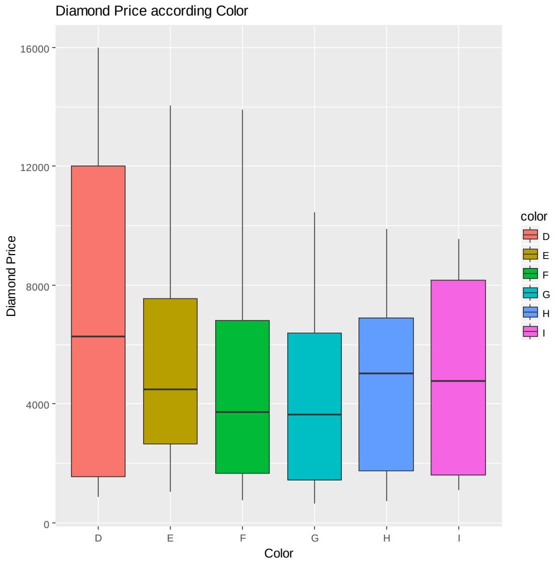

Figure 3. Diamond Prices and Color

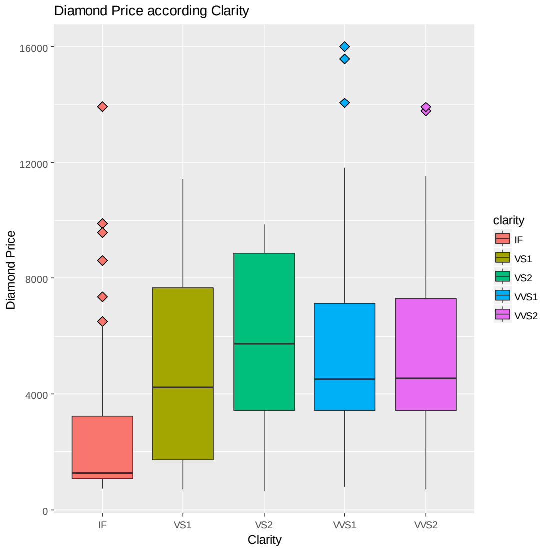

Figure 4. Diamond Prices and Clarity

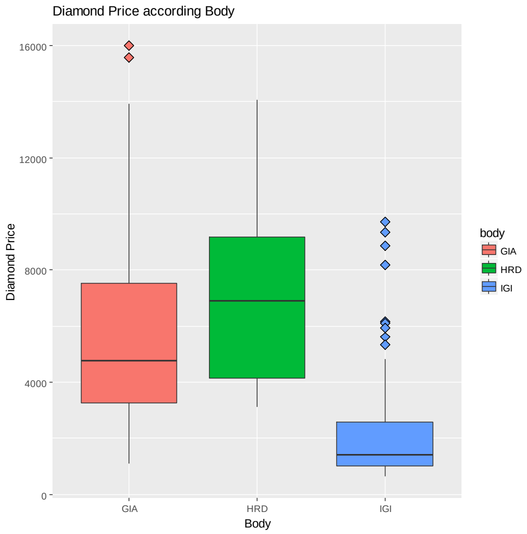

Figure 5. Diamond Price and Certification Bodies

Impact of diamond’s characteristics on its prices was first analyzed using a simply additive model. Prices were used as dependent variable together with carat, color, clarity and certification body as independent variables. All variables were statistically significant except certification body and F-statistics was 562.5 with a very small P-value less than 0.05 significance level. This was a good indicator that there was a relationship between predictor and response variables. Yet, when further analyses were done to test for assumptions of constant variance and normality, the tests had extremely low P-values that had to reject null hypothesis. Therefore, an interaction term was added to the simply additive model for additional model testing, yet a similar finding was found again that assumptions of constant variance and normality couldn’t be met with evidences of small P-values. Since the dependable variable was highly right skewered, a decision was made to use Box Cox power transformation to determine an estimate of λ to transform prices into a normal shape. Squared of prices was used for the transformation.

To select the best model using the transformed data, stepwise selection process was used to find a model with the most reduced AIC. As a result, prices with carat, color, clarity, body and interaction terms with carat and color, carat and clarity was found to be a better model but failed at the non-constant variance score and Shapriro-Wilk normality tests once more.

Since lack of non-constant variance and normality were a persisted problem, outliner test was used to test for extreme outliners in the data set and removed to yield the best model for this data set. The non-constant variance score test for this model was passed with P-value of 0.0736 and Shapiro-Wilk test was also passed with a P-value of 0.1303. The residual plot showed the points were randomly dispersed around the horizontal axis (figure 6).

