# Pricing the C’s of Diamond Stones

Ha Doan, Ha Pham, Duc Nguyen

## 1. *Introduction*

### *1.a. General background relating to the project:*

The formation of natural diamonds takes billions of years, contributing to their rarity and value. While diamonds are highly sought after for their brilliance and durability, their prices are determined by various factors such as carat weight, cut quality, clarity, color, market demand, and geographic origin, significantly influencing diamond prices.

**Carat weight:** Carat plays a significant role in determining the price of diamonds. Carats represent the unit of weight for diamonds, with one carat equaling 0.2 grams [1].

**Cut quality:** Diamond cut is often considered the most crucial factor in evaluating a diamond's price, surpassing its carat weight. The cut enhances the diamond's appeal, symmetry, and overall worth [1].

**Color:** The color of a diamond is a significant factor in determining its price. Contrary to intuition, in the case of traditional white diamonds, the absence of color is highly prized. The Gemological Institute of America (GIA) ranks diamond color on a scale from D (colorless) to Z (light yellow or brown). Diamonds closer to completely colorless are rarer and, therefore, more expensive [1].

**Certification Body (CB):**

Diamond certificate generation involves grading and testing diamonds by independent gemology labs. These labs assess diamonds based on the 4Cs: color, clarity, carat, and cut. Certification assures buyers of the diamond's quality, as experienced gemologists verify it [1].

**Clarity:** Diamond clarity is an important aspect that affects a diamond's value. Diamonds with fewer flaws are graded higher for clarity and more valuable. Diamonds with minimal or no inclusions are highly sought after and can fetch premium prices due to their rarity and superior clarity [2].

### *1.b. Research goals and objectives:*

The diamond industry uses the Four Cs (Caratage, Colour, Clarity, and Cut) to determine prices. In this project, we are developing an explanatory model for diamond prices to analyze the importance of each factor and how they contribute to determining the diamond's price.

### *1.c. Discussion of interesting questions/related problems:*

We are curious to see if the color, clarity, and grading of the diamond truly affect its price on the market based on the data collected. Additionally, we recognize that the weight of the diamond may also influence its price, but determining the extent and nature of this relationship is another aspect we aim to explore.

### *1.d. Description of the variables:*

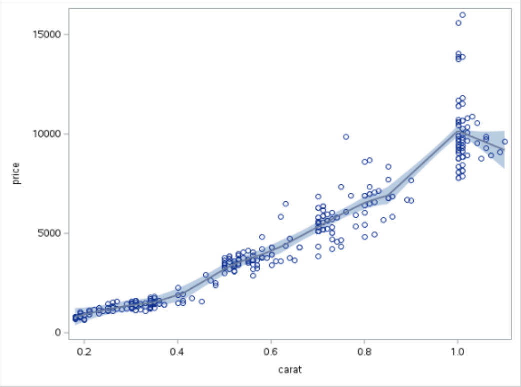
The dataset, named "diamond_price1" or "diamond_price2", contains information about diamond stones and their prices in Singapore dollars. It comprises the following variables:

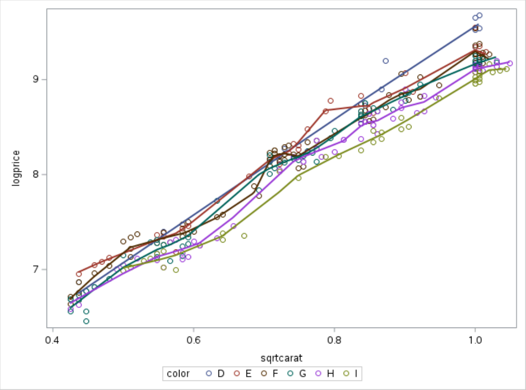
- Carat (carat): A continuous quantitative, independent variable representing diamond weight in carats.
- Colour (color): A nominal qualitative, independent variable indicating diamond color (D - absolutely colorless, E - colorless, F - , G+H - near colorless, or I - faint yellow hue).
- Clarity (clarity): A nominal qualitative, independent variable denoting diamond clarity (IF - Internally Flawless, VVS1, VVS2 - Very Very Slightly Included 1 and 2, VS1,VS2 - Very Slightly Included 1 and 2).
- Certification Body (cb): A nominal variable specifying the certification body for diamond grading (GIA, IGI, or HRD).
- Price: The continuous, dependent variable is measured in Singapore dollars, representing diamond prices.

Note: In "diamond_price1", Colour, Clarity, and Certification Body are directly nominal variables. Conversely, in "diamond_price2", they are encoded using indicator variables.

## 2. *Initial Data Exploring*

### *2.a. Scatter plot construction*





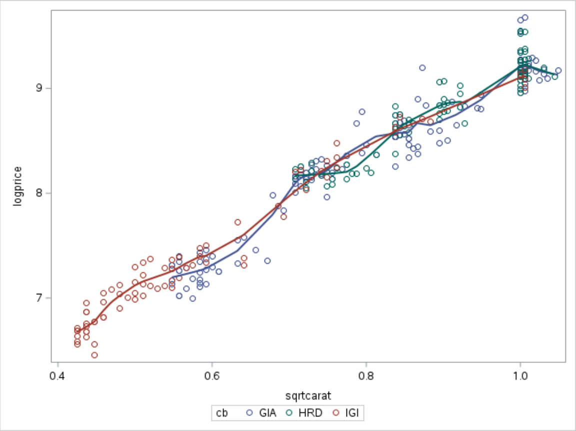
(figure 1 - LOESS curve)

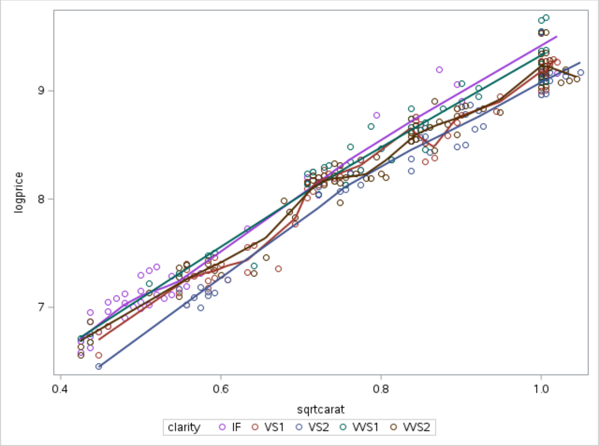
(Figure 2: price vs carat by color)

The initial scatter plot between price and carat suggests a potential linear relationship between these two variables. However, the spread of the data indicates that the model might benefit from some level of transformation. Additionally, the end of the loess curve trends downward instead of continuing its upward trajectory, which suggests a reduced likelihood that a simple linear regression will accurately model the data.

