DTP assignment 3

Ma Lok Sum, Zoe (a1819866)

2022-09-01

Executive Summary

Recently, the Boston Sun-Times has seen a recent decrease in readership. The report is written for Masthead Media to determine the problem of whether to continue to invest in the Sun-Times' investigative journalism or to encourage the newspaper to take a more populist, tabloid slant.

In the technical part, I prepared the dataset by changing the specific variable (change_0413) into the appropriate type of variable. I plot the histogram for average circulation and percentage change of circulation for visual and analytics purposes. In addition, I adjust the log skewness to make the data frame more effectively. Then, I build 2 models and check with the four assumptions that the assumption investigates whether we should use the model or not. After that, I predict the expected circulation for the three strategic directions using the built model. Last part, I list a few limitations for those two modellings and leave the recommendation in conclusion.

During that, I found that the Pulitzer Prizes significantly contribute to resolving the existing problem. Masthead Media may use the number of Pulitzer Prizes and the average circulation to estimate predicted circulation. As a result, we may use the forecasted number to compare it with the present circulation figure and choose the appropriate course of action.

Publications with a higher average circulation have won more Pulitzer Prizes due to a rise in the ratio of Pulitzer Prize winners to general circulation. Expect an increase in the average circulation as the number of Pulitzer Prizes rises.

A percentage rise in circulation observes for publications with more Pulitzer Awards during the award-winning era. The upward growing slop between the quantity of Pulitzer Prizes and the proportion in circulation is the cause. Expect the shift in circulation to increase as the number of Pulitzer Prizes rises.

The final recommendation to Masthead Media is If these relationships are authentic, the circulation trajectory of the Boston Sun-Times might alter depending on the editorial strategy of the third daily. Based on the third strategy, the newspaper's predicted circulation has increased from the existing circulation.

Question One: Reading and Cleaning

Get the data

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.8     v stringr 1.4.0
## v tidyr   1.2.0     v forcats 0.5.1
## v readr   2.1.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```r
library(readr)
pulitzer <- read_csv("pulitzer.csv")
```

```
## Rows: 45 Columns: 5
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (2): newspaper, change_0413
## dbl (3): circ_2004, circ_2013, prizes_9014
##
## 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.
```

```r
pulitzer
```

```
## # A tibble: 45 x 5
##    newspaper             circ_2004 circ_2013 change_0413 prizes_9014
##    <chr>                     <dbl>     <dbl> <chr>             <dbl>
##  1 USA Today               2192098   1674306 -24%                  3
##  2 Wall Street Journal     2101017   2378827 13%                  51
##  3 New York Times          1119027   1865318 67%                 118
##  4 Los Angeles Times        983727    653868 -34%                 86
##  5 Washington Post          760034    474767 -38%                101
##  6 New York Daily News      712671    516165 -28%                  7
##  7 New York Post            642844    500521 -22%                  1
##  8 Chicago Tribune          603315    414930 -31%                 39
##  9 San Jose Mercury News    558874    583998 4%                    7
## 10 Newsday                  553117    377744 -32%                 19
## # ... with 35 more rows
## # i Use `print(n = ...)` to see more rows
```

1(a): convert the percentage to a numerical number

```r
pulitzer$change_0413 <- as.numeric(substr(pulitzer$change_0413,0,nchar(pulitzer$change_0413)-1))
```

change the varaible change_0413 from character to integers

```r
pulitzer$change_0413 <- as.integer(pulitzer$change_0413)
pulitzer$change_0413
```

```
##  [1]  -24   13   67  -34  -38  -28  -22  -31    4  -32  -34  -23  -56  -37    4
## [16]  -45  -44  -14  -45  -20  -18  -15  -30   -2   22 -100  -33  -55   15  -34
## [31]  -41  -36  -31  -57  -40  -39 -100  -47  -38  -44  -26  -18  -20  -21  -60
```

1(b): append a new variable to the tibble which contains the average of circ_2004 and circ_2013

```
pulitzer$newcirc <- (pulitzer$circ_2004 + pulitzer$circ_2013)/2
pulitzer
```

```
## # A tibble: 45 x 6
##    newspaper            circ_2004 circ_2013 change_0413 prizes_9014  newcirc
##    <chr>                    <dbl>     <dbl>       <int>       <dbl>    <dbl>
##  1 USA Today              2192098   1674306         -24           3 1933202
##  2 Wall Street Journal    2101017   2378827          13          51 2239922
##  3 New York Times         1119027   1865318          67         118 1492172.
##  4 Los Angeles Times       983727    653868         -34          86  818798.
##  5 Washington Post         760034    474767         -38         101  617400.
##  6 New York Daily News     712671    516165         -28           7  614418
##  7 New York Post           642844    500521         -22           1  571682.
##  8 Chicago Tribune         603315    414930         -31          39  509122.
##  9 San Jose Mercury News   558874    583998           4           7  571436
## 10 Newsday                 553117    377744         -32          19  465430.
## # ... with 35 more rows
## # i Use `print(n = ...)` to see more rows
```

Question Two: Univariate Summary and Transformation

2(a) Describe the distribution of the variable representing average circulation

