Diamond Information: Plotting and Data Transformation

MINH CHAU

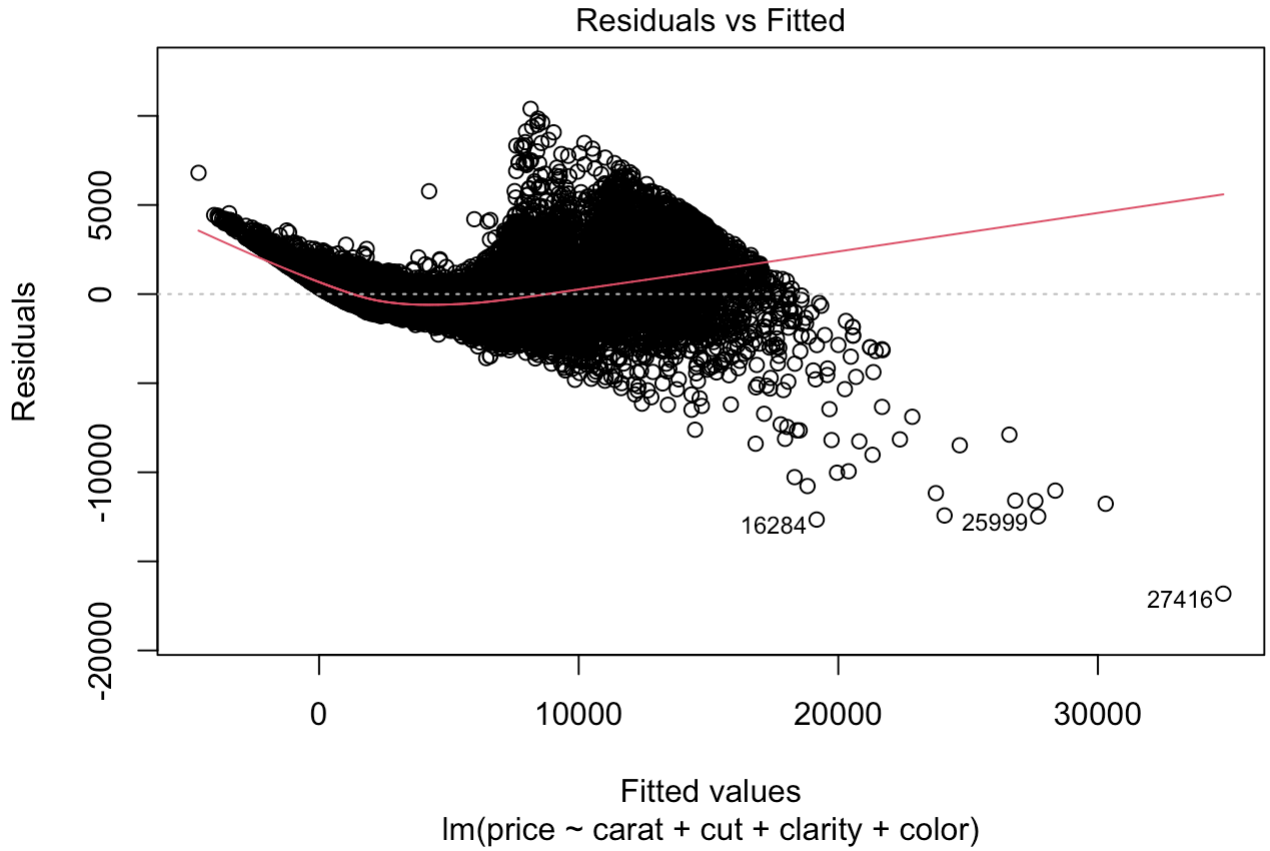
```
# load required packages
library(dplyr)
library(tidyr)
library(car)
library(ggplot2)
```

Investigate the assumptions in linear regression and use plots which are important diagnostic tools to check whether the assumptions underlying linear regression are met. Here, you will assess the validity conditions by examining the residuals and explore if applying transformations to variables can achieve linearity.

