Assignment3

aditya venugopalan a1899824

2023-10-15

#Executive Summary:

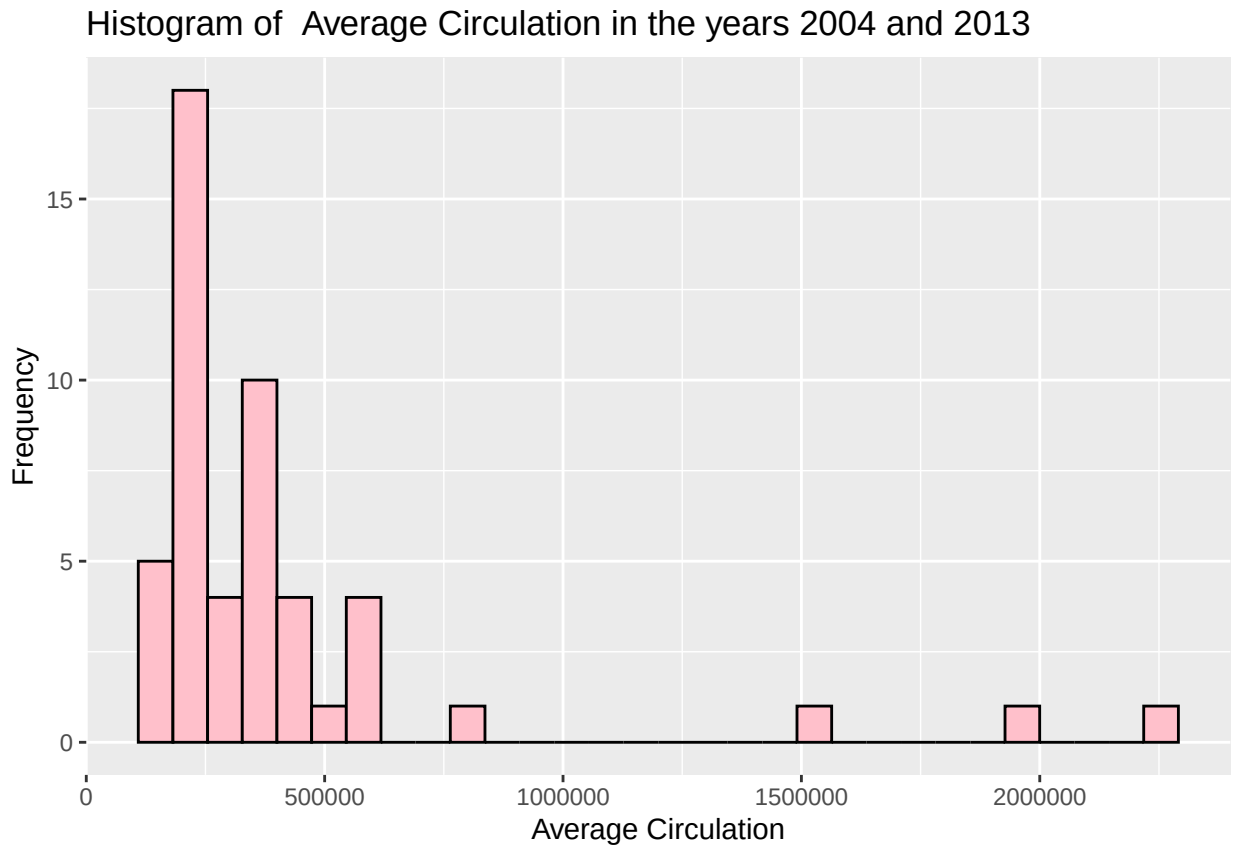
Due to decline in the recent readership and current circulation of investigative journalism, necessary choices are supposed to be made , whether to invest or not in the Boston Sun - times investigative journalism under the flagship newspaper of Masthead Media company. Such bold choices cannot be made without any investigation between the relationship between the number of Pulitzer prizes won and the average newspaper circulations and percentage of newspaper circulation between 2004 and 2013. predictions were derived from the help of some diagrams along with two models , telling us that there are positive relationship between both the numbers of pulitzer prizes won, and the average newspapers circulation as well as percentage change in newspaper's circulations. Which in other words or non technical words can be said that number of the prizes won by the publications increases , with the average newspapers' circulations and the percentage change in newspapers circulation also increases. Also in larger variations in the number of prizes won causes smaller variations , which means if there is a small decrease in investment will result in a larger decrease in the number of Pulitzer prizes won. So according to my analysis Masthead Media should invest more in investagative journalism .

```
#Q.1(b)
pulitzer <- pulitzer %>%
mutate(average_circ = (circ_2004+circ_2013)/2)
head(pulitzer)
```

```
## # A tibble: 6 x 6
##   newspaper           circ_2004 circ_2013 change_0413 prizes_9014 average_circ
##   <chr>                   <dbl>     <dbl>       <int>       <dbl>        <dbl>
## 1 USA Today             2192098   1674306         -24           2     1933202
## 2 Wall Street Journal   2101017   2378827          13          50     2239922
## 3 New York Times        1119027   1865318          67         117     1492172.
## 4 Los Angeles Times      983727    653868         -34          85      818798.
## 5 Washington Post        760034    474767         -38         100      617400.
## 6 New York Daily News    712671    516165         -28           6      614418
```

