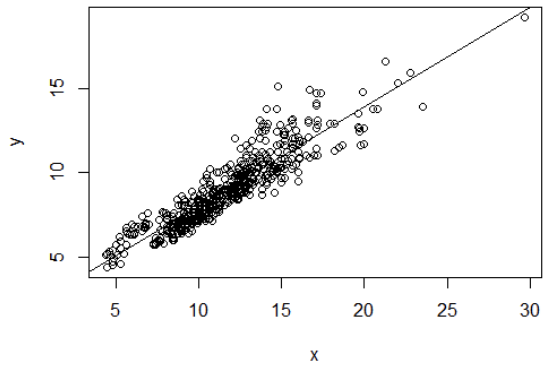
Canadian fuel efficiency

Determining if City fuel efficiency can be used to predict Highway Fuel

How many variables and how many observations does your dataset have?

- 15 Variables
- 602 Observations
- fuel | 602 obs. of 15 variables

Does it appear that there is a relationship between X and Y? Is that relationship linear?



- The scatter plot seems mostly linear, however, most of the points are clustered in the beginning, making it less linear. There is also one outlier at the end of the plot.

What is the form of the least-squares regression line?
Interpret the y-intercept and slope, if appropriate.

- Least Squares regression line: y=1.13+0.59x
- Interpretation for slope: For every 1 unit increase in City Fuel efficiency, Highway fuel efficiency is expected to increase by 0.59 units.
- Interpretation for y-intercept: Since City fuel efficiency was not sampled at 0, it is not appropriate to interpret the y-intercept

Conduct a hypothesis test of the slope to determine if X is useful for predicting Y.

Ho: $\beta_1 = 0$

Ha: $\beta_1 \neq 0$

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.13616    0.13122   16.28   <2e-16 ***
x            0.58798    0.01053   55.83   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

$t\ statistic: t = 55.83\,; \quad p\ value = 2 \times 10^{-16}$

Reject Ho. There is sufficient evidence at $\alpha = 0.05$ to indicate that β_1 is significant and therefore, useful to predict highway fuel efficiency.

Interpret the coefficient and coefficient of determination.

```
> cor(fuel_city, fuel_highway)
[1] 0.9119344
> cor(fuel_city, fuel_highway)^2
[1] 0.8316243
```

$R = 0.911934$ indicates a positive and strong linear relationship between highway fuel efficiency (y) and city fuel efficiency (x).

$R^2 = 0.8316243$ indicates that 83.16% the variation in highway fuel efficiency (y) can be explained by city fuel efficiency (x)

Find a 95% confidence interval for the correlation. Interpret the interval.
Does this agree with your results from the hypothesis test of slope?

```
        Pearson's product-moment correlation

data:  x and y
t = 55.826, df = 631, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.8978093 0.9241850
```

$Confidence\ Interval\ for\ correlation$: $(0.8978093,\ 0.9241850)$

We are 95% confident that the true value for the correlation coefficient is between 0.8978093 and 0.9241850. This result is consistent with d) because a significant slope indicates a strong linear relationship, while our correlation confidence interval confirms that the relationship is strong and positive. Since the confidence interval does not include 0, this further supports the conclusion.

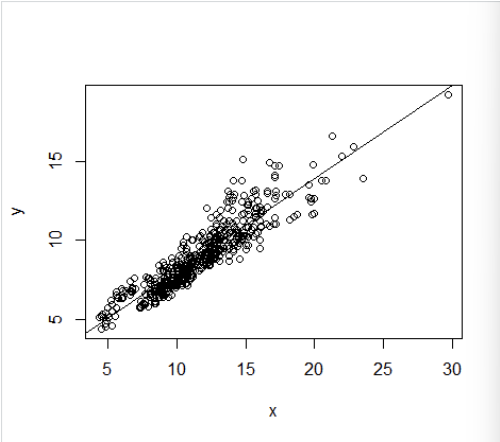
Check the assumptions of the model. Does it appear that any assumptions are violated?

