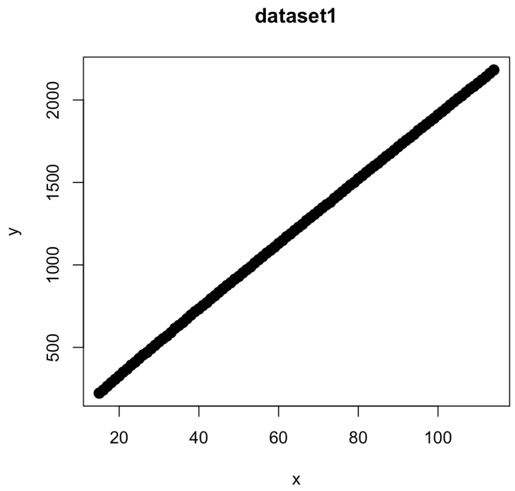
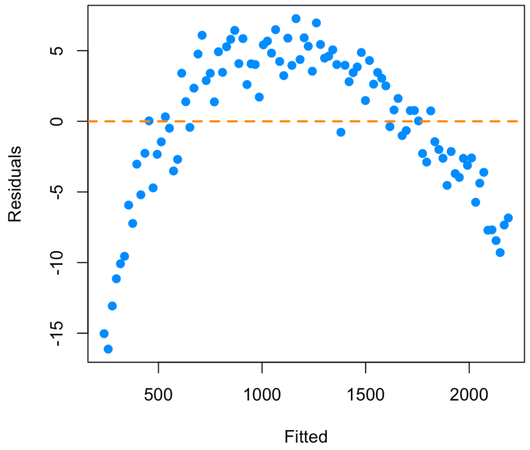
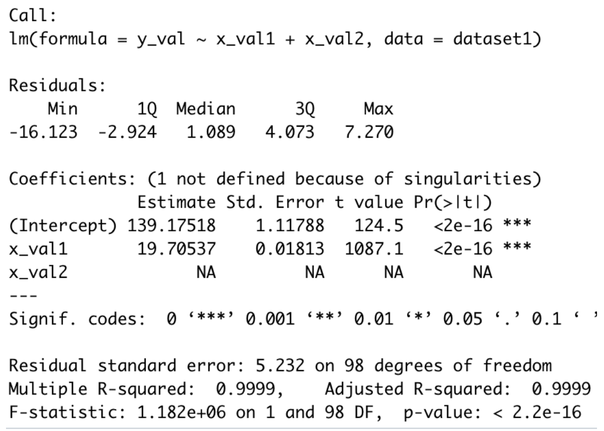
Kayla Benitez

①a Dataset 1: (unaltered analysis)

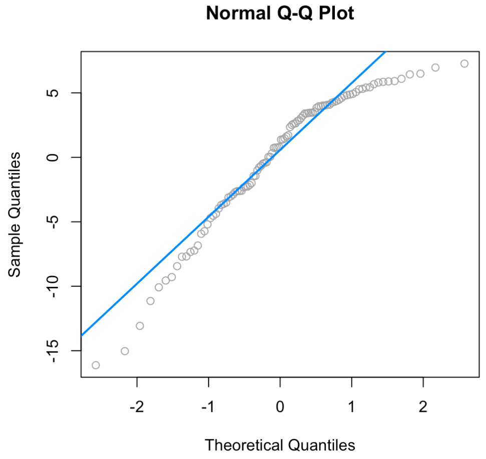




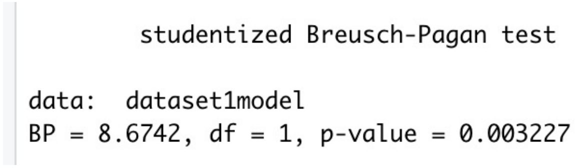
This dataset1 is not Normally distributed, we can see this clearly with the plotted vs. residuals graph where the pots do not have an even distribution.



