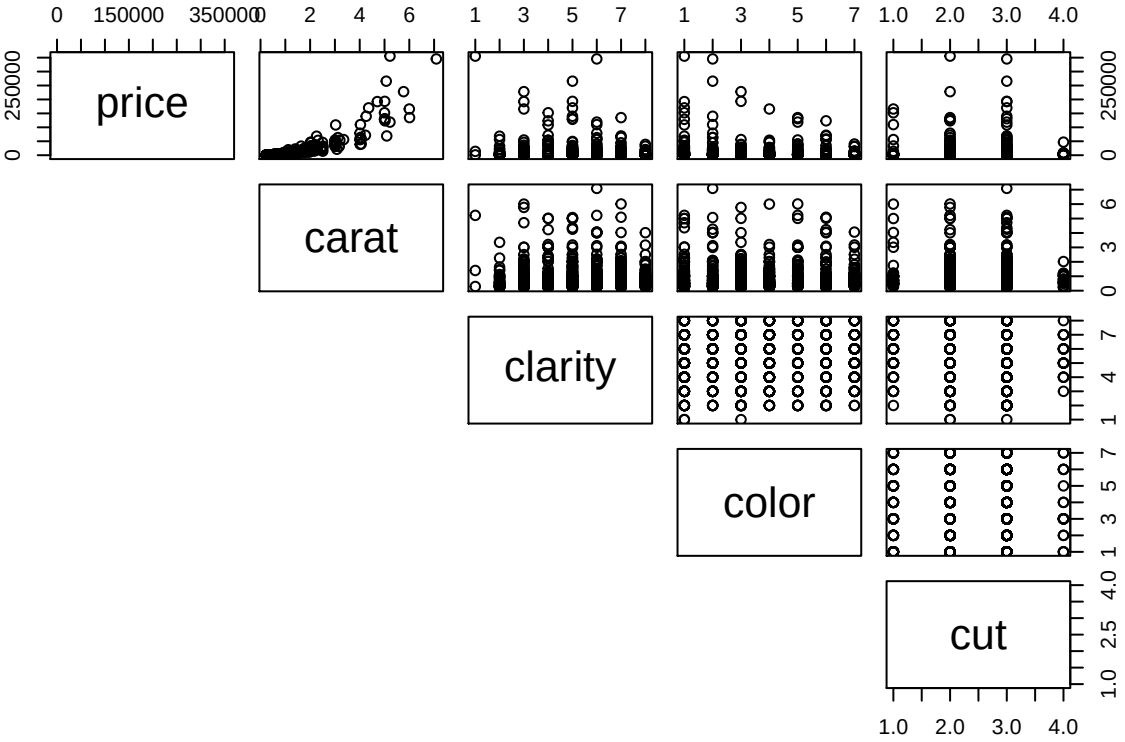
STAT 6021: Project 1

Greg Madden, Maxwell Levinson, Andrew Setaro, and Trey Hamilton

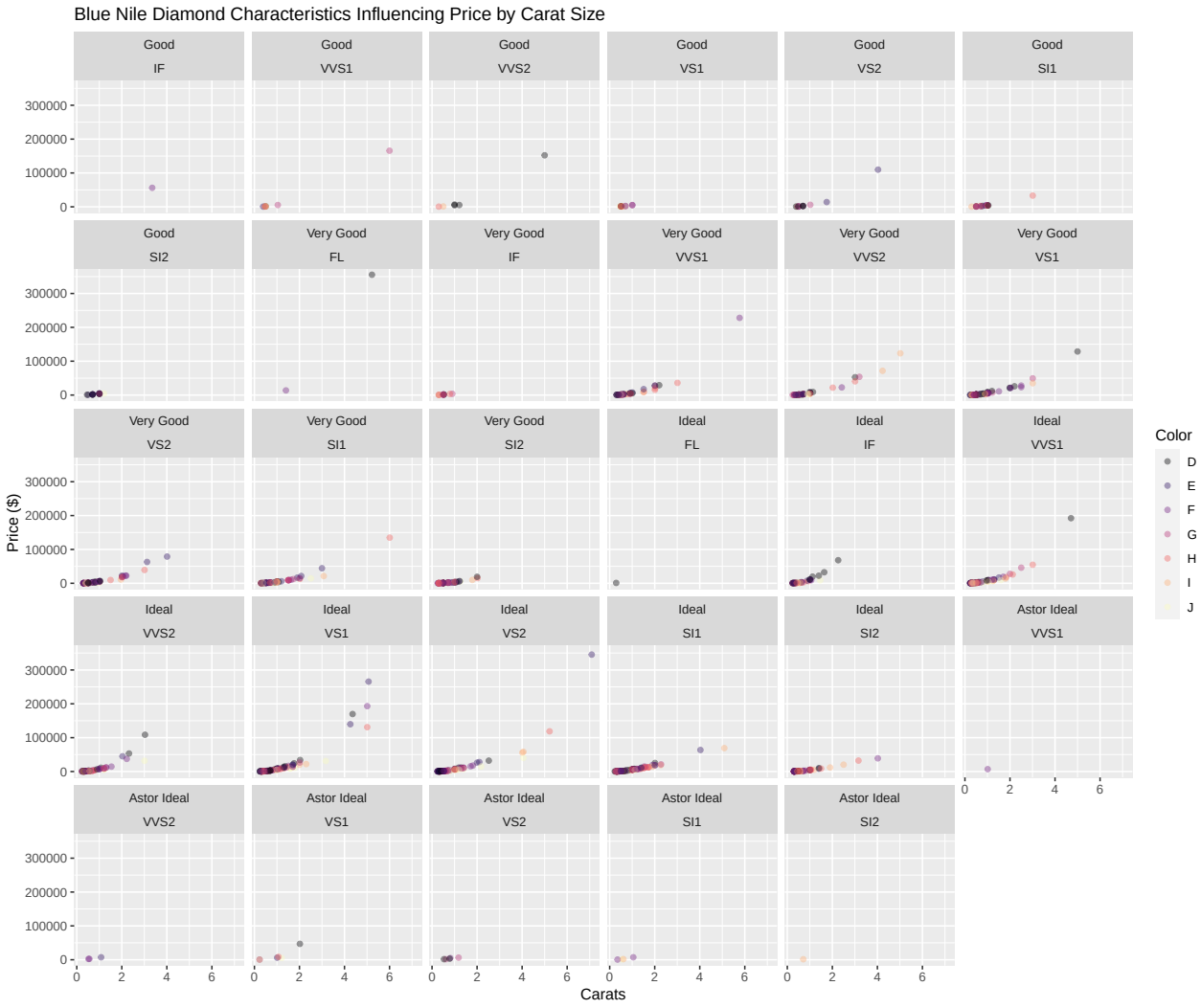
3/3/2022

We have been approached by Blue Nile to perform the following tasks:

1. Use data visualizations to explore how price is related to the other variables (carat, clarity, color, cut), as well as how the other variables may relate to each other.



Note that the x-axis above corresponds to increasing desireability of the factored categorical variables: clarity, color, and cut. In the above scatterplot matrix how price appears to have the clearest linear relationship with carats.



