Midterm 1

Luke Letizia

February 22, 2018

Question 1

Prove the following: $r^2 = R^2$ where $R^2 = \frac{RegressionSS}{TotalSS}$

$r = \frac{1}{(n-1)s_x s_y} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$

$R^2 = \frac{\sum_{i=1}^{n} (\hat{y} - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$

$\frac{1}{(n-1)s_x s_y} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \frac{\sum_{i=1}^{n} (\hat{y} - \bar{y})^2}{(n-1)s_y^2}$

$\frac{(n-1)s_y^2}{(n-1)s_x s_y} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} (\hat{y} - \bar{y})^2$

$\frac{s_y}{s_x} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} (\hat{y} - \bar{y})^2$

Question 2

a.

Mean = 2145.251

Standard Deviation = 2459.051

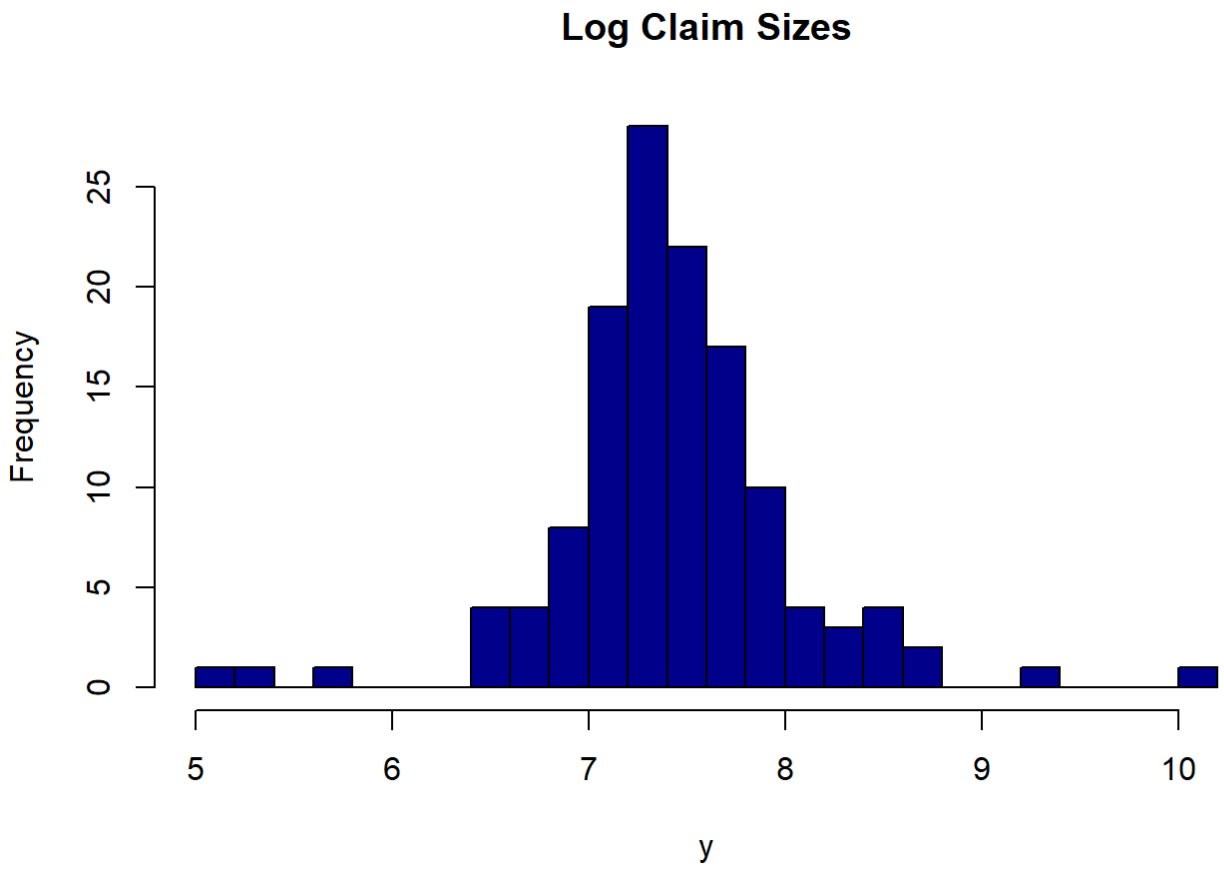
Minimum = 176.0436

Maximum = 25526.64

b.

