“Analysis and Prediction of Diamond Prices using Regression”

Course: "DADS6001 Applied Modern Statistical Analysis"

Student Name: "Miss Nutthida Yotaprasert"

Student ID: "6610422019"

Supervisor: "Asst. Prof. Dr. Ramidha Srihera"

Project Description: "The project focuses on analyzing and predicting diamond prices by utilizing a diverse dataset to create a statistical model that can effectively forecast diamond values based on key characteristics."

Date: "9 May 2024"

Analysis and Prediction of Diamond Prices using Regression

Introduction

Analyzing data to predict dependent variables from independent variables is crucial in understanding and forecasting data behavior across various fields. Predicting diamond prices based on their characteristics using regression is a compelling example. By utilizing regression, statistical models can be created to predict diamond prices from their various characteristics.

This research project focuses on analyzing diamond price data with a diverse set of characteristics like weight, cut quality, color, clarity, and size. The primary goal is to develop a suitable statistical model to predict diamond prices accurately, aiding analysts and stakeholders in better understanding and predicting diamond values.

The dataset used in the project comprises crucial information such as diamond prices in USD, carat weight, cut quality, color, clarity, and dimensions, essential for effective analysis and prediction of diamond prices. Through data analysis and the creation of appropriate statistical models, a better understanding and accurate prediction of diamond values can be achieved.

Therefore, analyzing data to predict diamond prices showcases the importance and benefits of regression in solving mathematical problems in daily life and predicting crucial variables for decision-making in the diamond industry today. The aim is for the results of this project to be beneficial and provide guidance for future research and development efforts.

Objectives

This project aims to analyze and predict diamond prices using a dataset with diverse characteristics like weight, cut quality, color, clarity, and size. The objective is to create a suitable statistical model to predict diamond prices accurately from these characteristics. By analyzing data and developing appropriate statistical models, a more efficient and precise prediction of diamond values can be achieved. Analyzing and predicting diamond prices demonstrates the importance and benefits of regression in solving mathematical problems in daily life and predicting essential variables for decision-making in the current diamond industry.

References

https://www.kaggle.com/datasets/shivam2503/diamonds
https://github.com/Jeniejean/Applied-Stat/blob/main/Diamond_Data.csv

The dataset consists of the following variables:

- Price: The price of the diamond in United States dollars (USD), ranging from $326 to $18,823.
- Carat: The weight of the diamond in carats, ranging from 0.2 to 5.01.
- Cut: The quality of the diamond cut (Fair, Good, Very Good, Premium, Ideal), with corresponding numerical values: "Fair" = 0, "Good" = 1, "Very Good" = 2, "Premium" = 3, and "Ideal" = 4.
- Color: The color grade of the diamond, ranging from J (worst) to D (best), with numerical values assigned as follows: J = 0, I = 1, H = 2, G = 3, F = 4, E = 5, D = 6.
- Clarity: The clarity grade of the diamond (I1 (worst), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (best)), with numerical values assigned as follows: I1 = 0, SI2 = 1, SI1 = 2, VS2 = 3, VS1 = 4, VVS2 = 5, VVS1 = 6, IF = 7.
- x: The length of the diamond in millimeters, ranging from 0 to 10.74.
- y: The width of the diamond in millimeters, ranging from 0 to 58.9.
- z: The depth of the diamond in millimeters, ranging from 0 to 31.8.
- Depth: The total depth percentage of the diamond, calculated as z / mean(x, y) or 2 * z / (x + y), with values ranging from 43 to 79.
- Table: The width of the top of the diamond compared to its widest point, with values ranging from 43 to 95.