In the above plot, you can see that certain cut/clarity combinations (e.g., Good cut and VVS2 clarity) have different slopes in terms of their price ~ carat relationships. For example, ideal cut with FL (Flawless) clarity appears to have a higher price per additional carat size than the Ideal Cut with SI2 (Slightly Included). In addition, more desireable colors (i.e.D-F) appear to cluster at the lower price ranges.

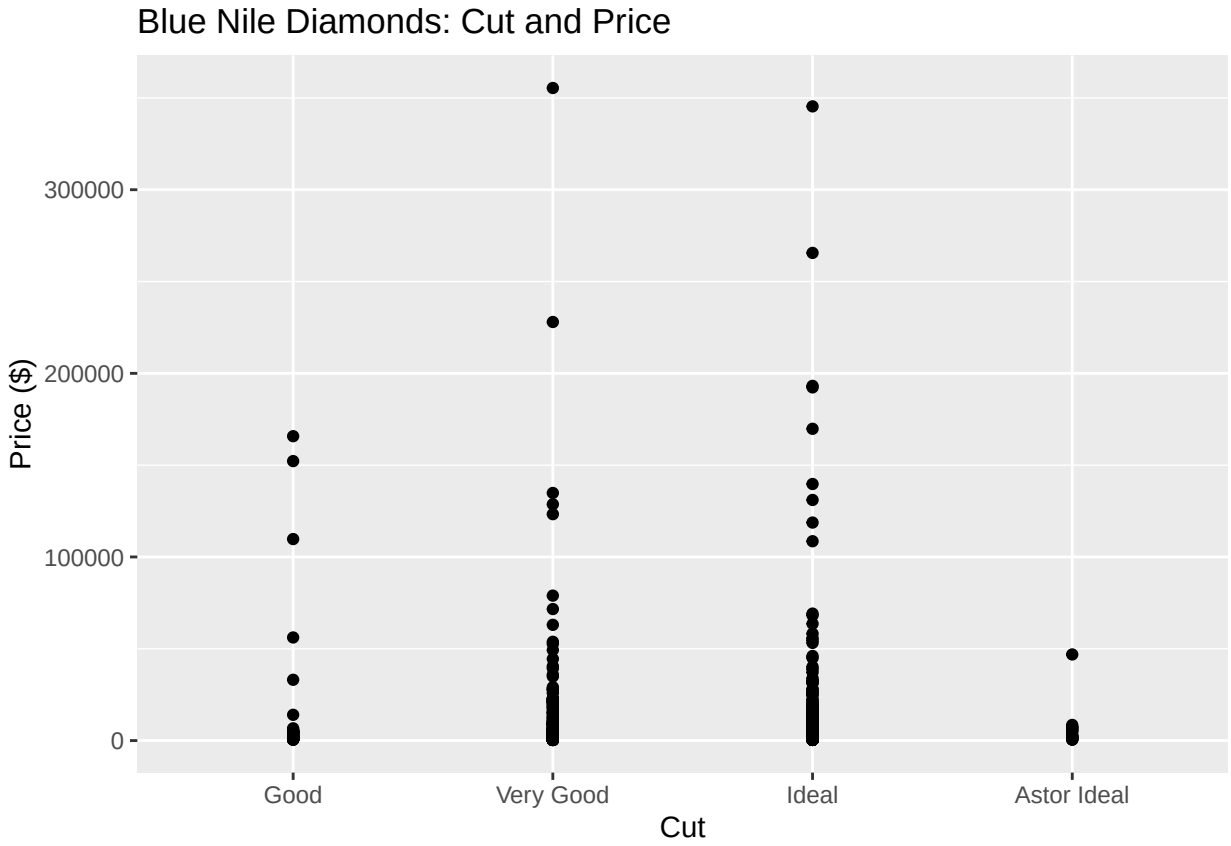
Address the various claims on the diamond education page on Blue Nile.

