Stockholms universitet

Pricing of Diamonds - A Study with Multiple Linear Regression

Dana Malas

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2021:2
Matematisk statistik
Januari 2021

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Mathematical Statistics
Stockholm University
Bachelor Thesis **2021:2**
http://www.math.su.se

Pricing of Diamonds - A Study with Multiple Linear Regression

Dana Malas*

January 2021

Abstract

In this Bachelor thesis the selling price of diamonds is empirically examined. We use a sample of 308 certified diamonds collected from brilliance.com in July 2001. The models that will be analysed are linear models and the method that is used is multiple linear regression. We find that the relationship between the price of diamonds and the explanatory variables is actually not linear, but is better explained when transforming it to a quadratic model. We also find out that the price of diamonds increases markedly with the carat weight. What is also interesting is that a diamonds certificate is proved to be a significant factor for the price, even though it is in theory believed that the price is independent of which certificate the diamond has.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: dana.malas@hotmail.com. Supervisor: Taras Bodnar and Pieter Trapman.

Contents

1 Introduction

Diamonds are undoubtedly one of the most well-recognized materials. They have since long been demanded assets and the use of a diamond is deeply ingrained in many cultures. It has many different attributes and works as a symbol of wealth and love. The crystal structure of diamond is what gives it its unusual physical and chemical properties. It is for example the hardest known naturally occurring mineral. The word "diamond" origins from the greek word adamas, which means invincible. (R. Tappert, M. C. Tappert, 2011, p. 1 [1])

During the 18th century, the first real diamond rush started in Brazil. But by the end of 19th century, South Africa made some great findings and took over the leading role from Brazil. Within the coming years, there would be tens of thousands of diamond hunters travelling to South Africa. Most of them left the country empty handed but a few of them became vastly wealthy. Two of them were Barney Barnato and Cecil Rhodes, which in year 1888 founded the firm De Beers Consolidated Mines Ltd. During the latest turn of century, this company controlled approximately 90% of the world's diamond production. (F. Schimanski, 2008 [2])

In this thesis, we are going to analyse how the crucial factors of diamond stones affect the price of it. The diamonds that are going to be used are from brilliance.com, where they guarantee that diamonds are not artificially made. The goal is to create a model that will in the best way explain the price of diamonds as well as give information about which variables that affect the price and how. Such a model could be used in practise, for example when evaluating a diamond. The main question that will be answered in this thesis is:

- How can the price of a diamond be explained in the best way?

The structure of the thesis is organised as follows. The theories of the concepts that will be used are explained in Section 2. In Section 3 the data will be processed, following by the execution of the statistical analysis in Section 4 and a presentation of the results in Section 5. After that, the results, possible sources of error and etcetera will be discussed in Section 6.

2 Theory

In this section the mathematical theory and methods will be analysed and explained.

2.1 Linear Regression

Given the information about the variables and the data set, a natural path to explore the relationships in the analysis is by a multiple linear regression. The definition of a linear regression model is

$$y_i = \sum_{j=0}^{k} x_{ij}\beta_j + e_j, \;\; i = 1, ..., n \tag{1}$$

where y_i is an observation of the dependent stochastic variable y. The value of y_i depends on the variables x_j and the errors e_j. The coefficients β_j are estimated from the equation, which can be written in matrix form as:

$$Y = X\beta + e \tag{2}$$

where

$$X = \begin{pmatrix} 1 & x_{11} & \dots & x_{1k} \\ 1 & x_{21} & \dots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{nk} \end{pmatrix}, \beta = \begin{pmatrix} \beta_0 \\ . \\ . \\ . \\ \beta_k \end{pmatrix}, e = \begin{pmatrix} e_1 \\ . \\ . \\ . \\ e_n \end{pmatrix}, Y = \begin{pmatrix} y_1 \\ . \\ . \\ . \\ y_n \end{pmatrix}$$

where n is the number of observations and k is the number of explanatory variables. (H. Lang, 2015 [3])

2.1.1 Parameter Estimation

In this section the point estimation theory will be explained.

- **Least Squares Method**

  To estimate the parameters we can use the least squares method, which is the most commonly used method in many mathematical fields. It was Carl Friedrich Gauss, a German mathematician who developed the basis for the least squares analysis by the end of the 18th century. The interpretation of the method is to fit a curve to points in the plane, corresponding to determined values $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$. The idea is to choose the curve that best fits the data. Since the relationship

between x and y is *approximately* linear, this approximate relationship is modeled through an error term e_j and takes the following form

$$y_j = \alpha + \beta x_j + e_j \tag{3}$$

where $j = 1, 2, ..., n$ and n is the number of data points. e_j is the difference between the data point y_j and the corresponding point on the straight line $\alpha + \beta x_j$. Obviously, the smaller the difference the better the fit. To minimize this, the least squares method uses the sum of squared errors

$$\sum_{j=1}^{n} e_j^2 = \sum_{j=1}^{n} (y_j - \alpha - \beta x_j)^2 \tag{4}$$

as an overall distance between observed data points and the fitted line. To determine the parameters α and β that minimize Equation 4, we differentiate with respect to α respectively β and set the derivatives equal to zero. (P. Andersson, K. Lindensjo, J. Tyrcha, 2019, p. 17 [4])

- **Maximum Likelihood Method Under the Normal-Theory Assumptions**

  Consider the linear regression model from Equation 2. Suppose that the errors in this model are normally and independently distributed with mean zero and constant variance σ^2. Then the observations Y are normally and independently distributed with mean $X\beta$ and variance σI. The likelihood function is found from the joint probability distribution of the observations. For the linear regression model the likelihood function is

$$L(y|\beta, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{(-1/2\sigma^2)(y-X\beta)'(y-X\beta)} \tag{5}$$

  The maximum likelihood estimators are the values of the parameters β and σ^2 that maximize the likelihood function. Maximizing the likelihood function L is equivalent to maximizing the log-likelihood, $\ln L$. The derivative of the log-likelihood is called the score function. Taking the partial derivatives of the log-likelihood with respect to the parameters β and equating to zero yields

$$\frac{1}{\sigma^2} X'(y - Xb) = 0 \tag{6}$$

  The solution to the score equations is the maximum likelihood estimator (MLE)

$$b = (X'X)^{-2} X'y \tag{7}$$

Notice that the MLE for the normal-theory linear regression model is identical to the ordinary least squares estimator. Maximizing the likelihood function involves minimizing the quantity in the exponent, which is the least squares function. (D. C. Montgomery, R. H. Myers, G. G. Vining, T. J. Robinson, 2010, p. 34 [5])

2.1.2 Hypothesis Testing

We may be interested in seeing whether one, a few or all the explanatory variables have any effect on the response variable. The objective in a hypothesis testing problem is to assess the validity of a claim against a counterclaim using sample data. The two competing claims are called the null and alternative hypothesis, denoted by H_0 respectively H_1 and a hypothesis test is a data-based rule to decide between H_0 and H_1. A test statistic calculated from data is used to make this decision. (P. Anderssion, K. Lindensjo, J. Tyrcha, 2019, p. 42 [4])

