MATHS 7107 Data Taming Assignment 03

Possakorn Kittipipatthanapong

2023-03-12

Executive Summary

The objective of this project is to determine the appropriate direction to Boston Sun-Times, the newspaper's unwavering dedication towards exposing the truth and seeking justice is the reason behind its excellent reputation. Based on information Over the last 25 years. Masthead Media could receive recommendations to decide whether to continue to invest in the Sun-Times' investigative journalism, or to encourage the newspaper to take a more populist and tabloid slant approach to counter the recent drop in readership.

Especially, project is analyzed following the relationship consideration, parameter transformation, and forecasting the circulation and percentage change of circulations based on amount of prizes.

In conclusion, our model proposes an investment in excess of 50 Pulitzer Prizes to achieve a rise in circulation to approximately 532,380.7 with minimal reductions in comparison to the present scenario. Nevertheless, the constraints pose a challenge to the effectiveness of the output. To address these issues, it is imperative to make prudent decisions to overcome them.

Question One: Reading and Cleaning

Load the data contained in pulitzer.csv into R. A summary of the variables as represented in the csv file are below

```
column_detail <- tibble(
  columns = c("newspaper", "circ_2004", "circ_2013", "change_0413", "prizes_9014")
  , Description = c("The name of one of the United States' 50 largest newspapers, as at 2004",
                    "The newspaper's circulation in 2004",
                    "The newspaper's circulation in 2013",
                    "The percentage change in the newspaper's circulation, between 2004 and 2013
                    "The number of Pulitzer Prizes won by the newspaper's journalists between 19
                    )
)
kable(column_detail, caption = "Columns informations")
```

Table 1: Columns informations

columns	Description
newspaper	The name of one of the United States' 50 largest newspapers, as at 2004
circ_2004	The newspaper's circulation in 2004
circ_2013	The newspaper's circulation in 2013
change_0413	The percentage change in the newspaper's circulation, between 2004 and 2013
prizes_9014	The number of Pulitzer Prizes won by the newspaper's journalists between 1990 and 2014

```
pulitzer <- read.csv("sc/pulitzer .csv")
kable(head(pulitzer,5), caption = "pulitzer dataset with 5 rows")
```

Table 2: pulitzer dataset with 5 rows

newspaper	circ_2004	circ_2013	change_0413	prizes_9014
USA Today	2192098	1674306	-24%	3
Wall Street Journal	2101017	2378827	13%	51
New York Times	1119027	1865318	67%	118
Los Angeles Times	983727	653868	-34%	86
Washington Post	760034	474767	-38%	101

For our analysis, we would like to predict either average circulation between 2004 and 2013, or change in circulation between 2004 and 2013, using the number of Pulitzer Prizes between 1990 and 2014.

(a) Recode the change_0413 variable so it represents the percentage change in circulation between 2004 and 2013 as an integer. This will require manipulating the strings in change_0413.

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

```
pulitzer <- pulitzer %>%
  mutate(
    change_0413 = as.integer(sub("*%","", change_0413)), #Q1-a
    average_circulation = ( circ_2004 + circ_2013 ) /2 #Q1-b
  )
kable(head(pulitzer,5), caption = "pulitzer dataset with 5 rows after generating related variabl
```

Table 3: pulitzer dataset with 5 rows after generating related variables

newspaper	circ_2004	circ_2013	change_0413	prizes_9014	average_circulation
USA Today	2192098	1674306	-24	3	1933202.0
Wall Street Journal	2101017	2378827	13	51	2239922.0
New York Times	1119027	1865318	67	118	1492172.5
Los Angeles Times	983727	653868	-34	86	818797.5
Washington Post	760034	474767	-38	101	617400.5

Question Two: Univariate Summary and Transformation

(a) Describe the distribution of the variable representing average circulation, including shape, location, spread and outliers (Reminder: plots and summary statistics are useful here).

```
pulitzer %>%
  ggplot(aes(x = average_circulation)) +
  geom_histogram()
```