Assumptions for the untransformed variables:

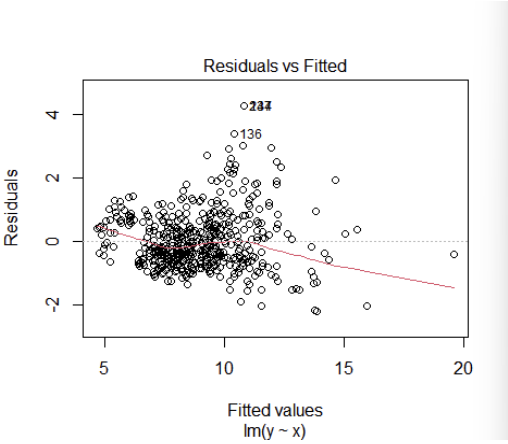
- Existence- Not violated
- Independence- Not violated; we assume all observations are independent.
- Linearity- Slightly violated; there is a slight wavy trend in the Residual vs fitted plot.
- Homoscedasticity- violation, there is a fanning shape as the plot progresses in the Residual vs Fitted plot.
- Normal Distribution- Violated; normal plot shows clear signs of violation

Fixing any assumptions and transforming variables

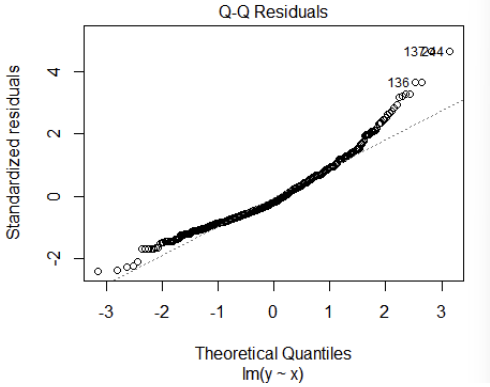
UNTRANSFORMED PLOTS
SCATTERPLOT



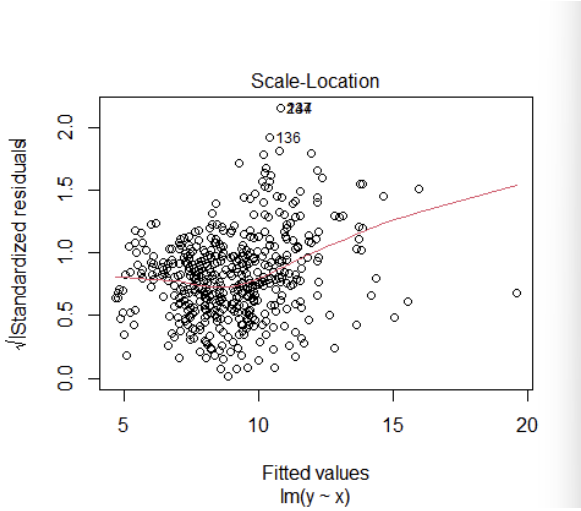
RESIDUAL VS FITTED



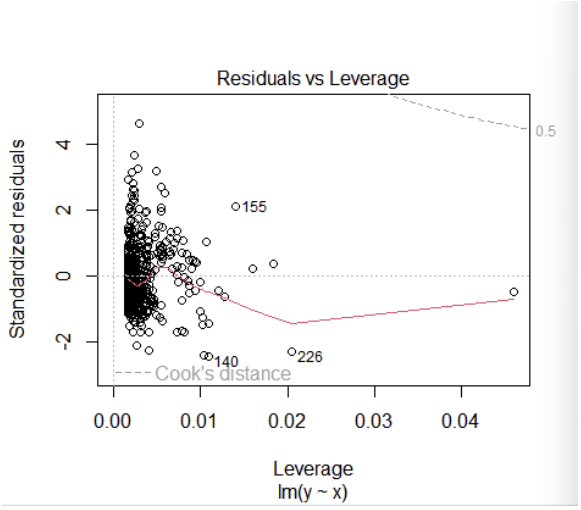
NORMAL PLOT



SCALE LOCATION

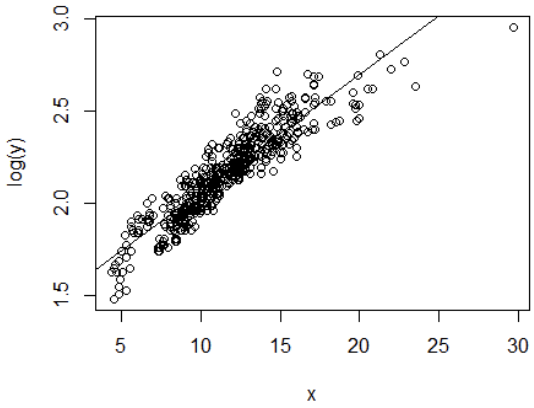


RESIDUALS VS LEVERAGE

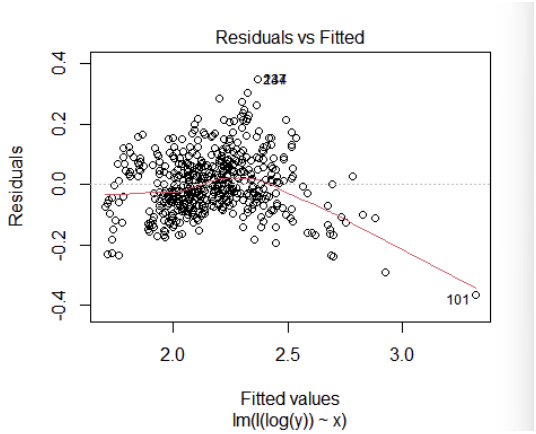


TRANSFORMED PLOTS (LOG(Y))

