Masthead Media

To: MastHeadMedia.com.au

CC:

Subject: Data Science Team Publication Analysis

Afternoon Mr Russel,

The data science team has completed their analysis and model forecasting with regards to the influence winning Pulitzer Prizes has on a publication's average circulation and percentage change in circulation over the desired period.

Based upon our modelling, we see a moderate positive linear relationship between the number of Pulitzer Prizes won and an increase in average circulation on average. For instance, when a publication has zero Pulitzer Prizes, we would expect their average circulation to be 258,626. However, if they were to increase the number of Pulitzer Prizes by one, then on average we expect average circulation to increase by 3,669 amounting to 262,295 average circulations. Conclusively, publications that win more Pulitzer Prizes have a larger average circulation on average (3.1). Based upon our modelling, we see a weak positive linear relationship between the number of Pulitzer Prizes won and a percentage increase in circulation on average. For instance, when a publication has zero Pulitzer Prizes, we would expect their percentage change in circulation to decrease by 35.4% on average. However, for every additional Pulitzer Prize won we expect percentage change in circulation to increase by 0.3870%. Notably, 91 Pulitzer Prizes would be the break-even percentage change based upon our fitted model values. Conclusively, publications that win more Pulitzer Prizes have an increase in percentage change in circulation on average (3.2).

The data science team proposed a null hypothesis that there is no relationship between Pulitzer Prizes and increases in average circulation or percentage change in circulation. Notably, both of the aforementioned models' results provide support or give evidence rejecting this null hypothesis as the results are not likely to occur due to chance but are instead likely attributable to a specific case. Therefore, pending our linear assumptions check, our model indicates that there is a statistically significant relationships between average circulation and Pulitzer Prizes as well as percentage change in circulation and Pulitzer Prizes (3.1&3.2).

With the 'Garbage in Garbage Out' principle in mind, the standard principle that more training data (especially more quality data) leads to better models is applicable as our model's limited by only being "trained" on 50 observations. I would argue that building a linear regression model based upon only 50 observations is insufficient especially since both models contain highly influential outliers. I would also like to draw your attention to the Cook's distance graphs under Q3.3. Evidentially our models contain observations that have high influence on our regression analysis specifically our assumption graphs. With a small dataset exhibiting 3-6% of observations being highly influential and outliers, bear in mind that the red line will suffer deviations attributed to a few end observations points.

There are four assumptions associated with our linear regression models: Linearity: the relationship between our dependant variable and Pulitzer Prize is linear. Constant variance (Homoscedasticity): the models residual/error values have constant variance. Normality: our models residuals/error values are normally distributed (i.e. follow a bell curve). Independence: observations are independent of each other (3.3).

The assumption of linear relationship is justified for model two but model one requires further investigation. Model two indicates there is certainly not a strong relationship, but the small size of the observations seems to be affecting the reference line. The downward trend towards the higher observation can be attributed to just a few tail observations. In Contrast, despite the fact that log transforming average circulation flattened out model ones' red line towards the higher values we still see a worrying curvature in the bulk portion of our data points indicating that the assumption is not satisfied and warrants further investigation.

For both models we can conclude that the residuals are normally distributed. Notably, we can see the aforementioned outliers towards the tails of both plots. Despite our model's normal quantile-quantile plot of the residuals and residuals histogram vaguely showing signs of distributions with right-skewness and heavy-skewness respectively, the bulk portion of our model's observations lie on the line with minimal deviations allowing us to make this conclusion.

The assumption of constant spread is justified for model one but model two requires further investigation. While the bulk portion of our observations in model two illustrate peaks and valleys with no apparent trend, there is a clear trend. As fitted values increase so too does spread in residuals (moving left to right). In Contrast, there is no clear trend in our data as we move from left to right in model one.

Notably, independence can only be assessed by looking at how the data was obtained. Information about the sampling process is not provided so it is difficult to make a judgement.

It is important to note that our two models offer inconsistencies with regards to their predictions. According to model one, only investing substantially more in journalism satisfies our condition of circulation being in excess of current circulation. Interestingly, investing the same amount in journalism will not yield desired results indicating the only way to increase circulation is to increase investments in journalism. Therefore, based upon model one, the trajectory of the Boston Sun-Times's circulation will only increase if they follow the strategic direction of investing substantially more in journalism (4.2-4.4).

For model two, all three instances have a prediction range (lower to upper) of 86% which is a significant amount and not ideal for our investigation as there is too much uncertainty in these predicted values. Interestingly, all upper values yield positive changes in circulation but the downside (lower) is heavily skewed. Therefore, based upon model two, our prediction intervals range is too wide and negatively skewed too justify investing substantially more in journalism (4.4).

Notably, model two would be insufficient for application because using Pulitzer Prize to explain the variability of percentage change in circulation does not seem to be well suited. Bascially, the accuracy of predicted values based upon this model would not yeild convinging results. Notably, both models are limited by their ability to explain increases in circulation variability with a single variable Pulitzer Prize. Additionally, time serious modelling would be ideal, but we are limited with our data and ability to also investigation autocorrelation further.

Please See attachment below.

Kind Regards Data Science Team

Loading packages and importing file.csv

```
pacman::p_load(pacman, tidyverse,rmarkdown,ggplot2,dplyr,mplot,
               gvlma,caret,rio,olsrr,car,stringr,gridExtra,hexbin,lmtest,ggpubr, skim
r)

df <- rio::import('/Users/garethbayvel/Desktop/rStudio/pulitzer.csv',setclass = 'tibb
le')
head(df)
```

```
## # A tibble: 6 x 5
##   newspaper           circ_2004 circ_2013 change_0413 prizes_9014
##   <chr>                   <int>     <int> <chr>             <int>
## 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
```

Question One

1.1 Recode % change in circulation between 2004 and 2013 as an integer

```
df$change_0413 <- str_split(df$change_0413, '%',simplify = TRUE)[,1] %>%
  as.integer()
knitr::kable(head(df))
```

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

1.2 Append a new variable containg the average of circ_2004 and circ_2013.

```
df<- df%>%
  rowwise()%>%
  mutate('Circ2004_Circ2013' = mean(c(circ_2004,circ_2013)))
knitr::kable(head(df))
```

newspaper	circ_2004	circ_2013	change_0413	prizes_9014	Circ2004_Circ2013
USA Today	2192098	1674306	-24	3	1933202.0

newspaper	circ_2004	circ_2013	change_0413	prizes_9014	Circ2004_Circ2013
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

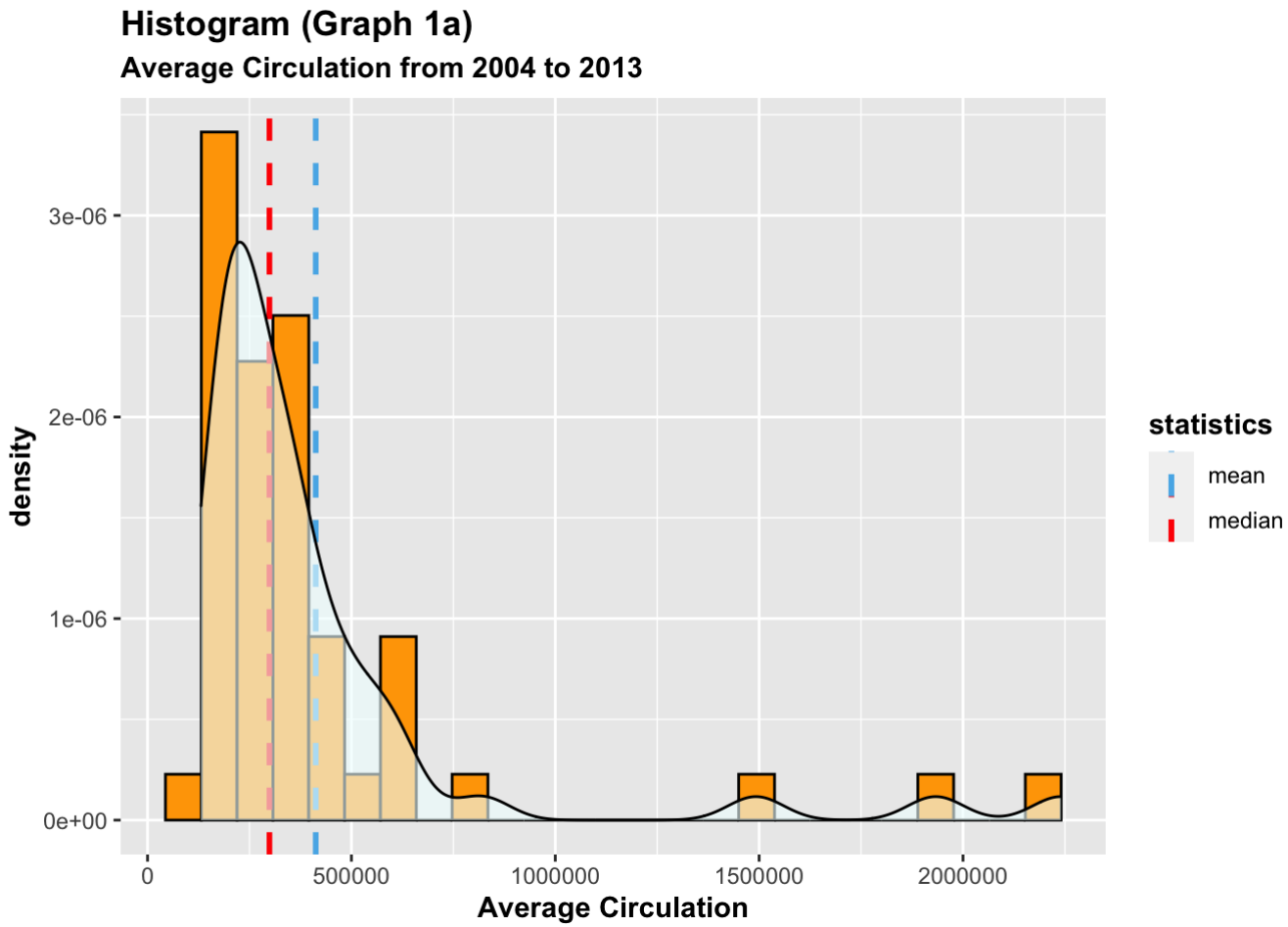
Question Two

2.1 Describe the distribution of the variable representing average circulation

Shape

This histogram clearly illustrates a right skewed unimodal distribution. Notably, our distribution has a mean (412,442.3) greater than its median (298,851.2) indicating a right skew (Graph 1a).

