December 6th, 2024

Predicting IMDB Movies Gross Revenue Using Multiple Linear Regression

STA302H1: Method of Data Analysis I

Professor: Austin Brown

Group 108

Contributions:

Ba Minh Dang Le:

- Prepared and finalized the R Markdown file for conducting data analysis.
- Wrote and finalized the Results section.
- Assisted with revisions to the Method session.
- Help revise the statistical poster.

Dang Trung Kien Nguyen:

- Revised the R Markdown file for tables and data analysis.
- Wrote the Introduction, Method, Limitations, and Conclusion sections.
- Designed and created the statistical results poster using Canva.

Abdelrahman Alkhawas:

- Composed the Demonstration of Editing Requirements

Session 1: Introduction

The financial success of movies has long been a topic of interest for researchers and decision-makers in the film industry. With production costs often reaching millions of dollars and substantial risks involved, understanding the factors that influence a movie's gross revenue is critical. This study aims to answer the research question: To what extent can IMDb rating, Metascore, release year, movie genre, and production cost predict a movie's gross revenue?

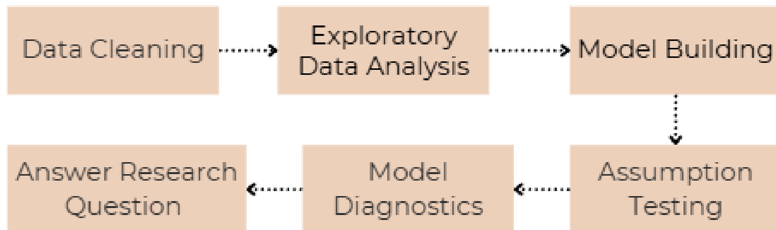
From previous studies, Eliashberg (1997, as cited in Pangarker, 2013) stated that critical reviews have a significant impact on box office revenue. De Vany and Walls (1999) argued that the film industry operates under stochastic and highly skewed dynamics, making outliers and influential points are important to interpret.

In addition, prior research suggests a positive correlation between production budgets and gross revenue (Litman, 1983, as cited in Pangarker, 2013), while higher Metascores are associated with better long-term performance (Kennedy, 2010). Moon et al. (2010) similarly found that early box office success often creates positive feedback loops, amplifying a film's overall financial performance.

This paper will emphasize both interpretability and prediction. Factors such as reviews and production costs influence gross revenue, aligning with findings that positive reviews (Eliashberg, 1997; Wallace et al., 1993) and release date (Litman, 1983) enhance revenue. From a predictive perspective, the model aims to estimate gross revenue for future films, offering insights for filmmakers. Since linear regression is a useful tool for quantifying relationships between response variables and predictors, it enables the testing of hypotheses related to individual and collective contributions. Overall, this study contributes to a deeper understanding of the factors driving box office success using regression.

Session 2: Methods

This session will further outline the methods, tools, and techniques that will be used to conduct the analysis and arrive at the final model.



2.1 Exploratory Data Analysis (EDA)

By using R and RStudio, descriptive statistics can be conducted and explore summary of the continuous variables.

Histograms will be used to visualize the distribution of continuous variables along with **Pie chart** for categorical variables (movie genre) to observe the distribution.

2.2 Model Building: Multiple Linear Regression

By constructing a preliminary model, we can evaluate the underlying assumptions and determine whether transformations are necessary.

Using R and RStudio, we establish a multiple linear regression model:

$$Gross\ Revenue \sim \beta_0 + \beta_1 I(x_i = Movie\ Genre\ i) + \beta_2 Release\ Year + \beta_3 IMDB + \beta_4 MetaScore + \beta_5 Budget + e$$

The model will be fitted using the ordinary least squares (OLS), which minimizes the sum of squared residuals to provide unbiased estimates.

2.3 Model Assumptions Check

Linearity: We will use the Residuals vs. Fitted plot to examine the relationship between the fitted preliminary model and the residuals, assessing whether they appear linear.

Independence of Errors: Check for autocorrelation in residuals, particularly important if data is time-dependent.

Homoscedasticity: The Scale-Location plot will be used to assess homoscedasticity.

Normality of Residuals: Q-Q plots will be used to assess whether the residuals follow a normal distribution.

Multicollinearity: The Variance Inflation Factor will be calculated for each predictor to evaluate the presence of multicollinearity. High VIF value (greater than 10) will be removed.

2.4 Model Refinement and Selection

Transformation: If the preliminary model violates several assumptions, a Box-Cox transformation can be applied to determine the optimal lambda, which will indicate the most suitable transformation for the data as well as a new assumption check will also follow.

For this studies, hypothesis testing method will be used for model selection, including:

Overall F-Test: This test evaluates whether the regression model significantly explains the variation in the response variable. It tests the null hypothesis that all regression coefficients, except the intercept, are equal to zero. A significant result indicates that at least one predictor contributes to the model.

T-Test on Each Predictor: The t-test assesses the individual significance of each predictor variable by testing the null hypothesis that its coefficient equals zero. A significant p-value indicates that the predictor contributes to explaining the response variable which helps identify which predictors are important in the model.

Partial F-Test: The partial F-test compares nested models, a full model with all predictors versus a reduced model with a subset of predictors, to evaluate whether excluding certain predictors significantly reduces the model's explanatory power.

2.5 Model Validation

Detecting Influential Observations: By using Cook's Distance and Box Plot, we can identify data points that have an outsized impact on the regression coefficients. Recognizing these points helps assess the robustness of the model.

