Regression

Meghna Bajoria

2022-11-20

```
library(readr)
DataModel <- read_csv("C:\\Users\\Meghna\\OneDrive\\Documents\\Fall'22\\ISO-201\\RegressionData.csv")
```

```
## Rows: 50 Columns: 2
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## dbl (2): x, y
##
## 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.
```

```
head(DataModel)
```

```
## # A tibble: 6 × 2
##       x     y
##   <dbl> <dbl>
## 1   679  0.79
## 2   292  0.44
## 3  1012  0.56
## 4   493  0.79
## 5   582  2.7
## 6  1156  3.64
```

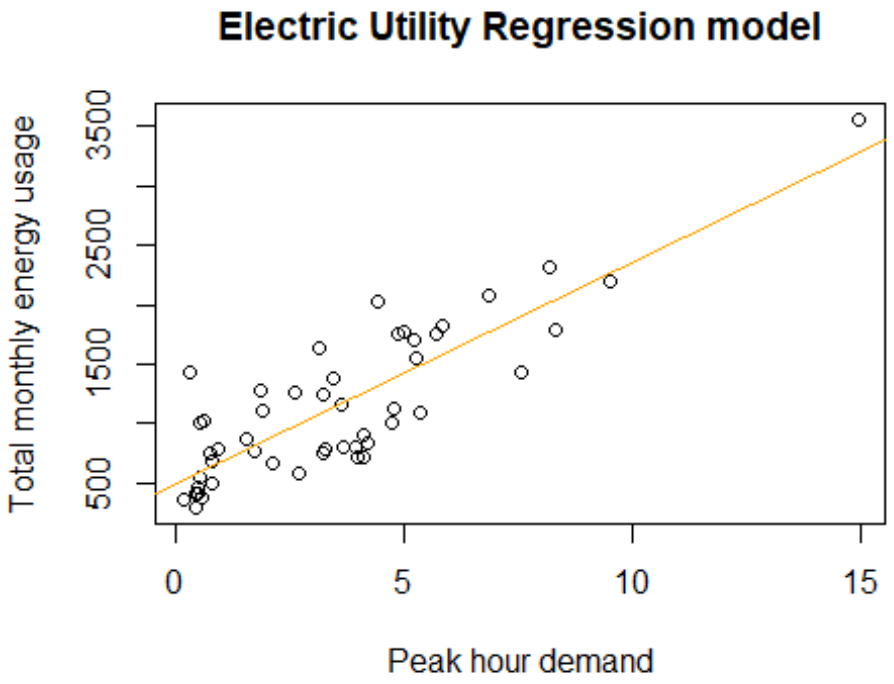
An electric utility is interested in developing a model relating peak-hour demand (y in kilowatts) to total monthly eneregy usage during the month (x, in kilowatt hours). Data for 50 residential customers

a. **Draw a scatter diagram of y versus x. Please record your observations.**

b. **Fit a simple linear regression model.**

```
plot(DataModel$y, DataModel$x, main = 'Electric Utility Regression model',
     xlab = 'Peak hour demand', ylab = 'Total monthly energy usage')

abline(lm(x ~ y, data = DataModel), col="orange")
```



From the above plot we can conclude that most of the consumers consume relatively constant and moderate quantity of energy monthly and at peak hours. However, we have one consumer at the top right who has high electricity consumption every month and specially at peak hours.

c. Test the significance of regression using α = 0.05 and 0.01 and record your observations.

```
utility.regression <- lm(x ~ y, data = DataModel)
summary(utility.regression)

##
## Call:
## lm(formula = x ~ y, data = DataModel)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -524.95 -268.45  -43.72  268.47  891.96
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   484.22      76.22   6.353 7.23e-08 ***
## y             186.52      16.88  11.047 8.81e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 343.9 on 48 degrees of freedom
```

```
## Multiple R-squared:  0.7177, Adjusted R-squared:  0.7118
## F-statistic:   122 on 1 and 48 DF,  p-value: 8.808e-15
```

```
qf(0.05, 1 , 48)
```

```
## [1] 0.003973475
```

```
qf(0.01, 1, 48)
```

```
## [1] 0.0001587329
```

qf value is used to check if the regression is statistically significant. The hypothesis is that there is no significant relationship between peak hour demand and total monthly energy consumption if the determined critical value was larger than the F statistic.

