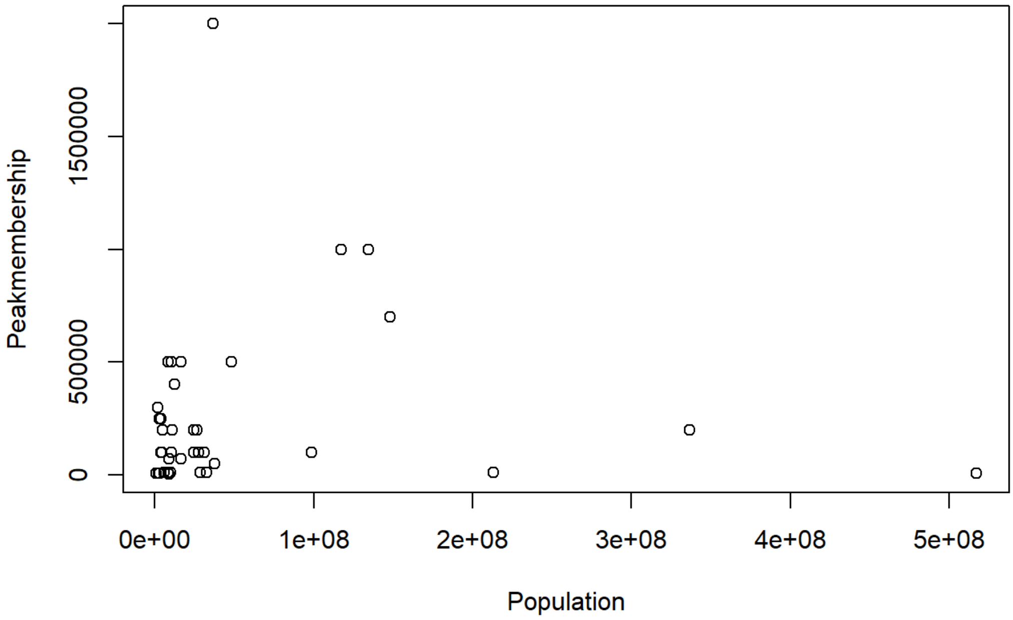
Step A:

Draw the scatter plot

```
#read in data in A2.csv and assign it to mydata
mydata <- read.csv("A2.csv",sep=",")
#Extract x and y from mydata and assign them to x1 and y1 respectively
x1 <- mydata$Population; y1 <- mydata$Peakmembership
#Draw the scatter plot with proper labels
plot(x1, y1,xlab = "Population", ylab = "Peakmembership")
```



Step B:

```
#fit a linear model and assign it to fit
fit <- lm(y1 ~ x1)
```

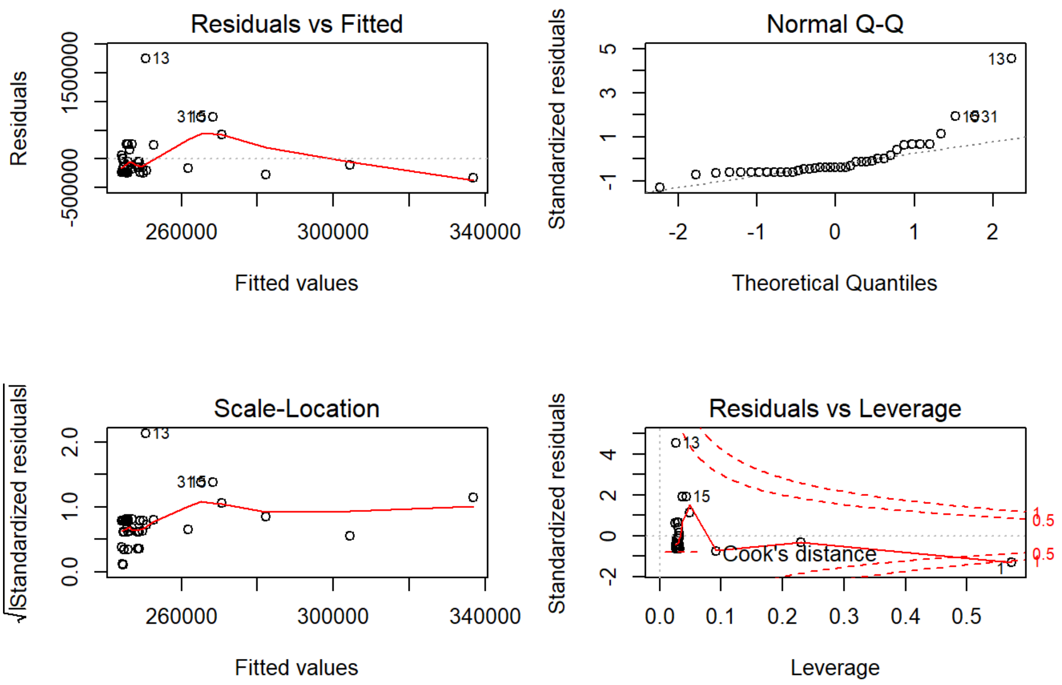
Step C:

From the Residuals vs Fitted plot below I can tell that the assumption of linearity is likely violated.