- Cut: https://www.bluenile.com/education/diamonds/cut +"A diamond's cut refers to how well-proportioned the dimensions of a diamond are, and how these surfaces, or facets, are positioned to create sparkle and brilliance. For example, what is the ratio of the diamond's diameter in comparison to its depth? These small, yet essential, factors determine the diamond's beauty and price."

Assertion above is that better cuts correlate with higher price. Let's check the scatterplot to see if that bears out in the data:

Increasing quality of diamond cut does not seem to have a linear relationship with price, contrary to the Blue Nile's claim.

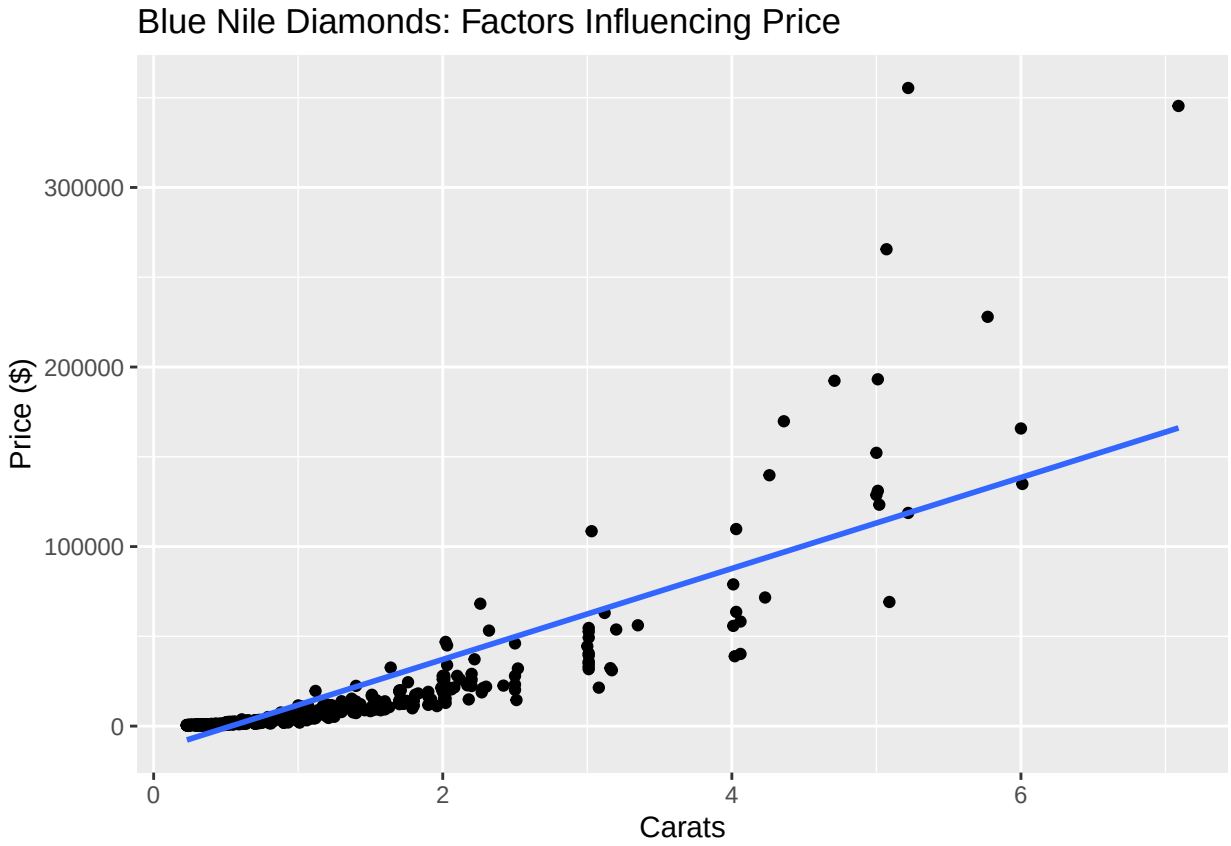
```
## `geom_smooth()` using formula 'y ~ x'
```



