Homework 2

A pricing model for diamond stones

Statistical Data Analysis - 8th of December, 2019

Group 5

Antoñanzas Martínez, Guillermo
Pueblas Núñez, Rodrigo
Martínez Capella, Ignacio
Burrell, David

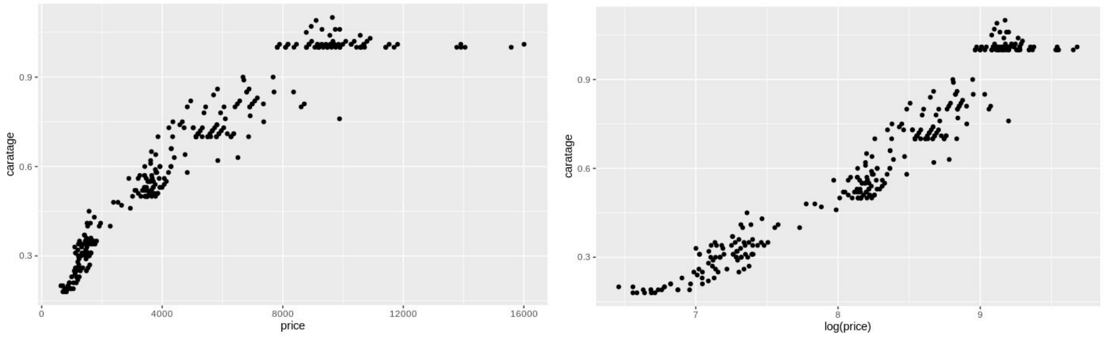
Brief data description, preparation and manipulation

The data provided consists on a dataset of diamond pricing containing 308 observations listed in an advertisement, with information related to caratage, purity, clarity, institution certificate and price. Following the instructions given, we first named all the variables. Then, ordered the purity and clarity categorical variables according to the given criterion (D, E, F, G, H, I for purity, and VS2, VS1, VVS2, VVS1, IF for clarity). No further preparation, manipulation or cleaning was needed before addressing the questions.

Research questions, plots and findings

Question 1

Plot price vs caratage and log(price) vs caratage. Decide on which response variable is better to use.



The above graphs show the caratage plotted against the price (left) and the caratage plotted against the logarithmic transformation of the price (right). It can be seen that the logarithmic transformation improves the normality of the data. Due to the linear pattern between costs and the logarithmic transformation, log(price) is the prefered response variable.

Question 2

Find a suitable way to include, besides caratage, the other categorical information available: clarity, color and certificate. Use the worst level of each categorical variable as the reference category and HRD for certification institution. Comment on the model fitted and perform a basic analysis of the residuals.

Our first approach, Model 1.1, was to try to fit all the explanatory variables in a linear model with the sum of all of them:

$$price = caratage + purity + clarity + certificate$$

With $R^2_a = 0.9564$ and $F = 562.5$, it can be seen that this initial model produces a decent estimate for the price with the selected predictor variables. Though as the logarithmic transformation of the price is not yet in use, there is a residual standard error of 710.4. By incorporating this logarithmic transformation we create Model 1.2:

$$log\ (price)\ \ = caratage + purity + clarity + certificate$$

The improvement of making this change can be seen in the values $R^2_a = 0.9712$ and $F = 863.6$. Thought the most drastic change is in the standard error of the residuals with a drop to 0.1382. For the next iteration of the model the reference values for the categorical variables to the “worst” values. This is I for purity, VS2 for clarity and HRD for certificate institution. This creates Model 1.3, in which, as expected, the statistical measurements do not change, as only the order used by R for the explanatory variables has been altered; the model is essentially the same as model 1.2.