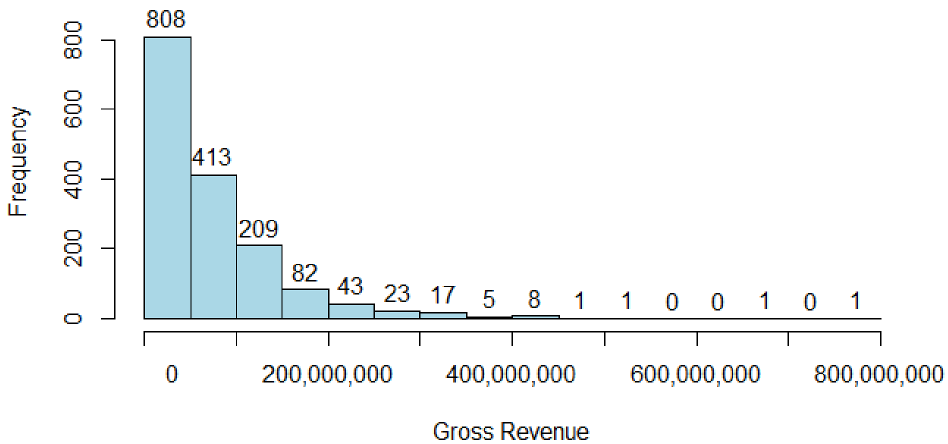
Model Fit and Performance: This study will use adjusted R-squared to report metrics for the refined model as well perform a comparison of metrics between initial and final models.

Interpret the Coefficients: Finally this study will explain the meaning of the regression coefficients (e.g., how much IMDb rating, etc., impact gross revenue).

Session 3: Result

3.1 Exploratory Data Analysis (EDA)

Figure 3.1.1
Distribution of Gross Revenue



Minimum	1st Q	Median	Mean	3rd Q	Maximum	S. D.
10000	22857500	49675000	71718319	95422500	760510000	75294763

Gross revenue has a right-skewed distribution, showing that it is highly concentrated at lower values.

Figure 3.1.2
Distribution of IMDB Rating

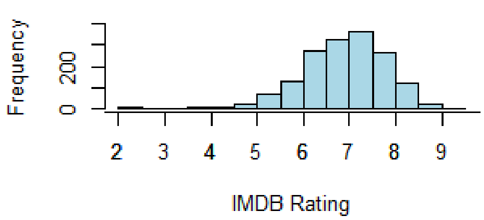


Figure 3.1.3
Distribution of Metascore

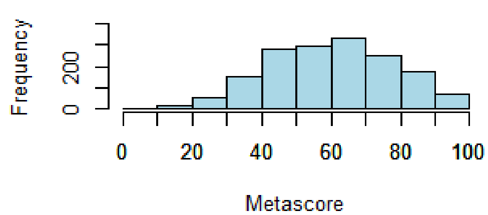


Figure 3.1.4
Distribution of Release Year

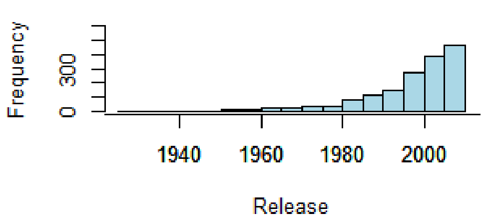
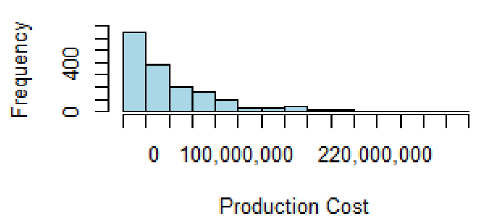
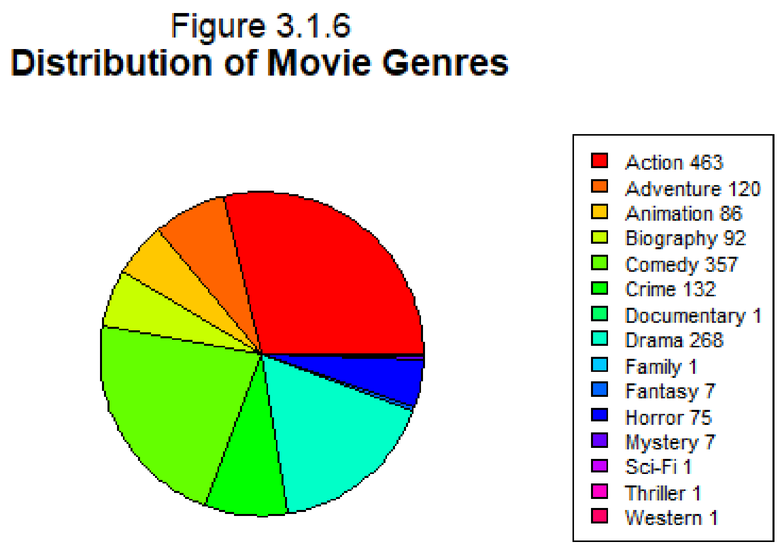


Figure 3.1.5
Distribution of Production Cost



IMDB rating (figure 3.1.2) and Metascore (figure 3.1.3) have an approximate normal distribution while production cost (figure 3.1.5) has a right-skewed distribution. The Release year (figure 3.1.4) shows a left-tail distribution indicating the increasing trends in movies over time.

