Boston Sun Times - To invest(igate) or not to invest(igate)

Satyaveer Pattanaik

2023-02-20

Executive summary

This project is aimed to find out whether you (Masthead Media) should continue to invest in the Boston Sun-Times' investigative journalism in order to prevent a recent decline in readership. The alternative is to nudge the newspaper to undertake a more populist, tabloid slant.

The problem in hand is to check whether:

- publications that win more Pulitzer Prizes have a smaller, or a larger, average circulation
- publications that win more Pulitzer Prizes see a percentage increase, or decrease, in circulation, during the period that they win the prizes

I have used a publicly available dataset to build two mathematical models to predict the expected circulation in the next 25 years.

Given the modelling results, I would suggest that you should **invest substantially more in investigative journalism than present**. This would lead to a higher average circulation in 25 years.

Question One: Reading and Cleaning

(a)

```
pacman::p_load(tidyverse, caret, inspectdf)
pul <- read_csv("pulitzer.csv")
```

```
## Rows: 50 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.
```

```
knitr::kable(head(pul), digits = 3, caption = "Pulitzer.csv")
```

Table 1: Pulitzer.csv

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
New York Daily News	712671	516165	-28%	7

```
#recode change_0413 to an integer
parsed <- str_match(pul$change_0413, "(\\D)(\\d+)")

pul <- pul %>%
  mutate(change_0413 = as.integer(parsed[,1]))

knitr::kable(head(pul), digits = 3, caption = "Pulitzer.csv tidied")
```

Table 2: Pulitzer.csv tidied

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
New York Daily News	712671	516165	-28	7

(b)

```
#apend variable avg_circ which is the average of circ_2004 and circ_2013
pul <- pul %>%
  mutate(avg_circ = (circ_2004 + circ_2013)/2)
knitr::kable(head(pul), digits = 3, caption = "Appended avg_circ to
             Pulitzer.csv")
```

Table 3: Appended avg_circ to Pulitzer.csv

newspaper	circ_2004	circ_2013	change_0413	prizes_9014	avg_circ
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
New York Daily News	712671	516165	-28	7	614418.0

Question Two: Univariate Summary and Transformation

(a)

```
#distribution of avg_circ
ggplot(pul, aes(avg_circ))+
  geom_histogram(col = "black")+
  geom_vline(xintercept = mean(pul$avg_circ), col = "red", lwd = 1) +
  annotate("text", x = 1000000, y = 15,
           label = paste("Mean =", mean(pul$avg_circ)),
           col = "red",
           size = 4)+
  geom_vline(xintercept = median(pul$avg_circ), col = "blue", lwd = 1) +
  annotate("text", x = 1000000, y = 10,
           label = paste("Median =", median(pul$avg_circ)),
           col = "blue",
           size = 4) +
  theme_bw()
```

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

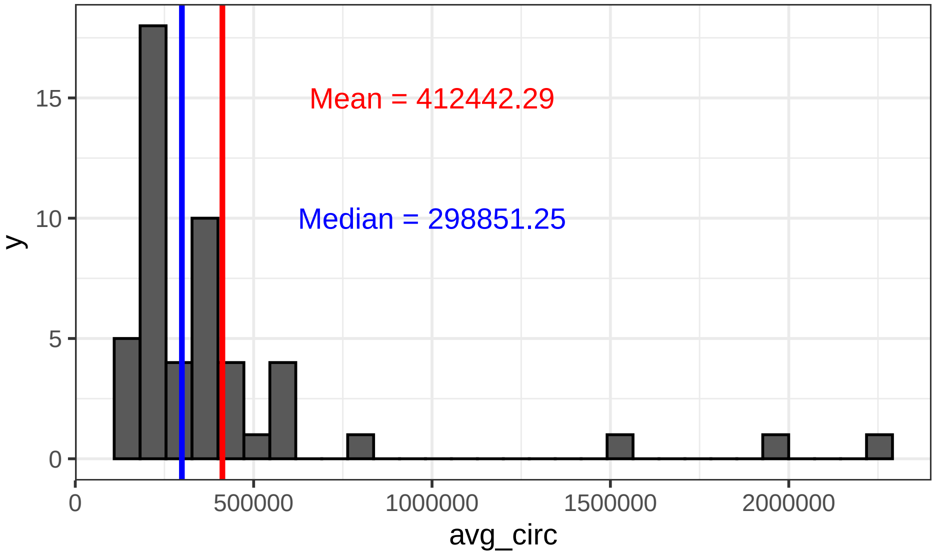


Figure 1: Histogram

```
summary(pul$avg_circ)
```

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

```
IQR(pul$avg_circ)
```

```
## [1] 222643.4
```

