Assignment 2 – Scaffolded Case Study

Kristen Osborne

## EMAIL

Good afternoon

This email (and attached documentation) details the work the data science team has undertaken following the request from the Masthead Media executive team to use data modeling to address key questions related to strategic planning.

Two models were created: one which predicted the average circulation based on the number of Pulitzer Prizes, and one which predicted the change in circulation based on the same factor. These were used to answer the key research questions, detailed below.

**Background**

**Research question 1:** *Whether publications that win more Pulitzer Prizes have a smaller, or a larger, average circulation.*

The model supports this assumption. The model has a positive slope, meaning that as the number of Pulitzer Prizes increased, the average circulation of the newspaper in question increased. See figure 4 in attached document for more details.

**Research question 2:** *Whether publications that win more Pulitzer Prizes see a percentage increase, or decrease, in circulation, during the period that they win the prizes.*

The model determined that, although the winning of more prizes seemed to result in an improved change in circulation, almost all newspapers are expected to see a fall in circulation. This means a negative prediction of the change in circulation. See figure 5 in attached document for more details

**Predictions**

**Research question 3:** *If these relationships exist, how might the trajectory of the Boston Sun-Times's circulation change depending on the newspapers strategic direction?*

Three strategic directions were evaluation using each model. Analysis is based on a current circulation of 453,869:

1. **Investing substantially less in investigative journalism, winning a predicted 3 Pulitzer Prizes.**
   1. Both models predicted that this would result in a decrease in circulation. The predicted average circulation was 269,789, and the predicted change in circulation was -34%.
2. **Investing the same amount in investigative journalism than present, winning a predicted 25 Pulitzer Prizes.**
   1. Both models predicted that this scenario would result in a decrease in circulation. The predicted average circulation was 367,775, and the predicted change in circulation was -26%
3. **Investing substantially more in investigative journalism, winning a predicted Pulitzer Prizes.**
   1. The model using average circulation predicted an increase in circulation under this scenario, rising to 522,984. The model using the percentage change in circulation predicted a fall of 16%

Each of these predictions is subject to a 90% confidence interval – a range in which the result is predicted to fall 90% of the time. Table 2 and 3 in the attached document these ranges, and they are summarized below:

|  |  |  |  |
| --- | --- | --- | --- |
| **Scenario** | **Predicted value** | **Lower confidence interval** | **Upper confidence interval** |
| Model 1 – Average circulation predicted by Pulitzer Prizes | | | |
| **Same level of investment** | 367,775 | 323,729 | 417,814 |
| **Reduced investment** | 269,789 | 235,516 | 309,050 |
| **Increased investment** | 522,984 | 425,950 | 642,124 |
| Model 2 – Change in circulation predicted by Pulitzer Prizes | | | |
| **Same level of investment** | -26 | -32 | -19 |
| **Reduced investment** | -34.25 | -41.14 | -27 |
| **Increased investment** | -16 | -26 | -6 |

**Assumptions**

There are four assumptions core to the linear models we have used. These are

* linearity (that the relationship between the variables is linear and well represented by a linear model),
* constant variance (‘noise’, or errors in the data are the same for different values),
* normality ('noise', or errors in the data are normally distributed), and
* independence (that the observations of data the models are based on are independent).

The following table summarises where the assumptions were and were not supported. For more details, see figures 6-11 in attached document.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Linearity** | **Constant variance** | **Normality** | **Independence** |
| Model 1 – Average circulation predicted by Pulitzer Prizes | Yes | Yes | Unclear | Unclear |
| Model 2 – Change in circulation predicted by Pulitzer Prizes | No | No | Unclear | Unclear |

**Limitations**

It is important to note the limitations of these models (and indeed of any similar models). Importantly, it is possible that a third factor is responsible for both any change in circulation and the winning of Pulitzer Prizes. For example, the funding available to a newspaper will determine its ability to advertise itself and to hire quality journalists. In this way, both the circulation and the number of prizes might both be determined by the funding available.