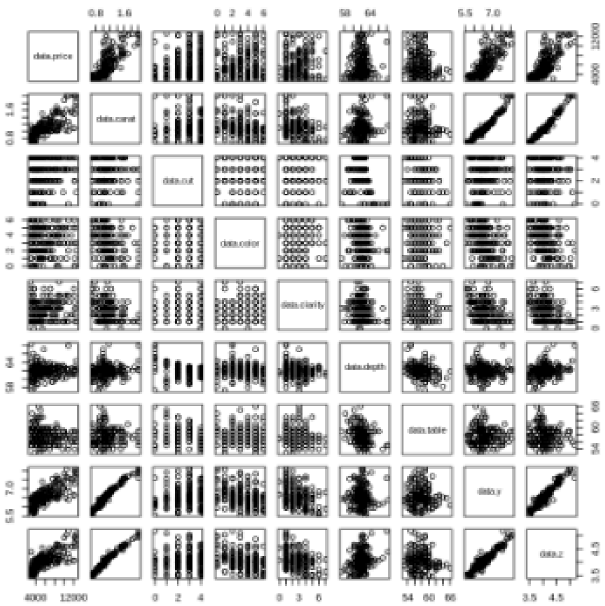
R Programming

1. **Plotting data and initiating data analysis**

- Using the readr package to load and read data from a CSV file from a specified URL
- Using the pairs() function to create a scatter plot matrix to analyze the data distribution in the created data frame

```
library(readr)
data <- read_csv("https://raw.githubusercontent.com/Jeniejean/Applied-Stat/main/Diamond_Data%20(3).csv")
dataf <- data.frame(data$price, data$carat, data$cut, data$color, data$clarity, data$depth, data$table, data$y, data$z)
pairs(dataf)
```

```
Rows: 203 Columns: 10
── Column specification ────────────────────────────────────────────────────
Delimiter: ","
dbl (10): carat, cut, color, clarity, depth, table, price, x, y, z

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```



2. Variable Selection

Selecting important variables to create the best Linear Regression model from the previously built Linear Regression model

Equation: The linear regression equation derived from the final model is:

Diamond Price = -50330.66 + 121.63 * cut + 477.39 * color + 777.89 * clarity + 304.82 * depth + 5106.73 * y

Adjusted R-squared: The adjusted R-squared value of this model is 0.8756, indicating that the 5 independent variables (cut, color, clarity, depth, y) can explain approximately 87.56% of the variance in diamond prices.

P-value: The p-value of this model is < 2.2e-16, which is less than the standard statistical significance value set at 0.05.

Hypotheses: H0: $\beta i = 0$

H1: $\beta i \neq 0$

Results Analysis: With a p-value less than 0.05, we can reject the null hypothesis H0 and accept the alternative hypothesis H1, indicating a response of diamond prices to the input independent variables.

Conclusion: The analysis reveals that the 5 independent variables (cut, color, clarity, depth, y) significantly influence diamond prices. The linear regression model developed has a strong ability to explain the data, with an adjusted R-squared value of approximately 87.56%.

```r
dataf <- data.frame(data)
model <- lm(price~ carat+cut+color+clarity+depth+y+z, data = data)
library(MASS)
step.model <- stepAIC(model, direction = "both", trace = FALSE)
summary(step.model)
```

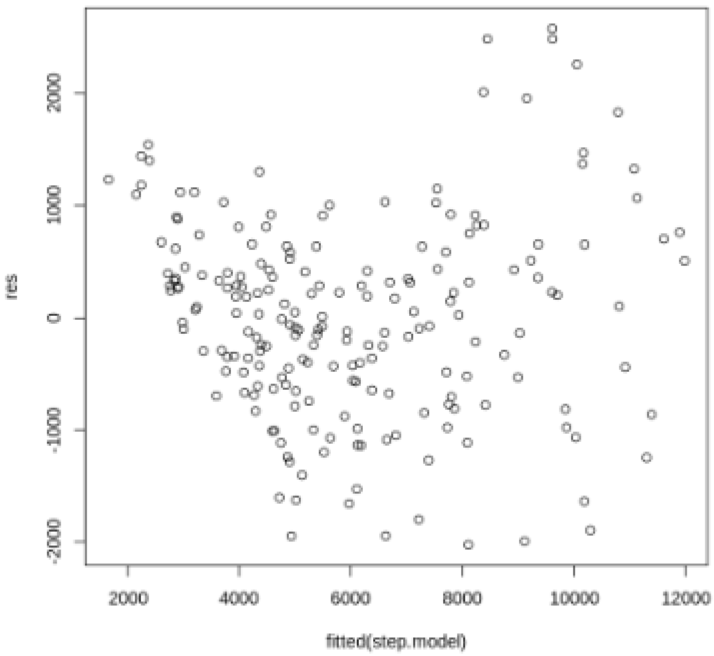
```
Call:
lm(formula = price ~ cut + color + clarity + depth + y, data = data)