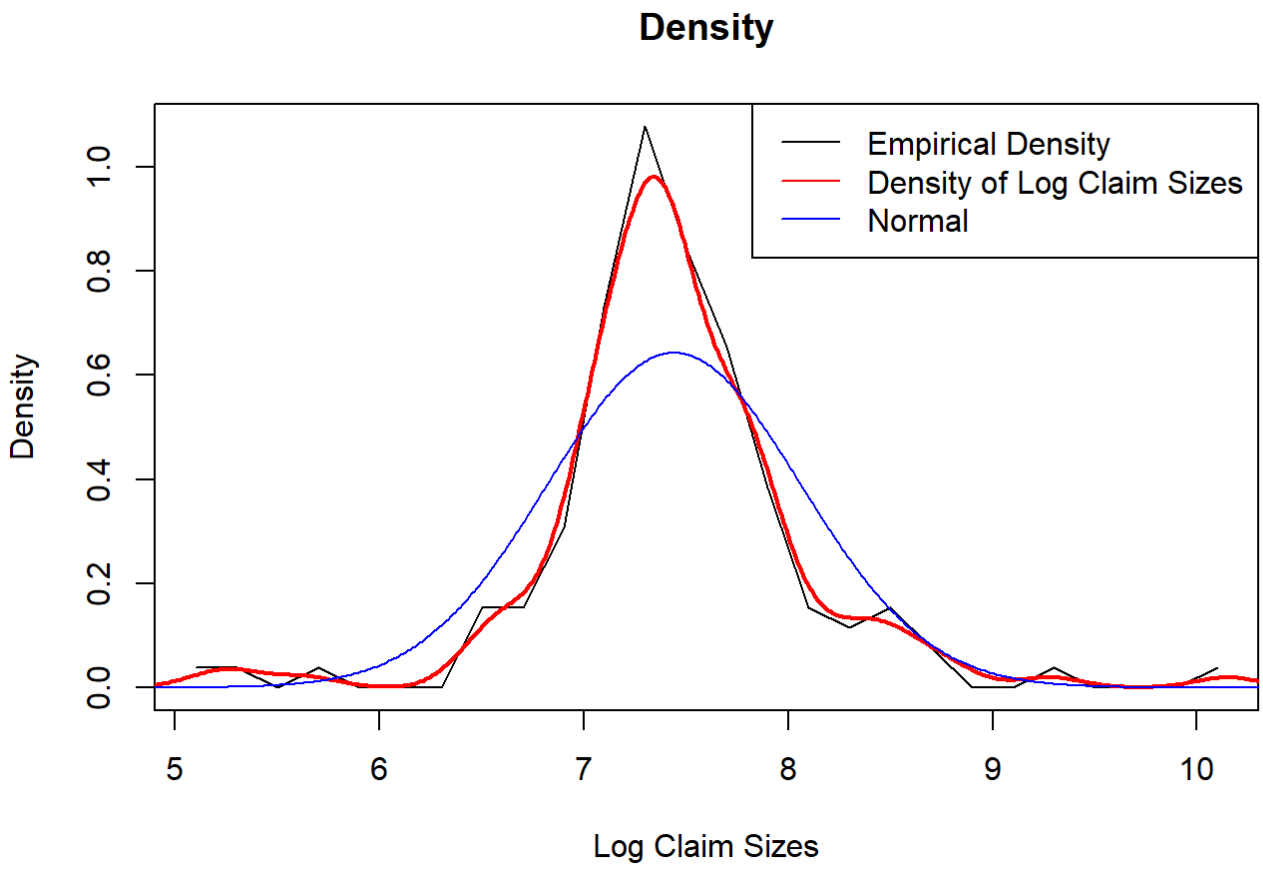
```
y = (log(data.1$V1))
```

c.



