# Assessment 2 - Scaffolded case study

Email response:

---

Lachlan,

Regarding the data in the Pulitzer.csv file you asked me to work on. All the details are in the appendix below, but to summaries what I've found:

- Using the average circulation model (pul_model) and the accompanying prediction intervals, the more prizes you win the larger the average circulation will be. But it shows that our current rate of award winning will not be enough to maintain our current circulation, we will need to almost double it.

$$avg\ \ circ = 12.463142 + 0.014083(Pulitzer\ prizes\ won) + \epsilon_i$$

- I would not recommend using the percentage change model (pul_model2) to predict how prizes influence percentage changes because although significance was found, it took some convincing. The assumption of linearity, homoscedasticity, and normal noise distribution were not well verified. Furthermore, the prediction intervals suggest we cannot use the model with any certainty.

$$\ln(\%\ change) = -35.4152 + 0.3870(Pulitzer\ prizes\ won) + \epsilon_i$$

- The average circulation model (pul_model) would indicate that investing substantially more in investigative journalism is the way to go.

**<u>However,</u>**

There are some fundamental issues with using this data that become apparent when considering the assumption of error independence and how the model applies to the real world, see appendix for a more in-depth response.

- **Customer dynamics** resulting from the introduction or loss of competition relates values to a degree as customer proportions change accordingly.
- **Contents/ marketing** relate values because there is a finite pool of readers. A big news story will be covered in varying levels of detail in each paper which influences a customer's choice.
- **The Reader to paper ratio** marks a dependency attached to individuals that buy multiple papers if/ when they get the paper. So, paper A will increase with paper B.
- **Readership and Distribution** will influence subject values and should be represented with a coefficient itself. The business reasonably available sets the maximum possible circulation.

- **The awards available** will relate values because there are only 24 Pulitzer Prizes awarded each year so when I take one, that leaves n-1 Pulitzer Prizes.

The models also have a few limitations:

- **Technological advancement** has influenced circulation and will continue to do so in ways that the model doesn't currently account for.
- **Public Awareness** of Pulitzer Prizes may be overestimated. People generally don't choose the paper based on award counts; they choose a paper based on good journalism.
- **The measure of good journalism** is quite poor in this model because it assumes an unchanging work force and that prize worthy work is guaranteed an award, but we can gain and loose talented writers who may or may not win prizes. The cohorts average work may still be prize worthy. I suggest an alternative measure in the Appendix (question 3.3 and 5).
- **The news needs** of society is shifting. Covid-19 is an excellent example of that. People needed accurate, real-time information on hot spots, infection totals, restriction announcements, etc. and they needed it available at home. The general niche that newspapers once dominated is diminishing.
- **Limited data** of just two years over a decade is a very poor representation of average changes for that period.

***Continue for Appendix***

# Appendix:

## Question One: Reading and Cleaning

### 1.1

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

```
pulitzer <- pulitzer%>%
  mutate(perc_change = str_replace_all(pulitzer$change_0413, "%", ""))
#remove % sign from change_0413 column
pulitzer$perc_change <- as.integer(pulitzer$perc_change)
#make percentage change an integer instead of a string
pulitzer <- pulitzer[, c(1,2,3,5,6)]
#Remove duplicate column
head(pulitzer)
```

```
## # A tibble: 6 x 5
##   newspaper           circ_2004 circ_2013 prizes_9014 perc_change
##   <chr>                   <dbl>     <dbl>       <dbl>       <int>
## 1 USA Today             2192098   1674306           3         -24
## 2 Wall Street Journal   2101017   2378827          51          13
## 3 New York Times        1119027   1865318         118          67
## 4 Los Angeles Times      983727    653868          86         -34
## 5 Washington Post        760034    474767         101         -38
## 6 New York Daily News    712671    516165           7         -28
```

### 1.2

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

```
colnames(pulitzer)
```

```
## [1] "newspaper"   "circ_2004"   "circ_2013"   "prizes_9014" "perc_change"
```

```
pulitzer$average_of_04_and_13 <- rowMeans(pulitzer[ , c(2,3)], na.rm=TRUE)
#Introducing a column for the average circulation over the decade
pulitzer <- pulitzer[, c(1,4,2,3,5,6)]
head(pulitzer)
```

```
## # A tibble: 6 x 6
##   newspaper       prizes_9014 circ_2004 circ_2013 perc_change average_of_04_and~
##   <chr>                 <dbl>     <dbl>     <dbl>       <int>              <dbl>
## 1 USA Today                 3   2192098   1674306         -24            1933202
## 2 Wall Street Jo~          51   2101017   2378827          13            2239922
## 3 New York Times          118   1119027   1865318          67            1492172.
## 4 Los Angeles Ti~          86    983727    653868         -34             818798.
## 5 Washington Post         101    760034    474767         -38             617400.
## 6 New York Daily~           7    712671    516165         -28             614418
```

## Question Two: Univariate Summary and Transformation

### 2.1

Describe the distribution of the variable representing average circulation, including shape, location, spread and outliers.

```
ggplot(pulitzer, aes(average_of_04_and_13))+
  geom_histogram()+
  labs( title = "Right Skewed Histogram of the Average Circulation \nbetween
2004 and 2013",
        x = "Average Circulation",
        y = "Count")
```

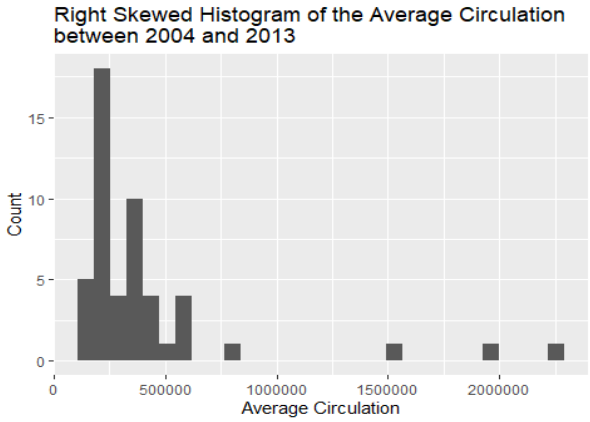


*Figure 1: Histogram of average circulation from 2004 and 2013 against the count, or frequency, of that value.*

```
ggplot(pulitzer, aes(average_of_04_and_13))+
  geom_boxplot()+
  labs( title = "Boxplot Demonstrating Spread and Outlier Positions \nfor the
Average Circulation between 2004 and 2013", x = "Average Circulation")
```

