POLITÉCNICA

REPORT:
MULTIPLE LINEAR REGRESSION

Statistical Analysis HW 2.1

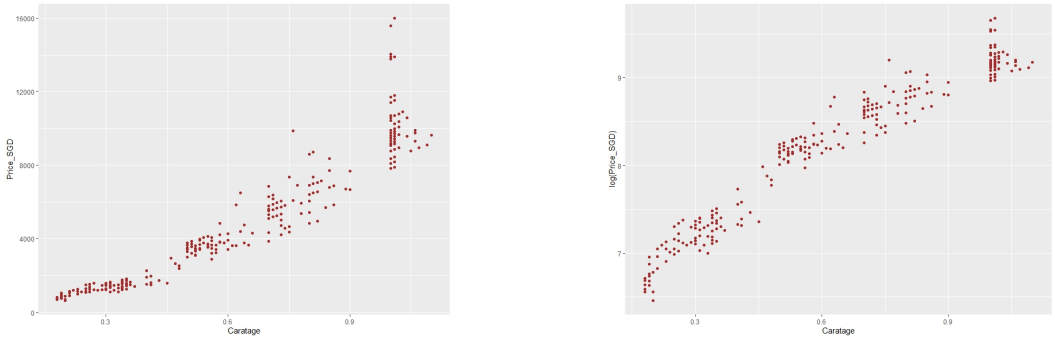
Junhui Liang, Miguel Pérez, Ángel Igareta

February 25, 2021

1 Answers

1.1 Plot Price vs Caratage and log(Price) vs Caratage. Decide on which response variable is better to use.

According to the first exercise, in the figure 1 are represented the plots of caratage against price and log(price).



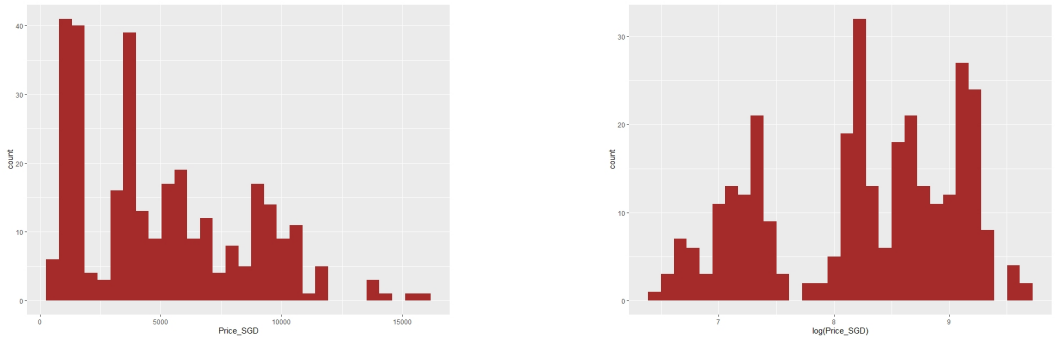
(a) Price vs Caratage (b) Log(Price) vs Caratage

Figure 1: Response Variable Comparision vs Caratage

In the first one, some outliers can be observed when the caratage value is approximately 1.1, with a significative distance in price. Apart from that, in the section where the price and caratage are low, a huge amount of points can be perceived. After plotting the density of the price observations (figure 2a), we can see the reason of the second problem is due the price distribution is right skewed, it does not follow a normal distribution.

Because of the high skewness of the data, the tail region could act as an outlier, affecting many statistical models performance, especially regression-based models [2]. It is necessary to transform the skewed data into a close enough normal distribution. This can be done through a log transformation, resulting in the distribution observed at 2b and 1a.

Hence, we can conclude the better response variable to use in our linear regression model is the transformed price through log operation, from now on refered as *Log_ Price_ SGD*, reducing the number of outliers and improving significatively the model.