Another issue is that of unknowns. The model is only able to predict circulation results in the context of the factors available in the data. There is always the possibility that unknown factors or events are key to a newspaper's circulation. These might be local or national news events, major investigative stories, or even the general state of the economy. A more robust model could explore additional factors and attempt to control for these.

For more details on this or any element of the analysis, see attached document.

## ATTACHMENT

## Question One: Reading and Cleaning

1.1

#create a pattern the change is represented in  
  
change\_pattern <- "(-?\\d+)%"  
  
#use str\_match to pull the change values as int, merge that with the original data, drop unneeded variables and rename the new variable to change\_0413 with the type int  
  
puli\_data <- as\_tibble(str\_match(raw\_data$change\_0413,change\_pattern),.name\_repair = "universal") %>%  
 merge(raw\_data, by = 0) %>%  
 select(newspaper, circ\_2004, circ\_2013, ...2, prizes\_9014, -...1, -change\_0413, -Row.names) %>%  
 rename(change\_0413 = ...2) %>%  
 mutate(change\_0413 = as.integer(change\_0413))

## New names:  
## \* `` -> ...1  
## \* `` -> ...2

#print the head of the table to check  
  
pander(head(puli\_data),caption = "Table 1. Sample of data")

Table 1. Sample of data

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| newspaper | circ\_2004 | circ\_2013 | change\_0413 | prizes\_9014 |
| USA Today | 2,192,098 | 1,674,306 | -24 | 3 |
| Newsday | 553,117 | 377,744 | -32 | 19 |
| Houston Chronicle | 549,300 | 360,251 | -34 | 6 |
| Dallas Morning News | 528,379 | 409,265 | -23 | 18 |
| San Francisco Chronicle | 499,008 | 218,987 | -56 | 10 |
| Arizona Republic | 466,926 | 293,640 | -37 | 8 |

1.2

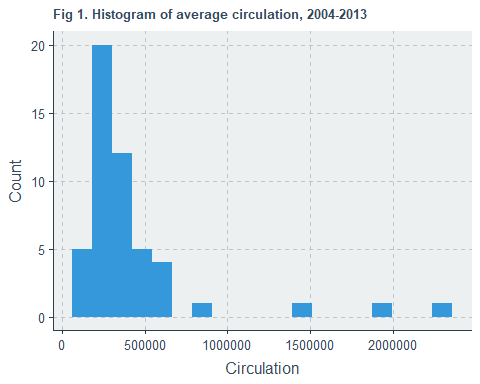
#use mutate to adda new variable that's the average of the circulation in 2004 and 2013  
  
puli\_data <- puli\_data %>%  
 mutate(circ\_av = (circ\_2004+circ\_2013)/2)

## Question Two: Univariate Summary and Transformation

2.1

The average circulation between 2004-13 is shown in figure 1. There was a medium right (positive) skew in circulation, with a median of 298851 and a mean of 412442. The average circulation has medium variance, with a standard deviation of 410340. There were 4 outliers - the Wall Street Journal, USA Today, New York Times, Los Angeles Times.

#generate a histogram to check distribution  
  
#to choose a bin width, one easy way is to use the Freedman-Diaconis rule (which is calculated based on the interquartile range and the number of observations), resulting in a bin width of around 120,869  
  
bw <- 2 \* IQR(na.exclude(puli\_data$circ\_av)) / length(na.exclude(puli\_data$circ\_av))^(1/3)  
  
#plot the histogram, with a title and appropriate x lable  
  
fig\_1 <- ggplot(puli\_data,aes(x = circ\_av)) +  
 geom\_histogram(binwidth = bw) +  
 labs(x = 'Circulation', y = 'Count', title = 'Fig 1. Histogram of average circulation, 2004-2013')+  
 theme(plot.title = element\_text(size=10))  
  
