Predicting Domestic Gross of Movies

Xiaoyu Hu

Advisor: David Aldous

I. Introduction

As watching movies becomes one of the most popular entertainments in the 21st century, the film industry is really lucrative. More and more people want to invest a movie or be a film producer. Thereupon, there are many questions we may consider. What kinds of movies can appeal audience? How to make a profit by producing movies? The goal of this research is to figure out how does factors from all aspects impact the box office of a movie based on the data from internet and predict the box office of movies which are going to release in 2017 fall.

II. Data Collection

Data were collected from various sources, basically from this website http://www.the-numbers.com/. I collected the detailed information of movies from 2010 to 2016 and used R to create a data frame include all these information.

Budget	Movie	ReleaseDate	Distributor	Genre	Gross	MPAA
5000000	10 Cloverfield Lane	3/11/2016	Para	Thri	72082999	PG-13
50000000	13 Hours: The Secret Soldiers of Benghazi	1/15/2016	Para	Dra	52853219	R
43000000	A Monster Calls	12/23/2016	Other	Dra	72618	PG-13
8000000	A Street Cat Named Bob	11/18/2016	Other	Dra	82703	R
170000000	Alice Through the Looking Glass	5/27/2016	Dis	Adv	77042381	PG
17000000	Almost Christmas	11/11/2016	Uni	Dra	42002805	PG-13
47000000	Arrival	11/11/2016	Para	Dra	92191690	PG-13
125000000	Assassin<U+2019>s Creed	12/21/2016	Fox	Act	39721564	PG-13
20000000	Bad Moms	7/29/2016	Other	Com	113257297	R
20000000	Barbershop: The Next Cut	4/15/2016	WB	Com	54030051	PG
250000000	Batman v Superman: Dawn of Justice	3/25/2016	WB	Act	330360194	PG-13

Figure 1: first several rows of data frame

III. Preliminary Analysis

There are four predictor variables that I want to include :

Budget: Theoretically, a high budget may bring high box office. I treat this predictor as a numeric variable.

Distributor: Distributor may affect the quality of a film, and then affect the box office of the film. I treat this predictor as a categorical variable. I choose some of the famous film factories, like Disney and 21st Century Fox, and put the less famous ones into other.

Genre: Audience may prefer some specific genre of movies. I treat this predictor as a categorical variable.

MAPP: MAPP stands for Motion Picture Association of America film rating system. I treat this predictor as a categorical variable.

I draw the scatter plot for the numeric variable and box plots for the categorical variables to see if these variables really have an effect on the box office and decide if I should include these variables in linear regression models.

