MATHS 7107 Data Taming Assignment 3

Ky Phong Mai

2023-03-14

Executive summary

The aim of the project is to determine on the direction of Boston Sun-Times, a flagship newspaper of Masthead Media. Having won on average one Pulitzer Prize every year for the past 25 years, Boston Sun-Times is known for their investigative journalism, which is believed to be the reason for their success. Due to a recent drop in readership (current circulation of 453869), Masthead Media wants to find out if continuing to focus on investigative articles or shifting the direction to a more populist, tabloid slant is better for their flagship newspaper.

In particular, Masthead wants know if winning more Pulitzer Prizes will lead to an increase in circulation. The project establishes two statistical models for evaluation. Following the prediction process using the two models, the results are subsequently analyzed and compared for interpretation.

While one model predicts a reduction in circulation regardless of the strategy, the other model projects an increase in circulation to round 532381 with the estimation range from 431398 to 657001 if the publication invests more in investigate journalism to achieve 50 Pulitzer Prizes. However, there are some drawbacks of the both models that make it unreliable to provide a good prediction of circulation. Hence, it is recommended to reassess and analyse the data to find the better ways to determine the newspaper's directions.

Question One: Reading and Cleaning

Load the data contained in pulizer.csv

```
pacman::p_load(tidyverse,readr,knitr)
```

```
pulitzer <- read_csv("pulitzer.csv")
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
```

(a) Recode the change_0413 variable as integer

```
pulitzer$change_0413 <- str_sub(pulitzer$change_0413, end =-2)
pulitzer$change_0413 <- as.integer(pulitzer$change_0413)
pulitzer
```

```
## # A tibble: 45 x 5
##    newspaper              circ_2004 circ_2013 change_0413 prizes_9014
##    <chr>                      <dbl>     <dbl>       <int>       <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
```

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

```
pulitzer <-pulitzer %>%
  mutate (avg_cir_0413 = (circ_2004 + circ_2013)/2)
pulitzer
```

```
## # A tibble: 45 x 6
##    newspaper              circ_2004 circ_2013 change_0413 prizes_9014 avg_cir_0~1
##    <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, and abbreviated variable name 1: avg_cir_0413
```

Question Two: Univariate Summary and Transformation

(a) Describe the distribution of the variable representing average circulation, including shape, location, spread, and outliers

```
pulitzer %>%
  ggplot(aes(x = avg_cir_0413)) +geom_histogram(col = "black", fill = "orange") +
  labs(y = "Frequency")
```

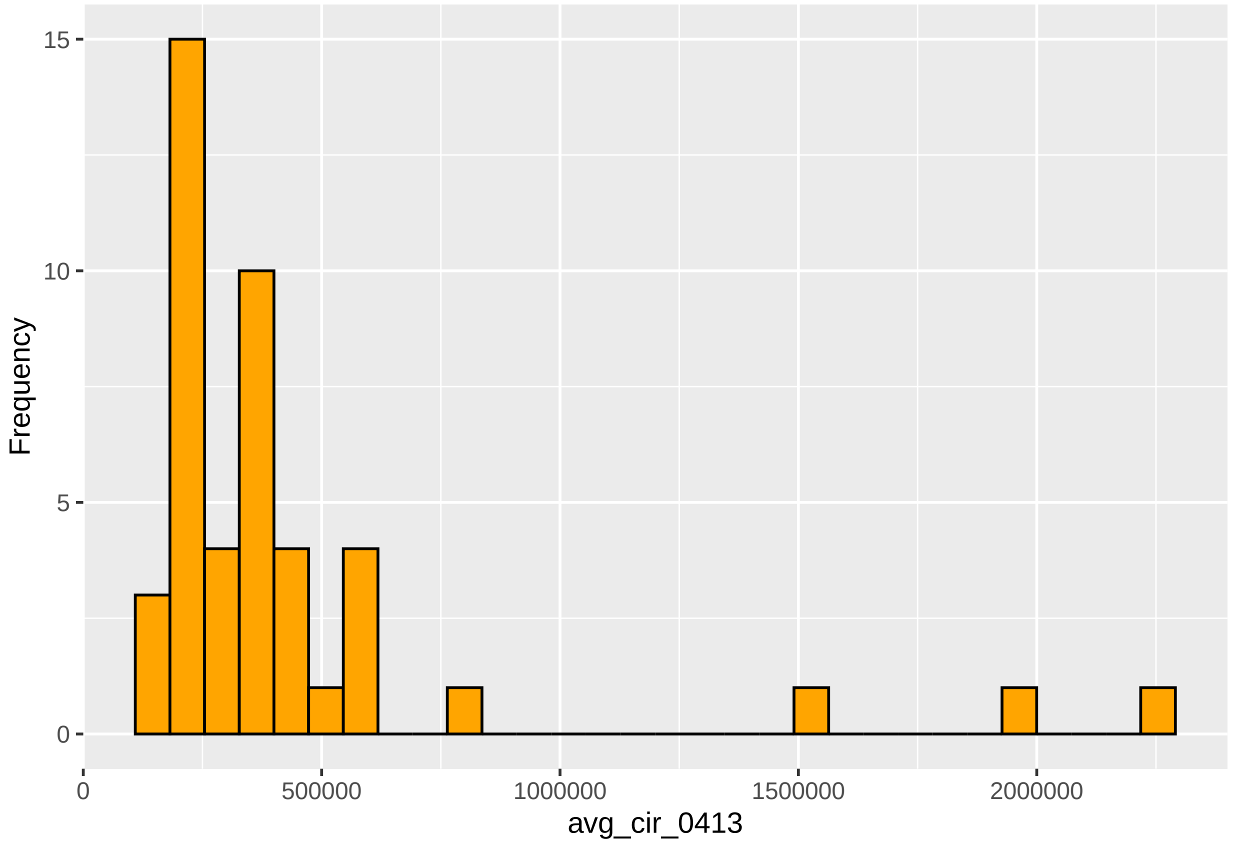

Figure 1: Distribution of average circulation

Shape:

- The distribution of average circulation is asymmetrical, right skewed, unimodal.