- **t-test**

  We are going to test whether the coefficients β_j, $j = 1, ..., k$ are significantly nonzero, with the purpose of examining if the explanatory variable x_j influences the response variable. In other words, we are going to test the null hypothesis $H_0 : \beta_j = 0$ against the alternative hypothesis $H_1 : \beta_j \neq 0$. If the assumptions that the residuals are i.i.d. and normally distributed with $E(e_j) = 0$ and $Var(e_j) = \sigma^2$ are satisfied, it follows that $\hat{\beta}_j$ is normally distributed with $E(\hat{\beta}_j) = \beta$ and $Var(\hat{\beta}_j) = (\sigma^2(X^TX)^{-1})_{jj}$, where $(\sigma^2(X^TX)^{-1})_{jj}$ is the j^{th} diagonal element in the matrix $Var(\hat{\beta}_j)$. Since we want to compare whether two mean values are significantly different, we are going to use the t-test. The t-statistic is given by

  $$T_j = \frac{\hat{\beta}_j}{\sigma\sqrt{(X^TX)^{-1})_{jj}}} \tag{8}$$

  and has a t-distribution with $(n - k)$ degrees of freedom, under the null hypothesis. The null hypothesis is rejected at significance level α if $|T_j| \leq t_{\alpha/2}(n - k)$ where $t_{\alpha/2}(n - k)$ is the critical value which is given by a table for the quantiles of the t-distribution. In this thesis we use $\alpha = 0.05$. (P. Andersson, J. Tyrcha, 2015, p. 48 [6])

- **p-value**

  When determining which variables that are significant for explaining the model, the p-value will be used. The p-value is defined as the probability that $|T_j| \leq t_{\alpha/2}(n - k)$ under the null hypothesis, i.e.

$P(|T_j| \leq t_{\alpha/2}(n-k))$. If the p-value for a variable is less than $\alpha = 0.05$ the coefficient β_j is significantly nonzero and we may reject the null hypothesis. (P. Andersson, J. Tyrcha, 2015, p. 49 [6])

2.1.3 Linearity

In order to use the linear regression model, we first have to examine some important assumptions of the model. One of the assumptions is linearity. This assumption specifies that the functional form of the relationship between the response variable and the explanatory variables is linear. Without linearity between them, parameter estimates will be biased and without any meaning when using the OLS method.
(P. Anderssion, K. Lindensjo, J. Tyrcha, 2019, p. 63 [4])

2.1.4 Residuals

The inspection of residuals is also an important aid in finding out whether a linear regression model is plausible. The residuals are the differences between what is observed and what is explained by the model, they are expressed as $\hat{e}_i = y_i - \hat{y}_i$. The assumption in the linear regression model is that the error terms are independent and normally distributed, with constant variance and $E(e_i) = 0$.

The easiest way to analyse the residuals is to examine graphical plots. Some of the most useful plots are histograms and plotting the residuals against fitted values $\hat{Y}_i$. From the histogram we can learn something about the shape of the probability density function of the error terms. In this case, the density function should be the bell shaped normal distribution. This can also be examined in the Normal Quantile plot where we want an approximate straight line to conclude that the residuals are approximately normally distributed. In the plot against $\hat{Y}_i$, we want it to have the form of a horizontal "band". (P. Anderssion, K. Lindensjo, J. Tyrcha, 2019, p. 67 [4])

- **Homoscedasticity**
  Homoscedasticity occurs when the residuals have constant variance, i.e. $Var(e_i) = \sigma_i^2$. When this is not the case, heteroscedasticity occurs and leads to inconsistent standard deviations of the coefficient estimates, which makes the F-test invalid. To prevent this occurrence, it is appropriate to transform or add more variables to the model. (H. Lang, 2015 [3])
- **Endogeneity**
  Endogeneity occurs when the residuals correlate with one or more variables. This violates the assumption that $E(e_i|x_i) = 0$ which makes the OLS estimator produce inconsistent estimates. (H. Lang, 2015 [3])

2.1.5 Non Multicollinearity

Another assumption for the linear regression model is that the matrix X has full rank. This means that the explanatory variables are not exactly linearly related. When having many potential explanatory variables, it is very common that there exists some kind of linear relationship between them. This is denoted as multicollinearity between the explanatory variables and may cause some problems for the model. If you expected that some variables were going to be statistically significant, but weren't, then this might be because of multicollinearity in the model. If two variables are multicollinear, the results might show that none of them are statistically significant and you may want to exclude both of them from the model. But they can in fact both be very explanatory, however multicollinearity has made their estimates misleading. You should therefore investigate further, and the conclusion is usually that you end up keeping one of them. One useful statistical measure that helps identifying multicollinearity is the Variance Inflation Factor (VIF) that is computed as follows:

$$VIF = \frac{1}{1 - R_j^2}, \tag{9}$$

where R_j^2 is the coefficient of determination explained in Section 2.7. (P. Anderssion, K. Lindensjo, J. Tyrcha, 2019, p. 69 [4]) (R. Sundberg, 2020, p. 92 [7])

2.2 Backward Elimination

The Backward Elimination is a method of fitting regression models where the choice of variables is carried out by an automatic procedure. This procedure starts with the model with all k variables included. One variable at a time is eliminated, until the procedure stops. In every step the hypothesis $\beta_h = 0$ is tested for all the remaining variables x_h. The procedure stops when all remaining parameters β_h are significantly nonzero. If one or more variables are not significant, the variable which gives the smallest decrease in R^2 when excluded will be eliminated. (R. Sundberg, 2020, p. 88 [7])

2.3 Cook's Distance

Cook's distance (after Dennis Cook) is a measure of the influence of an observation on the regression, that measures the effect on $\hat{\beta}$ when excluding that particular observation. It is defined as follows:

$$D_i = (\hat{\beta}_{(i)} - \hat{\beta})^T S(\hat{\beta}_{(i)} - \hat{\beta})/k\hat{\sigma}^2 \tag{10}$$

where $\hat{\beta}_{(i)}$ is the estimate when the i:th observation is excluded from the data set. k is the number of explanatory variables, $\hat{\sigma}^2$ is the estimated variance

and $S = X^T X$. Cook's distance has a threshold value, where observations above this value are seen as very influential for the estimate. Normally, there is a big proportion of observations above this value, and therefore it is reasonable to only control the observations with the highest values. In case we find out that the observation is incorrect or too extreme in any sense, we should investigate whether we want to keep the observation or not. (R. Sundberg, 2020, p. 100 [7])