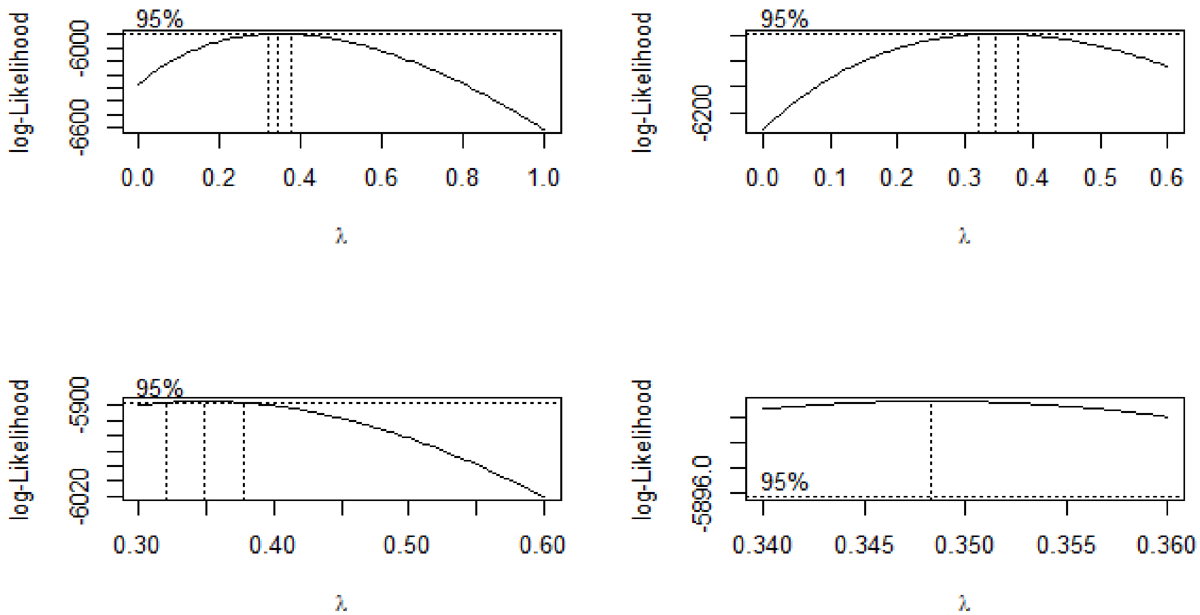


The Pie Chart (figure 3.1.6) shows the most popular genres which would be useful insights in predicting consumer preferences.

3.2 Transformation on the preliminary model

Since the preliminary model violates several assumptions regarding the linearity, homoscedasticity, and normality of residuals, as addressed in the proposal. We apply the Box-Cox transformation method, see figure 3.2.1 below:

Figure 3.2.1



The Box-Cox transformation suggests raising the preliminary model to the power of λ = 0.3482828.

<u>Transformed model:</u>

λ = 0.3482828

$$Gross\ Revenue^{\lambda} \sim Intercept + \beta_{Movie\ Genre} I(x_i = Movie\ Genre\ i) + \beta_{Release\ Year} Release\ Year^{\lambda} + \beta_{IMDB} IMDB^{\lambda} + \beta_{MetaScore} MetaScore^{\lambda} + \beta_{Budget} Budget^{\lambda}$$

3.2 Model Assumptions Check

Linearity Assumption

Figure 3.2.2 suggests that the transformed model is no longer violating Linearity Assumption since residuals are approximately evenly spread out with no patterns.

Figure 3.2.2

Residuals vs Fitted

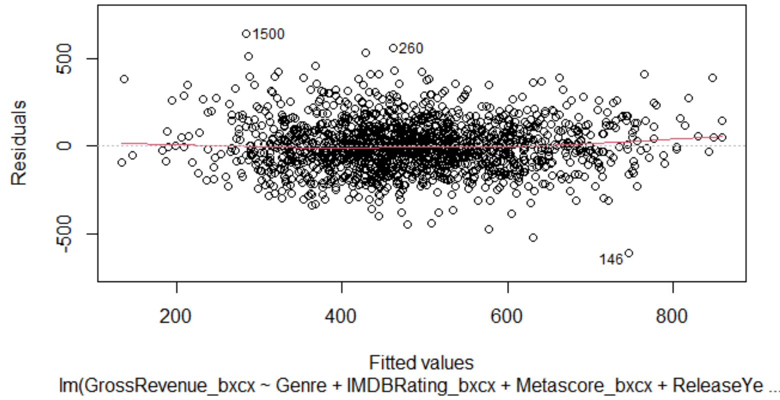
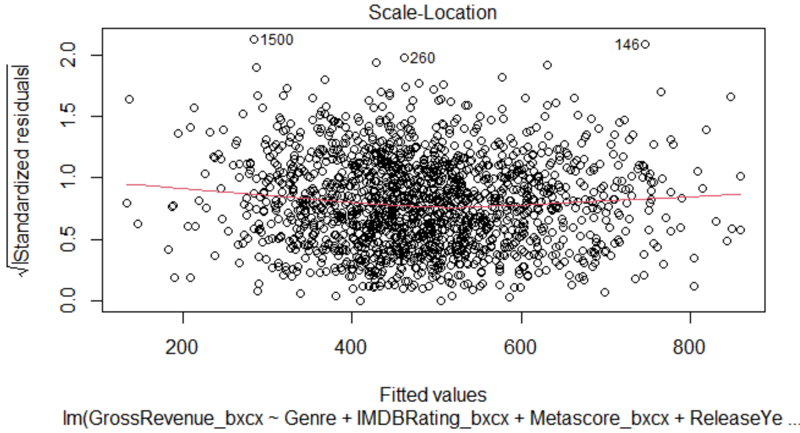


Figure 3.2.3

Scale-Location



Homoscedasticity

Figure 3.2.3 suggests a horizontal band of points with a relatively even spread and no discernible pattern across the entire range of fitted values.

Normality of Errors

Figure 3.2.4 shows that the model significantly improves the normality of errors assumption. However, slight deviations remain in the Q-Q plot, indicating that the transformation, while helpful, still has some limitations.