Shape: It is a **unimodal**, **right-skewed** distribution.

Location: It has a mean of **412442** and a median of **298851**.

Spread: It has an inter quartile range of **222643.4**.

Outliers: There are three outliers between **1500000** and **2250000**.

(b)

```
#distribution of change_0413
ggplot(pul, aes(change_0413))+
  geom_histogram(col = "black")+
  geom_vline(xintercept = mean(pul$change_0413), col = "red", lwd = 1) +
  annotate("text", x = 0, y = 6,
           label = paste("Mean =", mean(pul$change_0413)),
           col = "red",
           size = 4)+
  geom_vline(xintercept = median(pul$change_0413), col = "blue", lwd = 1) +
  annotate("text", x = 0, y = 5,
           label = paste("Median =", median(pul$change_0413)),
           col = "blue",
           size = 4) +
  theme_bw()
```

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

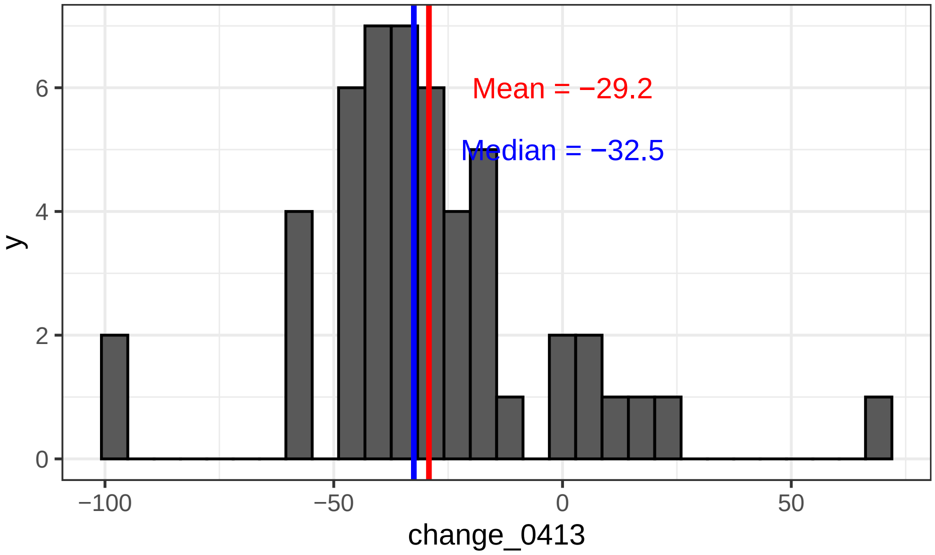


Figure 2: Histogram

```
summary(pul$change_0413)
```

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

```
IQR(pul$change_0413)
```

```
## [1] 20.75
```

Shape: It is a **unimodal**, **symmetric**(somewhat) distribution.

Location: It has a mean of **-29.2** and a median of **-32.5**.

Spread: It has an inter quartile range of **20.75**.

Outliers: There are two outliers with a percentage change of **-100%** and one outlier with a percentage change of **67%.**

(c)

```
pul %>%
  mutate(logavg_circ = log10(avg_circ)) %>%
  ggplot(aes(logavg_circ)) +
  geom_histogram(col = "black")+
  theme_bw()
```

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

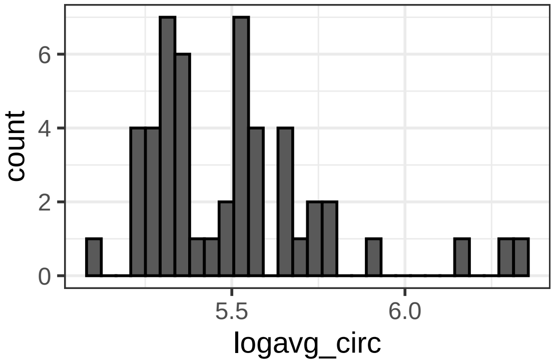


Figure 3: Histogram

The variable *avg_circ* has a right-skewed distribution which could be resolved by performing a log transform.