Location:

```
summary(pulitzer$avg_cir_0413)
```

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

- Median circulation is at 333083, which is smaller than the mean circulation of 437141, which is expected for a right skewed distribution

Spread:

```
sd(pulitzer$avg_cir_0413)
```

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

```
IQR(pulitzer$avg_cir_0413)
```

```
## [1] 246140
```

- Standard deviation is 425701.9
- IQR is 246140

Outliers:

```
pulitzer %>%
  ggplot(aes(y= avg_cir_0413)) + geom_boxplot()
```

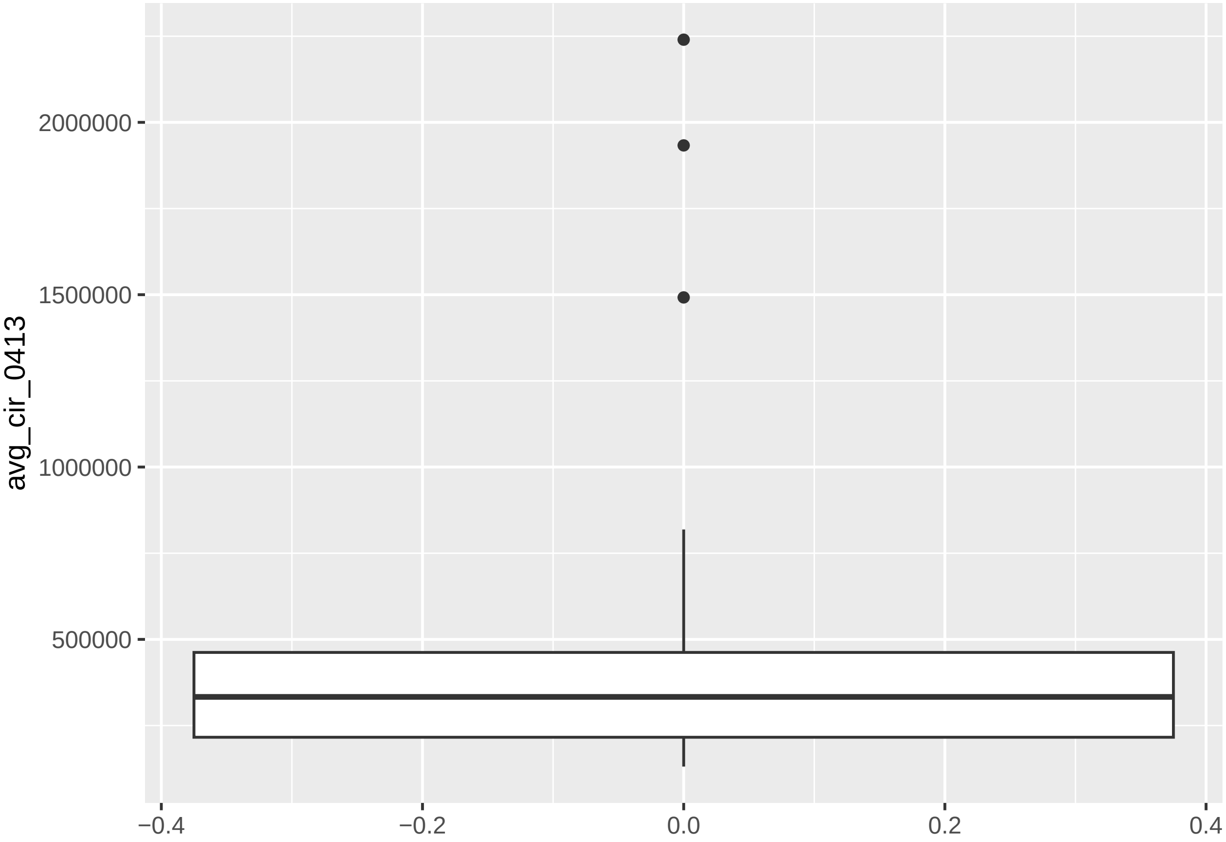

Figure 2: Box plot of average circulation

- There are 3 potentials outliers in the distribution of average circulation based on the boxplot
- Q3 + 1.5IQR = 462152 + 1.5*246140 = 831362
- The 3 outliers are as follows:

```
pulitzer %>%
  filter (avg_cir_0413 > 831362)
```

```
## # A tibble: 3 x 6
##   newspaper           circ_2004 circ_2013 change_0413 prizes_9014 avg_cir_0413
##   <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.
```

(b) Describe the distribution of change_0413, including shape, location, spread and outliers

```
pulitzer %>%
  ggplot(aes(change_0413)) + geom_histogram(col = "black", fill = "orange") +
  labs (y = "Frequency")
```