2.4 Box-Cox Transformation of the Response Variable

Generally, transformations are used for three purposes: stabilizing response variance, making the distribution of the response variable closer to the normal distribution, and improving the fit of the model to the data. Sometimes a transformation will be reasonably effective in simultaneously accomplishing more than one of these objects. We often find that the power family of transformations $y^* = y^\lambda$ is very useful, where λ is the parameter of the transformation to be determined. Box and Cox (1964) have shown how the transformation parameter λ may be estimated. The theory underlying their method uses the method of maximum likelihood, discussed in Section 2.1.1. The computation consists of performing for various values of λ, a standard analysis of variance on

$$y^{(\lambda)} = \begin{cases} \frac{y^\lambda - 1}{\lambda \dot{y}^{\lambda-1}}, & \lambda \neq 0 \\ \dot{y} \ln y, & \lambda = 0 \end{cases} \tag{11}$$

where $\dot{y} = \ln^{-1}\left[(1/n)\sum \ln y\right]$ is the geometric mean of the observations. The maximum likelihood estimate of λ is the value for which the error sum of squares is a minimum. However, there is a problem that arises in y as λ approaches zero, y^λ approaches unity. That is, when $\lambda = 0$, all the response values are a constant. This problem is alleviated by the component $(y^\lambda - 1)/\lambda$ from Equation 11 because as λ tends to zero, $(y^\lambda - 1)/\lambda$ goes to a limit of $\ln y$. Values of λ close to unity would suggest that no transformation is necessary. (D. C. Montgomery, R. H. Myers, G. G. Vining, T. J. Robinson, 2010, p. 43 [5])

2.5 Dummy Variables

Dummy (or indicator) variables are variables that are either 0 or 1. Usually 1 represents the presence of some attribute and 0 its absence. They allow qualitative characteristics to be introduced into the regression model. We create a regressor, u, that takes the values 0 and 1 dependent on which of the two categories that the observation belongs to. For example, a multiple linear model with this dummy variable might look like this:

$$y_i = \alpha + \beta_u u_i + \beta_x x_i + e_i \tag{12}$$

(R. Sundberg, 2020, p. 104 [7])

2.6 Akaike's Information Criterion

Akaike's Information Criterion (AIC) is an index used in a number of areas as an aid to choosing between different competing models. It is defined as follows:

$$-2L_k + 2k, \tag{13}$$

where L_k is the maximized log-likelihood and k is the number of parameters in the model. The index takes into account both the statistical goodness of fit and the number of parameters that have to be estimated to achieve this particular degree of fit, by imposing a penalty for increasing the number of parameters. Lower values of the index indicate the preferred model, that is, the one with the fewest parameters that still provides an adequate fit to the data. (B. S. Everitt, A. Skrondal, 1998 [8])

2.7 R^2 and R^2_{adj}

The most common goodness-of-fit measure associated with linear models in general, and multiple regression models in particular, is the coefficient of determination, denoted R^2. It can be defined as the proportion of the total variation that is explained by the model. The formula of the coefficient of determination is defined as follows:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}, \tag{14}$$

where $SSR = \sum_i (\hat{y}_i - \bar{y})^2$, $SST = \sum_i (y_i - \bar{y})^2$ and $SSE = \sum_i (y_i - \hat{y}_i)^2$. The value of R^2 varies between 0 and 1, where the higher value explains that a model has a better fit to the observations. However, R^2 can not exactly be used to determine if a model is good or bad, rather to compare different models with the same data set. However, R^2 increases when adding variables to the model, even though the variables do not explain the response variable. In that case, it is more appropriate to use the measure R^2_{adj}, that measures how much the variation decreases in the actual model and takes the number of parameters into consideration. The R^2_{adj} is defined as follows:

$$R^2_{adj} = 1 - \frac{\hat{\sigma}^2}{\hat{\sigma}^2_0} \tag{15}$$

where $\hat{\sigma}^2$ is the estimated variance and $\hat{\sigma}^2_0 = \sum_i (y_i - \bar{y})^2/(n-1)$ is the estimated variance when there is no explanatory variable in the model. (R. Sundberg, 2020, p. 86 [7])

3 Data

The data that are used in this analysis are from July 2001, obtained from the archive of the Journal of Statistics Education. It consists of 308 observations of round diamond stones from brilliance.com. The data set is also used in one of 14 projects in the course MT5001 at Stockholm University. However, I have not been in touch with this data set before and no project that contains this data.

3.1 Original data

The dependent variable is the *Price* of diamonds in Singaporean dollars and the independent variables are $Carat$, $Colour$, $Clarity$ and $Certification$.

3.1.1 Carat

The weight of a diamond stone is indicated in terms of carat units. One carat is equivalent to 0.2 grams. The larger the diamond stone is the higher the price, ceteris paribus.

3.1.2 Clarity

The majority of diamonds have inclusions that are only visible for a magnifying glass of a microscope. Diamonds with no inclusion, under a loupe with a 10 power magnification, are labelled IF ("internally flawless"). Diamonds that are not internally flawless, which are not free from inclusion, are categorised in descending order as "very very slightly imperfect" VVS1 or VVS2 and "very slightly imperfect" VS1 and VS2. The lesser visible inclusions under the loupe the higher is the price of the diamond, ceteris paribus.

3.1.3 Colour

The most priced diamonds show colour purity. They are not contaminated with neither yellow nor brown tones. Diamonds are colour graded on a scale from D-Z, where top colour purity is the grade of D, and lesser degrees of colour purity are E, F, G, etc. In this data set the colour grades D, E, F, G, H and I are observed.

3.1.4 Certification

Certification bodies examine diamond stones and provide them with a certificate, where they list caratage, grade of clarity, colour and cut. The certification bodies in this data set are New York based Gemmological Institute of America (GIA), the Antwerp based International Gemmological Institute (IGI) and Hoge Raad Voor Diamant (HRD). They are not ranked

in any sense, however their reputation could be a factor in the pricing or demand of the diamond stones.

3.2 Transformations and Data Processing

Since three out of the four independent variables are categorical, they will be transformed into becoming numerical variables. The variables colour and clarity are ranked, and therefore it is reasonable that some kind of rank is also used when transforming them into numerical. They will be transformed according as follows:

Colour	Value		Clarity	Value
D	1		IF	1
E	0.83		VVS1	0.8
F	0.67		VVS2	0.6
G	0.5		VS1	0.4
H	0.33		VS2	0.2
I	0.17			

Table 1: Transformation of Colour and Clarity

The calculations in Table 1 have been made with the same method. The assumed highest ranked categories of $Colour$ and $Clarity$, D and IF, are given the highest numerical values. Since there are six categories for $Colour$, the grade D gets the value $\frac{6}{6}$. The grade E will get the value $\frac{5}{6}$, the grade F will get the value $\frac{4}{6}$ and so on. This method is also applied to $Clarity$.

When transforming our third categorical variable, $Certification$, to numerical, this method will not be applied since the categories are not ranked. Instead we will be including two dummy variables into our models.

$$\begin{cases} u_i = 1, & \text{if certificate of observation } i \text{ is from GIA.} \\ u_i = 0, & \text{otherwise.} \end{cases}$$

$$\begin{cases} v_i = 1, & \text{if certificate of observation } i \text{ is from HRD.} \\ v_i = 0, & \text{otherwise.} \end{cases}$$

Table 2 shows the amount of observations for the three categories.