### *2.b. The linear relationship between Price vs Carat by grouping color, certification, or clarity*

### *2.c. Pattern of Price over Categories of Qualitative Variables:*





(Figure 3: Price vs carat by cb)

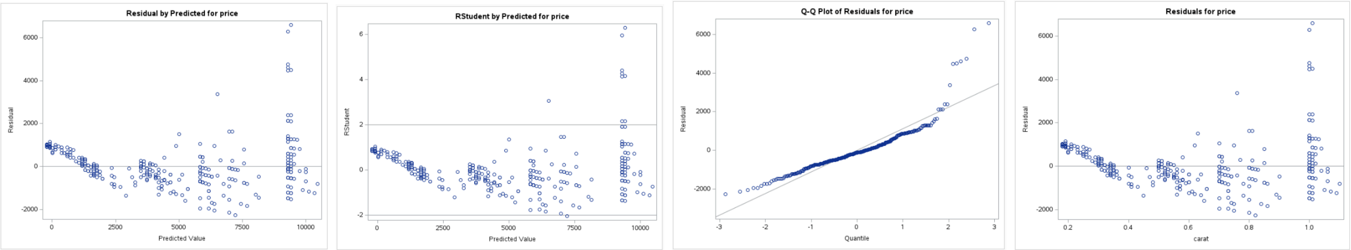
(Figure 4: Price vs carat by clarity)

Color: Figure 2 presents a nuanced delineation of price by diamond color grades, revealing a clear gradient of value where premium pricing correlates strongly with color grades D through F. These color grades exhibit a steeply ascending relationship between carat size and price, a reflection of market preference for colorless diamonds. The incremental price difference between each successive color grade becomes more pronounced with increased carat size. This observation suggests that the color attribute is not only a primary driver of value but also interacts significantly with carat size, intensifying the price differential in larger stones where the absence of color is particularly prized.

Certification Body: Figure 3 explores the influence of certification bodies (GIA, HRD, IGI) on diamond pricing across varying carat sizes. The alignment of the LOESS curve across these bodes implies a general market consensus on the added value of certification. Although slight price differentials can be observed among the certifiers, the overarching trend suggests that the market attributes a uniform value to the role of certification, recognizing it as a guarantor of quality. The plot hints at the possibility of certain certification bodies being marginally favored, potentially due to perceived stringency in grading standards or global recognition, but overall, the effect of certification on price appears to be systematic and indicative of the trust placed in these institutions by the market.

Clarity: Clarity's role in diamond valuation is intricately depicted in Figure 4. Here, the price trajectory for diamonds with superior clarity grades (Internally Flawless, Very Very Slightly Included 1 and 2) demonstrates a sharp upward curve relative to carat size. Conversely, diamonds with clarity grades slightly lower (Very Slightly Included 1 and 2) show a more moderate price progression. The visual spread of the LOESS curves suggests that clarity significantly imparts price, particularly for larger diamonds where clarity is more discernible and hence more valued. The graph indicates that the premium for higher clarity is a compounded effect of both the inherent scarcity of such diamonds and the amplified aesthetic appeal in larger sizes.

### *2.d. Transformation of Y or Quantitative X or Adding a Function of X:*



(figure 5 - Assessing residuals between price and carat)

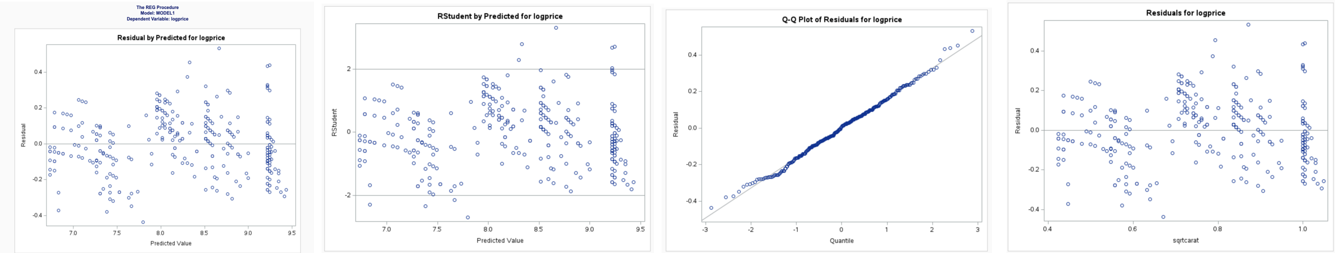
By examining the residual plots, we observe a downward trend in the residuals when plotted against the Price (Y) or the Carat (X). This pattern violates the first assumption of linear regression, which states that residuals should not exhibit systematic behavior relative to X and Y. Moreover, the spread of residuals increases across the range of X and Y, indicating a failure to meet the second assumption of linear regression, which requires that residuals be randomly distributed. In addition, the Q-Q plot reveals a noticeable deviation at the end of the regression line, suggesting a failure to meet the fourth assumption, which requires a linear relationship across all predicted values. Finally, the RStudent plot indicates the presence of multiple outliers, which violates the fifth assumption of linear regression by demonstrating that outliers should not significantly influence the regression model. Therefore, the regression might need a transformation.

Transformation:

Since price (Y) is considerably a large number ranging from 0 to 15000 Singaporean Dollars, in addition to the failure of the second, fourth, and fourth assumptions, it is best practice to transform this variable by logging it. Next, the failure to meet the first assumption suggests we might consider changing the carat values. Given that carat values are relatively small, we can apply a square root transformation, potentially stabilizing the data and improving the model fit.

### *2.e. Checking models' assumptions:*

Figure 6: Residuals plots after transformation.



Examining the residuals against predicted values (Y) or the independent variables (X) reveals that the residuals behave randomly and lack systematic patterns, satisfying the first and second assumptions of linear regression. The Q-Q plot demonstrates a linear relationship between Y and X, fulfilling the fifth assumption of linear regression. Finally, the RStudent plot shows an absence of significant outliers, satisfying the fourth assumption of linear regression. Note that the 6th assumption will be checked if further analysis is conducted during the multilinear regression model since we do not know if we are excluding any variables within the model. Also, the independent variables assumption (assumption 3) will not be checked since the data set does not measure over time, so this assumption can bee neglected.