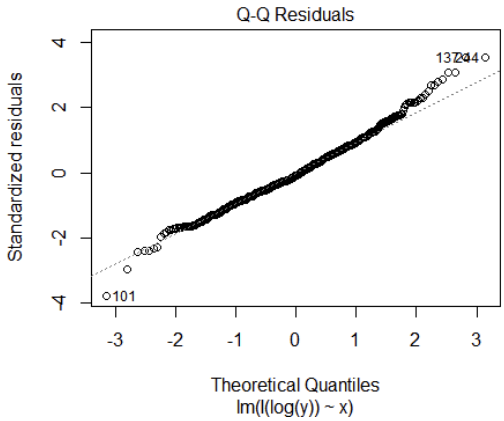
SCATTERPLOT



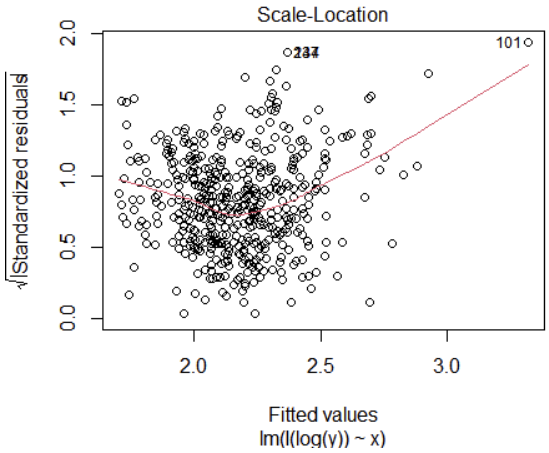
RESIDUAL VS FITTED



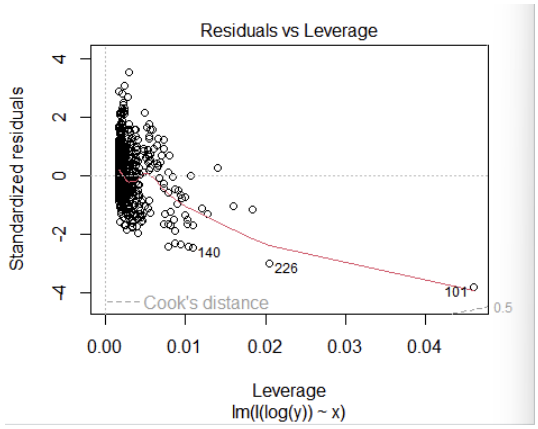
NORMAL PLOT



SCALE LOCATION



RESIDUALS VS LEVERAGE

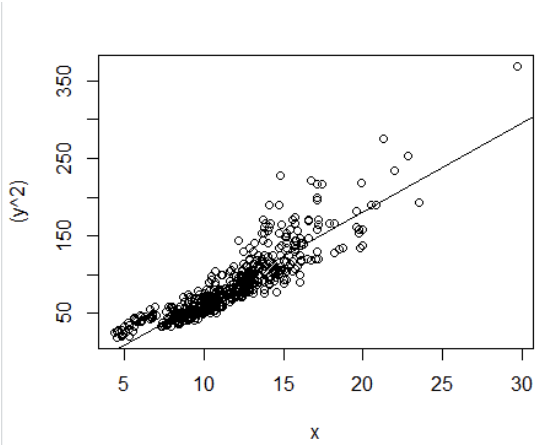


*Why **did** we choose this transformation?*

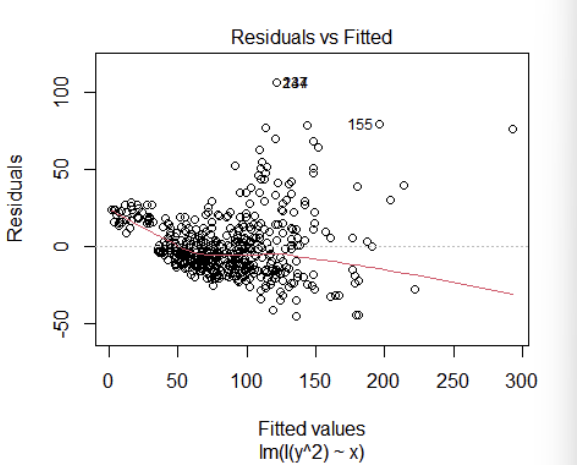
- The normal plot looks much better in comparison to the untransformed variable plots
- The residual vs fitted plot also shows a slightly better trend, however, at the end of the plot, the linearity becomes significantly worse which makes this an unideal transformation
- The scatter plot also looks much more linear and less clustered than the untransformed plot
- However, we concluded this was the best transformation since many plots improved rather than get worse even if it wasn't an ideal transformation.

TRANSFORMED PLOTS (Y^2)