(a) Price Distrbution (b) Log(Price) Distrbution

Figure 2: Response Variable Density Comparision

1.2 Find a suitable way to include, besides caratage, the other categorical information available: clarity, color and certificate

1.2.1 Comment on the model fitted.

The first approach was to create a linear model using all the available variables, resulting in a model with the response variable of *Log_ Price_ SGD* and the following explanatory variables: *Caratage*, *ColourPurity*, *Clarity* and *InstCert*.

According to the results of the model's summary, all the variables are significative except the certificate institution GIA dummy variable. Besides, if we plot the linear regression model per each certificate institution, it can be observed that keeping the lines for IGI and HRD, the model with GIA does not explain a high difference in the price amount and according to the summary table is not significative. However, only with this data we can not conclude that the model without InstCert is better, because the other two variables for InstCert are significative.

In order to see if the model with InstCer explains more error than the model without InstCert, we could do an anova test with the two models. The first one, *lm1*, will contain all the variables including InstCert, the second one *lm2* will not include InstCert. The result of this anova test is that the addition of the second model is significative, as it can be seen in the figure 3, so we can conclude to not remove InstCert in the model.

Apart from it, the rest of variables are significant and the model itself. The adjusted R^2 of this model is of 0.966. The next step would be to check the model residuals.

```
Analysis of Variance Table

Model 1: Log_Price_SGD ~ Caratage + ColourPurity + Clarity
Model 2: Log_Price_SGD ~ Caratage + ColourPurity + Clarity + InstCert
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    297 6.6891
2    295 5.6343  2    1.0548 27.613 1.018e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 3: Anova Test

1.2.2 Basic analysis of residuals.

First of all, we will plot the residuals distributed along the fitted value, and maybe via visual inspection we can extract some broad conclusions regarding the residuals.

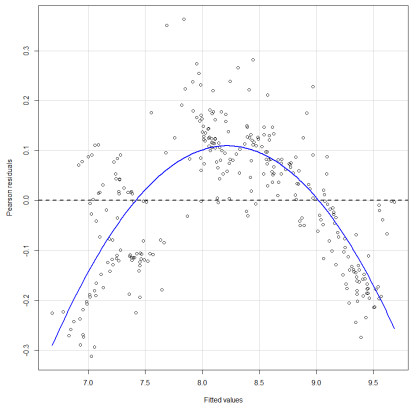


Figure 4: Residual Plot

As we can observe, there is a clear pattern in the residuals instead of a mean 0, which implies that this model is clearly missing something to fit. We have to mention that no outlier with Bonferroni p-value greater or equal than 0.5 was found via outlierTest [1]. The homoscedasticity and normality are harder to interpret visually, so we will perform some test to validate some of the following features:

1. **Independence:**
   To check the independence in the residuals, we will use the Durbin-Watson test. The DW statistic, was 0.31422, meaning the correlation between the residuals was positive (all DW statistic between 0 and 2 implies a positive autocorrelation), while values closer to 2 would have implied no correlation at all, and values ranging from 2 to 4 are from negative correlated residuals.
   The Box test was also applied to obtain a p-value lower than 0.05, implying that dependency exists between residuals

2. **Normality:**
   The normality in the distribution of the model will be compared using the Jarque Bera test. This

test tells us how close to the normal distribution are our data. A normal distribution has Skew 0 (related to symmetric data) and Kurtosis (related to how data is concentrated) equals to three, but in our case, Skewness was 0.15695 and Kurtosis 2.2722, with respective p-values of (0.2608, 0.009125). Global p-value for Jarque Bera test is 0.01775, meaning that the distribution of the residuals is not normal.