In Table 3 and Table 4 we have the domains of the data set before and after transformation. We are going to investigate these variables further in the coming sections.

Category	# of observations
GIA	151
HRD	79
IGI	78

Table 2: Number of observations of Certification

Price	Carat	Colour	Clarity	Certification
$16008	1.10 ct.	D	IF	GIA
.	.	E	VVS1	HRD
.	.	F	VVS2	IGI
.	.	G	VS1	
.	.	H	VS2	
$638	0.18 ct.	I		

Table 3: Original data

Variable	Type	Max. Value	Min. Value
Price	Numerical	16008	638
Carat	Numerical	1.10	0.18
Colour	Numerical	1	0.17
Clarity	Numerical	1	0.2
u_i	Dummy Variable	1	0
v_i	Dummy Variable	1	0

Table 4: Data after transformation

4 Statistical Analysis

In this section, we are going to put our theory into practice.

4.1 Analysis of Individual Variables

We start off by looking at plots of each of the explanatory variables against the response variable. From these plots, we will analyse whether the relationships are linear or not. Since we have transformed three out of the four variables - $Colour$, $Clarity$, and $Certificate$ - these three variables are now discrete and the fourth variable $Carat$ is continuous. Therefore, linearity between those three variables and the response variable will not be as clear to detect as for the continuous one.
In Figure 1 we can observe the relationships between Y and the explanatory x. Figure 2 shows us the residuals against fitted values of each explanatory variables in a simple regression against the response variable.

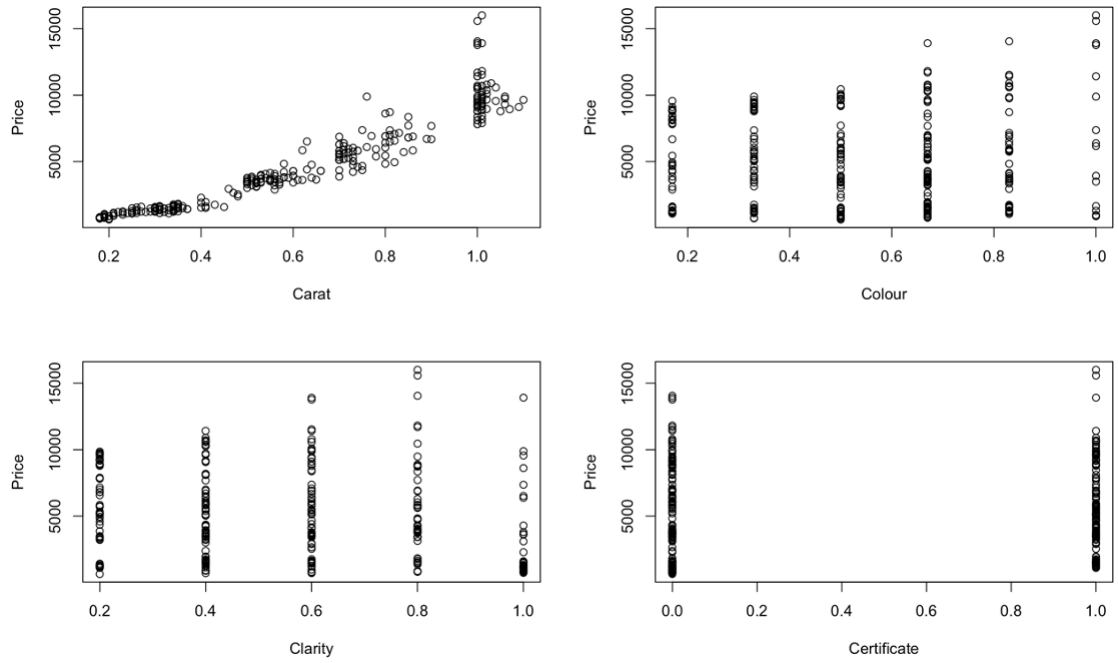

Figure 1: Relationships between Price and explanatory variables

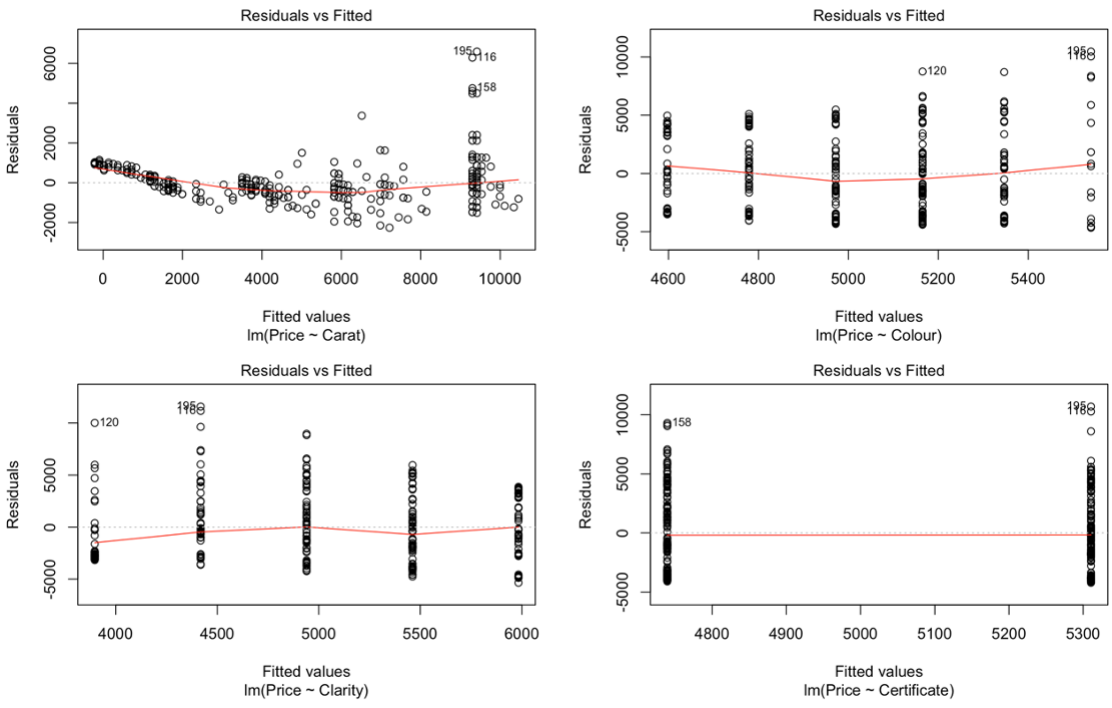

Figure 2: Residuals vs Fitted

As mentioned in 3.1.1, heavier stones are more priced than the lighter ones. This was a visible trend when looking at the first plot from the left in Figure 2, where we can see that the scatters have a somewhat exponential relationship with the price. Without transformation, the data would most likely violate the assumption of homoscedasticity mentioned in 2.1.3. Further investigation will therefore be made to determine which transformation of Y to employ in the analysis.