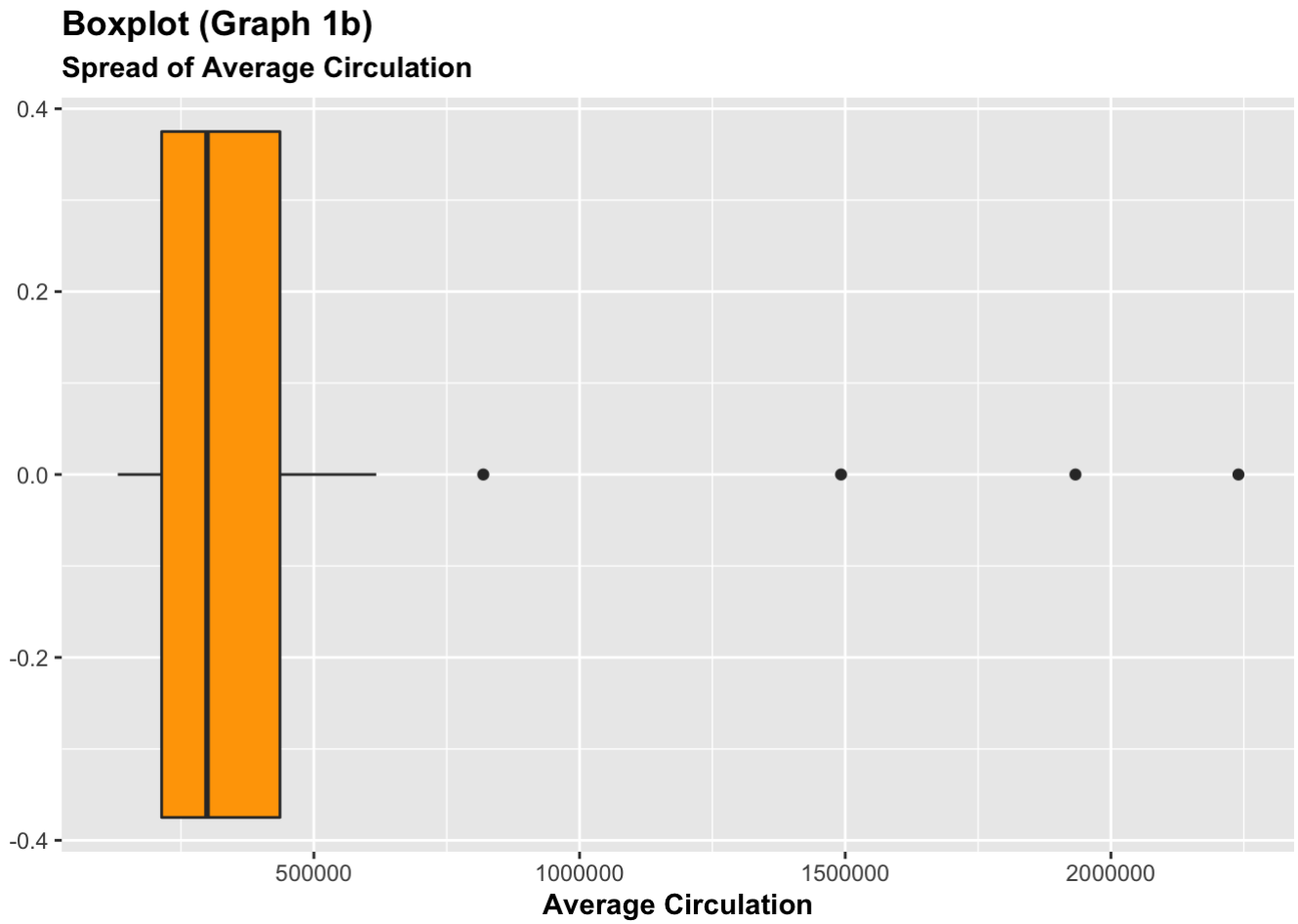
```
df%>%
  ggplot(aes(x = Circ2004_Circ2013, y = ..density..)) +
  geom_histogram(col = 'black', bins = 25, fill= 'orange') +
  labs(title = "Histogram (Graph 1a)", subtitle = "Average Circulation from 2004 to 2
013") +
  xlab("Average Circulation") +
  theme(title = element_text(face = 'bold')) +
  geom_vline(aes(xintercept=mean(Circ2004_Circ2013, na.rm =T),
                 color="mean"), linetype="dashed", size=1) +
  geom_vline(aes(xintercept=median(Circ2004_Circ2013, na.rm =T),
                 color="median"), linetype="dashed", size=1) +
  geom_density(alpha = 0.6,fill="azure") +
  scale_color_manual(name = "statistics", values = c(median = "red", mean = "#56B4E9"
))
```



Location

In contrast to normally distributed data the measures of central tendency for positively skewed data are disperse (mean > median > mode). Thus, the median is the preferred metric for central tendency as our mean seems to be distorted by 4 outliers. The location (median) of our distribution is 298,851 (Graph 1b).

```
df %>%
  ggplot(aes(x = Circ2004_Circ2013)) +
  geom_boxplot(fill = 'orange') +
  labs(title = "Boxplot (Graph 1b)", subtitle = "Spread of Average Circulation") +
  xlab('Average Circulation')+
  theme(title = element_text(face = 'bold'))
```



Outliers

Outliers are generally indicated by laying more than 1.5 times the IQR of the upper (Q3) and lower (Q1) quartiles. IQR (222,643.4 x 1.5 = 333,965) for the upper limit is Q3 plus 333,965 (770,117) indicating we have 4 outliers laying more than 1.5 times the upper quartile. Specific notice should be paid to the extreme outliers New York Times (1,492,172), USA Today (1,933,202), and, Wall Street Journal (2,239,922) as they all lie more than 3 times greater than our upper Q3 (Graph 1b).

Spread

Spread can be obtained from its interquartile range (Q3 – Q1). Our distribution has an IQR of 222,643.4 (436,152 – 213,509). Indicating the spread of the middle 50% of our set of values lies within 213,509 to 436,152 circulations. A range of 2,108,918 (min of 131,004 - max of 2,239,922) (Graph 1b&1c).

```
summary(df$Circ2004_Circ2013)
```

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

Interquartile Range

```
IQR(df$Circ2004_Circ2013)
```

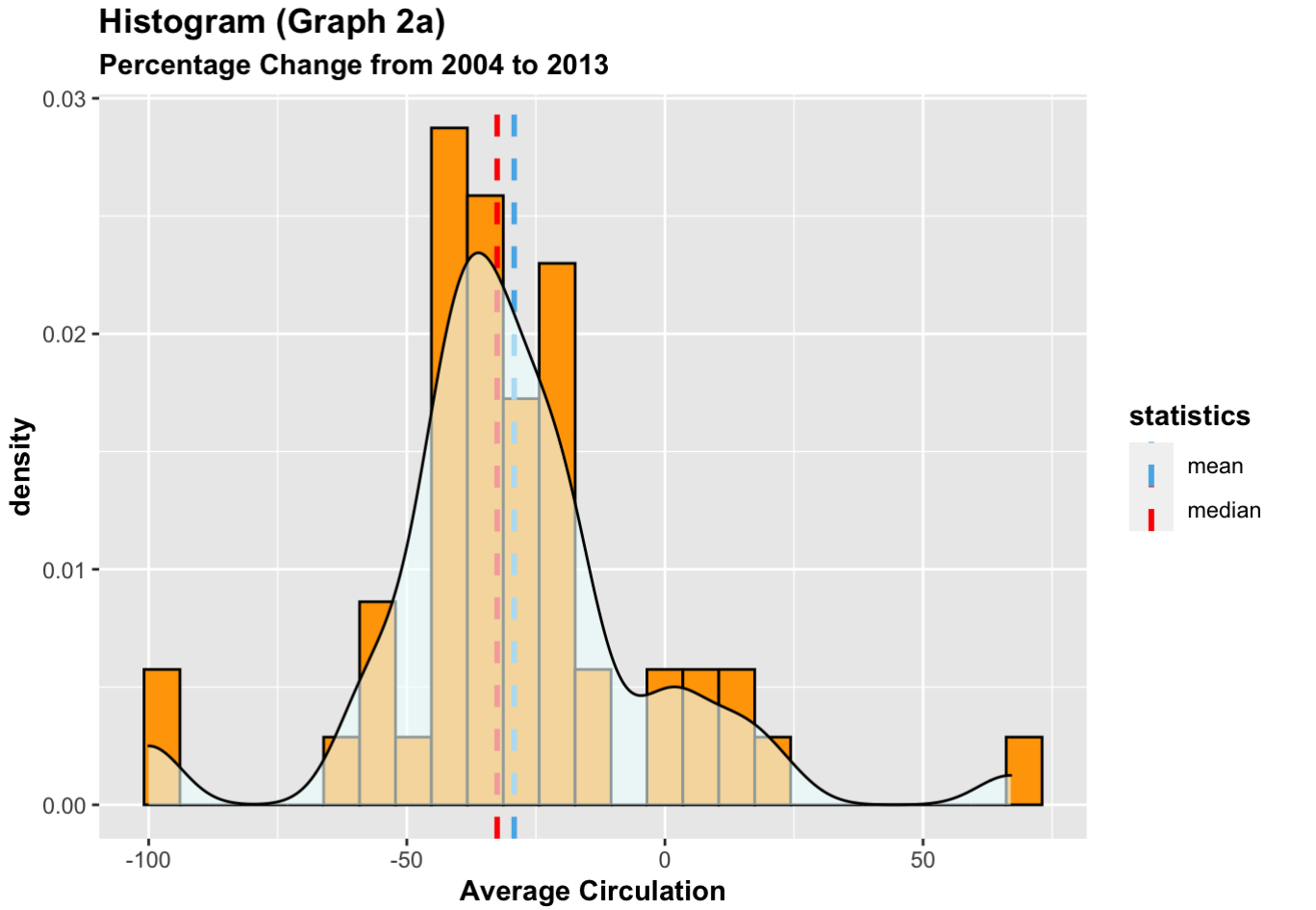
```
## [1] 222643.4
```

2.2 Describe the distribution of the variable representing change_0413

Shape

This histogram appears more normally distributed however, it is obviously not symmetric and illustrates a slight right skewed unimodal distribution with most values tailoring off with a positive right skew. Notably, our distribution has a mean (-29.20) greater than its median (-32.50) indicating a skewness to the right (Graph 2a).