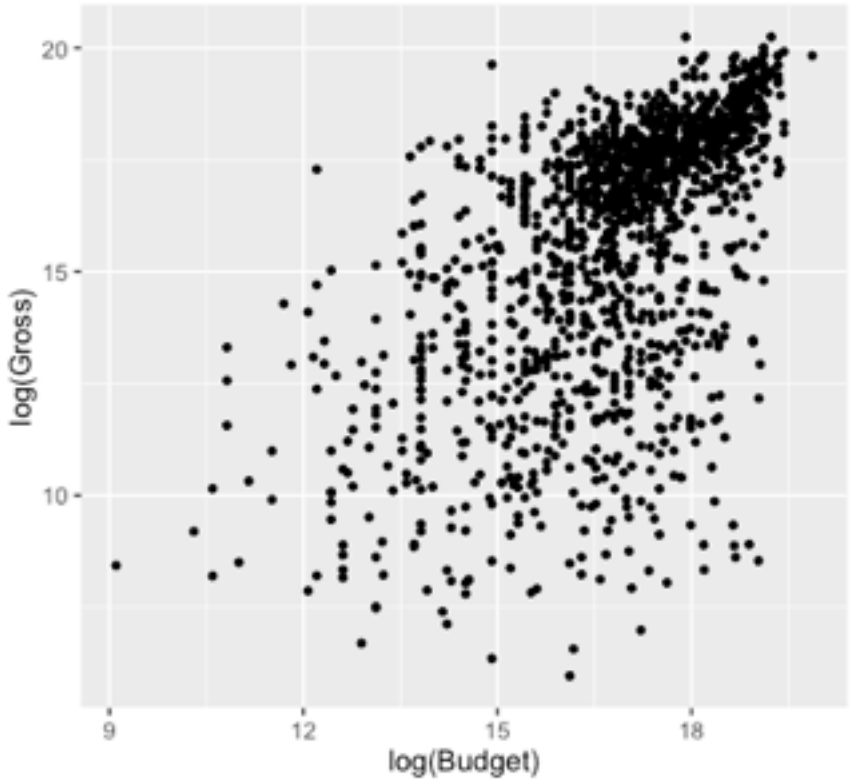

Figure 2

In general, budget and gross is positive correlation. However, for some films have very high budget got low gross.

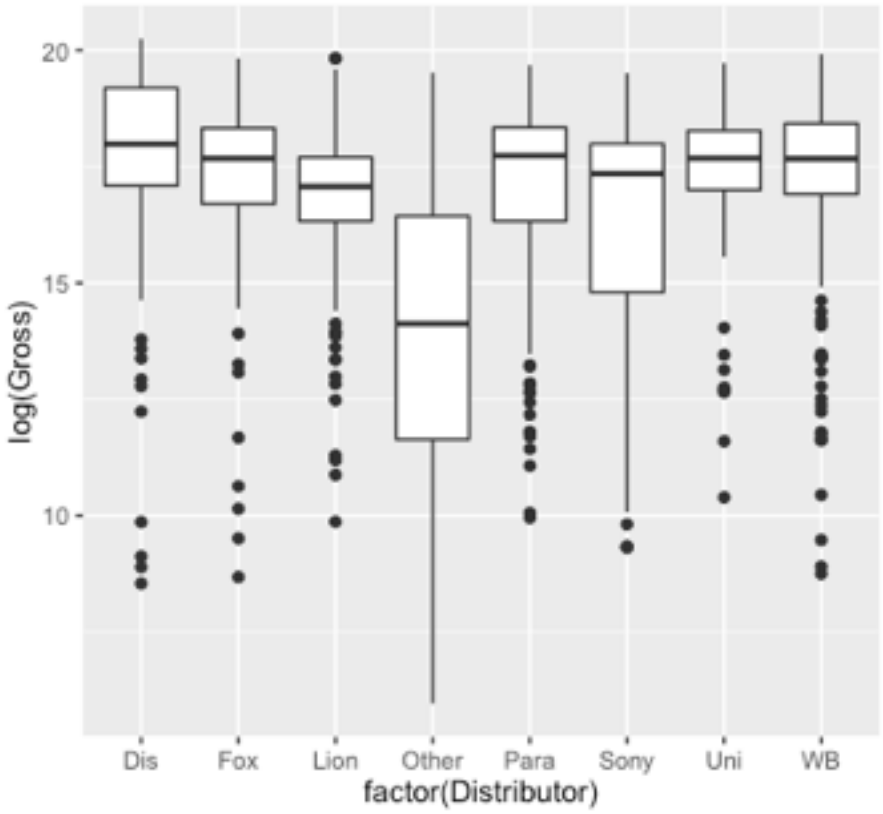

Figure 3: Dis = Disney, Fox = 21st Century Fox, Lion = Lionsgate, Para = Paramount Pictures, Sony = Sony Pictures, Uni = Universal Pictures, WB = Warner Bros.

There are apparent differences in grosses using the Distributor variables. It seems that medium of the big film factories are approximate and they are much higher than the other small film factories.