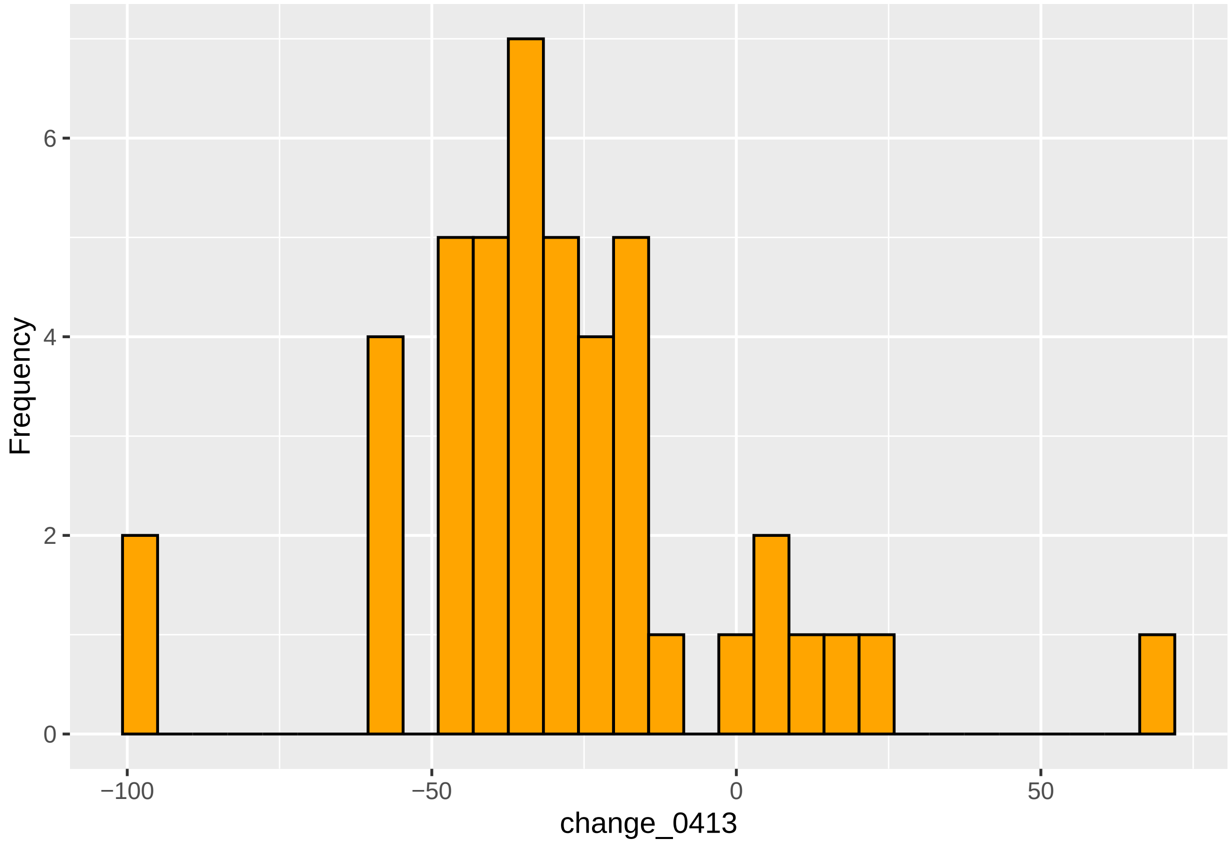

Figure 3: Distribution of circulation change in percentage

Shape:

- The distribution of circulation change is somewhat symmetrical and with one distinct peak (unimodal)

Location:

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

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

- Median change is -32% while mean change is -29.04%

Spread:

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

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

```
IQR(pulitzer$change_0413)
```

```
## [1] 21
```

- Standard deviation is 28.083%
- IQR is 21%

Outliers:

```
pulitzer %>%
  ggplot(aes(y = change_0413)) + geom_boxplot()
```

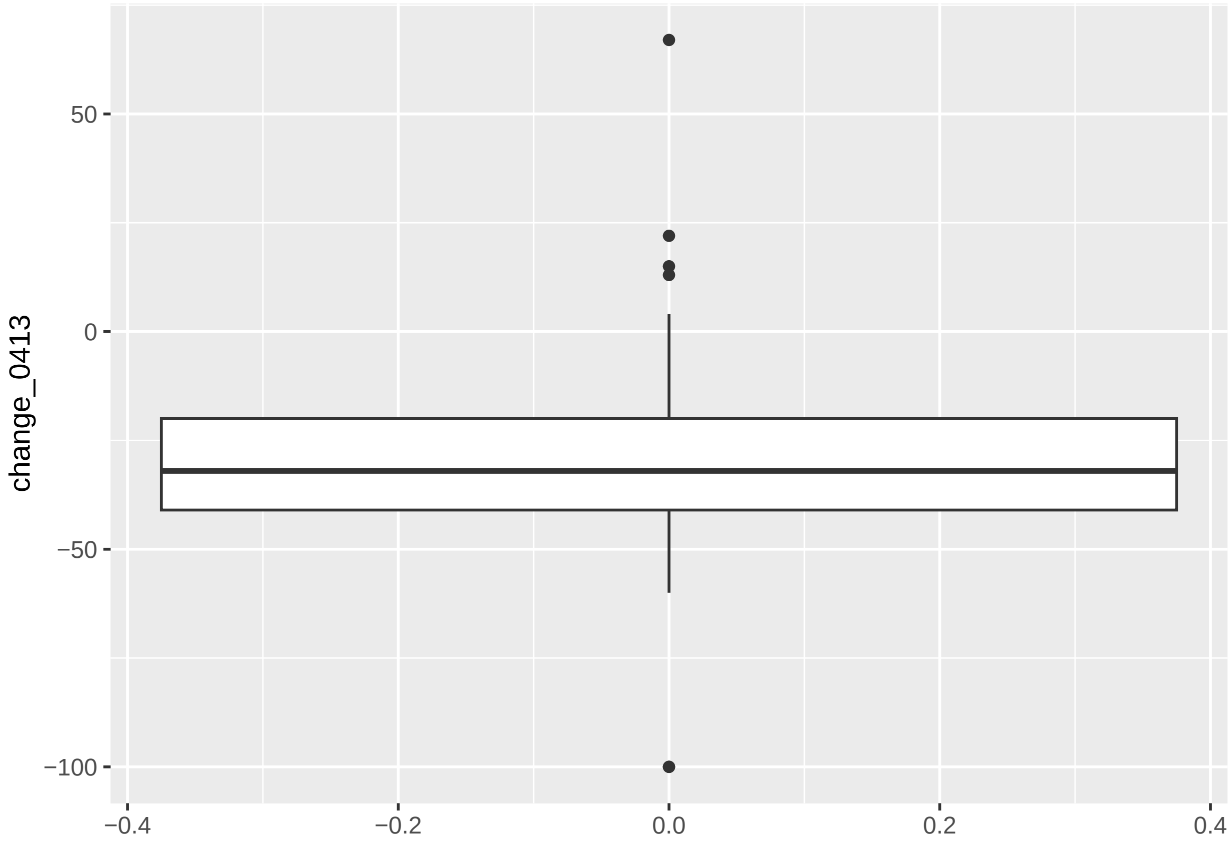

Figure 4: Box plot of circulation change

- There are 4 high outliers and 2 low outliers
- Q1 - 1.5*IQR = -72.5%
- Q3 + 1.5*IQR = 11.5%
- Outliers are as follows:

```
pulitzer %>%
  filter(change_0413 < -72.5)
```

```
## # A tibble: 2 x 6
##   newspaper                    circ_2004 circ_2013 change_0413 prizes_9014 avg_c~1
##   <chr>                            <dbl>     <dbl>       <int>       <dbl>   <dbl>
## 1 Rocky Mountain News             340007         0        -100           6 170004.
## 2 New Orleans Times-Picayune      262008         0        -100           9 131004
## # ... with abbreviated variable name 1: avg_cir_0413
```

```
pulitzer %>%
  filter(change_0413 > 11.5)
```

```
## # A tibble: 4 x 6
##   newspaper                circ_2004 circ_2013 change_0413 prizes_9014 avg_cir_0~1
##   <chr>                        <dbl>     <dbl>       <int>       <dbl>       <dbl>
## 1 Wall Street Journal        2101017   2378827          13          51     2239922
## 2 New York Times           1119027   1865318          67          118    1492172.
## 3 Denver Post               340168    416676          22           10     378422
## 4 Orange County Register    310001    356165          15            6     333083
## # ... with abbreviated variable name 1: avg_cir_0413
```

(c) Do either of change_0413 and the variable representing average circulation have a skew that could be resolved by a log transform? For each variable, select whether it should be transformed.