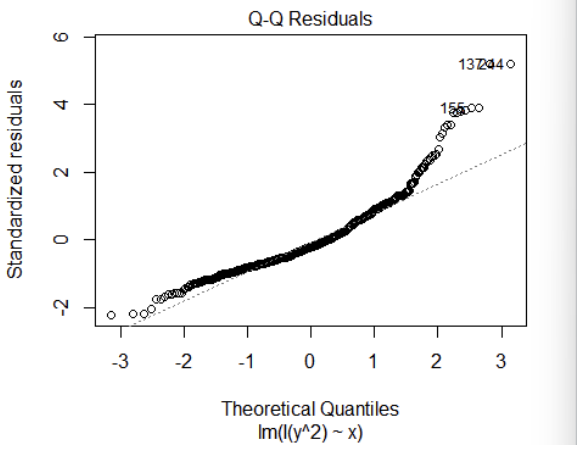
SCATTERPLOT



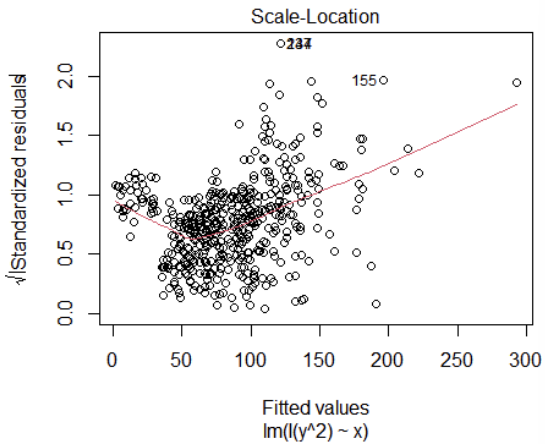
RESIDUAL VS FITTED



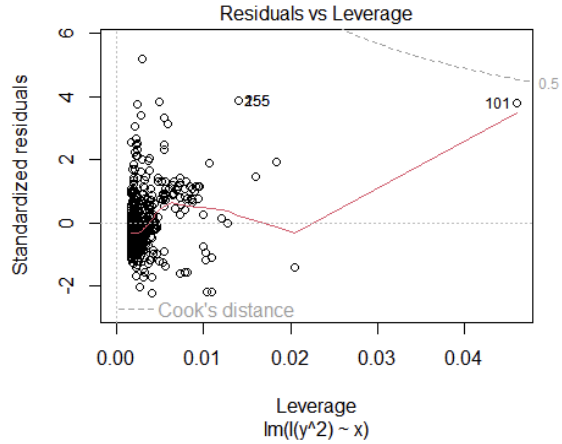
NORMAL PLOT



SCALE-LOCATION



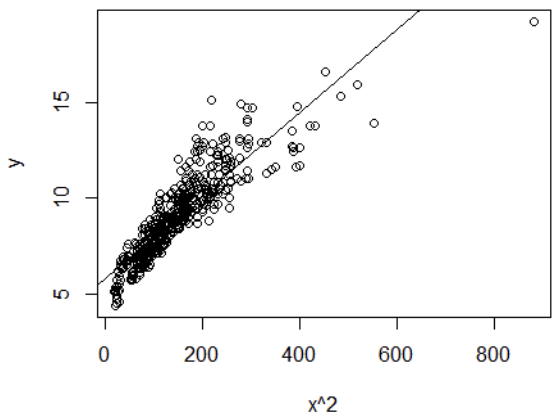
RESIDUALS VS LEVERAGE



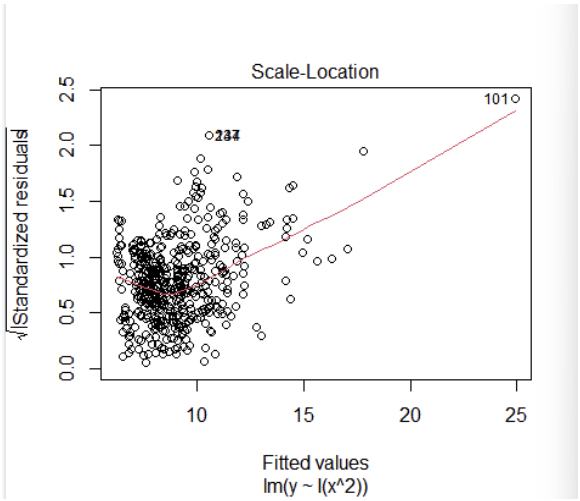
*Why **didn't** we choose this transformation?*

- The normal plot looks much worse than the untransformed variable plot
- The scatterplot also shows no major change from the untransformed plot
- Though linearity in the Residual vs fitted plot looks better compared to log(y), factors such as the normal plot looking worse made us lean away from choosing this transformation