### *2.f. Interpretation of the regression coefficients, and the associated confidence interval and hypothesis*

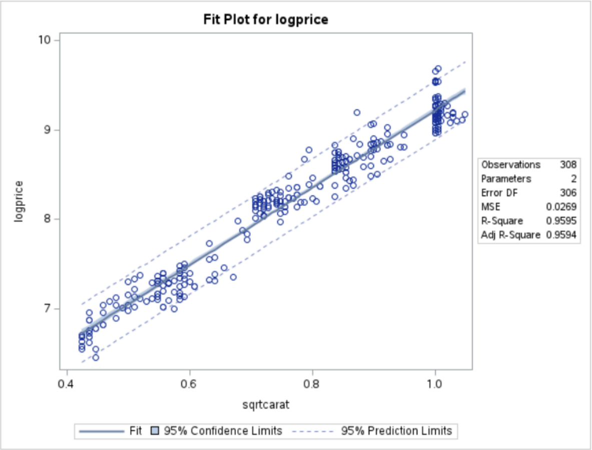


Figure 8: 95% Confidence Interval

Regression coefficient interpretation:

Regression Function: $log(Price) = 4.89629 + 4.32450\sqrt{Carat}$

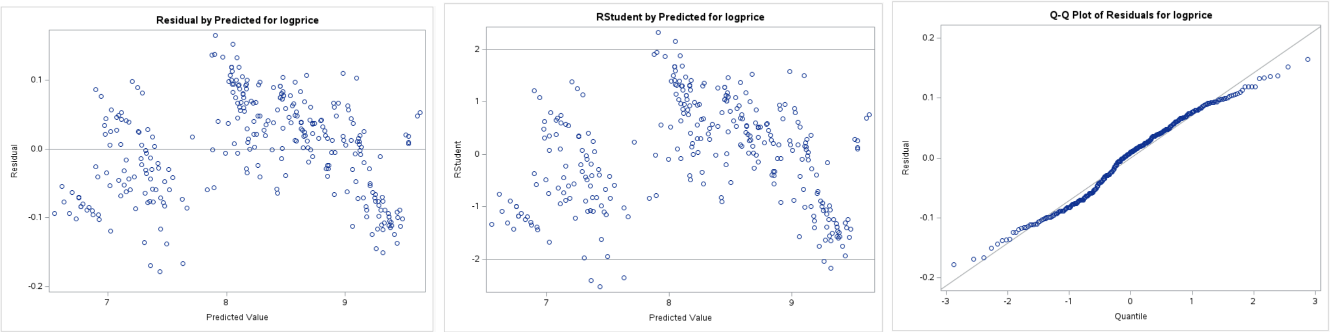
The regression equation depicts a linear relationship between the price's natural logarithm and the carat weight's square root. The coefficient of the square root of carat weight (4.32450) signifies the change in the estimated natural logarithm of the price for a one-unit increase in the square root of the carat weight.

For hypothesis testing, for the null hypothesis, we assume that there is no linear correlation between price and carat (Beta 1). For the alternative hypothesis, we assume that there is a linear relationship between price and carat. According to the ANOVA Table in Figure 7, the t statistic of beta 1 is 85.14 and the p-value of beta is smaller than 0.01, which is considered significant. Thus, p-value and t-test suggest that there is evidence of a linear relationship between carat and price which we can reject the null hypothesis and conclude the alternative hypothesis. In addition, according to Figure 8, most of the data fell into the predicted 95% interval, which is a good sign.

### 3. *Multilinear Regression Models*

#### *3.a. Multiple Regression between logprice vs sqrtcarat, color, and clarity*

*Figure 9: Residual*



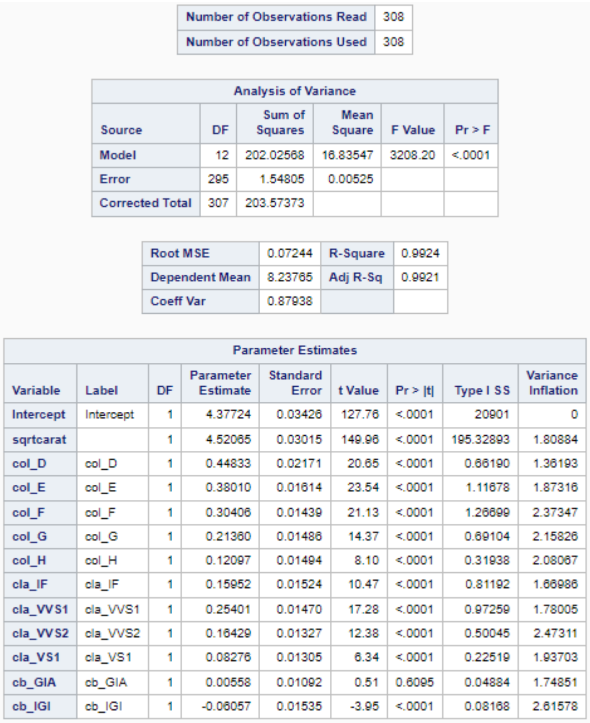
Checking Assumption:

Looking at the residual plots, we can observe that the model behaves as anticipated. In the residuals vs. predicted values plot, the residuals exhibit a somewhat random pattern, with minimal trend along the predicted values, supporting the first assumption of the model. Additionally, the variance of residuals appears consistently random, providing strong evidence for the second assumption that there is no systematic behavior in the residuals. In the Q-Q plot, there is a slight deviation from linearity at both the tail and head, but the majority of predicted values fall within the expected range, indicating that the error terms may be normally distributed. This satisfies the fourth assumption of the model. The RStudent plot of residuals validates the fifth assumption, as all residuals fall below 3 RStudent, indicating the absence of significant outliers. However, the seventh assumption may not hold, as section 1.d suggests that color could have some level of correlation with clarity since their definition of qualification seems closely related to one another. This should be further investigated by testing its significance through the variance inflation factor.

