GLASS (2019) REGRESSION ANALYSIS

Nicholas DeChant

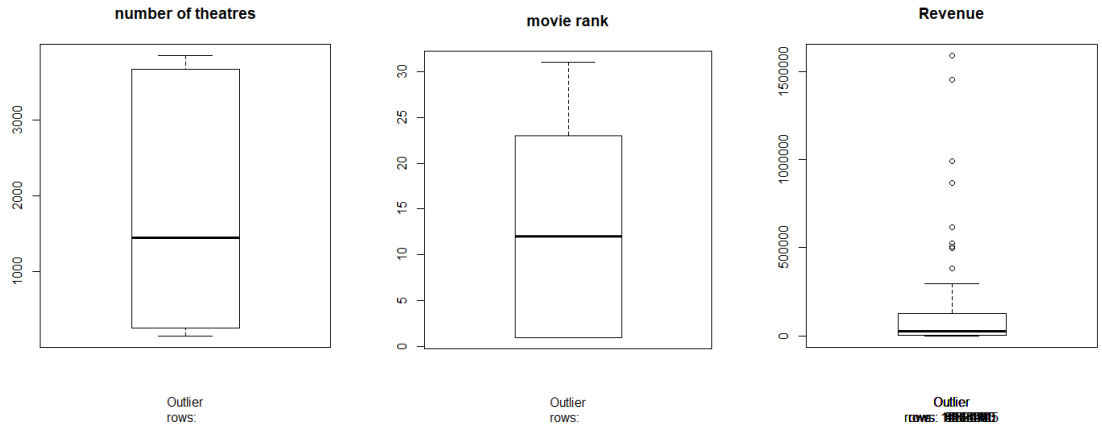
2020-04-04

Introduction

The following report presents a statistical analysis of the data collected for a popular movie called Glass, which came out in 2019 starring Bruce Willis and James McAvoy. The data was collected by Box Office Mojo, a website that tracks box office revenue. The site provides the collected daily domestic gross revenue, the number of theaters the movie was played at, the average ticket sales earned per theater in each day, and the movies rank compared to other feature films for 77 days, i.e. N = 77. The dataset was picked because there are several variables, such as the ones listed above, that can all be used to model linear regression equations. Out of the 4 variables above, we can pick two variables, label one as an independent predictor, the other as a dependent response variable, and then use RStudio to quickly determine a correlation between each predictor and response variable being modeled.

For this report, we're interested in looking at three of the variables by asking two simple questions: can we predict the number of movie theaters the movie is being played at based on how high it is ranked? And secondly, can we determine the daily gross revenue based off its rank?

First, let's check for outliers in our three variables. The following are boxplots for each variable we're interested in:

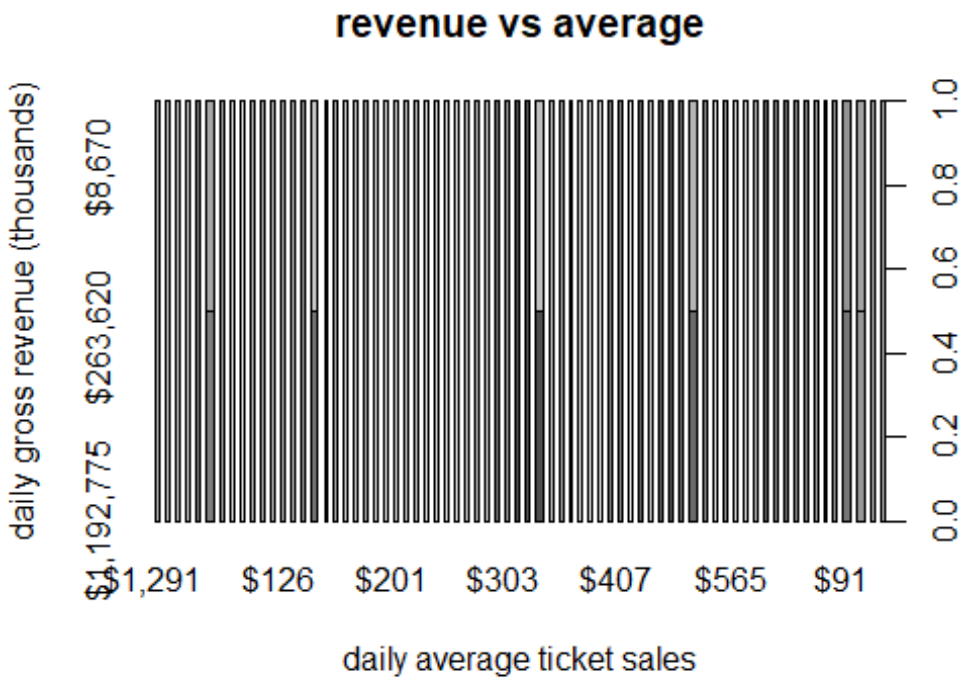


The only variable with outliers is the daily revenue, and there are clearly several. This may or may not be a poor response, but we will choose to leave the outliers alone because they are still valuable points in our dataset. The large outliers could mean the first opening days of a movie, a holiday where people chose to go to a movie because they didn't have work, or a Friday and Saturday night during the first few weeks Glass became available to watch.

Data Cleaning and Regression Analysis in R

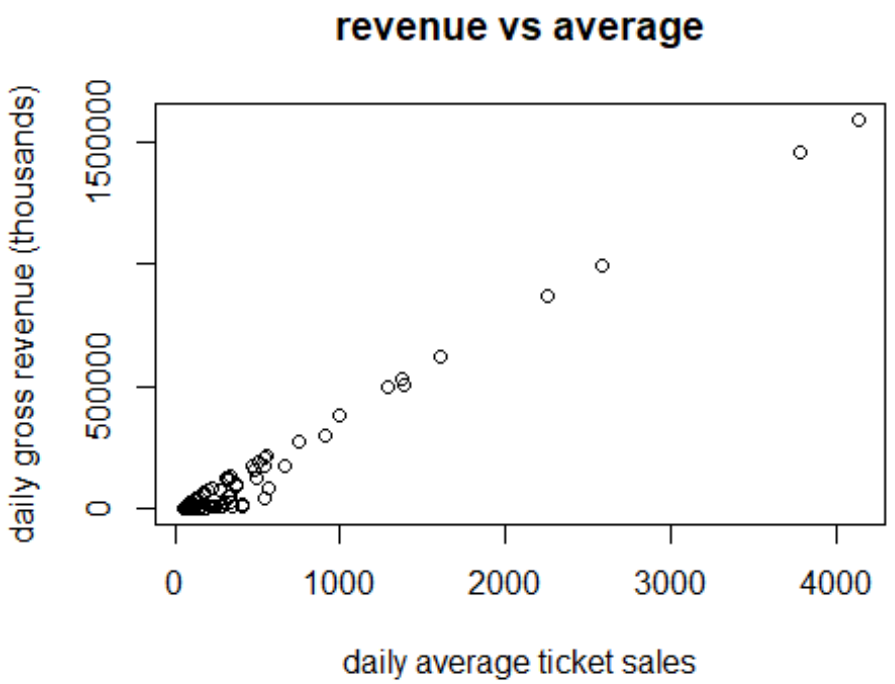
The data was extracted from Box Office Mojo's website, inserted into excel, and finally saved as a CSV file. This data caused problems in RStudio when executing commands such as "plot()", declaring variables, and it was determined that it needed cleaning. The following code shows the difference between the uncleaned and cleaned data in R:

```
##The data in the graph and code below shows the "uncleaned data". What does
this look like if we plot it? It looks like a bunch of bars, the data is out
of order on each axis, the range for both variables are incorrect, and there
is an extra label on the right y-axis going from 0.0 to ... 1.0.

##SCATTERPLOT of uncleaned data:
uc_movie <- read.csv("C:\\Users\\User\\Desktop\\uncleaned_moviedata.csv")
df <- data.frame(uc_movie)
y <- uc_movie$Daily.Revenue
x <- uc_movie$Avg

plot(x,y, xlab = "daily average ticket sales", ylab = "daily gross revenue (t
housands)", main = "revenue vs average")
```




```
##In fact y <- (uc_movie$Daily.Revenue / 10) doesn't even compile if we try
to reduce zeros on the y-axis, so we can't change the display of numbers for
daily gross revenue.
```

```
##List of UNCLEANED data (first 5 rows):
head(uc_movie, n = 5)

##   Date.of.Movie Daily.Revenue Rank Theaters     Avg Revenue.To.Date
## 1     18-Jan-19  $15,886,745     1     3841 $4,136     $15,886,745
## 2     19-Jan-19  $14,524,105     1     3841 $3,781     $30,410,850
## 3     20-Jan-19   $9,918,070     1     3841 $2,582     $40,328,920
## 4     21-Jan-19   $6,176,030     1     3841 $1,607     $46,504,950
## 5     22-Jan-19   $3,814,910     1     3841   $993     $50,319,860

##Now lets clean the data. We have to remove currency symbols to 'General' in
excel, and we can get a nice picture:


moviedata <- read.csv("C:\\Users\\User\\Desktop\\moviedata.csv")
df <- data.frame(moviedata)

y <- moviedata$Daily.Revenue/10 ## y <- (uc_movie$Daily.Revenue / 10) will
execute with cleaned data
x <- moviedata$Avg

plot(x,y, xlab = "daily average ticket sales", ylab = "daily gross revenue (t
housands)", main = "revenue vs average")
```