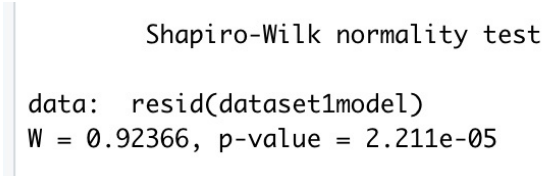
```
Call:
lm(formula = y_val ~ x_val1 + x_val2, data = dataset1)

Residuals:
    Min      1Q  Median      3Q     Max 
-16.123  -2.924   1.089   4.073   7.270 

Coefficients: (1 not defined because of singularities)
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 139.17518    1.11788   124.5   <2e-16 ***
x_val1       19.70537    0.01813  1087.1   <2e-16 ***
x_val2             NA         NA      NA       NA    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 

Residual standard error: 5.232 on 98 degrees of freedom
Multiple R-squared:  0.9999,	Adjusted R-squared:  0.9999 
F-statistic: 1.182e+06 on 1 and 98 DF,  p-value: < 2.2e-16
```



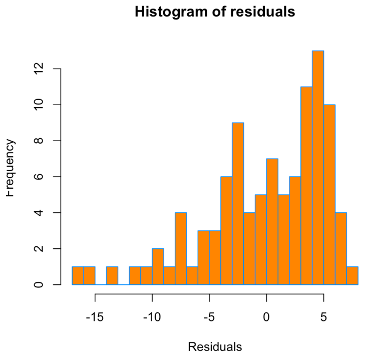
The quadrants 1Q and 3Q are unbalanced and the median is not equal to zero. The normal Q-Q plot shows the residuals are not on the regression line indicating the pts being violated. While the histogram also does not follow the bell curve since it is skewed to the right.



```
	studentized Breusch-Pagan test

data:  dataset1model
BP = 8.6742, df = 1, p-value = 0.003227
```



```
	Shapiro-Wilk normality test

data:  resid(dataset1model)
W = 0.92366, p-value = 2.211e-05
```

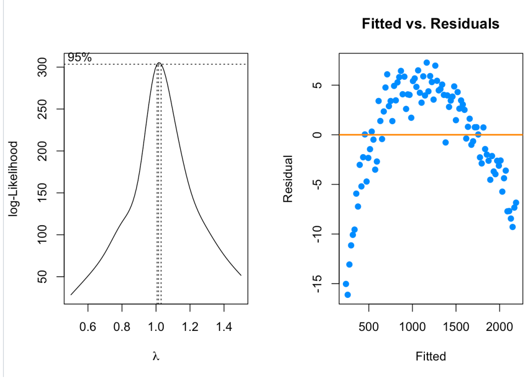


The Shapiro - Wilk Test for Normality Fails because the p-value < 0.05 and the Breush- Pagan tests for constant variance fails because the p-value < 0.05. We will have to transform the data so the dataset1 follows a Normal distribution.

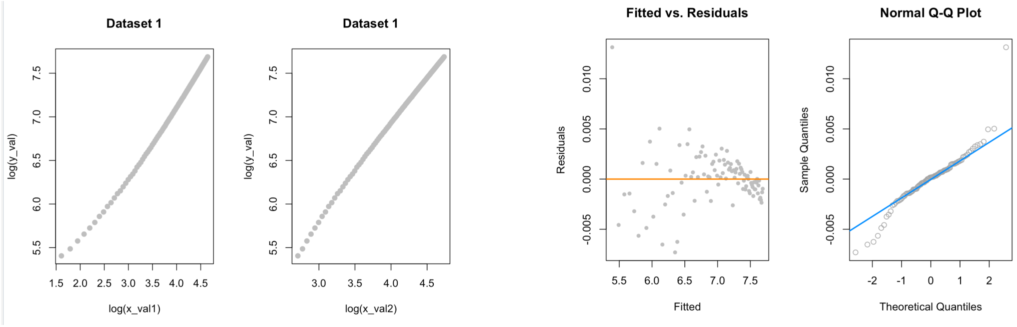
①b After Transformation:

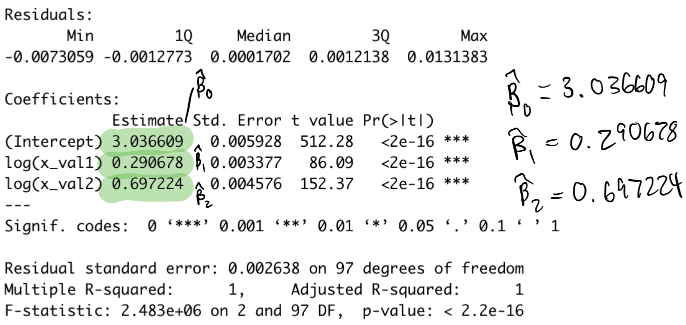
I couldn't find the correct transformation, the log transformation was the closest but since it failed the BPtest and SW test I know it is not a good fit.