Regression Analysis:

*Figure 10: ANOVA and Regression analysis*

t



| Number of Observations Read | 308 |
|---|---|
| Number of Observations Used | 308 |

| Analysis of Variance | | | | | |
|---|---|---|---|---|---|
| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
| Model | 12 | 202.02568 | 16.83547 | 3208.20 | <.0001 |
| Error | 295 | 1.54805 | 0.00525 | | |
| Corrected Total | 307 | 203.57373 | | | |

| Root MSE | 0.07244 | R-Square | 0.9924 |
|---|---|---|---|
| Dependent Mean | 8.23765 | Adj R-Sq | 0.9921 |
| Coeff Var | 0.87938 | | |

| Parameter Estimates | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Variable | Label | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| | Type I SS | Variance Inflation |
| Intercept | Intercept | 1 | 4.37724 | 0.03426 | 127.76 | <.0001 | 20901 | 0 |
| sqrtcarat | | 1 | 4.52065 | 0.03015 | 149.96 | <.0001 | 195.32893 | 1.80884 |
| col_D | col_D | 1 | 0.44833 | 0.02171 | 20.65 | <.0001 | 0.66190 | 1.36193 |
| col_E | col_E | 1 | 0.38010 | 0.01614 | 23.54 | <.0001 | 1.11678 | 1.87316 |
| col_F | col_F | 1 | 0.30406 | 0.01439 | 21.13 | <.0001 | 1.26699 | 2.37347 |
| col_G | col_G | 1 | 0.21360 | 0.01486 | 14.37 | <.0001 | 0.69104 | 2.15826 |
| col_H | col_H | 1 | 0.12097 | 0.01494 | 8.10 | <.0001 | 0.31938 | 2.08067 |
| cla_IF | cla_IF | 1 | 0.15952 | 0.01524 | 10.47 | <.0001 | 0.81192 | 1.66986 |
| cla_VVS1 | cla_VVS1 | 1 | 0.25401 | 0.01470 | 17.28 | <.0001 | 0.97259 | 1.78005 |
| cla_VVS2 | cla_VVS2 | 1 | 0.16429 | 0.01327 | 12.38 | <.0001 | 0.50045 | 2.47311 |
| cla_VS1 | cla_VS1 | 1 | 0.08276 | 0.01305 | 6.34 | <.0001 | 0.22519 | 1.93703 |
| cb_GIA | cb_GIA | 1 | 0.00558 | 0.01092 | 0.51 | 0.6095 | 0.04884 | 1.74851 |
| cb_IGI | cb_IGI | 1 | -0.06057 | 0.01535 | -3.95 | <.0001 | 0.08168 | 2.61578 |

7th assumption:

Variance Inflation of all variables is bigger than 1 and there exists some level of correlation between the variables. However, VIFs are also smaller than 10 which can be tolerable and, thus, neglectable. Therefore, the 7th assumption can be held true.

Regression function & interpretation:

$$log(price) \ = 4.37 \ + 4.52\sqrt{carat} + 0.45col_D + 0.38col_E + 0.3col_F + 0.21col_G + 0.12col_H + 0.15cla_{IF} + 0.25cla_{VVS1} + \ 0.16cla_{VVS2} + 0.08cla_{VS1} + 0.005cb_{GIA} - 0.06cb_{IGI}$$

For carat, given that all other variables are fixed, on average if sqrtcart increase by one unit, then logprice of diamond will increase by 4.52 log of Singaporean dollar.

For other categorical variables, given that all other variables are fixed (including) if $X_k = 1$, then logprice of diamond will increase by $\beta_k$ such that $k \in [2, 12]$.

F-test, t-test, and the conclusion of the model:

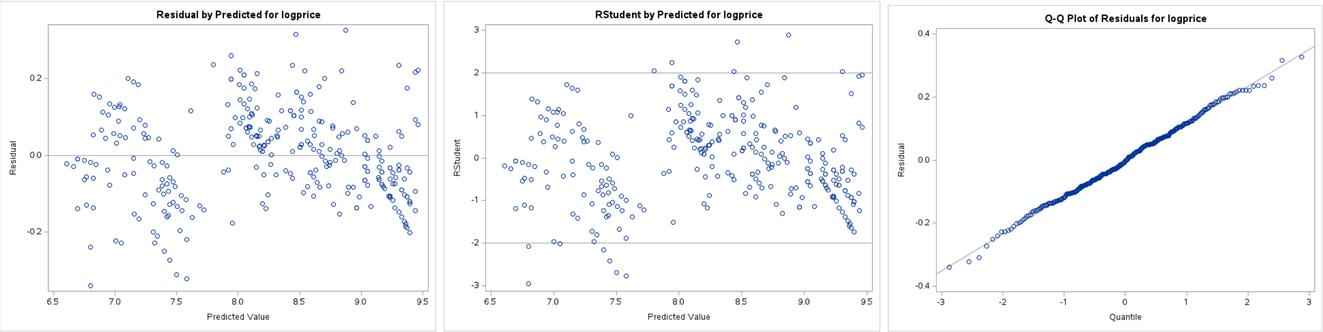
H0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = \beta_8 = \beta_9 = \beta_{10} = \beta_{11} = \beta_{12} = 0$

H1: $\exists\beta_k \neq 0$ such that $k \in [1, 12]$

From the table, the F-value is 3208.2, and the p-value <0.0001. This indicates that the overall regression model is statistically significant at any conventional significance level. In addition, the t-values of all the variables except cb_GIA are significantly different from zero as their p-values are all <0.0001. In summary, both the F-test and the T-test support the conclusion that the model and the individual predictors are statistically significant contributors to predicting the logarithm of the price of diamonds. Furthermore, Sum Square Total is 202 over 203 which over 99% of the data. Thus, this model is considerably great in predicting the price of diamonds, however, we still need to take a look at the interaction effect between sqrtcarat and other categorical variables.

### *3.b. Multiple Regression between logprice vs sqrtcarat and color*

*Figure 13: Residual plots*

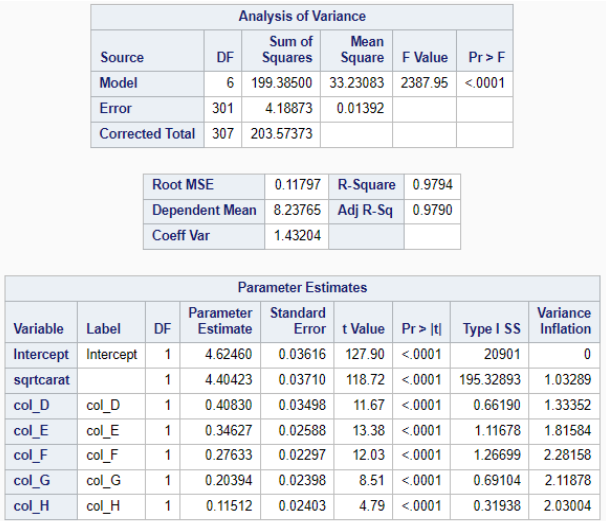


Checking assumption:

Taking a closer look at our residual plots (Figure 13), it's going in the right direction. The distribution of residuals is well-behaved, with no discernible pattern around the zero mark, affirming the linear relationship required for assumption one. Moving on, the residual plot stands up to scrutiny for assumption two, showing a consistent variance across the board—no problematic funnel patterns in sight. When we turn our attention to the Q-Q plot for the fourth assumption, it's a textbook example of normal distribution, with all points aligned neatly along the expected line. For assumption five, a thorough scan for outliers came up clean, ensuring that no rogue data points were skewing our predictions. And for that final check, assumption seven, Figure 14 backs us up with VIF values that are just above 1, putting any concerns about multicollinearity to rest. Stitching all these observations together, it's clear we've got a well-tuned model that's adhering to the essential linear regression assumptions.