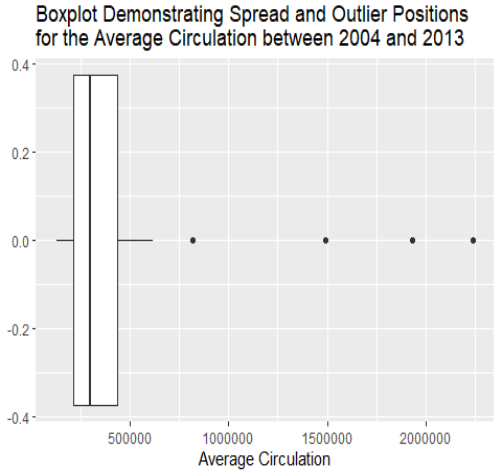


*Figure 2: Boxplot of the average circulation between 2004 and 2013.*

```
summary(pulitzer$average_of_04_and_13, na.rm = TRUE)

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

sd(pulitzer$average_of_04_and_13)

## [1] 410339.9
```

Average circulation has a **mean** of 412442, **standard deviation** of 410340, an **interquartile range**, IQR, of 222643, a **median** of 298851, and a **domain** of [131004, 2239922].

It is **right skewed** with **4 outliers**: The Wall Street journal, USA Today, New York Times, and Los Angeles Times *(figure 1 and 2)*. This indicates that the majority of newspapers reach a circulation of less than 600,000. These publishers are likely state specific, topic specific, or cannot penetrate the national market.

### 2.2

Describe the distribution of perc_change, including shape, location, spread and outliers.

```
ggplot(pulitzer, aes(perc_change))+
  geom_histogram() +
  labs( title = "Histogram Demonstrating Spread of the Percentage Change in C
irculation \nbetween 2004 and 2013", x = "Percentage Change of Circulation",
y = "Count"
```

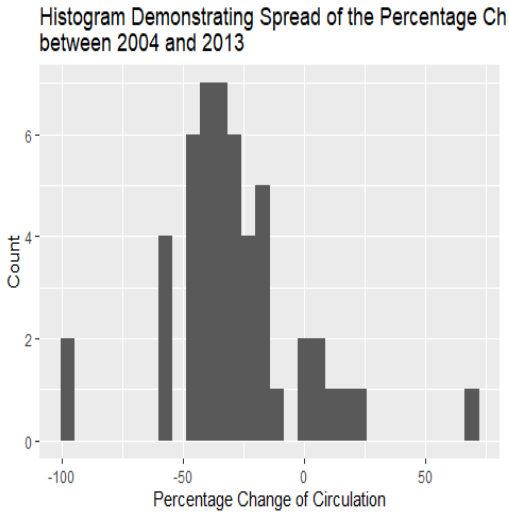


*Figure 3: A histogram of the percentage change in circulation between 2004 and 2013 showing the count, or frequency, of each value.*

```
ggplot(pulitzer, aes(perc_change))+
  geom_boxplot() +
  labs( title = "Boxplot Demonstrating Spread and Outlier Positions \nfor the
```

```
Percentage change in Circulation between 2004 and 2013", x = "Percentage Chan
ge")
```

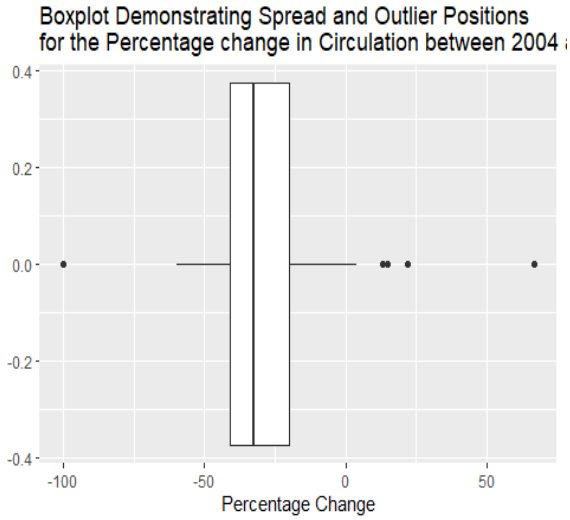


*Figure 4: Boxplot of the average circulation between 2004 and 2013.*

```
summary(pulitzer$perc_change)

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

unique(pulitzer$perc_change)

sd(pulitzer$perc_change)

## [1] 27.06681

#-------------#
#What about removing the publishers that appear to be out of print, surely th
eir data distorts the model too much?
```

The histogram is a little bit messy but it's vaguely **centered, and multimodal** *(figure 3)*. There are **5 outliers**; Rocky Mountain News, New Orleans Times-Picayune, Boston Herald, Detroit News, and San Francisco Chronicle *(Figure 2)*. The Rocky Mountain News and New Orleans Times-Picayune both appear to be out of print with a percentage change of -100%. The data has a **mean** of -29.2%, **IQR** of 20.75%, **domain** of [-100, 67]. The **median** is -32.50%, with three instances each, the data is **multimodal** at -34%, -40%, and -44%.

When you remove those two data that represent two publishers that appear to be out of print from the population you have improved significance in later sections. I think there is a strong case for removing them because they misrepresent themselves; they did not go out of print on the final day of the decade nor did they have a whole 25 years to obtain Pulitzer Prizes. Furthermore, papers distributed nationally also cause trouble, but the case for their removal requires a significant amount of research, more than is necessary for this assignment. To justify removing them you would need to show that there is a significant

difference between papers with national circulation and those with a more limited reach or niche. I think it's also possible that when those two publishers went out of print, it drastically increased the amount of readership in other papers that could fill that gap in the market.

I introduce a data set for a third model below that excludes publishing companies that are out of print just to see what would happen:

```
pul2 <- pulitzer %>%
   filter( perc_change != "-100")
#removes publishers that are out of print
ggplot(pul2, aes(perc_change))+
  geom_histogram() +
  labs( title = "Histogram Demonstrating Spread and Outlier Positions for the
percentage change \nbetween 2004 and 2013 when out-of-print publishers are ex
cluded", x = "Percentage Change", y =  "Count")
```