```
library(ggplot2)
mean(pulitzer$newcirc)
```

```
## [1] 437140.7
```

```
sd(pulitzer$newcirc)
```

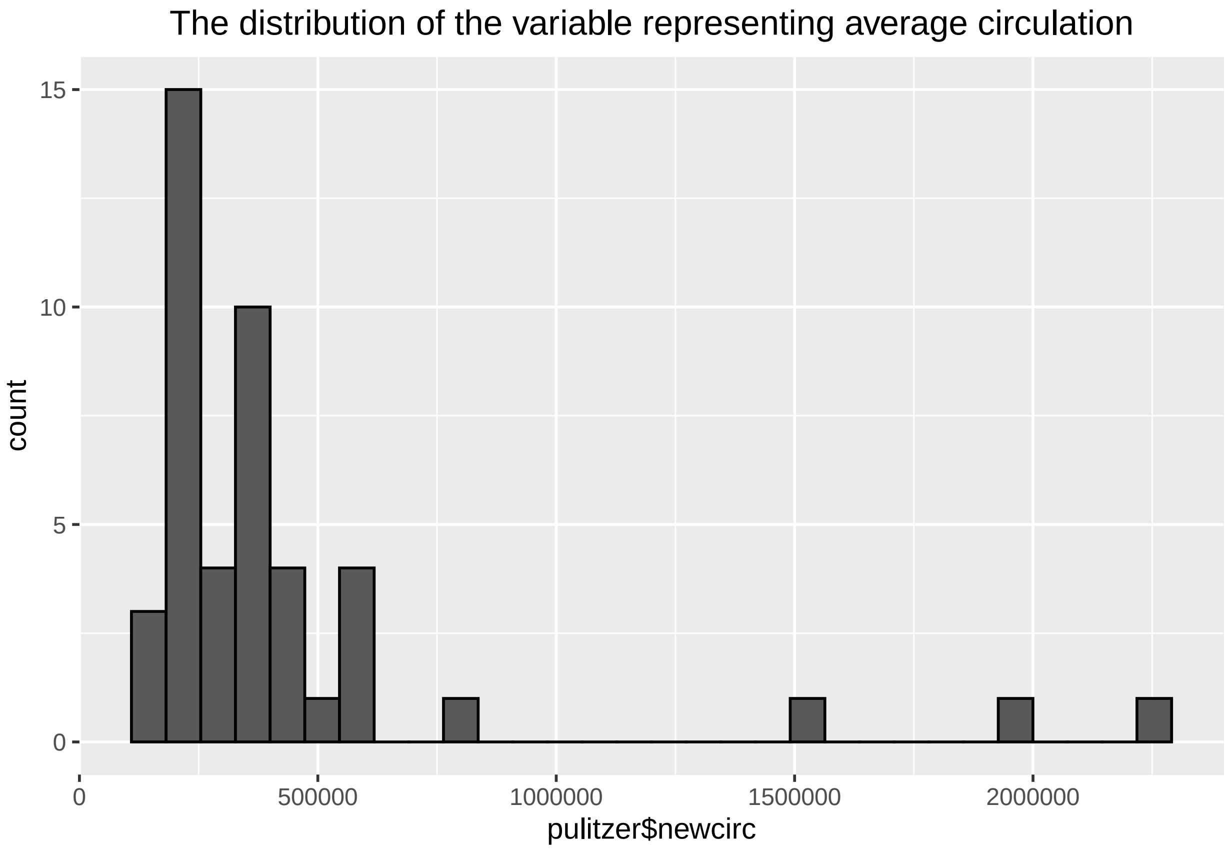
```
## [1] 425701.9
```

```
summary(pulitzer$newcirc)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  131004  216012  333083  437141  462152 2239922
```

```
ggplot(pulitzer,aes(x = pulitzer$newcirc)) + geom_histogram(col = "black")+ ggtitle("The distribution o
```

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



Shape: Unimodal and positive-skewed.

Location: The median (333083) larger than the mode, which the peak near the median.

Spread: The IQR range of value is 246140.

Outliers: Three potential outliers between 1500000 and 2239922.

2(b) Describe the distribution of change_0413