Regression Analysis:

*Figure 14: ANOVA, regression analysis*



Analysis of Variance

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 6 | 199.38500 | 33.23083 | 2387.95 | <.0001 |
| Error | 301 | 4.18873 | 0.01392 | | |
| Corrected Total | 307 | 203.57373 | | | |

| | | | |
|---|---|---|---|
| Root MSE | 0.11797 | R-Square | 0.9794 |
| Dependent Mean | 8.23765 | Adj R-Sq | 0.9790 |
| Coeff Var | 1.43204 | | |

Parameter Estimates

| Variable | Label | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| | Type I SS | Variance Inflation |
|---|---|---|---|---|---|---|---|---|
| Intercept | Intercept | 1 | 4.62460 | 0.03616 | 127.90 | <.0001 | 20901 | 0 |
| sqrtcarat | | 1 | 4.40423 | 0.03710 | 118.72 | <.0001 | 195.32893 | 1.03289 |
| col_D | col_D | 1 | 0.40830 | 0.03498 | 11.67 | <.0001 | 0.66190 | 1.33352 |
| col_E | col_E | 1 | 0.34627 | 0.02588 | 13.38 | <.0001 | 1.11678 | 1.81584 |
| col_F | col_F | 1 | 0.27633 | 0.02297 | 12.03 | <.0001 | 1.26699 | 2.28158 |
| col_G | col_G | 1 | 0.20394 | 0.02398 | 8.51 | <.0001 | 0.69104 | 2.11878 |
| col_H | col_H | 1 | 0.11512 | 0.02403 | 4.79 | <.0001 | 0.31938 | 2.03004 |

Regression function: $log(price) = 4.62 + 4.4\sqrt{carat} + 0.41col_D + 0.34col_E + 0.27col_F + 0.2col_G + 0.12col_H$

For carat, given that all other variables are fixed, on average if sqrtcart increase by one unit, then logprice of diamond will increase by 4.40 log of Singaporean dollar.

For other categorical variables, given that all other variables are fixed if $X_k = 1$, then logprice of diamond will increase by $\beta_k$ such that $k \in [2, 6]$.

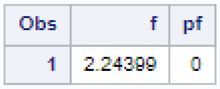
Interaction effect & Partial F-test:

H0: $\beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = 0$

H1: $\exists\beta_k \neq 0$ such that $k \in [2, 6]$

$$SSR(X_2, X_3, X_4, X_5, X_6|X_1) = 0.66190 + 1.11678 + 1.26699 + 0.69104 + 0.31938 = 4.05609$$

$$F^{*} = \frac{SSR(X_2,X_3,X_4,X_5,X_6|X_1)}{5} \div MSE(X_2, X_3, X_4, X_5, X_6, X_1) = 4.05609 \div 5 \div 0.01392 = 58.2771$$

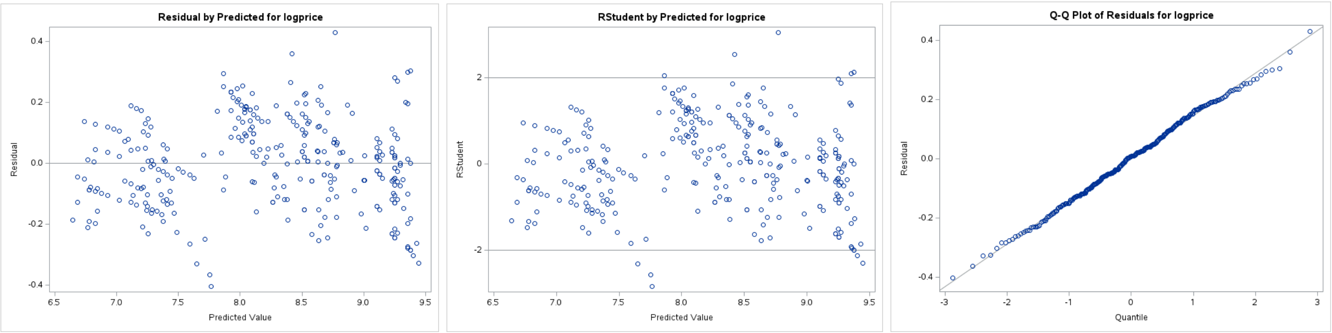


| Obs | f | pf |
|---|---|---|
| 1 | 2.24399 | 0 |

Partial f-test shows that F* is at 58.2771 where its p-value is at near 0 which is considerably significant. Thus, we can reject the null hypothesis and conclude that there is a correlation between log price of the diamond and the color of the diamond. Further inspection from the regression in Figure 14 also shows that the t-value of all color variables is substantially high along with the statistical significance of the p-value aiding the conclusion that color is a good predictor of log price. Therefore, we can keep the color in our model.