fig\_1

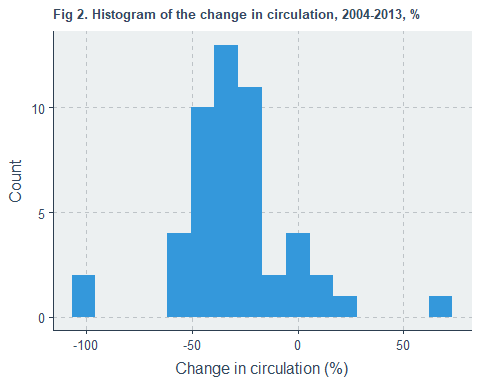


2.2

A histogram of the change in circulation between 2004 and 2013 is shown in figure 2. There was a right (positive) skew in the distribution of the change, but the median change was negative (-32). The mean change in circulation was a fall of 29%. The change in circulation had a standard deviation of 27.

There were two extremely negative outliers - the Rocky Mountain News and the New Orleans Times-Picayune, both of whom had a circulation of 0 in 2013 and therefore a change of -100%. There was one positive outlier - the New York Time, which experienced an increase in circulation of 67% between 2004 and 2013.

#generate a histogram to check distribution  
  
#choose bin width, using the Freedman-Diaconis rule, resulting in a bin width of around 11  
  
bw <- 2 \* IQR(na.exclude(puli\_data$change\_0413)) / length(na.exclude(puli\_data$change\_0413))^(1/3)  
  
#plot the histogram, with a title and appropriate x lable  
  
fig\_2 <- ggplot(puli\_data,aes(x = change\_0413)) +  
 geom\_histogram(binwidth = bw) +  
 labs(x = 'Change in circulation (%)', y = 'Count', title = 'Fig 2. Histogram of the change in circulation, 2004-2013, %')+  
 theme(plot.title = element\_text(size=10))  
  
fig\_2



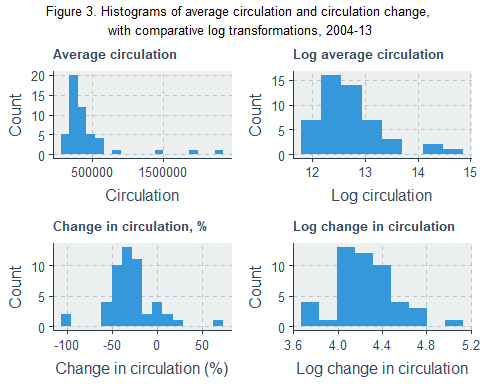
2.3

The average circulation represents a good candidate for a log transform to adjust skew. Circulation numbers are inherently positive (you cannot have a negative circulation). This often creates a rightward (positive) skew, which is present (figure 1). Additionally, the predicted values from any future model of the log-transformed regression will be positive. The circulation has a number of high-circulation outliers, which can be dealt with though a log transformation as well. On the other hand, the change in circulation has many valid negative values, and a less skewed distribution. It also has fewer outliers.

One way to investigate if the log function should be used is to visualise the change in distributions. Figure 3 shows a contrast between the histograms of average circulation and change in circulation, and their log transformation. From this we can see that the log of average circulation has a more normal distribution and less variance. Note that as the circulation contains 0s, 1 has been added to all values, and as the lowest circulation changes were -100%, 100 has been added to all these values. Figure 3 confirms the supposition that a log transformation is beneficial for average circulation, but does not meaningfully improve the distributions of the change in circulation.

puli\_data\_log <- puli\_data %>%   
 mutate(log\_circ\_av = log(circ\_av + 1),log\_change\_0413 = log(change\_0413+100))   
  
  
#generate a histogram to check distribution  
  
#choose a bin width, using the Freedman-Diaconis rule,   
bw1 <- 2 \* IQR(na.exclude(puli\_data\_log$circ\_av)) / length(na.exclude(puli\_data\_log$circ\_av))^(1/3)  
bw2 <- 2 \* IQR(na.exclude(puli\_data\_log$change\_0413)) / length(na.exclude(puli\_data\_log$change\_0413))^(1/3)  
bw3 <- 2 \* IQR(na.exclude(puli\_data\_log$log\_circ\_av)) / length(na.exclude(puli\_data\_log$log\_circ\_av))^(1/3)  
bw4 <- 2 \* IQR(na.exclude(puli\_data\_log$log\_change\_0413)) / length(na.exclude(puli\_data\_log$log\_change\_0413))^(1/3)  
  