```
library(ggplot2)
mean(pulitzer$change_0413)
```

```
## [1] -29.04444
```

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

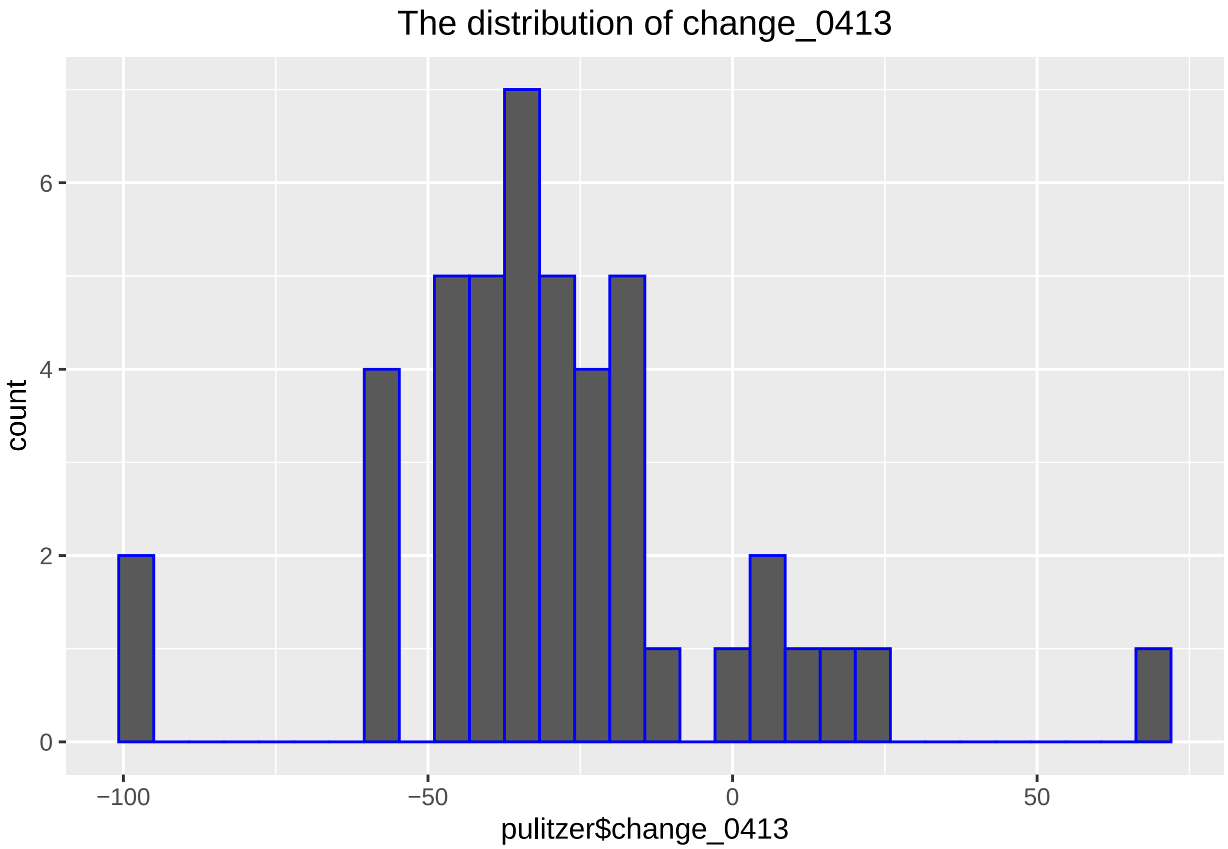
```
## [1] 28.08263
```

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

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -100.00  -41.00  -32.00  -29.04  -20.00   67.00
```

```
ggplot(pulitzer,aes(x = pulitzer$change_0413)) + geom_histogram(col = "blue") + ggtitle("The distributi
  theme(plot.title = element_text(hjust = 0.5))
```

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

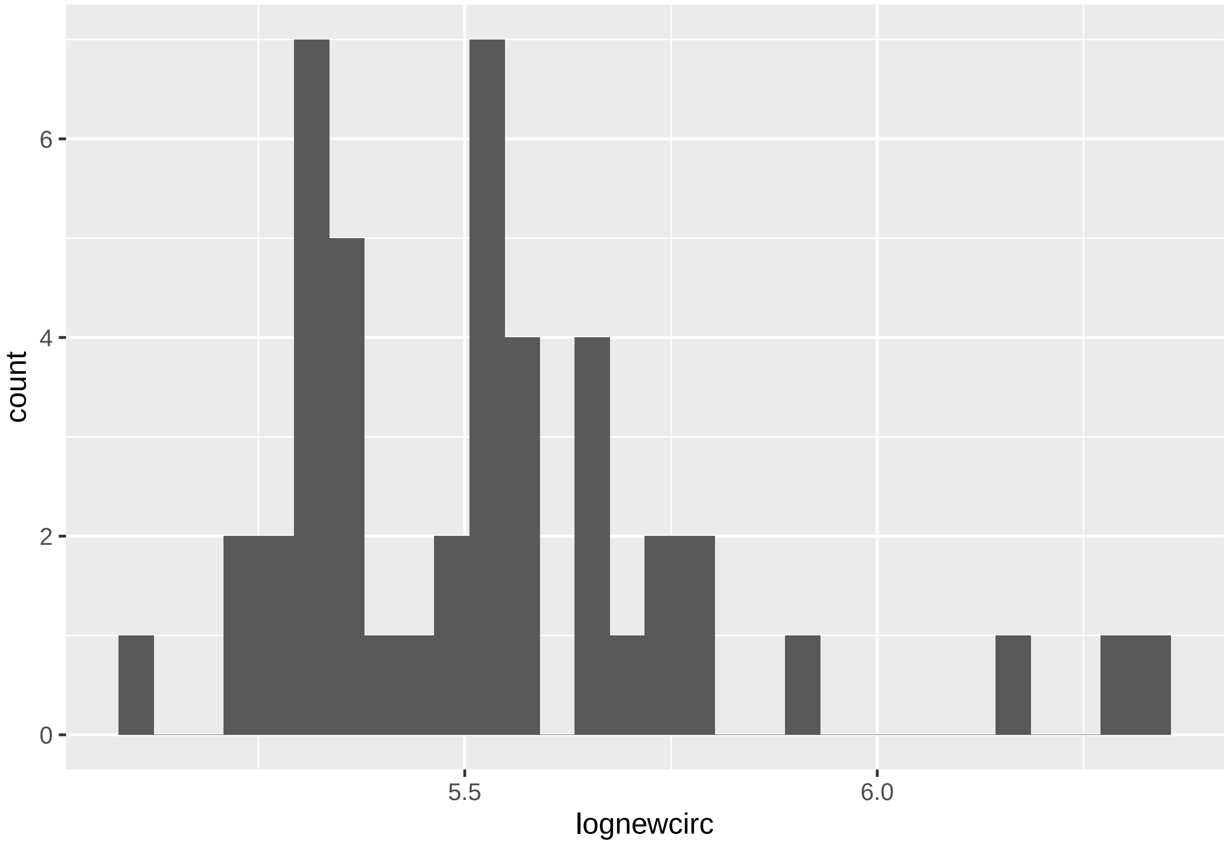


## Shape: Unimodal and positive-skewed. ## Location: The median is -32, which the peak near the median. ## Spread: The IQR range of value is 21. ## Outliers: Two potential outliers are founded in -100 and 60.