The first value is the significance level (α). The number of independent variables used in the regression analysis is represented by the second parameter, "df1," which is a degree of freedom. Only 'y' exists in this instance. Degrees of freedom are provided as the third number in the function "df2," which is the answer to the model equation. For this instance, that number is 48. Due to the fact that both the 0.05 and 0.01 levels of significance are smaller than the F-statistic, we may conclude that the model significantly explains fluctuations in the data and reject the null hypothesis.

d. Estimate the correlation coefficient and compute a 95% confidence interval for the correlation coefficient.

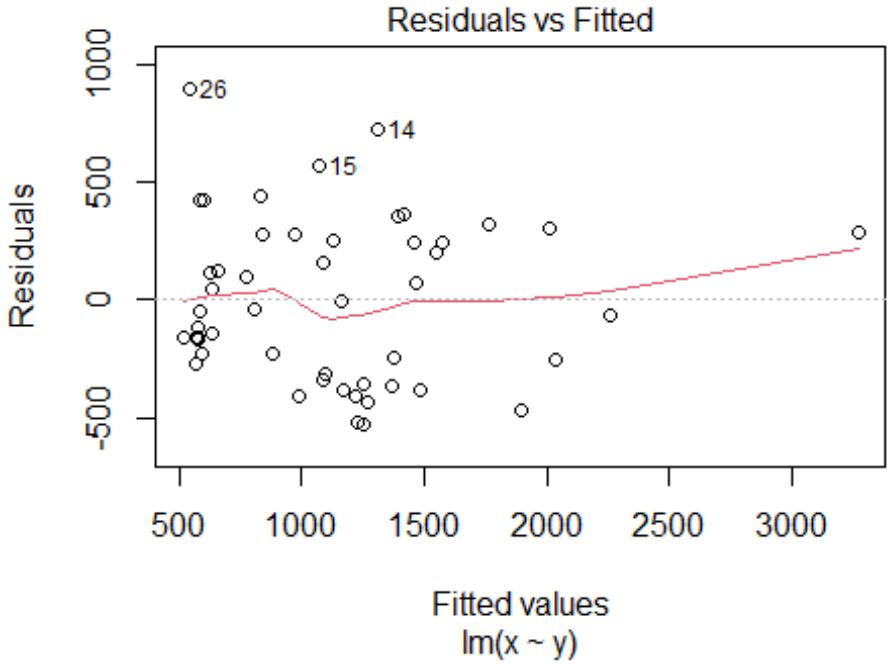
```
cor.test(DataModel$x, DataModel$y,
         method = "pearson")
```