The variable representing average circulation is right skewed, hence it could be resolved by a log transform. As the variable change_0413 is somewhat symmetrical, log transformation is not necessary

```
pulitzer <- pulitzer %>%
  mutate (log_avg_cir_0413 = log(avg_cir_0413))
pulitzer %>%
  ggplot(aes(log_avg_cir_0413)) + geom_histogram(col = "black", fill = "orange")
```

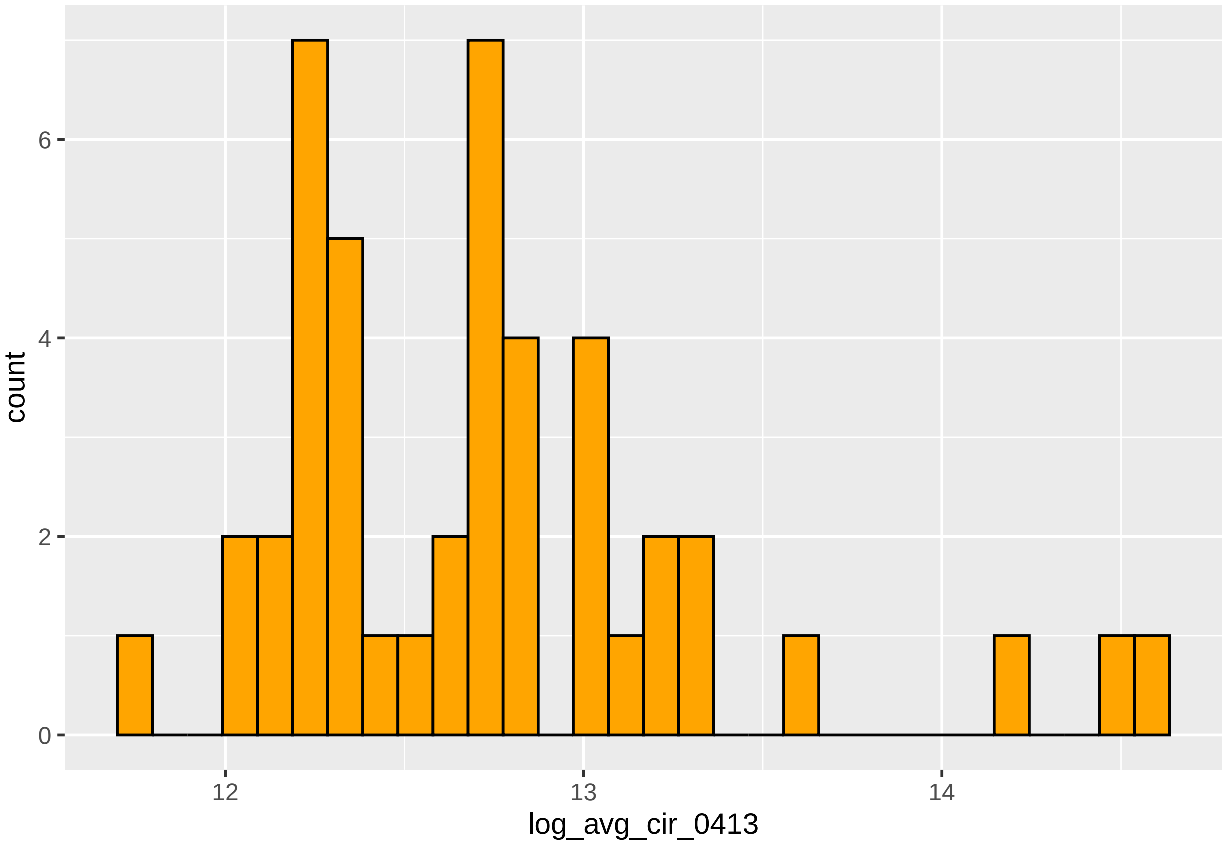

Figure 5: Distribution of log average circulation

Question Three: Model building and Interpretation

(a) Build a model predicting the variable representing a newspaper's circulation using prizes_9014, incorporating a log transform for the average circulation if you decided this was necessary. State and interpret the slope and intercept of this model in context. Is there a statistically significant relationship between the number of Pulitzer Prizes, and average circulation?

```
avg_cir_lm <- lm(log_avg_cir_0413 ~ prizes_9014 ,data = pulitzer)
summary(avg_cir_lm)
```

```
## 
## Call:
## lm(formula = log_avg_cir_0413 ~ 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
```

```
exp(12.520712)
```

```
## [1] 273953
```

- Intercept = 12.521. This means if there are 0 Pulitzer prize, the log of average circulation is expected to be 12.521, which is the same as the circulation of 273953
- Slope = 0.013. This means if the number of Pulitzer prizes increase by 1, the log of average circulation is expected to increase by 0.013
- There is a significant relationship between the number of Pulitzer Prizes as indicated by p value $<$ 0.001

(b) Build a model predicting change_0413 using prizes_9014, incorporating a log transform for change_0413 if you decided this was necessary. Is there a statistically significant relationship between the number of Pulitzer Prizes, and change in circulation?

```
change_0413_lm <- lm(change_0413 ~ prizes_9014 ,data = pulitzer)
summary(change_0413_lm)
```

```
##
## 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
```

- Intercept = -35.592. This means if there is 0 Pulitzer prize, the change in circulation is expect to decrease by 35.592%
- Slope = 0.381. This means if the number of Pulitzer prizes increases by 1, the circulation is expect to increase by 0.381%
- There is still a significant relationship between the number of Pulitzer Prizes and change in circulation as indicated by the p value < 0.05

(c) Check the assumptions of the linear models. Recall that there are four assumptions for each model.