#plot the histogram, with a title and appropriate x lable  
  
hist\_1 <- ggplot(puli\_data,aes(x = circ\_av)) +  
 geom\_histogram(binwidth = bw1) +  
 labs(x = 'Circulation', y = 'Count', title = 'Average circulation') +  
 theme(plot.title = element\_text(size=10)) +  
 scale\_x\_continuous(breaks = c(500000,1500000))  
  
hist\_2 <- ggplot(puli\_data,aes(x = change\_0413)) +  
 geom\_histogram(binwidth = bw2) +  
 labs(x = 'Change in circulation (%)', y = 'Count', title = 'Change in circulation, %') +  
 theme(plot.title = element\_text(size=10))  
  
hist\_3 <- ggplot(puli\_data\_log,aes(x = log\_circ\_av)) +  
 geom\_histogram(binwidth = bw3) +  
 labs(x = 'Log circulation', y = 'Count', title = 'Log average circulation') +  
 theme(plot.title = element\_text(size=10))  
  
hist\_4 <- ggplot(puli\_data\_log,aes(x = log\_change\_0413)) +  
 geom\_histogram(binwidth = bw4) +  
 labs(x = 'Log change in circulation', y = 'Count', title = 'Log change in circulation') +  
 theme(plot.title = element\_text(size=10))  
  
#grid presents the differences pre and post log transformation for each variable  
  
fig\_3 <- grid.arrange(hist\_1, hist\_3, hist\_2, hist\_4,nrow=2,top = text\_grob("Figure 3. Histograms of average circulation and circulation change,   
with comparative log transformations, 2004-13",size = 10))

## Warning: Removed 2 rows containing non-finite values (stat\_bin).



fig\_3

## TableGrob (3 x 2) "arrange": 5 grobs  
## z cells name grob  
## 1 1 (2-2,1-1) arrange gtable[layout]  
## 2 2 (2-2,2-2) arrange gtable[layout]  
## 3 3 (3-3,1-1) arrange gtable[layout]  
## 4 4 (3-3,2-2) arrange gtable[layout]  
## 5 5 (1-1,1-2) arrange text[GRID.text.238]

## Question Three: Model building and interpretation

3.1

The linear regression model predicting circulation (with a log transform for circulation, figure 4) based on pulitzer prizes was significant (p <0.001). The slope of this model was 0.014, which means that the model predicts that winning an additional prize results in an average increase in circulation of 0.014 on the log scale. More usefully, this would mean that winning a pulitzer prize increases your circulation on average by 1.418%.

The intercept of this model was 12.463. This means when a paper has no pulitzer prizes, the expected circulation in log terms is 12.463, or in the original units, a circulation of 258628.3.

#define the model - predicting the log average circulation by pulitzer prizes  
  
lm\_circ\_prize <- lm(log\_circ\_av ~ prizes\_9014, data = puli\_data\_log)  
  
#get summary of the data  
  
summary(lm\_circ\_prize)

##   
## Call:  
## lm(formula = log\_circ\_av ~ prizes\_9014, data = puli\_data\_log)  
##   
## 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) 12.463147 0.085500 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

#derive the non-log % increase per prize  
  
(exp(coef(lm\_circ\_prize)["prizes\_9014"]) - 1) \* 100

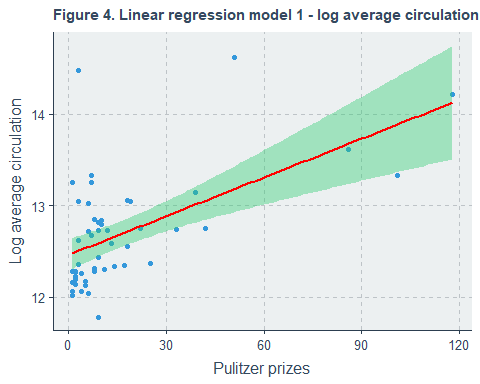
## prizes\_9014   
## 1.418284

#derive the intercept in non-log terms  
  
exp(12.463147)