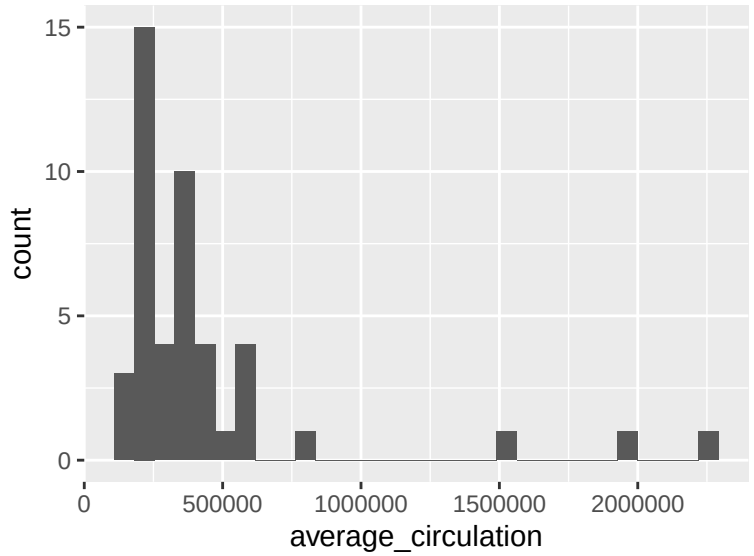

Figure 1: Histrogram of average circulation on publications

```
pulitzer %>%
  ggplot(aes(x = average_circulation)) +
  geom_boxplot()
```

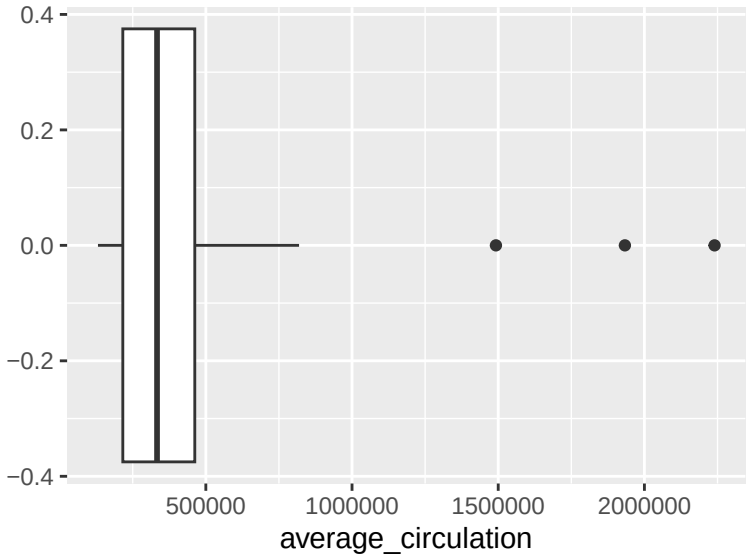

Figure 2: Boxplot of average circulation on publications

Statistic Summary

```
summary(pulitzer$average_circulation)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  131004  216013  333083  437141  462153 2239922
```

For shape of this data set, the right-skewed pattern is represented with bi-model distribution. For Location, mean and median equal to 437,141 and 333,083 respectively. The measuring of spread following the inter-quartile range is 2.4614×10^5. There are potential outliers in average circulation.

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

```
pulitzer %>%
  ggplot(aes(x = change_0413)) +
  geom_histogram()
```