For first model, avg_cir_lm

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

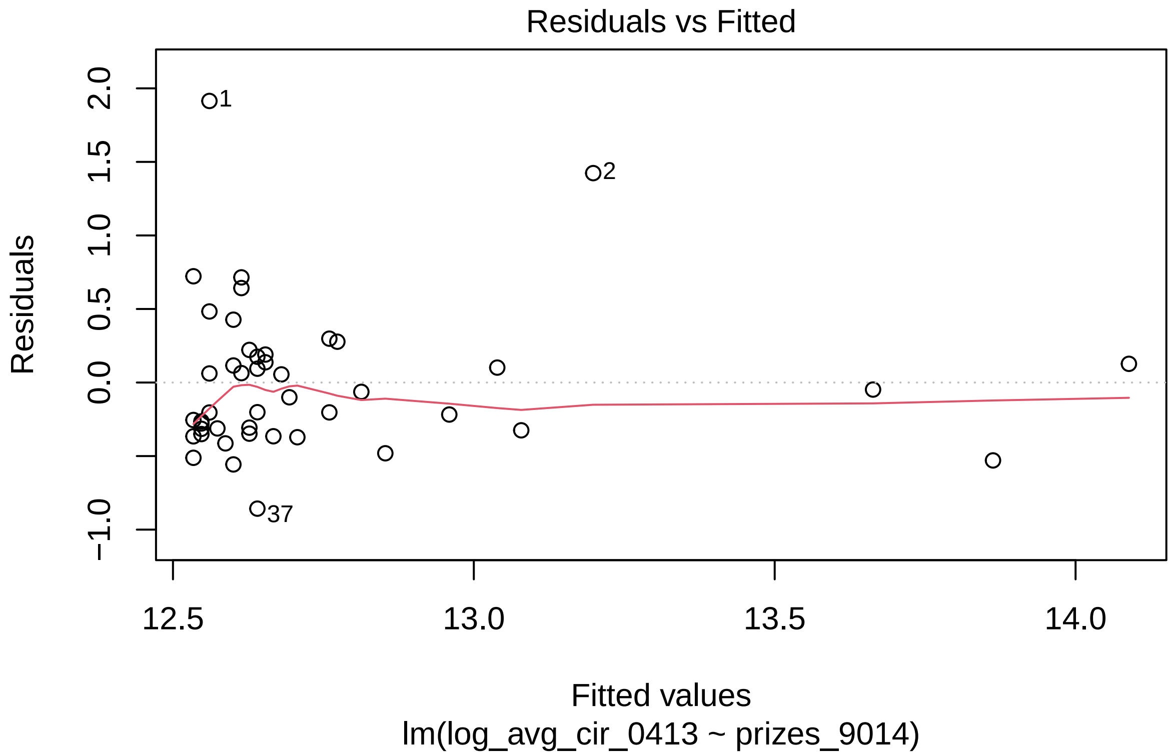

Figure 6: Plot of residuals against fitted value for average circulation lm

- **Linearity**: Based on the above plot, the red line is roughly straight with no trends in the residuals. The linearity assymption is satisfied

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

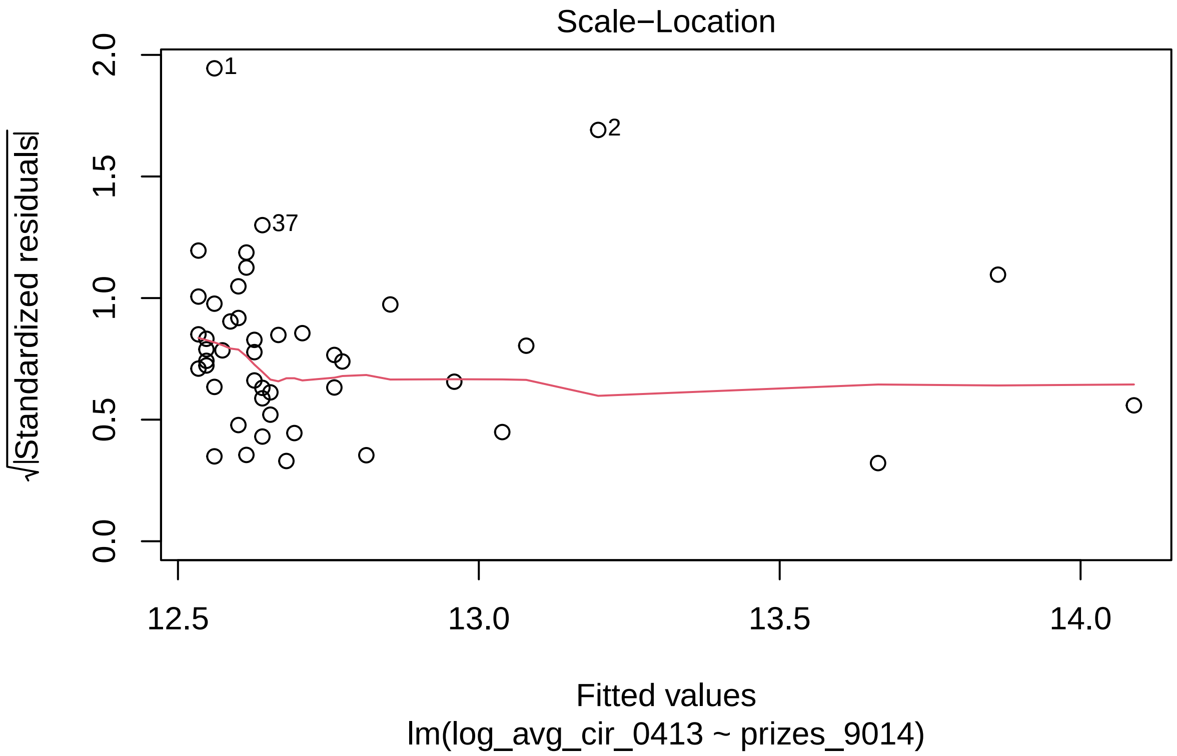

Figure 7: Plot of square root of standardized residuals against fitted values for average circulation lm

- **Homoscedasticity**: Based on the above plot, the points are evenly spread from left to right with no trends from the red line. Hence the assumption is satistifed