## [1] 258628.3

#plot the model  
  
ggplot(puli\_data\_log, aes(x = prizes\_9014, y = log\_circ\_av)) +   
 geom\_point() +  
 stat\_smooth(method = "lm", col = "red") +  
 labs(x = 'Pulitzer prizes', y = 'Log average circulation', title = 'Figure 4. Linear regression model 1 - log average circulation predicted by pulitzer prizes')+  
 theme(plot.title = element\_text(size=11))

## `geom\_smooth()` using formula 'y ~ x'



3.2

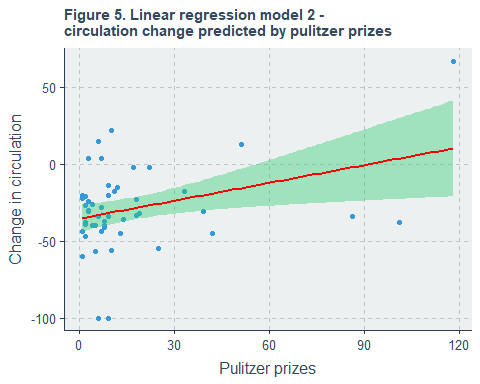
The linear model for predicting the change in circulation based on the number of pulitzer prizes a newspaper has won is presented in figure 5. This model shows a significant relationship between the number of Pulitzer Prizes and a change in circulation (P<0.01) and has a slope of 0.387. This means for every additional pulizter prize a newspaper wins, the model predicts an average increase in the change in circulation of 0.387 percentage points. For example, if a newspaper has a rate of change in circulation of 1% and it wins an additional prize, the model predicts the new rate of change in circulation will be 1.387%. The intercept of this model was -35.415, meaning that the predicted rate of change in circulation for a newspaper with no pulizter prizes is -35.415%.

#define the model - predicting the log average circulation by pulitzer prizes  
  
lm\_change\_prize <- lm(change\_0413 ~ prizes\_9014, data = puli\_data\_log)  
  
#get summary of the data  
  
summary(lm\_change\_prize)

##   
## Call:  
## lm(formula = change\_0413 ~ prizes\_9014, data = puli\_data\_log)  
##   
## 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

#plot the model  
  
ggplot(puli\_data\_log, aes(x = prizes\_9014, y = change\_0413)) +   
 geom\_point() +  
 stat\_smooth(method = "lm", col = "red") +  
 labs(x = 'Pulitzer prizes', y = 'Change in circulation', title = 'Figure 5. Linear regression model 2 -  
circulation change predicted by pulitzer prizes')+  
 theme(plot.title = element\_text(size=11))

## `geom\_smooth()` using formula 'y ~ x'

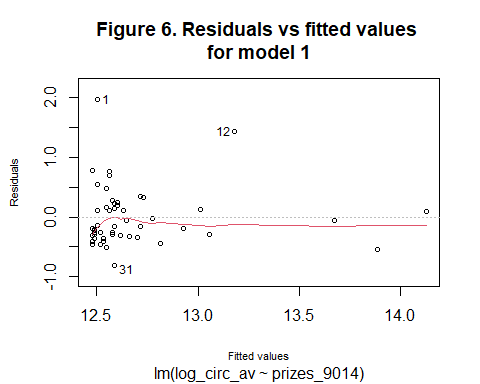


3.3

There are four assumptions core to the linear models we have used. These are linearity, constant variance (homoscedascity), normality, and independence.

This first assumption is linearity - that the relationship between the predicted variable and the predictor variable is linear and therefore well represented by a linear model. To assess linearity, the residual (the observed value of the predicted variable minus the predicted value of the variable) was plotted again the fitted values (the predicted values of the variable). This is shown in figure 6 for the average circulation predicting model (model 1) and figure 7 for the model predicting the change in circulation (model 2).