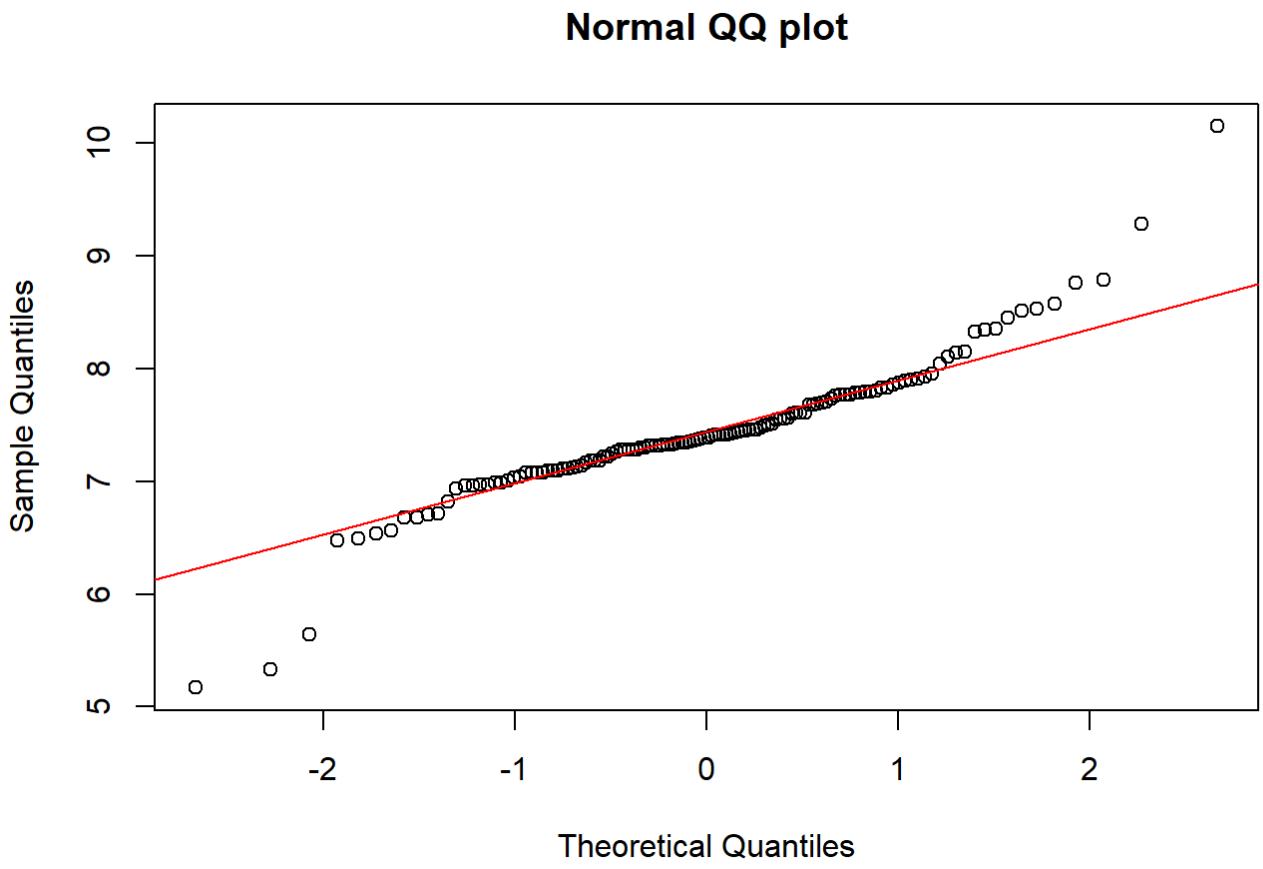
I chose to set the grid size to 30. Because the sample size is n=130 I wanted to choose a number of breaks that was less than 130. I decided to pick 30 because it is conveyed very nicely on the graph, you can see where it begins to incline and it is easy to read and analyse.

d.