From the Scale-Location plot below i can tell that the assumption of constant variance of residuals is likely violated too.

I’ll enter Step D.

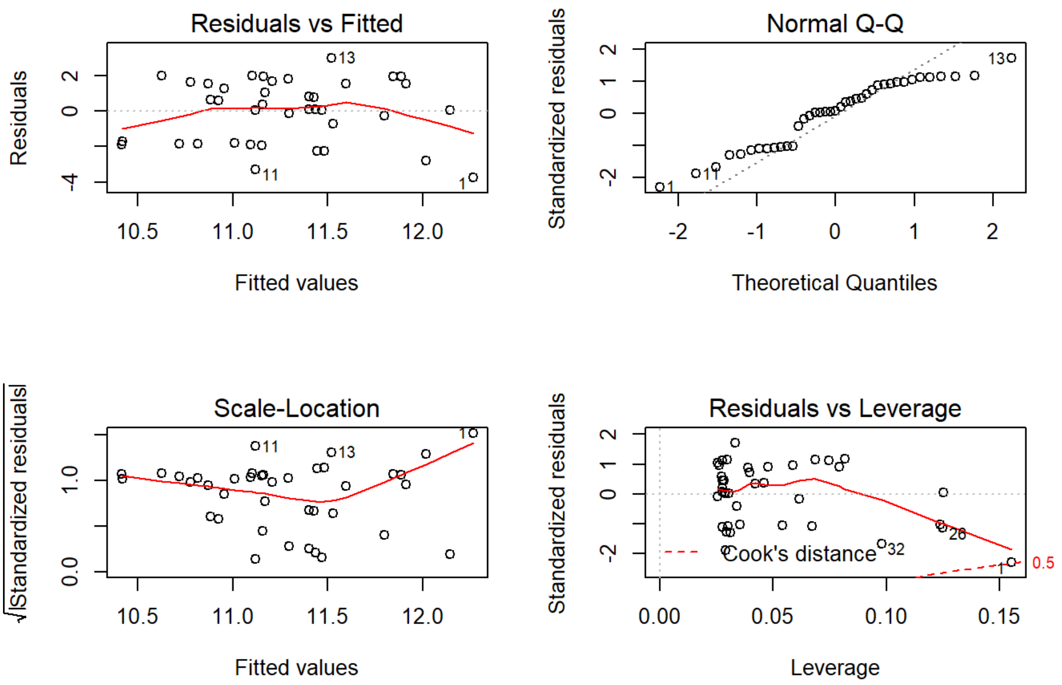
```
#produce the four plots 2 by 2 at the same time
par(mfrow=c(2,2));
plot(fit)
```



Step D (a):

Applied logarithmic transformation to both x and y

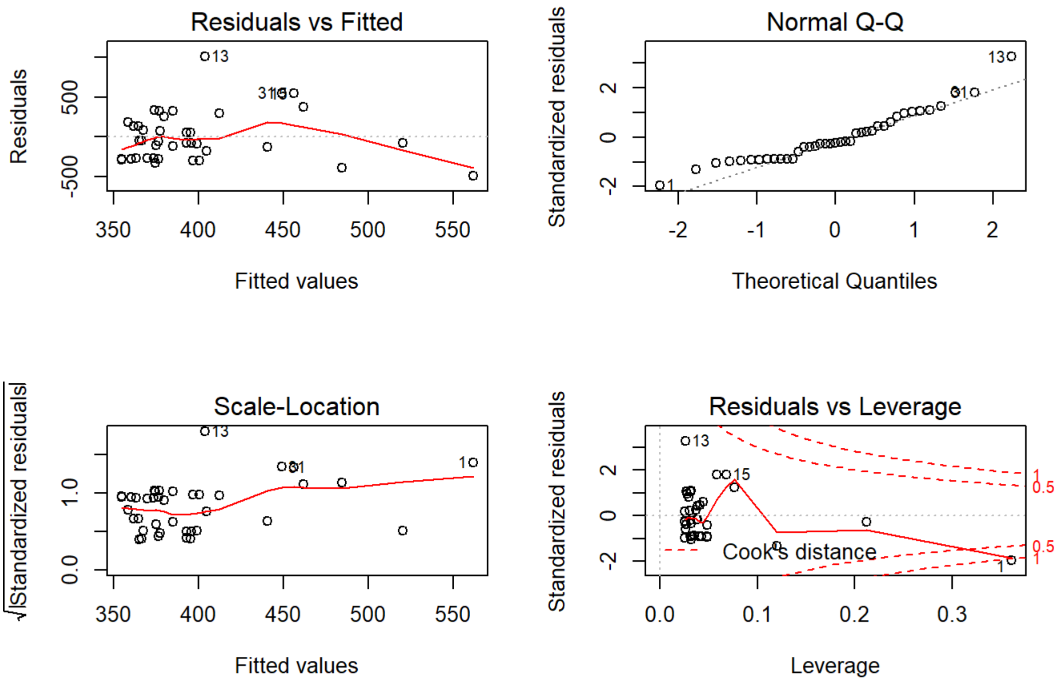
```
#Applied logarithmic transformation to both x and y
x<-log(x1); y<-log(y1)
#re-fit model and produce the four plots 2 by 2 at the same time
fit2 <- lm(y~x)
par(mfrow=c(2,2)); plot(fit2)
```