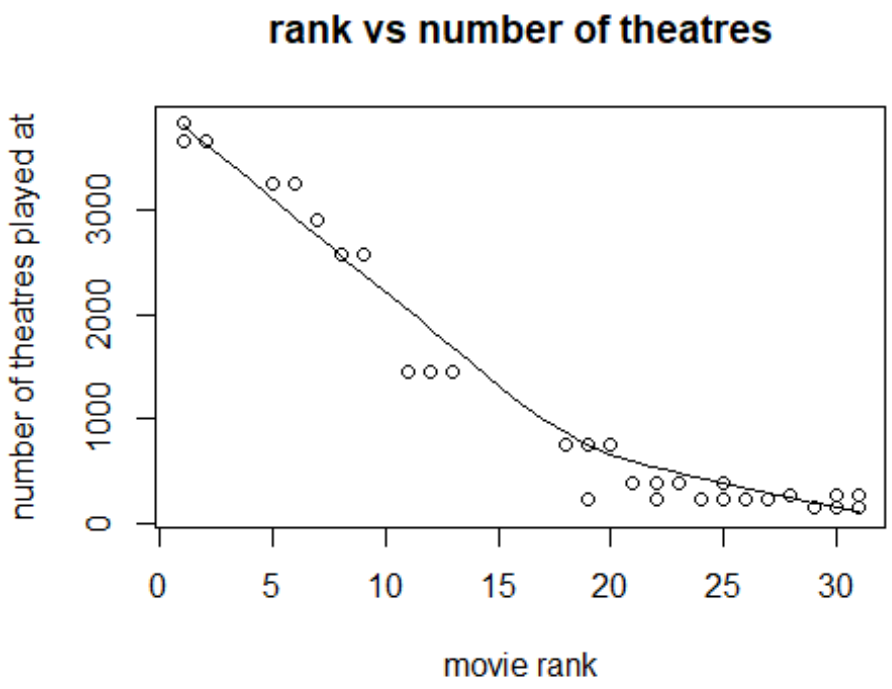



```
##The dataset we have contains the first 77 days of data collected for the
movie Glass. Let's check and see what the first 10 days, or rows, of cleaned
```

```r
data looks like:
head(moviedata, n = 10)
```

```
##    Date.of.Movie Daily.Revenue Rank Theaters  Avg Revenue.To.Date
## 1      18-Jan-19      15886745    1     3841 4136        15886745
## 2      19-Jan-19      14524105    1     3841 3781        30410850
## 3      20-Jan-19       9918070    1     3841 2582        40328920
## 4      21-Jan-19       6176030    1     3841 1607        46504950
## 5      22-Jan-19       3814910    1     3841  993        50319860
## 6      23-Jan-19       2144560    1     3841  558        52464420
## 7      24-Jan-19       2076715    1     3841  540        54541135
## 8      25-Jan-19       4963595    1     3844 1291        59504730
## 9      26-Jan-19       8652565    1     3844 2250        68157295
## 10     27-Jan-19       5268280    1     3844 1370        73425575
```

```r
##Can we predict the number of movie theaters the movie is being played at based on how high it is ranked?

linearMod <- lm(moviedata$Theaters ~ moviedata$Rank, data = moviedata)
scatter.smooth(moviedata$Rank,moviedata$Theaters, xlab = "movie rank", ylab = "number of theatres played at", main = "rank vs number of theaters")
```




```r
summary(linearMod)
```

```
##
## Call:
## lm(formula = moviedata$Theaters ~ moviedata$Rank, data = moviedata)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -944.43 -382.93   92.57  271.57  694.58
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3705.933     75.188   49.29   <2e-16 ***
## moviedata$Rank -133.500      4.249  -31.42   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 405.4 on 75 degrees of freedom
## Multiple R-squared:  0.9294, Adjusted R-squared:  0.9284
## F-statistic: 987.1 on 1 and 75 DF,  p-value: < 2.2e-16
```