As the graph suggests, the log of claims is at it's maximum density around 7.25. The empirical density and the density of the log claim sizes are nearly identical, besides the fact that the empirical function is much smoother. The normal curve is similar in shape but differs greatly from the maximum data points.

e.



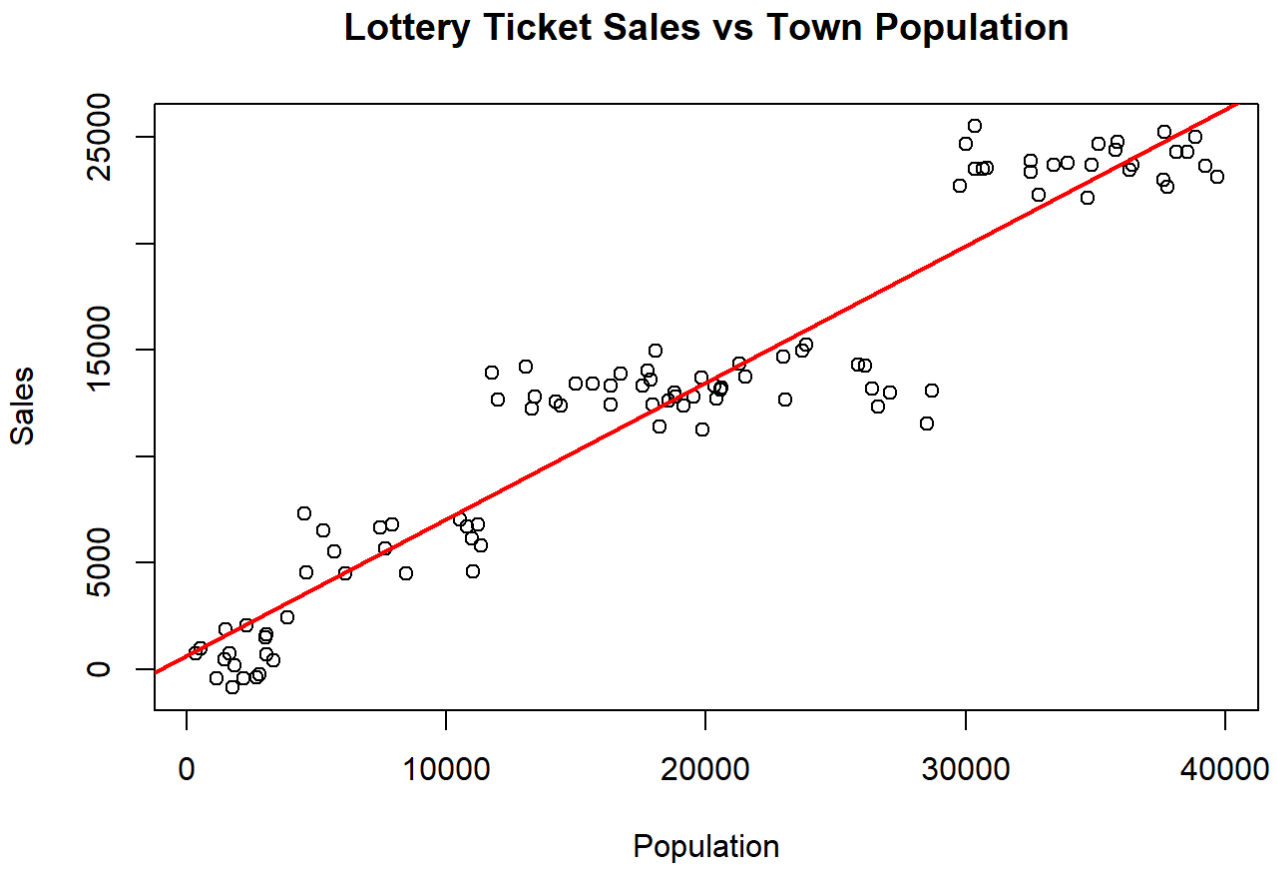
The qqplot shows that the data doesn't follow normal distribution because several data points on the extremes of the graph deviate greatly from the best fit line.

f.