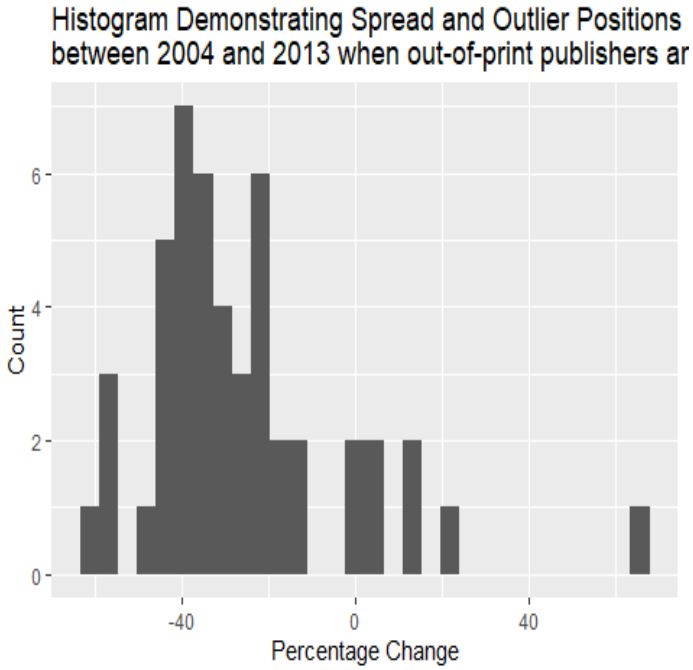


*Figure 5: A histogram of the percentage change in circulation between 2004 and 2013 showing the count, or frequency, of each value excluding subjects with changes of -100%.*

```
ggplot(pul2, aes(perc_change))+
  geom_boxplot() +
  labs( title = "Boxplot Demonstrating Spread and Outlier Positions for the A
verage Circulation\nbetween 2004 and 2013 when out-of-print publishers are ex
cluded", x = "Percentage Change")
```

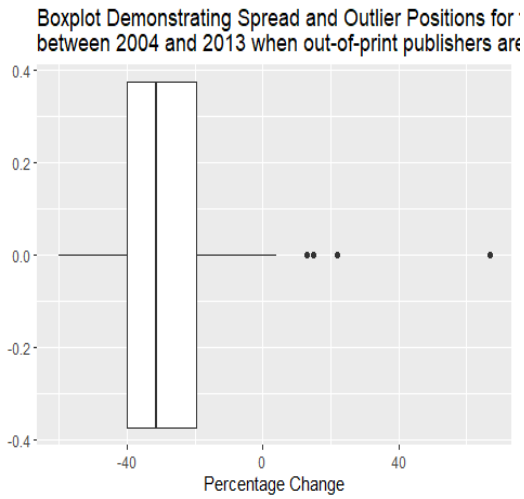


*Figure 6: Boxplot of the average circulation between 2004 and 2013 excluding subjects with values of -100%.*

```
summary(pul2$perc_change)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -60.00  -40.00  -31.50  -26.25  -19.50   67.00

sd(pul2$perc_change)

## [1] 23.27221
```

Now the graph, appears more right-skewed. You can also see how some of the publishers increase circulation by upwards of 10% while the majority reduced circulation. These are the four outliers: the New York Times, Denver Post, Orange County Register, and the Wall Street Journal. The mean is -26.25%, IQR of 20.5%, and a domain of [-60, 67].

### 2.3

Do either of perc_change and the variable representing average circulation have a skew that could be resolved by a log transform? For each variable, select whether it should be transformed.

```
#Right skew can be rectified with a log transformation but lets use BoxCoxTra
ns() to check it out.
pacman::p_load(caret)
BoxCoxTrans(pulitzer$prizes_9014, pulitzer$average_of_04_and_13)

## Estimated Lambda: 0
## With fudge factor, Lambda = 0 will be used for transformations

#lambda = 0, suggesting a ln(x+60) transformation
BoxCoxTrans(pulitzer$prizes_9014, pulitzer$perc_change)

## Estimated Lambda: -0.1
## With fudge factor, Lambda = 0 will be used for transformations
```

```
#lambda = 0, suggesting a ln(x+60) transformation
ggplot(pulitzer, aes(log(average_of_04_and_13)))+
  geom_histogram()

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

#Appears more normally distributed
ggplot(pulitzer, aes(log(perc_change+101)))+
         geom_histogram()

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

#accounted for x<=0, the distribution is impaired by the value of x=-100%.
ggplot(pul2, aes(log(perc_change+61)))+
         geom_histogram()

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

#accounted for x<=0, appears to be a near mirror image of the initial un-tran
sformed histogram, probably not helpful
#Now do they look more similar to our distribution of prizes_9014:
ggplot(pulitzer, aes(prizes_9014))+
  geom_histogram()

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

#heavy right skew, nothing lower than 0 so transform by log(x+1).
ggplot(pulitzer, aes(log(prizes_9014+1)))+
  geom_histogram()

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

ggplot(pulitzer, aes(prizes_9014))+
  geom_histogram()

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

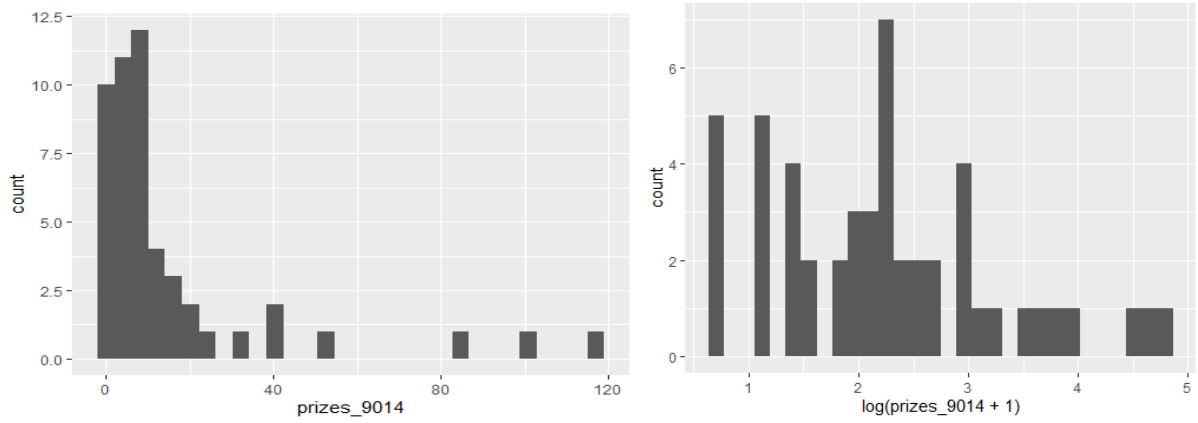


Figure 7: Histograms showing the effects of transforming the prizes by the natural log.