The variable *change_0413* has a somewhat symmetric distribution hence it does not require a log transform.

Question Three: Model building and interpretation

(a)

```
# build a model using prizes_9014 and log(avg_circ)
model_1 <- lm(log(avg_circ)~prizes_9014, data = pul)
summary(model_1)
```

```
##
## Call:
## lm(formula = log(avg_circ) ~ prizes_9014, data = pul)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.8069 -0.3147 -0.1556  0.1825  1.9693
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.463142   0.085501 145.767  < 2e-16 ***
## prizes_9014  0.014083   0.002928   4.811 1.53e-05 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.505 on 48 degrees of freedom
## Multiple R-squared:  0.3253, Adjusted R-squared:  0.3112
## F-statistic: 23.14 on 1 and 48 DF,  p-value: 1.532e-05
```

```
ggplot(pul, aes(prizes_9014, log(avg_circ)))+
  geom_point()+
  geom_smooth(method = "lm", se = F)+
  theme_bw()
```

```
## ‘geom_smooth()‘ using formula ’y ~ x’
```

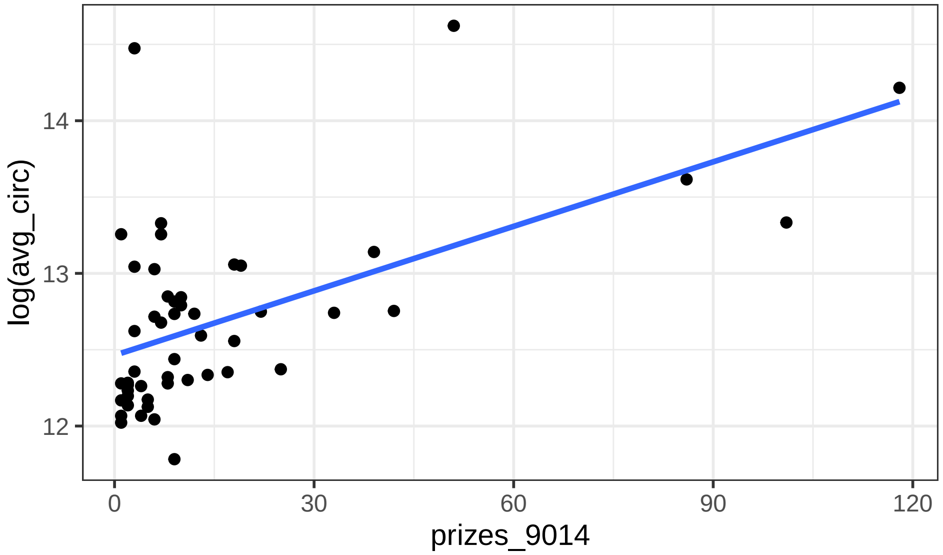


Figure 4: Scatterplot

Our model is:

$$\log(\mathit{average\ circulation}) = 12.463 + 0.014 \times \mathit{pulitzer\ prizes}$$

The slope of this model is **0.014**. This means if a newspaper gets one additional Pulitzer prize, then its average circulation will increase by 0.014083 on the log scale.

```
exp(0.014083)
```

```
## [1] 1.014183
```

The intercept of this model is **12.463**. This means if a newspaper has zero Pultizer prizes, it will have an average circulation of 12.463 on the log scale or **258627**

```
exp(12.463142)
```

```
## [1] 258627
```

Statistical significance: As the P-value for the slope is very close to zero in this model, this means that **there is a statistically significant relationship** between the number of Pulitzer prizes and the average circulation.