Residuals:
     Min       1Q   Median       3Q      Max 
-2024.04  -582.36    26.43   511.83  2578.76 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -50330.66    3099.14 -16.240  < 2e-16 ***
cut            121.63      57.03   2.133   0.0342 *  
color          477.39      39.46  12.100  < 2e-16 ***
clarity        777.89      46.47  16.741  < 2e-16 ***
depth          304.82      45.51   6.698 2.16e-10 ***
y             5106.73     140.37  36.381  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 895.7 on 197 degrees of freedom
Multiple R-squared:  0.8786,	Adjusted R-squared:  0.8756 
F-statistic: 285.2 on 5 and 197 DF,  p-value: < 2.2e-16
```

3. **Create a scatter plot to examine the relationship between the fitted values and the residuals in the Linear Regression model.**

The x-axis represents the predicted values, and the y-axis represents the residuals, showing the relationship between predicted and residual values. This aids in checking the completeness of the Linear Regression model, which should have residuals distributed randomly both vertically and horizontally. **If residuals scatter randomly around zero, it indicates the suitability of our Linear Regression model. However, non-random shapes or trends above or below may suggest inadequacy of the model.**

From examining the Scatter Plot, it's observed that the majority of residuals scatter above and below the zero level.

```
res <- resid(step.model)
plot(fitted(step.model), res)
```



4. **Durbin-Watson Test**

To check for autocorrelation in the Linear Regression model

- The lag Autocorrelation value is 0.2184589
- The D-W Statistic (Durbin-Watson Statistic) value is 1.549828
- The p-value is 0.002

The results of the Durbin-Watson test provide the D-W Statistic (Durbin-Watson Statistic), which is a value between 0 and 4, with the following interpretation:

<u>If the D-W Statistic is closer to 2 (between 1.5 and 2.5), it indicates that there is no time-related autocorrelation in the residuals.</u>

One of the conditions for multiple regression analysis is that the error terms must be independent. The Durbin-Watson statistic is used to check for this. If the Durbin-Watson value is close to 2, i.e., falls within the range of 1.5 to 2.5, it can be concluded that the error terms are independent.

From the data analysis, the Durbin-Watson value was found to be 1.549828, which is within the range of 1.5 to 2.5. Therefore, it can be concluded that the independent variables used in the test do not have any internal correlation.

```
install.packages("car")
library(car)
durbinWatsonTest(step.model)
```

```
Installing package into '/usr/local/lib/R/site-library'
(as 'lib' is unspecified)

 lag Autocorrelation D-W Statistic p-value
   1       0.2184589      1.549828   0.002
 Alternative hypothesis: rho != 0
```

5. Cook's distance

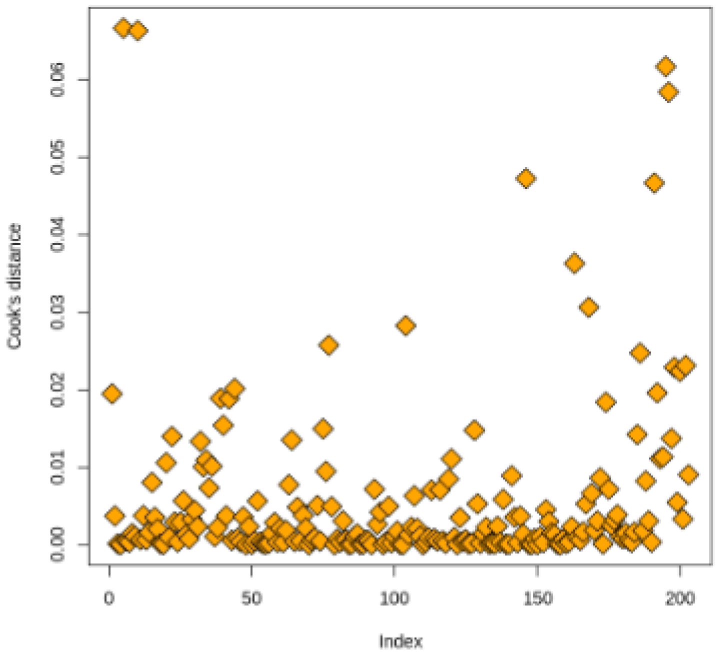
Cook's distance is a measure used to assess how each data point in a Linear Regression model influences the estimation within the model. It is primarily used to examine data points that have a significant impact on the model or may be outliers, or data points that have the most influence on the estimation within the model. A high Cook's distance for a data point indicates that the data point may have a significant impact on the estimation within the model, requiring further consideration or detailed analysis to decide whether that data point should be included in the model or not.

When the Cook's distance falls within the range of 0.00 to 0.07, it indicates that the data point has a minimal impact on the estimation within the Linear Regression model, which is considered to have a very low impact on the estimation within the model. Overall, the model can still provide reliable results in estimating the independent variables that affect the dependent variable efficiently.