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

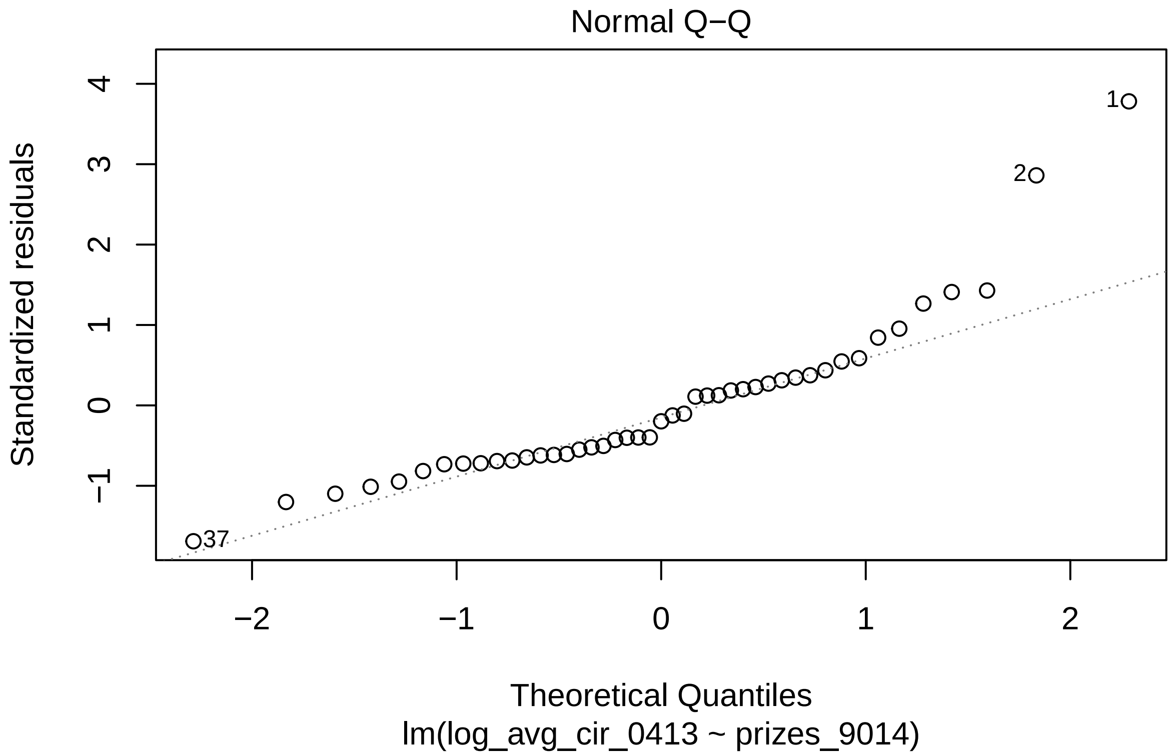

Figure 8: Normal QQ plot of the residuals for average circulation lm

- **Normality**: based on the above plot, most of the points lie along the dotted line. Hence the assumption is satisfied.
- **Independence**: Since there is no plot to assess independence, we need to justify by looking at how the data is obtained. The assumption can only be deemed valid if we must have full control of how data is collected to be certain that one observation/measurement is independent of another. As the circulation by each publication can be affected by the industry in the same way, they are not independent. Hence the assumption cannot be satisfied.

For second model, change_0413_lm

```
plot(change_0413_lm, which = 1)
```

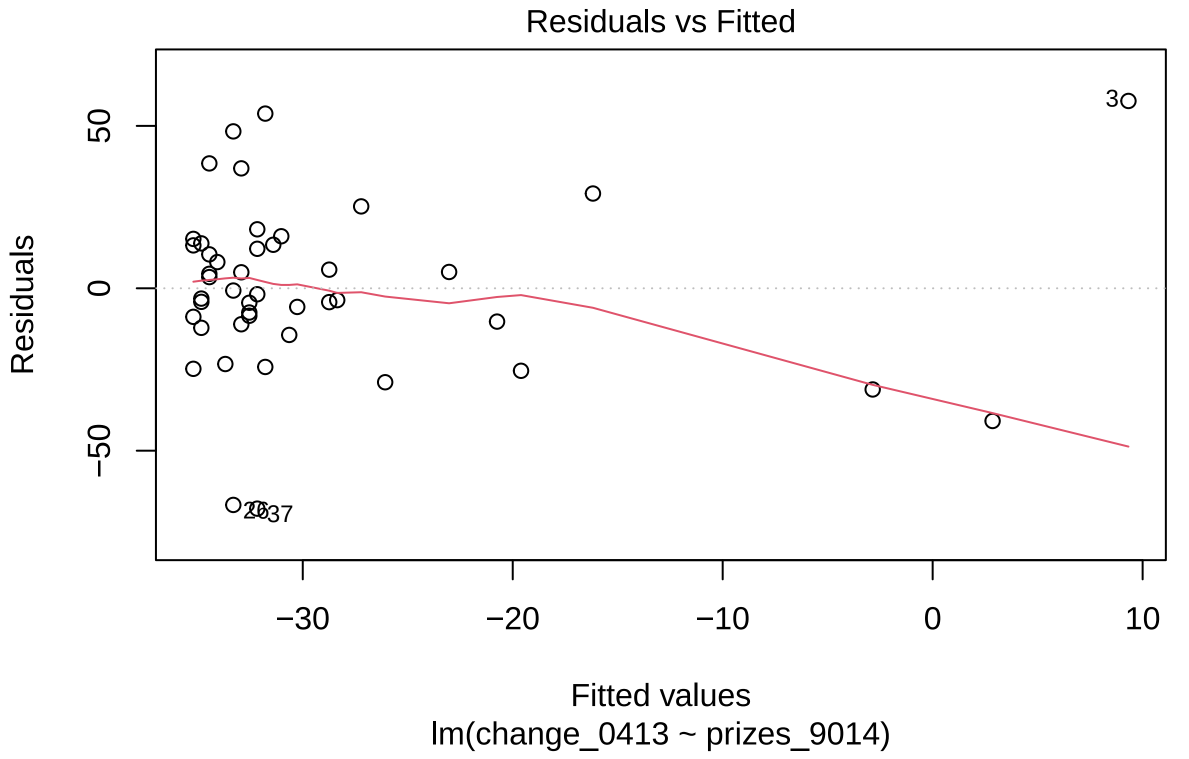

Figure 9: Plot of residuals against fitted value for circulation change lm

- **Linearity**: Based on the above plot, there seems to be no clear trend in the red line. We observe some slight downward movement, however it is due to some outliers towards the right of the plot. Hence, we can still conclude that the assumption is justified

```
plot(change_0413_lm, which = 3)
```