2(c) Log Transformations to Adjust for Skewness

```
pulitzer %>%
  mutate(lognewcirc = log10(newcirc + 1)) %>%
  ggplot(aes(lognewcirc)) +
  geom_histogram()
```

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



Since we observe the distribution between the average circulation and the change_0413, the variable of average circulation is skewed than the variable of change_0413.

The variable representing average circulation should be transformed.

Question Three: Model building and interpretation

3(a)

```
pulitzer_lm  <- lm(log(newcirc) ~ prizes_9014, data = pulitzer)
summary(pulitzer_lm)
```

```
##
## Call:
## lm(formula = log(newcirc) ~ prizes_9014, data = pulitzer)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.8573 -0.3249 -0.1005  0.1752  1.9141
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.520712   0.092499 135.361  < 2e-16 ***
## prizes_9014  0.013288   0.003017   4.405 6.91e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5137 on 43 degrees of freedom
## Multiple R-squared:  0.3109, Adjusted R-squared:  0.2949
## F-statistic:  19.4 on 1 and 43 DF,  p-value: 6.91e-05
```

The model model predicting the variable representing a newspaper's circulation using prizes_9014:

log(newcirc) = 12.520 + 0.0133 * prizes_9014

The intercept is 12.520. It implies that If the prizes is zero, the log of the news circulation is 12.520 in the log scale.

The slope is 0.0133. It implies that If the prizes increase by one, the circulation in the log news circulation will increase by 0.0133 in the log scale.

Talking to the statistically significant relationship between the number of Pulitzer Prizes, and average circulation, In this case, if p-value $<$ 0.05 for these two variables then it is significant and has some relationship with the predictor. Then, the prizes with p-value 6.91e-05 is below 0.05. Therefore, it is a statistically significant relationship between the number of Pulitzer Prizes, and average circulation.

3(b)

```
pulitzer_lm_2  <- lm(change_0413 ~ prizes_9014, data = pulitzer)
summary(pulitzer_lm_2)
```

```
##
## Call:
## lm(formula = change_0413 ~ prizes_9014, data = pulitzer)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -67.834 -11.073  -1.834  13.404  57.675
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -35.5915     4.7955  -7.422 3.17e-09 ***
## prizes_9014   0.3806     0.1564   2.434   0.0192 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 26.63 on 43 degrees of freedom
## Multiple R-squared:  0.1211, Adjusted R-squared:  0.1006
## F-statistic: 5.924 on 1 and 43 DF,  p-value: 0.01916
```

The model model predicting change__0413 using prizes__9014:

change__0413 = -35.591 + 0.381 * prizes__9014

The intercept is -35.591. It implies that If the prizes is zero, the percentage change in the newspaper's circulation is -35.591.

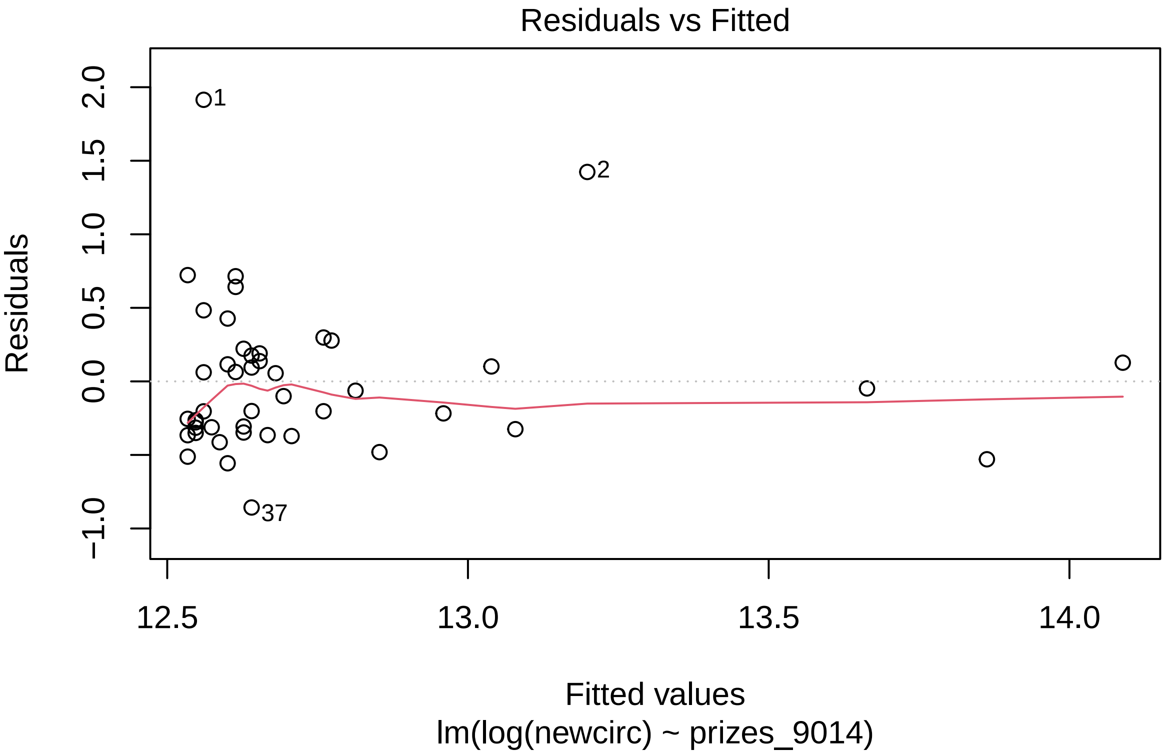
The slope is0.381. It implies that If the prizes increase by one, the percentage change in the newspaper's circulation will increase by 0.381.