```
out<-cooks.distance(step.model)
out
```

```
1:    0.019478838679321 2:   0.0037130909124606 3:   0.000115005072207719 4:   9.80643226451709e-06 5:   0.0666335577581792 6:   0.000483196807686789 7:   0.000310287599308864 8:   0.00142677051875871 9:   0.00107236061732534 10:
      0.0663018511125092 11:   0.000636554979191252 12:   0.00374975282427623 13:   0.000636841934684354 14:   0.00157199850326975 15:   0.00803637697860493 16:   0.0035351430647189 17:   0.00202045768652925 18:   0.000114226694304681 19:
      5.83276267378604e-05 20:   0.0105870132843604 21:   0.000940499973135732 22:   0.0139778375369013 23:   0.00292258290250594 24:   0.000358104182712075 25:   0.00288372834808071 26:   0.00566503812318891 27:   0.00131534075586025 28:
      0.000746650488034 29:   0.0033003291508467 30:   0.0044297175610939 31:   0.00231887768803256 32:   0.01333377631637 33:   0.010057440120989 34:   0.0109362631584591 35:   0.00733344235734592 36:   0.0101263586486837 37:   0.00105454575832691
38:   0.00220281082021339 39:   0.0188969692139048 40:   0.0154182723026542 41:   0.00373050253236186 42:   0.0188363570153133 43:   0.000558861188148635 44:   0.0201535136849239 45:   0.000808585213694767 46:   0.000367756593126729 47:
      0.00358995781789403 48:   5.2216751827397e-06 49:   0.00234464411464192 50:   5.65606953479142e-05 51:   0.000449808834848904 52:   0.00564370285249941 53:   0.000347699199903897 54:   0.000114185777812373 55:   8.9776532135295e-05 56:
      0.000212476963893511 57:   0.00145347721634043 58:   0.00285726901893939 59:   0.000311010961847488 60:   0.00210119822285202 61:   0.000298437254355057 62:   0.00183876699433231 63:   0.00773812323889919 64:   0.0135152468955668 65:
      0.000414362588122262 66:   0.00483933190731253 67:   0.00047567678739892 68:   0.00377826339047961 69:   0.00213553898041566 70:   2.29784692370471e-06 71:   0.000548835985071108 72:   0.000854397727338 73:   0.0050663870063159 74:
      0.000468491931867962 75:   0.0149621156596902 76:   0.00946788541167132 77:   0.0257606466237255 78:   0.00491842004578622 79:   3.47926875943907e-07 80:   0.000727550145669183 81:   0.000328936124348973 82:   0.00306174949403815 83:
      0.00060632882450128 84:   1.07315793820711e-05 85:   8.30862172491116e-05 86:   0.00107048504664634 87:   0.00137117024725735 88:   3.34307048603262e-05 89:   6.1783680028182e-05 90:   0.00057991717373639 91:   0.000508217834233316 92:
      2.34220075203673e-05 93:   0.0071596434018947 94:   0.00263846205993666 95:   0.00419591950341283 96:   1.7729721468691e-05 97:   0.000326160528118365 98:   0.00496224336725883 99:   0.00025115583770624 100:   0.00111655121010304 101:
      0.00173595353179802 102:   4.71334788836694e-05 103:   1.09941997918797e-05 104:   0.0282801646293699 105:   0.00129211854757949 106:   0.00225681208953637 107:   0.00631176686551784 108:   0.00211020193126746 109:   0.00108536348521703 110:
      3.5121787353621e-07 111:   0.000310276456585486 112:   0.000887993057388667 113:   0.00705499553339696 114:   0.000775426608195787 115:   0.000484993580721768 116:   0.00706343706409355 117:   0.000490292878283787 118:   0.00018784886098131 119:
      0.00847326144650538 120:   0.0110718528024455 121:   0.00092333963446622 122:   1.93295586916002e-05 123:   0.00349144172049764 124:   0.000393793867567723 125:   0.000643677761923412 126:   0.000104142413948941 127:   0.000174863218287601 128:
      0.0147688960811977 129:   0.00527332164888017 130:   0.000181076249414555 131:   0.0014327030639978 132:   0.00224940032378129 133:   9.57556910592051e-05 134:   0.000177737839665965 135:   9.6643038487228e-05 136:   0.00238080710329135 137:
      0.000313697693309199 138:   0.00584373037835742 139:   7.39539432229307e-05 140:   5.61803066301089e-05 141:   0.00889991562718008 142:   0.00353907563632441 143:   0.000261874578318196 144:   0.00367789042351576 145:   0.0013628731634665 146:
      0.0472168352789147 147:   5.13684119602324e-05 148:   1.62563061548312e-05 149:   0.000774080968745045 150:   2.91675414749147e-05 151:   0.000998282292176137 152:   0.000277624338562502 153:   0.00451525713103029 154:   0.00299700820512059 155:
      0.00150414860787534 156:   0.0012965743570287 157:   7.42191830632657e-05 158:   2.69597785813309e-06 159:   0.00067475403550923 160:   0.000145507036776428 161:   0.00034485227528041 162:   0.00230809321219072 163:   0.0363156430975114 164:
      0.000925904503011862 165:   0.000573893991457866 166:   0.00164670013971995 167:   0.00527923895565897 168:   0.0306528149877042 169:   0.00662742894835711 170:   0.00184806533312547 171:   0.00305831693264844 172:   0.00862001771345276 173:
      9.64733959082152e-05 174:   0.0184029660528 175:   0.00715556312732788 176:   0.00246812903179433 177:   0.00316531139257279 178:   0.00387047230353543 179:   0.00103993926405791 180:   0.000757781619877794 181:   0.00149369025473511 182:
      0.000575435687087654 183:   0.000331638300175998 184:   0.00162195100409655 185:   0.0142254759497093 186:   0.0247492972281852 187:   0.00180695796640826 188:   0.00822331957272751 189:   0.00306031441093705 190:   0.000334897470908548 191:
      0.0466394610364021 192:   0.0196075590478781 193:   0.011104875045564 194:   0.0113302289102217 195:   0.0616757916521861 196:   0.0583894448288548 197:   0.0137110712878143 198:   0.0229050468326967 199:   0.00546639829160562 200:
      0.0223723646661068 201:   0.00330496449778148 202:   0.0231308346066991 203:   0.00903187244171285
```

