MATHS 7107 Data Taming Assignment Three Questions

Chang Dong

2023-03-11

Due date: 5pm, Wednesday 15th March 2023.

Scenario

You are performing data science for a major media company, Masthead Media, in the United States, determining the direction of their flagship newspaper, the Boston Sun-Times. The Boston Sun-Times is known for investigative journalism, for which it has won a number of Pulitzer Prizes. The newspaper's editorial staff believe that this unrelenting focus on truth and justice is responsible for the newspaper's sterling reputation. The Boston Sun-Times has received, on average, a Pulitzer Prize every year for the past 25 years.

Masthead Media is deciding whether to continue to invest in the Sun-Times' investigative journalism, or to encourage the newspaper to take a more populist, tabloid slant, in a bid to arrest a recent decline in readership. The Boston Sun-Times currently has a circulation of 453,869.

Masthead Media wants to know:

- Whether publications that win more Pulitzer Prizes have a smaller, or a larger, average circulation;
- Whether publications that win more Pulitzer Prizes see a percentage increase, or decrease, in circulation, during the period that they win the prizes; and
- If these relationships exist, how the trajectory of the Boston Sun-Times's circulation might change depending on the newspapers strategic direction.

To this end, using provided data on newspaper circulation and number of Pulitzer Prizes in a 25 year period, you should answer Questions 1–5 below. You will also need to report the results to the company, using an executive summary (at the start of your assignmnet) and a conclusion (at the end of your assignmnet).

Some rules about your submissions:

- You must complete this assignment using R Markdown;
- Your assignment must be submitted as pdf only on MyUni with Rmd file (R markdown file);
- You must include units when providing solutions;
- Include any working when providing solutions;
- Provide all numerical answers to 3 decimal places;
- Make sure you include both your code and R output / plots in your answers;
- Make sure any tables or plots included have captions;
- Do not write directly on the question sheet;
- You can submit more than once if you find errors and your latest submission will be marked;
- Make sure you only upload one pdf document for your final submission. If you submit multiple pages (i.e. one per question) you will be deducted 10% per page submitted;

• Penalties for late submission - within 24 hours 40% of final mark. After 24 hours, assignment is not marked and you get zero; and

• Finally, make sure you check your submitted assignment is the correct one, as we cannot accept other submissions after the due date.

Load all needed pkgs

```
pacman::p_load("tidyverse","stringr")
```

Executive summary

You need to write an executive summary to report your findings to Masthead Media. This summary should include:

• A recap of the project outline including what the purpose of your project is and what problem you are solving.

• A very brief outline of what you did.

• Your final recommendation to Masthead Media.

Your executive summary should be written as if you are presenting it to Masthead Media. They will read the executive summary to get a quick overview; to evaluate the quality of the report; or even to make a decision. It should be no more than half a page, and contain no figures, plots or code. It should be written in plain English with no or minimal terminology. If terminology is included, it must be explained.

This section will be easier to write once you have completed questions 1-5.

SUMMARY HERE

The purpose of our project is to analyze the data and then build models to determine the development directions of the newspaper, the Boston Sun-Times. And we aim to solve the problem that "Whether publications that win more Pulitzer Prizes have a smaller or a larger average circulation, see a percentage increase, or decrease, in circulation". In this project, we cleaned, analyzed, and modified the data, then used the finalized data to build two models to predict future circulation. Finally, we find that winning more Pulitzer Prizes in 25 years can have some positive influence on circulation, but the relationship is not so strong and reliable. As a decision maker, don't pool too much money into this field, because it is risky, instead find another way to improve the circulation of news.

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.

Variable	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

Variable	Description
prizes_9014	The number of Pulitzer Prizes won by the newspaper's journalists between 1990 and 2014

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.

```
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.
```

```
pulitzer$change_0413 <-  str_match(pulitzer$change_0413,"-?(\\d+)")[,1]
pulitzer$change_0413 <- as.integer(pulitzer$change_0413)
head(pulitzer)
```

```
## # A tibble: 6 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
```

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

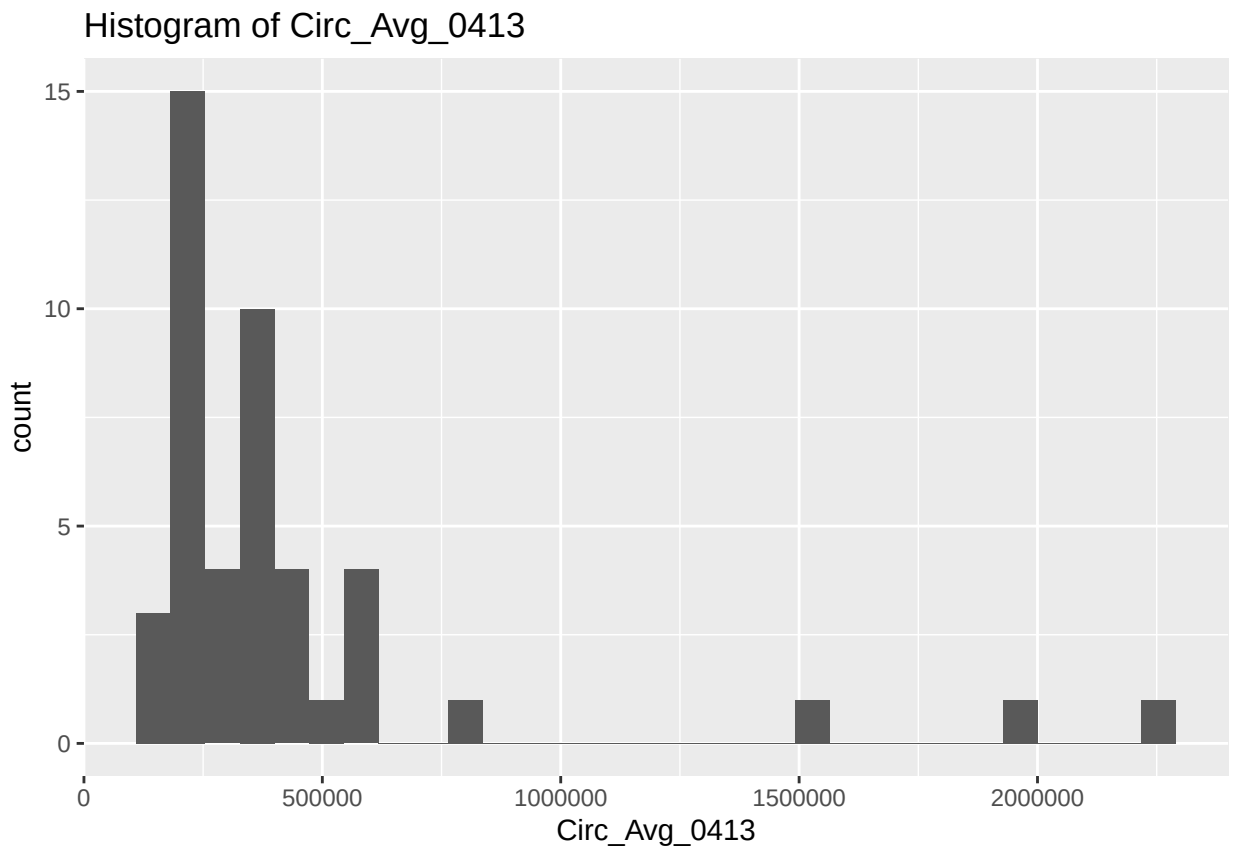
```
pulitzer$Circ_Avg_0413 <- (pulitzer$circ_2004 + pulitzer$circ_2013) /2
head(pulitzer)
```

```
## # A tibble: 6 x 6
##   newspaper           circ_2004 circ_2013 change_0413 prizes_9014 Circ_Avg_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.
## 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
```