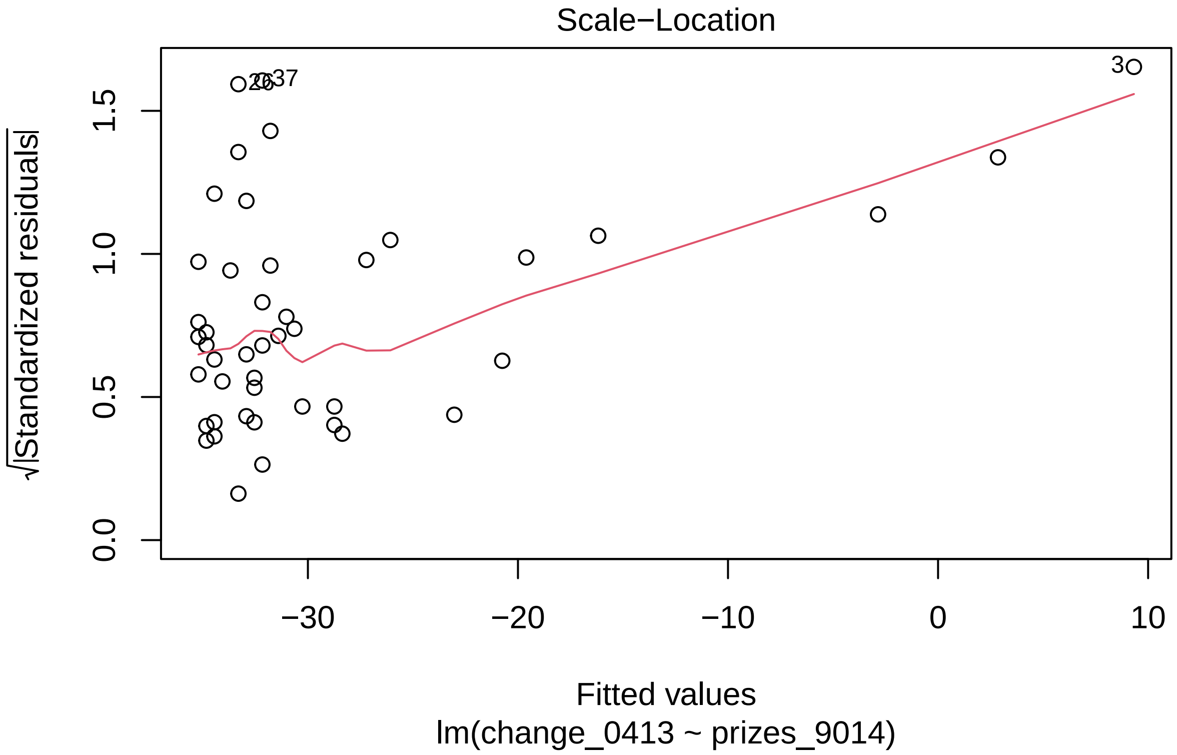

Figure 10: Plot of square root of standardized residuals against fitted values for circulation change lm

- **Homoscadesticity**: Based on the above plot, we can see an upward trend from fitted value of -20 onwards. Hence, for this case, the assumption is not satisfied

```
plot(change_0413_lm, which = 2)
```

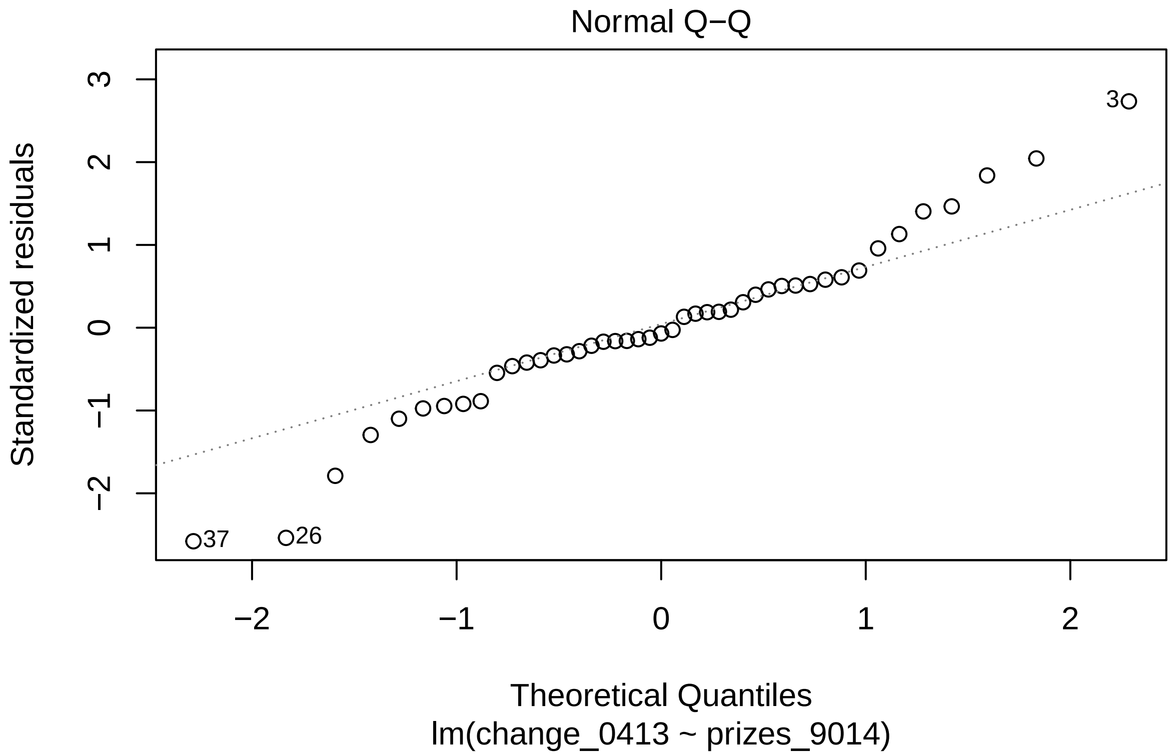

Figure 11: Normal QQ plot of the residuals for circulation change lm

- **Normality**: Based on the above plot, points within -1 to 1 range lie along the dotted line though outside that range, the points started drifting apart slightly. The assumption is still considered justified.
- **Independence**: Similar to average circulation, assessment needs to be done based on how the data is collected. As the circulation by each publication can be affected by the industry in the same way, they are not independent. Hence the assumption cannot be satisfied.

Question 4:

(a) Using the model from Question 3(a), calculate the expected circulation of the newspaper under each of the three proposed strategic directions and represent these in a table. How does this compare with the current circulation?

```r
new_prizes <- tibble (prizes_9014 =c(3,25,50))
new_prizes_avg <- new_prizes %>%
  mutate(predicted_circ = round(exp(predict(avg_cir_lm, new_prizes))))
colnames(new_prizes_avg)[1] <- "prizes"
kable(new_prizes_avg, caption = "Prediction of Circulation under 3 proposed directions")
```

Table 1: Prediction of Circulation under 3 proposed directions

prizes	predicted_circ
3	285095
25	381900
50	532381

The current circulation of Boston Sun-Times is 453869. Only the strategy that invests more in investigative journalism with 50 Pulitzer Prizes will result in an increase in circulation to 532381. Investing substantially less or same amount will both result in an reduction in circulation to 285095 and 381900 respectively.