- Dataset: `Dia.csv`
- There are 8 attributes in each case of the dataset.
- The description of the variables are :
    - row: row identifier
    - carat: weight of the diamond
    - cut: quality of the cut
    - color: diamond colour, from D (best) to J (worst)
    - clarity: a measurement of how clear the diamond is (I1 (worst), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (best))
    - depth: total depth percentage
    - table: width of top of diamond relative to widest point
    - price: in US dollars

(2a) First, fit the multivariate regression model with *price* as the response and the predictors: *carat, cut, clarity* and *color*. Next, plot the Residuals vs. Fitted plot for the fitted multivariate regression model. Interpret the diagnostic plot by relating it to the assumptions of regression analysis.

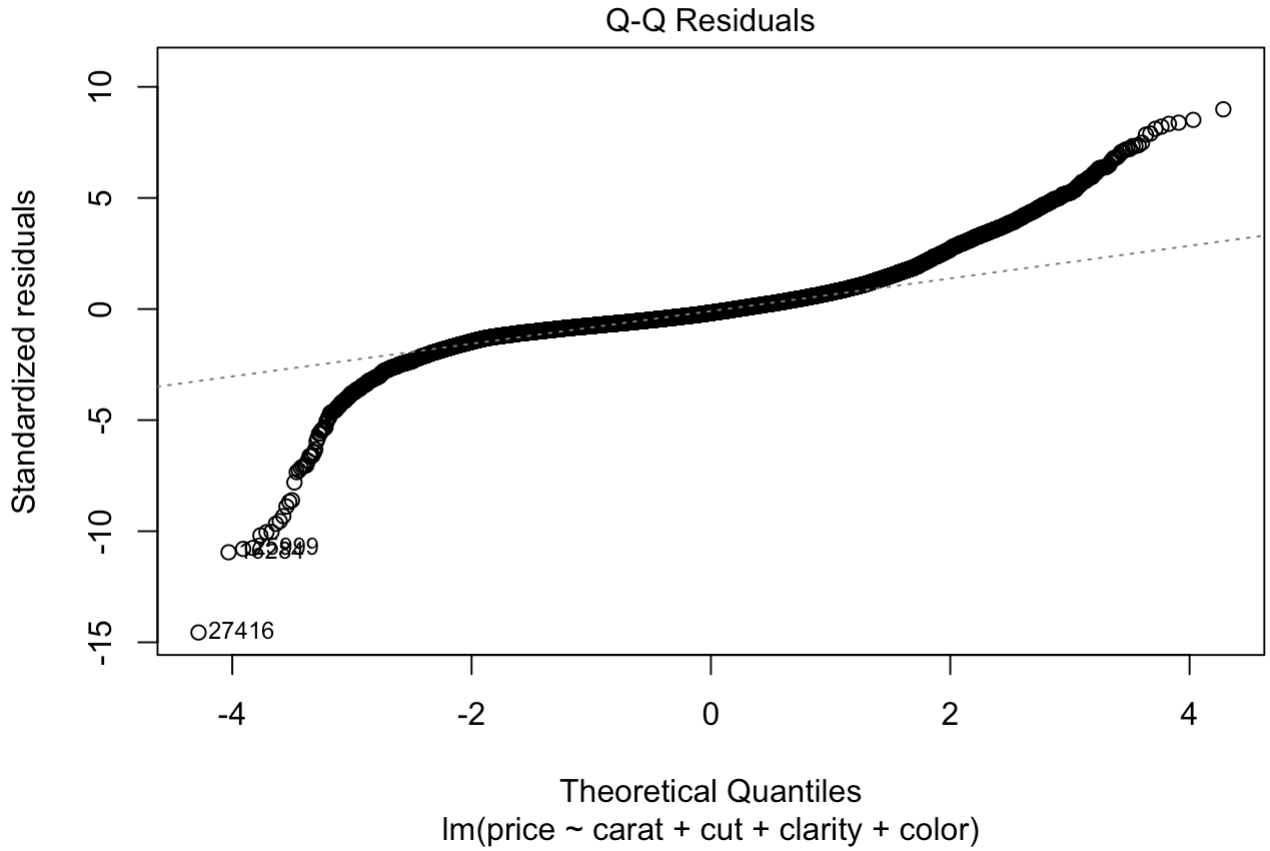
```
diamonds <- read.csv("/Users/minhchau/Downloads/Dia.csv")
diamonds.lm <- lm(price ~ carat +cut + clarity + color, data=diamonds)
plot(diamonds.lm,1)
```



(2b) Plot the Normal Q-Q plot. What is its purpose? Interpret the plot and relate it to the assumptions of regressession analysis.