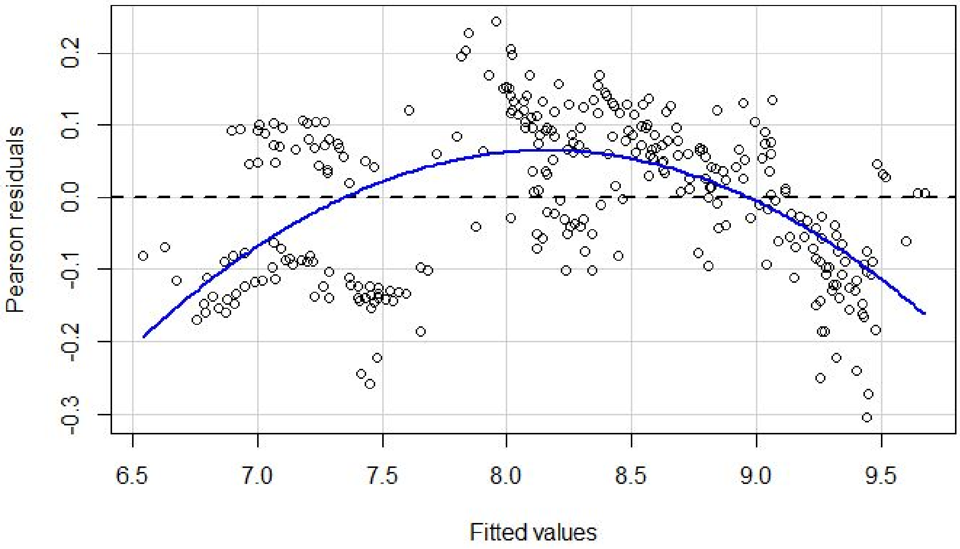
How could we further improve our model? For our model 1.4, we will see the outcome of adding all the interactions among the predictor variables, to see the impact of each of them in our results:

$$log\ (price)\ =\ .\hat{}2\ .$$

This improves even more our results, but there is a danger the model is overfitting the data. This could be occurring due to the fact that our model is too specific for our dataset, and thus unable to better generalize to unseen data. Also, the resulting model has a large number of variables, making it hard to interpret. So to reduce the number of predictors we will only consider the interactions with higher influence, or higher p-value: caratage against certificate institution. This leads to the final model 1.5:

$$log\ (price)\ = caratage + purity + clarity + certificate + caratage : certificate\ .$$

With this model, we achieve R^2_a = *0.9813* and *F = 1154*, with a residual standard error of 0.1113. All the components of our model have a p-value lower than 2^{-5}. Now that a suitable model has been created it needs to be confirmed it follows the LINE assumptions. This can be done by performing some tests. First, we plot the residuals against fitted values:



By inspecting this plot we see that the residuals are not following two LINE properties: Linearity, as there is a trend along the Y axis with an inverted U shape, and equal variances, as the residuals in the left of the plot are more spread than the ones in the right. Apart from that, after performing several tests:

Test Name	Result Value	P-value
Jarque-Bera	11.785	0.00276

Durbin-Watson	0.43495	$2.2e^{-16}$
Breusch-Pagan	84.825	$3.568e^{-12}$

The Jarque-Bera test determines the normality of the residuals, the Durbin-Watson for independence of the residuals and the Breusch-Pagan test for heteroscedasticity of the residuals. As can be seen in the table all of the tests reject their respective null hypothesis so, in addition to the former analysis, we can say that our model is built over failed assumptions for linear regression.

Question 3

Question 3a

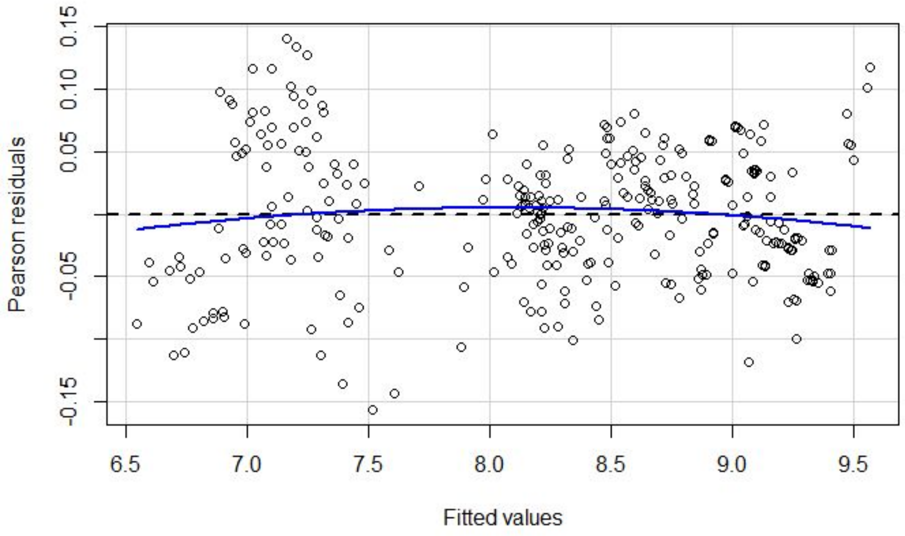
Create a new categorical variable to segregate the stones according to caratage: let's say less than 0.5 carats (small), less than 1 carat (medium) and 1 carat and over (large). Make small as the reference category. Add this new variable to the existing model as well as an interaction term between this new variable and caratage.