#Plot residual vs fitted  
  
#log average circulation predicated by prizes  
par(cex.lab=0.7)  
plot(lm\_circ\_prize, which = 1,cex=0.7, title("Figure 6. Residuals vs fitted values   
for model 1"))



#change in circulation predicated by prizes  
par(cex.lab=0.7)  
plot(lm\_change\_prize, which = 1, title("Figure 7. Residuals vs fitted values  
for model 2"))

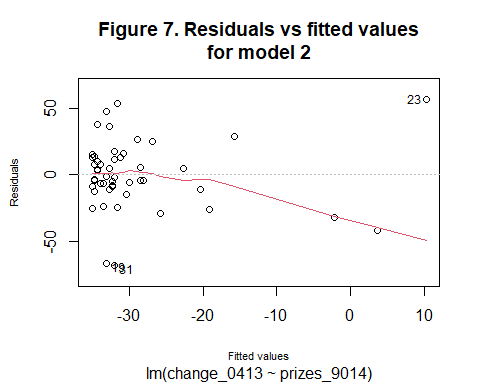
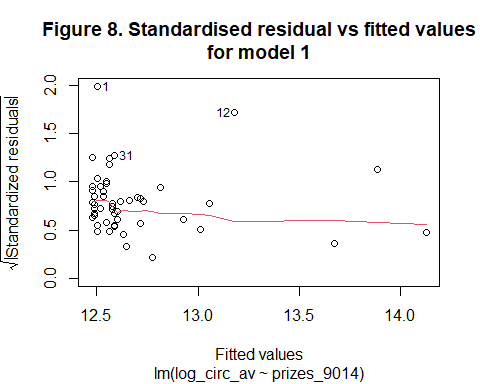


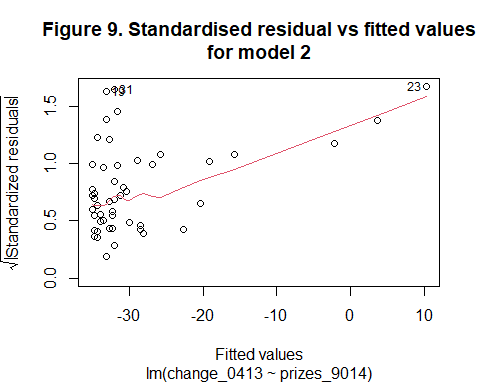
Figure 6 suggests that the relationship between the log of average circulation and number of pulitzer prizes is linear, and therefore well represented by a linear model. Alternatively, figure 7 shows a non-linear relationship - residuals are lower for higher fitted values. This means model 2 does not well represent the relationship between the change in circulation and the number of pulitzer prizes.

The next assumption is that of constant variance (homoscedascity). This is the assumption that the errors (or ‘noise’) present are uniform for different fitted values - there is not more error for certain types of value. This is plotted in figure 8 for model 1 and figure 9 for model 2.

#plot square root of standardised residual  
#log average circulation predicated by prizes  
plot(lm\_circ\_prize, which = 3, title("Figure 8. Standardised residual vs fitted values  
for model 1"))



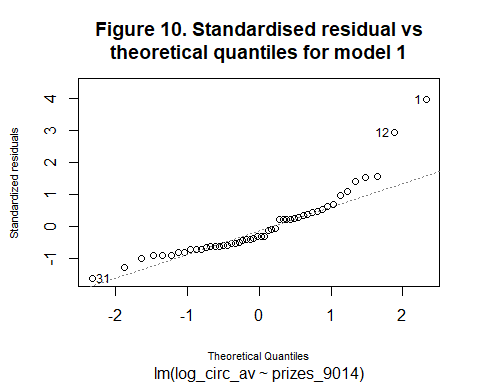
#change in circulation predicated by prizes  
plot(lm\_change\_prize, which = 3, title("Figure 9. Standardised residual vs fitted values  
for model 2"))