Plot the scatter plots of carat, cut, clarity, color `price` . Consider using the scatter plot matrix or pairs() function. What type of relationship exists between the variables? Is it linear?

```
plot(diamonds.lm,2)
```

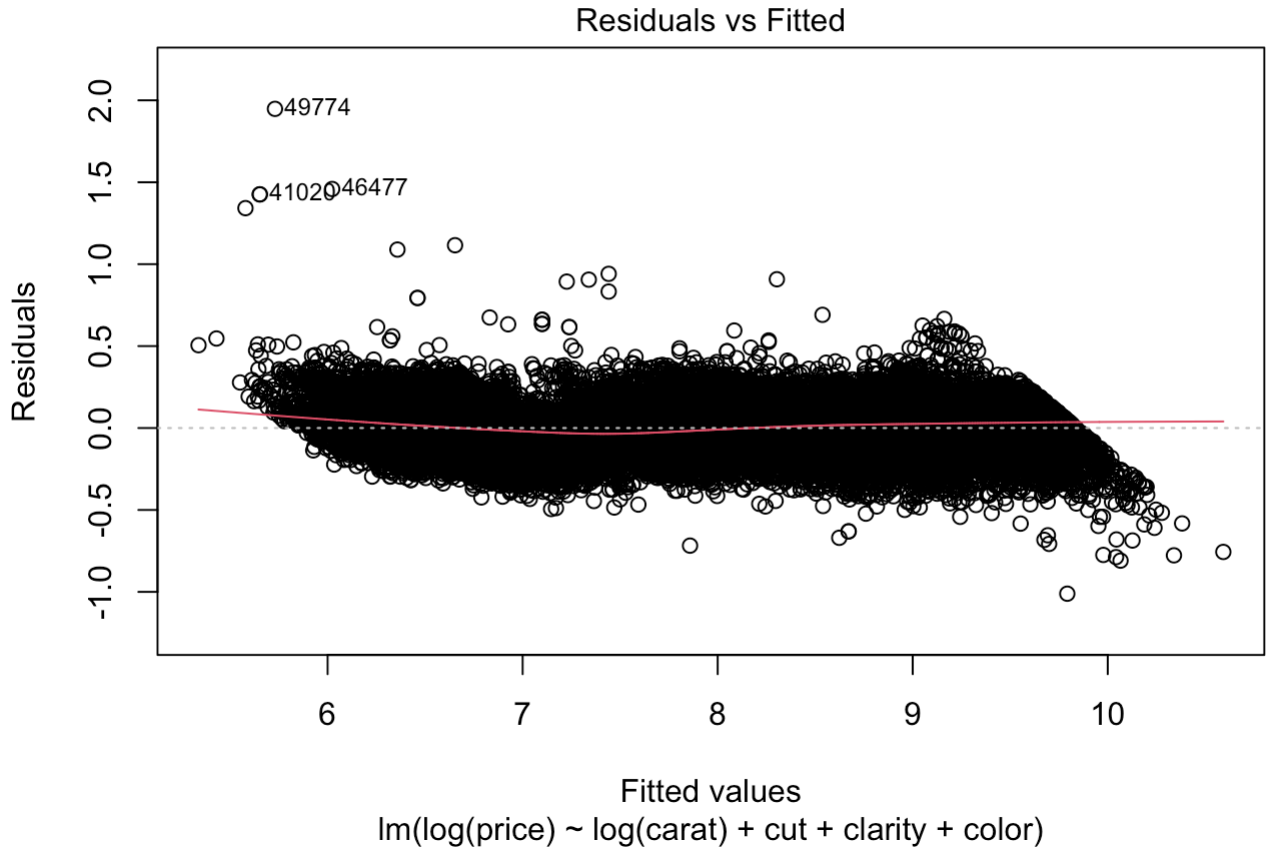


Perform a ***log-transformation*** on both the price and carat variables, such as log(price) and log(carat). In regression modeling, the natural log is preferred to maintain consistency with standard practices and interpretations.