(b)

```
# build a model using prizes_9014 and change_0413
model_2 <- lm(change_0413~prizes_9014, data = pul)
summary(model_2)
```

```
##
## Call:
## lm(formula = change_0413 ~ prizes_9014, data = pul)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -68.068 -10.251  -2.713  13.126  56.749
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -35.4152     4.3336  -8.172 1.21e-10 ***
## prizes_9014   0.3870     0.1484   2.608   0.0121 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 25.59 on 48 degrees of freedom
## Multiple R-squared:  0.1241, Adjusted R-squared:  0.1059
## F-statistic: 6.802 on 1 and 48 DF,  p-value: 0.0121
```

```
ggplot(pul, aes(prizes_9014, change_0413))+
  geom_point()+
  geom_smooth(method = "lm", se = F)+
  theme_bw()
```

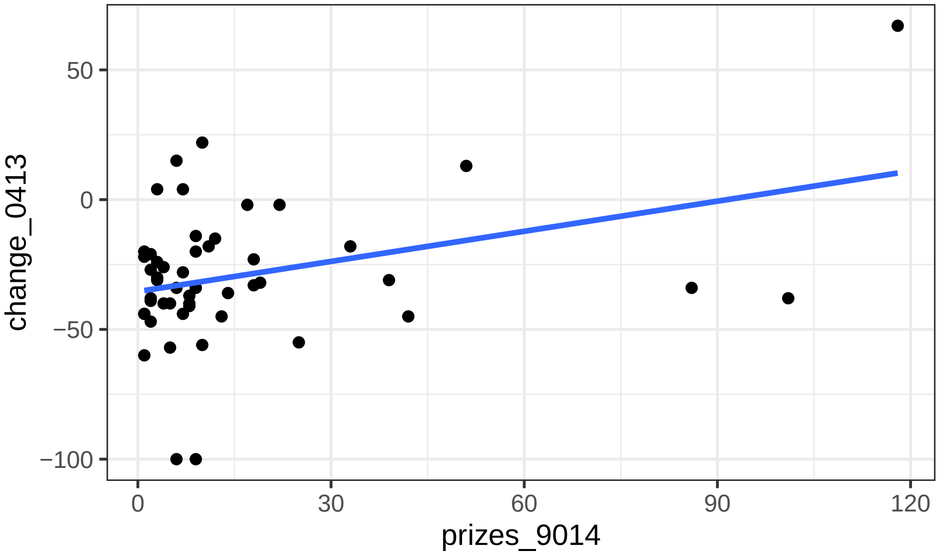


Figure 5: Scatterplot

```
## 'geom_smooth()' using formula 'y ~ x'
```

Our model is:

$$\textit{percentage change in circulation} = -35.415 + 0.387 \times \textit{pulitzer prizes}$$

The slope of this model is **0.387**. This means if a newspaper gets one additional Pulitzer prize, then on average its percentage change in circulation will increase by **0.387%.**

The intercept of this model is **-35.415**. This means if a newspaper has zero Pultizer prizes, we can expect a percentage change in circulation to be **-35.415%.**

Statistical significance: As the P-value for the slope in model_2 is less than 0.05, this means that **there is a statistically significant relationship** between the number of Pulitzer prizes and the percentage change in circulation.

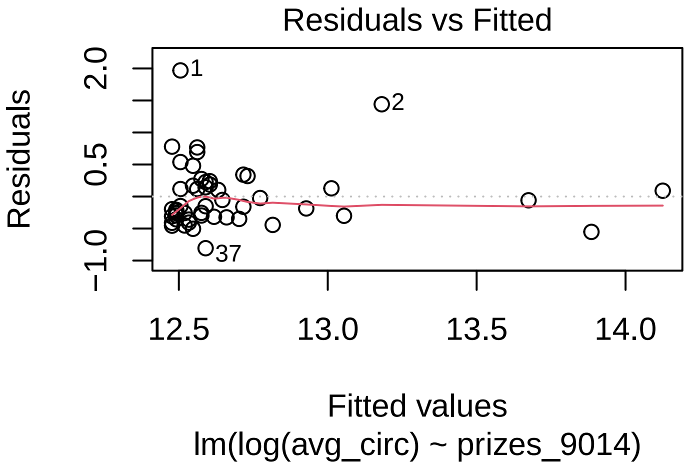
(c)