Here, I tried the box cox method, where my lambda = 1.0303, although since the variance is non constant in the fitted vs residuals plot this would not be a good fit.



These are graphs of the predictor variables after applying the log transform. The fitted vs residuals plot is better with the variance more constant, although it is still non constant variance. The Normal Q-Q plot is better than the original since the residuals lie on the regression line although there are some data pts that are still off.





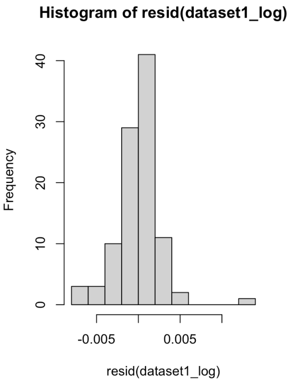
```
Residuals:
       Min         1Q     Median         3Q        Max
-0.0073059 -0.0012773  0.0001702  0.0012138  0.0131383

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.036609   0.005928  512.28   <2e-16 ***
log(x_val1) 0.290678   0.003377   86.09   <2e-16 ***
log(x_val2) 0.697224   0.004576  152.37   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.002638 on 97 degrees of freedom
Multiple R-squared:      1,     Adjusted R-squared:      1
F-statistic: 2.483e+06 on 2 and 97 DF,  p-value: < 2.2e-16
```

$\hat{\beta}_0$ $\hat{\beta}_1$ $\hat{\beta}_2$

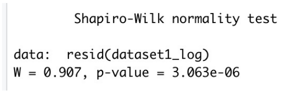
$\hat{\beta}_0 = 3.036609$

$\hat{\beta}_1 = 0.290678$

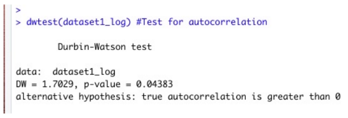
$\hat{\beta}_2 = 0.697224$



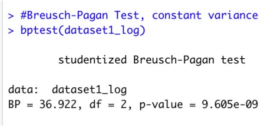
The summary of the residuals for this log transform has the quadrants more symmetric than the original and the median is closer to 0. The histogram also appears more normal with a bell curve distribution and not as skewed to the right as the original.



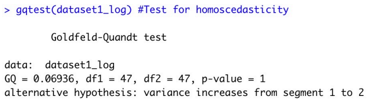
```
        Shapiro-Wilk normality test

data:  resid(dataset1_log)
W = 0.907, p-value = 3.063e-06
```



```
>
> dwtest(dataset1_log) #Test for autocorrelation

        Durbin-Watson test

data:  dataset1_log
DW = 1.7029, p-value = 0.04383
alternative hypothesis: true autocorrelation is greater than 0
```



```
> #Breusch-Pagan Test, constant variance
> bptest(dataset1_log)

        studentized Breusch-Pagan test

data:  dataset1_log
BP = 36.922, df = 2, p-value = 9.605e-09
```



```
> gqtest(dataset1_log) #Test for homoscedasticity

        Goldfeld-Quandt test

data:  dataset1_log
GQ = 0.06936, df1 = 47, df2 = 47, p-value = 1
alternative hypothesis: variance increases from segment 1 to 2
```

The tests however prove that this log transformed data is not a good fit since the p< 0.05 meaning it is not Normally distributed or have constant variance.