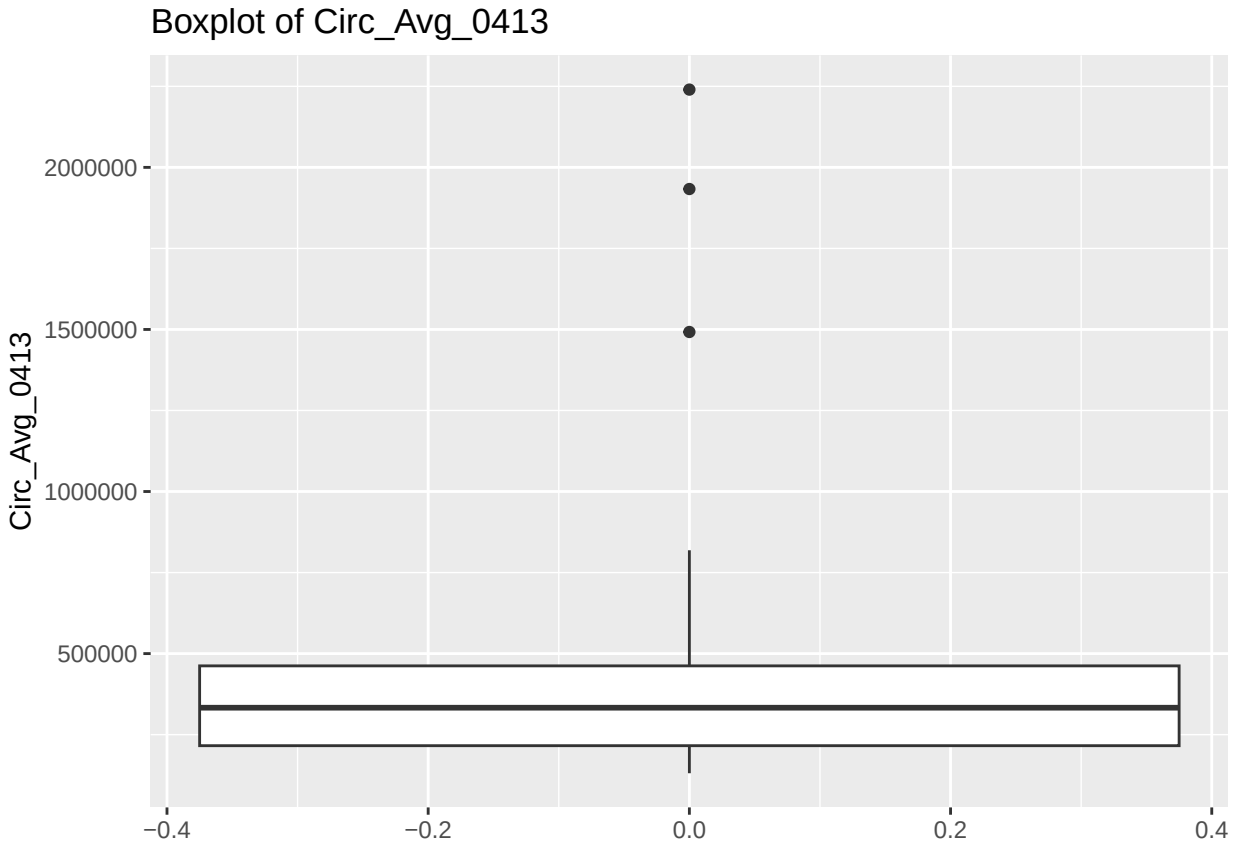
Question Two: Univariate Summary and Transformation

(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(Circ_Avg_0413)) + geom_histogram(bins = 30) +ggtitle("Histogram of Circ_Avg_0413")
```




```
pulitzer %>%
  ggplot(aes(y = Circ_Avg_0413)) + geom_boxplot() +ggtitle("Boxplot of Circ_Avg_0413")
```