```
#better distribution with the transformation.
```

The average_of_04_and_13 is right skewed and would benefit from a log transformation. As the data doesn't include any values of zero or less, it is sufficient to transform with the default ln(x).

The perc_change attribute is also right skewed but transforming by ln(x+101) did not alter the distribution in a way that could increase accuracy or predictive power. The same is true when considering the data of the pul2 model which, when transformed by ln(x+61), seemed to produce a distribution that almost mirrored the untransformed version.

To be clear, going forward:

average circulation —> transformation of ln(x)

perc_change —> leave as

## Question Three: Model building and interpretation

### 3.1

Build a model predicting the variable representing a newspaper's circulation using prizes_9014, incorporating a log transform for the average circulation if you decided this was necessary. State and interpret the slope and intercept of this model in context. Is there a statistically significant relationship between the number of Pulitzer Prizes, and average circulation? (Include this model in your email attachment, and include your interpretation in your email.)

```
pul_model <- lm(log(average_of_04_and_13) ~ prizes_9014, data = pulitzer)
summary(pul_model)
```

```
##
## Call:
## lm(formula = log(average_of_04_and_13) ~ prizes_9014, data = pulitzer)
##
## 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
```

```
#Just a quick look:
ggplot(pulitzer, aes(log(average_of_04_and_13), prizes_9014)) +
  geom_point()+
  labs( title = "A Visualization of the log Transformed \nAverage against the
Prize Count Attribute",
        x= "ln(Average Circulation between 2004 and 2013)",
        y = "Prize count")
```

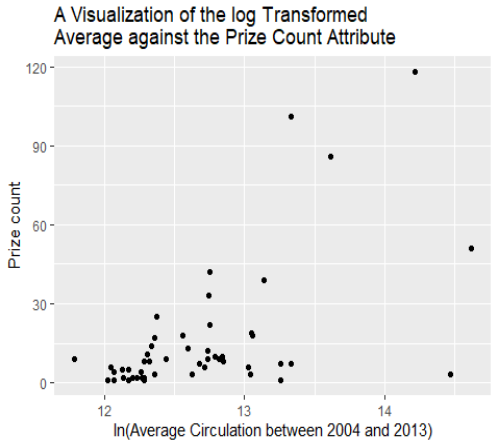


*Figure 8: A scatterplot of the natural log of average circulation from 2004 and 2013 against the Pulitzer prizes won by each subject over the last 25 years.*

```
plot(log(pulitzer$average_of_04_and_13))

plot(pulitzer$prizes_9014)

#Plots didn't reveal anything that we didn't already know
#now with pul2 data:
pul_model_without_outliers <- lm(log(average_of_04_and_13) ~ (prizes_9014), d
ata = pul2)
summary(pul_model_without_outliers)

##
## Call:
## lm(formula = log(average_of_04_and_13) ~ (prizes_9014), data = pul2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.5475 -0.3311 -0.1813  0.1669  1.9369
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.49669    0.08574 145.754  < 2e-16 ***
## prizes_9014  0.01370    0.00288   4.758 1.97e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4955 on 46 degrees of freedom
```

```
## Multiple R-squared:  0.3298, Adjusted R-squared:  0.3152
## F-statistic: 22.63 on 1 and 46 DF,  p-value: 1.973e-05

#Doesn't seem to improve the model in any meaningful way.
```

Generally, you want your distribution of y values to resemble your distribution of x values for better output. So, prizes_9014 will remain un-transformed to match the right skew that persists in the 'ln(average_of_04_and_13)'. (Interestingly this model has the best R^2 values which are measures of correlation and should not be taken to mean the model's predictive potential has improved, but it does indicate a stronger relationship.)

The intercept was at 12.463142, and the coefficient of the prizes_9014 was 0.014083. There does appear to be a significant relationship here with a p-value of 2x10^-16 for the intercept and 1.53x10^-5 for the coefficient of prizes_9014; well below 0.05.

In short, for every prize won, the average circulation will increase by (e^0.014083=) 1.

Linear regression equation: $y_i = \beta_0 + \beta_1 x_i + +\epsilon_i$

Pul_model eqn: $avg\ \ circ = 12.463142 + 0.014083(Pulitzer\ prizes\ won) + \epsilon_i$

### 3.2

Build a model predicting perc_change using prizes_9014, incorporating a log transform for perc_change if you decided this was necessary. State and interpret the slope and intercept of this model in context. Is there a statistically significant relationship between the number of Pulitzer Prizes, and change in circulation? (Include this model in your email attachment, and include your interpretation in your email.)

```
pul_model2 <- lm(perc_change ~ prizes_9014, data = pulitzer)
summary(pul_model2)

##
## Call:
## lm(formula = perc_change ~ prizes_9014, data = pulitzer)
##
## 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(pulitzer$perc_change)

plot(pulitzer$prizes_9014)

#log function on prizes_9014 was not significant.
```

The BoxCoxTrans() function suggested a transformation by the natural log. Making that transforamtion does however result in a p-value of 0.05185 which is of course above the acceptable significance threshold of 0.05. Without that transformation, pul_model2 is significant, with a p-value of 0.0121; which is below the acceptable limit of 0.05 (intercept's p-value is 1.21x10^-10, prize coefficient p-value is 0.0121). We'll use that model and see what comes of it going forward keeping in mind it misbehaves a little bit, the intercept is at -35.4152, and the coefficient of prizes_9014 is at 0.3870.

In short, for every prize won, the perc_change will increase by 0.3870%.

Pul_model2 eqn: $\ln(\%\, change) = -35.4152 + 0.3870(Pulitzer\ prizes\ won) + \epsilon_i$

```
#Model 3, from pul2 data (no -100% perc changes)
pul_model3 <- lm(perc_change ~ log(prizes_9014+1), data = pul2)
summary(pul_model3)

##
## Call:
## lm(formula = perc_change ~ log(prizes_9014 + 1), data = pul2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -35.83 -13.16  -4.44   9.65  75.68
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -41.636      7.540  -5.522  1.5e-06 ***
## log(prizes_9014 + 1)    6.896      3.056   2.257   0.0288 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.32 on 46 degrees of freedom
## Multiple R-squared:  0.09969,    Adjusted R-squared:  0.08011
## F-statistic: 5.093 on 1 and 46 DF,  p-value: 0.02881

#In the model that excludes the -100% percentage changes, pul_model3, we see
significance in the log transformed data. The p value of the intercept is 1.5
x10^6, and the p-value of the coefficient is 0.0288; both below 0.05.
#Intercept -> -41.636
#log(prizes_9014+1) -> 6.896
```

```
#The r^2 values are not excellent, and in the next section we see that the as
sumptions aren't well satisfied.
```

### 3.3

Check the assumptions of both linear models. Recall that there are four assumptions for each model. (Outline your assumptions in the email to your manager and include the assessment of these assumptions in the email attachment.)