```
plot(cooks.distance(step.model), pch=23, bg='orange', cex=2, ylab="Cook's distance")
```

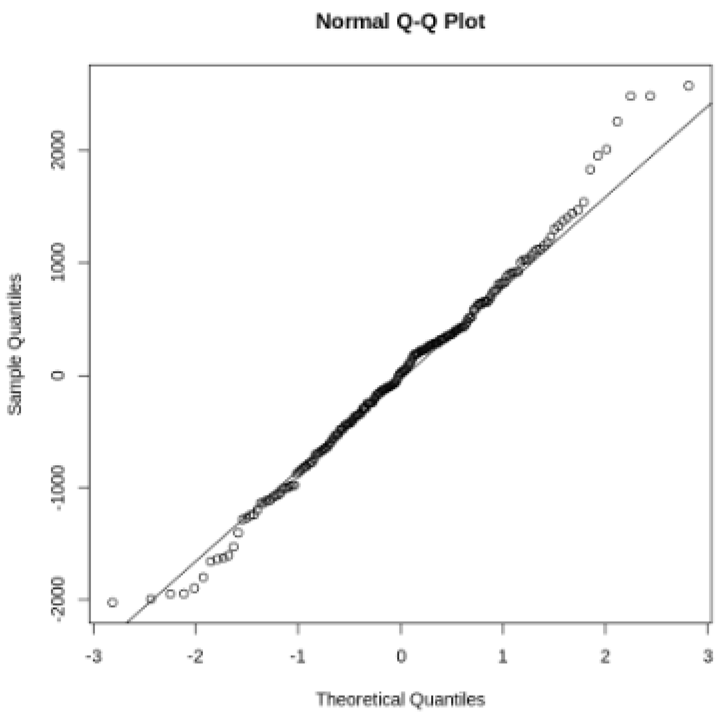


6. Quantile-Quantile (Q-Q) Plot

This is a method used to check whether the residuals of a Linear Regression model follow a normal distribution or not. It uses a Quantile-Quantile plot (QQ plot) to make predictions.

The test of residuals using a QQ plot is a method that creates a graph comparing the actual residuals with the expected values from a normal distribution. If the residuals follow a normal distribution, the data points will appear as a diagonal line with a consistent slope in the QQ plot and will be arranged in a straight line. The presence of data points on the straight line indicates a suitable distribution matching the normal distribution. However, if the residuals do not follow a normal distribution, there will be data points that deviate from the diagonal line or exhibit non-uniform clustering.