```
round(c(summary(pulitzer$Circ_Avg_0413)),3)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
##  131004.0  216012.5  333083.0  437140.7  462152.5 2239922.0
```

shape: The variable has a right-skewed and unimodal shape we can see from the histogram.

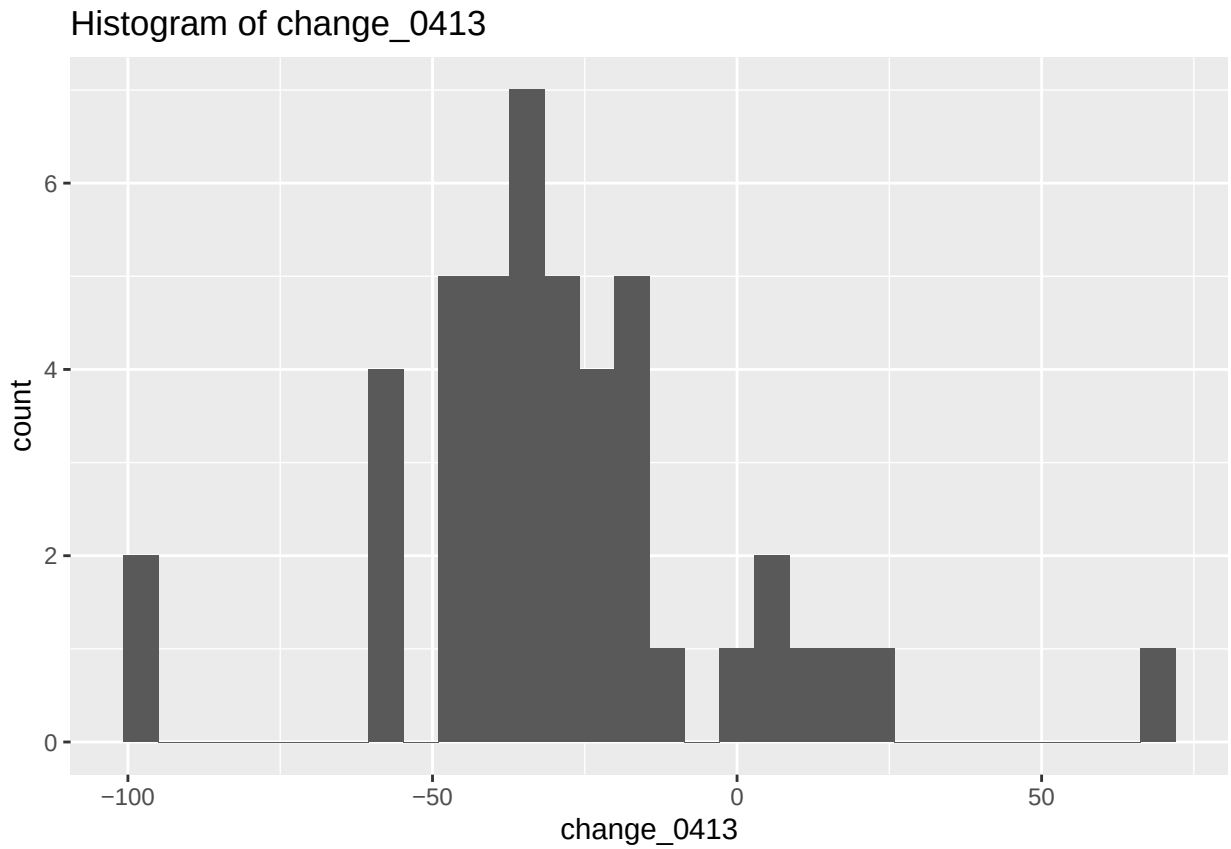
location: The median value is 333083.000 we can see from the summary table.

spread:The minimum value is 131004.000, and the maximum value is 2239922.000. 75% quantile value is 462152.000 , while 25% quantile value is 216012.000. IQR = Q3 - Q1 = 462152.000 - 216012.000 = 246140.000

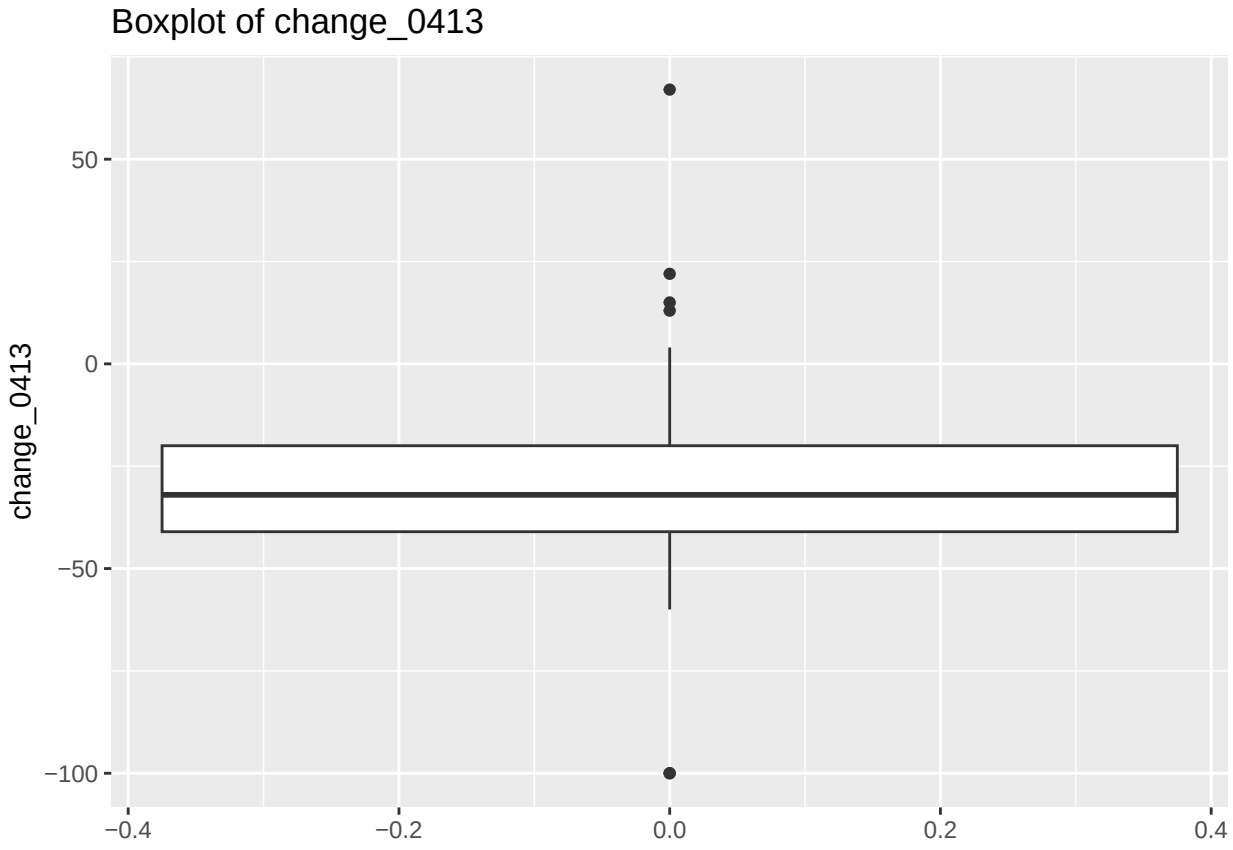
outliers: From the boxplot, we can see the outliers were only apeared in the right side, value bigger than Q3+1.5IQR= 831362.000 are outliers (3 outliers from the boxplot)

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

```
pulitzer %>%
  ggplot(aes(change_0413)) + geom_histogram(bins = 30) +ggtitle("Histogram of change_0413")
```




```
pulitzer %>%
  ggplot(aes(y = change_0413)) + geom_boxplot() +ggtitle("Boxplot of change_0413")
```