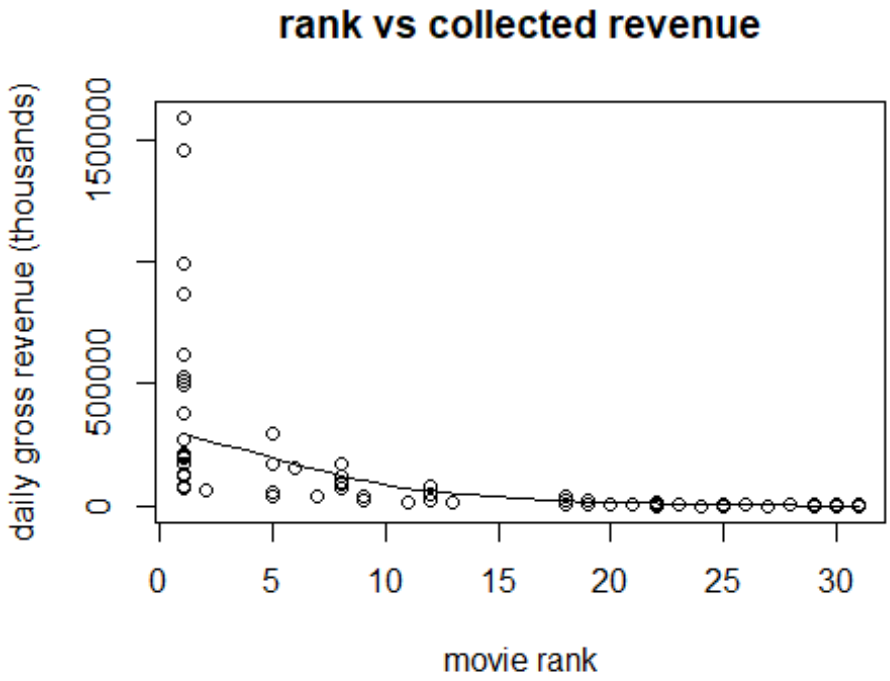
##We obtain r^2 = .9294, an adjusted r^2 = .9284, and p = < 2.2e-16 = 0.00000000000000022. With p < .05, and high r^2, this is a is a solid model and the data is significant.
*##This is indeed an excellent predictor model, and summary(linearMod) allows us to make y = 3705.933 – 133.5*x, or theaters = 3705.933 – 133.5x*rank.*

##Can we determine the daily gross revenue based off its rank?

```
linearMod <- lm(moviedata$Daily.Revenue ~ moviedata$Rank , data = moviedata)
scatter.smooth(moviedata$Rank,(moviedata$Daily.Revenue / 10 ), xlab = "movie rank", ylab = "daily gross revenue (thousands)", main = "rank vs collected revenue")
summary(linearMod)
```