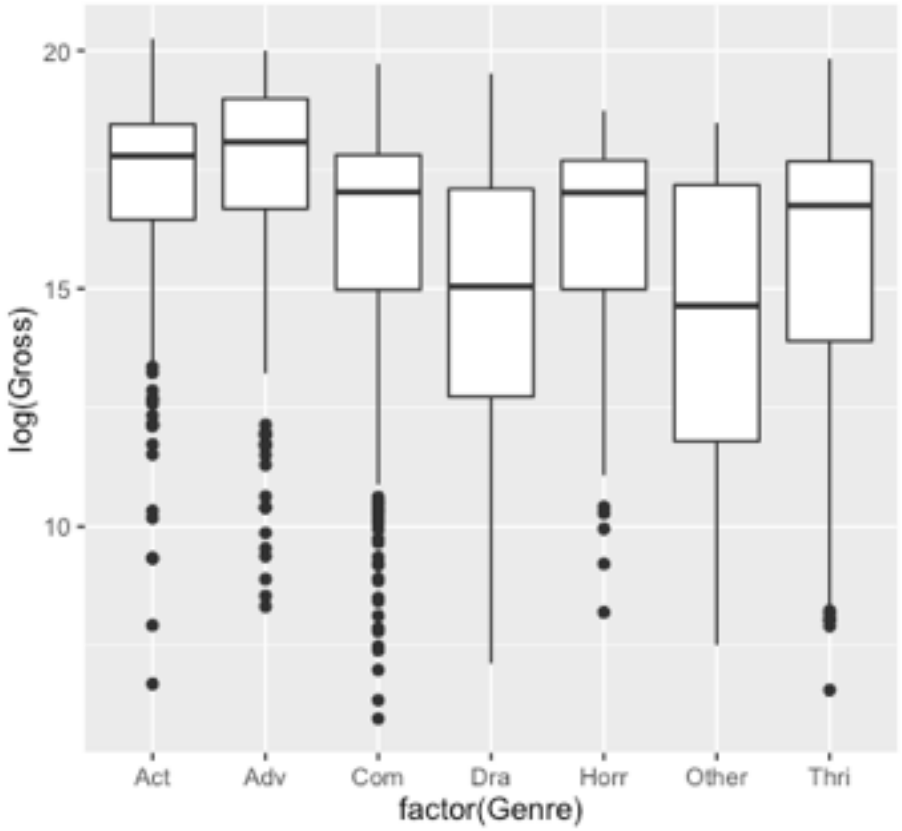

Figure 4: Act = Action, Adv = Adventure , Com = Comedy, Dra = Drama, Horr = Horror, Thri = Thriller

The difference in gross is significant between different genre.

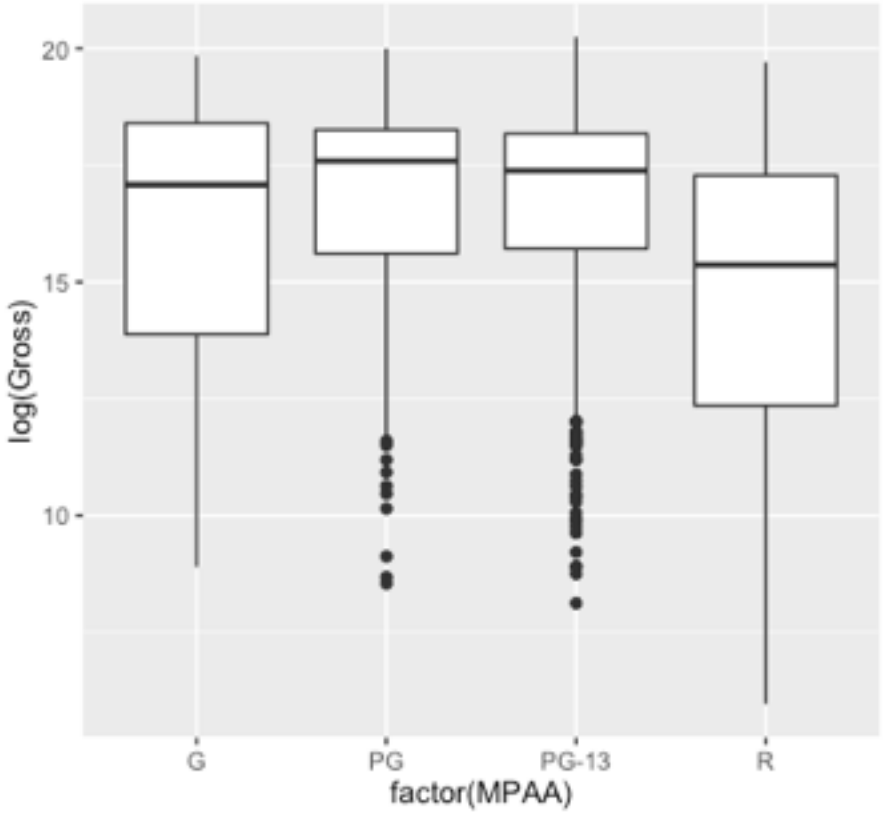

Figure 5: G – General Audiences, PG – Parental Guidance Suggested, PG-13 – Parents Strongly Cautioned, R – Restricted

In general, R-rated films have lower gross than other films. It is reasonable because R-rated films may lose some teenager audience.