```
round(c(summary(pulitzer$change_0413)),3)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -100.000  -41.000  -32.000  -29.044  -20.000   67.000
```

shape: The variable shape is almost symmetric unimodal(with slight right-skewed).

location: The median value is -32.000 we can see from the summary table.

spread:The minimum value is -100.000, and the maximum value is 67.000. 75% quantile value is -20.000, while 25% quantile value is -41.000. IQR = Q3 - Q1 = -20.000 - (-41.000) = 21.000.

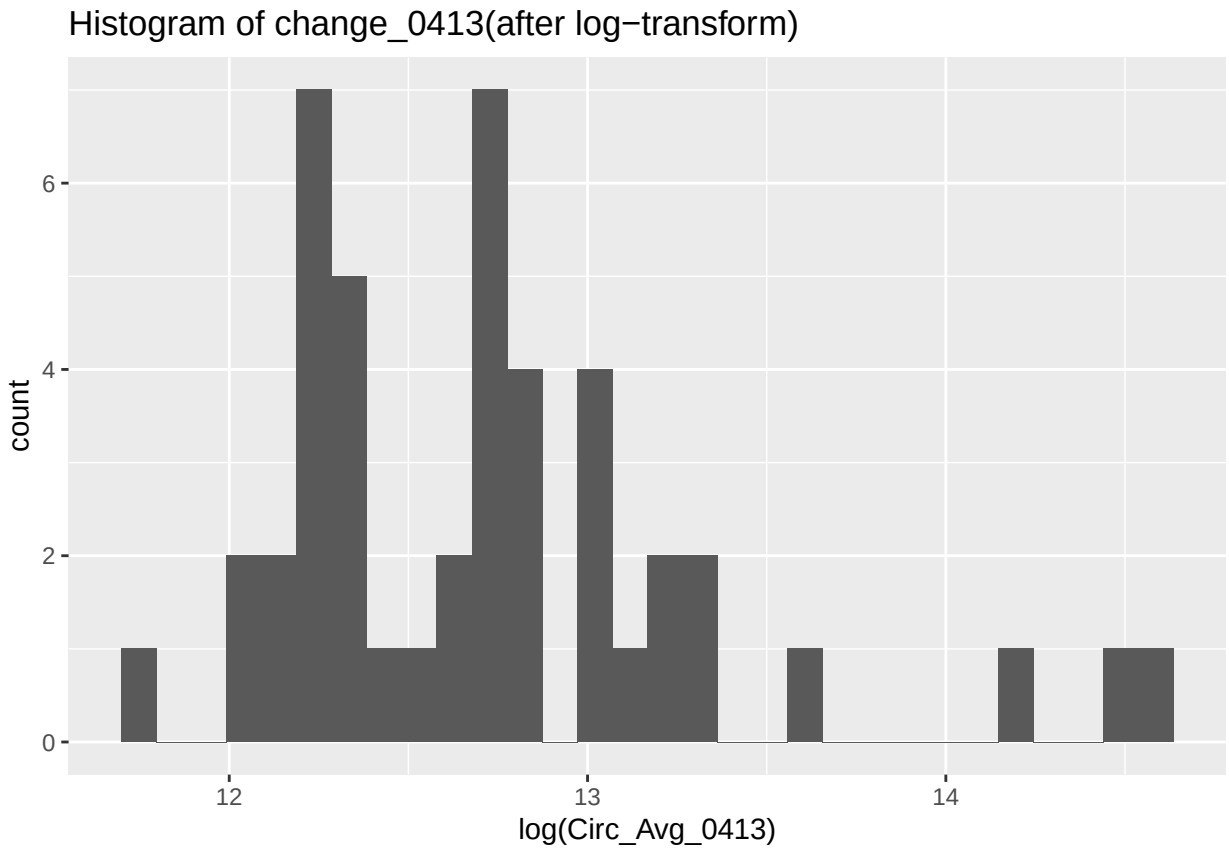
outliers: From the boxplot, we can see there are 1 lower outlier(at the bottom) and 4 upper outlier(at the top). Value outside this interval can be considered as outlier, which is [Q1-1.5IQR,Q3+1.5IQR] = [-72.500, 11.500].

(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.

According to the chart from (a) and (b), we can infer that only Circ_Avg_0413 should be transformed becuase it has a obvious right-skewed, while change_0413 is alomost symmetric, so it don't need to be transformed. But we can do log-transform both for the two variables to see what will happen.

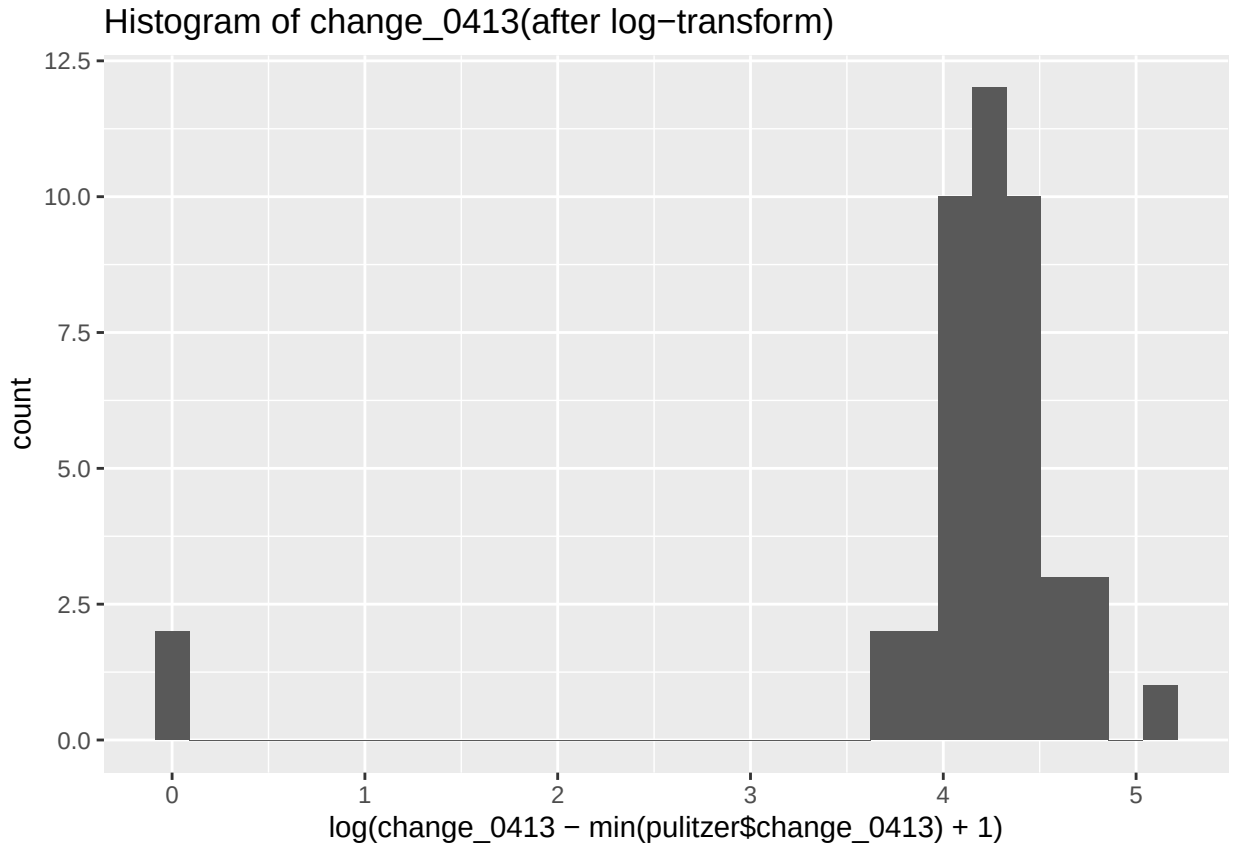
For Circ_Avg_0413: After log-transformation, the shape of Circ_Avg_0413 changed from right-skewed to almost symmetric(slight right-skewed). So it should be transformed because most of the value were crowded in a small value, after log-transform, interval between small values can be enlarged from log-scale.