(2c) Plot the Residuals vs. Fitted plot for the transformed multivariate regression model. Interpret the plot and how has the log-transformation changed the residuals?

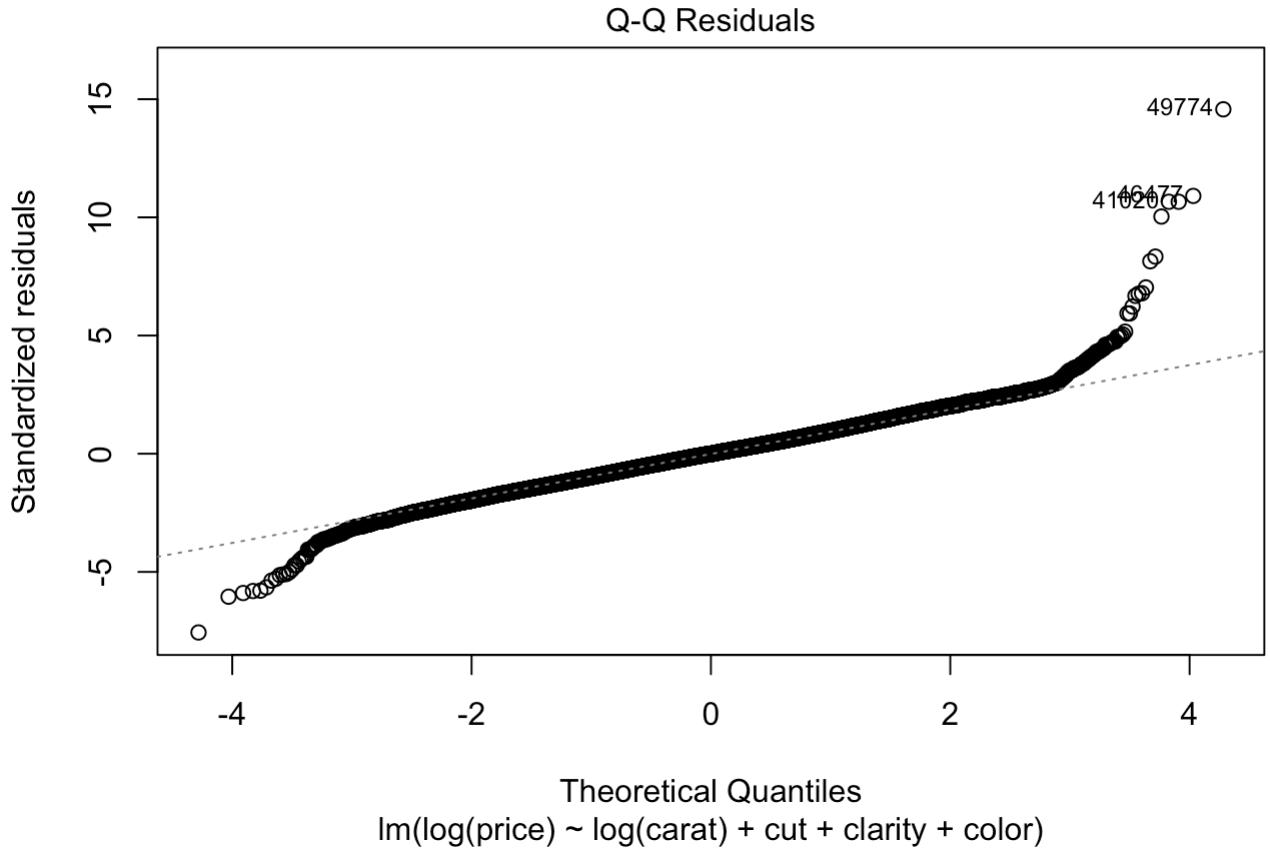
```
diamonds.lm2 <- lm(log(price) ~ log(carat) + cut + clarity +color, data=diamonds)

plot(diamonds.lm2,1)
```



(2d) Plot the Normal Q-Q for the transformed multivariate regression model. Interpret the plot and how has the log-transformation changed this plot?

```
plot(diamonds.lm2,2)
```


Q-Q Residuals
Standardized residuals
15
10
5
0
-5
49774
46477
41020
-4
-2
0
2
4
Theoretical Quantiles
lm(log(price) ~ log(carat) + cut + clarity + color)