Figure 3.2.4

Q-Q Residuals

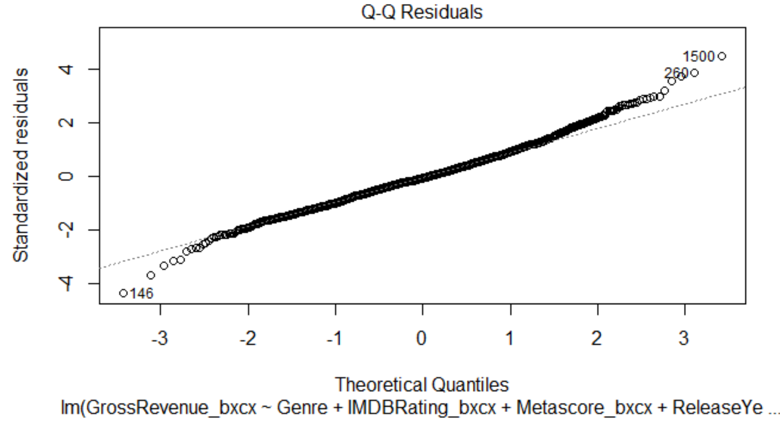


Figure 3.2.5

VIF Values for Each Predictor

Predictor	GVIF	Df	GVIF^(1/(2*Df))
Genre	1.530632	14	1.015319
IMDBRating_bxcx	2.442514	1	1.562854
Metascore_bxcx	2.461298	1	1.568853
ReleaseYear_bxcx	1.466445	1	1.210969
Budget_bxcx	1.632794	1	1.277808

Multicollinearity

Figure 3.2.5 shows demonstrates that all Box-Cox-transformed predictors exhibit no significant multicollinearity, as indicated by their adjusted GVIF values, which are all below the threshold of 5

3.3 Model's Variable Selecting

Overall F-Test

Null Model:

λ = 0.3482828

$$Gross\ Revenue^{\lambda} \sim\ Intercept$$

Full Model:

λ = 0.3482828

$$Gross\ Revenue^{\lambda} \sim\ Intercept\ +\ \beta_{Movie\ Genre} I(x_i = Movie\ Genre\ i)\ +\ \beta_{Release\ Year} Release\ Year^{\lambda} + \beta_{IMDB} IMDB^{\lambda} + \beta_{MetaScore} MetaScore^{\lambda} + \beta_{Budget} Budget^{\lambda}$$

Overall F-Test Hypothesis:

$$H_0:\ \beta_{Movie\ Genre} = \beta_{Release\ Year} = \beta_{IMDB} = \beta_{MetaScore} = \beta_{IMDB} = \beta_{Budget} = \beta^0 = 0$$

$$H_a:\ \beta_{Movie\ Genre} \neq \beta_{Release\ Year} \neq \beta_{IMDB} \neq \beta_{MetaScore} \neq \beta_{IMDB} \neq \beta_{Budget} \neq \beta^0 \neq 0$$

Figure 3.3.1

F Test on Overall Model

term	df.residual	rss	df	sumsq	statistic	p.value
GrossRevenue_bxcx ~ 1	1,611	59,372,232				NA
GrossRevenue_bxcx ~ Genre + IMDBRating_bxcx + Metascore_bxcx + ReleaseYear_bxcx + Budget_bxcx	1,593	35,452,371	18	23,919,861	59.71132	1.3e-163

From Figure 3.3.1, we reject the null hypothesis, indicating that at least one predictor significantly explains variance in Gross Revenue.

T-Test on all predictors

T-tests were performed on numerical predictors, excluding Genre, which was retained to capture its effects. To balance complexity and minimize error, we selected two significant numerical predictors at a 90% confidence level, ensuring a comprehensive yet controlled model.

Hypothesis:

$H_0: \beta_i = 0 \ - \ Predictor\ variable\ i\ has\ coefficient\ =\ 0$

$H_a: \beta_i \neq 0 \ - \ Predictor\ variable\ i\ has\ coefficient\ \neq\ 0$

Figure 3.3.2

T-Test on each predictor variable

Predictor	P_value
IMDBRating_bxcx	0.7796478377
Metascore_bxcx	0.7457768554
ReleaseYear_bxcx	0.0000000074
Budget_bxcx	0.0000000000
Votes	0.0000000000
Duration	0.0000000061

From Figure 3.3.2, transformed predictors are significant. We selected transformed Release Year, Budget, and Genre to ensure model complexity.

Partial F-Test on all subset of predictors
After the T-Test predictors process, the model thus far is:

$$Gross\ Revenue^{0.35} \sim\ Intercept\ +\ \beta_{Movie\ Genre} I(x_i\ =\ Movie\ Genre\ i)\ +\ \beta_{Release\ Year} Release\ Year^{0.35} + \beta_{Budget} Budget^{0.35}$$

Although IMDb Rating and Metascore failed individual T-tests, these tests ignore combined effects in a multiple regression model. To address this, Partial F-tests were conducted through following subsets tested:

1. Remove IMDb and Metascore
2. Remove IMDb
3. Remove Metascore
4. Remove Release Year and Budget
5. Remove Release Year
6. Remove Budget

<u>Subset 1: Removing IMDB and MetaScore</u>

Reduced Model

λ = 0.3482828

$$Gross\ Revenue^{\lambda} \sim\ Intercept\ +\ \beta_{Movie\ Genre} I(x_i\ =\ Movie\ Genre\ i)\ +\ \beta_{Release\ Year} Release\ Year^{\lambda} + \beta_{Budget} Budget^{\lambda}$$

Full Model

λ = 0.3482828

$$Gross\ Revenue^{\lambda} \sim\ Intercept\ +\ \beta_{Movie\ Genre} I(x_i\ =\ Movie\ Genre\ i)\ +\ \beta_{Release\ Year} Release\ Year^{\lambda} + \beta_{IMDB} IMDB^{\lambda} + \beta_{MetaScore} MetaScore^{\lambda} + \beta_{Budget} Budget^{\lambda}$$

Hypothesis

$$H_o:\ \beta_{Movie\ Genre} = \beta_{IMDB} = \beta_i^{0} =\ 0 :\ The\ removed\ predictors\ has\ coefficient\ equals\ 0$$

$H_a: \beta_{Movie\ Genre} \neq \beta_{IMDB} \neq \beta_i^0 \neq 0: At\ least\ one\ of\ the\ removed\ predictors\ has\ non\ zero\ coefficient$

The F-Test is conducted using ANOVA statistical method:

Partial F-Test Results

term	df.residual	rss	df	sumsq	statistic	p.value
GrossRevenue_bxcx ~ Genre + ReleaseYear_bxcx + Budget_bxcx	1,595	34,585,011				NA
GrossRevenue_bxcx ~ Genre + IMDBRating_bxcx + Metascore_bxcx + ReleaseYear_bxcx + Budget_bxcx	1,593	33,048,494	2	1,536,517	37.03152	1.9e-16

We reject the null hypothesis and conclude that at least one of the removed predictors is significant.