4.2 Models

In this section, we are going to construct four different models. These models are going to be compared later on to find the one with the best fit to the data. To construct the models, we start off with the basic model

$$y_i = \alpha + \beta_1 Carat_i + \beta_2 Colour_i + \beta_3 Clarity_i + \beta_4 u_i + \beta_5 v_i + e_i.$$

When plotting the residuals for this model, a quadratic trend can be observed. This is shown in Figure 6 in the Appendix. This clearly means that we need a transformation to our basic model. But which transformation is the most suitable? Since we observe a quadratic trend, we would suggest that raising the model to the power of 2 will be helpful. Doing this is the same thing as taking the square root of our response variable Y. We get the following model:

Model 1: $\sqrt{y_i} = \alpha + \beta_1 Carat_i + \beta_2 Colour_i + \beta_3 Clarity_i + \beta_4 u_i + \beta_5 v_i + e_i$

In the next model we are going to take something more into account, namely the observations we found in 4.1, where Figure 1 showed us that $Carat$ does not follow a linear trend. Figure 7 in the Appendix shows us what $Carat$ looks like after transforming it to $Carat^2$. It becomes more linear. Therefore the next model will be as follows:

Model 2: $\sqrt{y_i} = \alpha + \beta_1 Carat_i + \beta_2 Carat_i^2 + \beta_3 Colour_i + \beta_4 Clarity_i + \beta_5 u_i + \beta_6 v_i + e_i$

Instead of just estimating with the eye, there is another method that analyses which the optimal number is for a power transformation on Y, namely the Box-Cox method explained in Section 2.4. When applying this procedure to the basic model, we get $\lambda = 0.42$. This is quite close to our self estimated value 0.5, but since this has more statistic calculations behind it it will be used to construct a third model.

Model 3: $y_i^{0.42} = \alpha + \beta_1 Carat_i + \beta_2 Colour_i + \beta_3 Clarity_i + \beta_4 u_i + \beta_5 v_i + e_i$

In the fourth model the original data will be used, i.e. the one with the categorical variables. The grade IF is selected as the baseline for $Clarity$. The grade D is selected as the baseline for $Colour$ and the $Certification$ denoted by GIA is selected as the baseline for $Certification$. The Box-Cox results are the same as before for this model.

$$\text{Model 4: } y_i^{0.42} = \alpha + \beta_1 Carat_i + \beta_2 E_i + \beta_3 F_i + \beta_4 G_i + \beta_5 H_i + \beta_6 I_i$$

$$+\beta_7 VVS1_i + \beta_8 VVS2_i + \beta_9 VS1_i + \beta_{10} VS2_i + \beta_{11} HRD_i + \beta_{12} IGI_i + e_i$$

4.3 Non Multicollinearity Check

Now the models will be checked if they satisfy the assumption of non multicollinearity. This is done by calculating the VIF of every model.

Variable	Transformed Data	Original Data (df)	Original Data VIF$^{(1/(2df))}$
Carat	1.158288	1.687710 (1)	1.299119
Colour	1.036061	1.165812 (5)	1.015460
Clarity	1.358944	1.736131 (4)	1.071391
Certification	1.210990	2.129348 (2)	1.207985

Table 5: VIF of variables

As can be seen in Table 5, all the variables of the four different models have relatively low VIF values. None of the values exceed 5, which means that the assumption of non multicollinearity is satisfied. The values of the transformed data are in average a bit larger than the values of the original data when taking the degrees of freedom into account.

4.4 Residuals

In Figure 3 the residuals vs fitted plots can be observed for each model. They are all quite similar. What can be observed is a couple of outliers in all of the models. These outliers will be investigated further in Section 4.6.

4.5 Backward Elimination

A stepwise selection is now performed to see if some variable that is not significant could be removed from the models. The results in Table 11 in the Appendix are obtained from the backward elimination. All variables seem to be significant at a significant level of 5%.

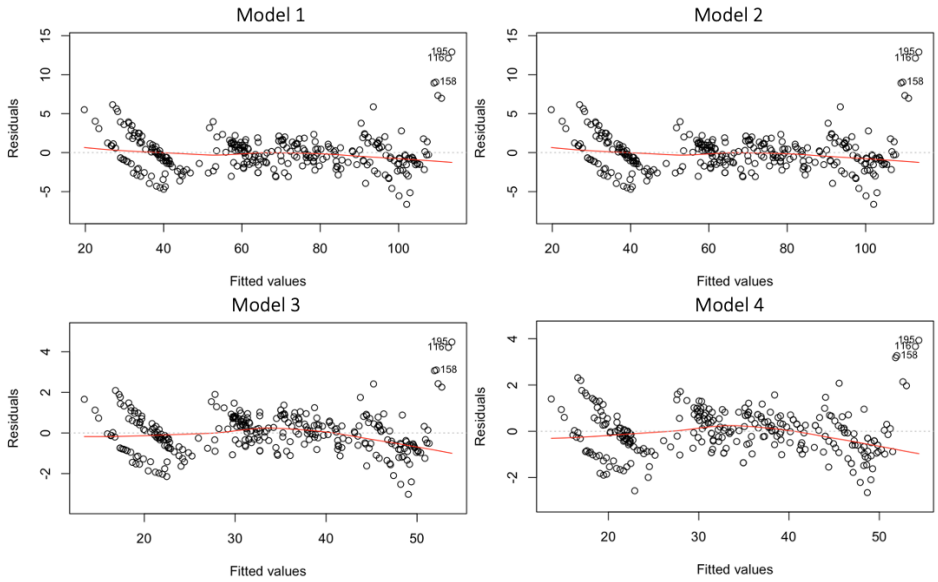

Figure 3: Residuals vs Fitted

4.6 Cook's Distance

In Figure 4, we can see that the most extreme outliers are identified as observations 195, 158, 120 and 116. In Table 6, we take a closer look at them to come to a conclusion of whether we should keep them or not.

Observation	Price	Carat	Colour	Clarity	Certificate
195	16008	1.01	D	VVS1	GIA
158	14051	1.00	E	VVS1	HRD
120	13913	1.00	F	IF	GIA
116	15582	1.00	D	VVS1	GIA

Table 6: Observations from Cook's Distance

As we can see in Table 6, a common thing for the observations is that their price is relatively high. However, we do not note anything that might be wrong with the observations. Therefore, the observations are decided to be kept since there is nothing wrong with the data.