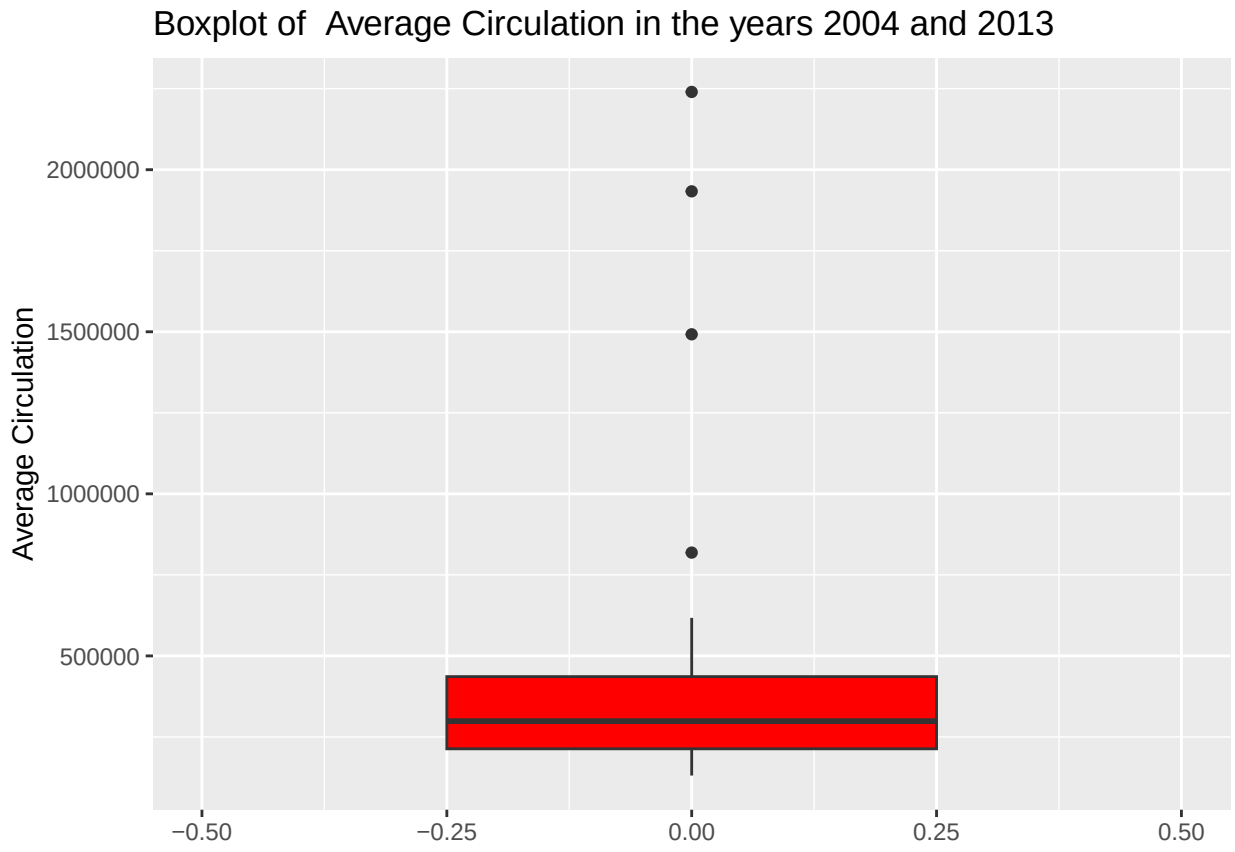
```
#Q.2(a)
ggplot(pulitzer, aes(x=average_circ)) +
geom_histogram(fill = "pink", color = "black") +
  labs(title = "Histogram of  Average Circulation in the years 2004 and 2013",
x = "Average Circulation", y = "Frequency")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
#theme_bw()
```

```
pulitzer %>%
  ggplot(aes(y = average_circ)) +
  geom_boxplot(show.legend = FALSE, fill = "red", width = 0.5) + xlim(-0.5, 0.5) +
  labs(title = "Boxplot of  Average Circulation in the years 2004 and 2013" , y = "Average Circulation")
```




```
summary(pulitzer$average_circ)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  131004  213509  298851  412442  436152 2239922
```

```
sd(pulitzer$average_circ)
```