2

2) In this problem work within the Simple Linear Regression (SLR) context:

$$Y_n = \beta_0 + \beta_1 x_n + \epsilon_n,\ n = 1, \ldots, N,$$
$$\epsilon_n \sim N(0, \sigma^2), n = 1, \ldots, N.$$

(a) Show that that $\mathrm{E}[\hat{\beta}_1] = \beta_1$, where E[ ] denotes expectation. Please do not simply quote a theorem statement for this part of this problem or those below, but instead give a mathematical argument. (Also note that showing $\mathrm{E}[\hat{\beta}_m] = \beta_m$ for $m = 0$ is similar to the case $m = 1$, and you only need to include in your answer the case for $m = 1$.)

(b) Show that Var($\hat{\beta}_1$)= $\frac{\sigma^2}{S_{xx}}$, where $S_{xx} = \sum_{n=1}^{N}(x_n - \bar{x})^2$ and Var( ) denotes the variance.

(c) Show that Var($\hat{\beta}_0$)= $\sigma^2\left(\frac{1}{N} + \frac{\bar{x}^2}{S_{xx}}\right)$, where $\bar{x} = \frac{1}{N}\sum_{n=1}^{N} x_n$.

(d) Are we able to conclude, directly from parts (a) and (b), that $\hat{\beta}_1$ is normally-distributed with mean β_1 and variance $\frac{\sigma^2}{S_{xx}}$? Why or why not?

$$S_{xx} = \sum_{n=1}^{N} (x_n - \bar{x})^2$$

From Lec2:

$$\hat{\beta}_1 = \frac{\sum_{n=1}^{N}(x_n - \bar{x})(y_n - \bar{y})}{\sum_{n=1}^{N}(x_n - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}$$

$$\hat{\beta}_0 = \frac{1}{n}\left(\sum_{n=1}^{N} y_n - \hat{\beta}_1 \sum_{n=1}^{N} x_n\right) \quad \text{where } \bar{x} = \frac{1}{n}\sum_{n=1}^{N} x_n$$

Appendix A.9

$$\hat{\beta}_0 = \bar{y} - \frac{S_{yx}}{S_{xx}}\bar{x}$$

a) $E(\hat{\beta}_1) = \beta_1$

$$= E\left[\frac{\sum_{n=1}^{N}(x_n-\bar{x})(y_n-\bar{y})}{\sum_{n=1}^{N}(x_n-\bar{x})^2}\right]$$

$$= \frac{\sum_{n=1}^{N}(x_n-\bar{x})E(y_n-\bar{y})}{\sum_{n=1}^{N}(x_n-\bar{x})^2}$$

$$= \frac{\sum_{n=1}^{N}(x_n-\bar{x})[E(y_n)-E(\bar{y})]}{\sum_{n=1}^{N}(x_n-\bar{x})^2}$$

$$= \frac{\sum_{n=1}^{N}(x_n-\bar{x})\left[\beta_0+\beta_1x_n-\frac{1}{n}\Sigma(\beta_0+\beta_1x_n)\right]}{\sum_{n=1}^{N}(x_n-\bar{x})^2}$$

$$= \frac{\sum_{n=1}^{N}(x_n-\bar{x})\left[\beta_0+\beta_1x_n-\beta_0-\beta_1\cdot\frac{1}{n}\Sigma x_n\right]}{\sum_{n=1}^{N}(x_n-\bar{x})^2}$$

$$= \frac{\sum_{n=1}^{N}(x_n-\bar{x})[\beta_1(x_n-\bar{x})]}{\sum_{n=1}^{N}(x_n-\bar{x})^2}$$

$$= \frac{\sum_{n=1}^{N}(x_n-\bar{x})^2\ \beta_1}{\sum_{n=1}^{N}(x_n-\bar{x})^2}$$

$$= \beta_1$$

b) $\mathrm{Var}(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}$

$$= \mathrm{Var}\left[\frac{\Sigma(x_n-\bar{x})(y_n-\bar{y})}{\Sigma(x_n-\bar{x})^2}\right]$$

$$= \mathrm{Var}\left[\frac{\Sigma(x_n-\bar{x})(y_n)}{\Sigma(x_n-\bar{x})^2}\right]$$

$$= \left[\frac{\Sigma(x_n-\bar{x})}{\Sigma(x_n-\bar{x})^2}\right]^2 \mathrm{Var}(y_n)$$