IV. Linear Regression Analysis

After preliminary analyzing the variables separately, to better predict the box office, I construct linear regression models to incorporate all of the variables at the same time.

First, I construct a linear regression model with interception:

Call:

lm(formula = log(Gross) ~ log(Budget) + factor(MPAA) + factor(Distributor) +

factor(Genre), data = data)

Variable		**Coefficients**	**t value**	**Pr(>|t|)**
(Intercept)		4.57978	4.499	0.00000734036 ***
Budget		0.67537	14.184	< 0.0000000000000002 ***
MPAA	PG	0.91558	1.930	0.05379 .
	PG-13	1.16904	2.426	0.01539 *
	R	0.29968	0.615	0.53878
Distributor	Fox	-0.13900	-0.462	0.64450
	Lion	-0.21089	-0.643	0.52027
	Other	-1.73985	-6.302	0.00000000038 ***
	Para	-0.12166	-0.391	0.69564
	Sony	-0.45063	-1.567	0.11736
	Uni	0.24829	0.779	0.43624

	WB	-0.16649	-0.568	0.57021
Genre	Adv	-0.24577	-0.942	0.34649
	Com	0.05647	0.264	0.79150
	Dra	-0.47472	-2.248	0.02472 *
	Horr	0.98761	3.260	0.00114 **
	Other	-0.51149	-1.612	0.10723
	Thri	0.02907	0.122	0.90277

—

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.215 on 1580 degrees of freedom

Multiple R-squared: 0.4224, Adjusted R-squared: 0.4162

F-statistic: 67.98 on 17 and 1580 DF, p-value: < 0.00000000000000022

Secondly, I construct a linear regression model without interception:

Call:

lm(formula = log(Gross) ~ log(Budget) + factor(MPAA) + factor(Distributor) +

factor(Genre) - 1, data = data)

Variable		**Coefficients**	**t value**	**Pr(>|t|)**
Budget		0.67537	14.184	< 0.0000000000000002 ***
MPAA	G	4.57978	4.499	0.00000734036 ***
	PG	5.49536	5.832	0.00000000662 ***
	PG-13	5.74882	6.129	0.00000000112 ***
	R	4.87946	5.308	0.00000012627 ***
Distributor	Fox	-0.13900	-0.462	0.64450

	Lion	-0.21089	-0.643	0.52027
	Other	-1.73985	-6.302	0.00000000038 ***
	Para	-0.12166	-0.391	0.69564
	Sony	-0.45063	-1.567	0.11736
	Uni	0.24829	0.779	0.43624
	WB	-0.16649	-0.568	0.57021
Genre	Adv	-0.24577	-0.942	0.34649
	Com	0.05647	0.264	0.79150
	Dra	-0.47472	-2.248	0.02472 *
	Horr	0.98761	3.260	0.00114 **
	Other	-0.51149	-1.612	0.10723
	Thri	0.02907	0.122	0.90277

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.215 on 1580 degrees of freedom

Multiple R-squared: 0.981, Adjusted R-squared: 0.9808

F-statistic: 4534 on 18 and 1580 DF, p-value: < 0.00000000000000022

The R-squared value of the model without interception is bigger than the model with interception(0.9808 > 0.4162). Therefore, interception should be excluded.

From the result of this model, we can see that PG rating movies and PG-13 rating movies are easier get higher box office. This is reasonable because R rating movies which require accompanying parent or adult guardian may lose some young audience while G

rating movies are generally considered only suitable for children.

There is no big difference between the influence of famous film factories. However, the small factories, which usually produce independent films, easily get lower gross.