```
plot(pul_model, which = 1)

plot(pul_model, which = 3)

plot(pul_model, which = 2)
```

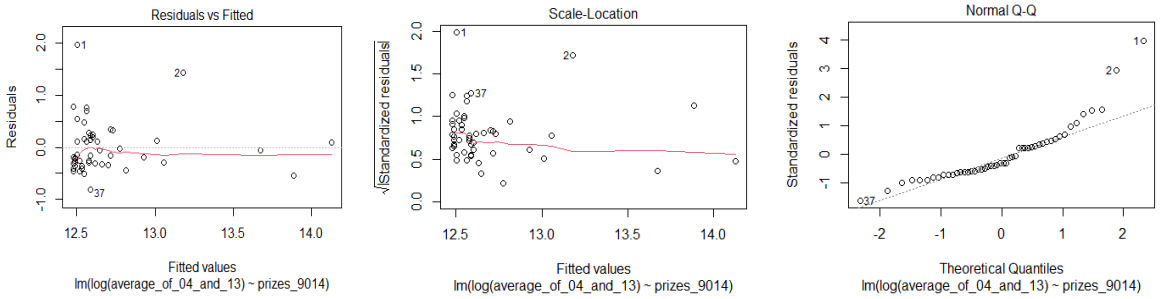


Figure 9, 10, and 11: From left to right we have a plot of fitted values against residuals to verify linearity, a plot of fitted values against the root or standardized residuals to verify homoscedasticity, and a plot of theoretical quantiles against standardized residuals to confirm noise is normally distributed for 'pul_model'.

```
plot(pul_model2, which = 1)

plot(pul_model2, which = 3)

plot(pul_model2, which = 2)
```

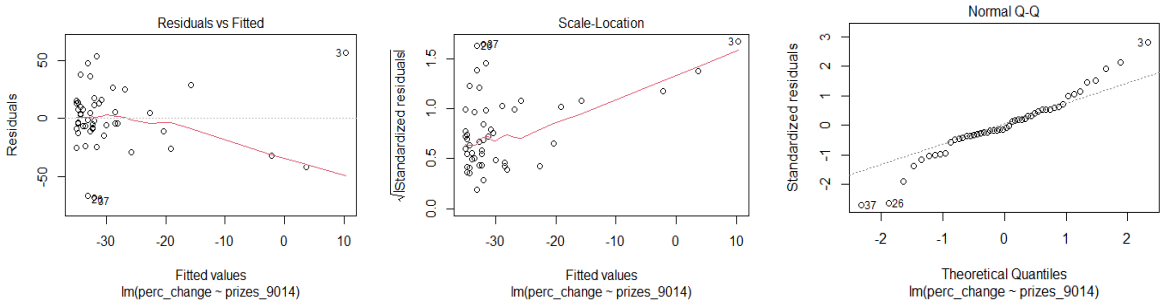


Figure 12, 13, and 14: From left to right we have a plot of fitted values against residuals to verify linearity, a plot of fitted values against the root or standardized residuals to verify

homoscedasticity, and a plot of theoretical quantiles against standardized residuals to confirm noise is normally distributed for 'pul_model2'.

```
#and for the model excluding out of print publishers:
plot(pul_model3, which = 1)

plot(pul_model3, which = 3)

plot(pul_model3, which = 2)
```

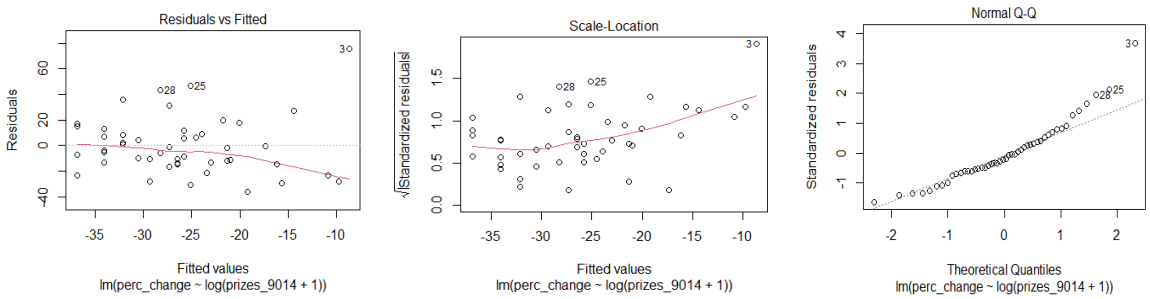


Figure 15, 16, and 17: From left to right we have a plot of fitted values against residuals to verify linearity, a plot of fitted values against the root or standardized residuals to verify homoscedasticity, and a plot of theoretical quantiles against standardized residuals to confirm noise is normally distributed for 'pul_model3'.

```
#the assumptions of linearity and homoscedasticity aren't well supported for
pul_model3.
ggplot(pulitzer, aes(log(average_of_04_and_13), prizes_9014))+
  geom_point()