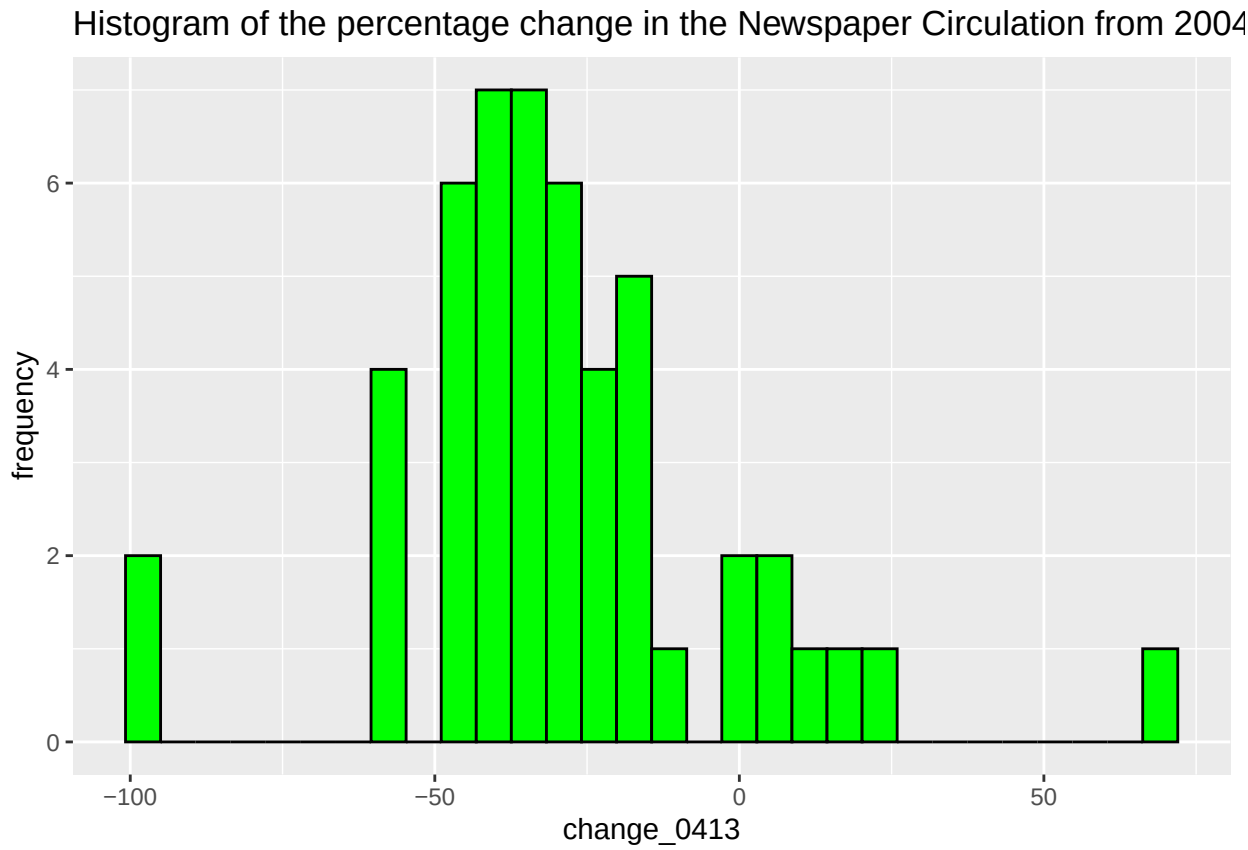
```
## [1] 410339.9
```

```
# According to my analysis with context to histogram and boxplot following conclusions can be made.
#Shape: Right skewed and unimodal
#Location : The Median circulation is around 298851 (from box plot or summary)
#Spread: The minimum and maximum of the data range from values 131004 to 2239922 . Also it is evident t
#Outliers: There are four outliers that can be observed from the boxplot , which approximately have the
```

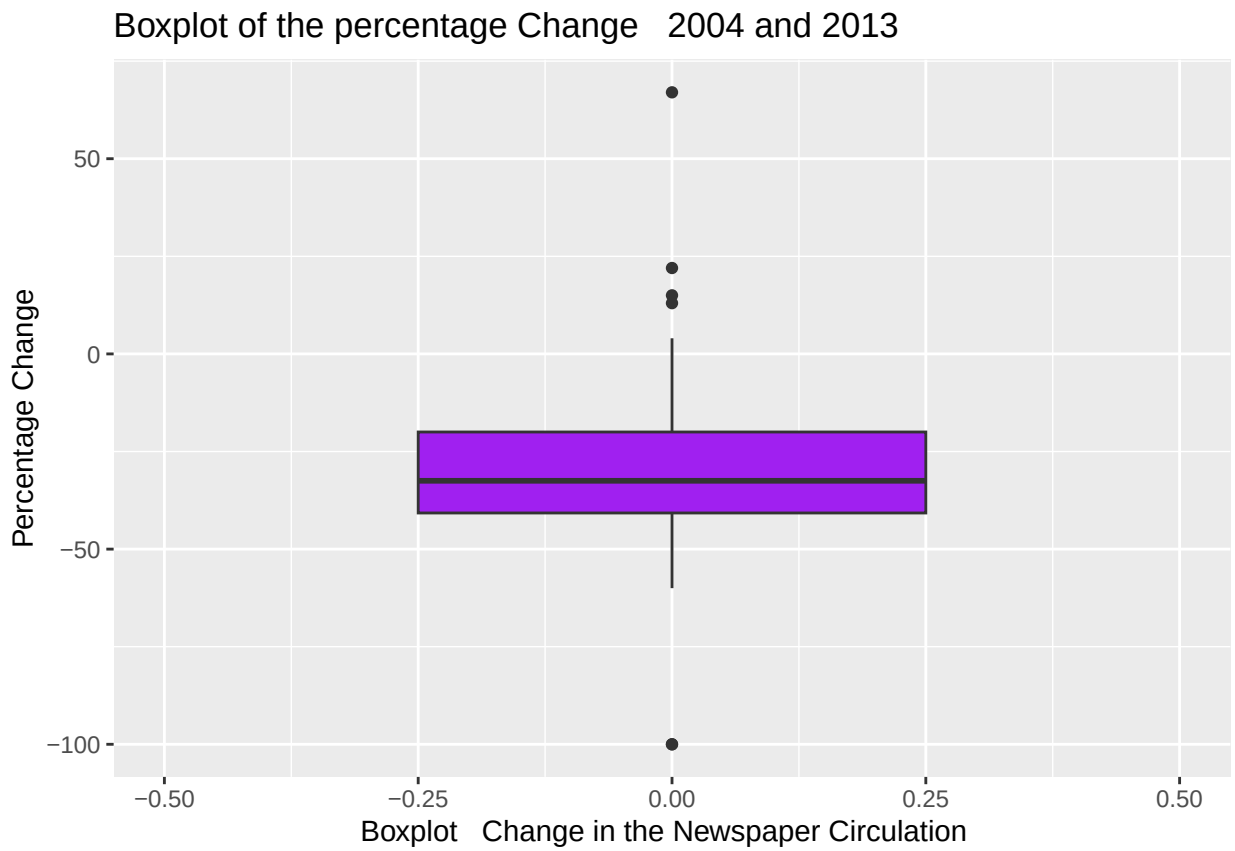
```
# Q.2(b)
pulitzer %>%
  ggplot(aes(x= change_0413))+
  geom_histogram(bin = 30 ,col = "black", fill= "green")+
  labs(title = "Histogram of the percentage change in the Newspaper Circulation from 2004 to 2013", y="
```

```
## Warning in geom_histogram(bin = 30, col = "black", fill = "green"): Ignoring
## unknown parameters: ‘bin‘
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
ggplot(pulitzer, aes(y=change_0413)) +
  geom_boxplot(show.legend = FALSE, fill = "purple", width = 0.5) + xlim(-0.5, 0.5) + labs(title = "Boxp
```




```
summary(pulitzer$change_0413)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -100.00  -40.75  -32.50  -29.20  -20.00   67.00
```

```
sd(pulitzer$change_0413)
```

```
## [1] 27.06681
```