Horror movies are the most popular types of movies while drama movies and other types of movies, like documentary movies, are niche.

In all the categorical variables, MAPP has the most significant influence on domestic gross.

V. Rate of Return Analysis

From the linear model, we got the coefficient of the budget variable is 0.67537. To make further analysis of the relationship between the budget and gross so I use R to draw another plot to analysis that the budget is around which amount that has the highest rate of return.

rate of return = gross / budget

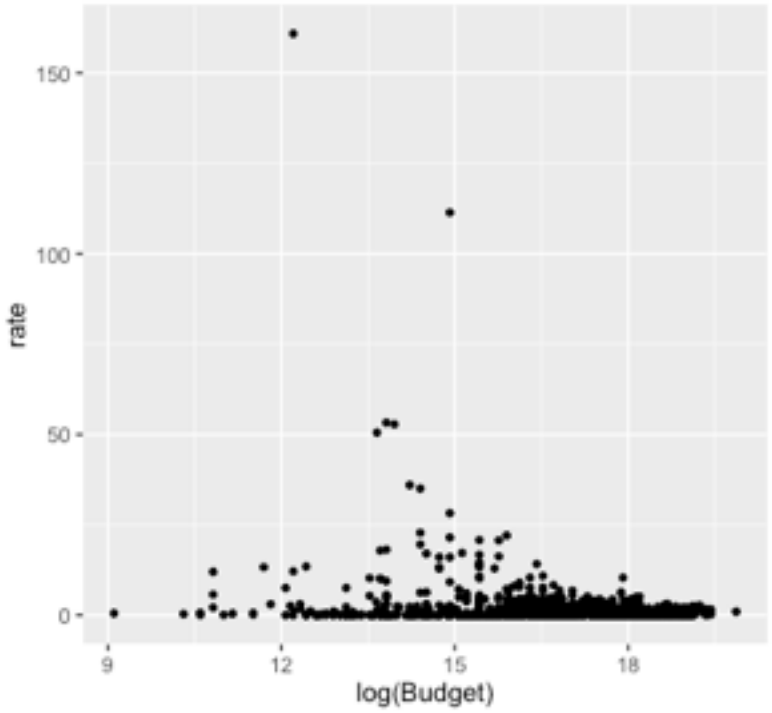

Figure 6

In this plot, except several outliers that get a really high rate of return, it seems that most of them have a similar return of rate. Therefore, I change the scale to get details.

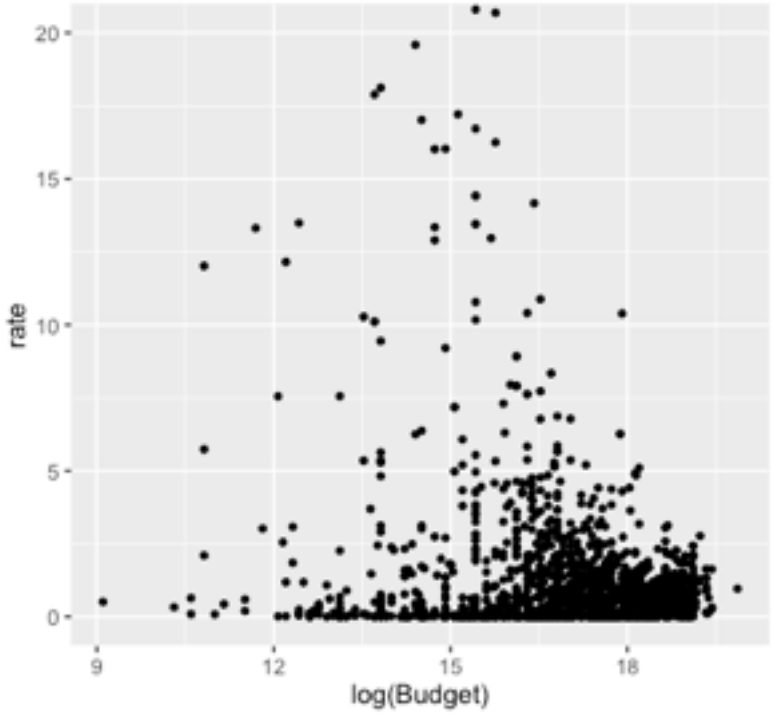

Figure 7

Movies which budget are around 15 get the highest rate of return. It is difficult to get a high rate of return for a high budget.