#Quick look at the above plot doesn't reveal anything we hadn't already ascer
tained.
```

The assumptions being made are:

1. **linearity** - we assume that a straight line is the best model to use to represent the relationship between variables.

To confirm this assumption, we should see a fairly horizontal line on a plot of fitted values against residuals. For pul_model, there is good linearity considering the scale of the y-axis (*Figure 9*). While the curve looks worse in pul_model2, the bulk of the data behaves well and it looks to be only a few data points towards the end that cause the trouble (*Figure 12*). I favor the pul_model over the pul_model2 because the assumption of linearity appears more accurate, that doesn't immediately disqualify pul_model2 but we must keep this in mind while exploring these models further.

2. **homoscedasticity** - We assume that the noise in the data is consistent over the domain with the same variance.

For this to be true we should see a fairly horizontal line on a plot of fitted value against (standardized residuals) ^0.5. For pul_model, there is a fairly horizontal line over the domain of interest (*Figure 10)*. In pul_model2, the line quite clearly has a positive linear trend but again the bulk of the data shows no specific trend (*Figure 13*).

3. **A normal distribution of noise** - we assume the noise is normally distributed and doesn't increase or decrease across the domain.

We would expect linearity on the domain of [-2,2]. On the plots of theorized quantiles against standardized residuals, we clearly see that both models veer away from linearity inside that domain, therefore the assumption of normality is not well supported (*Figure 12 and 15*). However, it again seems to be only the outliers causing the movement from linearity.

4. **Independence** - We assume all error is independent of the variable and not influenced by it in any way.

I think there is a lot to unpack when it comes to the assumption of independence. The assumption we're making is that noise and error act separately to the values. So, the question is what kind of noise or error can there be around circulation numbers and prize counts? How could observations from one subject's values give us more information about another?

- Customer dynamics - First of all when one paper goes out of print, it's highly likely that their proportion of readers will move to a different publisher. Customer dynamics relate these values over the period with the publishers included and perhaps many that are not.

- Contents/ marketing - Another thing that relates these papers are their contents. For example, if I'm the type of person that goes to the news agent to buy my daily paper, occasionally the bigger publishers will convince me to buy their paper instead with a story so big and recent that only a firm of their size and with their resources could provide. When I see the front-page story about the volcanic eruptions in Antarctica, or the discovery of Atlantis, I'm likely to either put down my usual newspaper and buy that one or buy both papers. That's one way we may have dependence in the data which relates to contents and marketing. If the bigger paper has a better story, less smaller papers will sell. This will happen multiple times a year, clearly adding noise to the data.

- reader to paper ratio - Another way there may be dependence has increased over the last decade. One individual that buys multiple papers from different publishers whenever they get the paper. This has likely become more confounding with online newspapers. It is not clear to me whether or not online readers that pay for the membership are included in circulation numbers. But if they are, that certainly opens up room for noise as an individual that no longer has to pay 2.50 for the paper each morning can now afford a monthly subscription to multiple papers at a lower monthly cost. There is a dependency attached to these individuals because the

circulation of one will go up with the circulation of another just because the individual decides to get news access.

- Readership and Distribution - Using circulation as a measure of success will distort the model because it does not consider the population size. For example, I imagine the 'USA Today' is available in multiple states, but good luck finding the "Daily Oklahoma' anywhere but in Oklahoma. The reach some papers can get is limited for geographical or content reasons. I would be surprised if *'(circulation)/ (population of news readers in areas of distribution)*' didn't produce a more accurate model. After all, the Pulitzer Prize is awarded for excellence in newspaper journalism, not the size of the publishing company. We should compare our market share against our competitors, not against national papers that don't fill the same niche as we do here in Boston.
- Available awards - There are only 24 Pulitzer Prizes awarded each year so when I take one that leaves n-1 Pulitzer Prizes. In this way, an observation for paper A can give you information about paper B.

## Question Four: Prediction

Incorporate your answers from this question in your email to your manager. Masthead Media is considering three strategic directions for the Boston Sun-Times. These are:

- Investing substantially less in investigative journalism than present. In this case, Masthead Media projects that the newspaper will be awarded 3 Pulitzer Prizes in the next 25 years.
- Investing the same amount in investigative journalism than present, leading to the award of 25 Pulitzer Prizes in the next 25 years.
- Investing substantially more in investigative journalism, leading to the award of 50 Pulitzer Prizes.

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

### 4.1

Using the model from Question 3.1, calculate the expected circulation of the newspaper under each of the three proposed strategic directions. How does this compare with the current circulation?

```
#pul_model <- lm(log(average_of_04_and_13) ~ prizes_9014, data = pulitzer)
predict(pul_model, newdata= tibble(prizes_9014 =3))

##        1
## 12.50539

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

```
##        1
## 12.81522

predict(pul_model, newdata= tibble(prizes_9014 =50))

##       1
## 13.1673

ggplot(pulitzer, aes(prizes_9014, log(average_of_04_and_13)))+
  geom_point()+
  geom_smooth(aes(group=1), method="lm", se=FALSE, col = "black")+
  labs( title = "The Linear Regression Model produced by the pul_model data",
        x ="Prize count",
        y = " ln( Average circulation between 2004 and 2013)")
```

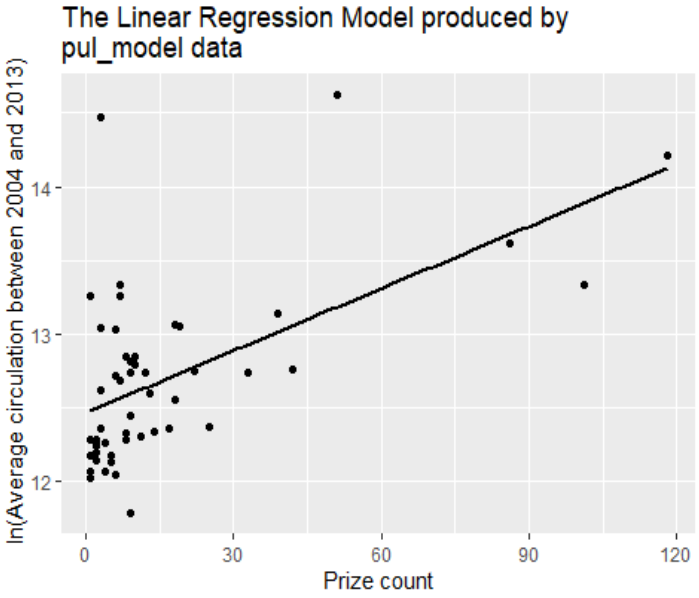


*Figure 18: The linear regression model produced by pul_model with prize count along the x - and the natural log of average circulation values along the y-axis.*

F(3) = log(average_of_04_and_13) = 12.50539, or a total average circulation of 269,788.
F(25) = log(average_of_04_and_13) = 12.81522, or a total average circulation of 367,773.
F(50) = log(average_of_04_and_13) = 13.1673 , or a total average circulation of 522,981.

Currently the circulation is 453,869, with a single Pulitzer Prize awarded every year; or 25 over 25 years. The model indicates that following that trend will be insufficient to maintain the current circulation average.

### 4.2

Using the model from Question 3(b), calculate the change in circulation of the newspaper, across the next decade, under each of the three proposed strategic directions. Comment on whether the projections of each of the two models are consistent.

```
#pul_model2 <- lm(perc_change ~ prizes_9014, data = pulitzer)
predict(pul_model2, newdata= tibble(prizes_9014 =3))

##         1
## -34.25423

predict(pul_model2, newdata= tibble(prizes_9014 =25))

##         1
## -25.74021

predict(pul_model2, newdata= tibble(prizes_9014 =50))

##         1
## -16.06518

ggplot(pulitzer, aes(prizes_9014, perc_change))+
  geom_point()+
  geom_smooth(aes(group=1), method="lm", se=FALSE, col = "black") +
  labs( title = "The Linear Regression Model produced by the pul_model2 data"
, x ="Prize count", y = "Percenatge change in circulation from 2004 to 2013")
```