```
pulitzer %>%
  ggplot(aes(log(Circ_Avg_0413))) + geom_histogram(bins = 30) +ggtitle("Histogram of change_0413(after
```



For change_0413: Before log-transformation, we need to subtract min(change_0413) then add 1 to every observations to make our math operation valid(log negative value is invalid), and the min value was transformed to log(1)). But the chart shows that the shape wasn't changed from almost symmetric to symmetric, so it is not suitable for log-transformation. Another reason is that the data don't have too many points crowded in small value intervals, so there is no need to do a log transformation to a almost symmetric distrubution data.

```
pulitzer %>%
  ggplot(aes(log(change_0413 - min(pulitzer$change_0413)+1))) + geom_histogram(bins = 30) +ggtitle("His
```



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?

```
lm_1 <- lm(log(Circ_Avg_0413) ~ prizes_9014,data = pulitzer)
summary(lm_1)
```

```
##
## Call:
## lm(formula = log(Circ_Avg_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
```

Intercept: the value of log(Circ_Avg_0413) is 12.521 when prizes_9014 equals to 0.

Slope: with increasing one unit of prizes_9014, the increment of log(Circ_Avg_0413) is 0.013

which can be interpreted as : $log(Circ\_Avg\_0413) = 0.013 \times prizes\_9014 + 12.521$

There is a statistically significant relationship between the number of Pulitzer Prizes, and average circulation, because our model shows the p-value of coefficient is lower than 0.05 which is 6.91e-05, so we should reject our alternative assumption(they don't have linear relationship).

(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?

```
lm_2 <- lm(change_0413 ~ prizes_9014, data = pulitzer)
summary(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
```

Intercept: the value of change_0413 is -35.592 when prizes_9014 equals to 0.

Slope: with increasing one unit of prizes_9014, the increment of change_0413 is 0.381.

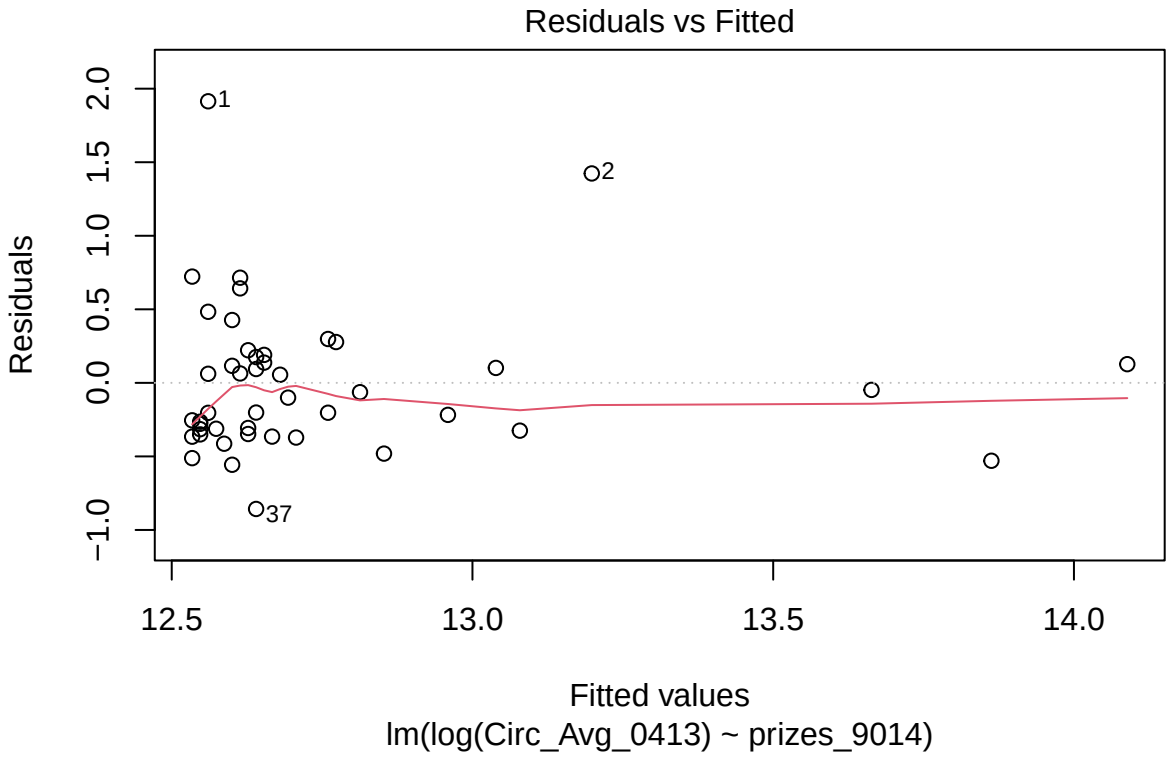
which can be interpreted as : $change_0413 = 0.381 \times prizes\_9014 - 35.592$

There is a statistically significant relationship between the number of Pulitzer Prizes, and change in circulation, because our model shows the p-value of Intercept is lower than 0.05 which is 0.0192, so we should reject our alternative assumption(they don't have linear relationship).

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

For lm_1: log(Circ_Avg_0413) ~ prizes_9014 **Linearity:**

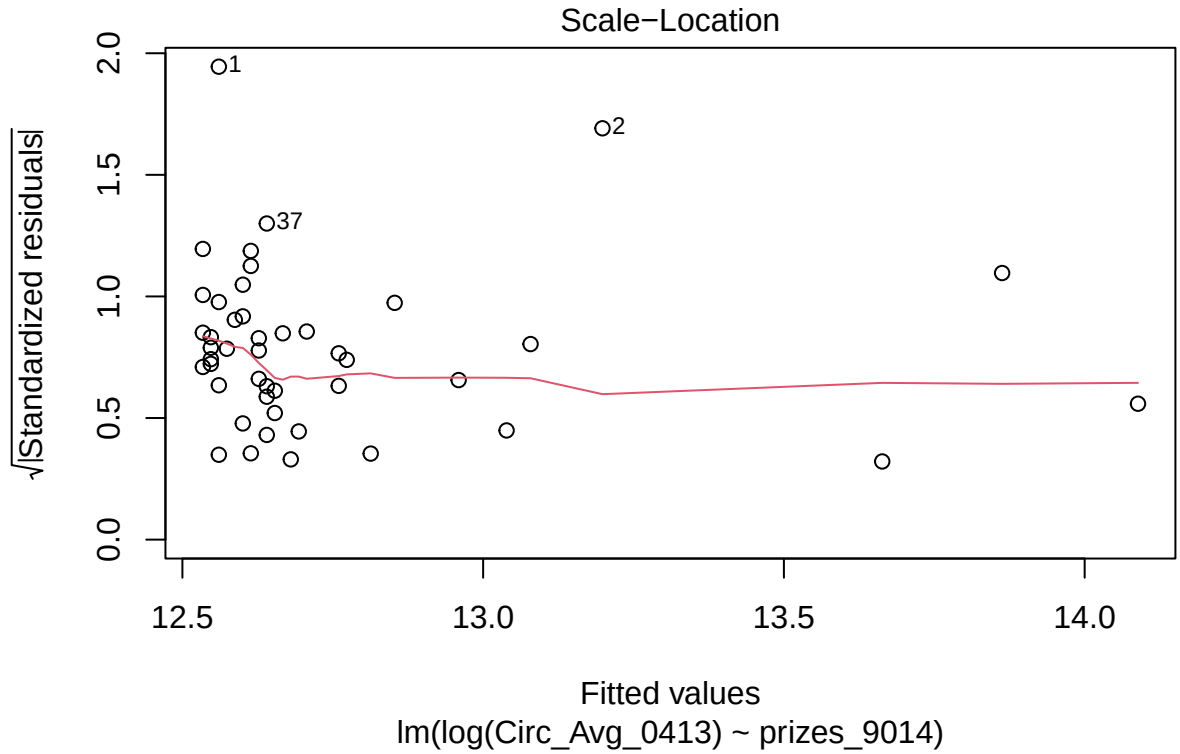
```
plot(lm_1, which = 1)
```



We can see these dots roughly spread around 0, the trend of red line looks straight and slightly bent by some outliers. So it basically linear.

Homoscedasticity:

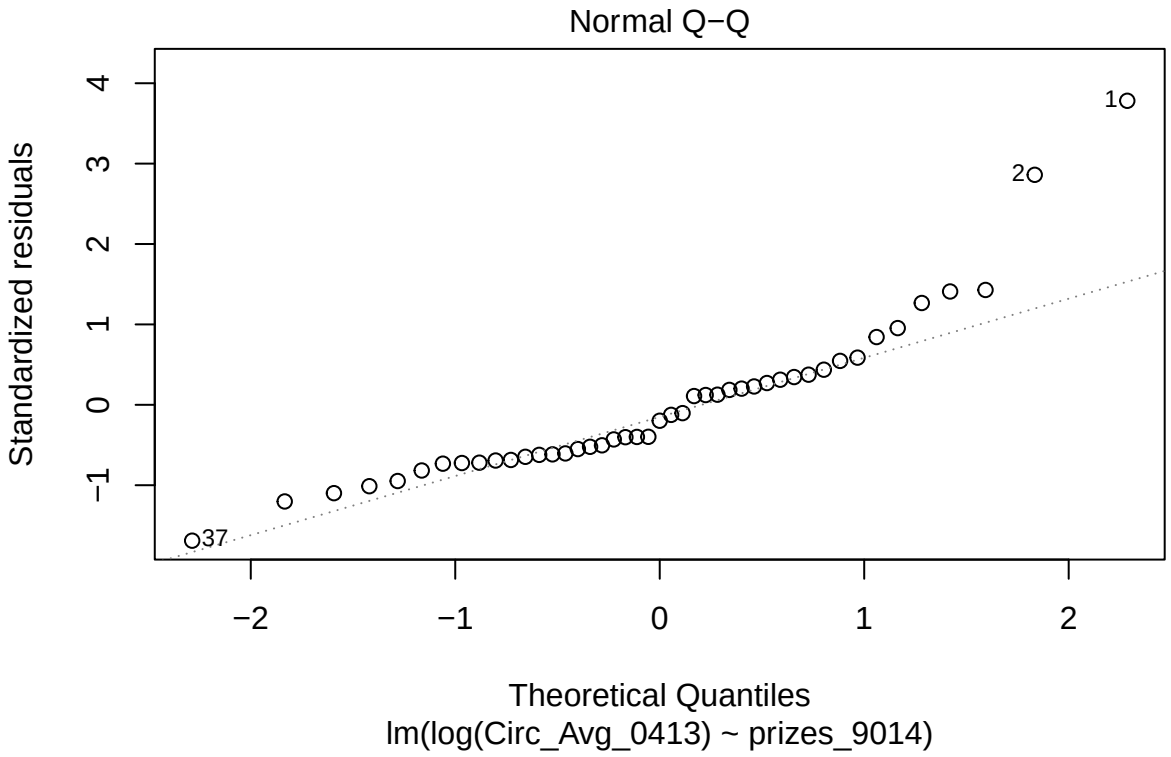
```
plot(lm_1, which = 3)
```



We can see the red line is almost straight, the $\sqrt{standardized\ residual}$ do not have an apparent trend from left to right,it has constant. So the $\sqrt{standardized\ residual}$ is roughly homogeneous.

Normality:

```
plot(lm_1, which = 2)
```

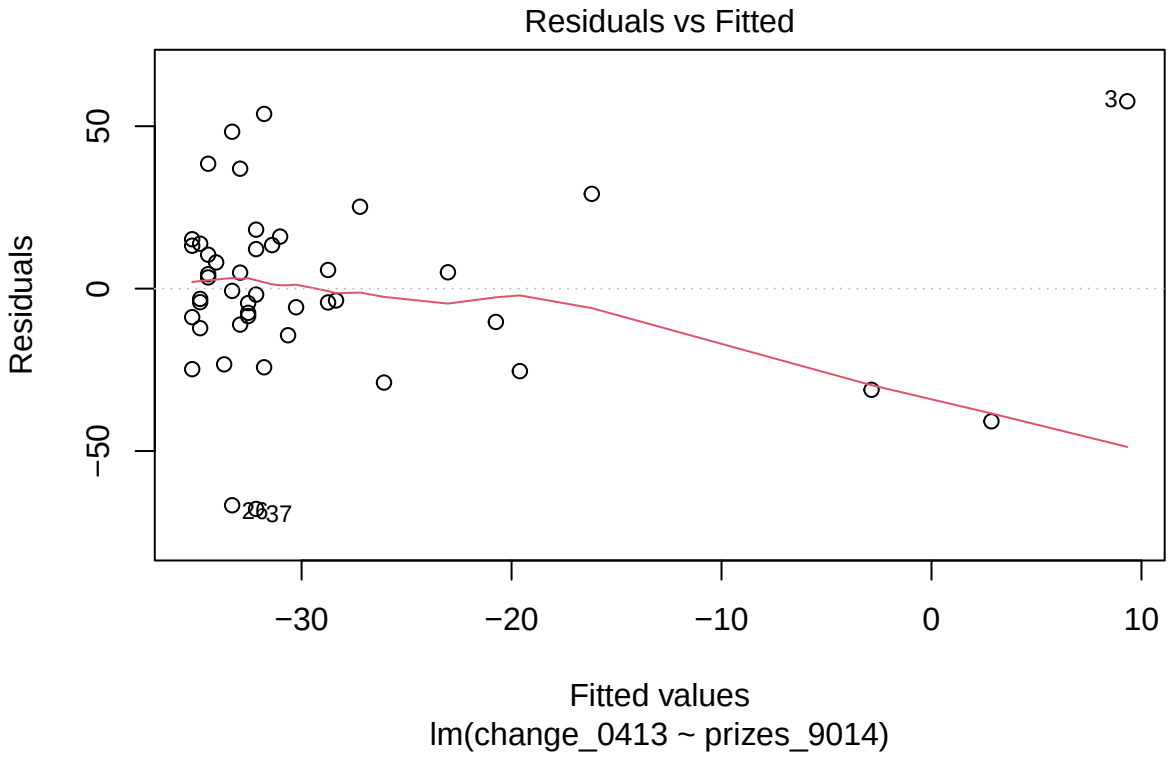


This looks roughly good, most of the points lie quite close to the normal quantile reference line, and only small amount points deviate from the line. So it roughly obey the assumption of Normality.

Noise independent Independence relies on the subjects being independent of each other; if these subject are independent with each other, noise can be independent, othewise they can be not.

For lm_2 model **Linearity:**

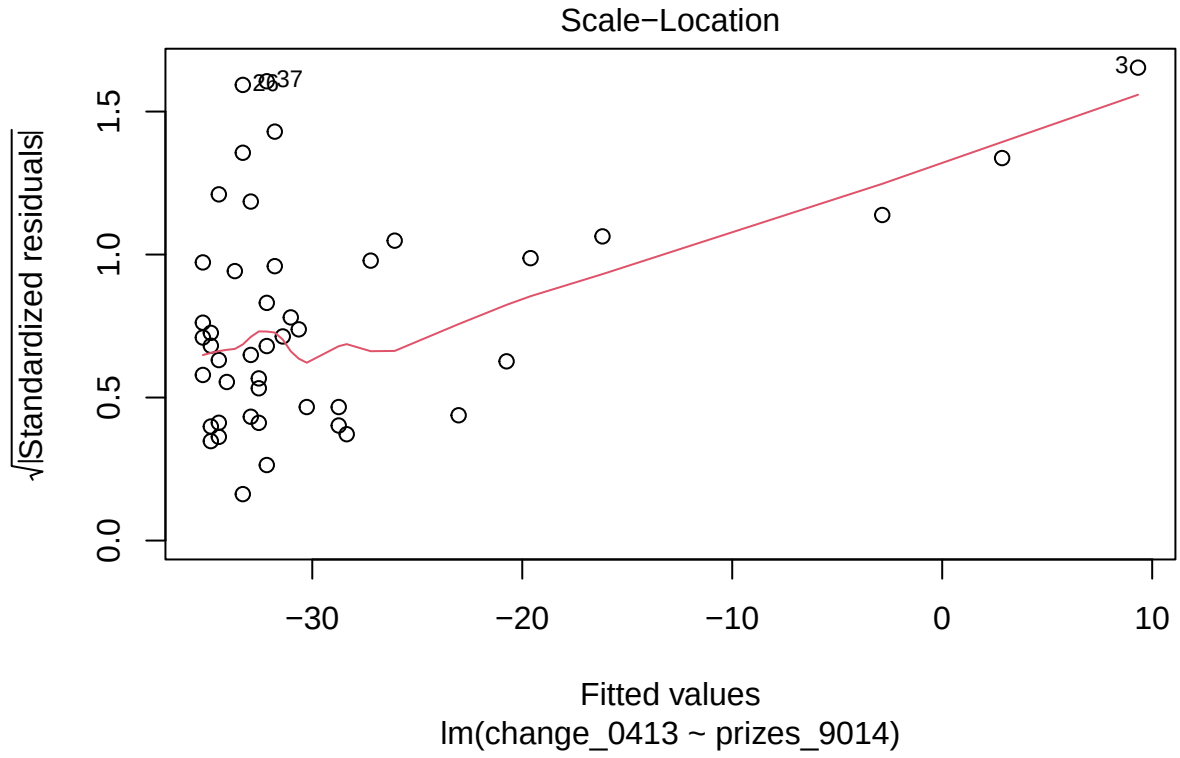
```
plot(lm_2, which = 1)
```



We can see these dots evenly spread around 0 before it is greater than -20 but after that, there is a trend downward from left to right which is influenced by a small amount of dots whose value is greater than -20. So it may not obey the assumption.

Homoscedasticity:

```
plot(lm_2, which = 3)
```

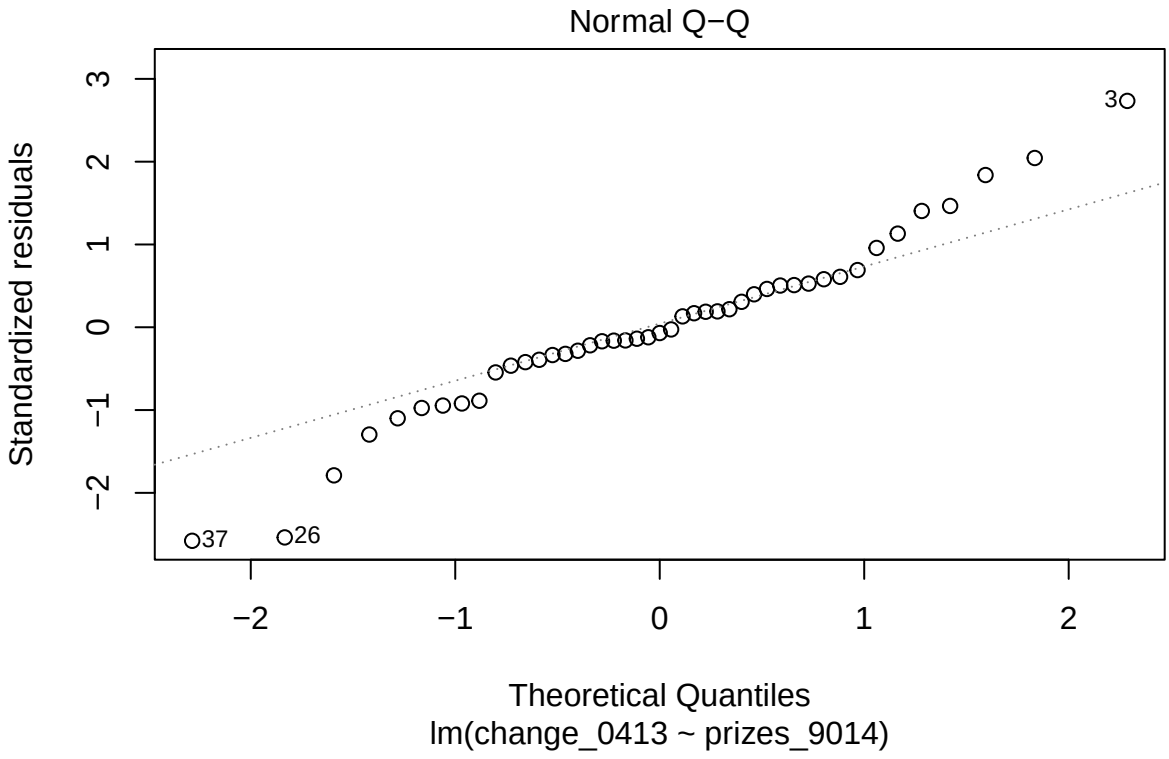


We

can see the value has a trend upward from left to right. So it disobeys the assumption.

Normality:

```
plot(lm_2, which = 2)
```



Most of the points lie quite close to the normal quantile reference line, but there are still an unignorable amount of points on the dual sides that deviate from the line.

Noise independent Independence relies on the subjects being independent of each other; if these subject are independent with each other, noise can be independent, othewise they can be not.

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?