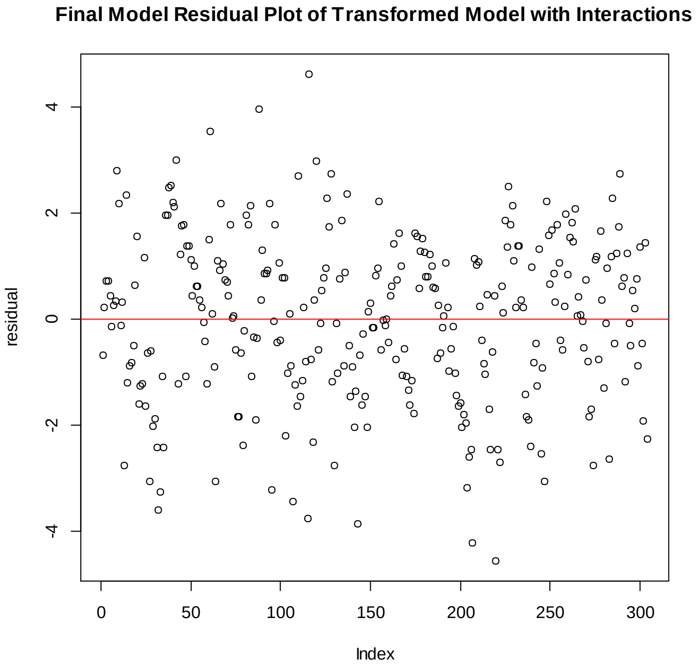

Figure 6. Residual Plot of Transformed Model with Interactions

3. Results

The final model for the diamond data set was found to be prices with carat, color, clarity, body and interaction terms with carat and color, carat and clarity using the transformed data on prices and outliners removed. The model yielded a strong adjusted R-squared at 0.9955 that almost 100% of variance found in the response variable can be explained by the predictor variables. The actual prices of diamonds deviated from the true regression line was found to be approximately on average 1.595 SGD. Given that the mean prices of diamond dataset were $5019 SGD and the residual standard error at 1.595, percentage error was calculated to be 0.03178%.

Characteristics that drive high increase in prices of diamond were carat and color. For every one unit increase in carat, price was increased by $12,561.93 SGD and equivalent of $1,256.193 SGD per 0.1 size of carat. Color with the highest grading had better prices. Getting a better certification body also increased the prices of diamond. Among three certification bodies, IGI was the cheapest and HRD had a medium price. Clarity had negative slopes implying that it brought diamond prices down.

4. Discussion

The model gave a good prediction on prices of diamond using its characteristics however few factors are not studied in this analysis. One is the income on prices. People is more willing to purchase goods when their incomes are high. Examples are that people less likely to buy used cars and more likely to buy new cars, or less likely to rent an apartment and more likely to own a home [8]. Ability to purchase suggests that income is important. Therefore, income on effect of pricing in diamonds is unknown for this study but it’s an important factor that one should investigate when studying the pricing of diamonds. Second is industrial diamond price movement. A paper from The Global Diamond Industry Journal suggests that diamond manufacturing will influence market pricing. That means it would be important to take a close look at the amount of diamond production in years to study how prices change per amount produced [9].

References

1. Harlow, G. E., & American Museum of Natural History. (1998). Chapter 1: What is Diamond. In *The Nature of Diamonds* (p. 5, 84). Cambridge, U.K
2. Clayton, J. (2008). Canon Law and Christian Marriage. In *Pope Innocent III and his times* (pp. 29-31). U.S., CA: Kessinger Pub.
3. *Cawley, L. "De Beers Myth: Do People Spend a Month's Salary on a Diamond Engagement Ring?" BBC News Magazine, May 16, 2014. Accessed October 17, 2014. http://www.bbc.com/news/magazine-27371208*
4. *Sullivan, J. C. "How Diamonds Became Forever." The New York Times, May 3, 2013. Accessed July 11, 2014. http://www.nytimes.com/2013/05/05/fashion/weddings/how-americans-learned-to-love-diamonds.html?_r=3&*
5. Chu, S. (2001). Pricing the C's of Diamond Stones. *Journal of Statistics Education*, *9*(2)
6. KUMAR, U. (2005). Comparison of neural networks and regression analysis: A new insight. *Expert Systems with Applications*, *29*(2), 424-430. doi:10.1016/j.eswa.2005.04.034
7. Chapter 3: Demand and Supply. (n.d.). In *Principles of Economics* (Vol. 1, pp. 50-52). Textbook Equity.
8. Ariovich, G. (1985). The Economics of Diamond Price Movements. *The Global Diamond Industry*, 123-136. doi:10.1057/9781137537584_7