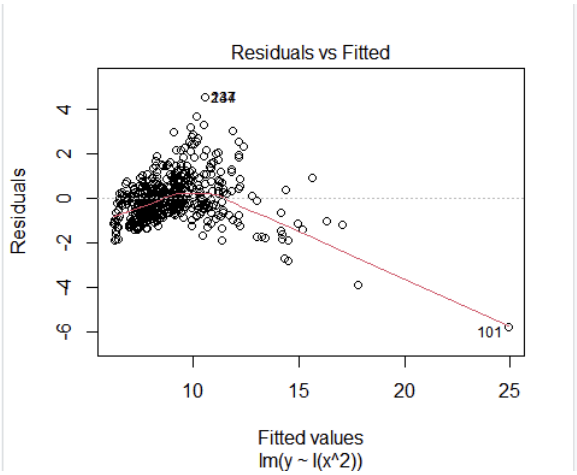
TRANSFORMED PLOT (X^2)
SCATTERPLOT



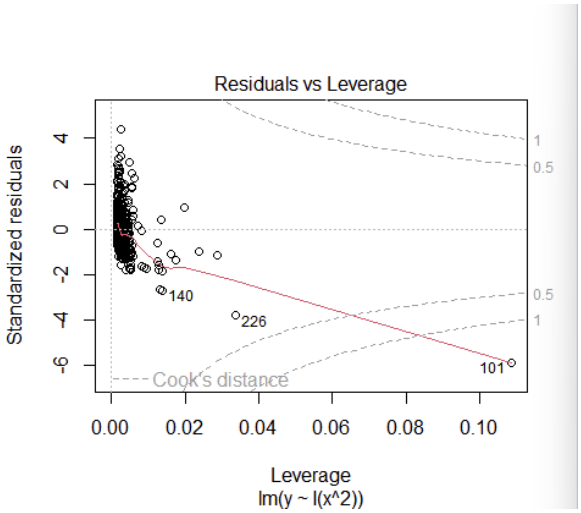
SCALE LOCATION



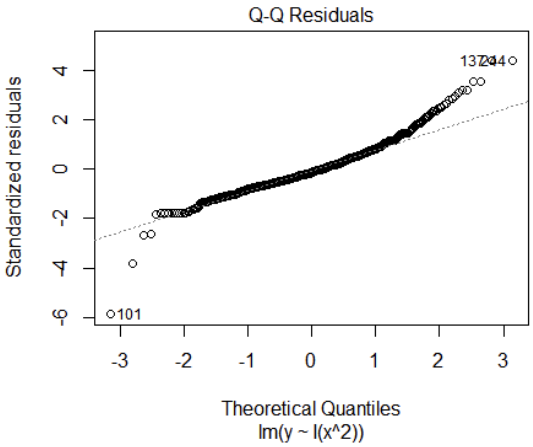
RESIDUAL VS FITTED



RESIDUALS VS LEVERAGE



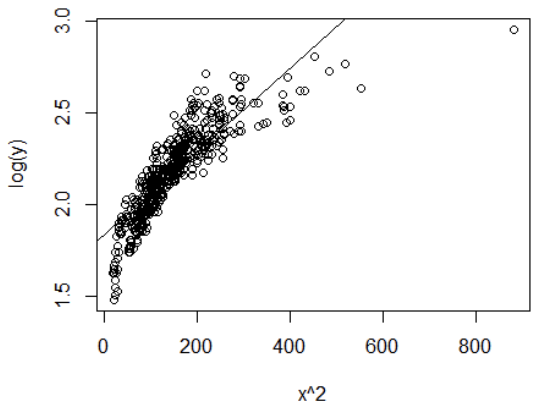
NORMAL PLOT



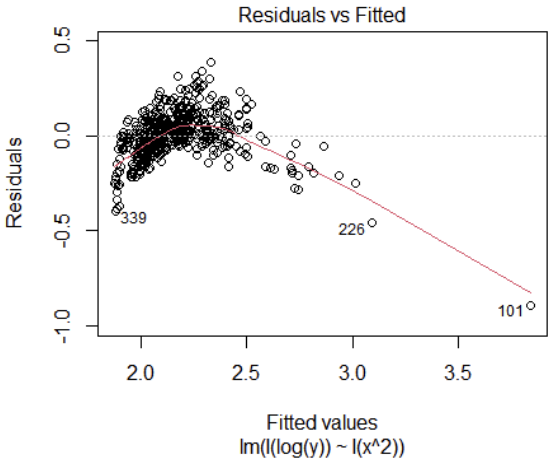
*Why **didn't** we choose this transformation?*

- The normal plot looks significantly worse
- The residual vs fitted plot shows poor linearity and a cluster on the left of the plot.
- The scatterplot also looks similar to the untransformed variable plots, showing that this transformation made the variable plots worse.

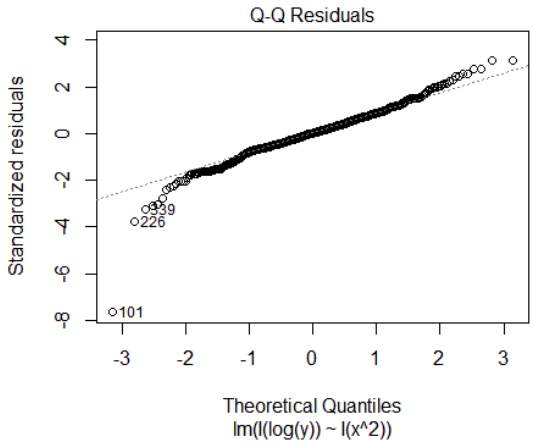
TRANSFORMED PLOTS (LOG(Y) , X^2)
SCATTERPLOT