```
prizes_next25 = tibble(prizes_9014 = c(3,25,50))
log_Cir_Avg = as.data.frame(predict(lm_1, prizes_next25))
Circ_Avg= exp(log_Cir_Avg)
Circ_now = c(453869,453869,453869)
```

```
result1 = round(data.frame(prizes_next25,log_Cir_Avg,Circ_Avg,Circ_now),3)
colnames(result1) = c("prizes_next25","log_Cir_Avg","Circ_Avg","Circ_now")
result1
```

```
##   prizes_next25 log_Cir_Avg Circ_Avg Circ_now
## 1             3      12.561 285094.6   453869
## 2            25      12.853 381899.6   453869
## 3            50      13.185 532380.7   453869
```

Using model lm_1 from 3(a), we can see only the third strategy that led to the award of 50 Pulitzer Prizes can make the expected circulation bigger than the present. Which is 532380.700, bigger than 453869.000.

(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.

```
prizes_next25 = tibble(prizes_9014 = c(3,25,50))
Circ_change = as.data.frame(predict(lm_2, prizes_next25))
Circ_now = c(453869,453869,453869)
result2 = round(data.frame(prizes_next25,Circ_change,Circ_now),3)
colnames(result2) = c("prizes_next25","Circ_change","Circ_now")
result2
```

```
##   prizes_next25 Circ_change Circ_now
## 1             3     -34.450   453869
## 2            25     -26.075   453869
## 3            50     -16.559   453869
```

In this model, we can find all the strategies will not make a positive change. Which means in the next decade, the circulation will decrease even when we invest substantially more in investigative journalism.

(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.

```
prizes_next25 = tibble(prizes_9014 = c(3,25,50))
log_prediction <- predict(lm_1, prizes_next25, interval = "confidence",level = 0.9)
result3 = round(data.frame(prizes_next25,exp(log_prediction)),3)
colnames(result3) = c("prizes_next25","fit_Circ","lwr_Circ","upr_Circ")
result3$CI = result3$upr_Circ - result3$lwr_Circ
result3
```

```
##   prizes_next25 fit_Circ lwr_Circ upr_Circ        CI
## 1             3 285094.6 245996.8 330406.3  84409.47
## 2            25 381899.6 333781.7 436954.2 103172.54
## 3            50 532380.7 431398.5 657000.9 225602.42
```

In this model, if we have 3 prizes in the next 25 years, the best-fit line gives the prediction which is 285094.600 average circulation in the next decade, we can have 90% confidence that the average circulation will be within [245996.800, 330406.300]; Similarly, for 25 prizes, best-fit prediction is 381899.600, and the confidence interval will within [333781.700, 436954.200]. For 50 prizes, the best-fit prediction is 532380.700, and the confidence interval will be within [431398.500, 657000.900]. Besides, the Confidence interval width was shown in the table, and the confidence interval is wider with the prizes_next25 increase.

** Note: all the predicted avg_circ is converted from log scale to the normal scale.**

(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.

```
prizes_next25 = tibble(prizes_9014 = c(3,25,50))
prediction <- predict(lm_2, prizes_next25, interval = "prediction",level = 0.9)
result4 = round(data.frame(prizes_next25,prediction),3)
colnames(result4) = c("prizes_next25","fit_Change","lwr_Change","upr_Change")
result4$PI = result4$upr_Change - result4$lwr_Change
result4
```

```
##   prizes_next25 fit_Change lwr_Change upr_Change     PI
## 1             3    -34.450    -79.868     10.969 90.837
## 2            25    -26.075    -71.387     19.236 90.623
## 3            50    -16.559    -62.638     29.520 92.158
```

In this model, if we have 3 prizes in the next 25 years, the best-fit line gives the prediction which is -34.450% change in the next 10 years, we can have 90% probability that the change will be within [-79.868%,10.969%], even when we decrease the investment, there is still some chance to make our circulation increase; Similarly, for 25 prizes, best-fit prediction is -26.075%, and the prediction interval will within [-71.387%,19.236%]. For 50 prizes, the best-fit prediction is -16.559%, and the prediction interval will be within [-62.638%, 29.520%]. Besides, the prediction interval are almost the same width for the three prizes_next25.

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.

For model lm_1 from 3(a)

1. It has a low explainability, because the $R^2 = 0.311$, is a small value. We assume the circulation can be explained by the prize number owned in the next 25 years, though the p-value shows they have a relationship, the relationship is not credible to trust, due to the low explainability.
2. Wide confidence interval for the large independent variable, which will make our model low trusty.

For model lm_1 from 3(b)

1. The same reason as lm_1, low credibility due to low $R^2 = 0.121$.
2. The two variables even disobey the assumptions of the linear model, **Linearity**, **Normality** and **Homoscedasticity**, so they are not suitable to use in the linear model.
3. Though the prediction interval are the same for all the three strategy, they still shows a big uncertainty. Because the wide prediction interval covers from the big negative change to the positive change. It can't give us a reliable prediction interval, a narrow interval.

Conclusion

You need to write conclusion to recap your findings to Masthead Media. This will go at the very end of a report and should include:

- A recap of the project outline including what the purpose of your project is and what problem you are solving.
- A very brief outline of what you did.
- Your final recommendation to Masthead Media.
- Any limitations that need to be addressed or future work to be done.

Your conclusion should be written as if you are presenting it to Masthead Media. They will read the conclusion to get a recap of everything outlined in the full report. It should be no more than 3/4 of a page, and contain no figures, plots or code. It should be written in plain English with no terminology.

This section will be easier to write once you have completed questions 1-5.

CONCLUSION HERE

The purpose of this project is to find that "will win more Pulitzer prizes and the future circulation", and to help the decision maker to determine which strategy should we take. To solve this problem, We built 2 models after cleaning and analyzing the data and find that winning more Pulitzer prizes has some positive influence on future circulation. Both of the two models show win more Pulitzer prizes can improve future circulation in different ways. The first model shows that only increasing the investment can help while the second model display that all the 3 strategies fail to help improve circulation. Actually, our model can't give a trustworthy recommendation, they are not reliable. As a decision maker Pooling money to win more Pulitzer prizes can to some extent help improve circulation, but it is risky. Finally, we can not determine our future direction only through a single or a few variables, instead, I also recommend having more factors to help make our model more convincing.