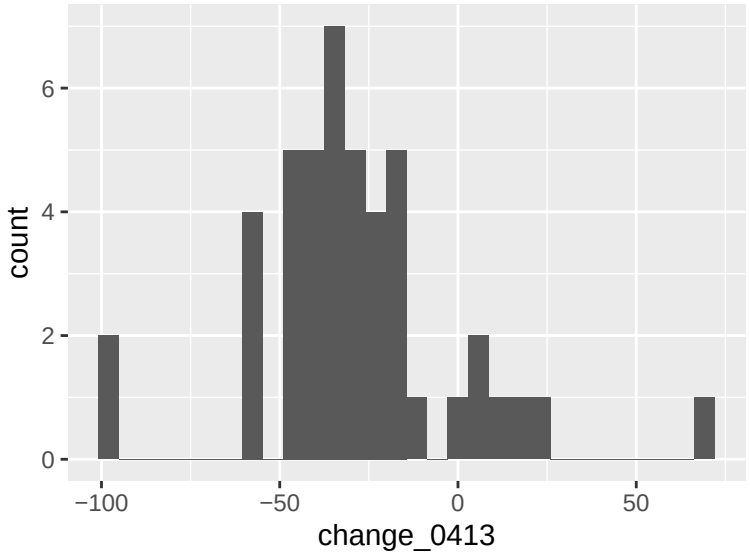

Figure 3: Histrogram of change_0413 on publications

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

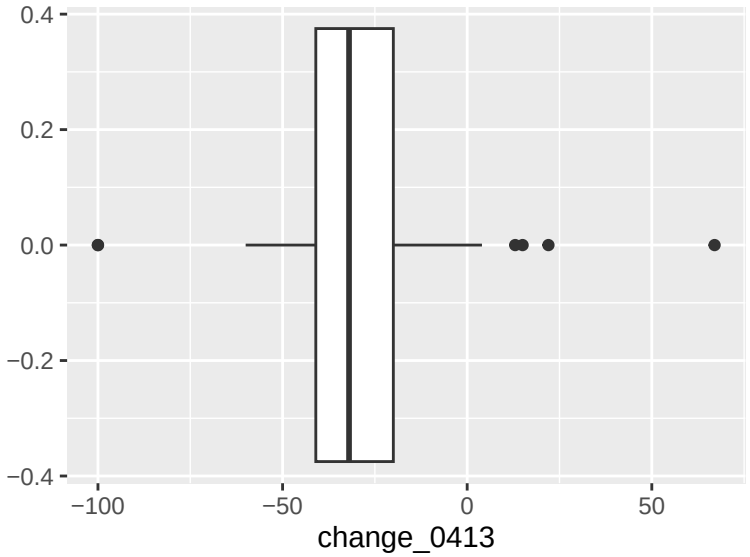

Figure 4: Boxplot of change_0413 on publications

Statistic Summary

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

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

For shape of this data set, symmetric distribution is represented without skewness. For Location, mean and median equal to -29.040 and -32.000 respectively. The measuring of spread following the inter-quartile range is 21. There are potential outliers in change_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.

For average circulation, there are potential skewed pattern in the distribution. From my hypothesis, we should resolved by a log transformation.

```
pulitzer <- pulitzer %>%
  mutate(
    log_average_circulation = log(average_circulation)
  )
pulitzer %>%
  ggplot(aes(x = log_average_circulation)) +
  geom_histogram()
```

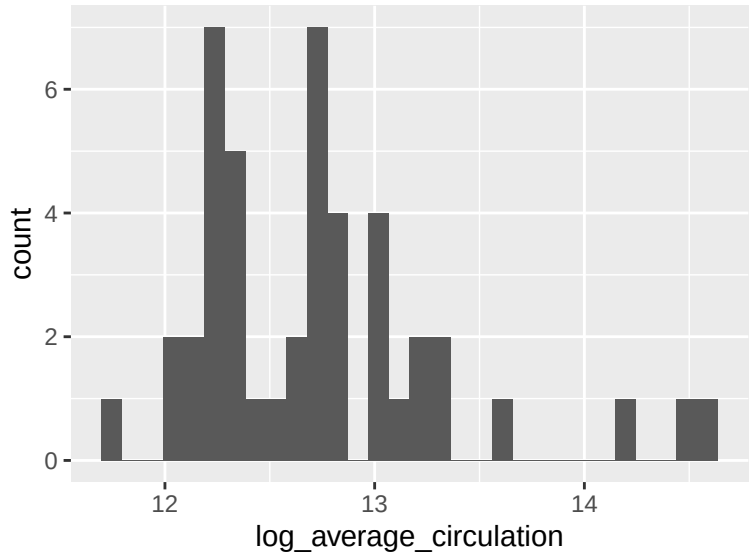

Figure 5: Histrogram of average circulation on publications after log transformation

Therefore, we could resolve the skew with log transformation.

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?

```
pulitzer_lm_avg <- lm(log_average_circulation ~ prizes_9014, data = pulitzer)
summary(pulitzer_lm_avg)
```

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

Answer: slope = 0.0133 and Intercept = 12.521. Following the non-zero slope, there is statistically significant relationship between the number of Pulitzer Prizes, and average circulation by p value less than 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?

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

```
##
## 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
```

Answer: slope = 0.381 and Intercept = -35.592. Following the non-zero slope, there is statistically significant relationship between the number of Pulitzer Prizes, and change_0413 by p value less than 0.05 .