Appendix

```
diamonds.url <- "http://www.amstat.org/publications/jse/v9n2/4cdata.txt"

diamonds <- read.table(diamonds.url)

colnames(diamonds)<-c("carat","color","clarity","body","price")

summary(diamonds$price)


#Data Visualization

 library(ggplot2)

ggplot(aes(x = price), data = diamonds) +geom_histogram(binwidth = 250) +
scale_x_continuous(breaks = seq(0, 20000, 2000)) + scale_y_continuous(breaks = seq(0, 15000,
1000)) + labs(title = "Diamond Price Distribution", x = "Diamond Price Singapore Dollar$", y =
"Count")

library(car)
scatterplot(diamonds$price~diamonds$carat, xlab ="Carat",

            ylab="Diamond Price Singapore Dollar$",

            main="Diamond Price vs Carat")


#boxplot with jitter of each variable

ggplot(diamonds, aes(x=clarity, y=price, fill = clarity)) +

geom_boxplot(outlier.size = 3, outlier.color = "black", outlier.shape = 23) +

ggtitle("Diamond Price according Clarity") + xlab("Clarity") +  ylab("Diamond Price ")

ggplot(diamonds, aes(x=body, y=price, fill = body)) +  geom_boxplot(outlier.size = 3,
outlier.color = "black", outlier.shape = 23) + ggtitle("Diamond Price according Body") +
xlab("Body") +  ylab("Diamond Price ")

ggplot(diamonds, aes(x=color, y=price, fill = color)) +  geom_boxplot(outlier.size = 3,
outlier.color = "black", outlier.shape = 23) + ggtitle("Diamond Price according Color") +
xlab("Color") + ylab("Diamond Price ")


#Simple Additive Model with all predictors variables

diamonds_lm1 = lm(price~carat+factor(color)+factor(clarity)+factor(body), data = diamonds)

summary(diamonds_lm1)

residual = resid(diamonds_lm1)
```

```
plot(residual, main = "Residual Plot of Simple Additive Model")

abline(0,0, col="red")

shapiro.test(residual)

ncvTest(diamonds_lm1)

anova(diamonds_lm1)


# simple Additive model with interactions

diamonds_lm2 = lm(price~carat+factor(color)+factor(clarity)+factor(body)+
carat*factor(color)+ carat*factor(clarity) + carat*factor(body),

                data = diamonds)

summary(diamonds_lm2)

residual2 = resid(diamonds_lm2)

plot(residual2, main = "Residual Plot of Simple Additive Model with interactions")

abline(0,0, col="red")

shapiro.test(residual2)

ncvTest(diamonds_lm2)


#boxcox

library(MASS)

boxcox(price~carat+factor(color)+factor(clarity)+factor(body),

      data = diamonds, lambda = seq(.2, .5, length = 10))

#On transformed data

trans = sqrt(diamonds$price)


#using step() to select models

null=lm(trans~  1, data=diamonds)

summary(null)

full = lm(trans~carat+factor(color)+factor(clarity)+ factor(body) + carat*factor(clarity) +
carat*factor(color) , data = diamonds)

summary(full)
```

```
step(null, scope=list(lower=null, upper=full),

    direction="forward")

diamonds_lm3 = lm(trans~carat+factor(color)+factor(clarity)+ factor(body) +
carat*factor(clarity) + carat*factor(color) , data = diamonds)

summary(diamonds_lm3)

plot(residuals(diamonds_lm3))

#Checking for constancy for variance

ncvTest(diamonds_lm3)

#Shapiro Wilk Test

shapiro.test(resid(diamonds_lm3))


#Outliers test

outlierTest(diamonds_lm3)

#Final Model- on outliers removed from transformed data

trans = sqrt(diamonds_remove$price)

diamonds_lm4 = lm(trans~ carat+factor(color)+factor(clarity)+ factor(body) +
carat*factor(clarity) + carat*factor(color) , data = diamonds)

summary(diamonds_lm4)

residual5 = resid(diamonds_lm4)

plot(residual5, main = "Final Model Residual Plot of Transformed Model with Interactions",
ylab = "residual" )

abline(0,0, col="red)


#predictions

predict(diamonds_lm5, data.frame(carat=.18, color = "I", clarity = "VS2", body = "IGI"))

predict(diamonds_lm5, data.frame(carat=1.01, color = "D", clarity = "IF", body = "HRD"))
```