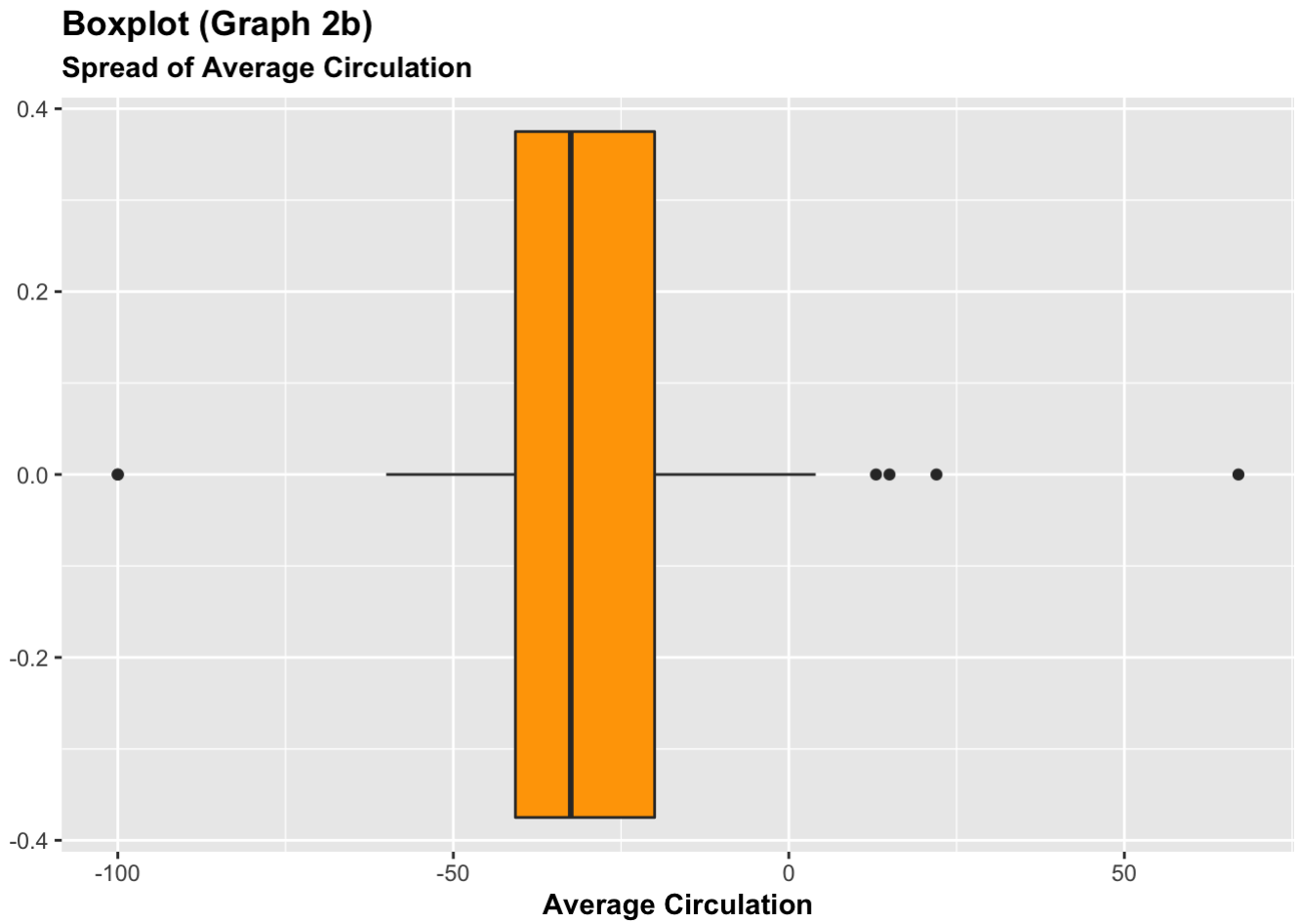
```
df%>%
  ggplot(aes(x = change_0413, y = ..density..)) +
  geom_histogram(col = 'black', bins = 25, fill= 'orange') +
  labs(title = "Histogram (Graph 2a)", subtitle = "Percentage Change from 2004 to 201
3") +
  xlab("Average Circulation") +
  theme(title = element_text(face = 'bold')) +
  geom_vline(aes(xintercept=mean(change_0413, na.rm =T),
                 color="mean"), linetype="dashed", size=1) +
  geom_vline(aes(xintercept=median(change_0413, na.rm =T),
                 color="median"), linetype="dashed", size=1) +
  geom_density(alpha = 0.6,fill="azure") +
  scale_color_manual(name = "statistics", values = c(median = "red", mean = "#56B4E9"
))
```



Location

The median is the preferred metric for central tendency as the value of the mean seems to be distorted by 4 rightward outliers and two leftward outliers. Therefore, the location of our distribution is -32.50 (Graph 2b).

```
df %>%
  ggplot(aes(x = change_0413)) +
  geom_boxplot(fill = 'orange') +
  labs(title = "Boxplot (Graph 2b)", subtitle = "Spread of Average Circulation") +
  xlab('Average Circulation')+
  theme(title = element_text(face = 'bold'))
```



Outliers

Our upper limit is Q3 plus 31.13 (11.3) indicating we have 4 outliers laying more than 1.5 times the upper quartile. Additionally, our lower limit is Q1 minus 31.13 (71.88) indicating we have 2 outliers laying more than 1.5 times the lower quartile. Specific notice should be paid to the two extreme lower outliers Rocky Mountain News and New Orleans Times-Picayune as a 100% decline seems to indicate they are no longer in operation. Special notice should also be paid towards New York Times (67%) as it lays more than 4 times the upper Q3 (Graph 2b).

Spread

Our distribution has an IQR (spread) of 20.75 (-20.00 – (-40.75)). Therefore, indicating the spread of the middle 50% of our set of percentage change in circulations lies within -40.75 to -20.00. A range of 167 (min of -100, max of 67) (Graph 2b&c).

```
summary(df$change_0413)
```

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

Interquartile Range

```
IQR(df$change_0413)
```

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

2.3 A skew that could be resolved by a log transform?

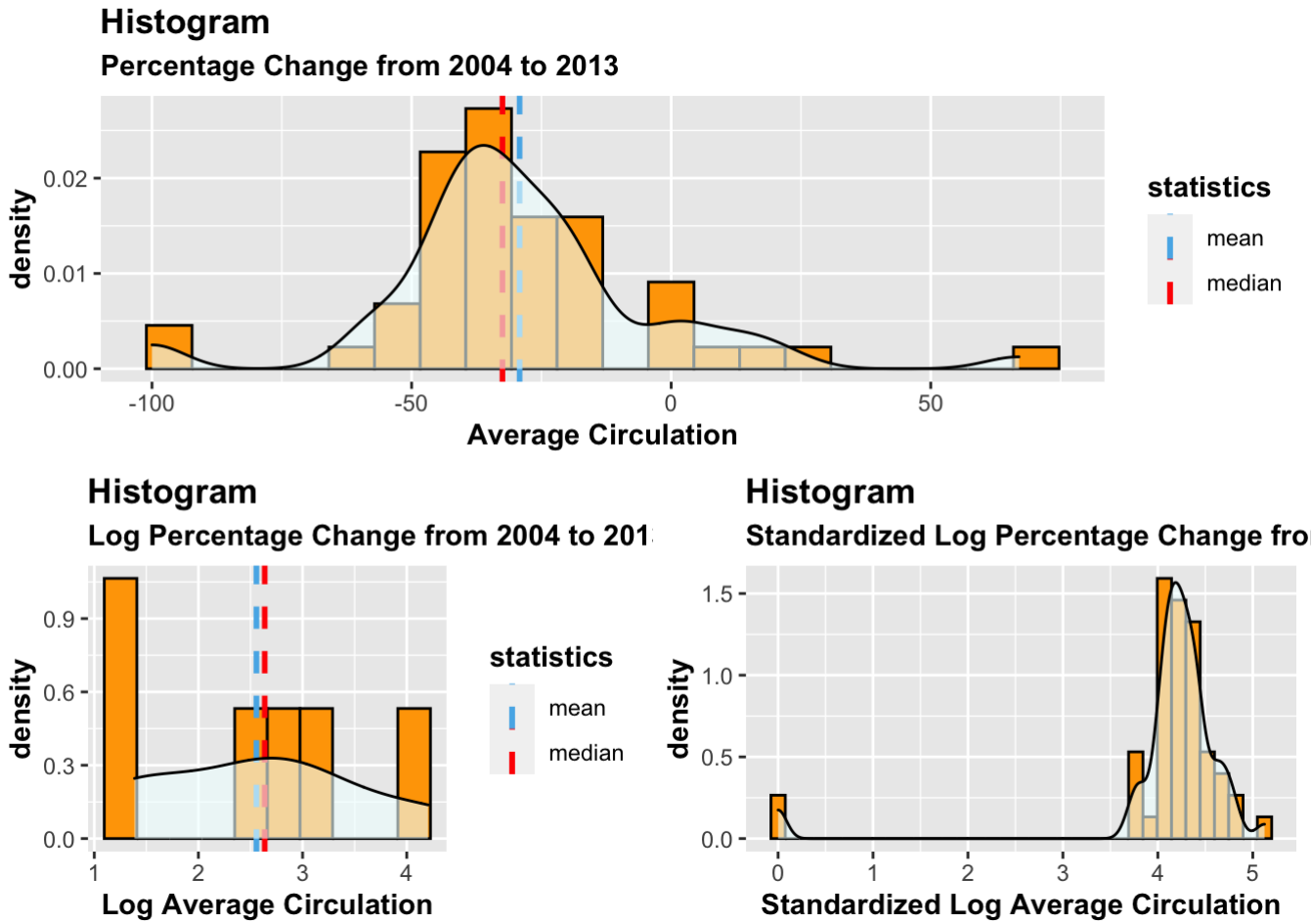
Standardized and Log transformed to avoid undefined values

```
df$mutated <- log(df$change_0413 + 1 - min(df$change_0413))
```

Change Variabe

Notably, the natural logarithm is defined only for data points greater than 0. With regards to our dataset, 6 out of the 50 observations contained in change_0413 are non-negative values. Therefore, a log transformation of this column is unnecessary as it would reduce our dataset by 88%. Additionally, standardising change_0413 and then implementing a log transformation does not resolve the skewness and actually creates a model that explains less than 3% of model variance and generates a staistically insignificant coefficient (Graph 3a). Interestingly, standardising and log transforming change_0413 and log transforming Prizes seems to statisfy all linear assumption but is the worst performing model in terms of explaining less than 0.8% model variance. Already this prediction model seems to have a few limitations.