$$\mathrm{var}(\beta_0+\beta_1x_n+\epsilon_n)$$

$$= \frac{1}{\Sigma(x_n-\bar{x})^2}\ \mathrm{var}\cancel{(\beta_0)}^{\,0} + \mathrm{var}\cancel{(\beta_1x_n)}^{\,0} + v(\epsilon_n)$$

$$= \left(\frac{1}{\Sigma(x_n-\bar{x})^2}\right)\sigma^2$$

$$= \frac{\sigma^2}{S_{xx}}$$

c) $(\mathrm{var}\,\hat{\beta}_0) = \sigma^2\left(\frac{1}{N}+\frac{\bar{x}^2}{S_{xx}}\right)$ $\quad \bar{x} = \frac{1}{N}\sum_{n=1}^{N} x_n$

$$= \mathrm{var}(\bar{y} - \hat{B}_1\bar{x})$$

$$= \mathrm{var}(\bar{y}) + \mathrm{var}(B_1\bar{x}) - 2\,\mathrm{co}\cancel{\mathrm{v}(\bar{y}, B_1\bar{x})}^{\,0}$$

$$= \frac{1}{N}\sigma^2 + \mathrm{var}\left(\frac{S_{yx}}{S_{xx}}\bar{x}\right)$$

$$= \frac{\sigma^2}{N} + \frac{\bar{x}^2}{S_{xx}^2}\ \mathrm{var}\cancel{(S_{yx})}^{\,\sigma^2}$$

$$= \frac{\sigma^2}{N} + \frac{\bar{x}^2\sigma^2}{S_{xx}}$$

$$= \sigma^2\left(\frac{1}{N}+\frac{\bar{x}^2}{S_{xx}}\right)$$

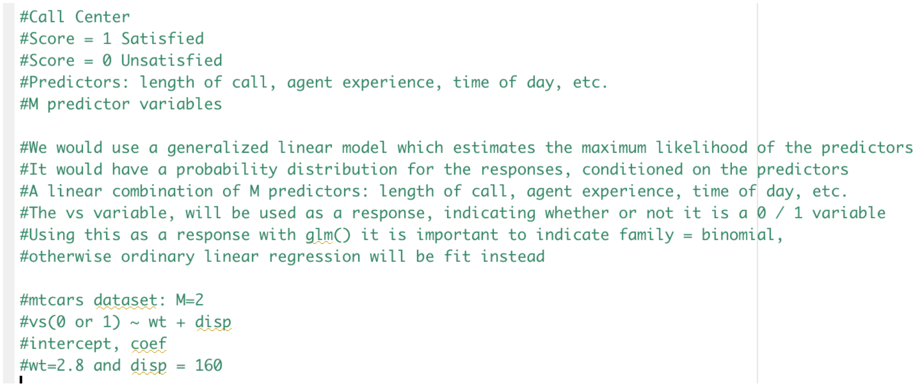
d) Part A: $E(\hat{\beta}_1) = \beta_1$ mean

Part B: $\mathrm{Var}(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}$ variance

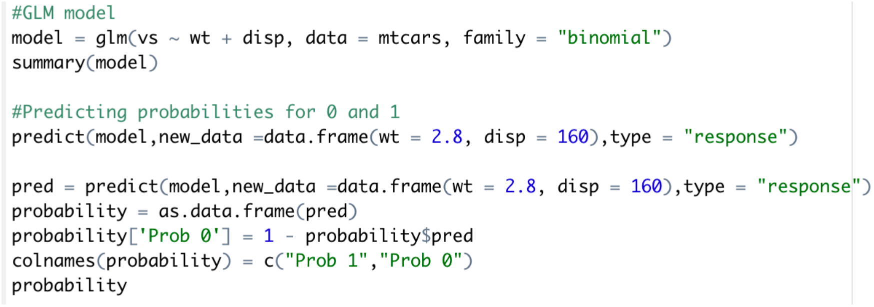
$$\hat{\beta}_1 \sim N\left(\hat{\beta}_1, \frac{\sigma^2}{S_{xx}}\right)$$

, yes, we can conclude $\hat{\beta}_1 \sim N\left(\hat{\beta}_1, \frac{\sigma^2}{S_{xx}}\right)$ b/c our SLR model had β_1 Normally distributed