According to the qqplot, the values on the extremes of the plot are either far under or above the abline, meaning they deviate very far from the abline and would create scenarios where the insurance company would under-price significantly.

Question 3

a.



β_0 = 653.3349

β_1 = 0.6410505

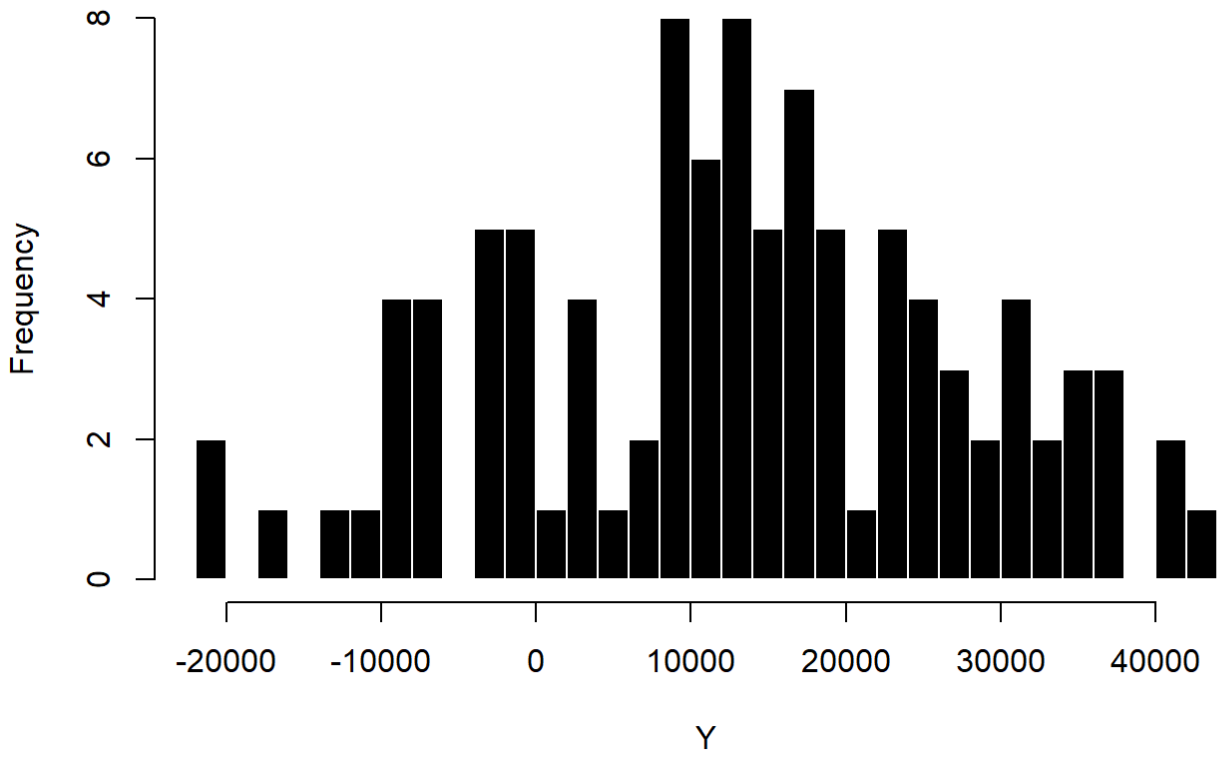
The model shows that the more the population increases, the more the sales also increase. This assumption is very reasonable and sensical.