#Shape : The shape is symmetric and unimodal # Location : The distribution of variable is seen between from -100% and 67%, while yhe median is located at -32.50% with the mean located at -29.20 which can help to conclude that the shape is roughly or very close to symmetric. # Spread : The interquartile range is between -40.75% and -20% # Outliers : There are 5 outliers located approximately at -100%, 10%, 15%, 25%, 70% respectively # Q.2(c) #after concluding the variable change_0413 it is almost symmetric . Therefore the only varaible which required a transformation according to my analysis is "average_circ".

```
#Q.3(a)
puli_circ <- lm(log(average_circ) ~ prizes_9014, data = pulitzer)
summary(puli_circ)
```

```
##
## Call:
## lm(formula = log(average_circ) ~ prizes_9014, data = pulitzer)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.8154 -0.3191 -0.1600  0.1787  1.9597
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.487127   0.085510 146.032  < 2e-16 ***
## prizes_9014  0.013911   0.002953   4.711 2.22e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5076 on 47 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.3207, Adjusted R-squared:  0.3063
## F-statistic: 22.19 on 1 and 47 DF,  p-value: 2.22e-05
```

```
exp(puli_circ$coefficients[1])
```

```
## (Intercept)
##    264905.1
```

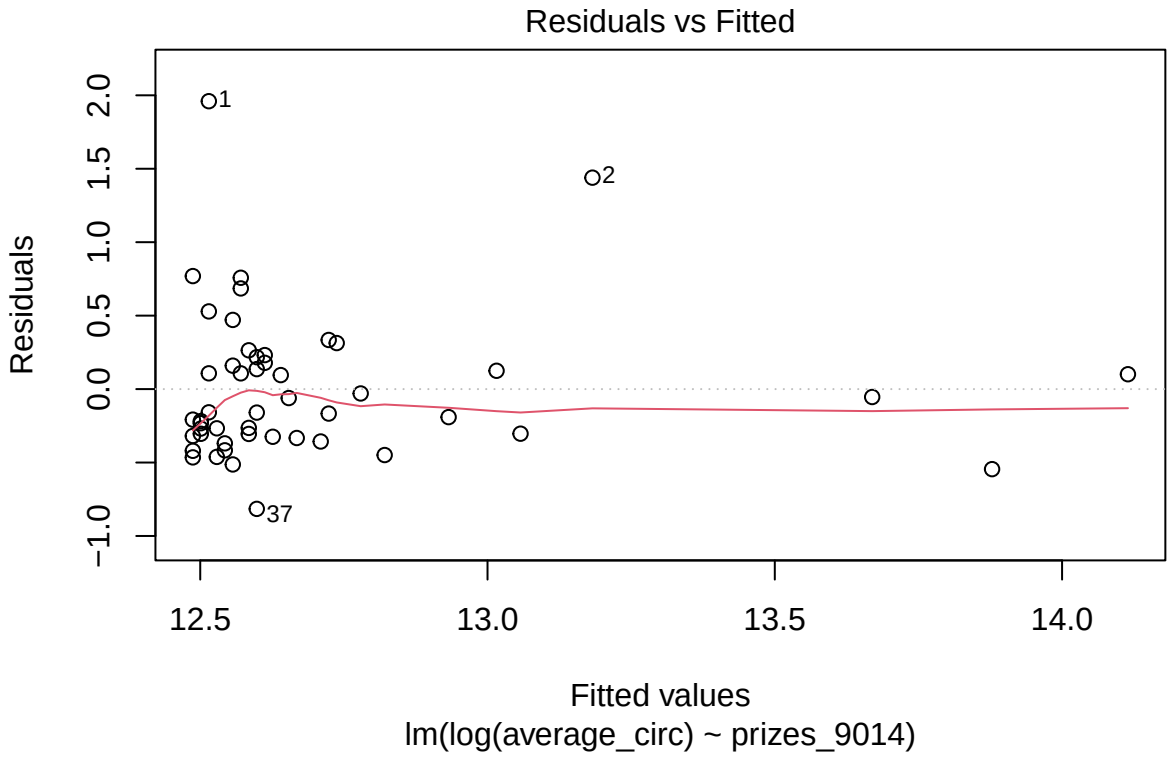
#From above we can conclude The intercept is 12.487127 and slope is 0.013911 #Interpretation of intercept: In cases where if a newspaper wins 0 Pulitizer Prizes , we can expect it to have a log circulation of 12.415 towards the the end of that 25 years of period of time , which means actual circulation is of 264905.1 . #Interpretation of Slope: If the number of Pulitizer Prizes won by a newspaper in a define period of time increases by one , the log circulation is expected to increase by 0.017. There is a statistically significant relationship between Pulitzer Prizes and newspaper circulation

```
# Q.3(b)
pr_change <- lm(change_0413 ~ prizes_9014, data = pulitzer)
summary(pr_change)
```