4.7 R^2_{adj}, VIF and Residuals

As mentioned in Section 4.3, the VIF values are very similar for the two data sets, however the average VIF value is a bit lower for the original data than the transformed data, when taking the degrees of freedom into consid-

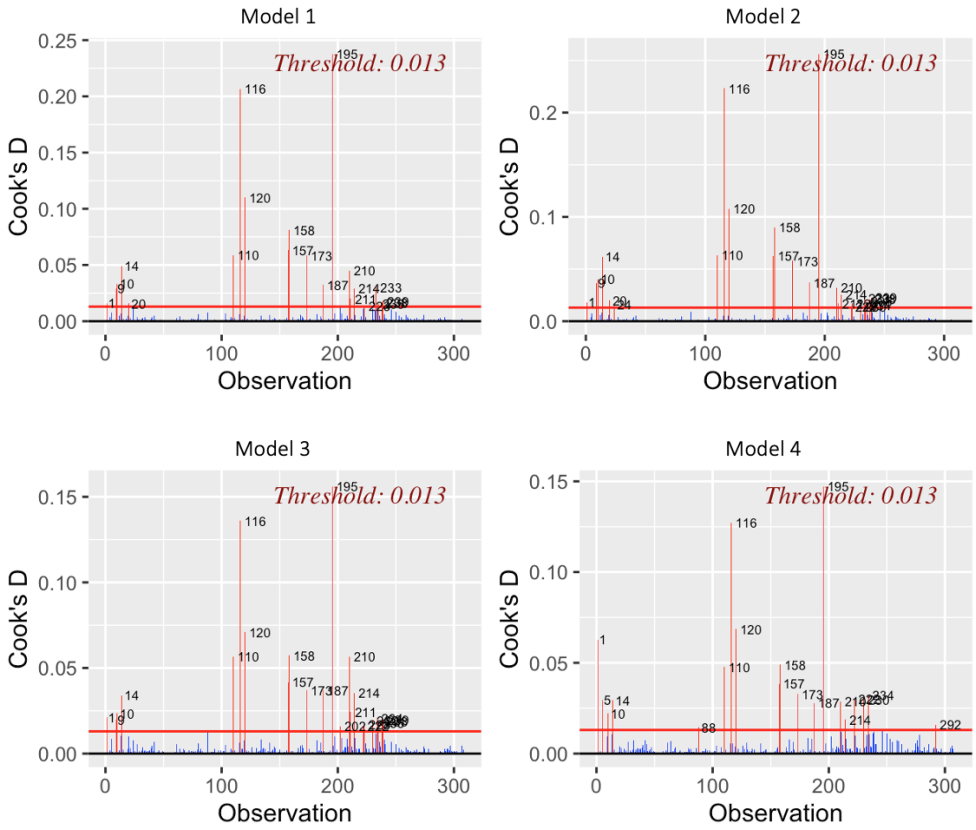

Figure 4: Cook's Distance

eration.

Figure 5 shows the QQ-plots of the different models. It is clear that the standardized residuals of Model 4 have the most linear trend of all models. The remaining models deviate from the line just before point 2 of the theoretical quantiles on the x-axis. In Table 7 we can see that the value of R^2_{adj} is high on all of the models, namely 0.99.
From the VIF and the QQ-plots, it is reasonable to say that Model 4 has the best fit to the data of all four models. This model uses the original data set. From the models with the transformed data, Model 3 has the best fit.

4.8 Model Selection

To determine which model that represents data in the best way, the four models will be compared with each other. The *AIC* values of the four models will be observed in Table 8. In Table 8 we can clearly see that Model 3 and Model 4 have the lowest AIC values. These two models are chosen as

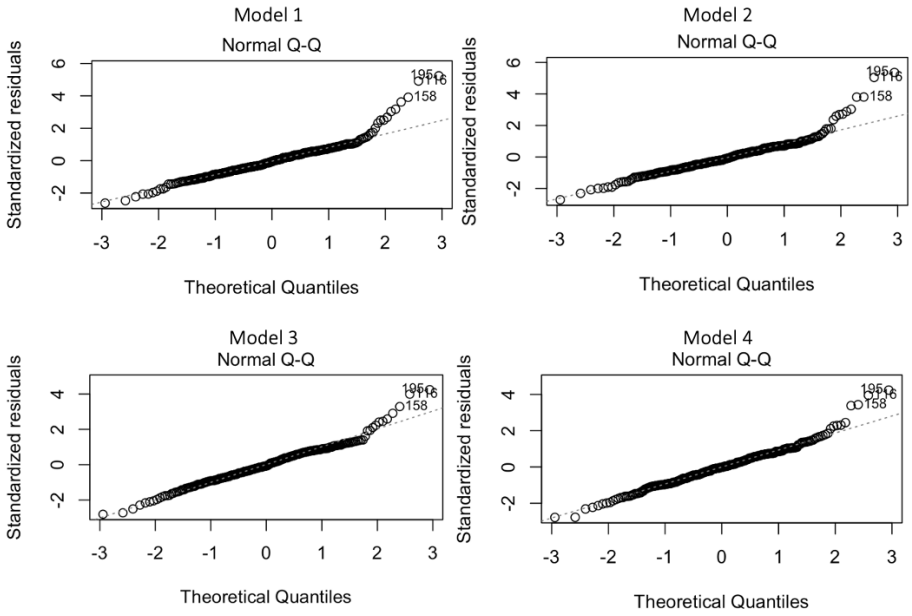

Figure 5: Normal QQ-plots

Model	R^2_{adj}
Model 1	0.9905
Model 2	0.9900
Model 3	0.9913
Model 4	0.9916

Table 7: Adjusted coefficients of determination

Model	AIC
Model 1	1422.638
Model 2	1436.381
Model 3	876.8033
Model 4	874.6925

Table 8: Akaike's Information Criterion

our final models to explain the price of diamonds.

4.9 Parameter Estimates

Table 9 and Table 10 shows the parameter estimates with standard errors of our final models.

Model 3					
Variable	DF	Parameter Estimate	Std. Error	t value	Pr> |t|
Intercept	1	0.2888	0.3132	0.922	0.357
Carat	1	39.2028	0.2592	151.236	<2e-16 ***
Colour	1	8.2091	0.2467	33.277	<2e-16 ***
Clarity	1	5.8892	0.2595	22.696	<2e-16 ***
u_i	1	1.0390	0.1737	5.983	6.18e-09 ***
v_i	1	1.0962	0.1981	5.533	6.83e-08 ***
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					

Table 9: Parameter estimations and std. error of Model 3

Model 4					
Variable	DF	Parameter Estimate	Std. Error	t value	Pr> |t|
Intercept	1	15.724141	0.350861	44.816	<2e-16 ***
Carat	1	39.253360	0.261444	150.141	<2e-16 ***
E	1	-2.051235	0.286150	-7.168	6.12e-12 ***
F	1	-3.173580	0.268608	-11.815	<2e-16 ***
G	1	-4.292499	0.275710	-15.569	<2e-16 ***
H	1	-5.706503	0.279178	-20.440	<2e-16 ***
I	1	-7.473688	0.292662	-25.537	<2e-16 ***
VVS1	1	-0.984697	0.220032	-4.475	1.09e-05 ***
VVS2	1	-2.364316	0.204668	-11.552	<2e-16 ***
VS1	1	-3.488792	0.219688	-15.881	<2e-16 ***
VS2	1	-4.593414	0.235524	-19.503	<2e-16 ***
HRD	1	0.002652	0.147556	0.018	0.986
IGI	1	-1.018152	0.176463	-5.770	2.01e-08 ***
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					

Table 10: Parameter estimations and std. error of Model 4

5 Results

In this section, we are going to interpret the results presented in Section 4.