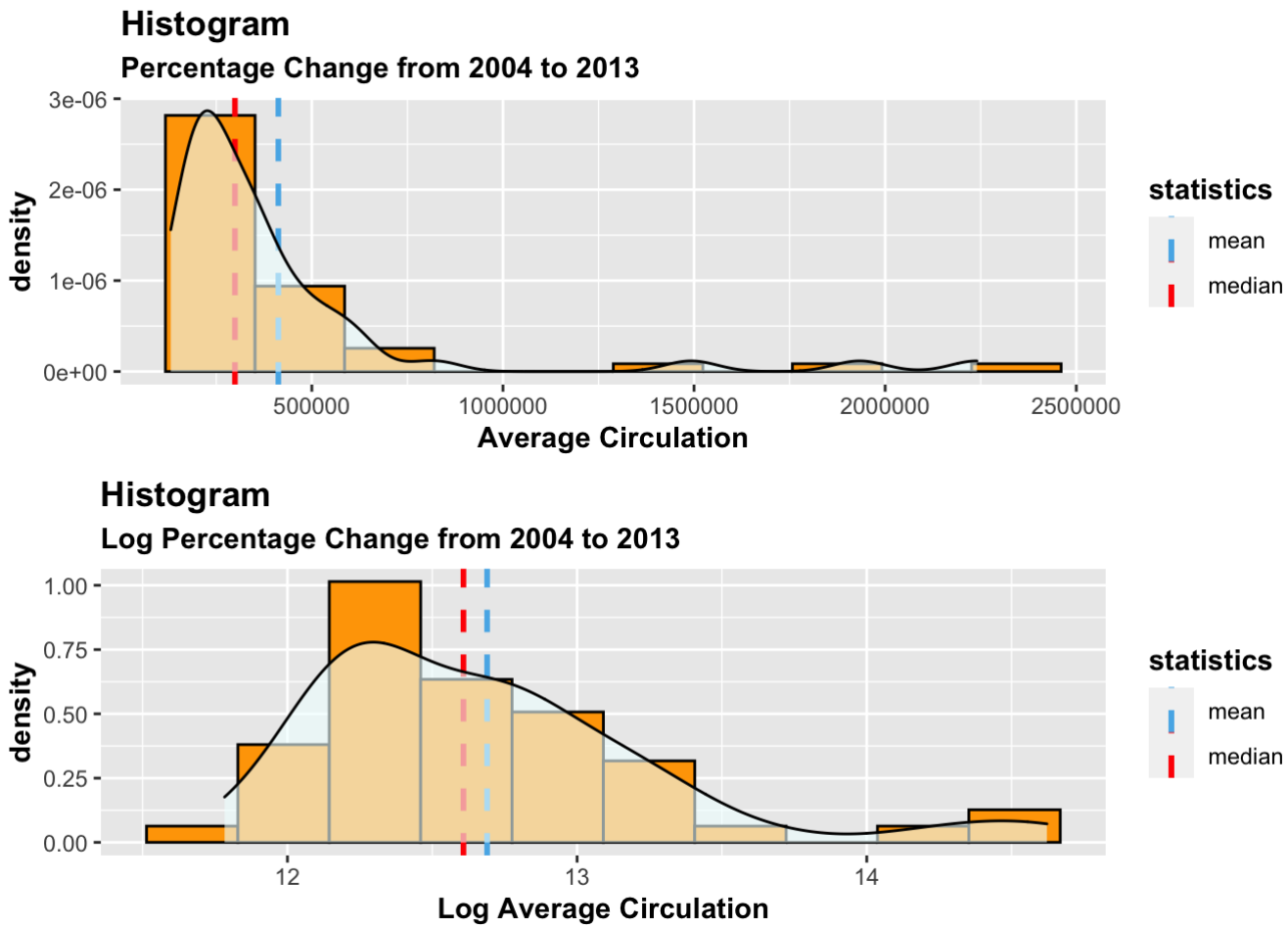
```
plot1 <- df%>%
  ggplot(aes(x = change_0413, y = ..density..)) +
  geom_histogram(col = 'black', bins = 20, fill= 'orange') +
  labs(title = "Histogram", subtitle = "Percentage Change from 2004 to 2013") +
  xlab("Average Circulation") +
  theme(title = element_text(face = 'bold')) +
  geom_vline(aes(xintercept=mean(change_0413, na.rm =T),
                 color="mean"), linetype="dashed", size=1) +
  geom_vline(aes(xintercept=median(change_0413, na.rm =T),
                 color="median"), linetype="dashed", size=1) +
  geom_density(alpha = 0.6,fill="azure") +
  scale_color_manual(name = "statistics", values = c(median = "red", mean = "#56B4E9"
))

plot2 <- df%>%
  ggplot(aes(x = log(change_0413), y = ..density..)) +
  geom_histogram(col = 'black', bins = 10, fill= 'orange') +
  labs(title = "Histogram", subtitle = "Log Percentage Change from 2004 to 2013") +
  xlab("Log Average Circulation") +
  theme(title = element_text(face = 'bold')) +
  geom_vline(aes(xintercept=mean(log(change_0413), na.rm =T),
                 color="mean"), linetype="dashed", size=1) +
  geom_vline(aes(xintercept=median(log(change_0413), na.rm =T),
                 color="median"), linetype="dashed", size=1) +
  geom_density(alpha = 0.6,fill="azure") +
  scale_color_manual(name = "statistics", values = c(median = "red", mean = "#56B4E9"
))

plot14 <- df%>%
  ggplot(aes(x = mutated , y = ..density..)) +
  geom_histogram(col = 'black', bins = 35, fill= 'orange') +
  labs(title = "Histogram", subtitle = "Standardized Log Percentage Change from 2004
 to 2013") +
  xlab("Standardized Log Average Circulation") +
  theme(title = element_text(face = 'bold')) +
  geom_density(alpha = 0.6,fill="azure") +
  scale_color_manual(name = "statistics", values = c(median = "red", mean = "#56B4E9"
))
grid.arrange(plot1,
             arrangeGrob(plot2, plot14, ncol = 2),
             nrow = 2)
```



Average Circulation Variable

Average circulation has a right skew that cannot be completely resolved (transformed into a Gaussian normal distribution). However, performing a log transformation definitely improves the skewness of our data and makes it appear more normally distributed. There is strong visual evidence that average circulation should be log transformed (Graph 3b).

```
require(gridExtra)
plot3 <- df%>%
  ggplot(aes(x = Circ2004_Circ2013, y = ..density..)) +
  geom_histogram(col = 'black', bins = 10, fill= 'orange') +
  labs(title = "Histogram", subtitle = "Percentage Change from 2004 to 2013") +
  xlab("Average Circulation") +
  theme(title = element_text(face = 'bold')) +
  geom_vline(aes(xintercept=mean(Circ2004_Circ2013, na.rm =T),
                 color="mean"), linetype="dashed", size=1) +
  geom_vline(aes(xintercept=median(Circ2004_Circ2013, na.rm =T),
                 color="median"), linetype="dashed", size=1) +
  geom_density(alpha = 0.6,fill="azure") +
  scale_color_manual(name = "statistics", values = c(median = "red", mean = "#56B4E9"
))
plot4 <- df%>%
  ggplot(aes(x = log(Circ2004_Circ2013), y = ..density..)) +
  geom_histogram(col = 'black', bins = 10, fill= 'orange') +
  labs(title = "Histogram", subtitle = "Log Percentage Change from 2004 to 2013") +
  xlab("Log Average Circulation") +
  theme(title = element_text(face = 'bold')) +
  geom_vline(aes(xintercept=mean(log(Circ2004_Circ2013), na.rm =T),
                 color="mean"), linetype="dashed", size=1) +
  geom_vline(aes(xintercept=median(log(Circ2004_Circ2013), na.rm =T),
                 color="median"), linetype="dashed", size=1) +
  geom_density(alpha = 0.6,fill="azure") +
  scale_color_manual(name = "statistics", values = c(median = "red", mean = "#56B4E9"
))
grid.arrange(plot3, plot4, ncol=1)
```



Question Three

3.1 Build a model predicting the variable representing a newspaper's circulation using prizes_9014

```
df_log_lm <- lm(log(Circ2004_Circ2013)~prizes_9014,data = df)
summary(df_log_lm)
```

```
##
## Call:
## lm(formula = log(Circ2004_Circ2013) ~ prizes_9014, data = df)
##
## 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
```

Based on the log transformation of our dependent variable (average circulation), our linear models' output is:

$$Log(AverageCirculation) = 12.463142 + 0.014083x(no.PulitzerPrizes)$$

$$AverageCirculation = e^{(12.463142+0.014083x(no.PulitzerPrizes))}$$

Slope

If we increase the number of Pulitzer Prizes by one, then on average the average circulations will increase by 0.014083 on the log scale.

Intercept

The intercept provides us with the value of our dependent variable when our independent variable is zero. In our model this is the average circulation on the log scale when the number of Pulitzer prizes is zero. Therefore, if a news agency had zero Pulitzer Prizes, we would expect their average circulation to be 12.463142 on the log scale.

Statistical Significance

The p value associated with the coefficient for Pulitzer Prizes is 1.53e-05 (0.0000153). This is notably below a significance level of 0.01. Generally, p values < 0.05 are statistically significant and indicate strong evidence against the null hypothesis which, for our model, would indicate that there is no relationship between the average circulation and Pulitzer Prizes. Therefore, pending our linear assumptions check, our model indicates that there is a statistically significant relationship between average circulation and Pulitzer Prizes.

3.2 Build a model predicting change_0413 using prizes_9014

```
df_change_lm <- lm(change_0413~prizes_9014,data = df)
summary(df_change_lm)
```

```
##
## Call:
## lm(formula = change_0413 ~ prizes_9014, data = df)
##
## 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
```

Based on our model predicting change_0413 using prizes_9014 the linear model's output is:

$$PercentageChangeinCirculation = -35.4152 + 0.3870x(no.PulitzerPrizes)$$

This model was the best fitting model in terms of explaining the variance in our dependent variable using the r-squared metric.

Slope

If we increase the number of Pulitzer Prizes by one, then on average the percentage change in circulation will increase by 0.3870.

Intercept

The intercept provides us with the value of our percentage change of circulation when the number of Pulitzer Prizes is zero. Therefore, if a news agency had zero Pulitzer Prizes, we would expect their average percentage change of circulation to be -35.4152.

Statistical Significance

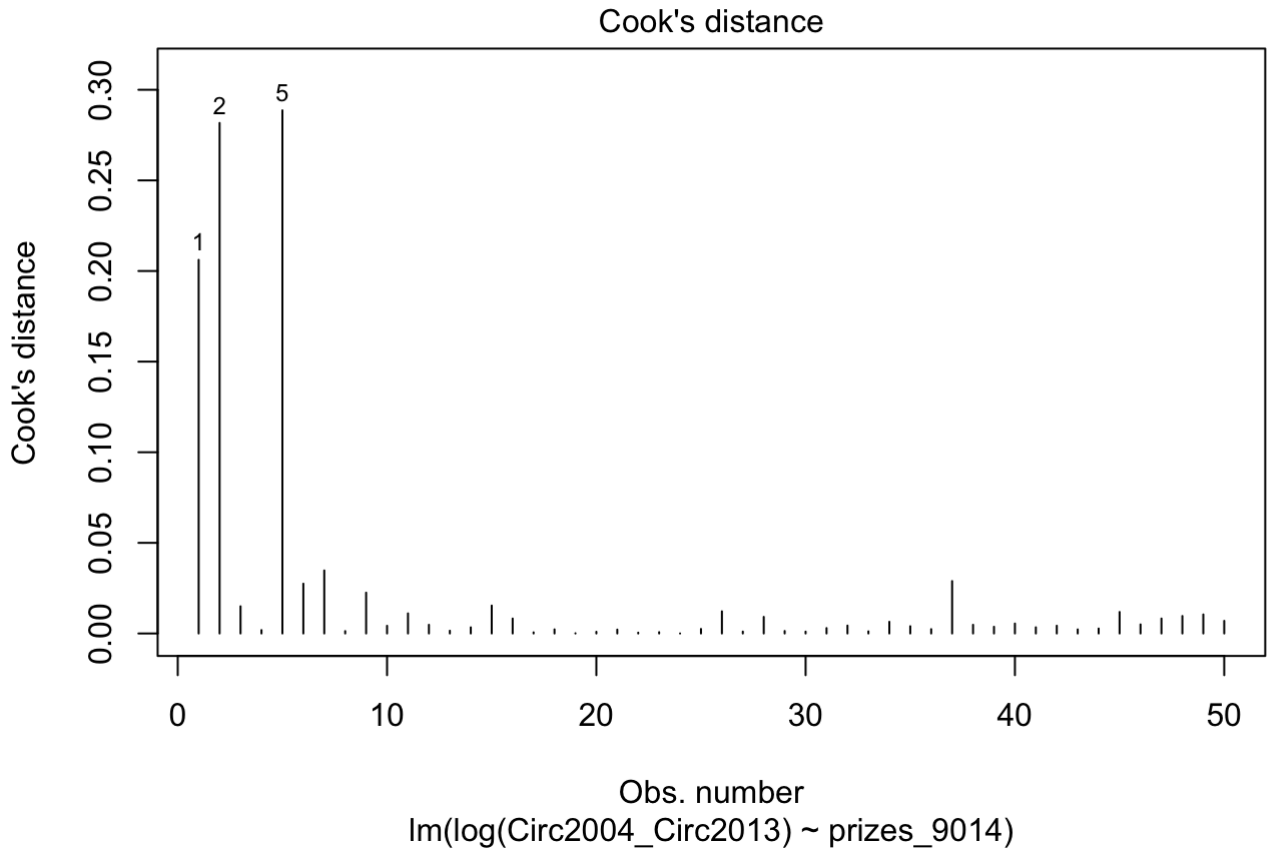
The p value associated with the coefficient for Pulitzer Prizes is 0.0121. Generally, p values < 0.05 are statistically significant and indicate strong evidence against the null hypothesis which, for our model, would indicate that there is no relationship between the percentage change in circulation and Pulitzer Prizes. Therefore, pending our linear assumptions check, our model indicates that there is a statistically significant relationship between percentage change in circulation and Pulitzer Prizes.