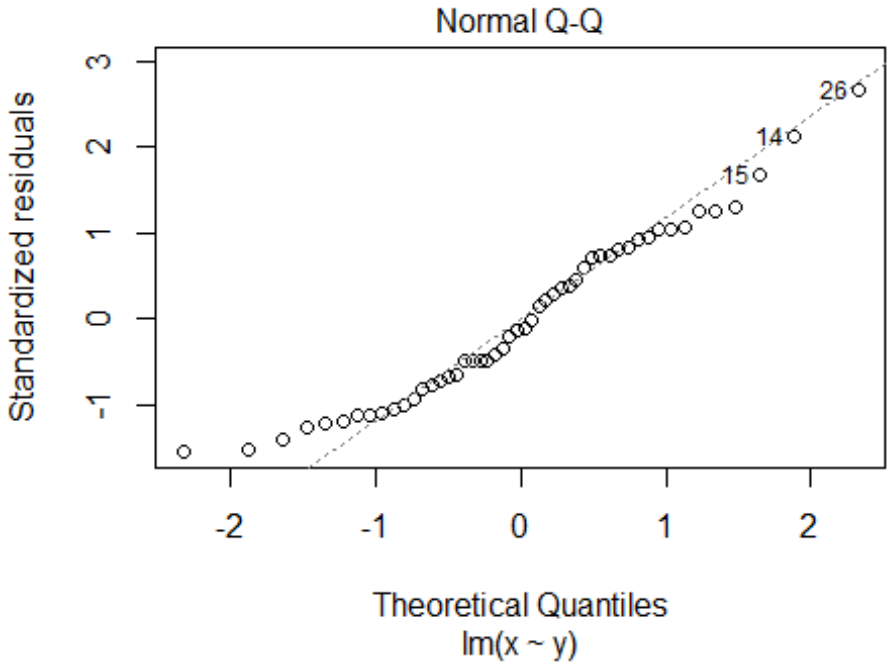
```
##
##  Pearson's product-moment correlation
##
## data:  DataModel$x and DataModel$y
## t = 11.047, df = 48, p-value = 8.808e-15
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.7443450 0.9107548
## sample estimates:
##       cor
## 0.8471703
```

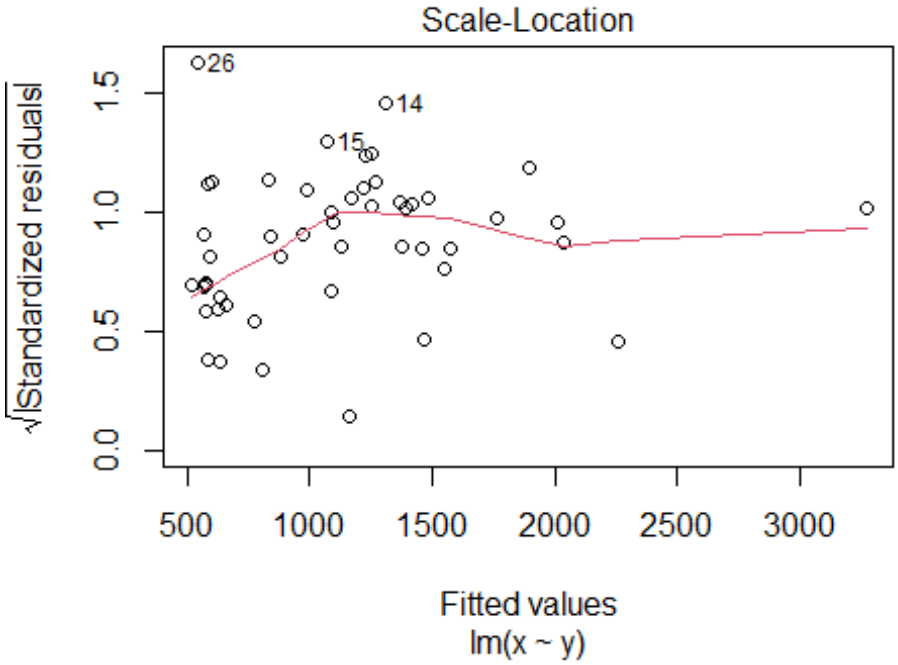
Using Pearson correlation coefficient, the correlation coefficient is 0.8471703, and the 95% confidence interval is [0.7443450,0.9107548]