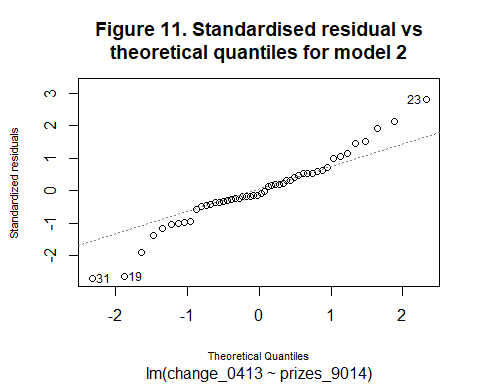
The relatively horizontal line in figure 8 suggests there is constant variance in the values of model 1, whereas the strong upward trend present in figure 9 suggests the variance is not constant in model 2.

The third assumption is normality, that the ‘noise’ (residuals) are normally distributed. A measure of this is a normal QQ plot of the standardised residuals against the theoretical quantiles. This is shown in figure 10 for model 1 and figure 11 for model 2.

#plot normal QQ  
#log average circulation predicated by prizes  
par(cex.lab=0.7)  
plot(lm\_circ\_prize, which = 2, title("Figure 10. Standardised residual vs  
theoretical quantiles for model 1"))



#change in circulation predicated by prizes  
par(cex.lab=0.7)  
plot(lm\_change\_prize, which = 2, title("Figure 11. Standardised residual vs  
theoretical quantiles for model 2"))



The closest the points on each figure lay to the dotted line, the more normally distributed the data is. Both figure 10 and figure 11 show data that diverges from normality outside the -1 to 1 quantile range, but the issue is more pronounced in figure 11. From this it appears that model 1 and 2 may both breach the assumption of normality.

The final assumption is independence- that the observations of data the models are based on are independent. One issue might arise if several of the newspapers are all owned by the same parent company, which would reduce the independence. This could be investigated in future work.

## Question 4: Prediction

4.1

The following calculations are based on model 1, which predicts circulation based on the number of pulitzer prizes. The continuing investment in journalism to the same degree as present, and an accompanying award of 25 pulitzer prizes, would result in a predicted circulation of 367,775. A reduction in the investment in investigative journalism, resulting in only 3 prizes, has a predicted circulation of 269,789. An increased investment and the receipt of 50 prizes has a predicted circulation result of 522,984.

As the current circulation is 453,869, only an increased investment in investigative journalism would result in an circulation increased according to this model.

#create tibble with values to prediction based on  
predict\_25 <-   
 tibble(  
 prizes\_9014 = 25  
 )  
#run prediction  
same1 <- as\_tibble(exp(predict(lm\_circ\_prize, newdata = predict\_25, interval = "confidence",level = 0.90)))  
  
predict\_3 <-   
 tibble(  
 prizes\_9014 = 3  
 )  
  
reduce1 <- as\_tibble(exp(predict(lm\_circ\_prize, newdata = predict\_3, interval = "confidence",level = 0.90)))  
  
predict\_50 <-   
 tibble(  
 prizes\_9014 = 50  
 )  
  
increase1 <- as\_tibble(exp(predict(lm\_circ\_prize, newdata = predict\_50, interval = "confidence",level = 0.90)))

4.2

Using model 2, which predicts the change in circulation based on the number of puliterzer prizes won, no change in investment in journalism will result in a reduction in circulation of 25.740%. A decreased in investigative funding is predicted to bring a drop in circulation of 34.254%, and an increase in funding is predicted to result in a fall of 16.065% in circulation.

The predictions of the two models are not consistent. None of the proposed strategic options is predicted to bring an increase in circulation under model 2, whereas the previously discussed results from model one suggested that an increase in investigative journalistic funding would result in an overall increase in circulation.