VI. Prediction

Using model without the interception, I start predicting the box office of 2017 movies. I choose 53 movies which already released and use R studio to predict the domestic gross of them:

```
predicted <- predict(model2, newdata=sample)
```

I plot the predicted gross and actual gross in the same graph. The red points are

predicted values and blue points are observed values. The x-axis is the index of the point which means the two points have the same index are fro the same movie. The y-axis is the log of gross.

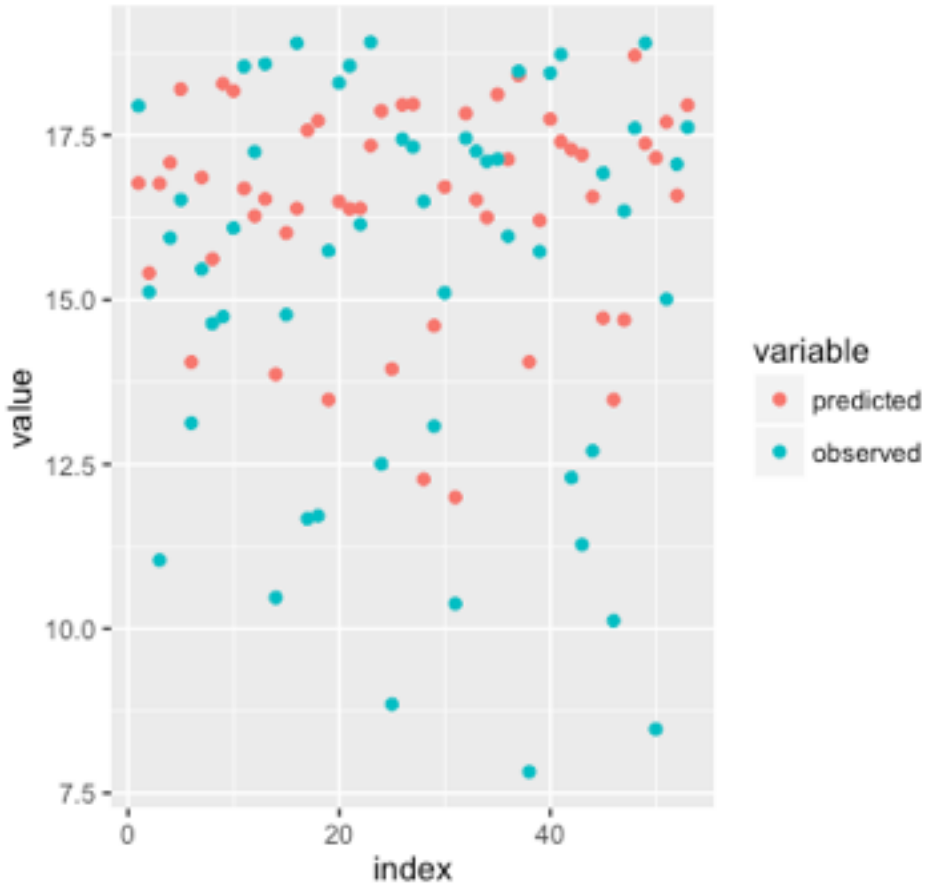

Figure 8

From the plot, we can see that the predicted values are close to the observed values and they are conservative.

VII. Future Developments

This project can be improved in these ways:

There are other predictors can be considered. For instance, the popularity of the leading characters may have an influence on the domestic gross of a movie. I did not include this predictor to my project because it is hard to decide how popular an actor is. Therefore, I

need to collect more data and figure out a method to calculate the popularity of an actor. Another example is that if a movie won an Oscar Award before it released, like *La La Land,* the box office of this movie might increase.

One possible way to improve this project is to collect the data from 2000 to 2009 and consider the inflation.

Another point is that linear model may not be the most suitable model for this project. Constructing some nonlinear models may improve the accuracy.

VIII. References

1. Data: http://www.the-numbers.com/.

2. http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0071226