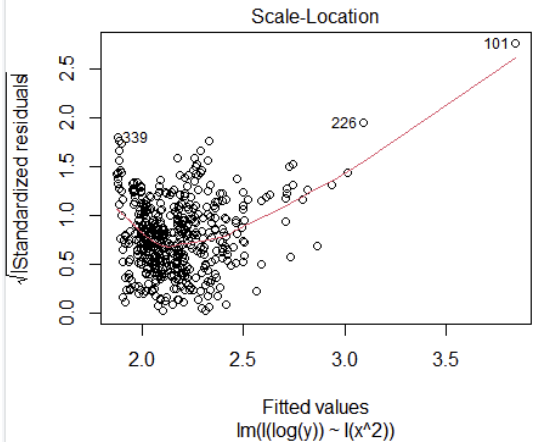
RESIDUAL VS FITTED



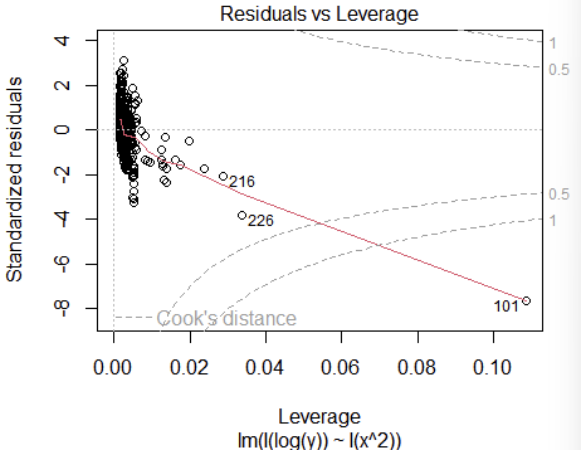
NORMAL PLOT



SCALE LOCATION



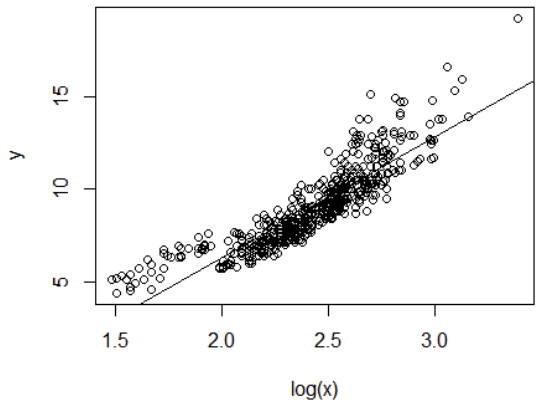
RESIDUALS VS LEVERAGE



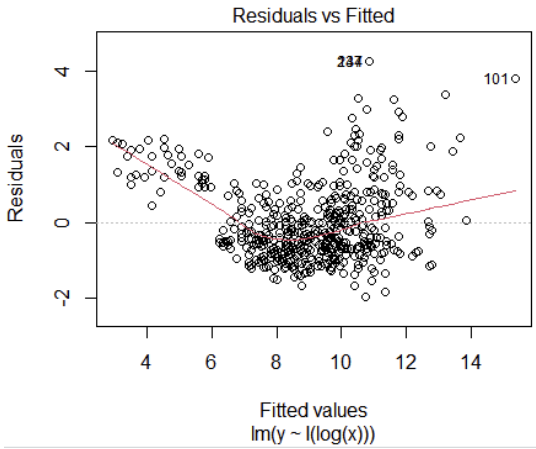
*Why **didn't** we choose this transformation?*

- Though the Normal plots look a little better, the Residual vs fitted plot shows worse linearity and homoscedasticity as the cluster is more compact and centers towards the beginning of the plot rather than the end.
- The scatter plot also shows a more quadratic trend than linear, which causes us to lean away from this transformation.