Step D (b):

Applied square-root transformation to both x and y

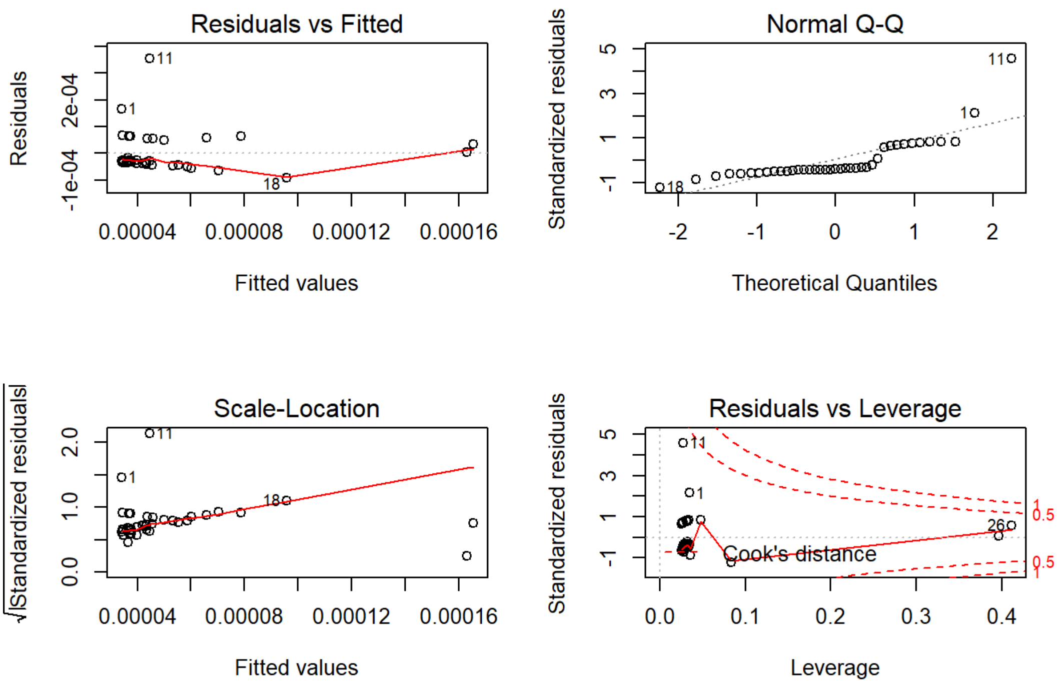
```
#Applied square-root transformation to both x and y
x2<-sqrt(x1); y2<-sqrt(y1)
#re-fit model and produce the four plots 2 by 2 at the same time
fit3 <- lm(y2~x2)
par(mfrow=c(2,2)); plot(fit3)
```



Step D (c):

Applied reciprocal transformation to both x and y

```
#Applied reciprocal transformation to both x and y
x3<-1/(x1); y3<-1/(y1)
#re-fit model and produce the four plots 2 by 2 at the same time
fit4 <- lm(y3~x3)
par(mfrow=c(2,2)); plot(fit4)
```



Step C (re-visited):

After comparing the four plots for three ways of transformation above. I conclude that logarithmic transformation gives a better result in terms of overcoming problems with violated assumptions of linearity and constant variance of residuals. I'll work with fit2 from now on.

Step E:

Firstly, there are 39 observations in this data set. I'll treat it as a small data set.