(b) Using the model from Question 3(b), calculate the change in circulation of the newspaper, across the next decade, under each of the three proposed strategic directions and represent these in a table. Comment on whether the projections of each of the two models are consistent.

```
new_prizes_change <-new_prizes %>%
  mutate("predicted_circ_change(%)" = round(predict(change_0413_lm, new_prizes),3))
colnames(new_prizes_change)[1] <-"prizes"
kable(new_prizes_change, caption = "Prediction of Circulation Change under 3 proposed directions")
```

Table 2: Prediction of Circulation Change under 3 proposed directions

prizes	predicted_circ_change(%)
3	-34.450
25	-26.075
50	-16.559

While based on the first model, we see an increase in circulation when the number of Pulitzer prizes is 50, the second model provides a different prediction. Even with the highest number of Pulitzer prizes, the model predict that there will still be a drop of 16.599% in circulation across the next decade.

(c) Using the model from Question 3(a), calculate 90% confidence intervals for the expected circulation of the newspaper under each of the three proposed strategic directions. Place these confidence intervals in a table, and contrast them in context.

```
predicted_avg_circ <- round(exp(predict(avg_cir_lm, new_prizes,
                                        interval = "confidence", level = 0.9)))
predicted_avg_circ <- tibble(prizes = new_prizes$prizes_9014,
                                "lower limit" = predicted_avg_circ[,2],
                                "predicted circulation" = predicted_avg_circ[,1],
                                "upper limit" = predicted_avg_circ[,3])
kable(predicted_avg_circ,
      caption = "Prediction of circulation with 90% confidence interval under 3
                 proposed directions ")
```

Table 3: Prediction of circulation with 90% confidence interval under 3 proposed directions

prizes	lower limit	predicted circulation	upper limit
3	245997	285095	330406
25	333782	381900	436954
50	431398	532381	657001

- For newspapers with 3 Pulitzer prizes, we are 90% confidence that on average, the circulation is within the range from 245997 and 330406 (investing less in investigative journalism)
- For newspapers with 25 Pulitzer prizes, we are 90% confidence that on average, the circulation is within the range from 333782 and 436954 (investing equally in investigative journalism)
- For newspaper with 50 Pulitzer prizes, we are 90% confidence that on average, the circulation is within the range from 431398 and 657001 (investing more in investigative journalism)
- Note that the range of predictions for circulation based on each different strategy are quite distinct from each other

(d) Using the model from Question 3(b), calculate 90% prediction intervals for the expected change in circulation of the newspaper under each of the three proposed strategic directions. Place these prediction intervals in a table, and contrast them in context.

```
predicted_circ_change <- round(predict(change_0413_lm, new_prizes,
                                       interval = "prediction", level = 0.9),3)

predicted_circ_change <- tibble(prizes = new_prizes$prizes_9014,
                             "lower limit (%)" = predicted_circ_change[,2],
                             "predicted circulation (%)" = predicted_circ_change[,1],
                             "upper limit (%)" = predicted_circ_change[,3])
kable(predicted_circ_change,
      caption = "Prediction of circulation with 90% confidence interval under
                3 proposed directions ")
```

Table 4: Prediction of circulation with 90% confidence interval under 3 proposed directions

prizes	lower limit (%)	predicted circulation (%)	upper limit (%)
3	-79.868	-34.450	10.969
25	-71.387	-26.075	19.236
50	-62.638	-16.559	29.520

- If a newspaper has 3 prizes, we are 90% confident that the newspaper will increase circulation in the range from -79.868% to 10.969%
- If a newspaper has 25 prizes, we are 90% confident that the newspaper will increase circulation in the range from -71.387% to 19.236%
- If a newspaper has 50 prizes, we are 90% confident that the newspaper will increase circulation in the range from -62.638% to 29.520%
- Prediction for a single newspaper is generally wider than prediction for the mean. As can be seen from the table, there isn't a clear distinction between the range of circulation change for 3 strategies.

Question Five: Limitations

(a) Discuss what limitations there might be to each of the models. Why might this model be insufficient for its application? You should discuss at least two limitations of these models in application.

The first limitation to both models is failing to satisfy the independence assumption for linear regression model. As mentioned earlier, as the measurement of circulation for each newspaper is influenced by the newspaper industry in the same time period, they are not independent of one another.

The second limitation to both models is failing to take into account the complexity of the problem. While the contribution from number of Pulitzer prizes could play a role in determining in the newspaper's circulation (assuming that it does), yet it is not the sole factor. Further exploratory data analysis should be done to select more relevant features for building the model. For example, transition from physical paper to online paper or the amount of investigative articles in proportion to the total number article produced are some potential factors. They can be used to build a better and more sophisticated model for prediction.

Lastly, there is not enough evidence to see a causation between 2 variables just because there is a correlation between them. It could potentially be the case that the high number of circulation is a factor that helps Boston Sun-Times achieve high number of Pulitzer Prizes over the past 25 years.

Conclusion

The objective of the project is to assess the future direction of Boston Sun-Times, which is the leading newspaper of Masthead Media. The newspaper has a reputation for producing outstanding investigative journalism and has been awarded an average of one Pulitzer Prize per year over the past 25 years. However, due to a decline in readership (currently at 453,869), Masthead Media is exploring whether maintaining the current investigative focus or shifting towards a more popular, tabloid-style approach would be more beneficial for their flagship publication.

Masthead Media is specifically interested in determining if increasing the number of Pulitzer Prizes won would result in a rise in circulation. To accomplish this, the project has developed two statistical models for evaluation. The predicted outcomes from the two models are subsequently interpreted and compared to one another.

One of the models forecasts a decline of 16.6% in circulation, while the other model anticipates an increase in circulation to approximately 532,381 with an estimated range of 431,398 to 657,001 if Boston Sun-Times invests more in investigative journalism and achieves 50 Pulitzer Prizes. Results from both models also indicate that investing less or maintaining the same level of investment in investigative journalism would lead to a decrease in circulation.

Three limitations for these models have been addressed. The first one concerns about the influence of the publication industry to the overall circulation that may result in the models' invalidity. Secondly, other important factors such as transitioning to online newspaper were not used for model prediction. Lastly, while there may be a correlation between the number of Pulitzer Prizes a newspaper receives and its circulation, it should be noted that an increase in the number of prizes does not necessarily result in an increase in circulation. In fact, the relationship could be bidirectional, or even influenced by other factors.

Based on the results and limitations of the two models, it is recommended that further data analysis should be done to select better and more relevant factors to use for building the models. Results from these models can only be used as first step in establishing a model for predicting circulation.