```
qqnorm(res)
qqline(res)
```



7. **Shapiro-Wilk Test**

Upon evaluating the p-value, which is greater than the significance level of 0.05, we cannot reject the null hypothesis H0. The null hypothesis states that the data follows a normal distribution.

- W (statistic) = 0.99047
- H0: Data follows a normal distribution
- H1: Data does not follow a normal distribution
- p-value = 0.2003

Therefore, we do not have sufficient statistical evidence to reject the hypothesis that the data follows a normal distribution. Hence, we cannot conclude that the data does not follow a normal distribution.

```
shapiro.test(res)
```

```
	Shapiro-Wilk normality test

data:  res
W = 0.99047, p-value = 0.2003
```

8. **Testing VIF (Variance Inflation Factor)**

The hypothesis used in testing VIF is:

- H0: There are no significant multicollinearity factors in the Linear Regression model
- H1: There are significant multicollinearity factors in the Linear Regression model

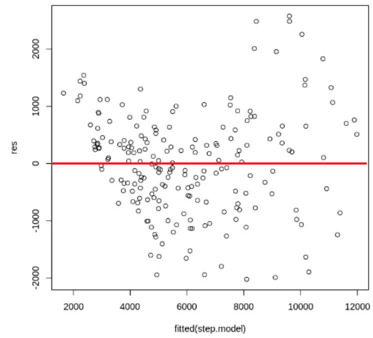
Therefore, when the VIF value of each independent variable is less than 10, indicating no significant multicollinearity issues in the Linear Regression model.

Thus, we can conclude that there are no problems regarding the interrelation between variables in the Linear Regression model, and this model can be trusted.

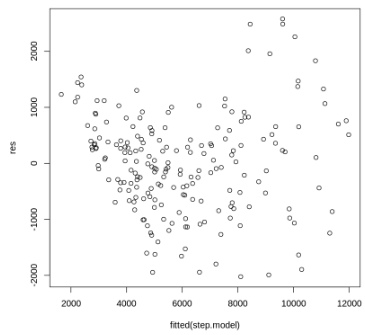
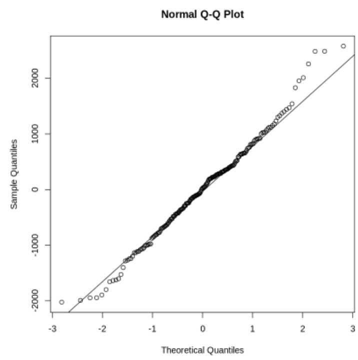
```
install.packages("car")
library(car)
vif(step.model)
```

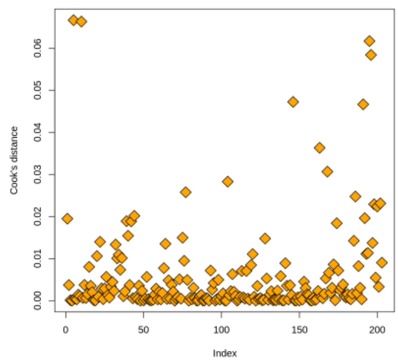
```
Installing package into '/usr/local/lib/R/site-library'
(as 'lib' is unspecified)