3.3 Check the assumptions of both linear models.

Model One

Notably, data points 1 and 2 appear in every plot and according to Cook's distance rule of thumb exceeding 4/(n-p-1), these observations have high influence on the regression analysis (including our graphs).

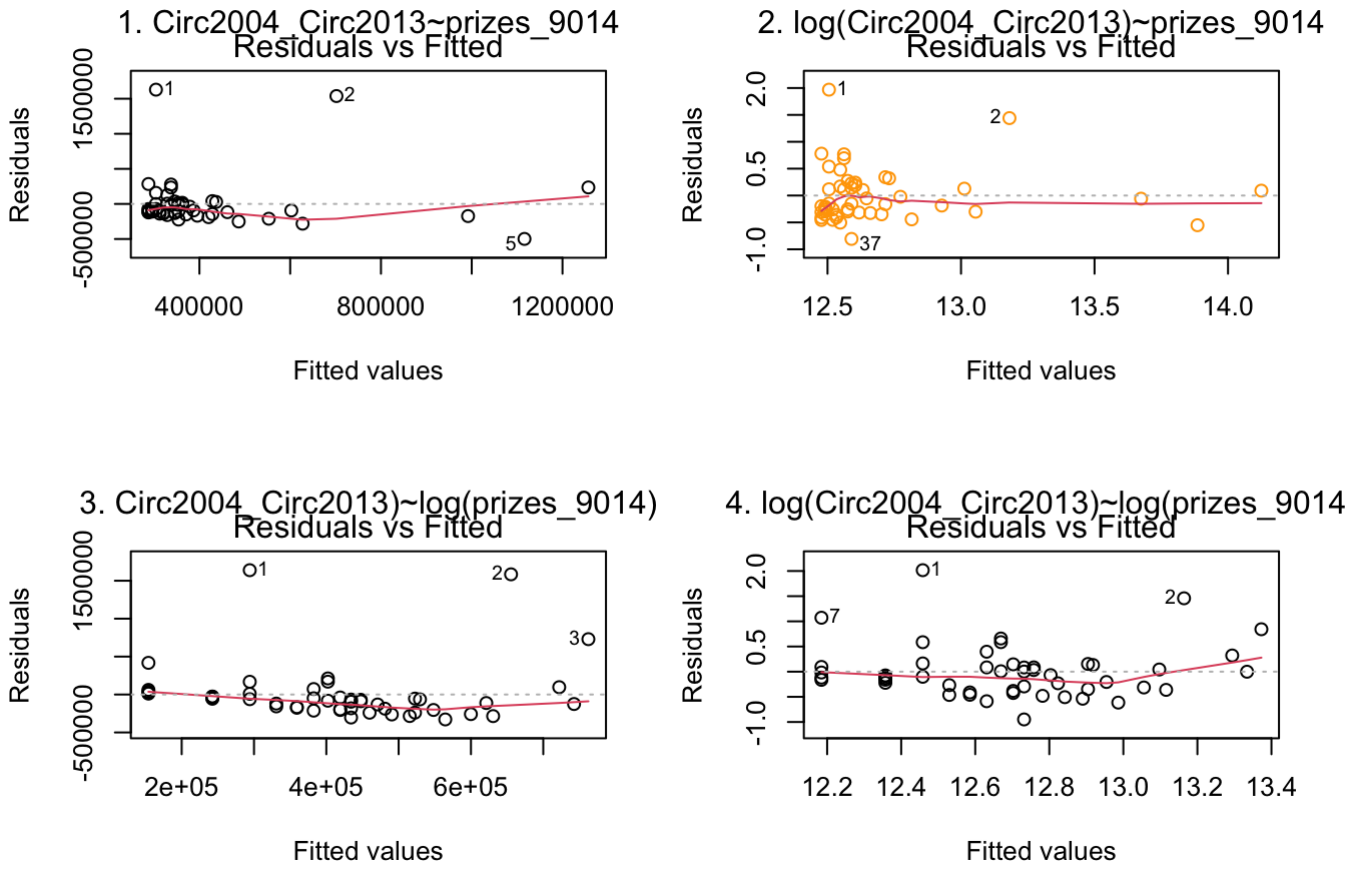
```
plot(df_log_lm, which =4)
```



Linear Relationship

In our model the red line is roughly straight which at first instance may indicate the presence of a linear relationship, however, despite the fact that log transforming average circulation flattened out the red line towards the higher values we still see a worrying curvature in the bulk portion of our data points indicating that the assumption is not satisfied and warrants further investigation. Interestingly, log transforming average circulation (orange) seems to have exacerbated the curvature slightly present from our standard model (plot 1).

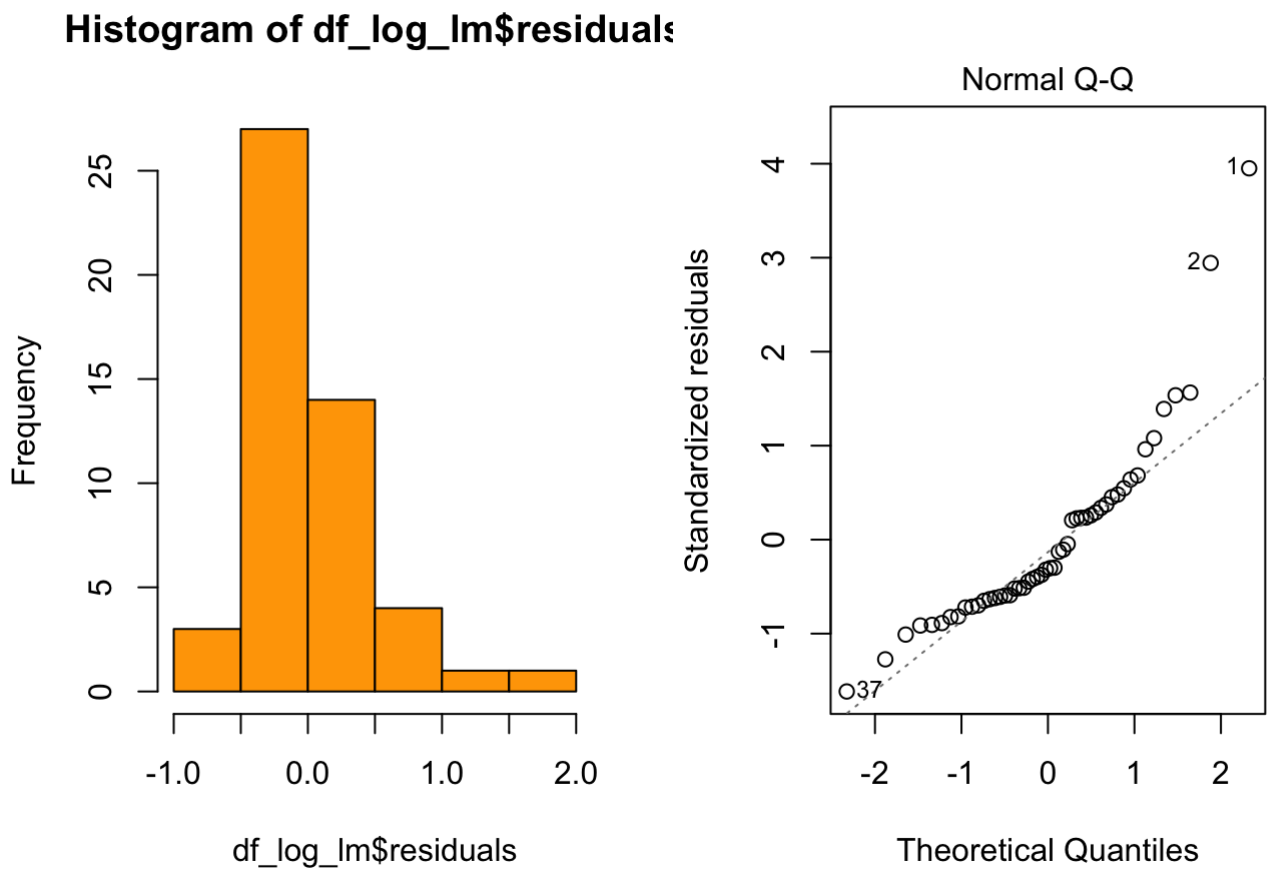
```
par(mfrow=c(2,2))
plot(lm(Circ2004_Circ2013~prizes_9014,data=df), which = 1)
mtext("1. Circ2004_Circ2013~prizes_9014", side=3, line=1, cex=1, adj=0.5)
plot(lm(log(Circ2004_Circ2013)~prizes_9014,data=df), which = 1,col = 'orange')
mtext("2. log(Circ2004_Circ2013)~prizes_9014", side=3, line=1, cex=1, adj=0.5)
plot(lm(Circ2004_Circ2013~log(prizes_9014),data=df), which = 1)
mtext("3. Circ2004_Circ2013)~log(prizes_9014)", side=3, line=1, cex=1, adj=0.5)
plot(lm(log(Circ2004_Circ2013)~log(prizes_9014),data=df), which = 1)
mtext("4. log(Circ2004_Circ2013)~log(prizes_9014)", side=3, line=1, cex=1, adj=0.5)
```



Noise Normally Distribution

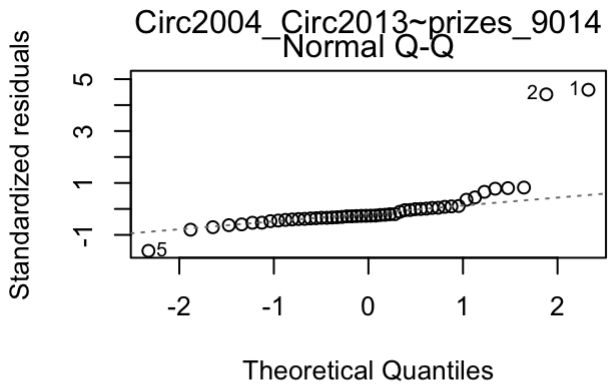
The normal quantile-quantile plot of the residuals and residuals histogram vaguely indicate a right-skewed distribution. Notably, the fact that the bulk proportion of our observations lie on the line with minimal deviations allows us to conclude that the residuals are normally distributed.