Talking to the statistically significant relationship between the number of Pulitzer Prizes, and change in circulation, In this case, if p-value < 0.05 for these two variables then it is significant and has some relationship with the predictor. Then, the prizes with p-value 0.0192 is below 0.05. Therefore, it is a statistically significant relationship between the number of Pulitzer Prizes, and change in circulation.

3(c)

for 1st linear model:

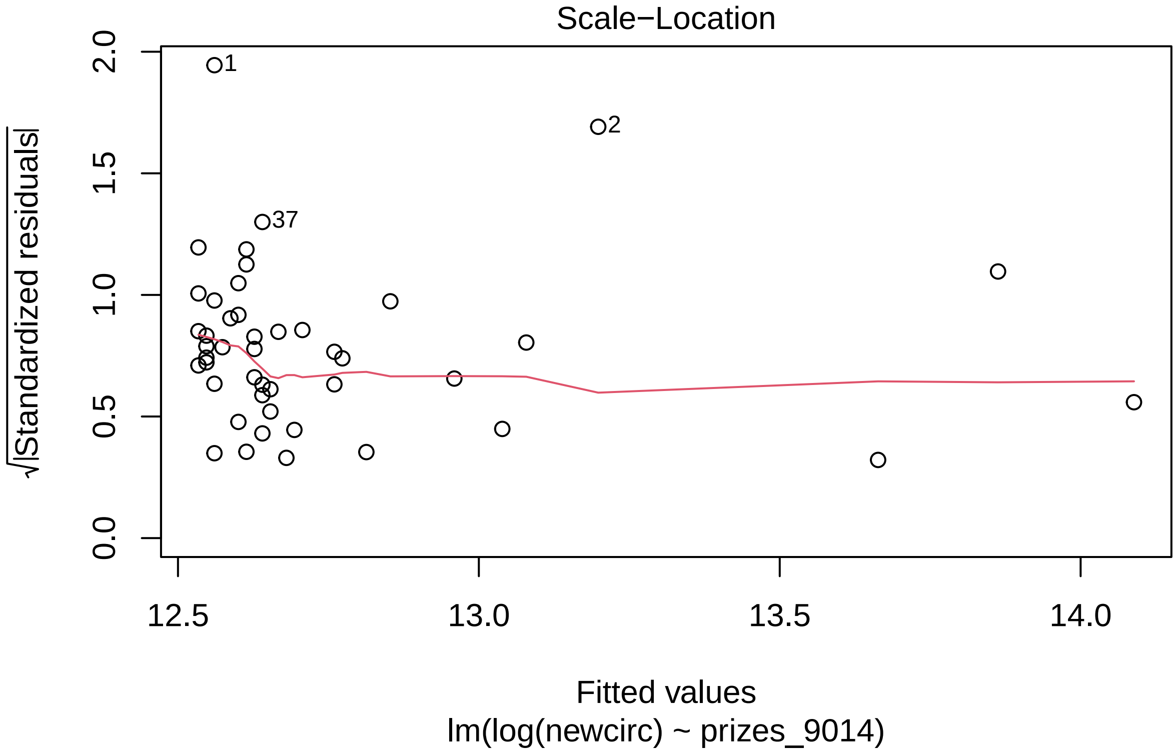
```
plot(pulitzer_lm, which = 1)
```



##

1st: Linearity: The linearity assumptions are satisfied since the residuals plot appears flat, indicating that there is no residual trend.