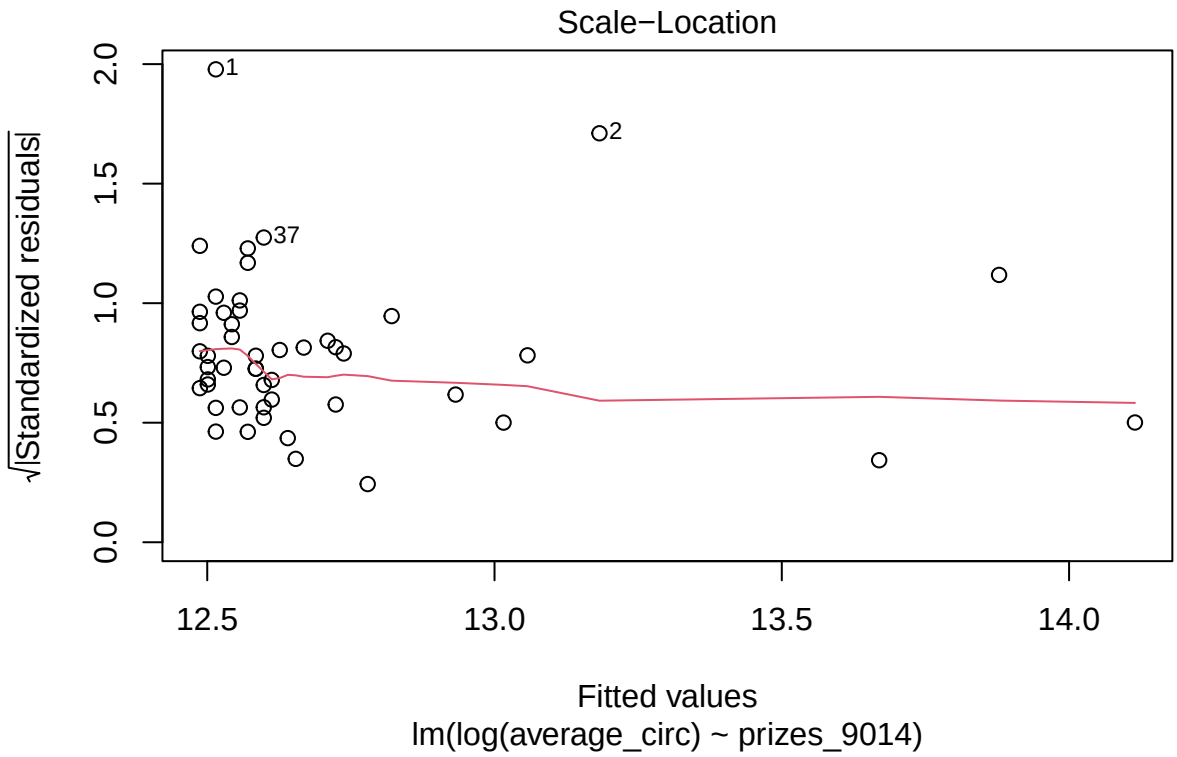
```
##
## Call:
## lm(formula = change_0413 ~ prizes_9014, data = pulitzer)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -67.89 -10.61  -3.15  13.33  56.53
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -35.2411     4.3530  -8.096 1.83e-10 ***
## prizes_9014   0.3907     0.1503   2.599   0.0124 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 25.84 on 47 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.1257, Adjusted R-squared:  0.1071
## F-statistic: 6.756 on 1 and 47 DF,  p-value: 0.01245
```

```r
# the value of slope is 0.3907 and the value of intercept is -35.2411
#Interpretation of intercept: If a newspaper wins 0 pulitzer prizes, we can expect its circulation to d
#Interpretation of the slope: If the number of pulitzer Prizes won by a newspaper increases by 1 , the
```