③

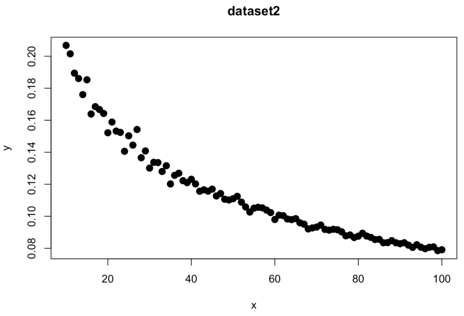


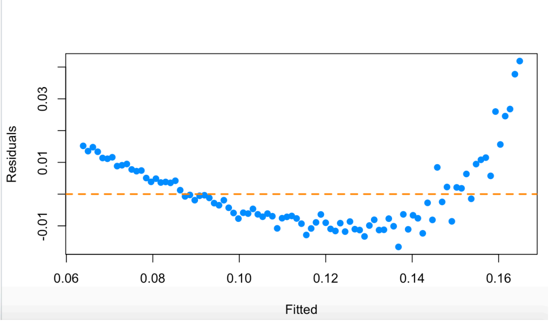
```
#Call Center
#Score = 1 Satisfied
#Score = 0 Unsatisfied
#Predictors: length of call, agent experience, time of day, etc.
#M predictor variables

#We would use a generalized linear model which estimates the maximum likelihood of the predictors
#It would have a probability distribution for the responses, conditioned on the predictors
#A linear combination of M predictors: length of call, agent experience, time of day, etc.
#The vs variable, will be used as a response, indicating whether or not it is a 0 / 1 variable
#Using this as a response with glm() it is important to indicate family = binomial,
#otherwise ordinary linear regression will be fit instead

#mtcars dataset: M=2
#vs(0 or 1) ~ wt + disp
#intercept, coef
#wt=2.8 and disp = 160
```

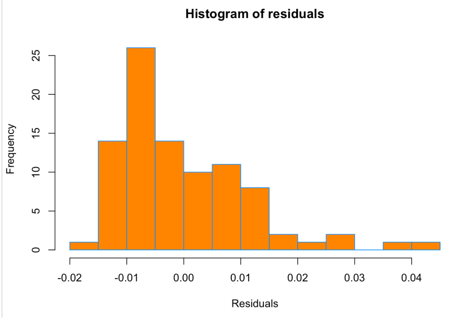


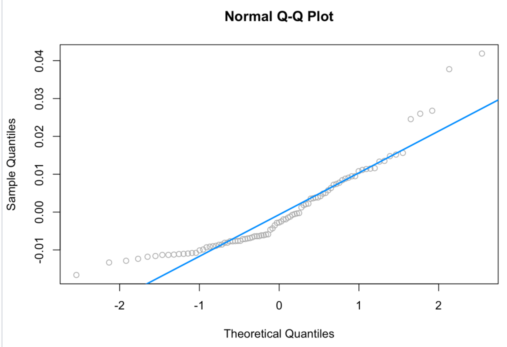
```
#GLM model
model = glm(vs ~ wt + disp, data = mtcars, family = "binomial")
summary(model)

#Predicting probabilities for 0 and 1
predict(model,new_data =data.frame(wt = 2.8, disp = 160),type = "response")

pred = predict(model,new_data =data.frame(wt = 2.8, disp = 160),type = "response")
probability = as.data.frame(pred)
probability['Prob 0'] = 1 - probability$pred
colnames(probability) = c("Prob 1","Prob 0")
probability
```

④ Dataset 2 (unaltered analysis)