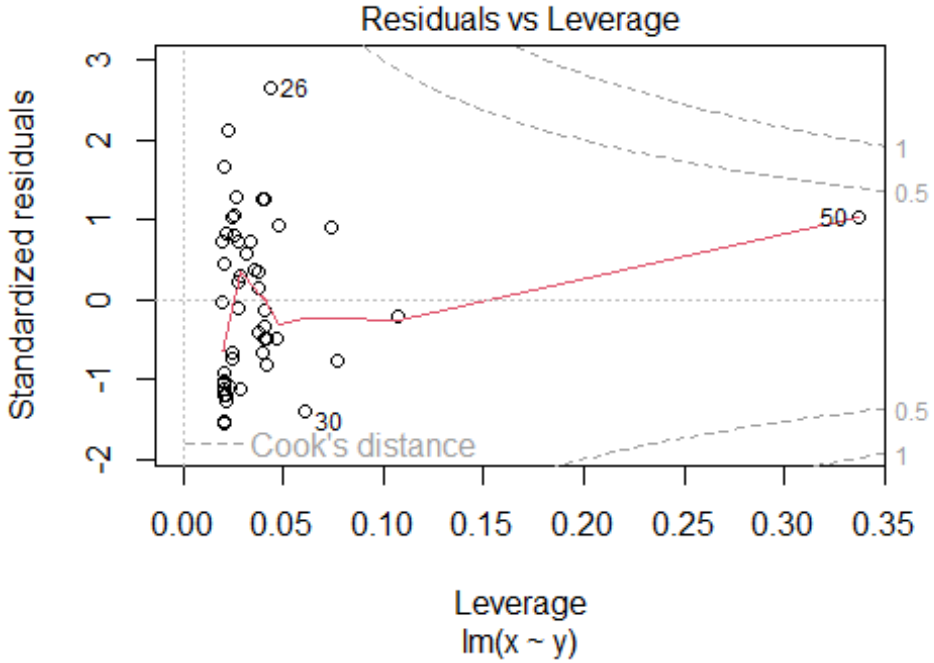
e. Plot the residuals versus y^i and comment on the underlying regression assumptions. Specifically, does it seem that the equality of variance assumption is satisfied.

```
plot(utility.regression)
```


Residuals vs Fitted
Residuals
Fitted values
lm(x ~ y)
26
14
15
1000
500
0
-500
500
1000
1500
2000
2500
3000


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(x ~ y)
26
14
15
3
2
1
0
-1
-2
-1
0
1
2





Few assumptions that are made in the simulation analysis are :

There is no correlation between ‘e’ errors and Y values.

The Y-values are linearly related to the x-values

The residual standard error, which measures the spread of observations around the regression line, is constant. Homoscedasticity, describes this feature.

The y-values (the errors) follow a normal distribution for any given value of x.

The validity of the first assumption can be checked by knowing how the study was set up and what data were collected. The model's residuals are used to check the second, third, and fourth assumptions. We can make four graphs that can help with this diagnosis by using the plot function in r. The first plot, which compares the residuals to the value that was fitted, is going to be the primary one that we look at while conducting the analysis of the second and third assumptions.

If linearity is met, the red line on the graph should appear to be relatively flat. The red line on our graph is mostly horizontal; however, as the line moves to the right along the fitted values axis, it gradually ascends. This might be read as generic linearity in this model, which is a valid assumption.

There should be no pattern if the fluctuation is constant. The points should be scattered and uniformly spaced, like a cloud of points. The majority of the points in this plot are off to the left and below, indicating a clustering effect. This assumption is flawed since the model's variation is not constant. One further factor that drives this point home is the fact that there is just one dot at the very end of the graph's y axis.

The fourth assumption can be examined using the second plot, a QQ or quantile-quantile plot, in which the y axis represents the actual residuals and the x axis represents the predicted residuals if the model were actually normally distributed. If the errors have a normal distribution, these points should generally fall along a diagonal line. The middle values on our graph are generally aligned along a diagonal line; however, the diagonals vary at the upper and lower boundaries of the line. This suggests that the majority of errors for the provided values of x in the model are normally distributed, with a few outliers demonstrating considerable deviations from the norm. This means that the model's fourth assumption holds true.