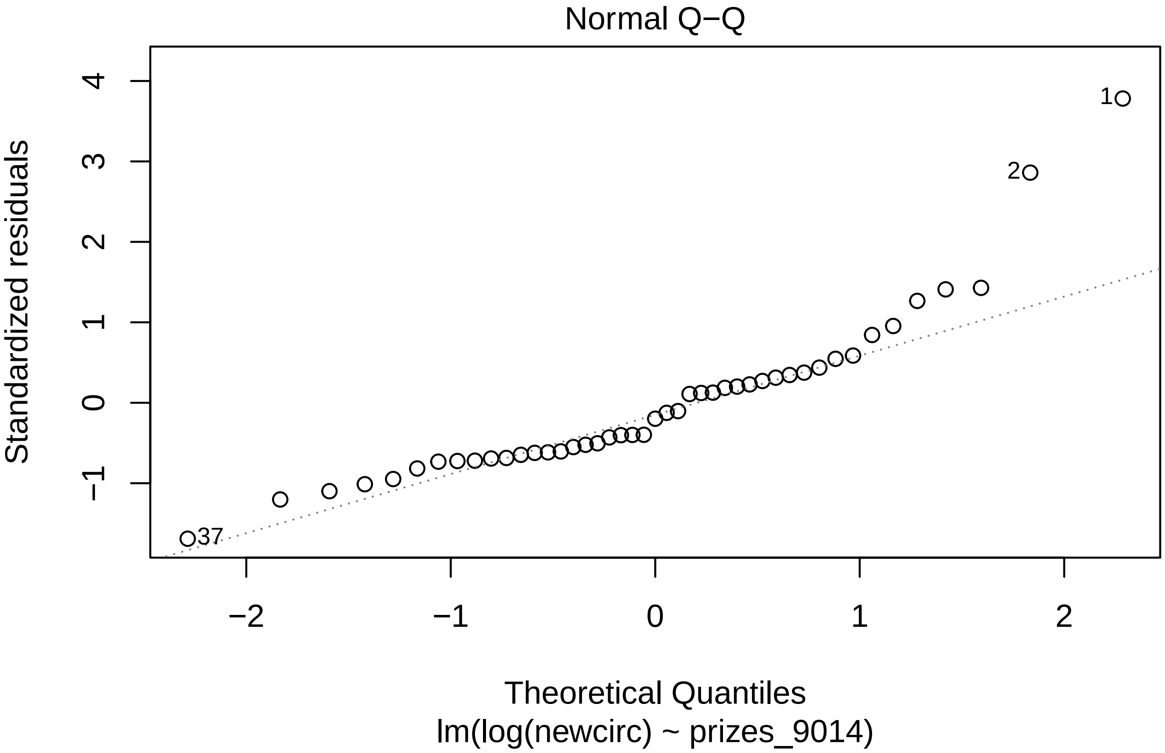
```
plot(pulitzer_lm, which = 3)
```



##

2nd: Constant variance: The plot of the standardized residuals is flat. It indicates that the homoscedasticityal assumption is valid.

```
plot(pulitzer_lm, which = 2)
```



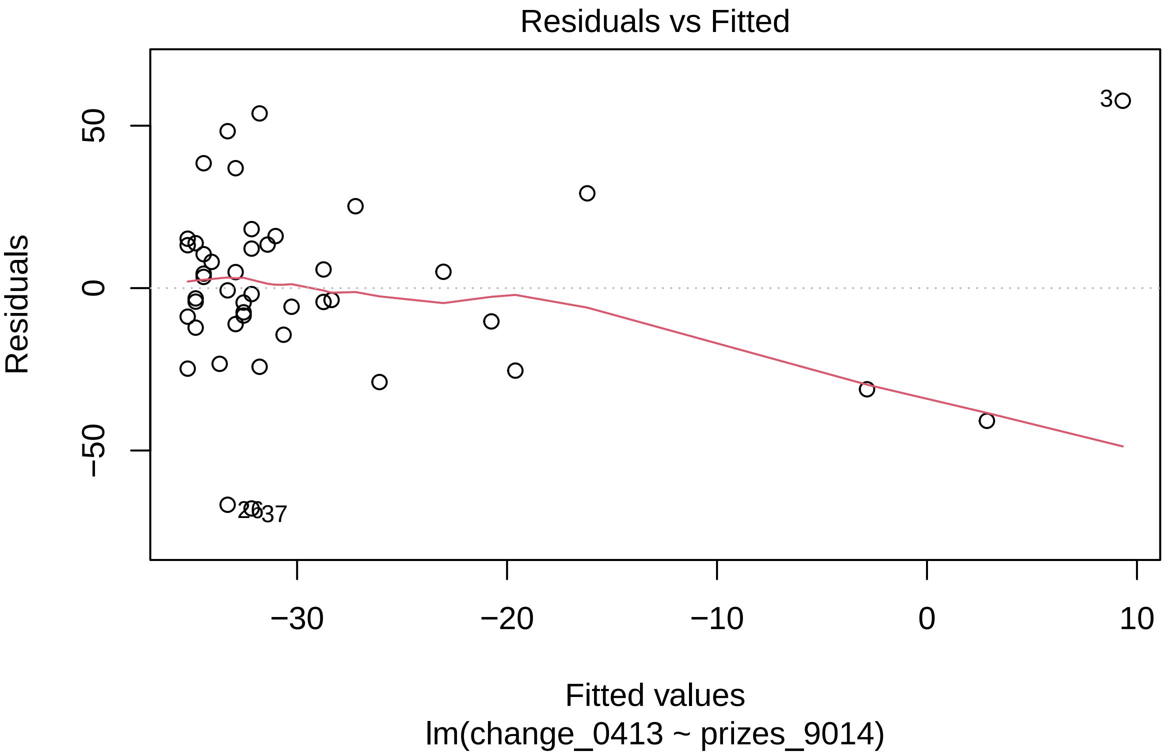
##

3rd: Normality: Since the residuals have a normal distribution, the points should fall along the dotted line. The normality assumption is met as a result.

4th Independence: The overall number of newspaper subscribers has remained constant notwithstanding the population. The population won't change if individuals go from one newspaper to another. As a result, it is dependent, which contradicts the independent assumption.