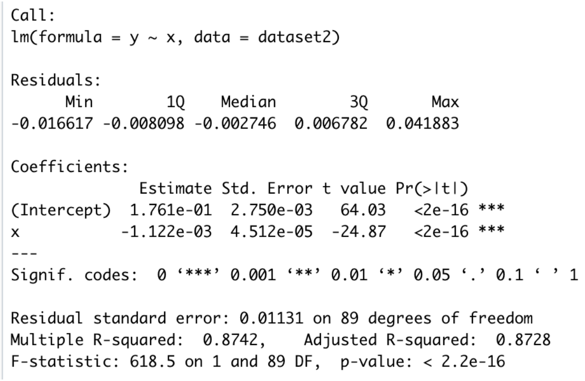




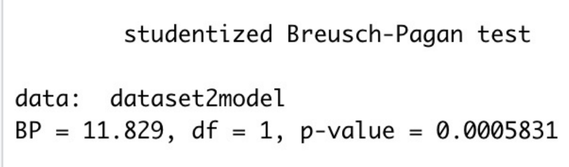




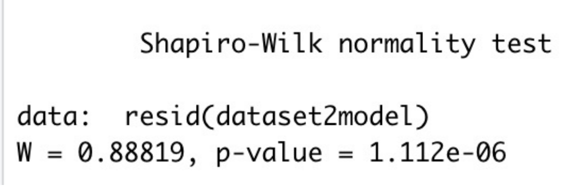
Dataset 2 is not Normally distributed. The fitted vs residuals plot shows non-constant variance because the data pits are not evenly distributed.



```
Call:
lm(formula = y ~ x, data = dataset2)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.016617 -0.008098 -0.002746  0.006782  0.041883 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.761e-01  2.750e-03   64.03   <2e-16 ***
x           -1.122e-03  4.512e-05  -24.87   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.01131 on 89 degrees of freedom
Multiple R-squared:  0.8742,	Adjusted R-squared:  0.8728 
F-statistic: 618.5 on 1 and 89 DF,  p-value: < 2.2e-16
```



```
	studentized Breusch-Pagan test

data:  dataset2model
BP = 11.829, df = 1, p-value = 0.0005831
```

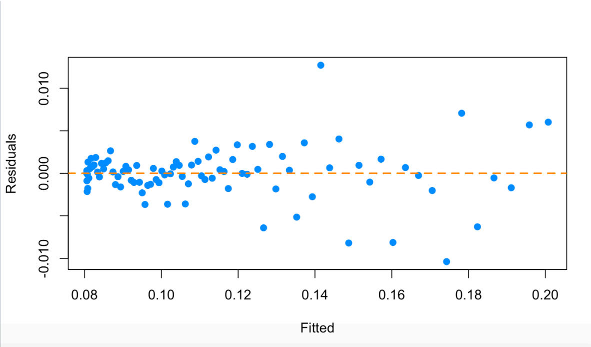


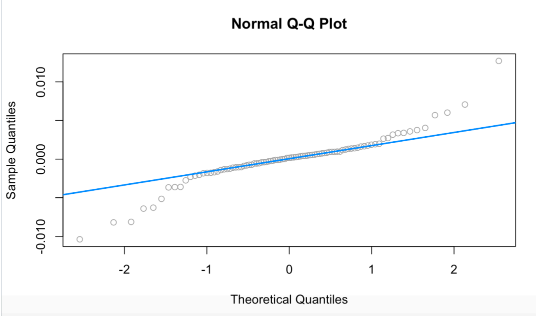
```
	Shapiro-Wilk normality test

data:  resid(dataset2model)
W = 0.88819, p-value = 1.112e-06
```

Dataset 2’s histogram doesn’t follow a normal distribution since the pats are skewed left and the Q-Q plot shows the residuals are not all on the regression line, meaning non-normal distribution. The quadrants 1Q and 3Q are unbalanced and the median is not equal to zero. While the Normality Shapiro - Wilk test fails with a p-value < 0.05 and the BP Test also fails with a p-value < 0.05. Dataset 2 also needs to be transformed.

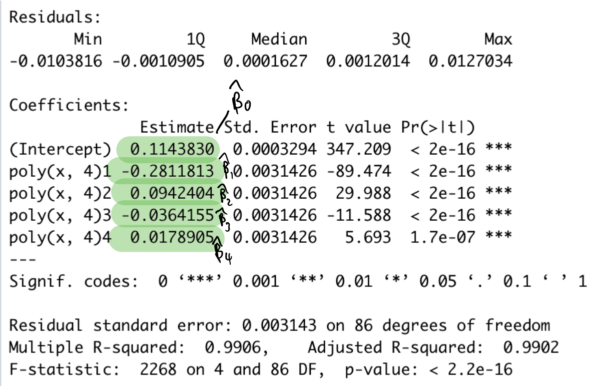
②b After Transformation

I couldn't find the correct transformation, the polynomial transformation was the closest but since it failed the diagnostic tests I know it is not a good fit.