3. **Equal Variance:**
The homoscedasticity is asessed via the Breusch-Pagan test. For our linear model, it yielded a p-value of 4.265e-06. For the BP test, the null hypothesis is homoscedasticity, so since the p-value is lower than 0.05, heteroscedasticity is assumed.

1.3 Two different remedial actions:

1.3.1 a) Create a new categorical explanatory variable Size with values Small, Medium, Large

1. **Is this regression model satisfactory? Are the standard assumptions of linear regression validated? Are the numerical estimates sensible?**
The regression model is satisfactory, because the R^2 is 0.9956, meaning that 99.56% of the response variable can be explained using the explanatory variables. Besides, the residual standard error equals 0.0554.

   The institute certificates seem to be non significant to the linear model, since the Pr(>|t|) values are greater than 0.05, so not every estimate is significant. But the p-value of the overall model is 2.2e-16, meaning a great significance of this model. Compared to the original linear model, which had $R^2 = 0.9723$, p-value 2.2e-16 (same value) and Residual standard error of 0.1382, we can certainly say that this is a better model, because of the increase on R-Squared and the decrease on Residual Standard Error.

   Finally we will analyze if the standard assumptions of linear regression models are validated.

   (a) **Linearity:** The linearity can be explain through a scatter-plot of the residuals, which can be observed in the figure 5. Compared to the last example, the residuals in this model does not seem to follow a significance pattern, instead they are close to the mean 0. So we could say the linearity assumption is satisfied.

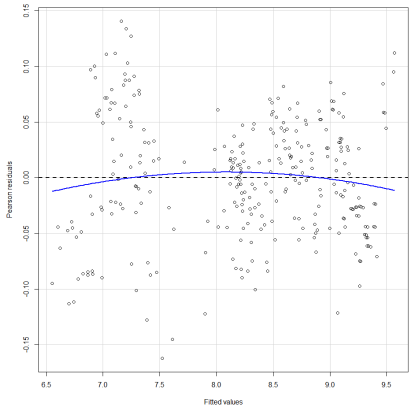


Figure 5: Residual Plot

   (b) **Independence:** In order to analyze independence of the residuals, first we will use the Durbin-Watson test, which results on a DW statistic of almost 1. This means that there is still a positive autocorrelation despite it is much smaller than in the previous model. Besides, Box-Ljung test also provided the same conclusion.

   (c) **Normality:** As in the previous exercise, we perform the Jarque bera test. Skewness was 0.10387 with a p-value of 0.45 and the Kurtosis was 2.7521 with a p-value of 0.37. On top of that, global p-value was 0.5111, which means the normality hypothesis should be rejected.

(d) **Equal Variance:** After performing the Studentized Breusch-Pagan test, we obtained a p-value of 4.117e-07, which means we reject the null hypothesis of homoscedasticity.

2. **Interpret the interaction parameter med*carat. What can we infer on the incremental pricing of caratage in the 3 clusters?**
   According to the summary of the line regression, firstly, the interaction parameter med*carat is vitally significant, contributing to an increment of the R^2 when comparing to the previous model with the same variables. The segmentation of 3 clusters for caratage helps capturing the effect of all the price variation in every interval, which can reduce the unwelcoming effect of outliers, making it more suitable for line regression on each cluster.

   Based on the scatter-plot as figure 6 and coefficient of diverse clusters of caratage, the incremental pricing of caratage is reducing with the higher rank of caratage. In other words, caratage can make the directly huge contribution to the price when the diamond is small, but less contribution when it is enough huge.

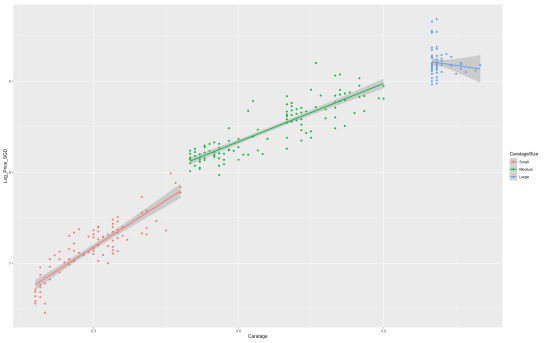


Figure 6: Log(Price) vs Caratage(on various ranks)

3. **Which is more highly valued: colour or clarity?**