2. Fit an appropriate simple linear regression for price against carat.

First a scatterplot

```
## `geom_smooth()` using formula 'y ~ x'
```

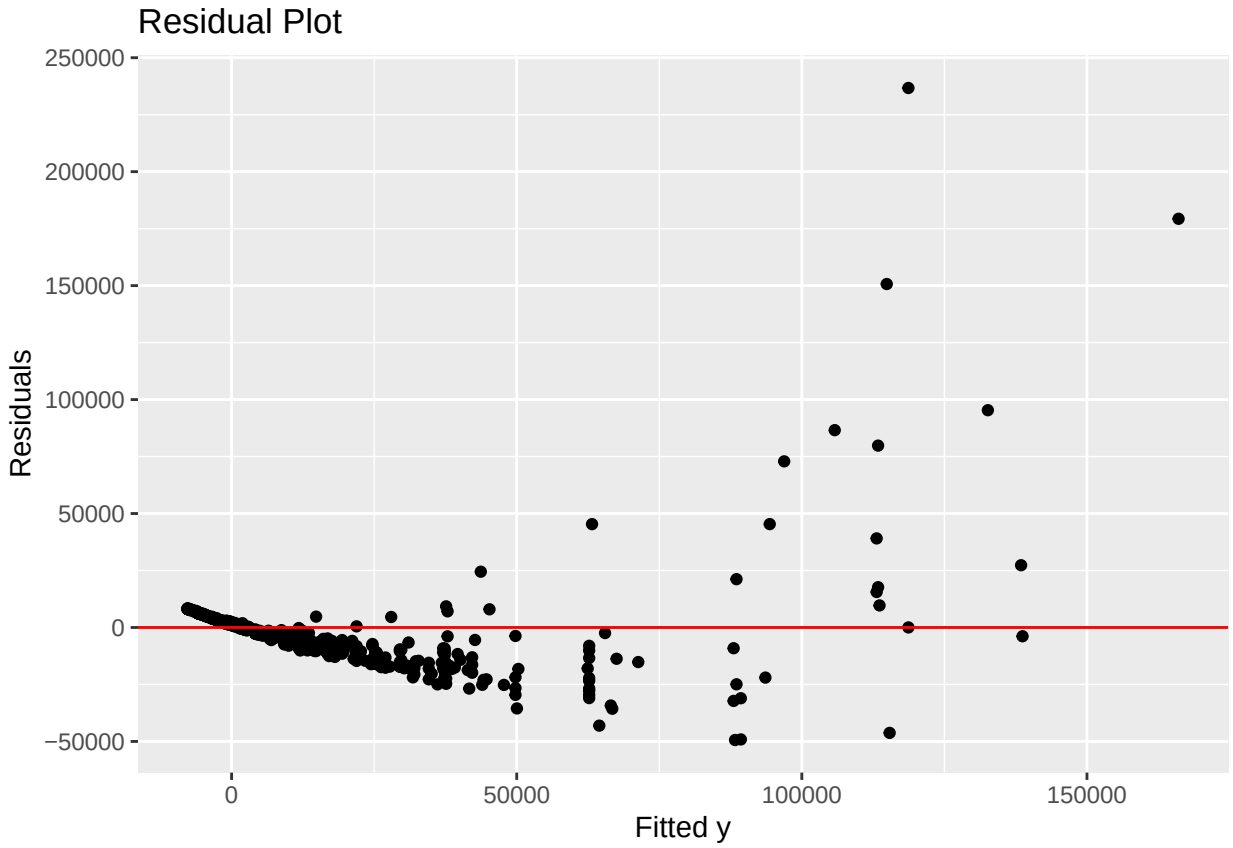


Model Summary

```
##
## Call:
## lm(formula = price ~ carat, data = Data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -49375  -5048   1867   4965 236711
##
## Coefficients:
##             Estimate Std. Error t value            Pr(>|t|)
## (Intercept) -13550.9      559.7  -24.21 <0.0000000000000002 ***
## carat        25333.9      494.4   51.24 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13560 on 1212 degrees of freedom
## Multiple R-squared:  0.6842, Adjusted R-squared:  0.6839
## F-statistic:  2625 on 1 and 1212 DF,  p-value: < 0.00000000000000022
```

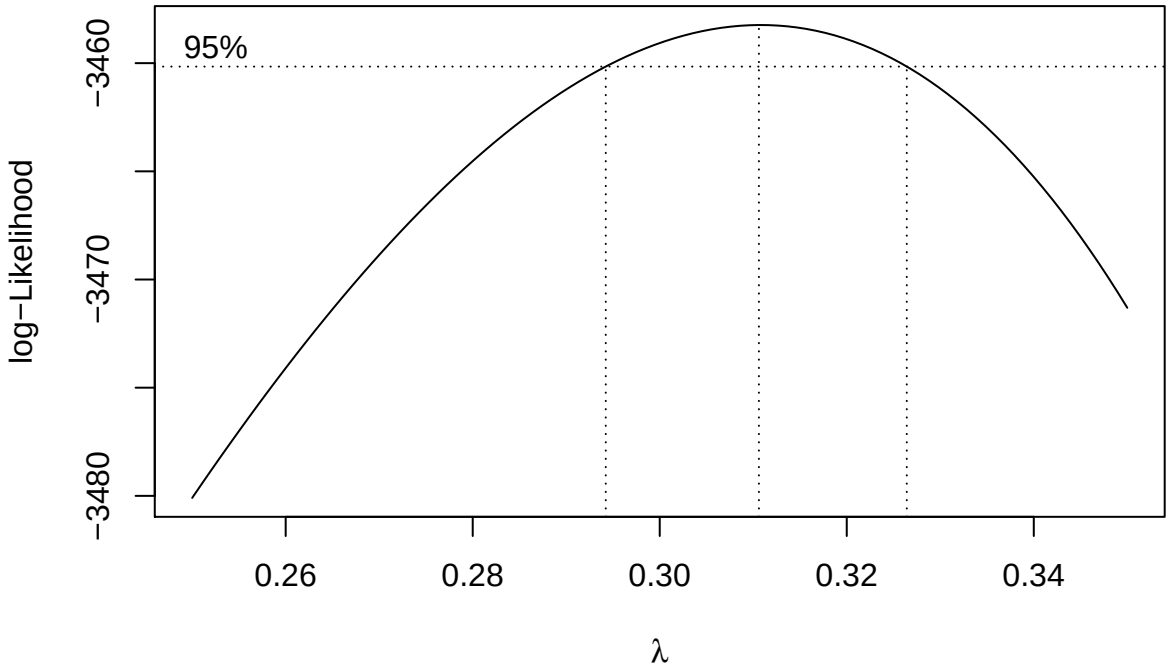
Plotting the residuals.

Constant variance and mean of error = 0 assumptions do not appear to be met.