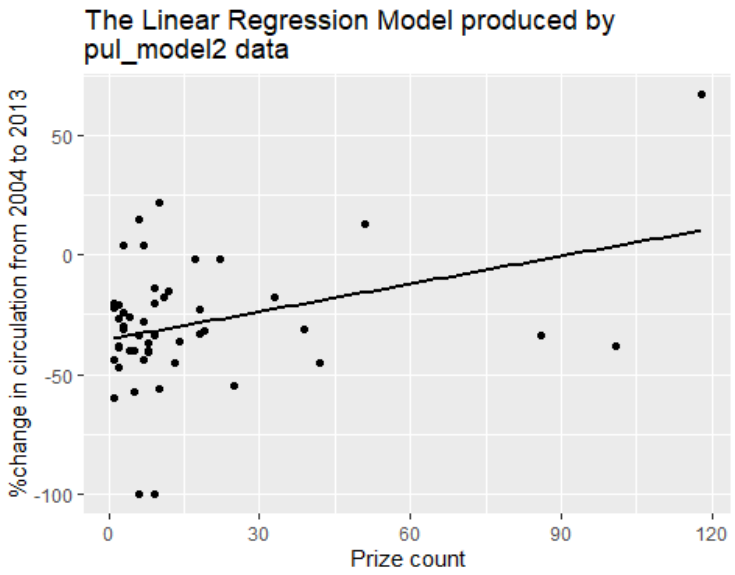


*Figure 19: The linear regression model produced by pul_model2 with prize count along the x - and the percentage change in circulation along the y-axis.*

F(3) = log(average_of_04_and_13) = -34 .25423, or a total average circulation of 1x10^-15

F(25) = log(average_of_04_and_13) = -25.74021, or a total average circulation of 6.6x10^-12

F(50) = log(average_of_04_and_13) = -16.06518, or a total average circulation of 1.1x10^-7.

The model is not consistent with the previous model. The coefficients are considerably different, so the outputs differ markedly. This model suggests around 90 Pulitzers need to be won over 25 years in order to prevent a reduction in circulation (*Figure 19*).

### 4.3

Using the model from Question 3(a), calculate 90% confidence intervals for the expected circulation of the newspaper under each of the three proposed strategic directions. Place these confidence intervals in a table and contrast them in context.

```
predict(pul_model, newdata= tibble(prizes_9014 =3), interval = "confidence")
```

```
##        fit      lwr      upr
## 1 12.50539 12.34252 12.66826
```

```
predict(pul_model, newdata= tibble(prizes_9014 =25), interval = "confidence")
```

```
##        fit     lwr      upr
## 1 12.81522 12.6623 12.96815
```

```
predict(pul_model, newdata= tibble(prizes_9014 =50), interval = "confidence")
```

```
##       fit      lwr      upr
## 1 13.1673 12.92128 13.41333
```

```
predict(pul_model, newdata= tibble(prizes_9014 =3), interval = "prediction")
```

```
##        fit      lwr      upr
## 1 12.50539 11.47712 13.53367
```

```
predict(pul_model, newdata= tibble(prizes_9014 =25), interval = "prediction")
```

```
##        fit      lwr      upr
## 1 12.81522 11.78848 13.84197
```

```
predict(pul_model, newdata= tibble(prizes_9014 =50), interval = "prediction")
```

```
##       fit      lwr      upr
## 1 13.1673 12.12262 14.21198
```

```
condifence_interval_pul_model <- tibble(
  model = c("pul_model", "pul_model", "pul_model"),
  prize_count = c(3, 25, 50),
  fit = c(2.71828^12.50539, 2.71828^12.81522, 2.71828^13.1673),
  lwr = c(2.71828^12.34252, 2.71828^12.6623, 2.71828^12.92128),
  upr = c(2.71828^12.66826, 2.71828^12.96815, 2.71828^13.41333),
  range = c(upr-lwr),
)
head(condifence_interval_pul_model)
```

```
## # A tibble: 3 x 6
##   model     prize_count     fit     lwr     upr   range
##   <chr>           <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
```

```
## 1 pul_model           3 269785. 229237. 317506.  88269.
## 2 pul_model          25 367769. 315619. 428541. 112922.
## 3 pul_model          50 522976. 408919. 668854. 259935.
```

```
prediction_interval_pul_model <- tibble(
  model = c("pul_model", "pul_model", "pul_model"),
  prize_count = c(3, 25, 50),
  fit = c(2.71828^12.50539, 2.71828^12.81522, 2.71828^13.1673),
  lwr = c(2.71828^11.47712, 2.71828^11.78848, 2.71828^12.12262),
  upr = c(2.71828^13.53367, 2.71828^13.84197, 2.71828^14.21198),
  range = c(upr-lwr),
)
head(prediction_interval_pul_model)
```

```
## # A tibble: 3 x 6
##   model     prize_count     fit     lwr      upr    interval range
##   <chr>           <dbl>   <dbl>   <dbl>    <dbl>    <dbl>
## 1 pul_model           3 269785.  96482.  754387.  657905.
## 2 pul_model          25 367769. 131725. 1026803.  895078.
## 3 pul_model          50 522976. 183985. 1486553. 1302567.
```

*Table 1: A table of Prediction intervals for pul_model*

| **Model** | **Prize count** | **Fit** | **Lower Bound** | **Upper bound** | **Interval range** |
|---|---|---|---|---|---|
| Pul-model | 3 | 269785 | 96482 | 754387 | 657905 |
| Pul-model | 25 | 367769 | 131725 | 1026803 | 895078 |
| Pul-model | 50 | 522976 | 183985 | 1486553 | 1302567 |

Table 2: A table of **Confidence intervals** for pul_model

| **Model** | **Prize count** | **Fit** | **Lower Bound** | **Upper bound** | **Interval range** |
|---|---|---|---|---|---|
| Pul-model | 3 | 269785 | 229237 | 317506 | 88269 |
| Pul-model | 25 | 367769 | 315619 | 428541 | 112922 |
| Pul-model | 50 | 522976 | 408919 | 668854 | 259935 |

In looking at the above table, what's clear about the pul_model is that the more Pulitzer Prizes you input, the less sure of the prediction you can be. The confidence interval

expresses the range of values you'll need to include in order to be 95% certain the true mean is present within the interval range. The range is the difference between the upper and lower bounds, when it gets larger, it means the certainty of the prediction is reducing.

Confidence intervals are about the population mean, while prediction intervals are about an individual subject. In comparing the two, our certainty around prediction intervals is much lower than confidence intervals as demonstrated by the magnitude of range values required to have a probability of including the true value at 95%.