2nd linear model

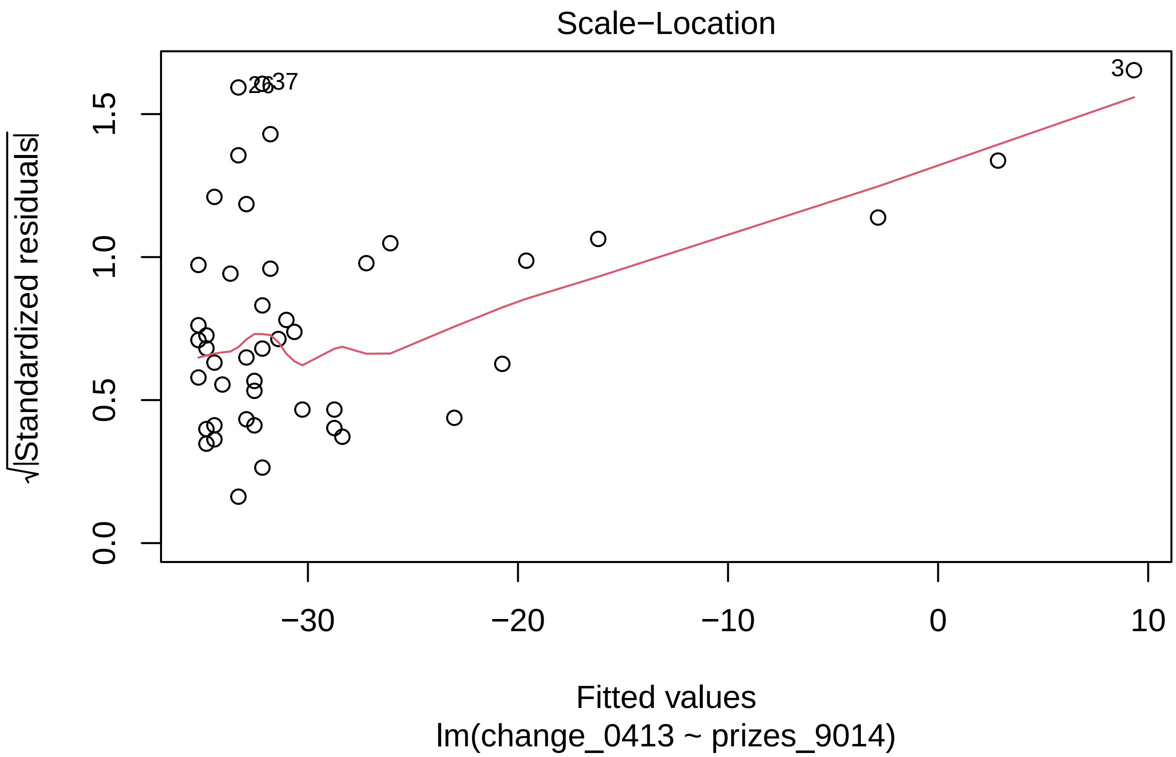
```
plot(pulitzer_lm_2, which = 1)
```



##

1st: Linearity: The linearity assumption are not satisfied since the residual plot are decreasing starting when the fitted values are -20 and appear not flat.

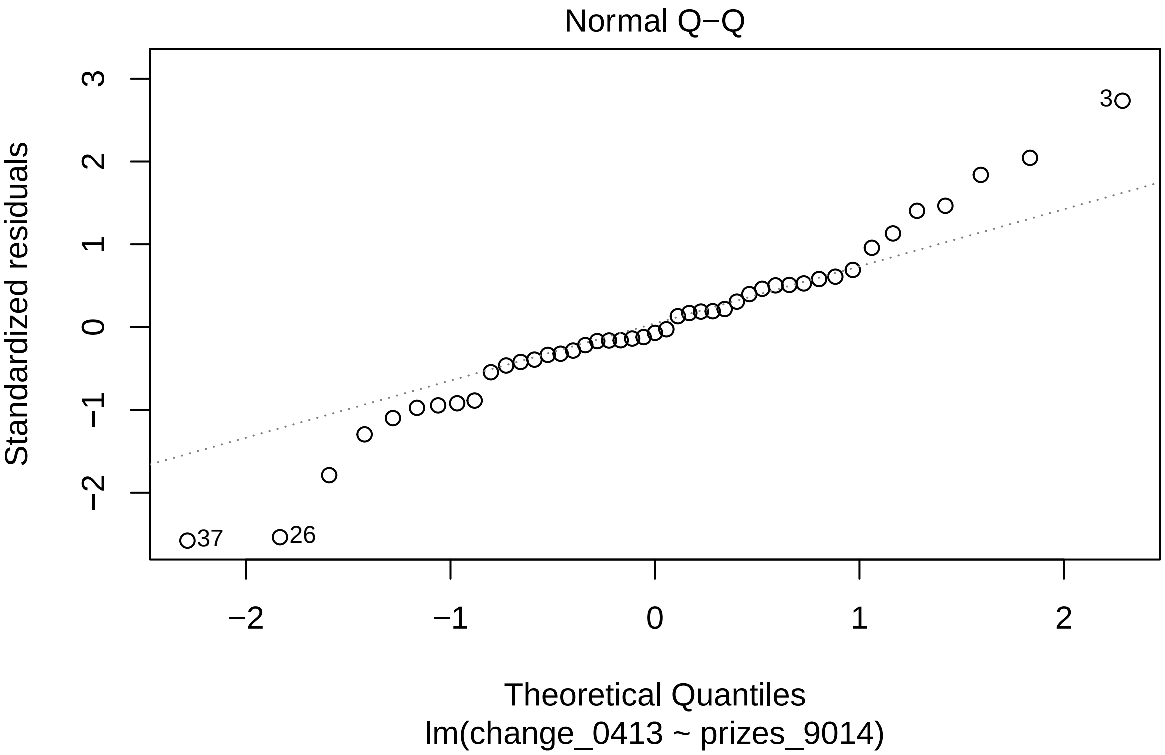
```
plot(pulitzer_lm_2, which = 3)
```



#

2nd: Constant variance: The plot of the standardized residuals is not flat. It indicates that the homoscedasticityal assumption is invalid.

```
plot(pulitzer_lm_2, which = 2)
```



# 3rd:

Normality: Since the residuals have a normal distribution, the points should fall along the dotted line. The normality assumption is met as a result.