### *3.c. Multiple Regression between logprice vs sqrtcarat and clariy*

*Figure 15: Residual plots*

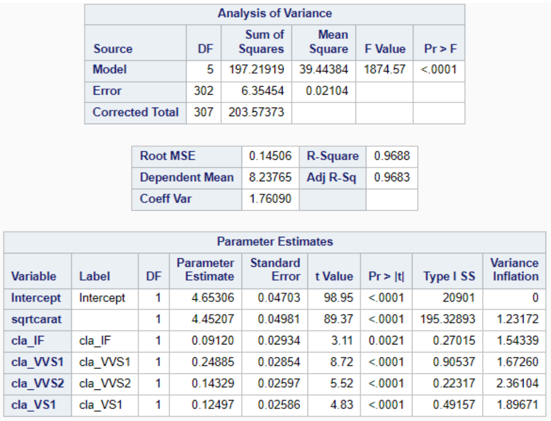


<u>Checking assumptions:</u>

Figure 15 provides clear evidence supporting several key assumptions necessary for the validity of a multiple linear regression analysis. The scatter plot of residuals against predicted values demonstrates linearity, as no systematic pattern is observable (Satisfy assumption 1). Additionally, the constant variance of errors is evident from the left scatter plot where residuals are randomly distributed (Satisfy assumption 2). The QQ plot on the right side further confirms the normal distribution of errors, with residuals following a linear trend along the reference line (Satisfy assumption 4). Moreover, the middle plot reveals no outliers, as all points lie within an acceptable range without any extreme deviations (Satisfy assumption 5). Lastly, the ANOVA table (Figure 16) shows a variance inflation factor (VIF) for all variables below 10, affirming low multicollinearity among the predictors (Satisfy assumption 7).

<u>Regression Analysis</u>:

*Figure 16: ANOVA, regression analysis*



| Analysis of Variance | | | | | |
|---|---|---|---|---|---|
| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
| Model | 5 | 197.21919 | 39.44384 | 1874.57 | <.0001 |
| Error | 302 | 6.35454 | 0.02104 | | |
| Corrected Total | 307 | 203.57373 | | | |

| | | | |
|---|---|---|---|
| Root MSE | 0.14506 | R-Square | 0.9688 |
| Dependent Mean | 8.23765 | Adj R-Sq | 0.9683 |
| Coeff Var | 1.76090 | | |

| Parameter Estimates | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Variable | Label | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| | Type I SS | Variance Inflation |
| Intercept | Intercept | 1 | 4.65306 | 0.04703 | 98.95 | <.0001 | 20901 | 0 |
| sqrtcarat | | 1 | 4.45207 | 0.04981 | 89.37 | <.0001 | 195.32893 | 1.23172 |
| cla_IF | cla_IF | 1 | 0.09120 | 0.02934 | 3.11 | 0.0021 | 0.27015 | 1.54339 |
| cla_VVS1 | cla_VVS1 | 1 | 0.24885 | 0.02854 | 8.72 | <.0001 | 0.90537 | 1.67260 |
| cla_VVS2 | cla_VVS2 | 1 | 0.14329 | 0.02597 | 5.52 | <.0001 | 0.22317 | 2.36104 |
| cla_VS1 | cla_VS1 | 1 | 0.12497 | 0.02586 | 4.83 | <.0001 | 0.49157 | 1.89671 |

<u>Regression function:</u> $log(price) = 4.65 + 4.45\sqrt{carat} + 0.09cla_{IF} + 0.24cla_{VVS1} + 0.14cla_{VVS2} + 0.12cla_{VS1}$

For carat, given that all other variables are fixed, on average if sqrtcart increase by one unit, then logprice of diamond will increase by 4.45 log of Singaporean dollar.

For other categorical variables, given that all other variables are fixed if $X_k = 1$, then logprice of diamond will increase by $\beta_k$ such that $k \in [2, 5]$.

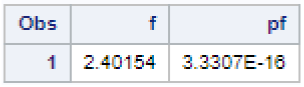
Interaction & Partial F-test:

H0: $\beta_7 = \beta_8 = \beta_9 = \beta_{10} = 0$

H1: $\exists\beta_k \neq 0$ such thats $k \in [7, 10]$

$SSR(X_7, X_8, X_9, X_{10}|X_1) = 0.49157 + 0.22317 + 0.90537 + 0.27015 = 1.89026$

$F^* = \frac{SSR(X_7,X_8,X_9,X_{10}|X_1)}{4} \div MSE(X_7, X_8, X_9, X_{10}, X_1) = 1.89026 \div 4 \div 0.02104 = 22.46031369$

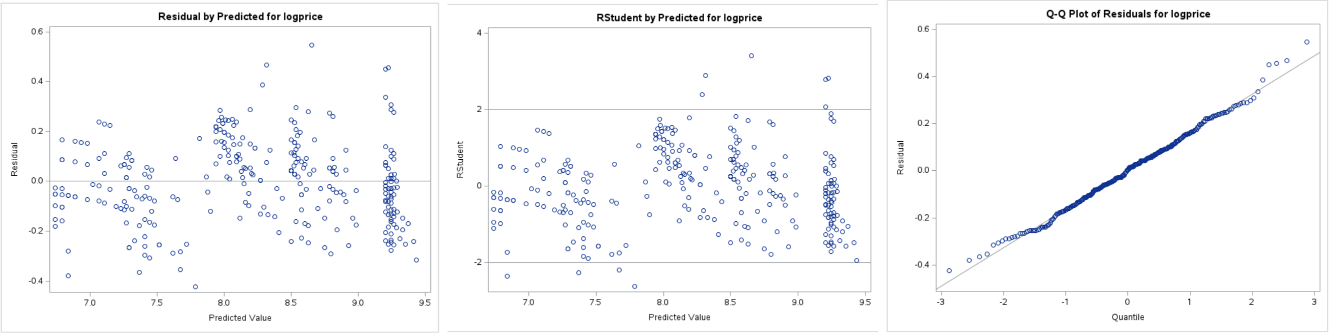


| Obs | f | pf |
|---|---|---|
| 1 | 2.40154 | 3.3307E-16 |

Partial f-test shows that F* is at 22.46 where its p-value is at near 0 which is considerably significant. Thus, we can reject the null hypothesis and conclude that there is a correlation between log price of the diamond and the clarity of the diamond. Further inspection from the regression in Figure 16 also shows that the t-value of all clarity variables is substantially high along with the statistical significance of the p-value aiding the conclusion that clarity is a good predictor of log price. Therefore, we can keep the clarity in our model.