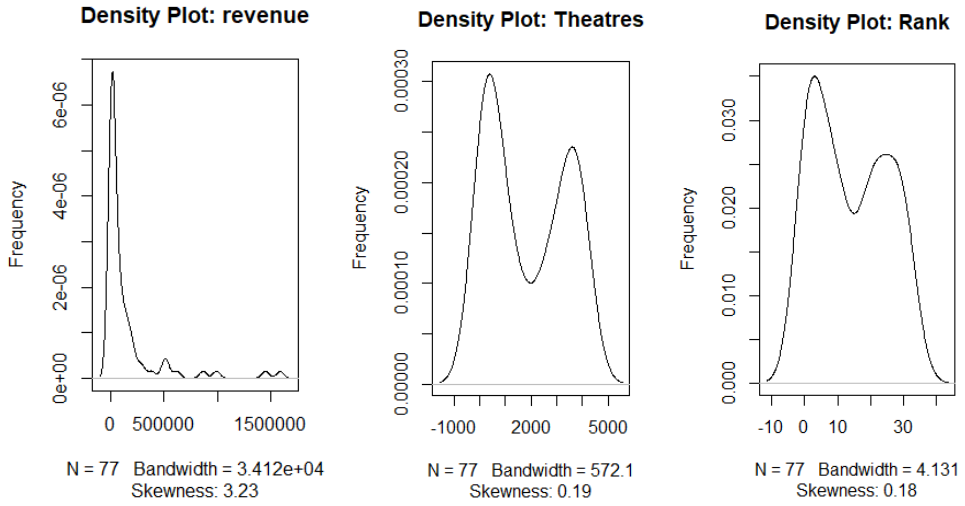
```
##
## Call:
## lm(formula = moviedata$Daily.Revenue ~ moviedata$Rank, data = moviedata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2530068 -1341348  -477293   685880 12619087
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     3408514     469912   7.254 3.10e-10 ***
## moviedata$Rank  -140856      26557  -5.304 1.11e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2533000 on 75 degrees of freedom
## Multiple R-squared:  0.2728, Adjusted R-squared:  0.2631
## F-statistic: 28.13 on 1 and 75 DF,  p-value: 1.106e-06
```

```
##We obtain r^2 = .2728, an adjusted r^2 = .2631, and p = 1.106e-06 = 0.00000
1106. P is low so the data is important, however it is a poor model for our
data since r^2 is low.
##These two variables, revenue and rank do not model a reliable relationship
due to low r^2. summary(linearMod) returns y = 3408514 - 140856*x, or revenue
= 3408514 - 140856*rank.
```



Findings

Our analysis in R shows that we can use linear regression to predict the number of movie theaters the movie is currently being played at based on how high it is ranked. However, the second model does not provide a reliable equation to predict daily revenue based off movie rank. This may be caused by Daily Revenue having several extremely large outliers in the dataset. Revenue's density curve has high skewness equal to 3.23 and is extremely skewed to the right. Rank and Theaters have low skewness, which allows us to take two better datasets to accurately model an equation. The following curves show the density plots of each variable below:



The equations we found were:

1) Number of theaters = 3705.933 – 133.5*rank, or y = 3705.933 – 133.5*x
2) Daily Revenue = 3408514 – 140856*rank, or y = 3408514 – 140856*x

Both equations have p small, however r^2 is high for equation (1), but low for equation (2). Therefore, equation (2) will be eliminated, and we can't use linear regression to accurately ask, "can we predict revenue by movie rank?".

We can take equation (1) and test cases for a given x, or rank value, and determine the number of movie theaters Glass is being played at, for example:

If x = rank 1, then y = -133.5 + 3705.933(1) = 3572, very close the actual number of real theaters with rank 1 in the dataset in a range between [3665, 3841] when rank was 1.

If x = rank 10, then y = -133.5 + 3705.933(10) = approximately 2370 theaters. There was no real time when Glass's rank was 10, however the data in the dataset jumped from rank = 9 to rank = 12, decreasing from 2575 to 1446 theaters. Rank = 10 is between ranks [9,12], and y(10) = 2370 theaters is between [1446,2575] real theaters, so this is a good estimate.

The multiple r^2 for equation (1) was r^2 = .9294 with an adjusted r^2 = .9284. Multiple r^2 = .9294 means that roughly 93% of the decrease in the number movie theaters Glass is played at is due to its increase in rank. Adjusted r^2 = .9284 ~ 93%. Both are identical because there is only one predictor variable, rank. If we added another predictor to 3705.933 – 133.5*rank, such as daily revenue, adjusted r^2 may either increase or decrease depending if it is useful to the model. Multiple r^2 will increase no matter how useful the variable is being added to the model. Therefore, an adjusted r^2 = .9284 for equation (1) is excellent.

Closing Remarks

This report has shown how we can use linear regression to predict the number of movie theaters a movie is played at by rank. We successfully modeled two variables (the number of theaters and rank) to determine a strong response based off rank as a predictor. We determined that you cannot make accurate predictions for revenue based off a movies rank. Perhaps by removing outliers and large values from daily revenue, we could model a better equation for predicting revenue. However, it was noted in the introduction that we will not do this, because such large outliers (or days with very high gross revenue) impact the total domestic earned gross revenue for a movies entire showtime period, and these outliers are crucial to knowing how successful a movie is over its total days of showtime duration.

Source:

"Glass." *Box Office Mojo*, www.boxofficemojo.com/release/rl1518241281/?ref_=bo_tt_gr_1.