4th Independence: The percentage change in the newspaper's circulation has remained constant notwithstanding the population. The population won't change if individuals go from one newspaper to another. As a result, it is dependent, which contradicts the independent assumption.

Question Four: Prediction

4(a)

the model from q3(a): log(newcirc) = 12.520 + 0.0133 * prizes_9014

The Boston Sun-Times currently has a circulation of 482,622.

```r
exp(12.520 + 0.0133 * 5)
```

```
## [1] 292581.9
```

Therefore, for the first strategic, the expected subscribe is 292581.9. It decrease nearly a half of the current subscriber.

```r
exp(12.520 + 0.0133 * 30)
```

```
## [1] 407990.8
```

For the second strategic, the expected subscribe is 407990.8. It slightly decreases compared to the current circulation.

```r
exp(12.520 + 0.0133 * 60)
```

```
## [1] 608042.5
```

For the third strategic, the expected subscribe is 608042.5. It simply increases compared to the current subscribers.

4(b)

the model from q3(b): change_0413 = -35.591 + 0.381 * prizes_9014

```r
round(-35.591 + 0.381* 5, 3)
```

```
## [1] -33.686
```

For the first strategic, the percentage change is -33.686.

```r
round(-35.591 + 0.381 * 30, 3)
```

```
## [1] -24.161
```

For the second strategic, the percentage change is -24.161.

```
round(-35.591 + 0.381 * 60, 3)
```

```
## [1] -12.731
```

For the third strategic, the percentage change is -12.731.

Despite the unfavorable outcomes, each of the proposed strategic initiatives saw an increase in the predicted change in circulation from getting 5 prizes (-33.686 percentage change) to 60 prizes (-12.731 percentage change) because the number of changes in newspaper circulation is declining.

It is inconsistent when compared to the first model, which shows an upward trend in the predicted subscribers. The second model, on the other hand, only gets the negative subscription numbers; it only changes to the positive when there are larger prizes, like 100.

4(c) calculate 90% confidence intervals

Create a new data set for calculating three strategics respectively

```
new_data  <- tibble(
  prizes_9014 = c(5,30, 60))
```

calculate the 90% confidence intervals

```
exp(predict(pulitzer_lm, newdata = new_data, interval = "confidence", level = 0.90))
```

```
##        fit      lwr      upr
## 1 292772.8 253806.3 337721.8
## 2 408135.0 353339.2 471428.6
## 3 608039.1 472432.1 782570.8
```

The 90% confidence interval for the first strategic is (253806.3, 337721.8) and the predicted value is 292772.8 expected subscribers.

The 90% confidence interval for the second strategic is (353339.2, 471428.6) and the predicted value is 408135.0 expected subscribers.

The 90% confidence interval for the tgird strategic is (472432.1, 782570.8) and the predicted value is 608039.1 expected subscribers.

4(d) calculate 90% prediction intervals

```
predict(pulitzer_lm_2, newdata = new_data, interval = "prediction", level=0.90)
```

```
##         fit       lwr      upr
## 1 -33.68831 -79.06670 11.69008
## 2 -24.17219 -69.56200 21.21762
## 3 -12.75284 -59.39535 33.88967
```

The 90% prediction interval for the first strategic is (-79.068, 11.690) and the fitted value is -33.688 expected subscribers.

The 90% prediction interval for the second strategic is (-69.562, 21.218) and the fitted value is -24.172 expected subscribers.

The 90% prediction interval for the third strategic is (- 59.395, 33.890) and the fitted value is -12.753 expected subscribers.

Question Five: Limitations

5(a)

Model 1: It left one of the assumption which is not satisfied–Independence.

Model 2: It left three of the assumption which is not satisfied–Linearity, Constant variance and Independence. Besides, it is not common to see those results are mostly tend to be a negative number, for example, there is no sense that we can predict a negative number of expected subscriber.

For both of the models: It has a few possible outliers that might skew the results. This is because outliers, or extreme numbers, have a significant negative impact on the mean. At the same time, it could be biassed.

Conclusion

Finally, Masthead Media is taking the Boston Sun-Times in one of three strategic ways. In an effort to reverse the recent dip, Masthead Media is deciding whether to continue supporting the Sun-Times' investigative reporting or to push it in the direction of a more populist, tabloid bent.

The conclusion is that publications with a higher average circulation have won more Pulitzer Prizes. Additionally, publications with a higher number of Pulitzer Prize wins saw a rise in circulation during the award-winning season. The third strategic strategy, which has led to the awarding of 60 Pulitzer Prizes and a major rise in investment in investigative journalism, may be employed to address the issue.