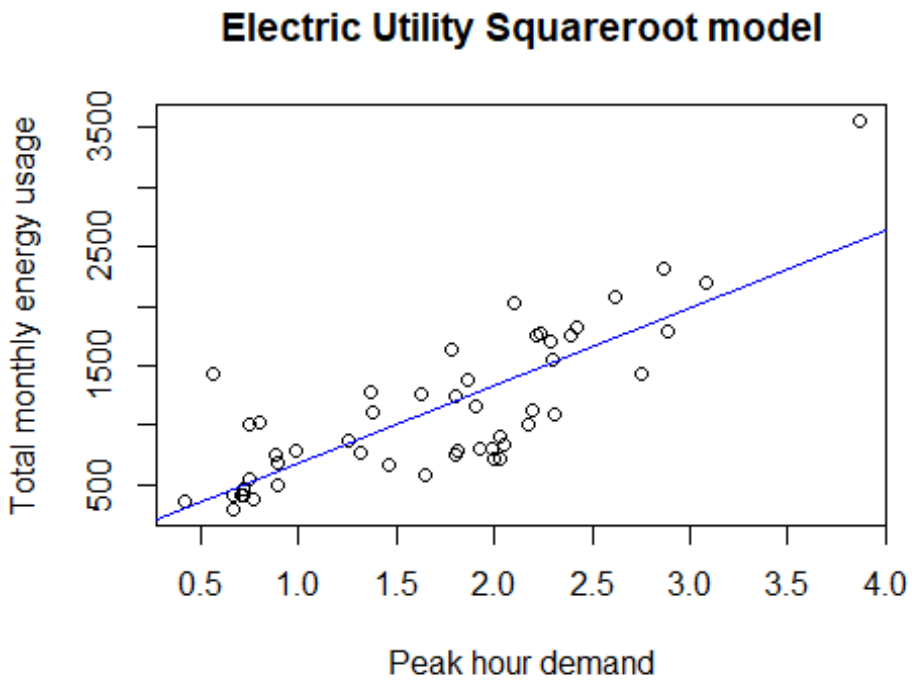
f. Find a simple linear regression model using $y\sqrt{}$ as the response, Does this transformation on y stabilize the inequality of variance problem noted in part (d)?

```
NewDataModel <- DataModel
NewDataModel[,"x"] <- DataModel[,"x"]
NewDataModel[,"rootY"] <- sqrt(DataModel[,"y"]) #creating a new column for
calculated y√
NewDataModel
```

```
## # A tibble: 50 × 3
##        x     y rootY
##    <dbl> <dbl> <dbl>
##  1   679  0.79 0.889
##  2   292  0.44 0.663
##  3  1012  0.56 0.748
```

```
##  4   493  0.79 0.889
##  5   582  2.7  1.64
##  6  1156  3.64 1.91
##  7   997  4.73 2.17
##  8  2189  9.5  3.08
##  9  1097  5.34 2.31
## 10  2078  6.85 2.62
## # … with 40 more rows
```

```
plot(NewDataModel$rootY, NewDataModel$x, main = 'Electric Utility Squareroot
model',
     xlab = 'Peak hour demand', ylab = 'Total monthly energy usage')
abline(lm(x ~ rootY, data = NewDataModel), col = "blue")
```