### *3.d. Multiple Regression between logprice vs sqrtcarat and certification body*

*Figure 17: Residual plots*

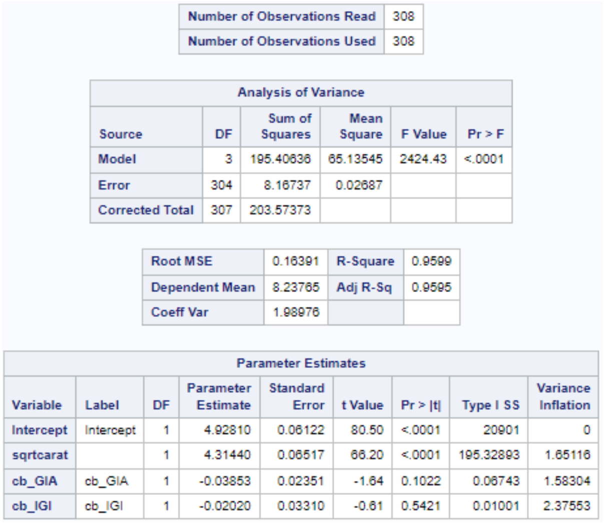


Checking Assumption:

Looking at Figure 17 we can see that it provides clear evidence supporting several key assumptions necessary for the validity of a multiple linear regression analysis. The scatter plot of residuals against predicted values demonstrates linearity, as no systematic pattern is observable (Satisfy assumption 1). Additionally, the constant variance of errors is evident from the left scatter plot where residuals are randomly distributed (Satisfy assumption 2). The QQ plot on the right side further confirms the normal distribution of errors, with residuals following a linear trend along the reference line (Satisfy assumption 4). Moreover, the middle plot reveals maybe only one hard-to-visible outlier, as all other points lie within an acceptable range without any extreme deviations (Satisfy assumption 5). Lastly, the ANOVA table (Figure 18) shows a variance inflation factor (VIF) for all variables below 10, affirming low multicollinearity among the predictors (Satisfy assumption 7).

Regression Analysis:

*Figure 18: ANOVA, regression analysis*



| | |
|---|---|
| Number of Observations Read | 308 |
| Number of Observations Used | 308 |

| Analysis of Variance | | | | | |
|---|---|---|---|---|---|
| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
| Model | 3 | 195.40636 | 65.13545 | 2424.43 | <.0001 |
| Error | 304 | 8.16737 | 0.02687 | | |
| Corrected Total | 307 | 203.57373 | | | |

| | | | |
|---|---|---|---|
| Root MSE | 0.16391 | R-Square | 0.9599 |
| Dependent Mean | 8.23765 | Adj R-Sq | 0.9595 |
| Coeff Var | 1.98976 | | |

| Parameter Estimates | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Variable | Label | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| | Type I SS | Variance Inflation |
| Intercept | Intercept | 1 | 4.92810 | 0.06122 | 80.50 | <.0001 | 20901 | 0 |
| sqrtcarat | | 1 | 4.31440 | 0.06517 | 66.20 | <.0001 | 195.32893 | 1.65116 |
| cb_GIA | cb_GIA | 1 | -0.03853 | 0.02351 | -1.64 | 0.1022 | 0.06743 | 1.58304 |
| cb_IGI | cb_IGI | 1 | -0.02020 | 0.03310 | -0.61 | 0.5421 | 0.01001 | 2.37553 |

Regression function: $log(price) = 4.93 + 4.31\sqrt{carat} - 0.03cla_{GIA} - 0.02cla_{IGII}$

For carat, given that all other variables are fixed, on average if sqrtcart increase by one unit, then logprice of diamond will increase by 4.31 log of Singaporean dollar.

For other categorical variables, given that all other variables are fixed if $X_k = 1$, then logprice of diamond will increase by $\beta_k$ such that $k \in [2, 6]$.

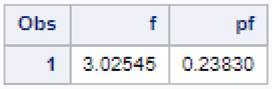
Interaction effect & Partial F-test:

H0: $\beta_{11} = \beta_{12} = 0$

H1: $\exists\beta_k \neq 0$ such that $k \in [11, 12]$

$SSR(X_{11}, X_{12}|X_1) = 0.06743 + 0.01001 = 0.07744$

$F^{*} = \frac{SSR(X_{11},X_{12}|X_1)}{2} \div MSE(X_{11}, X_{12}, X_1) = 0.05661 \div 2 \div 0.02687 = 1.441012281$



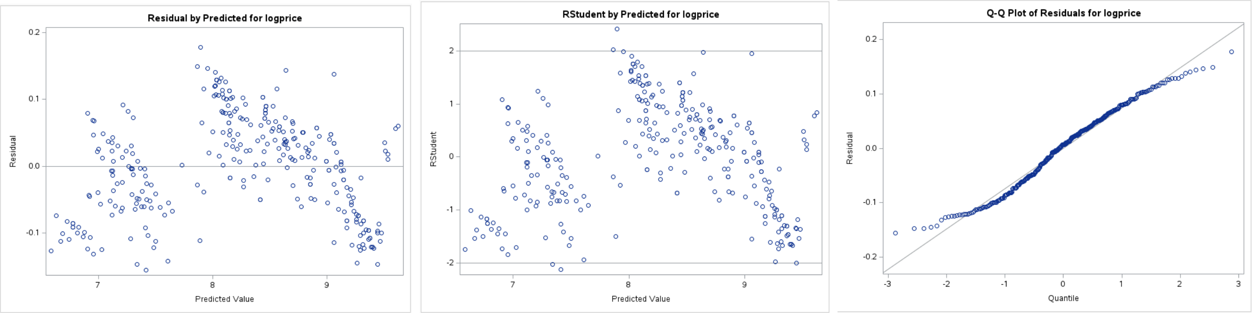
| Obs | f | pf |
|---|---|---|
| 1 | 3.02545 | 0.23830 |

Partial f-test shows that F* is at 1.44 where its p-value is at near 0.23 which is not significant. Thus, we fail to reject the null hypothesis and conclude that there might be no correlation between the certification body and log price. Further inspection from the regression in Figure 14 also shows that the t-value of all cb variables is considerably low along with their statistical significance of the p-value aiding the conclusion that certification might not be a good predictor of log price. Therefore, we exclude this variable from the model.

### 4. *Conclusion & Final Diamond Price Predicting Model*

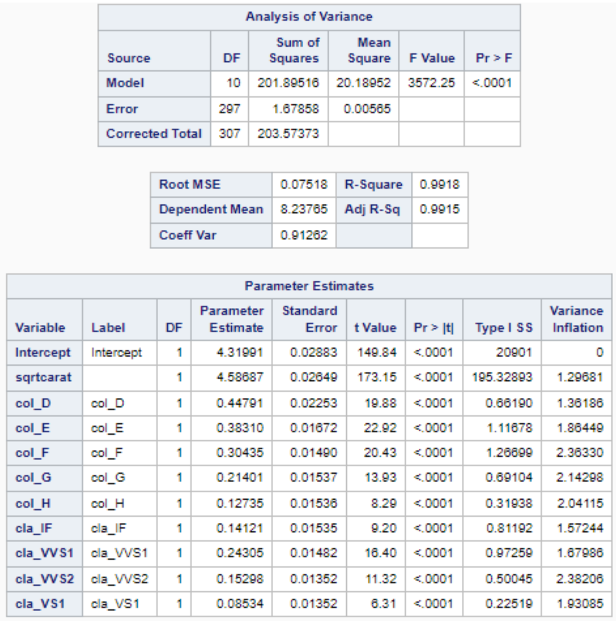
After running regression and testing, we conclude that diamond price can be predicted using Multilinear regression from variables of the square root of carat, colors, and clarities. Due to the absence of the certification body in our model, we can see that the sixth assumption in our final model might not hold true.