When looking at the QQ-plots in Figure 5 we can see that the residuals of Model 4 seem to have the best fit to the line. Otherwise, the plots look relatively similar. The related R^2_{adj} of the models also have very similar values. The differences between the R^2_{adj} of the models will not be relevant to draw any conclusions from when comparing the models, since we would for example prefer having lower multicollinearity for the cost of a higher R^2_{adj}.

The VIF values are as well similar and only the variable $Certification$ from the original data set exceeds the value 2, when not taking the degree

of freedom into consideration. Otherwise, all of the VIF-values are below 2, which is relatively low and indicates that the variables are non multicollinear.

When analysing the results of the AIC values for each model, there is a difference. It is clear that Model 4 and Model 3 have the lowest values. Since Model 4 also had the best fit of the QQ-plot, this model is determined to be the model that explain the data in the best way. Model 3 has the lowest AIC value from the models with the transformed data, and therefore we proceed with these two models to compare the parameter estimations.

Let us begin with examining the parameter estimate for $Carat$. The coefficient for $Carat$ has very similar estimates in both models, 39.25 in Model 4 and 39.20 in Model 3. Both of the related t-values are relatively high and the related $\Pr > |t|$ is low. This means that we may reject the null hypothesis that says that the beta estimate equals zero.

The estimate of the variable $Colour$ in Model 3 is 8.21. This means the price of the diamond stone increases when the colour of the stone has a higher rank. The same trend goes for the estimates in Model 4. However, the values of the estimates are negative and becomes *less* negative the higher ranking the colour has. The null hypothesis can be rejected in this case as well. The same reasoning applies for the variable $Clarity$, where the parameter estimate is positive in Model 3, and in Model 4 it becomes *less* negative the higher ranking the Clarity has.

As for the variable $Certification$, Model 3 shows a slightly positive trend for the dummy variables. However, the t-values are relatively low so the null hypothesis can not be rejected. For Model 4 the variable HRD is clearly insignificant and the null hypothesis can not be rejected. However, the certificate body IGI has a significant estimate, which is negative. Since GIA is selected as the baseline, it means that GIA has a higher ranking than IGI and influences the price positive.

The results of the intercepts are very different between the models. In Model 3, the parameter estimate is approximately 0.29 and the null hypothesis can not be rejected. In Model 4, the parameter estimate is 15.72 and the null hypothesis is rejected.

6 Discussion

Apart from $Carat$, the variables $Colour$ and $Clarity$ were earlier believed to affect the price of diamonds positively. This theory was correctly reflected in the models. However, according to brilliance.com, the certification body

of the diamonds should not affect the attributes of the diamonds. However, some of our results show different. In Model 4 the null hypothesis can not be rejected for IGI and HRD. In Model 3 the null hypothesis for β_4 and β_5 can be rejected, but u_i and v_i have approximately the same influence on the price. This means that IGI has a smaller influence on the price than the other two certification bodies. One reason of that could be that, according to diamonds.pro, when comparing the institutes GIA, HRD, and IGI, GIA is the most well-known and respected entity. They grade diamonds against strict guidelines, while IGI and HRD are looser in their grading. HRD is based in Europe and is the leading authority in Europe when it comes to diamond grading. However, since they are less strict than GIA in grading diamonds, that might affect the price of the diamond stone. For that reason, HRD might have the same influence on the price as GIA has, even though GIA is more well known. Another reason for these results might have been the method for collecting the data, which will be discussed further later on.

When pricing diamonds, it is often talked about the "Four C's", which are Carat, Colour, Clarity and Cut. In this thesis, the variable Cut is not examined. The reason for this is simply because this variable is not included in the data set. Diamonds also have different shapes and depths and another property is also which shine it gives when exposed to UV-light. These variables are not used in the study and they might have had an influence on the price of the diamonds. The data used in this thesis are only a small sample of 308 observations among over 80 000 diamonds at brilliance.com. When having a sample size like that, the method of collecting data could be crucial for the results. For that reason, we can not be entirely sure if the parameter estimates actually fit with the total pool of diamonds. A collection method that could have improved the precision of the model even more could have been if all the variables were collected orthogonal to each other, i.e. that for each variable there are equal proportions of the values of the other variables. That way, we can avoid correlations between variables that might not exist in reality.

Since the response variable, y_i, is defined as price, we always have that $y_i \geq 0$. On the other hand, one of the classical assumptions in the regression analysis is that the errors e_j are normally distributed. Therefore, the right hand sides of the models could be negative. This problem could be solved by using $\ln y_i$ instead of $y^{0.5}$ and $y^{0.42}$. However, when applying the logarithmic transformation to the models, the assumption of normality is not satisfied, since the residuals clearly follow a quadratic trend which can be observed in Figure 8 in the Appendix. To investigate this further, we will look at the predicted values of y_i and the upper and lower prediction intervals. This is observed in Figure 9, where the blue line is the prediction line, and the green and red lines are the upper and lower prediction inter-

vals. As we can see, the probability that $\hat{y}_i$ is negative is quite small.

When transforming *Colour* and *Clarity* to numerical variables, it is done in such a way that subsequent classes have the same numerical distance. However, there are other possible methods to use rather than this method. For example we could have transformed all the categories to dummy variables. This is the kind of method that is used in Model 4, when for example a diamond has the colour E the other colour categories in the model become 0. There are also other ways of transforming categorical variables into numerical, and the transformation method that is used has have an impact on the outcome.

The purpose of this thesis was to examine models that could explain the pricing of diamonds. The models we have constructed could be used in practice, however we have to examine the predictability and also compute prediction intervals to investigate how well the models fit for this purpose.