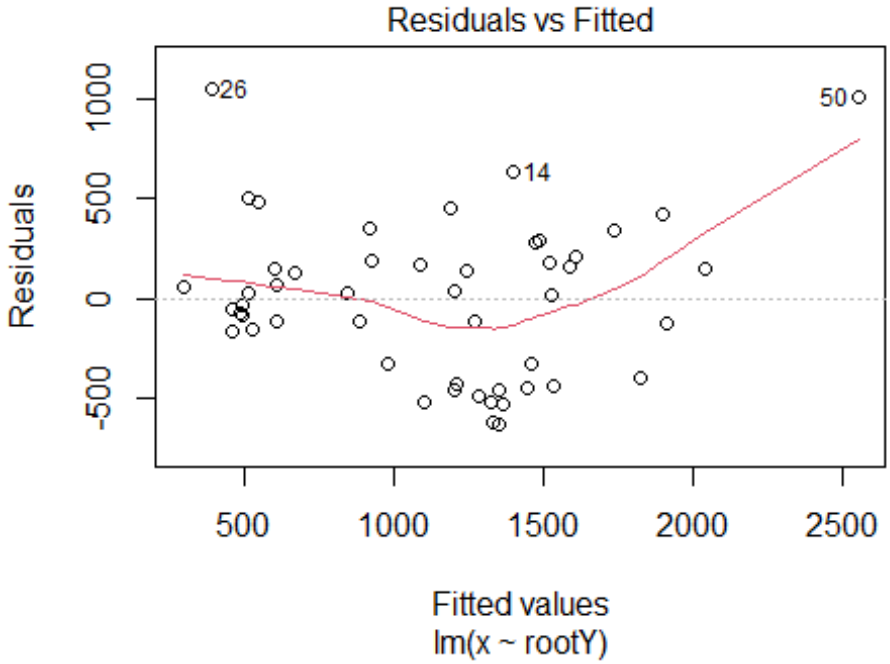


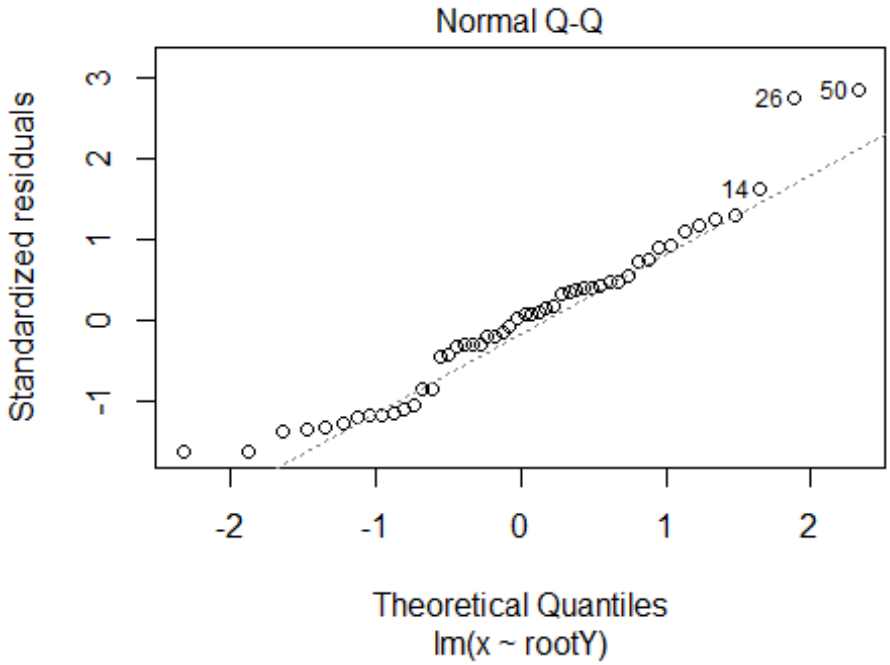

```
rootY.regression <- lm(x ~ rootY, data = NewDataModel)
summary(rootY.regression)
```