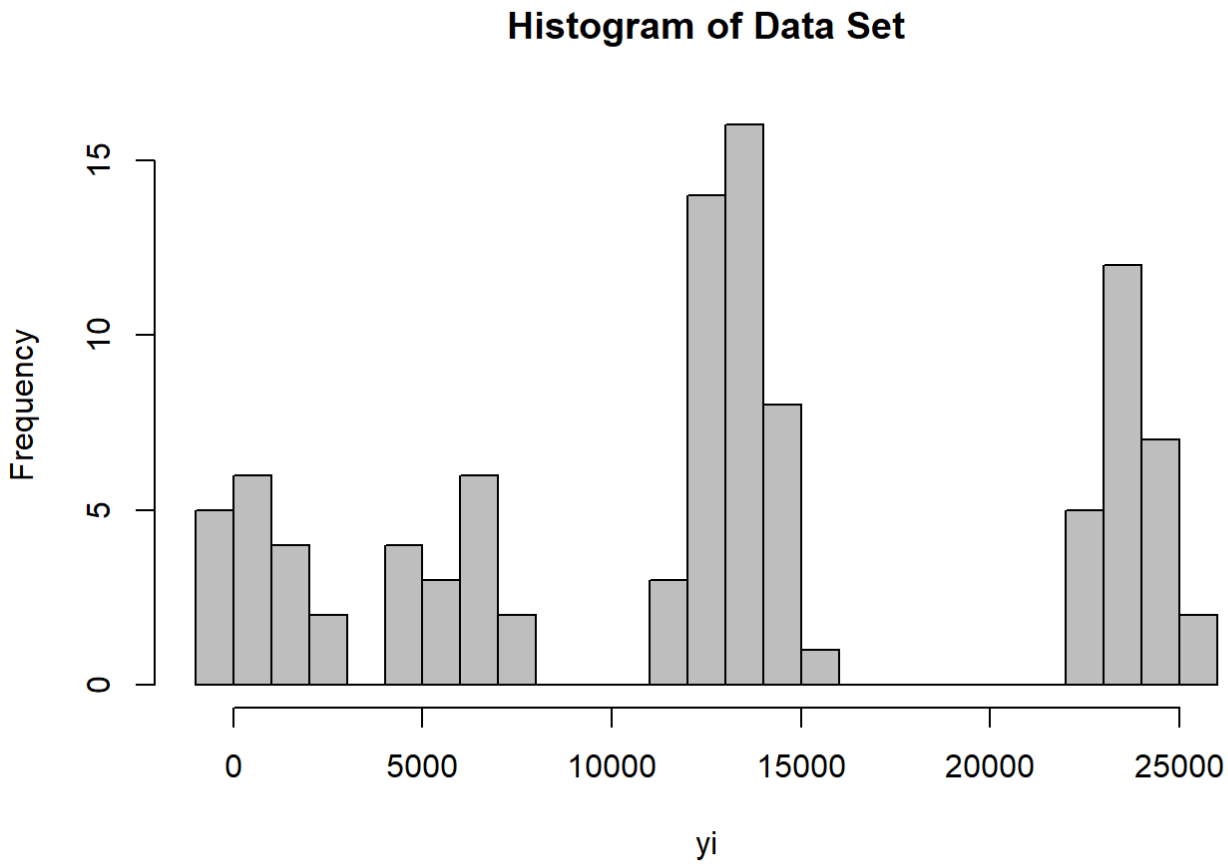
b.

$\hat{y}$ = 7063.839

c. The random variable $\hat{Y}$ is an expected value of the linear regression model. Y differs from $\hat{Y}$ in that $\hat{Y}$ is an expected output value from plugging in a given x value.

d.