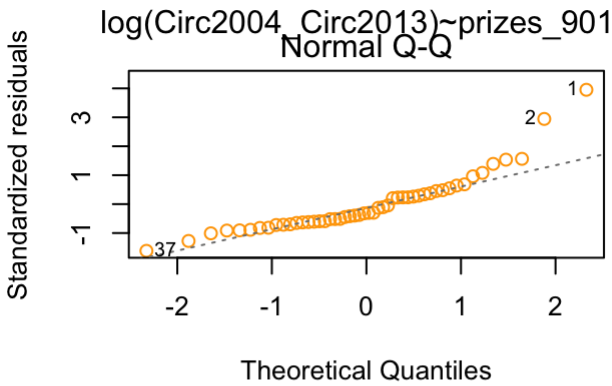
```
par(mfrow=c(1,2))
hist(df_log_lm$residuals, col = 'orange')
plot(df_log_lm, which = 2)
```

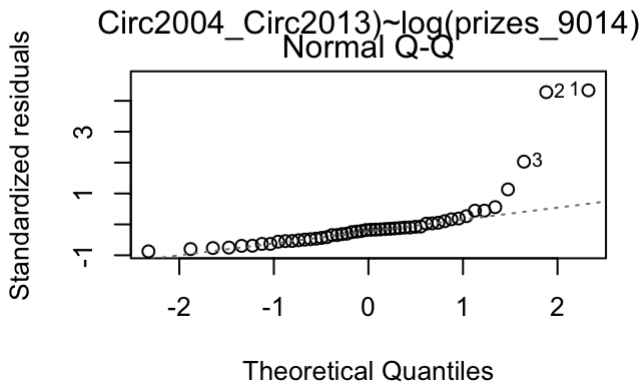


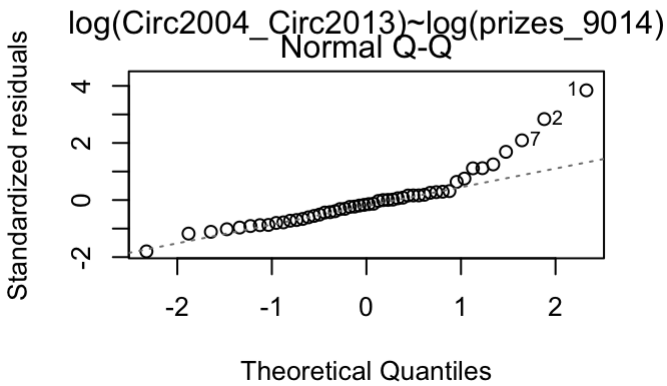
It is evident that our model's errors are noramlly distributed.

```
par(mfrow=c(2,2))
plot(lm(Circ2004_Circ2013~prizes_9014,data=df), which = 2)
mtext("Circ2004_Circ2013~prizes_9014", side=3, line=1, cex=1, adj=0.5)
plot(lm(log(Circ2004_Circ2013)~prizes_9014,data=df), which = 2,col = 'orange')
mtext("log(Circ2004_Circ2013)~prizes_9014", side=3, line=1, cex=1, adj=0.5)
plot(lm(Circ2004_Circ2013~log(prizes_9014),data=df), which = 2)
mtext("Circ2004_Circ2013)~log(prizes_9014)", side=3, line=1, cex=1, adj=0.5)
plot(lm(log(Circ2004_Circ2013)~log(prizes_9014),data=df), which = 2)
mtext("log(Circ2004_Circ2013)~log(prizes_9014)", side=3, line=1, cex=1, adj=0.5)
```





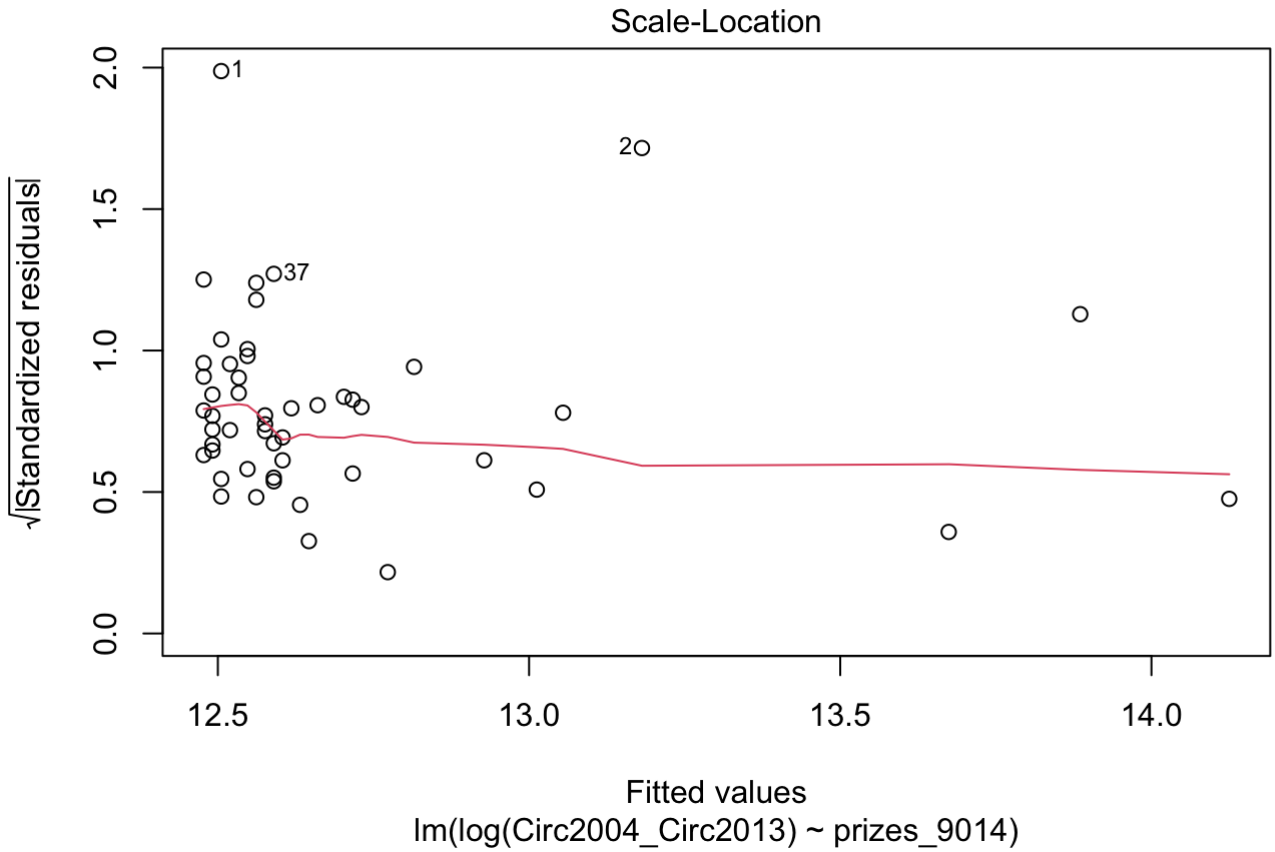




Homoscedasticity

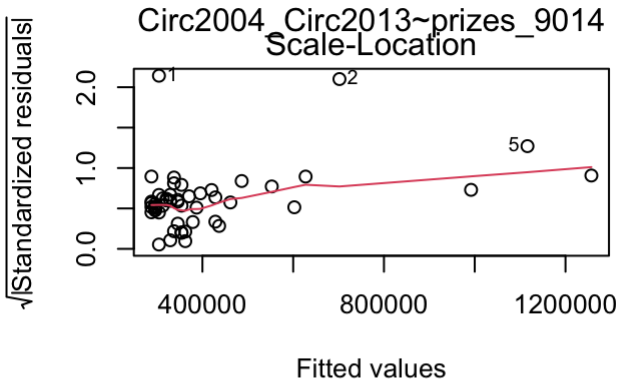
There is no clear trend in our data as we move from left to right that an increase in fitted values increases spread in residuals. Additionally, using the studentized Breusch-Pagan test the corresponding p-value is 0.9873. Since the p-value > 0.05, we cannot reject the null hypothesis that heteroscedasticity is not present. We have sufficient evidence to say that the assumption of constant spread is justified.

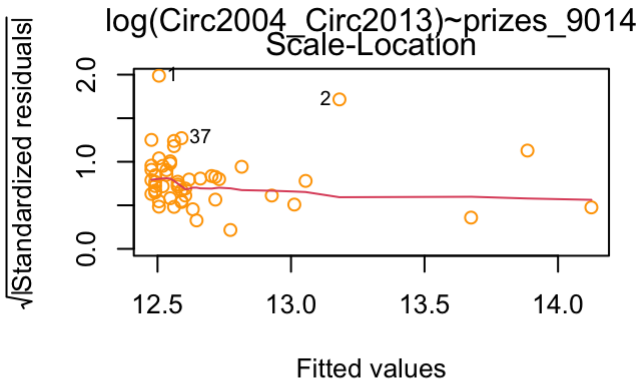
```
plot(df_log_lm, which =3)
```

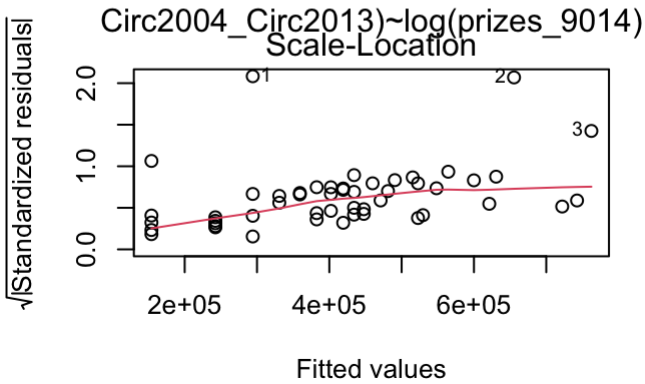


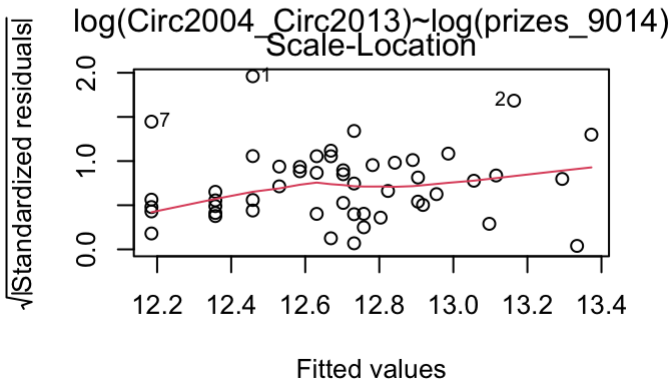
It is evident that heteroscedasticity is prevalent in the other models.

```
par(mfrow=c(2,2))
plot(lm(Circ2004_Circ2013~prizes_9014,data=df), which = 3)
mtext("Circ2004_Circ2013~prizes_9014",side=3, line=1, cex=1, adj=0.5)
plot(lm(log(Circ2004_Circ2013)~prizes_9014,data=df), which = 3,col = 'orange')
mtext("log(Circ2004_Circ2013)~prizes_9014", side=3, line=1, cex=1, adj=0.5)
plot(lm(Circ2004_Circ2013~log(prizes_9014),data=df), which = 3)
mtext("Circ2004_Circ2013)~log(prizes_9014)", side=3, line=1, cex=1, adj=0.5)
plot(lm(log(Circ2004_Circ2013)~log(prizes_9014),data=df), which = 3)
mtext("log(Circ2004_Circ2013)~log(prizes_9014)", side=3, line=1, cex=1, adj=0.5)
```









Breusch-Pagan test (used as justification but may be affected by slight non-normality of observations)

```
bptest(df_log_lm)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  df_log_lm
## BP = 0.00025223, df = 1, p-value = 0.9873
```