After creating the new categorical variable, caratage_category, it was incorporated into the existing model to create Model 1.5.1:

$$log\,(price) \;= purity + clarity + certificate + caratage : certificate + \; caratage * caratage\_category$$

Looking at this new model it can be said it is satisfactory. All the p-values for the included predictor variables except for certificate are low, well below the normal 0.05 threshold, while the model as a who has a $R^2_a = 0.9957$. Comparing it with the previous model, we can see a clear improvement.

To check the validation of the LINE assumptions, we will use the same set of tests performed in the previous section. First, we see a much improved plot of residuals against fitted values, resulting in a more random point distribution:



Then after completing the tests the following values are calculated:

Test Name	Result Value	P-value
Jarque-Bera	0.90364	0.6365
Durbin-Watson	1.0615	$2.2e^{-16}$
Breusch-Pagan	63.085	$6.431e^{-7}$

These results show that not all of the assumptions are met. The Durbin-Watson test for independence rejects the null hypothesis, which is due to the lack of correlation between the residuals. The Breusch-Pagan test also rejects the null hypothesis, thus the residuals do not have equal variances. These results lead to the conclusion that the model meets two of the four LINE properties: linearity and normality. Though even with the model not meeting all of the LINE properties it is able to closely predict the values of the given data.

When looking at the effect on price by caratage_category it can be seen an increment of one unit in any of the categories would result in an increase in the dependent variable. The amount depends on the category for small (S): for medium (M) it is S+0.84 and for large (L), S+2.18.

To investigate if either color or clarity has the higher influence on the dependent variable we can perform an ANOVA test on Model 1.5.1. As seen in the results, (which can be viewed in the accompanying R script) although both perform well in the F-test, color (purity) describes more of the variance within its Sum of Square value, 3.294, compared to clarity which has a value of 1.706. Further evidence of purity being more highly valued in the model is the mean value of its categories coefficients is 0.2411, versus the categories in clarity having a mean of 1.1575. So with this evidence purity is the better predictor variable.

When comparing the average price between the grades it can be seen that on a D grade diamond is on average has a price 0.294240 units higher than an I graded one. This is to be expected a D is the highest grading while I is the lowest. Though interestingly when D is compared to the E grade it is the later that has higher price, 0.046429 units on average.

When looking at the different certificate institutions there is no obvious difference in price from each institution. Both the coefficient and the p-values show that these categories do not contribute significantly to the model. P-values are much higher than 0.05, and the coefficients are almost zero. This lack of difference may be due to all the institutions using the same criteria to evaluate the diamonds.

Question 3b

Include the square of carat as a new explanatory variable. It avoids the subjectivity of clusters definition.

$$log\ (price) = purity + clarity + certificate + caratage : certificate + caratage : purity + caratage : clarity + caratage^2$$

This model gives very similar results than the one proposed on 3a. We achieve $R^2_a = 0.9959$ and $F = 2831$. After performing the same tests that were done in the previous model, we reach the same conclusions about the LINE properties of the residuals. Full results can be checked in the attached script document.

Question 4

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

By taking the first action more variables are introduced into the model and they also are more subjective (small, medium and large thresholds being randomly chosen). This model has an $R^2_a = 0.9954$ and a residual standard error of 0.05 Following the second action on the model, seems a bit far fetched to understand why it would make the model better, but it seems as good as the other option without having to "simulate" the categorizations of diamonds. This model has a $R^2_a = 0.9955$ and a residual standard error 0.05513.

Taking this into account, we prefer the first approach. Both models perform almost the same in terms of R_a^2 and residual error. However, even if it adds more variables (which is not ideal), the first approach provides an easier interpretability of the results.