Checking assumptions:

1. Linearity -

model_1:

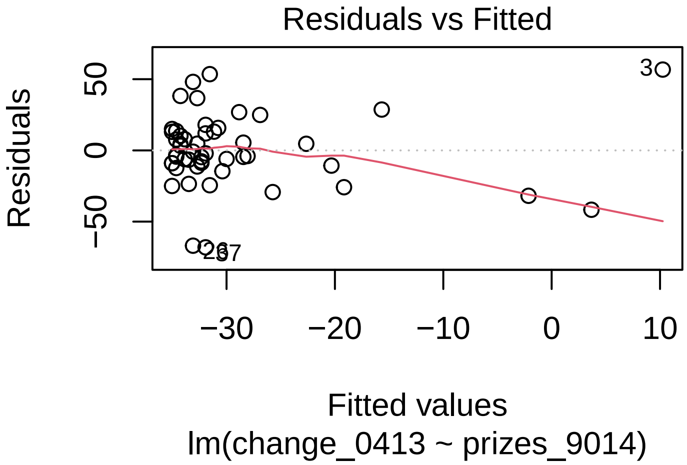
```
plot(model_1, which = 1)
```



There are no trends in the residual versus fitted plot. Hence, model_1 satisfies the linearity assumption.

model_2:

```
plot(model_2, which = 1)
```

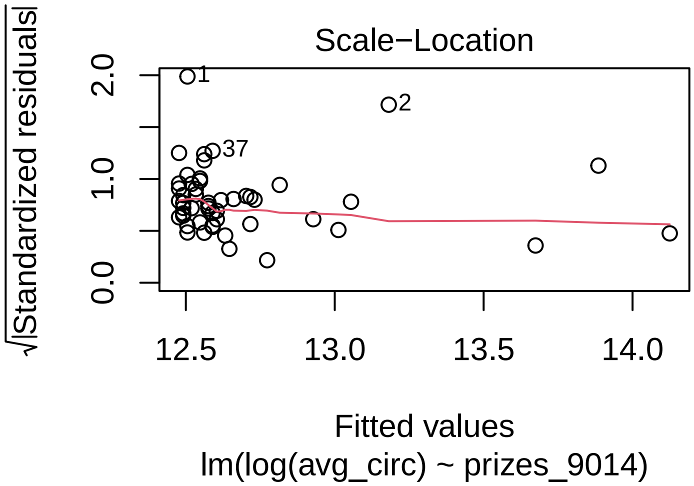


Here, if we look at the points, we observe no strong relationship, but there aren't many points after -20 and that's affecting the red reference line. Hence, model_2 too satisfies the linearity assumption.

2. Homoscedasticity -

model_1:

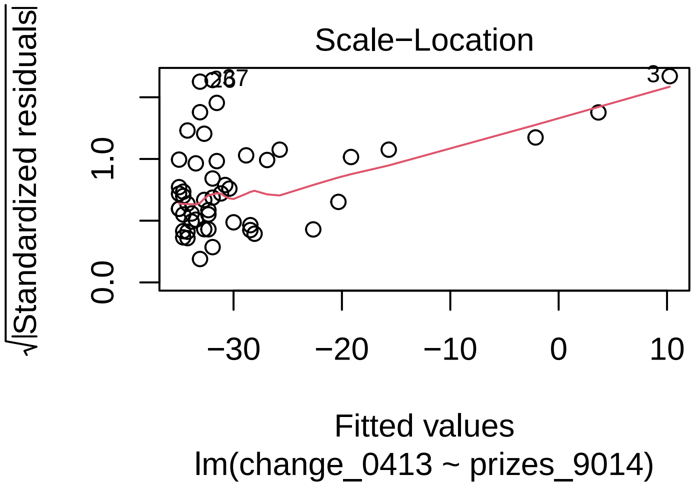
```
plot(model_1, which = 3)
```



There are no trends in the standardised residual versus fitted plot. Hence, model_1 is evenly spread, satisfying the homoscedasticity assumption.