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

For model average circulation vs number of Pulitzer Prizes

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

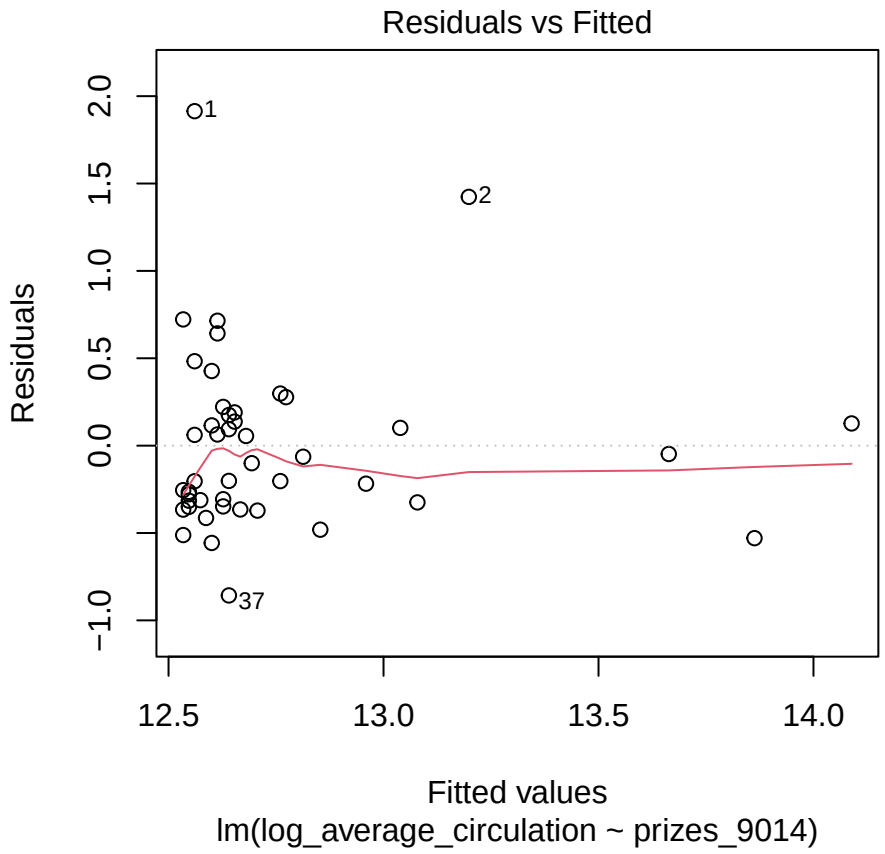

Figure 6: residual versus fitted plot - check Linearity on average circulation

01 Linearity: we could see the scatter point following the zero residual line and there is no significant trends out off this line. Following plot, it meet linearity.

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

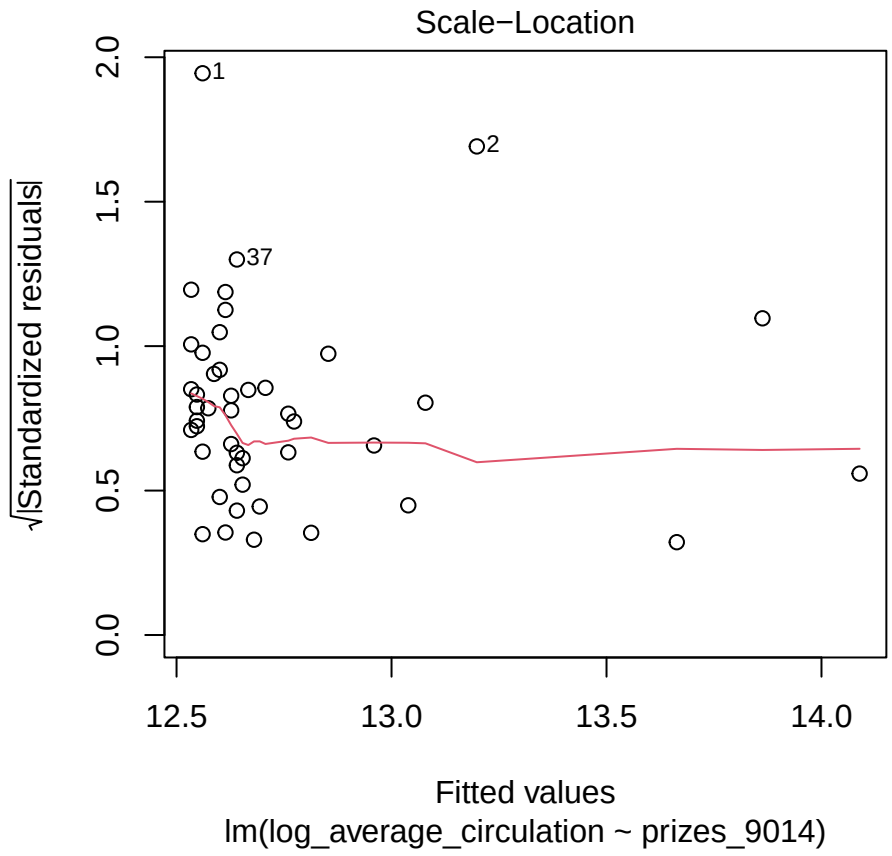

Figure 7: square root of the standardised residual versus fitted plot - check homoscedasticity on average circulation

02 homoscedasticity: Most of data is dense in the range of 12.500 to 13.000. And, Most of points evenly spread from left to right. So, no apparent trends and we could indicate that this model has constant spread. Following plot, it meet homoscedasticity.