7 Appendix

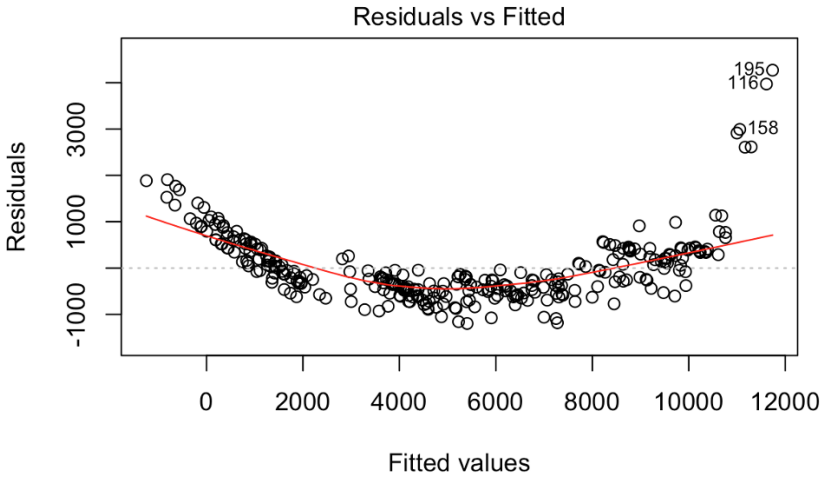

Figure 6: $y = \alpha + \beta_1 Carat_i + \beta_2 Colour_i + \beta_3 Clarity_i + \beta_4 u_i + \beta_5 v_i + e_i$

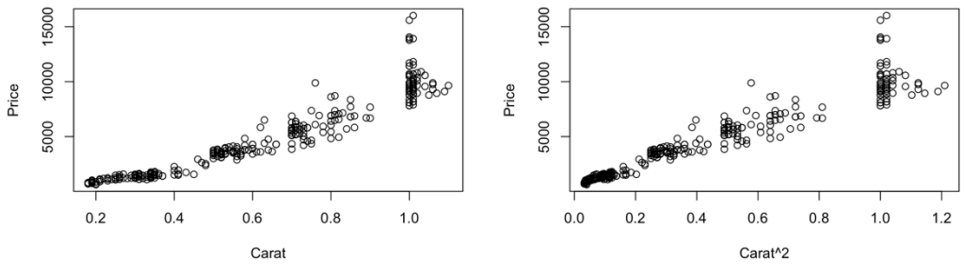

Figure 7: *Price* plotted against *Carat* respectively $Carat^2$

Model 1					Model 2			
Variable	AIC	RSS	Sum of Sq		Variable	AIC	RSS	Sum of Sq
Intercept	546.57	1747	-		Intercept	546.57	1747	-
Carat	1856.99	123849	122102		Carat	1856.99	123849	122102
Colour	1009.38	7902	6155		Colour	1009.38	7902	6155
Clarity	839.39	4550	2803		Clarity	839.39	4550	2803
u_i	562.64	1853	106		u_i	562.64	1839	106
v_i	560.32	1839	92		v_i	560.32	1853	92
Model 3					Model 4			
Variable	AIC	RSS	Sum of Sq		Variable	AIC	RSS	Sum of Sq
Intercept	-1204.26	5.937	-		Intercept	-1.37	281.8	-
Carat	-267.18	125.231	119.294		Carat	1336.17	21816.0	21534.1
Colour	-1023.10	10.760	4.823		Colour	469.97	1344.9	1063.0
Clarity	-1095.63	8.503	2.566		Clarity	307.06	787.3	505.5
u_i	-1147.95	7.174	1.237		Certification	29.29	315.4	33.6
v_i	-1158.81	6.926	0.989					

Table 11: Backward Elimination Results

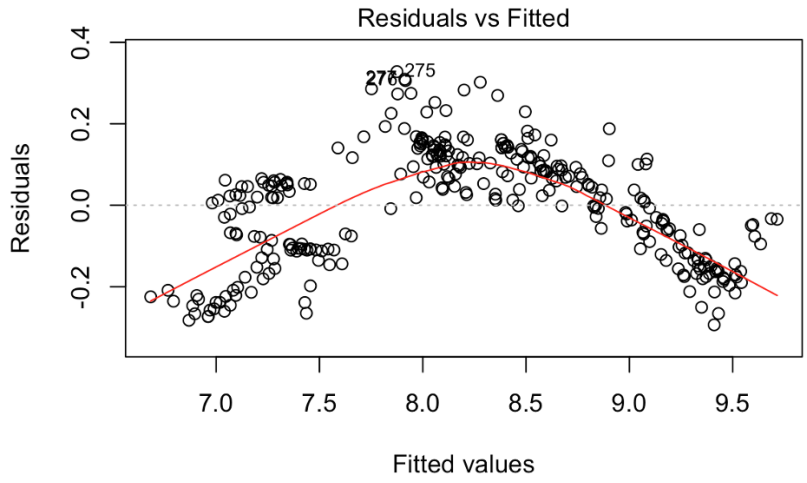

Figure 8: $\ln y = \alpha + \beta_1 Carat_i + \beta_2 Colour_i + \beta_3 Clarity_i + \beta_4 u_i + \beta_5 v_i + e_i$

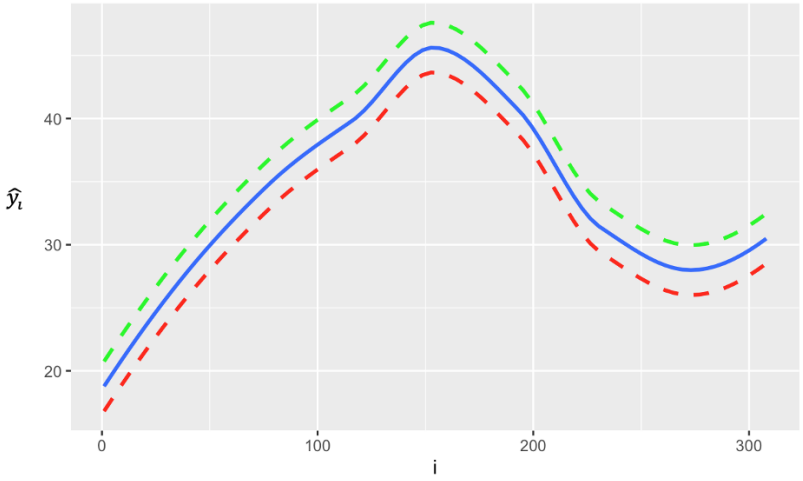

Figure 9: Prediction Line and the Prediction Intervals for Model 3

References

[1] *Diamonds in Nature: A Guide to Rough Diamonds* by Ralph Tappert, Michelle C. Tappert, 2011.

[2] *Diamanternas Historia* by Folke Schimanski, 2008.

[3] *Elements of Regression Analysis* by Harald Lang, 2015.

[4] *Notes in Econometrics* by Patrik Andersson, Kristoffer Lindensjo, Joanna Tyrcha, 2019.

[5] *Generalized Linear Models: with Applications in Engineering and the Sciences* by Douglas C. Montgomery, Raymond H. Myers, G. Geoffrey Vining, Timothy J. Robinson, 2010.

[6] *Econometrics* by Patrik Andersson, Joanna Tyrcha, 2010.

[7] *Kompendium i Lineara Statistiska Modeller* by Rolf Sundberg, 2020.

[8] *The Cambridge Dictionary of Statistics* by B. S. Everitt, A. Skrondal, 1998.

[9] *Brilliance.com*

[10] *Journal of Statistics Education*