<u>Subset 2: Removing IMDB</u>

Reduced Model

λ = 0.3482828

$$Gross\ Revenue^{\lambda} \sim Intercept + \beta_{Movie\ Genre} I(x_i = Movie\ Genre\ i) + \beta_{Release\ Year} Release\ Year^{\lambda} + \beta_{MetaScore} MetaScore^{\lambda} + \beta_{Budget} Budget^{\lambda}$$

Full Model

λ = 0.3482828

$$Gross\ Revenue^{\lambda} \sim Intercept + \beta_{Movie\ Genre} I(x_i = Movie\ Genre\ i) + \beta_{Release\ Year} Release\ Year^{\lambda} + \beta_{IMDB} IMDB^{\lambda} + \beta_{MetaScore} MetaScore^{\lambda} + \beta_{Budget} Budget^{\lambda}$$

Hypothesis

$H_0: \beta_{IMDB} = \beta_i^0 = 0: The\ removed\ predictors\ has\ coefficient\ equals\ 0$

$H_a: \beta_{IMDB} \neq \beta_i^0 \neq 0: The\ removed\ predictors\ has\ non\ zero\ coefficient$

The F-Test is conducted using ANOVA statistical method:

Partial F-Test Results

term	df.residual	rss	df	sumsq	statistic	p.value
GrossRevenue_bxcx ~ Genre + Metascore_bxcx + ReleaseYear_bxcx + Budget_bxcx	1,594	33,363,767				NA
GrossRevenue_bxcx ~ Genre + IMDBRating_bxcx + Metascore_bxcx + ReleaseYear_bxcx + Budget_bxcx	1,593	33,048,494	1	315,272.5	15.19673	1e-04

We reject the null hypothesis and conclude that the transformed IMDB rating is significant.

Subset 3: Removing Metascore

Reduced Model

λ = 0.3482828

$$Gross\ Revenue^{\lambda} \sim\ Intercept\ +\ \beta_{Movie\ Genre} I(x_i = Movie\ Genre\ i)\ +\ \beta_{Release\ Year} Release\ Year^{\lambda} + \beta_{IMDB} IMDB^{\lambda} + \beta_{Budget} Budget^{\lambda}$$

Full Model

λ = 0.3482828

$$Gross\ Revenue^{\lambda} \sim\ Intercept\ +\ \beta_{Movie\ Genre} I(x_i = Movie\ Genre\ i)\ +\ \beta_{Release\ Year} Release\ Year^{\lambda} + \beta_{IMDB} IMDB^{\lambda} + \beta_{MetaScore} MetaScore^{\lambda} + \beta_{Budget} Budget^{\lambda}$$

Hypothesis

$$H_0: \beta_{Metascore} = \beta_i^0 = 0 : The\ removed\ predictors\ has\ coefficient\ equals\ 0$$

$$H_a: \beta_{Metascore} \neq \beta_i^0 \neq 0 : The\ removed\ predictors\ has\ non\ zero\ coefficient$$

The F-Test is conducted using ANOVA statistical method:

Partial F-Test Results

term	df.residual	rss	df	sumsq	statistic	p.value
GrossRevenue_bxcx ~ Genre + IMDBRating_bxcx + ReleaseYear_bxcx + Budget_bxcx	1,594	33,226,107				NA
GrossRevenue_bxcx ~ Genre + IMDBRating_bxcx + Metascore_bxcx + ReleaseYear_bxcx + Budget_bxcx	1,593	33,048,494	1	177,612.6	8.561263	3.5e-03

We reject the null hypothesis and conclude that the transformed Metascore is significant.

Subset 4: Removing Release Year and Budget

Reduced Model

λ = 0.3482828

$$Gross\ Revenue^{\lambda} \sim\ Intercept\ +\ \beta_{Movie\ Genre} I(x_i = Movie\ Genre\ i)\ +\ \beta_{IMDB} IMDB^{\lambda} + \beta_{Metascore} Metascore^{\lambda}$$

Full Model

λ = 0.3482828

$$Gross\ Revenue^{\lambda} \sim\ Intercept\ +\ \beta_{Movie\ Genre} I(x_i = Movie\ Genre\ i)\ +\ \beta_{Release\ Year} Release\ Year^{\lambda}$$
$$+\ \beta_{IMDB} IMDB^{\lambda} + \beta_{MetaScore} MetaScore^{\lambda} + \beta_{Budget} Budget^{\lambda}$$

Hypothesis

$$H_o:\ \beta_{Release\ Year} = \beta_{Budget} = \beta_i^0 = 0:\ The\ removed\ predictors\ has\ coefficient\ equals\ 0$$

$$H_a:\ \beta_{Release\ Year} \neq \beta_{Budget} \neq \beta_i^0 \neq 0:\ At\ least\ one\ of\ the\ removed\ predictors\ has\ non\ zero\ coefficient$$

The F-Test is conducted using ANOVA statistical method:

Partial F-Test Results

term	df.residual	rss	df	sumsq	statistic	p.value
GrossRevenue_bxcx ~ Genre + IMDBRating_bxcx + Metascore_bxcx	1,595	48,404,092				NA
GrossRevenue_bxcx ~ Genre + IMDBRating_bxcx + Metascore_bxcx + ReleaseYear_bxcx + Budget_bxcx	1,593	33,048,494	2	15,355,598	370.0845	9.9e-133

The test yields p-value less than $\alpha = 0.05$ $(95\%\ significance\ level)$. We reject the null hypothesis and conclude that at least one of the removed predictors is significant.