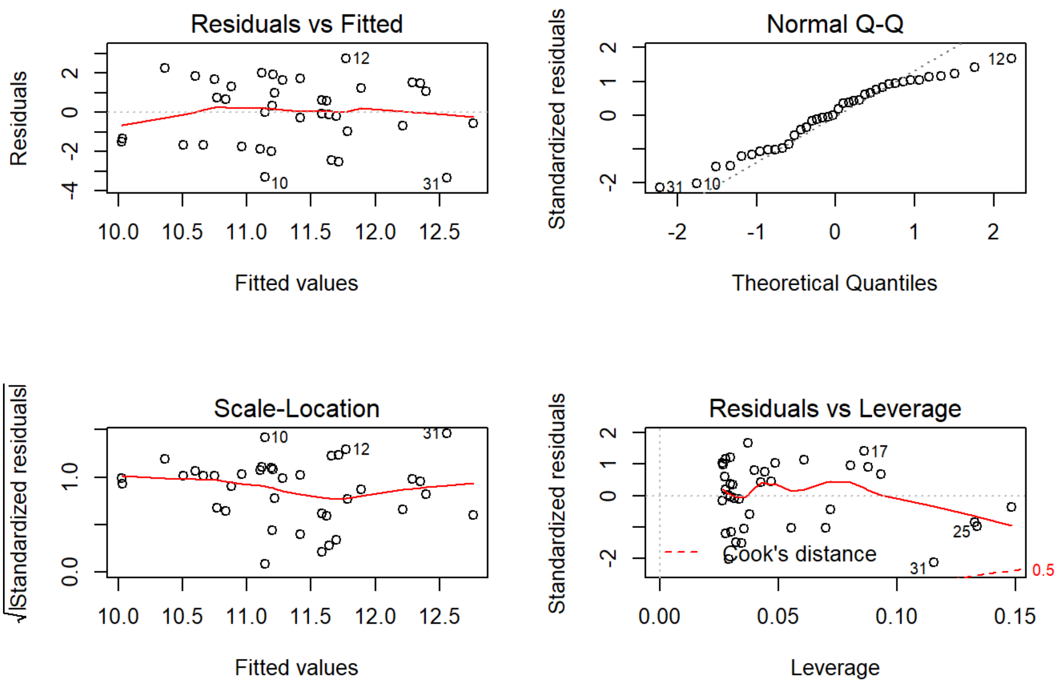
Outlier: From Residual vs Leverage plot, i can tell there is one outlier with |standardized residual| > 2. It has label 1.

Leverage: From Residual vs Leverage plot, I see no point with Cook's Distance > 1. So there's no leverage point.

Step F:

I'll report results with and without the outlier (point with label 1).

```
#eliminated the outlier with label 1 from mydata and assigned the result to data.new.
data.new <- mydata[-c(1), ]
#extracted x and y from data.new as x1.new and y1.new respectively
x1.new <- data.new$Population; y1.new <- data.new$Peakmembership
#applied logarithmic transformation to both x and y.
x.new<-log(x1.new); y.new<-log(y1.new)
#re-fit model and produce the four plots 2 by 2 at the same time
fit2.new <- lm(y.new~x.new)
par(mfrow=c(2,2)); plot(fit2.new)
```



Step J:

The sample size is 39, 39 < 200. So the sample size is not large. From the Normal Q-Q plot i can tell that there is heavy tails for model with and without the outlier, but I'll ignore step L and move to step M.

```
#produce sumary for model with outlier
summary(fit2)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.748 -1.814  0.111  1.552  2.990
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   6.6028     3.0452   2.168   0.0366 *
## x             0.2822     0.1830   1.542   0.1315
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.776 on 37 degrees of freedom
## Multiple R-squared:  0.0604, Adjusted R-squared:  0.03501
## F-statistic: 2.378 on 1 and 37 DF,  p-value: 0.1315
```

```
#produce sumary for model without outlier
summary(fit2.new)
```

```
##
## Call:
## lm(formula = y.new ~ x.new)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3476 -1.4655  0.1638  1.4288  2.7371
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.9899     3.0521   1.307   0.1994
## x.new         0.4468     0.1845   2.421   0.0206 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.668 on 36 degrees of freedom
## Multiple R-squared:  0.1401, Adjusted R-squared:  0.1162
## F-statistic: 5.864 on 1 and 36 DF,  p-value: 0.02062
```

Step M:

The model with outlier:

From the output of summary() command above, the slope has a p-value of 0.1315, so i can tell that the slope is not statistically significant at alpha = 0.05.

The R-square value is 0.064, which is super low. The amount of variance in Peakmembership explained by this model is only 6.4%.

Scale-Location plot shows sign of non-constant variance of residuals, the amount of variance tends to be larger as fitted value increases.

The model without outlier:

From the output of summary() command above, the slope has a p-value of 0.0206, so the slope is statistically significant at alpha = 0.05. On the other hand, the intercept has a p-value of 0.1994 and is not statistically significant at alpha = 0.05.

The R-square value is 0.1401, which is higher than before, but is still too small. The amount of variance in Peakmembership explained by the model is only 14.01%.

Scale-Location plot shows sign of non-constant variance of residuals, the amount of variance tends to be larger at the middle.

For both models, the Normal Q-Q plot shows that the residuals have heavy tails.

Putting it together, I conclude that there does not exist a (linear) relationship between the predictor (Population) and the response (Peakmembership) variable.