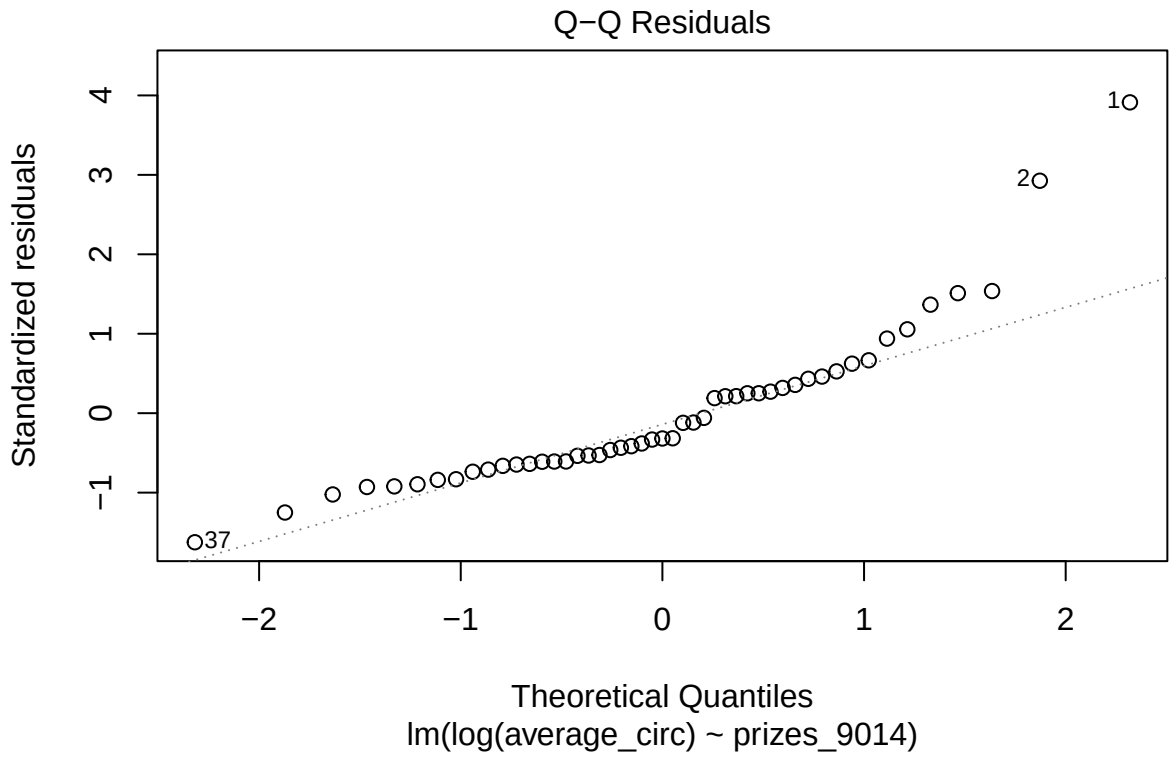
```r
#Q.3(c)
# Linearity
plot(puli_circ, which = 1)
```



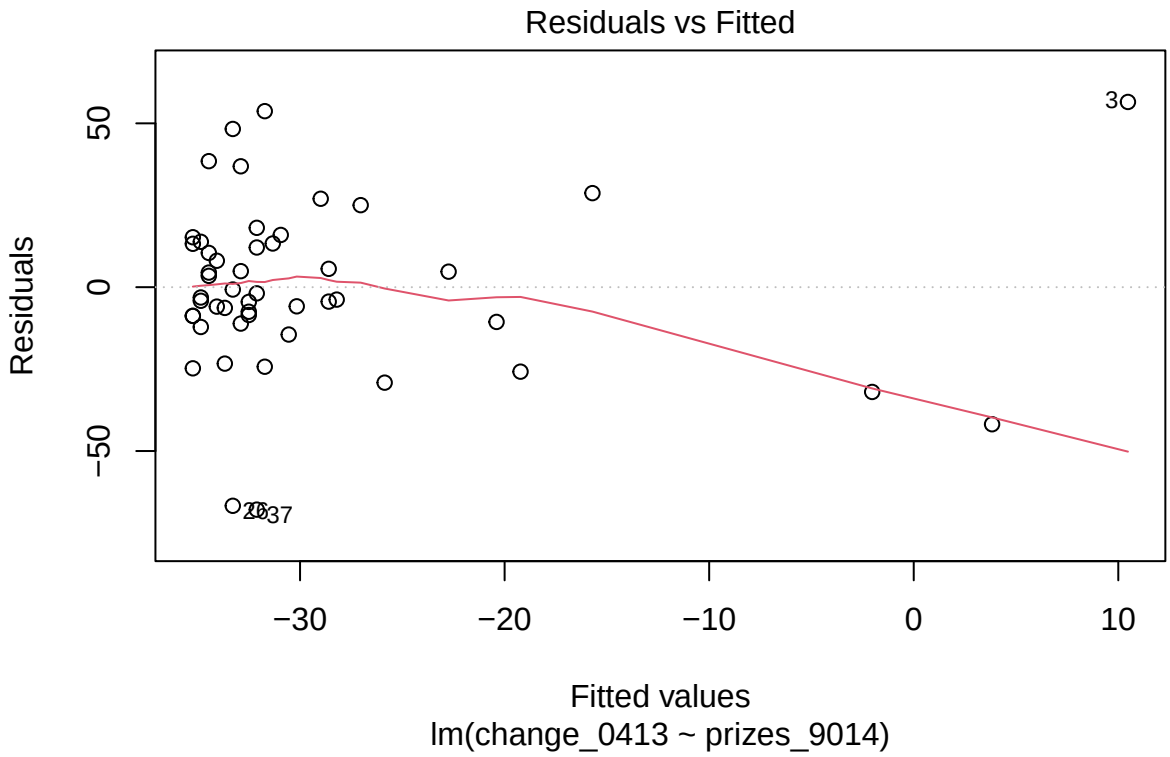

```r
#homoscedasticity
plot(puli_circ, which = 3)
```