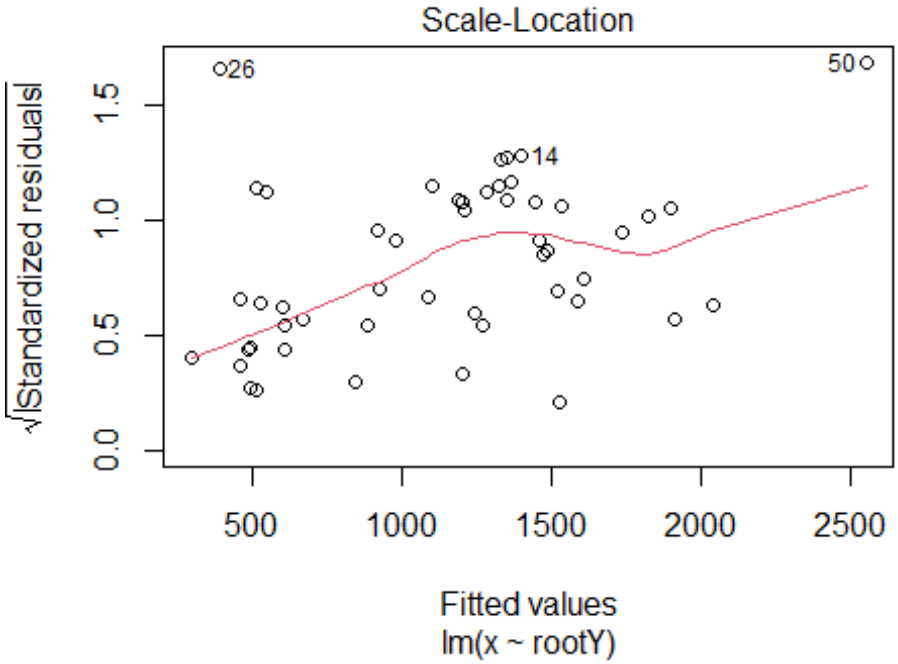
```
##
## Call:
## lm(formula = x ~ rootY, data = NewDataModel)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -623.08 -323.53   22.36  188.18 1046.39
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

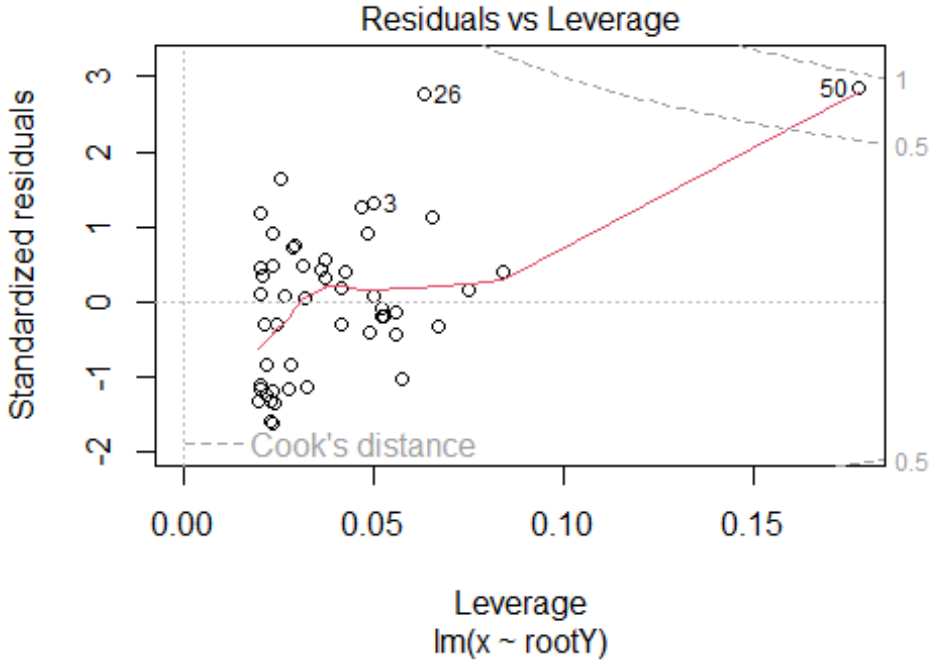
```
## (Intercept)    23.72     133.42   0.178     0.86
## rootY         653.56      71.56   9.132 4.53e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 391.2 on 48 degrees of freedom
## Multiple R-squared:  0.6347, Adjusted R-squared:  0.6271
## F-statistic:  83.4 on 1 and 48 DF,  p-value: 4.532e-12

plot(rootY.regression)
```


Residuals vs Fitted
Residuals
Fitted values
lm(x ~ rootY)
26
14
50
-500
0
500
1000
1500
2000
2500


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(x ~ rootY)
26
50
14
-2
-1
0
1
2
3





The modification of y, which took the form of a square root, was responsible for the observed change in the inequality of variance in part (d.). When compared to the prior plot, there has been a noticeable change in the manner in which the points are grouped together. The points have been moved toward the center, and there are fewer patterns visible than

there were in the earlier form. The points that are considered to be outliers are still present, and there is still a skew and a cluster to the left of the graph. The modification of y has led to a more significant change in the assumption, which has also been affected.