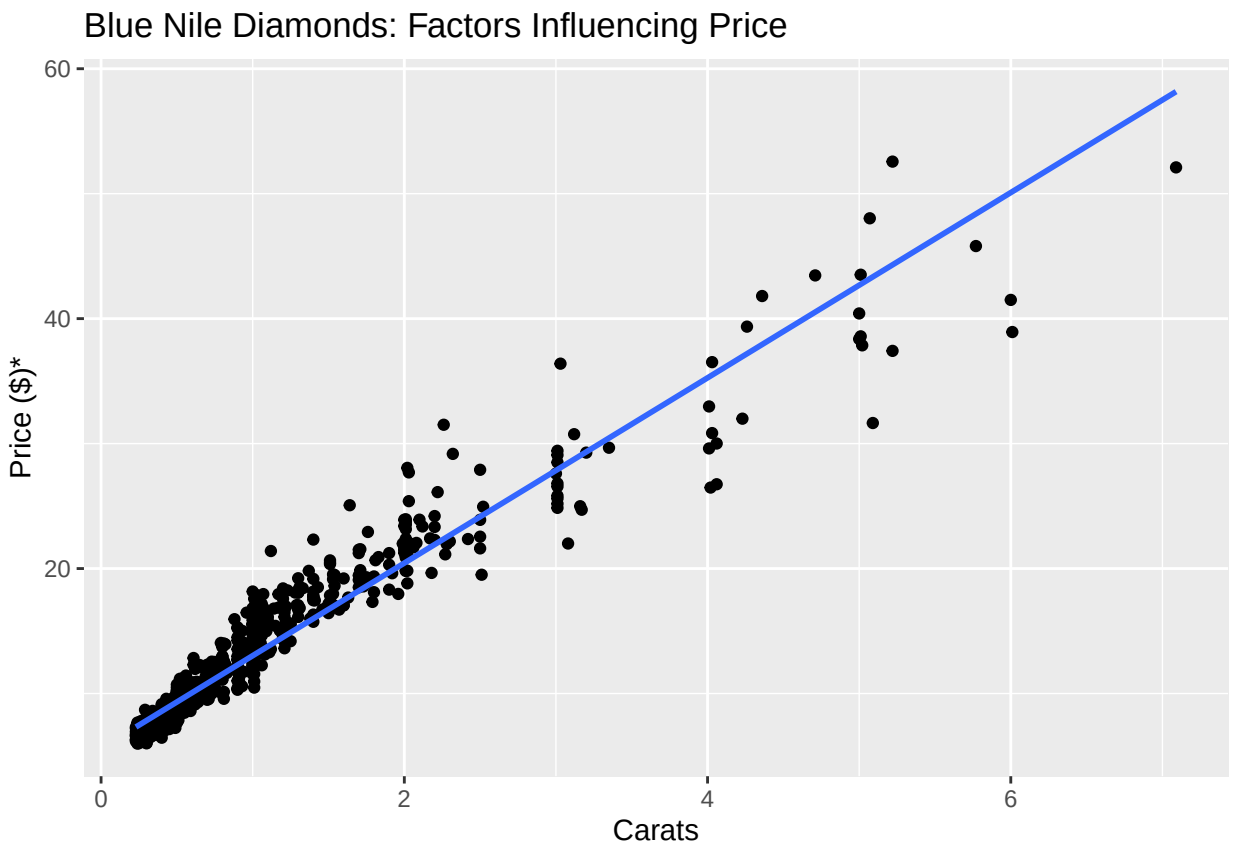
Variance is definitely not contant so attempting to transform y first. Will start with boxcox plot to see what the optimal lambda may be.

Looked like a lambda of 0.31 would be appropriate.



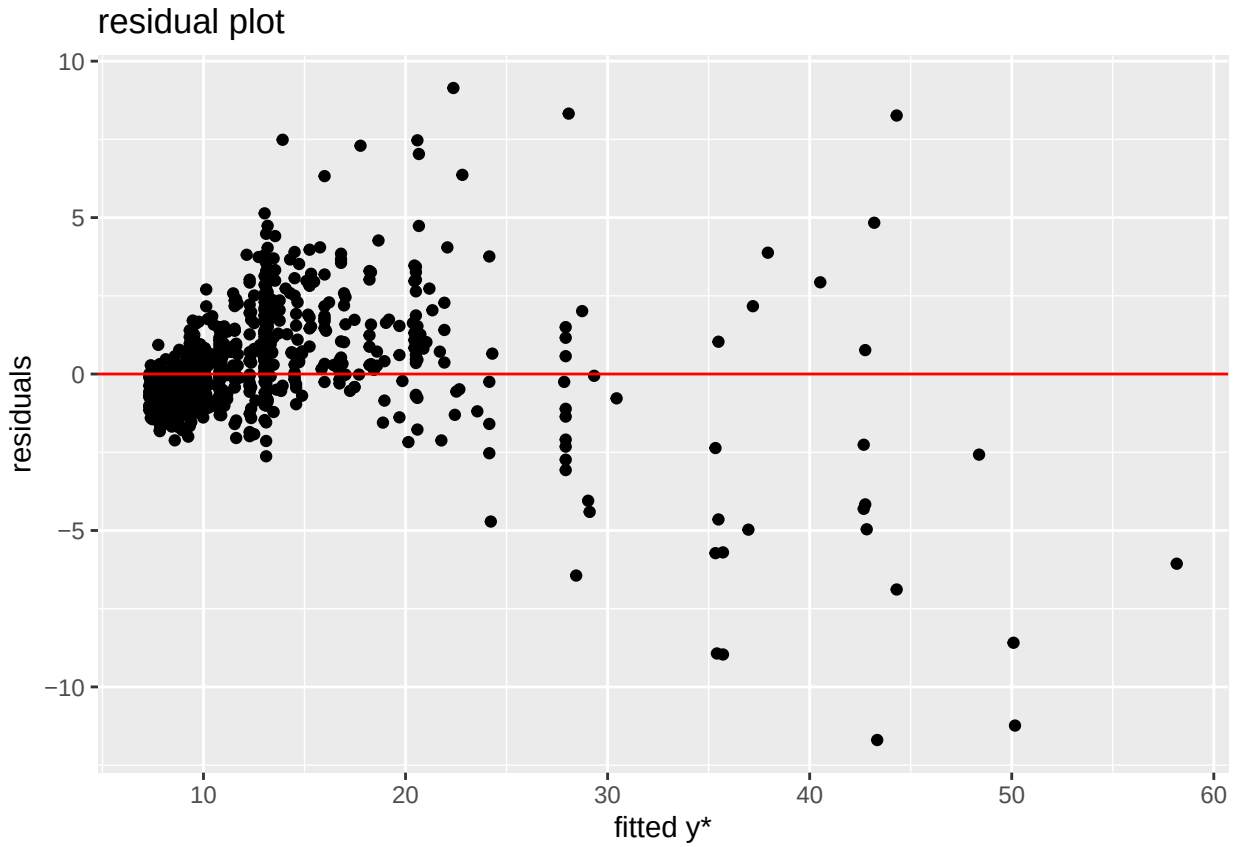
Replotting the scatter plot with the transformed y variable.

```
## `geom_smooth()` using formula 'y ~ x'
```