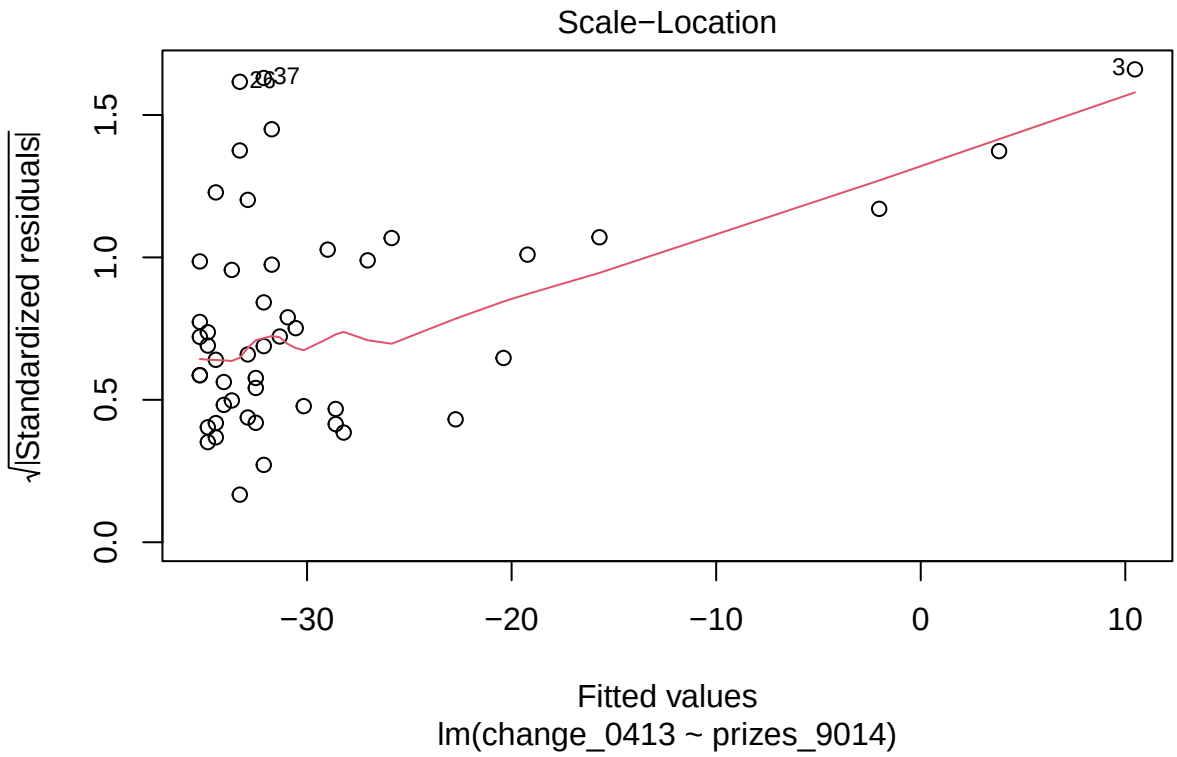

```
#normality
plot(puli_circ , which = 2)
```



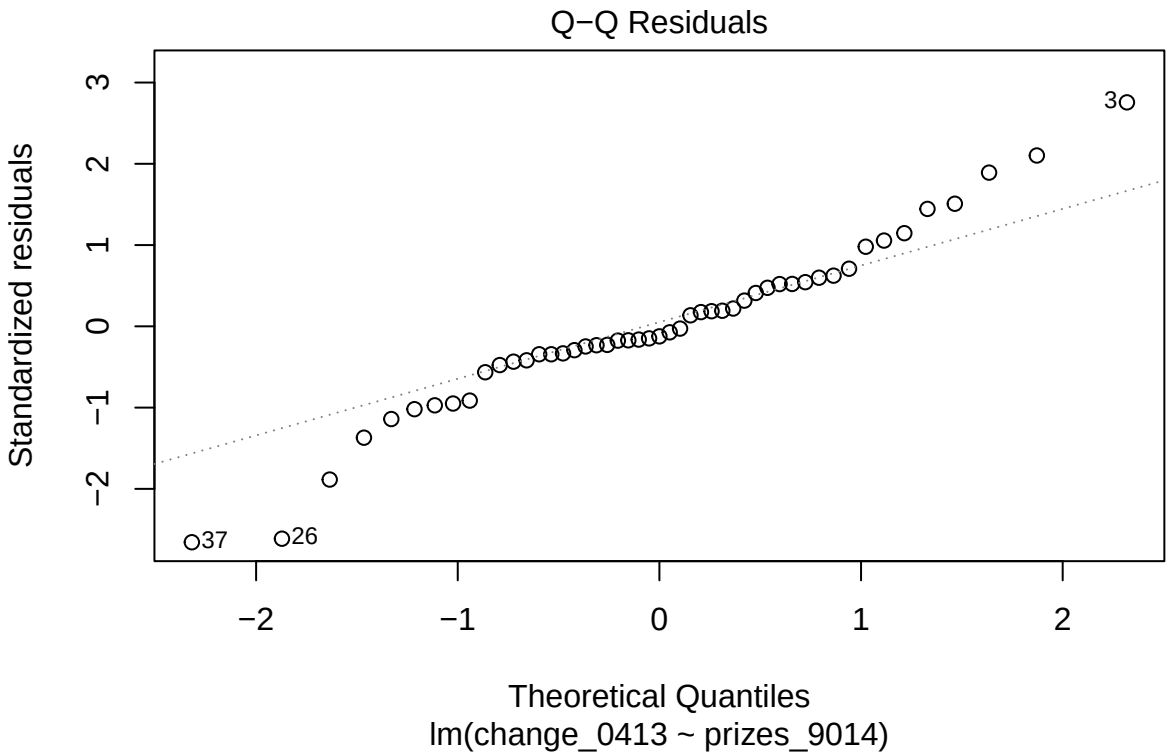

```
# Independence
#Linearity
plot(pr_change, which =1)
```




```
#Homoscedasticity
plot(pr_change, which = 3)
```




```
#normality
plot(pr_change, which=2)
```



#Independence: It is quiet impossible or difficult to tell the independence since the circulation of newspaper of one state will unlikely affect other states , but it is also possible it might affect since time measurement for all is same so therefore determing the independence is not possible by just working on the obtained data set without knowing how it was obtained. This holds for both the developed models # For model 1 #Linearity : Overall linearity can be observed also very less changes in trends , therefore the assumption of linearity can be satisfied. #Homoscedasticity: The scatter plot shows almost no obvious trend between the standardized residuals and fitted values . Which help us to conclude that this model fulfills the homoscedasticity assumption. #Normality: Even though there are very few outliers majority of the data points lie on the dotted line which indicates that the residulas are normally distributed . Thus , fulfilling the normality assumption.