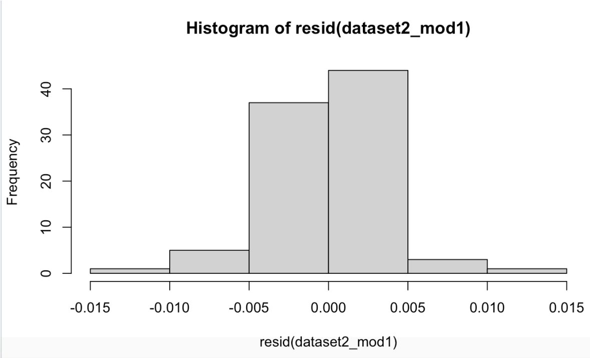




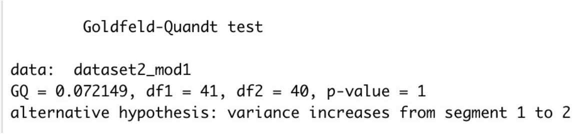
These are graphs after applying the polynomial transformation. The fitted vs residuals plot is better with the variance more constant, although it is still non constant variance. The Normal Q-Q plot is better than the original since the residuals lie on the regression line although there are still some data pts that are off.



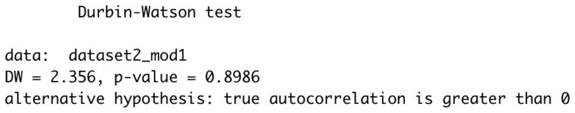
```
Residuals:
       Min         1Q     Median         3Q        Max 
-0.0103816 -0.0010905  0.0001627  0.0012014  0.0127034 
                                          ^
Coefficients:                             β0
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.1143830  0.0003294 347.209  < 2e-16 ***
poly(x, 4)1 -0.2811813 β1 0.0031426 -89.474  < 2e-16 ***
poly(x, 4)2  0.0942404 β2 0.0031426  29.988  < 2e-16 ***
poly(x, 4)3 -0.0364155 β3 0.0031426 -11.588  < 2e-16 ***
poly(x, 4)4  0.0178905  0.0031426   5.693  1.7e-07 ***
---                   β4
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.003143 on 86 degrees of freedom
Multiple R-squared:  0.9906,	Adjusted R-squared:  0.9902 
F-statistic:  2268 on 4 and 86 DF,  p-value: < 2.2e-16
```



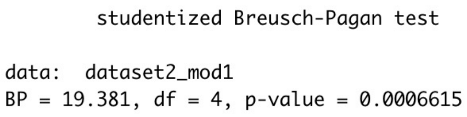
The summary of the residuals for this polynomial transform has the quadrants more symmetric than the original and the median is closer to 0. The histogram also appears more normal with a bell curve distribution and not as skewed to the left as the original.



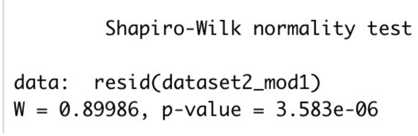
```
	Goldfeld-Quandt test

data:  dataset2_mod1
GQ = 0.072149, df1 = 41, df2 = 40, p-value = 1
alternative hypothesis: variance increases from segment 1 to 2
```



```
	Durbin-Watson test

data:  dataset2_mod1
DW = 2.356, p-value = 0.8986
alternative hypothesis: true autocorrelation is greater than 0
```



```
	studentized Breusch-Pagan test

data:  dataset2_mod1
BP = 19.381, df = 4, p-value = 0.0006615
```



```
	Shapiro-Wilk normality test

data:  resid(dataset2_mod1)
W = 0.89986, p-value = 3.583e-06
```

The diagnostics tests here prove though that this is not a good fit since it fails to have a normal distribution and constant variance. The Normality Shapiro - Wilk test fails with a p-value < 0.05 and the BP Test also fails with a p-value < 0.05.