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

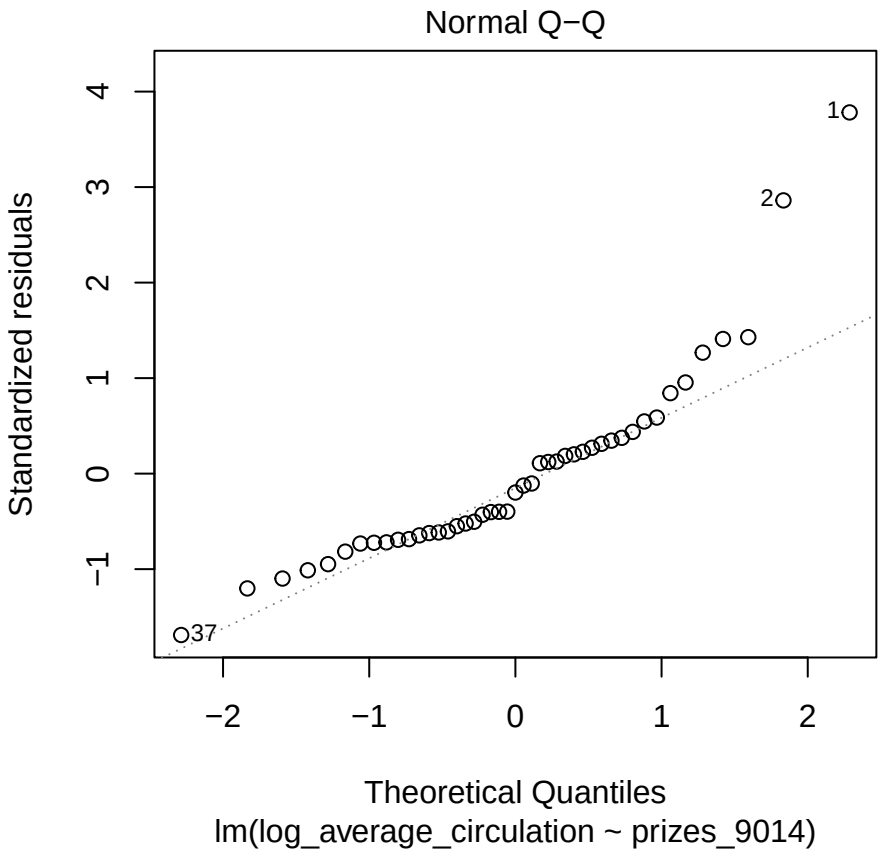

Figure 8: normal QQ-plot - check Normality on average circulation

03 Normality: Following the QQ plot. Most of points locate along the dotted line as normally distributed. However, there are points drift away from the line for point less than -1 and more than +1. Following plot, it meet Normality.
04 Independence: Cannot summarize the output with this information. It require some of knowledge of where the data comes from and related domain expertise.