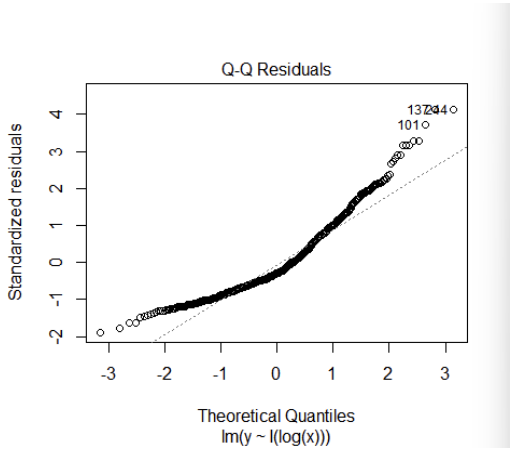
TRANSFORMED PLOT (LOG(X)) SCATTERPLOT



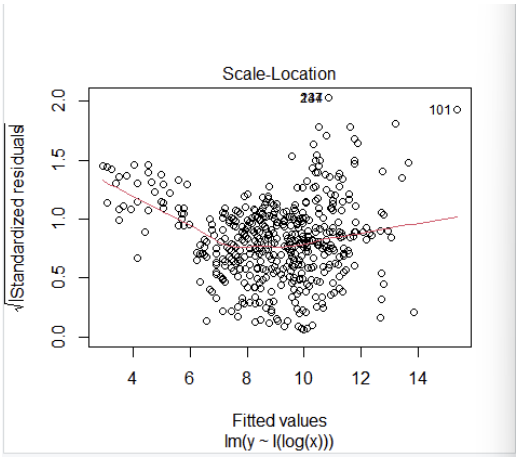
RESIDUALS VS FITTED



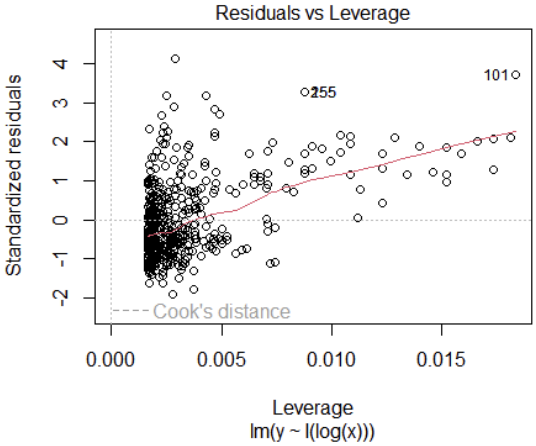
NORMAL PLOT



SCALE-LOCATION



RESIDUALS VS LEVERAGE



*Why **didn't** we choose this transformation?*

- The normal plot looks much worse compared to the untransformed variable plots.
- Linearity also looks worse as the residual vs fitted plot seems to look more quadratic than linear.
- The scatter plot also looks a little worse, as it shows a tiny bit of curvature.

Final model hypothesis test to determine if there is a significant linear relationship between the transformed variables

```
Call:
lm(formula = y ~ x, data = fuel)

Residuals:
    Min      1Q  Median      3Q     Max
-2.2045 -0.6229 -0.1876  0.5276  4.2572

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.12854    0.13485   15.79   <2e-16
x            0.58880    0.01089   54.08   <2e-16

(Intercept) ***
x           ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9197 on 600 degrees of freedom
Multiple R-squared:  0.8298,    Adjusted R-squared:  0.8295
F-statistic:  2925 on 1 and 600 DF,  p-value: < 2.2e-16
```

Ho: $\beta_1 = 0$

Ha: $\beta_1 \neq 0$

T.S.: 54.60

P-Value: <2e^-16 < α = 0.05 (Reject Ho!)

Conclusion: There is sufficient evidence at the 0.05 significance level to indicate there is a significant linear relationship between our transformed variables. Comparing this to the results we concluded from part D, we concluded that there is a linear relationship between our untransformed variables; we get a similar result when we transform our variables to log(y).

Abstract

The objective of this study was to analyze Canadian fuel consumption to determine the best linear model for predicting **highway fuel efficiency** using **city fuel efficiency**. First, we generated a **scatter plot** and obtained the **least squares regression line.** We then conducted a hypothesis test for the **slope**. Since the **p-value** obtained from **R** (a programming language) was lower than **α = 0.05**, falling into the **rejection region**, we concluded that the **slope** is significant, confirming that **city fuel efficiency** is a useful predictor of **highway fuel efficiency**.

This conclusion was further evident through the **correlation coefficient (R)** and the **coefficient of determination (R^2)**. The 95% **confidence interval** for **R**, $0.898 \leq R \leq 0.924$, indicates a strong positive relationship between **city fuel efficiency** and **highway fuel efficiency**, reinforcing our previous conclusion.

Additionally, we assessed the **five assumptions of linear regression** to check for potential violations. **Linearity**, **Homoscedasticity** (variance), and **Normal Distribution** showed clear signs of violation through their plots. To fix this, we transformed the variables and ultimately concluded that a logarithm transformation showed the best results within the plots. Nonetheless, the current model remains a **reasonable option** for predicting **highway fuel efficiency** based on **city fuel efficiency**.