Independence

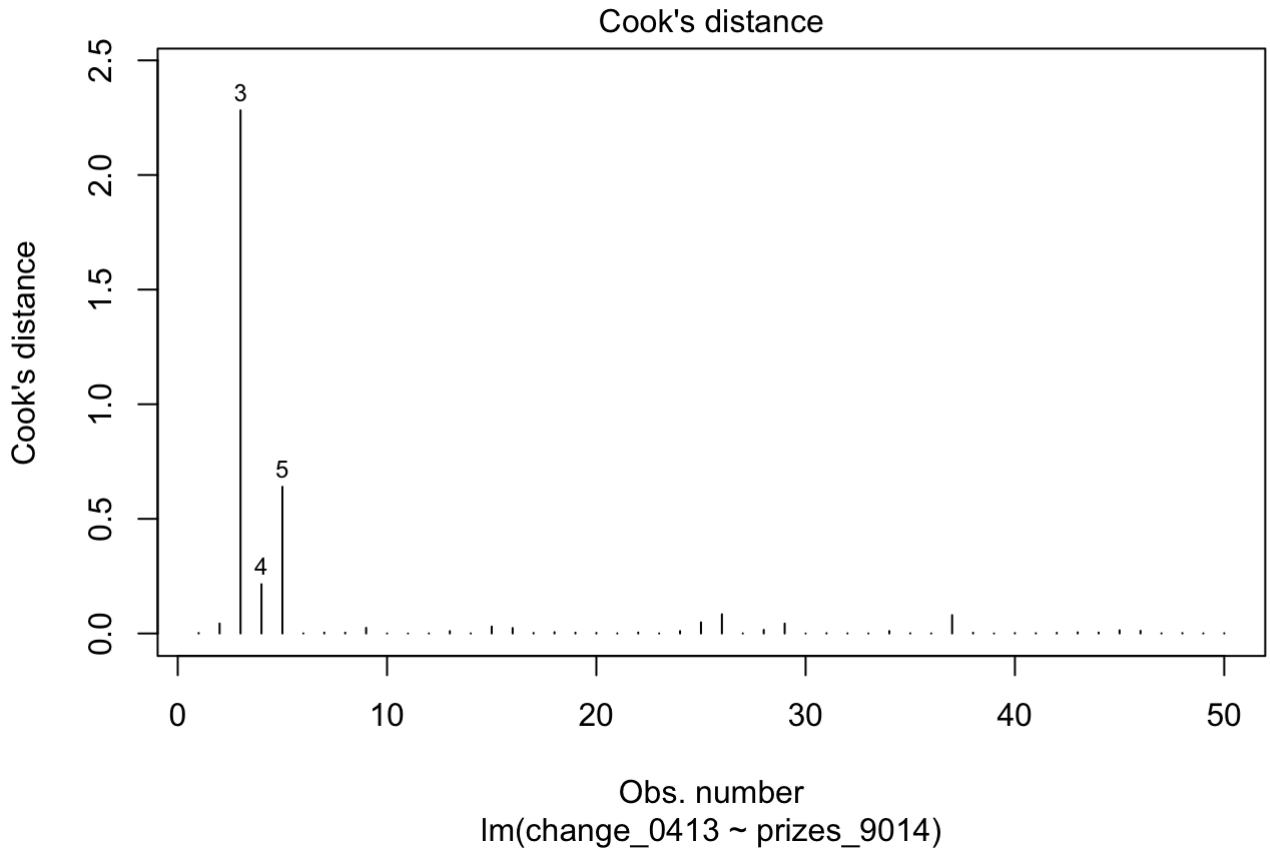
Notably, independence can only be assessed by looking at how the data was obtained. Information about the sampling process is not provided so it is difficult to make a judgement.

Each model seems to require further investigation under certain model assumptions. However, our selected model satisfies a majoirty of these linear regression assumptions and explains a greater amount of variance in our dependent varialble.

Model Two

Notably, observation 3 appear in every plot and according to Cook's distance rule of thumb exceeding 4/(n-p-1), this observation has high influence on the regression analysis specifically our graphs.

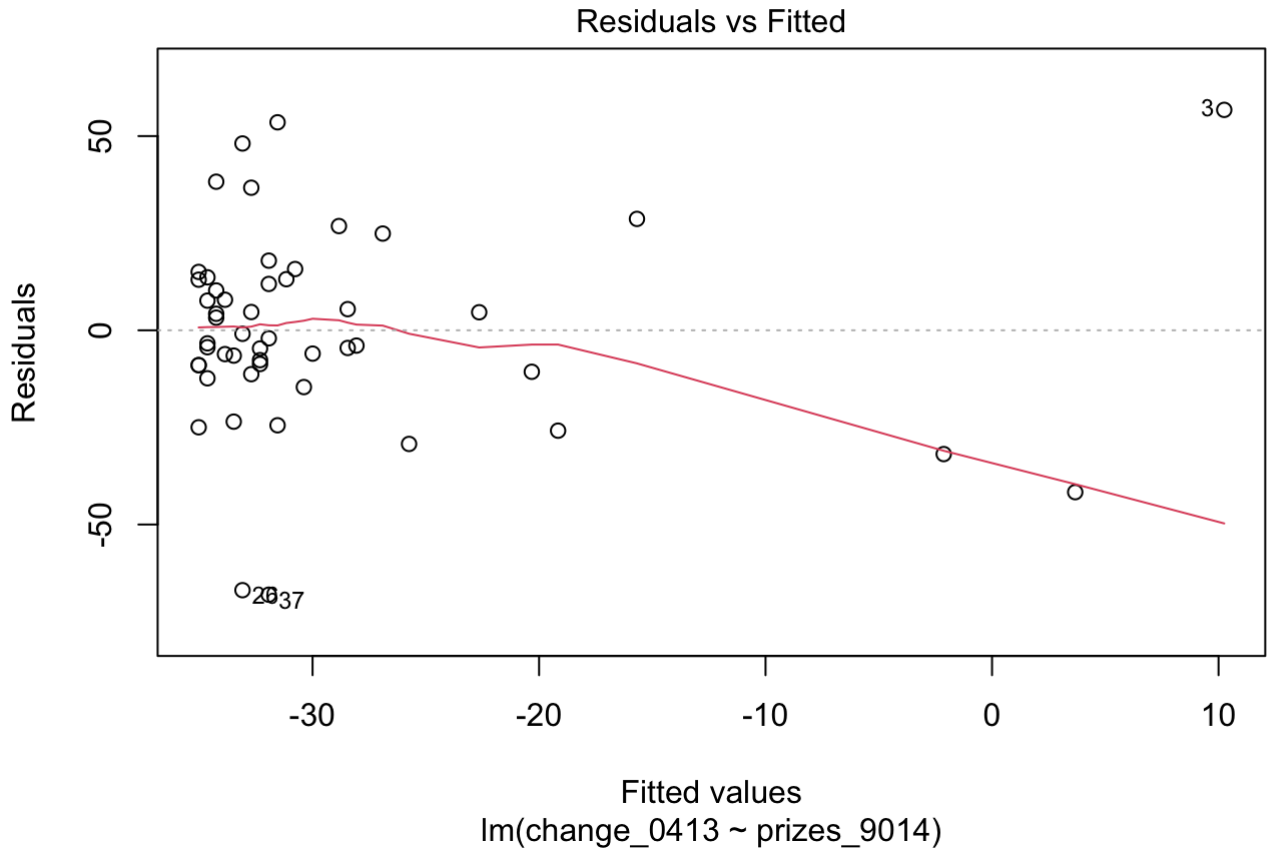
```
plot(df_change_lm, which =4)
```



Linear Relationship

Looking at the plot the points indicate that there is certainly not a strong relationship, but the small size of the observations seems to be affecting the reference line. There is clearly some trend on the red reference line towards the higher observations, but that can be attributed based upon just a few points towards the end. Arguably, the assumption of linearity is justified.

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



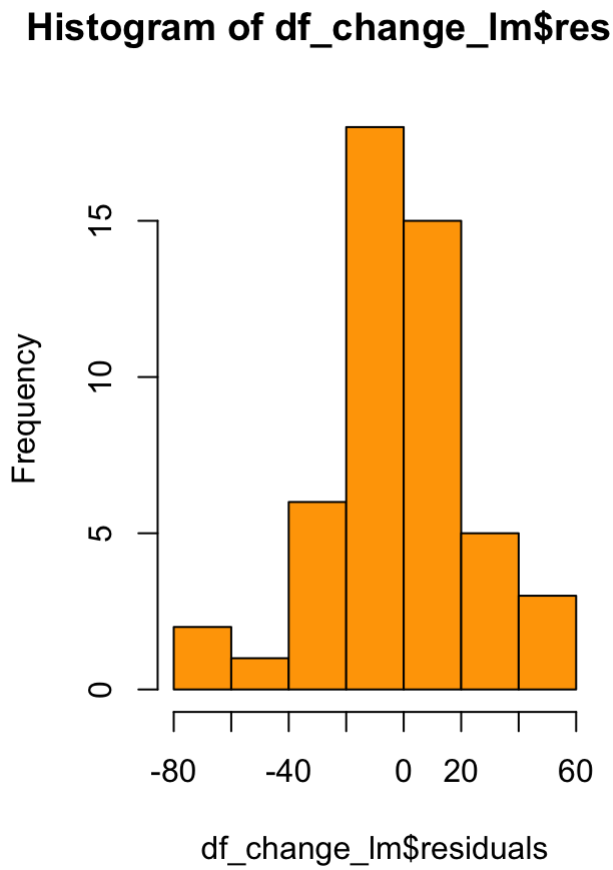

```
cor(df$change_0413, df$prizes_9014,use="complete.obs")
```

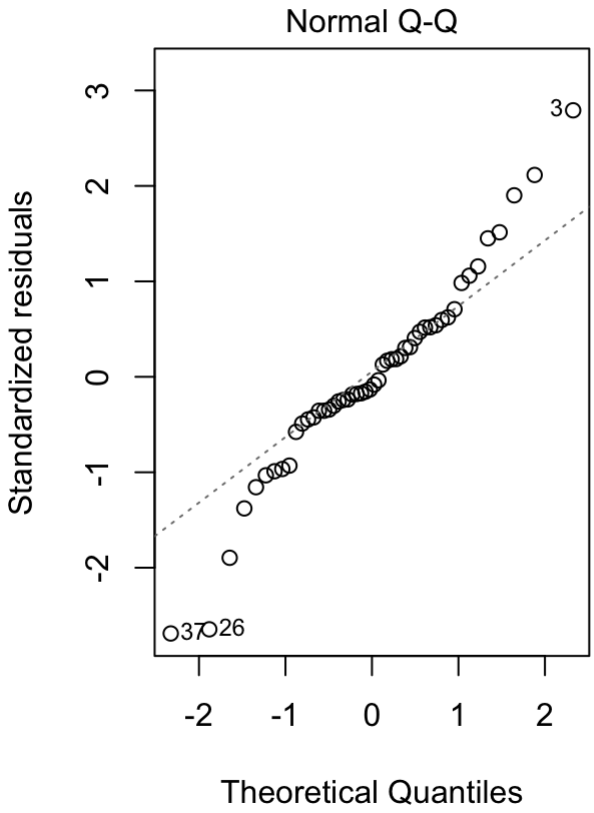
```
## [1] 0.3523113
```

Noise Normally Distribution

The normal quantile-quantile plot of the residuals and residuals histogram vaguely follow the similar shape as a heavy-skewed distribution. However, the fact that points that lie less than -1 and greater than 1 on the x-axis drift away from the line is somewhat acceptable because it is minimal (barring three labelled points) and it is the bulk portion of our data that lie between -1 and 1 that concerns us the most. Therefore, we can conclude that the residuals are normally distributed.

```
par(mfrow=c(1,2))
hist(df_change_lm$residuals, col = 'orange')
plot(df_change_lm, which =2)
```