*Figure 19: Residual plots*



Taking a closer look at our residual plots (Figure 11), it's going in the right direction. The distribution of residuals is well-behaved, with no discernible pattern around the zero mark, affirming the linear relationship required for assumption one. Moving on, the residual plot stands up to scrutiny for assumption two, showing a consistent variance across the board—no problematic funnel patterns in sight. When we turn our attention to the Q-Q plot for the fourth assumption, it's a textbook example of normal distribution, with all points aligned neatly along the expected line. For assumption five, a thorough scan for outliers came up clean, ensuring that no rogue data points were skewing our predictions. For assumption 6, however, we did not include the certification body variable which might fall into the second type error for the conclusion we made in section 3.d.  Lastly, the ANOVA table (Figure 18) shows a variance inflation factor (VIF) for all variables below 10, affirming low multicollinearity among the predictors (Satisfy assumption 7). Here, we can see that the residual assumptions are good enough to further regression analysis.

Regression function & Interpretation:



Analysis of Variance

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 10 | 201.89516 | 20.18952 | 3572.25 | <.0001 |
| Error | 297 | 1.67858 | 0.00565 | | |
| Corrected Total | 307 | 203.57373 | | | |

| | | | |
|---|---|---|---|
| Root MSE | 0.07518 | R-Square | 0.9918 |
| Dependent Mean | 8.23765 | Adj R-Sq | 0.9915 |
| Coeff Var | 0.91262 | | |

Parameter Estimates

| Variable | Label | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| | Type I SS | Variance Inflation |
|---|---|---|---|---|---|---|---|---|
| Intercept | Intercept | 1 | 4.31991 | 0.02883 | 149.84 | <.0001 | 20901 | 0 |
| sqrtcarat | | 1 | 4.58687 | 0.02649 | 173.15 | <.0001 | 195.32893 | 1.29681 |
| col_D | col_D | 1 | 0.44791 | 0.02253 | 19.88 | <.0001 | 0.66190 | 1.36186 |
| col_E | col_E | 1 | 0.38310 | 0.01672 | 22.92 | <.0001 | 1.11678 | 1.86449 |
| col_F | col_F | 1 | 0.30435 | 0.01490 | 20.43 | <.0001 | 1.26699 | 2.36330 |
| col_G | col_G | 1 | 0.21401 | 0.01537 | 13.93 | <.0001 | 0.69104 | 2.14298 |
| col_H | col_H | 1 | 0.12735 | 0.01536 | 8.29 | <.0001 | 0.31938 | 2.04115 |
| cla_IF | cla_IF | 1 | 0.14121 | 0.01535 | 9.20 | <.0001 | 0.81192 | 1.57244 |
| cla_VVS1 | cla_VVS1 | 1 | 0.24305 | 0.01482 | 16.40 | <.0001 | 0.97259 | 1.67986 |
| cla_VVS2 | cla_VVS2 | 1 | 0.15298 | 0.01352 | 11.32 | <.0001 | 0.50045 | 2.38206 |
| cla_VS1 | cla_VS1 | 1 | 0.08534 | 0.01352 | 6.31 | <.0001 | 0.22519 | 1.93085 |

$$log(price) \ = 4.31 \ + 4.58\sqrt{carat} + 0.45col_{D} + 0.38col_{E} + 0.3col_{F} + 0.21col_{G} + 0.13col_{H} + 0.14cla_{IF} + 0.24cla_{VVS1} + \ 0.15cla_{VVS2} + 0.08cla_{VS1}$$

For carat, given that all other variables are fixed, on average if sqrtcart increase by one unit, then logprice of diamond will increase by 4.58 log of Singaporean dollar.

For other categorical variables, given that all other variables are fixed if $X_k = 1$, then logprice of diamond will increase by $\beta_k$ such that $k \in [2, 10]$.

R-square:

R-square of the model is not as large as the model in 3.a, however, it is still at a good level and the model con explain over 99% of the data variability.

F-test:

H0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = \beta_8 = \beta_9 = \beta_{10} = 0$

H1: $\exists\beta_k \neq 0$ such that $k \in [1, 10]$

Since f-value is at 3572.25 and the model p-value is near 0, we can see that there exists a strong correlation between log price and at least one of the variables. Thus, we can reject the null hypothesis and conclude the alternative hypothesis.

t-test:

H0: $\beta_k = 0$ such that $k \in [1, 10]$

H1: $\beta_k \neq 0$ such that $k \in [1, 10]$

Overall t-values of all variables are greater than 5 which their p-values are all smaller than 0.0001 and considerably statistical significant. Thus, all of the variables within the model have a strong correlation with the price of diamonds.

Model conclusion:

This is a good model for predicting diamond prices, as it can explain 99% of the variance in the data and eliminate other sources of error and irrelevant variables, such as the certification body. Therefore, this is the most efficient model for predicting diamond prices.

**Peer assessment**:

Ha Doan: Responsible for the correctness of the SAS ODA code along with the MLR interpretation in 3.a

Duc Nguyen: Responsible for the information in 3b: Multi Linear Regression between logprice vs sqrtcarat and color.

Ha Pham: Completed information in 3c: Multi Linear Regression between logprice vs sqrtcarat and clarity

**Reference:**

Shah, Chirag. “Diamond Certificates- A Quick Guide on IGI, Gia & HRD.” *ANITA DIAMONDS*, 20 Jan. 2024, www.anitadiamonds.com/news-and-blog/diamond-certificates--a-quick-guide-on-igi-gi-8/.

“Diamond Prices: What Determines the Price of a Diamond?” *Dover Jewelry Blog*, 22 June 2023, www.doverjewelry.com/blog/diamond-prices-what-determines-the-price-of-a-diamond/#:~:text=Generally%20speaking%2C%20the%20heavier%20the,diamonds%20of%20the%20same%20quality.