<u>Subset 5: Removing Release Year</u>

Reduced Model

λ = 0.3482828

$$Gross\ Revenue^{\lambda} \sim\ Intercept\ +\ \beta_{Movie\ Genre} I(x_i = Movie\ Genre\ i)\ +\ \beta_{IMDB} IMDB^{\lambda} + \beta_{Metascore} Metascore^{\lambda} +\ \beta_{Budget} Budget^{\lambda}$$

Full Model

λ = 0.3482828

$$Gross\ Revenue^{\lambda} \sim\ Intercept\ +\ \beta_{Movie\ Genre} I(x_i = Movie\ Genre\ i)\ +\ \beta_{Release\ Year} Release\ Year^{\lambda} + \beta_{IMDB} IMDB^{\lambda} + \beta_{MetaScore} MetaScore^{\lambda} + \beta_{Budget} Budget^{\lambda}$$

Hypothesis

$$H_o:\ \beta_{Release\ Year} = \beta_{Budget} = \beta_i^{0} = 0 :\ The\ removed\ predictors\ has\ coefficient\ equals\ 0$$

$$H_a:\ \beta_{Release\ Year} \neq \beta_{Budget} \neq \beta_i^{0} \neq 0 :\ The\ removed\ predictors\ has\ non\ zero\ coefficient$$

The F-Test is conducted using ANOVA statistical method:

Partial F-Test Results

term	df.residual	rss	df	sumsq	statistic	p.value
GrossRevenue_bxcx ~ Genre + IMDBRating_bxcx + Metascore_bxcx + Budget_bxcx	1,594	33,402,222				NA
GrossRevenue_bxcx ~ Genre + IMDBRating_bxcx + Metascore_bxcx + ReleaseYear_bxcx + Budget_bxcx	1,593	33,048,494	1	353,727.8	17.05035	3.8e-05

The test yields p value less than $\alpha = 0.05$ $(95\%\ significance\ level)$, we reject the null hypothesis and conclude that the transformed Release Year is a significant predictor.

<u>**Subset 6: Removing Budget**</u>

Reduced Model

λ = 0.3482828

$$Gross\ Revenue^{\lambda} \sim Intercept + \beta_{Movie\ Genre} I(x_i = Movie\ Genre\ i) + \beta_{IMDB} IMDB^{\lambda} + \beta_{Metascore} Metascore^{\lambda} + \beta_{Release\ Year} Release\ Year^{\lambda}$$

Full Model

λ = 0.3482828

$$Gross\ Revenue^{\lambda} \sim Intercept + \beta_{Movie\ Genre} I(x_i = Movie\ Genre\ i) + \beta_{Release\ Year} Release\ Year^{\lambda} + \beta_{IMDB} IMDB^{\lambda} + \beta_{MetaScore} MetaScore^{\lambda} + \beta_{Budget} Budget^{\lambda}$$

Hypothesis

$$H_o:\ \beta_{Release\ Year} = \beta_{Budget} = \beta_i^0 = 0 :\ The\ removed\ predictors\ has\ coefficient\ equals\ 0$$

$$H_a:\ \beta_{Release\ Year} \neq \beta_{Budget} \neq \beta_i^0 \neq 0 :\ The\ removed\ predictors\ has\ non\ zero\ coefficient$$

The F-Test is conducted using ANOVA statistical method:

Partial F-Test Results

term	df.residual	rss	df	sumsq	statistic	p.value
GrossRevenue_bxcx ~ Genre + IMDBRating_bxcx + Metascore_bxcx + ReleaseYear_bxcx	1,594	47,217,754				NA
GrossRevenue_bxcx ~ Genre + IMDBRating_bxcx + Metascore_bxcx + ReleaseYear_bxcx + Budget_bxcx	1,593	33,048,494	1	14,169,259	682.9851	1.4e-125

The test yields p value less than $\alpha = 0.05$ (95% $significance\ level$), we reject the null hypothesis and conclude that the transformed Budget is a significant predictor.

<u>Subset 7: Removing Genre</u>

Reduced Model

λ = 0.3482828

$$Gross\ Revenue^{\lambda} \sim Intercept + \beta_{IMDB} IMDB^{\lambda} + \beta_{Metascore} Metascore^{\lambda} + \beta_{Release\ Year} Release\ Year^{\lambda} + \beta_{Budget} Budget^{\lambda}$$

Full Model

λ = 0.3482828

$$Gross\ Revenue^{\lambda} \sim Intercept + \beta_{Movie\ Genre} I(x_i = Movie\ Genre\ i) + \beta_{Release\ Year} Release\ Year^{\lambda} + \beta_{IMDB} IMDB^{\lambda} + \beta_{MetaScore} MetaScore^{\lambda} + \beta_{Budget} Budget^{\lambda}$$

Hypothesis

$$H_o:\ \beta_{Release\ Year} = \beta_{Budget} = \beta_i^0 = 0 :\ The\ removed\ predictors\ has\ coefficient\ equals\ 0$$

$$H_a:\ \beta_{Release\ Year} \neq \beta_{Budget} \neq \beta_i^0 \neq 0 :\ The\ removed\ predictors\ has\ non\ zero\ coefficient$$

The F-Test is conducted using ANOVA statistical method:

Partial F-Test Results

term	df.residual	rss	df	sumsq	statistic	p.value
GrossRevenue_bxcx ~ IMDBRating_bxcx + Metascore_bxcx + ReleaseYear_bxcx + Budget_bxcx	1,607	35,433,473				NA
GrossRevenue_bxcx ~ Genre + IMDBRating_bxcx + Metascore_bxcx + ReleaseYear_bxcx + Budget_bxcx	1,593	33,048,494	14	2,384,979	8.211465	3e-17

We reject the null hypothesis and conclude that the Genre is significant.