For model average circulation vs number of Pulitzer Prize

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

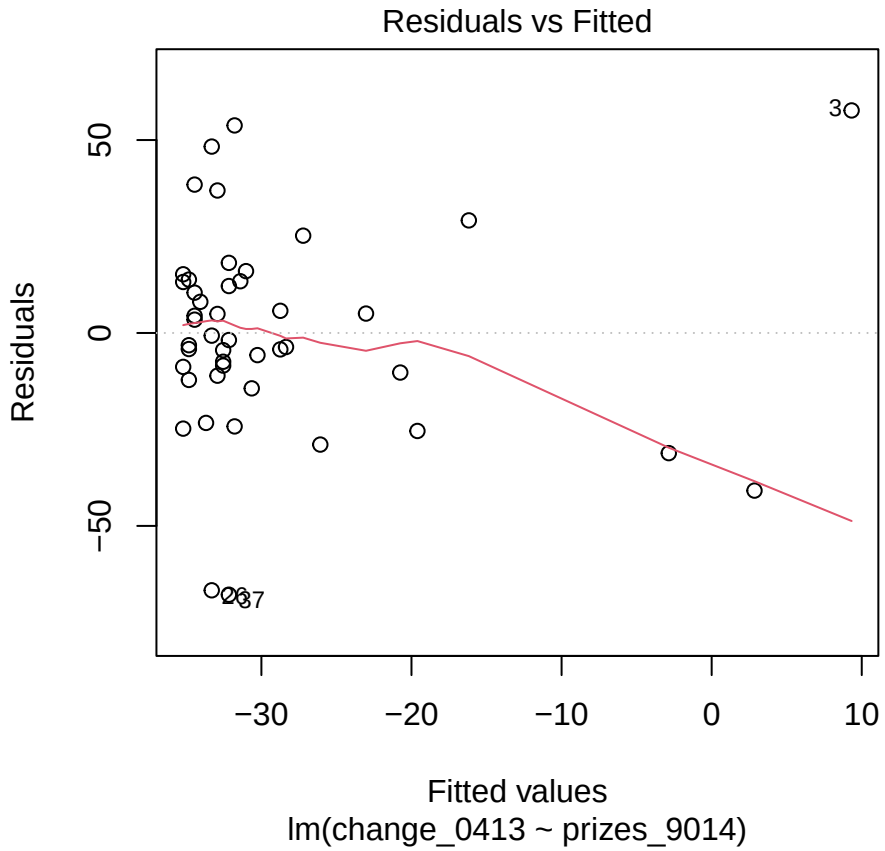

Figure 9: residual versus fitted plot - check Linearity on change_0413

01 Linearity: we could see the scatter point following the zero residual line and there is some significant trends out off this line. However, this trend is the minority of our data set. So we could summarize that it meet linearity.

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

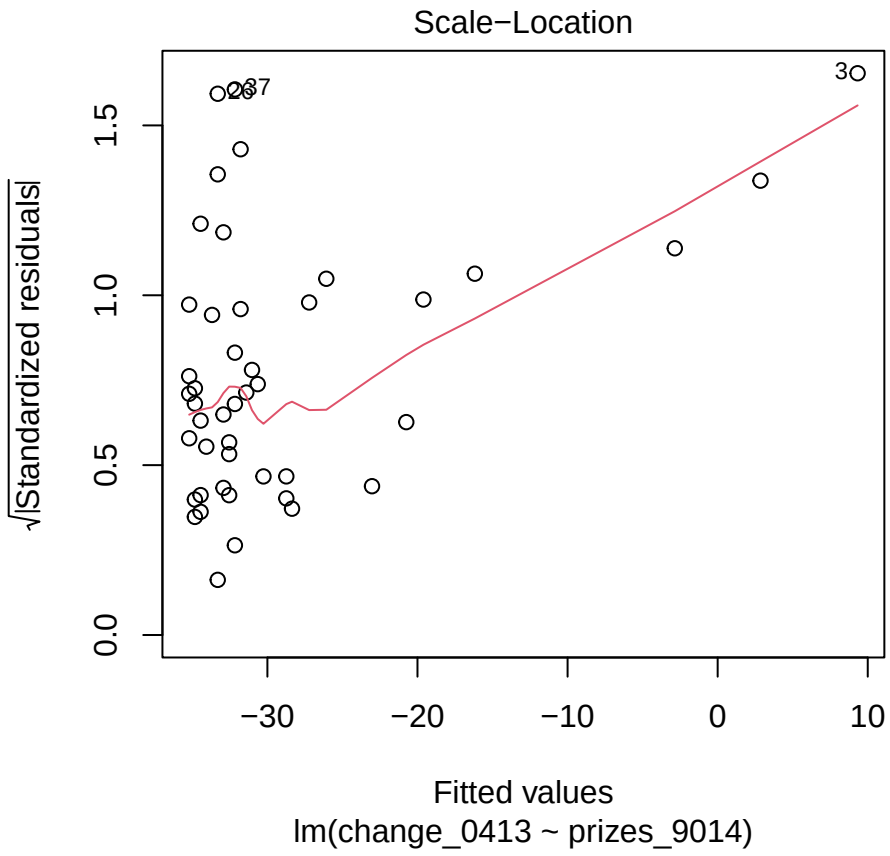

Figure 10: square root of the standardised residual versus fitted plot - check homoscedasticity on change_0413

02 homoscedasticity: Most of data is dense in the range of -40 to -20. And, Most of points evenly spread from left to right but only small amount of point represent the outliers for this trends. So we could indicate that this model has constant spread. Following plot, it meet homoscedasticity.