model_2:

```
plot(model_2, which = 3)
```

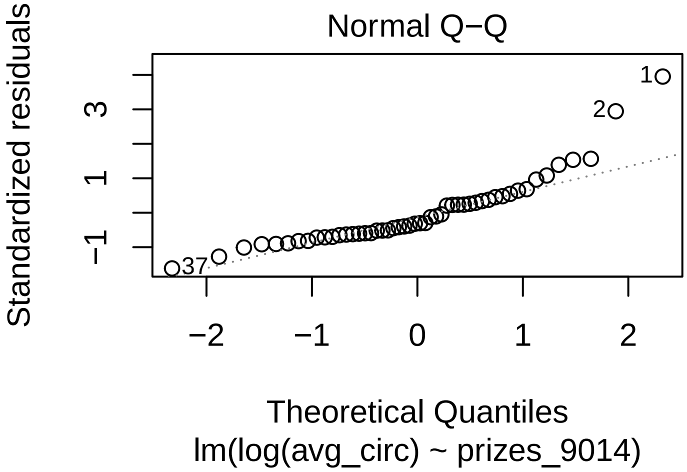


There is a bend in the red line due to the effect of outliers but there is no noticable trend in the spread of the data. Hence, model_2 too satisfies the homoscedasticity assumption.

3. Normality -

model_1:

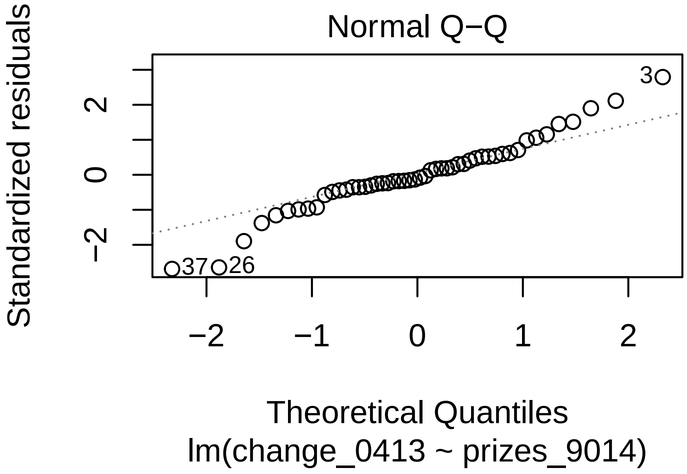
```
plot(model_1, which = 2)
```



The points between -1 and 1 (which is most of the data) lie along the dotted line. The residuals are mostly normally distributed, with extreme values at both ends.

model_2:

```
plot(model_2, which = 2)
```



The points between -1 and 1 (which is most of the data) lie along the dotted line. The residuals are mostly normally distributed, with extreme values at both ends.

4. Independence -

model_1:

Independence relies the subjects are independent of one another. This is not necessarily true here, since, for example, if a lot of subscribers of one newspaper change their subscription to another newspaper, the average circulations of the two newspapers will be related.

model_2:

Similarly, model_2 may not necessarily satisfy the independence assumption here.

Question Four: Prediction

These strategic directions were proposed:

- **Strategy 1** - 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.
- **Strategy 2** - Investing the same amount in investigative journalism than present, leading to the award of 25 Pulitzer Prizes in the next 25 years.
- **Strategy 3** - Investing substantially more in investigative journalism, leading to the award of 50 Pulitzer Prizes.