   To answer this question, we calculated an average on all the estimates of the dummy variables belonging to those categorical variables, which would mean how each categorical variable in average makes the price increase more (obtain a higher value). For the color purity explanatory variable we obtained a mean of 0.2701673 and for clarity we obtained 0.1817872. Thus, color is more highly valued than clarity on average.

4. **All other things being equal, what is the average price difference between a grade D diamond and another one graded (a) I (b) E?**
   a) The average price difference between D and I would be the estimate of D (0.433562), as I is the reference variable and it is already present in the intercept.
   b) The average price difference between D and E would be the difference between the estimates of D and E which is of 0.08488306.

5. **All other things being equal, are there price differences amongst the stones appraised by the GIA, IGI and HRD?**
   The price difference among the Certificate Institution is very low (~ 0.09). Apart from that, they are not significant, so as we stated before. Hence, it would be better to exclude this variable from the model.

1.3.2 b) Include the square of carat as a new explanatory variable

When we modified the previous model by introducing the square of caratage, the model is satisfactory with the enhancement of the R^2 and the decrease of the residual standard error. With reference to validation of the standard assumption of linear regression, we can check it as the following.

1. **Linearity:** It is feasible to check the linearity of model via the scatter-plot of the residual as figure 7, which is nearly flat and close to the mean 0, indicating the assumption of linearity is satisfied.

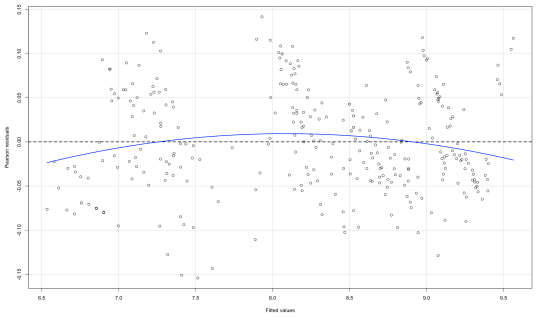


Figure 7: Residual Plot

2. **Independence:** With the attempt to test the independence of the residuals, Durbin-Waston(DW) test is used at first, resulting in a low p-value of 2.2e-16, which can be suggested to reject the null hypothesis of the autocorrelation is 0. Meanwhile, the same conclusion can be reached by Box-Ljung test.

3. **Normality:** Based on the consideration of the normality, Jarque Bera (JB) test is preference. We can come to the conclusion that the probability of the residuals being normal in this model is much lower than the previous one, as skewness was 0.08048 with a p-value of 0.56 and the Kurtosis was 2.4735 with a p-value of 0.59 while the global p-value was 0.143, manifesting there is enough evidence to reject the normality hypothesis.

4. **Equal Variance:** Considering the low p-value of 0.2441 obtained in the Studentized Breusch-Pagan test, it is clear that the model fitted the requirement of constant variance.

1.4 Which of the two remedial actions do you prefer and why? Think on terms of interpretability and validity of the assumptions.

When comparing two remedial actions from assorted aspects, the latter method in terms of validity of the assumptions is better as the equal variance rule is satisfied and the normality probability is much higher than the previous models, but both models are not valid in the independence assumption.

Besides, in respect of the interpretability, the second model without interaction term is much interpretable, possessing fewer and simpler explanatory variables. On top of that, it could be explained as the price increases exponentially with the square of caratage.

References

[1] QUALTRICS.COM. Interpreting Residual Plots to Improve Your Regression. *https://www.qualtrics.com/support/stats-iq/analyses/regression-guides/interpreting-residual-plots-improve-regression/*.

[2] SCIENCE, R. S. T. D. Skewed Data: A problem to your statistical model. *https://towardsdatascience.com/skewed-data-a-problem-to-your-statistical-model-9a6b5bb74e37*.