### 4.4

Using the model from Question 3(b), calculate 90% prediction intervals for the expected change in circulation of the newspaper under each of the three proposed strategic directions. Place these prediction intervals in a table and contrast them in context.

```
predict(pul_model2, newdata= tibble(prizes_9014 =3), interval = "prediction",
level=0.9)
```

```
##         fit       lwr      upr
## 1 -34.25423 -77.72971 9.221238
```

```
predict(pul_model2, newdata= tibble(prizes_9014 =25), interval = "prediction"
, level= 0.9)
```

```
##         fit       lwr      upr
## 1 -25.74021 -69.15107 17.67065
```

```
predict(pul_model2, newdata= tibble(prizes_9014 =50), interval = "prediction"
, level=0.9)
```

```
##         fit       lwr      upr
## 1 -16.06518 -60.23418 28.10381
```

```
prediction_interval_pul_model2 <- tibble(
  model = c("pul_model", "pul_model", "pul_model"),
  prize_count = c(3, 25, 50),
  fit = c(2.71828^-34.25423 , 2.71828^-25.74021 , 2.71828^-16.06518 ),
  lwr = c(2.71828^-77.72971  , 2.71828^-69.15107 , 2.71828^-60.23418 ),
  upr = c(2.71828^9.221238  , 2.71828^17.67065 , 2.71828^28.10381),
  range = c(upr-lwr),
)
prediction_interval_pul_model2
```

```
## # A tibble: 3 x 6
##   model     prize_count      fit      lwr     upr   range
##   <chr>           <dbl>    <dbl>    <dbl>   <dbl>   <dbl>
## 1 pul_model           3 1.33e-15 1.75e-34 1.01e 4 1.01e 4
## 2 pul_model          25 6.62e-12 9.29e-31 4.72e 7 4.72e 7
## 3 pul_model          50 1.05e- 7 6.93e-27 1.60e12 1.60e12
```

Table 2: A table of **Prediction intervals** for pul_model2

| Model | Prize count | Fit | Lower Bound | Upper bound | Interval range |
|---|---|---|---|---|---|
| Pul-model | 3 | 1.33e-15 | 1.75e-34 | 1.01e 4 | ~1.01e4 |
| Pul-model | 25 | 6.62e-12 | 9.29e-31 | 4.72e 7 | ~4.72e7 |
| Pul-model | 50 | 1.05e- 7 | 6.93e-27 | 1.60e12 | ~1.60e12 |

This model has had its issues from the start (outlier concerns, difficulty in reaching significance, poor verification of assumptions, etc.). It's saying here that to be 90% certain of an individual within the population achieving any of those prize inputs will sit between 0 and over 10,000.

## Question Five: Limitations

Incorporate your answers from this question in your email to your manager. ### 5.1 Discuss what limitations there might be to each of the models. Why might this model be insufficient for its application? You should discuss at least two limitations of these models in application.

```
filter(pulitzer, perc_change >= 0)
```

```
## # A tibble: 6 x 6
##   newspaper        prizes_9014 circ_2004 circ_2013 perc_change average_of_04_an~
##   <chr>                  <dbl>     <dbl>     <dbl>       <int>             <dbl>
## 1 Wall Street Jou~          51   2101017   2378827          13           2239922
## 2 New York Times           118   1119027   1865318          67           1492172.
## 3 San Jose Mercur~           7    558874    583998           4            571436
## 4 Chicago Sun-Tim~           3    453757    470548           4            462152.
## 5 Denver Post               10    340168    416676          22            378422
## 6 Orange County R~           6    310001    356165          15            333083
```

There are a few things that make these models insufficient for our purposes: The data doesn't consider distribution, geography, or the reasonable expected reach of each paper as I spoke too earlier. It also hasn't considered the many outside factors that come along with the passage of time such as social media and online news outlets which offer real-time and economic advantages to readers. Furthermore, people generally aren't aware that a newspaper even won an award, perhaps there are better measures of good journalism and marketing.

Of the 50 publishers, only 6 papers had a positive change in circulation over the last decade:

- Wall street Journal, with 51 prizes won
- New York Times, with 118 prizes won
- San Jose Mercury News, with 7 prizes won
- Chicago Sun-Times, with 3 prizes won
- Denver post, with 10 prizes won
- Orange County Register, with 6 prizes won

There were several publishers that won more than those listed above so there is clearly a better measure of change in circulation than the Pulitzer Prize count over the last 25 years. Adding weighting to the more recent Pulitzer winners may help, but even then, I expect a better variable is available. Perhaps the average education level of the staff at each publishing firm, the average amount of time spent on each article, or the average amount of experience of the work force at each publishing company would be a better variable for predicting future circulation numbers; particularly in conjunction with the suggestion made earlier about using '*(circulation)/ (population of news readers in areas of distribution)*' to make publishers more comparable.

In terms of limitations, the model assumes these businesses are machines that operate the same way every time. However, a prize-winning journalist can always resign and new talent can always arise. It cannot account for these changes or their influence on readership over the long term.

It also assumes the world will continue to act with the same motivations today as they did yesterday over long periods of time. But the type of news people really need today is changing, Covid19 has made that quite clear. In Australia, Television, radio, and the internet provided people with news on restrictions. Newspapers are incapable of providing the same type of support or reaching the same magnitude of people as restrictions prevented them from doing so. This was a trend the model could not hope to account for.

*References:*

Kuhn, M 2021, 'caret: Classification and Regression Training', R package version 6.0-90,< https://CRAN.R-project.org/package=caret>.

R Core Team, 2021, *R: A language and environment for computing and statistical computing* R foundation for statistical computing Vienna, Austria.

Wickham, H 2019, 'stringr: Simple, Consistent Wrappers for Common String Operations', R package version 1.4.0, <https://CRAN.R-project.org/package=stringr>.

Wickham, H, Averick, M, Bryan, J, Winston, C, McGowan, LDA, François, R, Grolemund, G, Hayes, A, Henry, L, Hester, J, Kuhn, M, Pedersen, TL, Miller, E, Bache, SM, Müller, K, Ooms, J, Robinson, D, Seidel, DP, Spinu, V, Takahashi, K, Vaughan, D, Wilke, C, Woo, K & Yutani, H 2019, 'Welcome to the {tidyverse}', *Journal of Open Source Software*, vol. 4, no. 43, p. 1686.

Wickham, H, François, R, Henry, L & Müller, K 2021, 'dplyr: A Grammar of Data Manipulation', R package version 1.0.7, <https://CRAN.R-project.org/package=dplyr>.

```
citation()

citation("tidyverse")

citation("stringr")

citation("dplyr")

citation("caret")
```