(a)

```
dir_1 <- tibble(prizes_9014 = 3)
dir_2 <- tibble(prizes_9014 = 25)
dir_3 <- tibble(prizes_9014 = 50)
circ_1 <- exp(predict(model_1, dir_1))
circ_2 <- exp(predict(model_1, dir_2))
circ_3 <- exp(predict(model_1, dir_3))

#table of expected circulations
tab <- matrix(nrow = 3, ncol=2, byrow = T)
colnames(tab) <- c("Pulitzer Prizes", "Expected Circulation")
rownames(tab) <- c('Strategy 1', 'Strategy 2', 'Strategy 3')
tab[1,1] <- 3
tab[2,1] <- 25
tab[3,1] <- 50
tab[1,2] <- circ_1
tab[2,2] <- circ_2
tab[3,2] <- circ_3
tab <- as.table(tab)
knitr::kable(tab, digits = 3, caption = "Expected Circulation with
             each Strategy")
```

Table 4: Expected Circulation with each Strategy

	Pulitzer Prizes	Expected Circulation
Strategy 1	3	269788.1
Strategy 2	25	367773.8
Strategy 3	50	522983.2

The Boston Sun-Times currently has a circulation of **453,869.**

When comparing the current circulation with the predicted future circulations, we see that only **Strategy 3 will yield a higher average circulation (522,983.1)**.

(b)

```
dir_1 <- tibble(prizes_9014 = 3/25*10)
dir_2 <- tibble(prizes_9014 = 25/25*10)
dir_3 <- tibble(prizes_9014 = 50/25*10)
circ_1 <- predict(model_2, dir_1)
circ_2 <- predict(model_2, dir_2)
circ_3 <- predict(model_2, dir_3)

#table of % change expected in circulations
tab_2 <- matrix(nrow = 3, ncol=2, byrow = T)
colnames(tab_2) <- c("Pulitzer Prizes", "% Change Expected in Circulation")
rownames(tab_2) <- c('Strategy 1', 'Strategy 2', 'Strategy 3')
tab_2[1,1] <- 3
tab_2[2,1] <- 25
tab_2[3,1] <- 50
```

```r
tab_2[1,2] <- round(circ_1, 3)
tab_2[2,2] <- round(circ_2, 3)
tab_2[3,2] <- round(circ_3, 3)
tab_2 <- as.table(tab_2)
knitr::kable(tab_2, digits = 3, caption = "% Change Expected in Circulation over
             next decade with each Strategy")
```

Table 5: % Change Expected in Circulation over next decade with each Strategy

	Pulitzer Prizes	% Change Expected in Circulation
Strategy 1	3	-34.951
Strategy 2	25	-31.545
Strategy 3	50	-27.675

[**NOTE: These projections are over the next decade (10 years), as asked in 4(b)**]

We observe that from **model_1**, an increase in the number of Pulitzer prizes leads to an **increase in the average circulation of a newspaper**.

In contrast, with **model_2**, there is a **percentage decrease in circulation** regardless of the strategy. This is **inconsistent** with the previous model.

But we also observe that as we keep increasing the number of Pulitzer prizes, there seems to be a **less negative percentage change expected in the circulation of a newspaper**.

(c)

```r
dir_1 <- tibble(prizes_9014 = 3)
dir_2 <- tibble(prizes_9014 = 25)
dir_3 <- tibble(prizes_9014 = 50)
circ_1 <- exp(predict(model_1, dir_1, interval = "confidence", level = 0.90))
circ_2 <- exp(predict(model_1, dir_2, interval = "confidence", level = 0.90))
circ_3 <- exp(predict(model_1, dir_3, interval = "confidence", level = 0.90))
tab_3 <- matrix(nrow = 3, ncol = 2, byrow = T)
colnames(tab_3) <- c("Expected Circulation (lower)",
                     "Expected Circulation (upper)")
rownames(tab_3) <- c('Strategy 1', 'Strategy 2', 'Strategy 3')

tab_3[1,1] <- circ_1[2]
tab_3[2,1] <- circ_2[2]
tab_3[3,1] <- circ_3[2]
tab_3[1,2] <- circ_1[3]
tab_3[2,2] <- circ_2[3]
tab_3[3,2] <- circ_3[3]
tab_3 <- as.table(tab_3)
knitr::kable(tab_3, digits = 3, caption = "90% Confidence Intervals with each
             Strategy")
```

Table 6: 90% Confidence Intervals with each Strategy

	Expected Circulation (lower)	Expected Circulation (upper)
Strategy 1	235515.2	309048.6
Strategy 2	323727.6	417813.0
Strategy 3	425949.0	642122.4

- The estimated average circulation increases progressively with each strategy (i.e. from Strategy 1 to Strategy 3).
- We can also observe that the width of the intervals itself keeps increasing progressively with each strategy.
- We can observe that the confidence interval of each strategy never overlaps with one another. This should give us convincing evidence of that winning more Pulitzer awards would guarantee a higher circulation in the future.