According to the variables selecting process, all predictors in the preliminary model are significant.

Full Model:

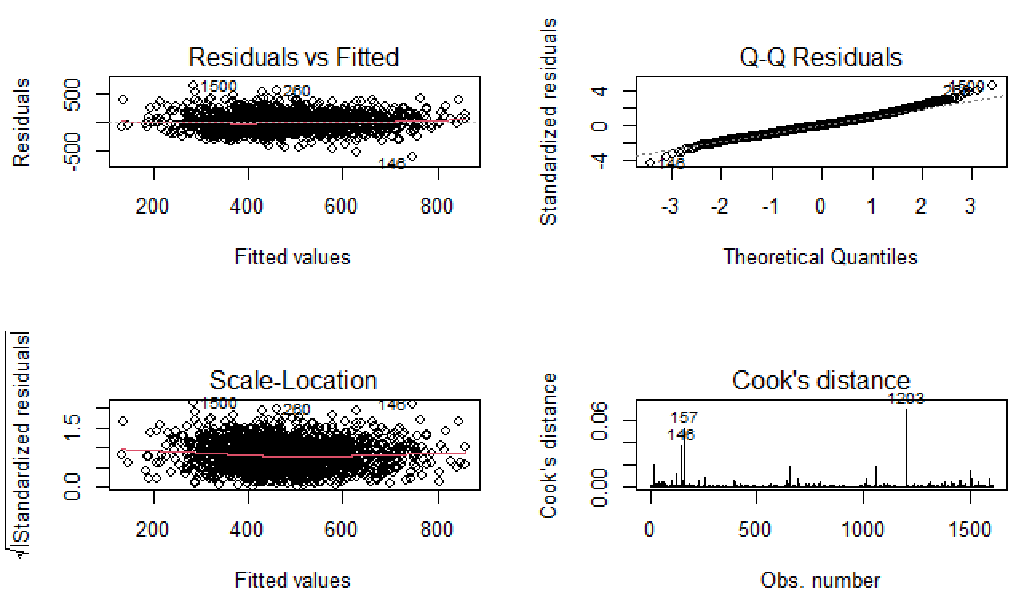
λ = 0.3482828

$$Gross\ Revenue^{\lambda} \sim Intercept + \beta_{Movie\ Genre} I(x_i = Movie\ Genre\ i) + \beta_{Release\ Year} Release\ Year^{\lambda} + \beta_{IMDB} IMDB^{\lambda} + \beta_{MetaScore} MetaScore^{\lambda} + \beta_{Budget} Budget^{\lambda}$$

3.4 Model Validation

Detecting Influential Observations

Figure 3.4.1



From the diagnostic plots (Figure 3.4.1), observations 146, 157, 260, 1203, and 1500 require further investigation. Additionally, R automatically identified and excluded high-leverage points (139, 980, 1379). Despite this exclusion, these points should still be reviewed to understand the underlying issues.

Detecting Outliers

The standardized residual is being compute using this formula:

$$r_i = \frac{\hat{e}_i}{\hat{\sigma}\sqrt{1-h_{i,i}}},\ where\ \hat{e}_i = Y_i - \widehat{Y_i}$$

The cutoff rule:

$$\left|r_i\right| \geq 4$$

Standardized residual of all problematic observations

Observations Standardized Residuals

Observation	Standardized_Residuals
146	4.343728
157	0.000000
260	3.899846
1,203	0.000000
1,500	4.521234

According to the cutoff rules, observations 146 and 1500 are outliers.

Detecting High Leverage Points

The leverage values are calculate using the formula:

$$h_{i,i} = x_i^T (X^T X)^{-1} x_i$$

The cutoff rule is:

$$h_{i,i} \geq 2(p + 1)/n$$

$, where\ 2(p + 1)/n\ is\ 2\ times\ the\ average\ leverage\ value\ = 0.0235732$

Leverage of all problematic observations

Leverage Values for Problematic Observations

Observation	Leverage_Value
146	0.03655793
157	1.00000000
260	0.01002138
1,203	1.00000000
1,500	0.01333521
139	1.00000000
980	1.00000000
1,379	1.00000000

According to cutoff rules, observations 146, 157, 1203, 139, 980, and 1379 are high-leverage points, aligning with R's automatic detection.

Detecting Influential Points

The Cook's distance are calculate using the formula:

$$D_i = \frac{r_i^2}{(p+1)} \times \frac{h_{i,i}}{(1-h_{i,i})}$$

The cutoff rules is

$$D_i > m\ ,\ where\ \ m\ =\ median\ of\ F(p\ +\ 1, n\ -\ p\ -\ 1)\ =\ 0.8917277$$

Cook's distance of all problematic observations

Cook's Distance for Problematic Observations

Observation	Cook_Distance
146	0.037681449
157	0.060974819
260	0.008102948
1,203	0.057547859
1,500	0.014540880

Figure 3.4.3 shows none of the observations are influential points

Investigating individual observations

Investigating Problematic Observations

MovieName	ReleaseYear	Duration	IMDBRating	Metascore	Votes	Genre	GrossRevenue	Budget
Metropolis	1,927	153	8.3	98	184,838	Drama	1,240,000	92,620,000
The Outlaw Josey Wales	1,976	135	7.8	69	79,553	Western	31,800,000	3,700,000
E.T. the Extra-Terrestrial	1,982	115	7.9	92	435,984	Adventure	435,110,000	10,500,000
Red Eye	2,005	85	6.5	71	135,739	Thriller	57,890,000	26,000,000
Alice in Wonderland	2,010	108	6.4	53	439,658	Adventure	334,190,000	3,000,000
Willy Wonka & the Chocolate Factory	1,971	100	7.8	67	225,965	Family	4,000,000	3,000,000
Super Size Me	2,004	100	7.2	73	113,030	Documentary	11,530,000	65,000
The Invasion	2,007	99	5.9	45	82,519	Sci-Fi	15,070,000	80,000,000

Upon investigation, no input errors or abnormalities were found, so all observations were retained except for 139, 980, and 1379, which were removed by R's automatic detection.