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

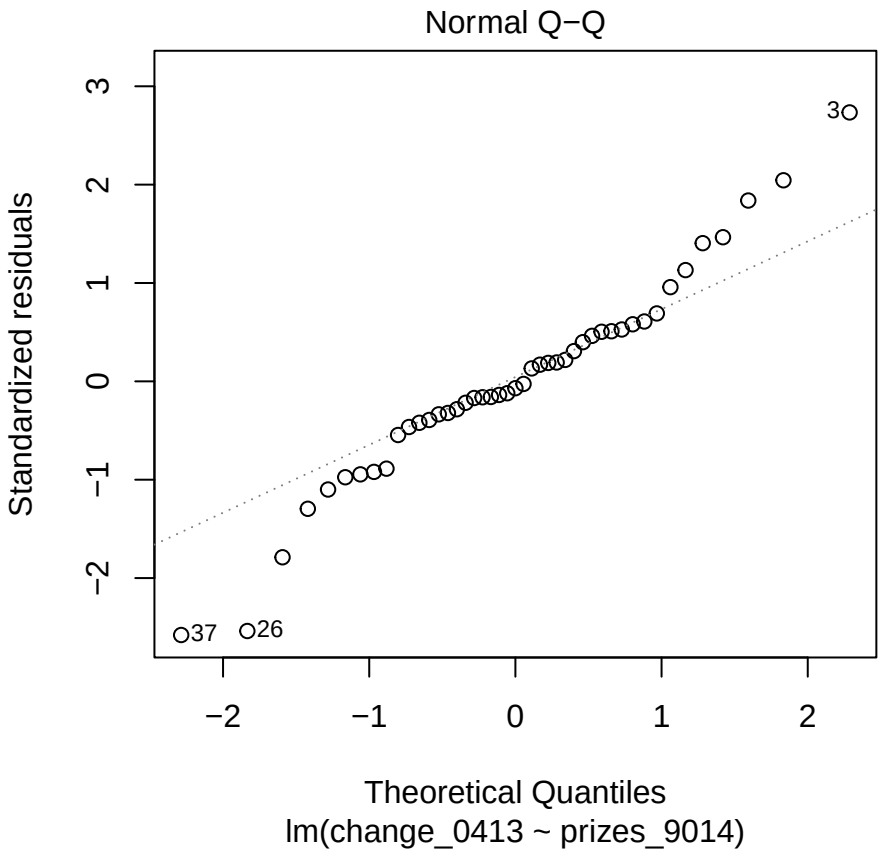

Figure 11: normal QQ-plot - check Normality on change_0413

03 Normality: Following the QQ plot. half of points locate along the dotted line as normally distributed. However, there are amount of points drift away from the line for point less than -1 and more than +1. Following plot, it doesn't meet Normality.
04 Independence: Cannot summarize the output with this information. It require some of knowledge of where the data comes from and related domain expertise.

Question Four: Prediction

Masthead Media is considering three strategic directions for the Boston Sun-Times. These are:
* Investing substantially less in investigative journalism than present. In this case, Masthead Media projects that the newspaper will be awarded 3 Pulitzer Prizes in the next 25 years.
* Investing the same amount in investigative journalism than present, leading to the award of 25 Pulitzer Prizes in the next 25 years.
* Investing substantially more in investigative journalism, leading to the award of 50 Pulitzer Prizes.

For the following questions, assume that the projected number of prizes under each possible strategic direction is known; that is, do not incorporate any uncertainty in the number of Pulitzer Prizes.

(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?

```
strategic <- tibble(
  prizes_9014 = c(3,25,50)
)
predict_avgcir = exp(predict(pulitzer_lm_avg, strategic)) # using the exponential to calculate t
predict_avgcir_table <- data.frame(direction = c(1,2,3),pulitzer_prizes = c(3,25,50),predict_avg
kable(predict_avgcir_table, caption = "Prediction of circulation by three different Pulitzer Pri
```

Table 4: Prediction of circulation by three different Pulitzer Prizes strategics

direction	pulitzer_prizes	predict_avgcir
1	3	285094.6
2	25	381899.6
3	50	532380.7

Answer: Following the three strategic directions, 3rd directions to invest substantially more in investigative journalism provide the highest circulation as 532,380.700 circulations compared to first and second direction that equal to 285,094.600 circulations and 381,899.600 circulations 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.

```
predict_change <- predict(pulitzer_lm_change, strategic)
predict_change_table <- data.frame(direction = c(1,2,3),pulitzer_prizes = c(3,25,50),predict_cha
kable(predict_change_table, caption = "Prediction of percentage change in circulation by three d
```

Table 5: Prediction of percentage change in circulation by three different Pulitzer Prizes strategics

direction	pulitzer_prizes	predict_change
1	3	-34.44960
2	25	-26.07541
3	50	-16.55929

Answer: Following the three strategic directions, 3rd directions to invest substantially more in investigative journalism provide the least negative changed of circulation as -16.559 % compared to first and second direction that equal to -34.450 % and -26.075 % respectively.