cut:	1.15260356973037 color:	1.15289761167837 clarity:	1.14244322826251 depth:	1.13794821510814 y:	1.27898539846846
```

Assumptions checking

<table>
<tr>
<td>1. Linear Model</td>
<td>According to the plot of residuals, they bounce randomly to ground zero. It is sufficient to suggest that the regression function is linear.</td>
<td>
</td>
</tr>
<tr>
<td>2. Independence</td>
<td>The Durbin-Watson test yielded a statistic of 1.549828, if the Durbin-Watson statistic is close to 2, typically falling between 1.5 and 2.5, it suggests that the residuals are independent. In our analysis, the Durbin-Watson statistic is 1.549828, falling within the range of 1.5 to 2.5, indicating that the independent variables used in the test do not exhibit internal correlation.</td>
<td>
<pre>
install.packages("car")
library(car)
durbinWatsonTest(step.model)

Installing package into '/usr/local/lib/R/site-library'
(as 'lib' is unspecified)

 lag Autocorrelation D-W Statistic p-value
   1   0.2184589   1.549828   0.002
 Alternative hypothesis: rho != 0
</pre>
</td>
</tr>
<tr>
<td>3. Homogeneity of variance</td>
<td>The random pattern in the residuals vs. fitted values plot shows the random errors have constant variance.</td>
<td>
</td>
</tr>
<tr>
<td>4. Normality</td>
<td>The points on the Q-Q plot fall about the straight line. That means the random errors follow a normal distribution.</td>
<td>
</td>
</tr>
</table>

<table>
<tr>
<td>5. Outliers</td>
<td>There are no outliers because the Cook's distance (Di) is less than 0.5.</td>
<td>
</td>
</tr>
<tr>
<td>6. Multicollinearity</td>
<td>A VIF < 5 explains that collinearity does not affect the variable.</td>
<td>install.packages("car")
library(car)
vif(step.model)

Installing package into '/usr/local/lib/R/site-library'
(as 'lib' is unspecified)

cut: 1.15260356973037 color: 1.15289761167837 clarity: 1.14244322826251 depth: 1.13794821510814 y: 1.27898539846846</td>
</tr>
</table>

Summary and Analysis

The report on analyzing and predicting diamond prices using regression emphasizes the significance of regression in forecasting data behavior across various domains. This analysis focuses on creating a statistical model to predict diamond prices using diverse characteristics such as weight, cut quality, color, clarity, and size. The research aims to develop an appropriate statistical model to enhance understanding and accurate prediction of diamond prices, highlighting the importance of regression in solving mathematical problems and aiding decision-making in the diamond industry.

BLUE (Best Linear Unbiased Estimators) Analysis:

- Unbiasedness: The model meets the assumption of unbiasedness as the regression coefficients are estimated without bias.
- Linearity: The residuals scatter randomly around zero, indicating that the regression function is linear.
- Efficiency: Given that the model satisfies the assumptions, the estimators are efficient and achieve the smallest variance among all unbiased estimators.

ANOVA & t-test Assumptions:

- Independence: The Durbin-Watson statistic falling within the range of 1.5 to 2.5 indicates that the residuals are independent, meeting the assumption for ANOVA and t-tests.
- Homogeneity of Variance: The variance of the residual value compared to the filled plot value shows that random error has constant variance.
- Normality: Normal distribution of residual values can be observed in the Q-Q plot, indicating that the random error values follow a normal distribution.

Model Effectiveness:

- Outliers: The absence of outliers, indicated by Cook's distance being less than 0.5, ensures that influential data points are not skewing the results.
- Multicollinearity: If the VIF value is less than 5, it indicates no issue with multicollinearity, suggesting that the variables do not have high interrelationships.

The results of the analysis and prediction of diamond prices using regression indicate that the 5 independent variables (cut, color, clarity, depth, y(width)) significantly influence diamond prices. The linear regression model developed has a strong ability to explain the data, with an adjusted R-squared value of approximately 87.56%. Additionally, the p-value of the model is less than the standard statistical significance level of 0.05, allowing the rejection of the null hypothesis (H0: independent variables have no effect on diamond prices) and acceptance of the alternative hypothesis (H1: at least one independent variable affects diamond prices), suggesting a response of diamond prices to the input independent variables.