#create tibble with values to prediction based on  
predict\_25 <-   
 tibble(  
 prizes\_9014 = 25  
 )  
#run prediction  
same2 <- as\_tibble(predict(lm\_change\_prize, newdata = predict\_25, interval = "confidence",level = 0.90))  
  
  
predict\_3 <-   
 tibble(  
 prizes\_9014 = 3  
 )  
  
reduce2 <- as\_tibble(predict(lm\_change\_prize, newdata = predict\_3, interval = "confidence",level = 0.90))  
  
  
predict\_50 <-   
 tibble(  
 prizes\_9014 = 50  
 )  
  
increase2 <- as\_tibble(predict(lm\_change\_prize, newdata = predict\_50, interval = "confidence",level = 0.90))

4.3

Table 2 below shows the upper and lower confidence limits for each management scenario

#bind all of the predictions and confidence intervals for model 1 to a tibble  
model\_1 <-tibble()  
  
model\_1 <- model\_1 %>%  
 bind\_rows(same1,reduce1,increase1) %>%  
 mutate(scenario = c("Same level of investment","Reduced investment","Increased investment")) %>%  
 transform(predicted\_value = fit, lower\_confidence\_interval = lwr, upper\_confidence\_interval = upr) %>%  
 select(scenario, predicted\_value,lower\_confidence\_interval,upper\_confidence\_interval)  
  
#table of predicted values and confidence intervals  
pander(model\_1,caption = "Table 2. Predicted values and confidence intervals for model 1")

Table 2. Predicted values and confidence intervals for model 1

|  |  |  |  |
| --- | --- | --- | --- |
| scenario | predicted\_value | lower\_confidence\_interval | upper\_confidence\_interval |
| Same level of investment | 367,775 | 323,729 | 417,814 |
| Reduced investment | 269,789 | 235,516 | 309,050 |
| Increased investment | 522,984 | 425,950 | 642,124 |

The confidence intervals in table 2 set out what model 1 predicts the circulation will be under each scenario 90% of the time. For example, 9 times out of ten, the circulation if the investment in journalism is increased will be between 425,950 and 642,124. Similarly, under the same level of investment the circulation could be as low as 323,729 or as high as 417,814 (in 90% of cases). Finally, the model predicts that in most cases, a reduced investment in investigative journalism will result in a circulation of between 235,516 and 309,050.

4.4

#bind all of the predictions and confidence intervals for model 2 to a tibble  
model\_2 <-tibble()  
  
  
model\_2 <- model\_2 %>%  
 bind\_rows(same2,reduce2,increase2) %>%  
 mutate(scenario = c("Same level of investment","Reduced investment","Increased investment")) %>%  
 transform(predicted\_value = fit, lower\_confidence\_interval = lwr, upper\_confidence\_interval = upr) %>%  
 select(scenario, predicted\_value,lower\_confidence\_interval,upper\_confidence\_interval)  
  
#table of predicted values and confidence intervals  
pander(model\_2,caption = "Table 3. Predicted values and confidence intervals for model 2")

Table 3. Predicted values and confidence intervals for model 2

|  |  |  |  |
| --- | --- | --- | --- |
| scenario | predicted\_value | lower\_confidence\_interval | upper\_confidence\_interval |
| Same level of investment | -25.74 | -32.21 | -19.275 |
| Reduced investment | -34.25 | -41.14 | -27.368 |
| Increased investment | -16.07 | -26.47 | -5.663 |

Table 3 sets out the 90% confidence intervals for each scenario as predicted by model - the change in circulation predicted by the model in each situation. The model predicts that in 9 out of ten cases, a reduced investment in journalism will result in a reduction in circulation of between 41.14% and 27.368%, and an increased investment will results in a drop of between 26.47% and 5.663%. If the investment in investigative journalism is increased, the fall in circulation should be between 32.21% and 19.275%.

## Question Five: Limitations

5.1

There are limitations to this modeling that must be considered. Importantly, it’s possible that a third factor is responsible for both any change in circulation and the winning of pulitzer prizes. For example, the funding available to a newspaper will determine it’s ability to advertise itself and to hire quality journalists. In this way, both the circulation and the number of prizes might both be determined by the funding available.

Another issue is that of unknowns. The model is only able to predict circulation results in the context of the factors available in the data. There is always the possibility that unknown factors or events are key to a newspaper’s circulation. These might be local or national news events, major investigative stories, or even the general state of the economy. A more robust model could explore additional factors and attempt to control for these.