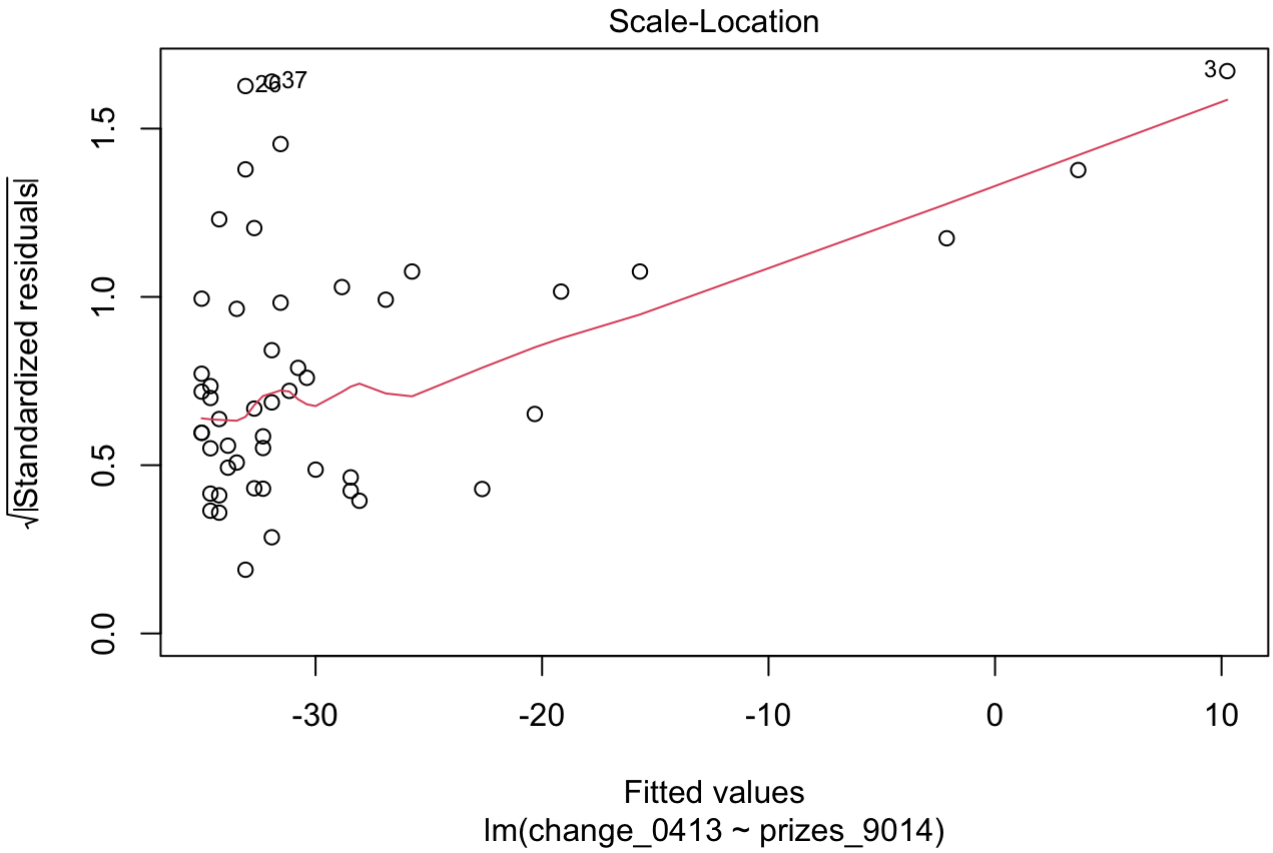




Homoscedasticity

An argument can be made that initially the bulk portion of our observations illustrate peaks and valleys with no apparent trend. However, there is a clear trend as we move from left to right. As fitted values increase so too does spread in residuals. Additionally, using the studentized Breusch-Pagan test the corresponding p-value is 0.02623. Since the p-value < 0.05, we reject the null hypothesis that heteroscedasticity is not present. We have sufficient evidence to say that heteroscedasticity is present in the regression model.

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



Independence

Notably, independence can only be assessed by looking at how the data was obtained. Information about the sampling process is not provided so it is difficult to make a judgement.

Question 4

4.1 Calculate the expected circulation of the newspaper under each of the three proposed strategic directions

Invest Substantially less

```
exp(predict(df_log_lm, newdata = tibble(prizes_9014=3)))
```

```
##        1
## 269788.1
```

On average we can predict average circulation would drop to 269,788 from the current circulation of 453,869. This equates to a percentage drop in circulation of 40.56% (-184,081)

Invest Same amount

```
exp(predict(df_log_lm, newdata = tibble(prizes_9014=25)))
```

```
##        1
## 367773.8
```

On average we can predict average circulation would drop to 367,773.8 from the current circulation of 453,869. This equates to a percentage drop in circulation of 18.96% (-86,095)

Invest Substanitally more

```
exp(predict(df_log_lm, newdata = tibble(prizes_9014=50)))
```

```
##        1
## 522983.1
```

On average we can predict average circulation would increase to 522,983 from the current circulation of 453,869. This equates to a percentage uptick in circulation of 15.23% (69,114)

4.2 change in circulation under each of the three proposed strategic directions.

Invest Substantially less

```
predict(df_change_lm, newdata = tibble(prizes_9014=3))
```

```
##         1
## -34.25423
```

On average we can predict that investing substantially less in investigative journalism resulting in 3 Pulitzer Prizes being awarded will result in percentage change in average circulation to drop 34.25% resulting in a predicted average circulation value of 298,399. This equates to a 6.3% and 28,611 difference in our two models which shows signs of inconsistency.

Invest Same amount

```
predict(df_change_lm, newdata = tibble(prizes_9014=25))
```

```
##         1
## -25.74021
```

On average we can predict that investing the same amount in investigative journalism resulting in 25 Pulitzer Prizes being awarded will result in a drop of percentage change in average circulation of 25.74% resulting in a predicted average circulation value of 337,042. This equates to a 6.78% and 30,732 difference in our two models which shows signs of inconsistency.

Invest Substanitally more

```
predict(df_change_lm, newdata = tibble(prizes_9014=50))
```

```
##         1
## -16.06518
```

On average we can predict that investing the substantially more in investigative journalism resulting in 50 Pulitzer Prizes being awarded will result in a drop of percentage change in average circulation of 16.07% resulting in a predicted average circulation value of 380,954. This equates to a 31.29% and 142,029 discrepancy between our two models. Therefore, there are clear signs that our two models are inconsistent.

4.3 Using the model from Question 3(a), calculate 90% confidence intervals

Looking at the table, only two outputs satisfy our condition of circulation being in excess of current circulation. According to our model, only investing substantially more in journalism yields fitted and upper confidence values greater than our current circulation. Interestingly, investing the same amount in journalism will not yield desired results indicating the only way to increase circulation is to increase investments in journalism.

```
one_CI <- exp(predict(df_log_lm, newdata = tibble(prizes_9014=3),
                      interval= 'confidence', level = 0.9))
two_CI <- exp(predict(df_log_lm, newdata = tibble(prizes_9014=25),
                      interval= 'confidence',level = 0.9))
three_CI <- exp(predict(df_log_lm, newdata = tibble(prizes_9014=50),
                        interval= 'confidence',level = 0.9))

Confidence_table <- tribble(
  ~Intervals,~less, ~Same, ~more,
  'lower', one_CI[2], two_CI[2], three_CI[2],
  'fit', one_CI[1], two_CI[1], three_CI[1],
  'upper', one_CI[3], two_CI[3], three_CI[3],
)

Confidence_table <- Confidence_table %>%
  mutate('Substantially_less' = format(less, big.mark=","))%>%
  mutate('Same_amount' = format(Same, big.mark=","))%>%
  mutate('Substantially_more' = format(more, big.mark=","))

Confidence_table <- Confidence_table%>%
  select(c(Intervals,Substantially_less,Same_amount,Substantially_more ))

knitr::kable(Confidence_table)
```

Intervals	Substantially_less	Same_amount	Substantially_more
lower	235,515.2	323,727.6	425,949.0
fit	269,788.1	367,773.8	522,983.1
upper	309,048.6	417,813.0	642,122.4

4.4 Using the model from Question 3(b) calculate 90% prediction intervals for the expected change in circulation

Prediction intervals are wider because they provide a range for a single observation that meets the Pulitzer Price requirement (2,25,50) and as such contain more uncertainty than confidence intervals. Notably, all three instances have a range (lower to upper) of 86% which is a significant amount and not ideal for our

investigation as there is too much uncertainty in these predicted values. Interestingly, all upper values yield positive changes in circulation but the downside (lower) is heavily skewed. This seems to contradict our previous model which stated only investing substantially more in journalism yields fitted and upper confidence values greater than our current circulation. Now our prediction table seems to state only if we achive the upper bound of our prediction interval will we increase our % change in circulation.

```
one_PI <- predict(df_change_lm, newdata = tibble(prizes_9014=3),
                  interval= 'prediction',level = 0.9)
two_PI <- predict(df_change_lm, newdata = tibble(prizes_9014=25),
                  interval= 'prediction',level = 0.9)
three_PI <- predict(df_change_lm, newdata = tibble(prizes_9014=50),
                    interval= 'prediction',level = 0.9)

new_prediction <- tribble(
  ~Intervals,~Substantially_less, ~Same_amount, ~Substantially_more,
  'lower', one_PI[2], two_PI[2], three_PI[2],
  'fit', one_PI[1], two_PI[1], three_PI[1],
  'upper', one_PI[3], two_PI[3], three_PI[3],
)


new_prediction <- new_prediction%>%
  select(c(Intervals,Substantially_less,Same_amount,Substantially_more ))

knitr::kable(new_prediction)
```

Intervals	Substantially_less	Same_amount	Substantially_more
lower	-77.729707	-69.15107	-60.23418
fit	-34.254234	-25.74021	-16.06518
upper	9.221238	17.67065	28.10381

Question 5

With the ‘Garbage in Garbage Out’ principle in mind, the standard principle that more training data (especially more quality data) leads to better models is applicable as our model is only trained on 50 observations. I would argue that building a linear regression model based upon only 50 observations is insufficient especially with the highly influential outliers contained in our observations. Both models suffer from influential outliers. Additionally, in practice, straight linear regression without any regularization (Ridge) would seldomly be used as our model would not generalize well (high variance from overfitting).

Model two would be insufficient for application because Pulitzer Prize only explains 12.4% of the proportion of variance in the model’s dependent variable (% change in circulation). Therefore, based upon an r-squared value of 0.124 the goodness of fit for this linear model is weak. The same limitation can be stated for Model One with an r-squared value of 0.325. Time serious modelling would be ideal but we are limited with our data and ability to also investigation auto-correlation further.