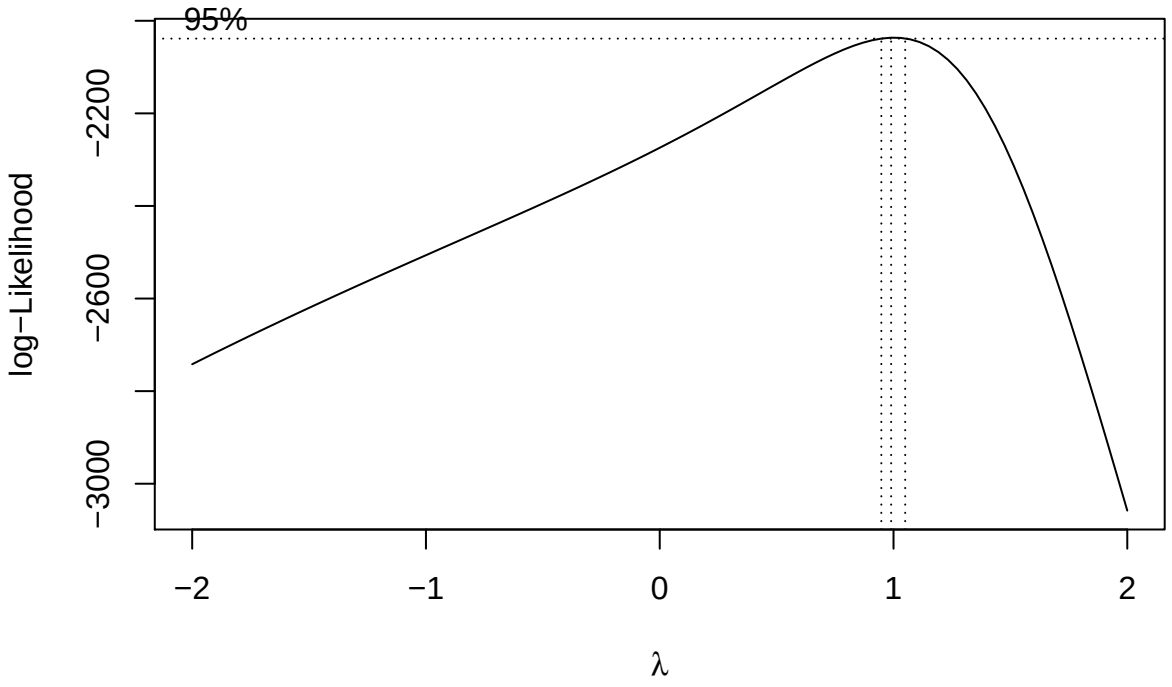


New residual plot:

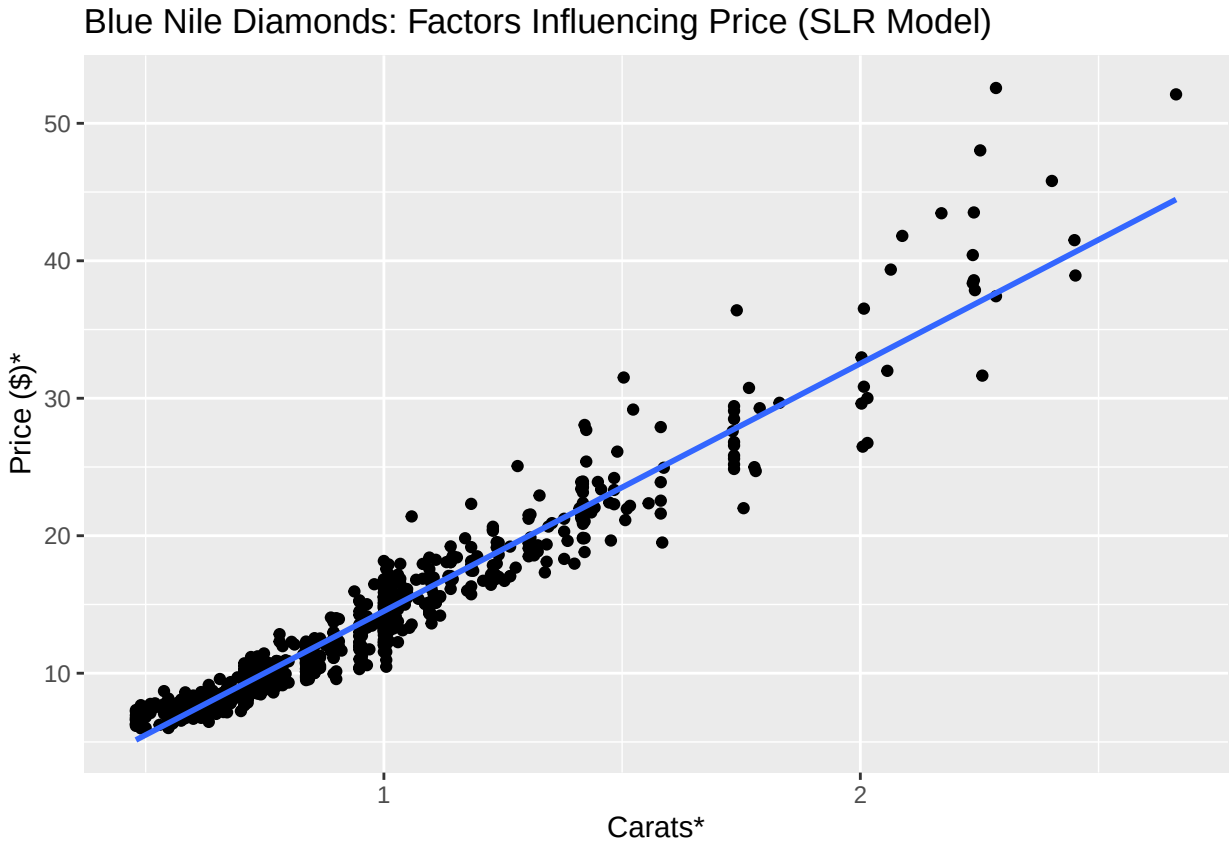
variance looks better but mean of errors still not equal to zero over x so will attempt to transform the x variable. Given the curved appearance, will try a square root transformation.



Confirming that the box cox plot looks better. Now I see that confidence interval includes 1. Next step is to consider whether or not to transform x.

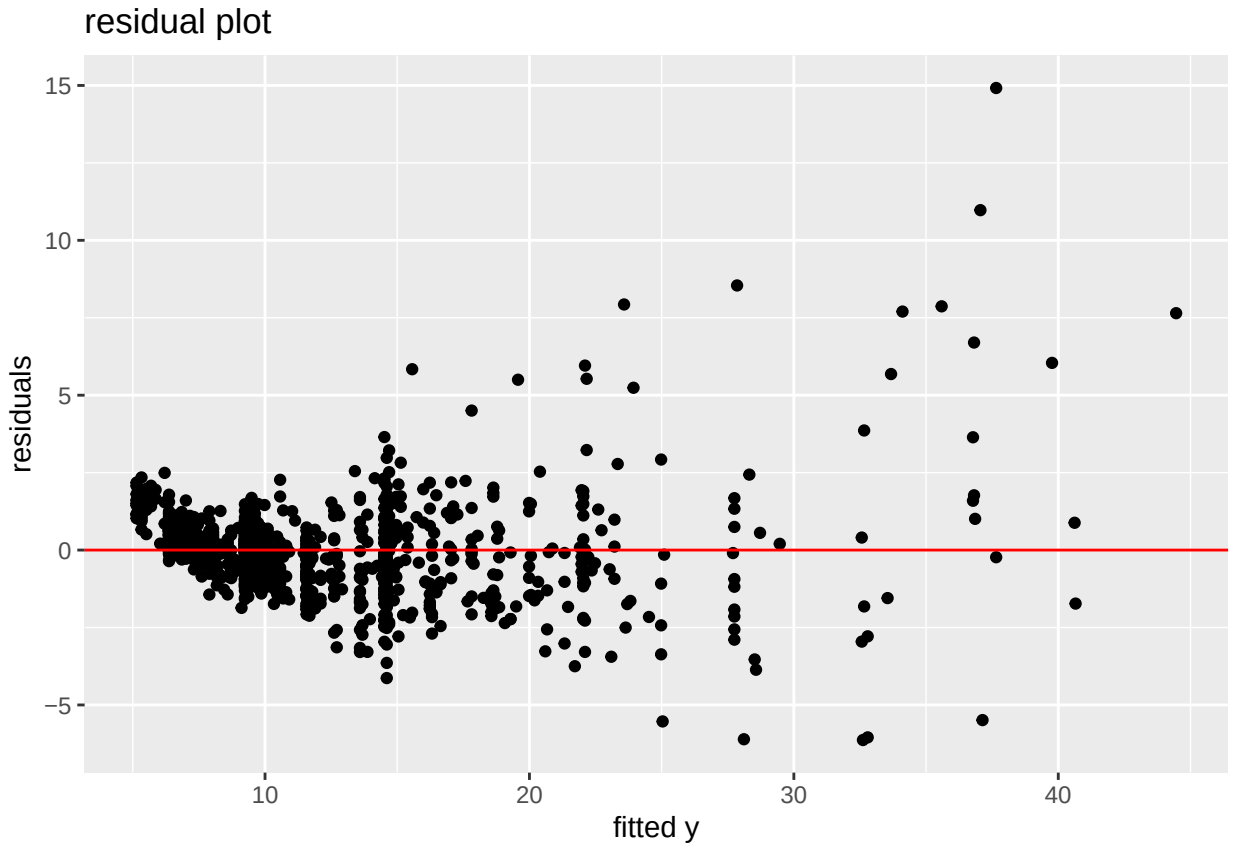



```
## `geom_smooth()` using formula 'y ~ x'
```



Fitting new model and creating yet another residual plot.