(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.

```
pulitzer_lm_avg_con <- data.frame(direction = c(1,2,3),
                                  pulitzer_prizes = c(3,25,50),
                                  as.tibble(exp(predict(pulitzer_lm_avg, strategic, interval = "
                                                        ))
                                            )
kable(pulitzer_lm_avg_con, caption = "Prediction of circulation with 90% confidence intervals")
```

Table 6: Prediction of circulation with 90% confidence intervals

direction	pulitzer_prizes	fit	lwr	upr
1	3	285094.6	245996.8	330406.3
2	25	381899.6	333781.7	436954.2
3	50	532380.7	431398.5	657000.9

For awarded 3 Pulitzer Prizes, the 90% confidence interval represent of average circulation in the range from 245996.800 to 330406.3.

For awarded 25 Pulitzer Prizes, the 90% confidence interval represent of average circulation in the range from 333781.7 to 436954.2.
For awarded 50 Pulitzer Prizes, the 90% confidence interval represent of average circulation in the range from 431398.5 to 657000.9.

(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.

```
pulitzer_lm_change_pre <- data.frame(direction = c(1,2,3),pulitzer_prizes = c(3,25,50), as.tibbl
kable(pulitzer_lm_change_pre, caption = "Prediction of Prediction of percentage change with 90%
```

Table 7: Prediction of Prediction of percentage change with 90% prediction intervals

direction	pulitzer_prizes	fit	lwr	upr
1	3	-34.44960	-79.86819	10.96898
2	25	-26.07541	-71.38673	19.23591
3	50	-16.55929	-62.63825	29.51967

For awarded 3 Pulitzer Prizes, the 90% prediction interval represent of change in circulation in the range from -79.868 to 10.969.
For awarded 25 Pulitzer Prizes, the 90% prediction interval represent of change in circulation in the range from -71.387 to 19.236.
For awarded 50 Pulitzer Prizes, the 90% prediction interval represent of change in circulation in the range from -62.638 to 29.520.

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.

First limitation indicate that these model don't involve with independence assumption for linear regression. Acquiring an understanding of the source, nature, and collection methodology of data is essential and cannot ignore to control the bias, standardization, and noise of the results. Therefore, if there are domain experts to review and provide the standard method to collect the data points, it will lead to better performance and prevent the mislead situations.

Secondly, the chosen prediction model has distinct time frames for the responses and predictors - 2004 to 2013 and 1990 to 2014, respectively. Consequently, this inconsistency can result in erroneous outcomes and inappropriate directions.

Lastly, to create an effective model, a substantial amount of data points, i.e., more than 45 transactions, are necessary to feed the algorithm and produce consistent and accurate output for better decision-making.

Conclusion

Related to the objective of this projects, Masthead Media is currently deliberating on whether to persist with investing in the investigative journalism of the Sun-Times or to prompt the newspaper to adopt a more populist, sensationalist approach to attract more circulations and increase growth rate of readers. Presently, the Boston Sun-Times has a readership of 453,869.

Our analysis explore the relationship on both of average circulation and percentage change of circulation during year 2004 and 2013 with number of Pulitzer prices between 1990 and 2014. Then, predict the forecasting models in three difference strategics to compared the performance for determining the most appropriate solutions to our business.

The relationship between two interesting parameters and number of Pulitzer prices represent the statistically important in the positive correlation. Moreover, One of the prediction models show the impressive growth of circulation to 532,380.7 in the confidence range of 431,398.5 and 657,000.9, while the growth decrease 16.559 percent in circulation. Therefore, results of 3rd direction from forecasting model indicate that allocation of significant funds substantially more investing journalism resulted in the achievement of 50 Pulitzer Prizes be the best recommendations.

There are few limitations following Comprehension of the origin and data gathering approach, consistent of time frame, and significant quantity of data observations. If the inconsistent output is utilized to make decisions, it can lead to inefficiency and misguided conclusions. Hence, these results will serve as an initial stage to carry out an exploratory analysis and evaluate input parameters to develop a better model in the future.