(d)

```
circ_1 <- predict(model_2, dir_1, interval = "prediction", level = 0.90)
circ_2 <- predict(model_2, dir_2, interval = "prediction", level = 0.90)
circ_3 <- predict(model_2, dir_3, interval = "prediction", level = 0.90)

#table of % change expected in circulations
tab_4 <- matrix(nrow = 3, ncol=2, byrow = T)
colnames(tab_4) <- c("% Change Expected in Circulation (lower)",
                     "% Change Expected in Circulation (upper)")
rownames(tab_4) <- c('Strategy 1', 'Strategy 2', 'Strategy 3')
tab_4[1,1] <- circ_1[2]
tab_4[2,1] <- circ_2[2]
tab_4[3,1] <- circ_3[2]
tab_4[1,2] <- circ_1[3]
tab_4[2,2] <- circ_2[3]
tab_4[3,2] <- circ_3[3]
tab_4 <- as.table(tab_4)
knitr::kable(tab_4, digits = 3, caption = "90% Prediction Intervals with each
             Strategy")
```

Table 7: 90% Prediction Intervals with each Strategy

	% Change Expected in Circulation (lower)	% Change Expected in Circulation (upper)
Strategy 1	-77.730	9.221
Strategy 2	-69.151	17.671
Strategy 3	-60.234	28.104

- Here, the estimated percentage change in circulation keeps increasing (gets less negative) with each strategy.
- We observe that the prediction interval of each strategy overlaps with one another. This observation gives a weak evidence that a higher number of Pulitzer awards may guarantee a more positive percentage change in circulation.
- We can also observe that the width of the intervals itself remains almost the same with each strategy.

Question Five: Limitations

(a)

model_1:

- With respect to the dataset, the Pulitzer prizes are counted between 1990 and 2014, but the circulation of newapapers is only first collected in 2004 and then in 2013. There exists a data gap in the period 1990 - 2003, and we don't know if most of the awards were won in these 14 years or the remaining period. This gap in data would significantly impact the model's accuracy.
- This model does not necessarily satisfy the independence assumption.
- The model relies heavily on past data. This data may not be relevant in 2022.

model_2:

- With respect to the dataset, the Pulitzer prizes are counted between 1990 and 2014, but we only have the percentage change in circulation of newspapers between 2004-2013. There exists a data gap in the period 1990 - 2003, and we don't know if most of the awards were won in these 14 years or the remaining period. This gap in data would impact the model's accuracy.
- This model too does not necessarily satisfy the independence assumption.
- The predictor variable **change_0413** does not provide us with much information as it calculates the percentage change in circulation between 2004 and 2013.It would have been more useful if it was calculated on a yearly basis.
- The predictor variable **change_0413** contains mostly negative values, which may be due to the industry wide trend of a negative growth in circulation from 2004 to 2013. Also there are outliers on both extremes, which significantly impact the accuracy of this model.

Conclusion

This project was aimed to find out whether you (Masthead Media) should continue to invest in the Boston Sun-Times' investigative journalism in order to prevent a recent decline in readership. The alternative is to nudge the newspaper to undertake a more populist, tabloid slant. I had to check whether:

- publications that win more Pulitzer Prizes have a smaller, or a larger, average circulation, and
- publications that win more Pulitzer Prizes see a percentage increase, or decrease, in circulation, during the period that they win the prizes.

I have used a publicly available dataset to build two mathematical models to predict the expected circulation in the next 25 years.

From the models, I concluded that the **more Pulitzer Prizes you win**, the **greater your average circulation**.

Given the modelling results, I would suggest that you should **invest substantially more in investigative journalism than present**. This would lead to a higher average circulation in 25 years.

However, I would also like to add that there is high chance of a percentage decrease in circulation after 25 years. This may be due to a quirk in the dataset used for this project.