Not a perfect fit but overall improved and adequate for prediction.

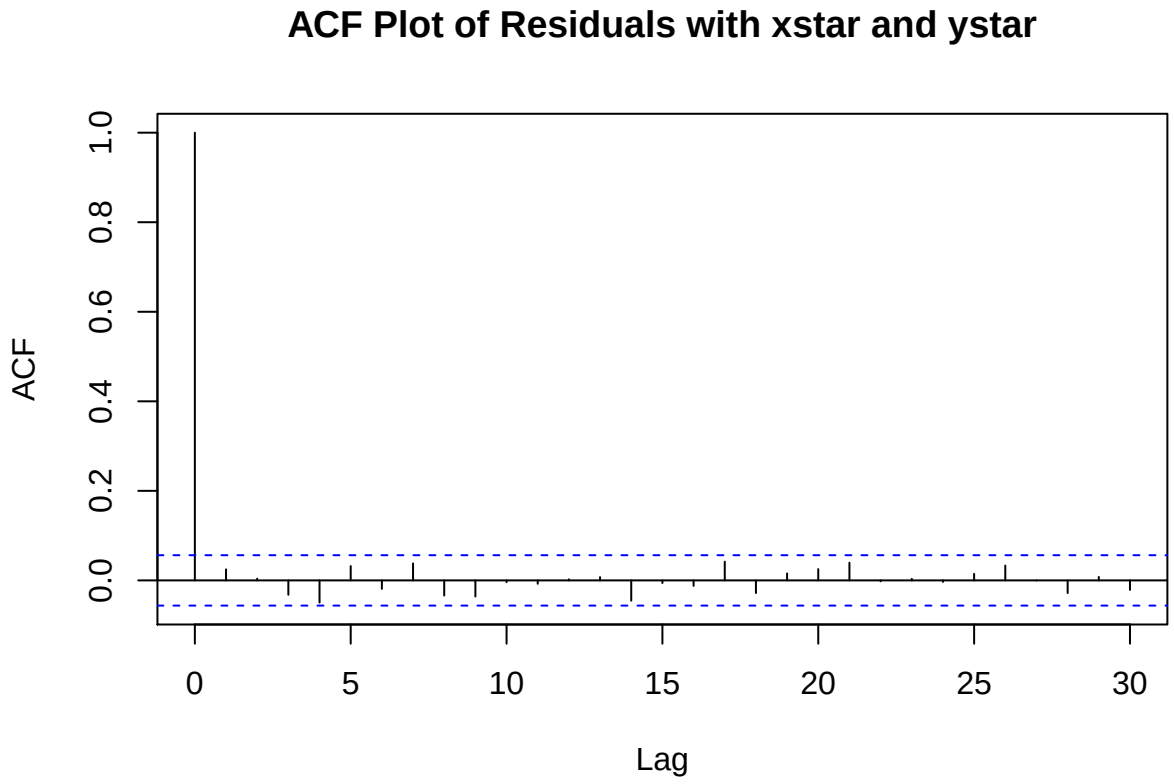


Summarizing the model below.

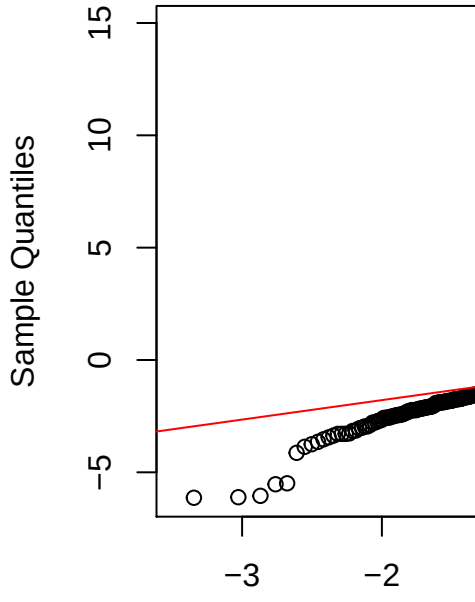
```
##
## Call:
## lm(formula = ystar ~ xstar, data = Data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.1320 -0.6377  0.0373  0.5315 14.9172
##
## Coefficients:
##             Estimate Std. Error t value            Pr(>|t|)
## (Intercept)  -3.4936     0.1137  -30.73 <0.0000000000000002 ***
## xstar        18.0085     0.1261  142.87 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.436 on 1212 degrees of freedom
## Multiple R-squared:  0.9439, Adjusted R-squared:  0.9439
## F-statistic: 2.041e+04 on 1 and 1212 DF,  p-value: < 0.00000000000000022
```

ACF Plot

Errors do not appear correlated to each other.



QQ



Plot ***Normality assumption is not met but this may be the least important.***