For model 2

#Linearity : Once again here the plot is roughly straight with no curvature , which tells us the assumption of linearity is fulfilled . # Homoscedasticity : The plot tells us how there is increasing trend as we go from left to right.Therefore assumption of Homoscedasticity is not justified. # Normality : In the plot we can see that majority of points lie in the dotted line, which is a very important observation for us to conclude that that the normality assumption is satisfied.

```
# Q.4(a)
directions <- tibble(prizes_9014 = c(3,25,50))
tibble("WON PRIZES"  = directions$prizes_9014,
"Expected Circulation" =
predict(puli_circ, directions) %>% exp())
```

```
## # A tibble: 3 x 2
##   `WON PRIZES` `Expected Circulation`
##          <dbl>                  <dbl>
## 1            3                276194.
## 2           25                375080.
## 3           50                531076.
```

#If we compared the current circulation which is 457,258, it is clearly evident that only the cases where 50 Pulitzer Prizes are won will lead to an increase in circulation, which turns out to be 531076.4

```
# Q.4(b)
tibble("Won Prizes" = directions$prizes_9014,
       'Change in Circulations(%)'=
        predict(pr_change, directions)) %>%
  knitr::kable(digits = 0, format.args = list(big.mark = " ,"))
```

Won Prizes	Change in Circulations(%)
3	-34
25	-25
50	-16

```
# All possible strategic directions lead to an expected decrease in the circulation. However this is di
```

```
# Q.4(c)
circ_conf <- predict(puli_circ, directions, interval = "confidence", level=0.9) %>% exp()
circ_conf <- tibble(Prizes = directions$prizes_9014,
`Lower Bounds  ` = circ_conf[,2],
`Expected circulation` = circ_conf[,1],
`Upper Bounds  ` = circ_conf[,3] )

circ_conf %>% knitr::kable(digits = 0, format.args = list(big.mark = ","))
```

Prizes	Lower Bounds	Expected circulation	Upper Bounds
3	241,029	276,194	316,489
25	329,114	375,080	427,466
50	430,294	531,076	655,463

With the help of the following tibble we can make a conclusion that with 90% confidence , the average circulation of newspapers are between 241,029 and 316,489 when there are 3 Pulitzer Prizes. The average circulation is between 329,114 and 427,466 and that for 50 pulitzer prizes it is between 430,294 and 655,463. All of the above conclusions are for the last 25 years. Hence we can say that with 90% confidence that, on average , newspaper following each of the three strategy directions differ in their average circulations respectively .

```
# Q.4(d)
change_conf <- predict(pr_change, directions , interval = "prediction", level = 0.9)
change_conf <- tibble(Prizes = directions$prizes_9014,
                      'Lower bound  newspaper' = change_conf[,2], 'Expected change for circulation' = cl
change_conf %>% knitr :: kable(digits = 1, format.args = list(big.mark = ","))
```

Prizes	Lower bound newspaper	Expected change for circulation	Upper bound newspaper
3	-78.0	-34.1	9.8
25	-69.3	-25.5	18.4
50	-60.4	-15.7	29.0

#With the following analysis we can say with a rate of 90% confidence that the circulation of newspaper with 3 pulitzer Prizes in the last 25 years would have increased between -78 and 9.8 , and in the case of circulation of newspaper with 25 Pulitzer Prizes in the last 25 years would have increased between -69.3 and 18.4. And finally, talking about the case where the circulation of a newspaper with 50 Pulitzer Prizes in the last 25 years would have increase between from -60.4 and 29.0. If observed closely we can see a substanstial overlap between the prediction intervals , telling us that there is a need for change in circulation.

#Q.5(limitations) # Limitations: (1) We are basically assuming the newspaper are going to hit the exact targets for the Pulitzer Prizes it will win. Which is very bold of us to make such derivations since it under the control of the Pulitzer committee. # (2) The data that is being used for the analysis is very limited , in the other words they are specifically being from a particular time period, there are high chances this might not be useful for new time period. Also assuming that it to have same results in future just because of its past is also incorrect. This model is overall built for predicting the performance of the Boston Sun-Times in the past , than predicting the future performance #(3) A limitation that i could figure out was that the sample size of the dataset is very small. The problem with this is the outliers / bias would have bigger effects on the prediction results. This statement can be supported by observing the range of confidence interval , and prediction interval, since they are very large as there are more chances of for errors.

#Q.5(conclusions) # The purpose of this project is to provide data analysis support to Masthead Media company. To be more specific , Masthead Media were interested to know that, will winning more Pulitzer prizes would lead to a change in the circulation numbers in order to determine the Boston Sun Times strategic direction. # In order to achieve this I developed two statistical models in order to understand if the numbers of prizes won are affecting the average circulation of 2004 and 2013 as well as the percentage change in the newspaper circulations between these years. # After implementation of both models we concluded that the most appropriate number of expected wins is 50 which makes Boston Sun Times to have an overall average circulation between 430,294 and 655,463 . But while doing this process we realized that there was an overlap between the differing number of prizes won within the range for 50 prizes being between a 60 % decrease and

a 29 % increase in circulation. # Another reason for decline can be , because people prefer reading online more than offline. # last but not the least , for future work , it is recommended that larger datasets are retrieved , with additional information such as the exact number Pulitzer won each other. And the main goal behind this is to minimize the limitation faced during this project for the future work.