Histogram of Simulation
Frequency
0
2
4
6
8
-20000
-10000
0
10000
20000
30000
40000
Y


Histogram of Data Set
Frequency
0
5
10
15
0
5000
10000
15000
20000
25000
yi

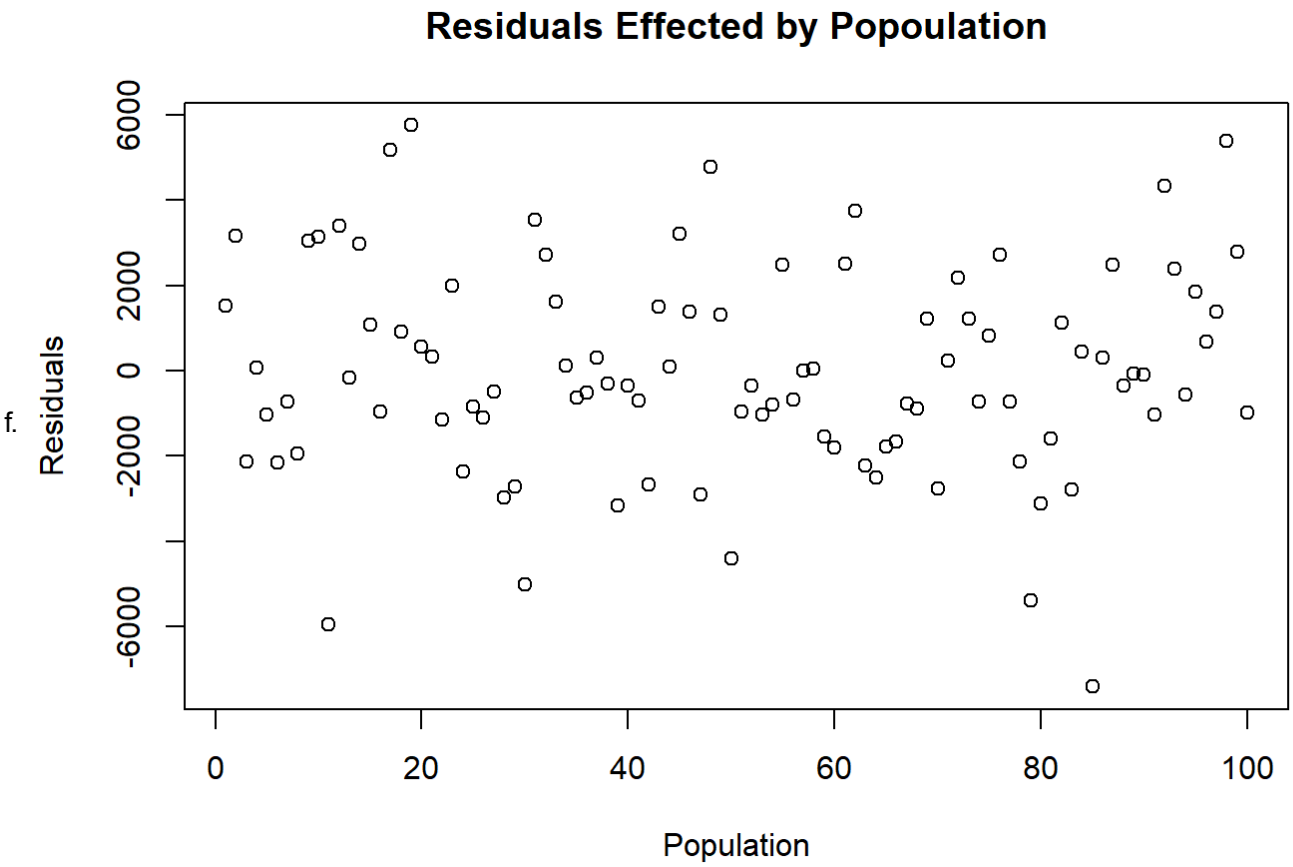
e.

Confidence Bounds at 95%:

Lower:6456.396

Upper:7671.583

f.



Throughout the entirety of the plot, the residuals are scattered everywhere but tend to be more populated around the point where residuals equals 0. There doesn't seem to be much relationship or correlation between the population size and the outputted residuals.

g.

I think the regression model is a valid estimator for future sales because we are able to clearly read and analyze what the data is telling us and how we can expect it to act in the future. It would just depend on how precisely and consistently we can predict the future sales.