Model Fit and Performance

Summary of Initial Model:

Initial Model Coefficients

Predictor	Estimate	Std_Error	T_Value	P_Value
(Intercept)	-34,084,712.884645	269,105,898.04609740	-0.12665911	9.0e-01
GenreAdventure	-1,258,169.545376	6,118,988.23829901	-0.20561725	8.4e-01
GenreAnimation	14,771,187.264145	7,121,403.45144576	2.07419610	3.8e-02
GenreBiography	-24,976,468.719290	7,058,577.67609678	-3.53845631	4.1e-04
GenreComedy	6,100,188.047321	4,426,302.24875676	1.37816798	1.7e-01
GenreCrime	-30,892,354.317049	6,109,480.96132373	-5.05646134	4.8e-07
GenreDocumentary	-29,052,132.709451	59,122,361.54429536	-0.49138992	6.2e-01
GenreDrama	-19,252,633.769349	4,827,178.61618345	-3.98838230	7.0e-05
GenreFamily	-48,375,104.135244	59,135,362.92262572	-0.81804020	4.1e-01
GenreFantasy	-13,231,375.629816	22,482,505.34262254	-0.58851874	5.6e-01
GenreHorror	7,174,859.514103	7,634,326.94474853	0.93981559	3.5e-01
GenreMystery	-25,750,935.552996	22,546,761.34842595	-1.14211239	2.5e-01
GenreSci-Fi	-81,700,163.895556	59,031,715.37031154	-1.38400457	1.7e-01
GenreThriller	1,345,271.800385	59,100,781.19690199	0.02276234	9.8e-01
GenreWestern	-21,767,679.679126	59,108,698.49680772	-0.36826525	7.1e-01
IMDBRating	15,768,152.775181	2,527,986.90873305	6.23743451	5.7e-10
Metascore	299,246.050218	129,877.38550019	2.30406586	2.1e-02
ReleaseYear	-30,328.374239	133,394.62439546	-0.22735829	8.2e-01
Budget	1.065322	0.04258071	25.01888613	8.6e-117

Initial Model Fit Statistics

Metric	Value
Residual Standard Error	58951554
Multiple R-squared	0.3938
Adjusted R-squared	0.387
F-statistic	57.5 (df1 = 18, df2 = 1593)

Summary of Final Model:

Final Model Coefficients

Predictor	Estimate	Std_Error	T_Value	P_Value
(Intercept)	7,475.7917525	1,923.54993670	3.88645577	1.1e-04
GenreAdventure	-11.1734165	14.92288134	-0.74874391	4.5e-01
GenreAnimation	34.7458915	17.37833630	1.99937962	4.6e-02
GenreBiography	-54.2701657	17.09738488	-3.17417933	1.5e-03
GenreComedy	22.8310986	10.74027888	2.12574542	3.4e-02
GenreCrime	-86.3707471	14.85961409	-5.81244887	7.4e-09
GenreDocumentary	50.9822898	144.84707414	0.35197321	7.2e-01
GenreDrama	-56.0222577	11.72345596	-4.77864701	1.9e-06
GenreFamily	-192.8701771	144.50194326	-1.33472376	1.8e-01
GenreFantasy	1.2565569	54.97068445	0.02285867	9.8e-01
GenreHorror	56.9608598	18.99100515	2.99935993	2.7e-03
GenreMystery	-52.3093523	55.07016125	-0.94986743	3.4e-01
GenreSci-Fi	-256.6678581	144.23088429	-1.77956240	7.5e-02
GenreThriller	31.5070194	144.35726865	0.21825724	8.3e-01
GenreWestern	12.9535839	144.43995040	0.08968145	9.3e-01
IMDBRating_bxcx	229.3876928	58.84302706	3.89829865	1.0e-04
Metascore_bxcx	36.1101418	12.34128181	2.92596364	3.5e-03
ReleaseYear_bxcx	-558.7204884	135.30938919	-4.12920708	3.8e-05
Budget_bxcx	0.7631775	0.02920249	26.13398433	1.4e-125

Final Model Fit Statistics

Metric	Value
Residual Standard Error	144
Multiple R-squared	0.4028
Adjusted R-squared	0.396
F-statistic	59.69 (df1 = 18, df2 = 1593)

Answering Research Question

The final model's adjusted R-squared (0.396) shows little improvement over the initial model (0.387) but does not violate assumptions.

IMDb Rating: A coefficient of 229.39 indicates gross revenue increases by 229.39 units for each one-unit IMDb rating increase, with a highly significant p-value.

Metascore: A coefficient of 36.11 shows an average gross revenue increase of 36.11 units per one-unit Metascore increase, also statistically significant.

Release Year: A negative coefficient (-558.72) suggests later releases earn less, reflecting industry trends or rising competition.

Production Cost: A coefficient of 0.76 implies gross revenue rises by 0.76 units for every unit increase in transformed budget, a highly significant result.

Movie Genres:

- **Crime**: Largest negative impact (-86.37), with lower gross revenue compared to the reference category.
- **Horror**: Positive impact (56.96), indicating higher gross revenue compared to the reference category.
- Other genres like Animation and Comedy also show significant effects, reflecting audience preferences.

Intercept: At 7,475.79, it represents baseline gross revenue when all predictors are at reference values or zero (transformed scale).

Session 4: Limitations and Conclusion

In conclusion, the multiple linear regression model shows that the mentioned predictors collectively explain some of the variability in gross revenue. Higher IMDB ratings, Metascores and production budgets are positively associated with increased gross revenue. In the meanwhile, the release year showed a significant negative relationship, reflecting changes in the industry over time. These results are consistent with existing literature that emphasizes the importance of critics, production budgets, and genre choice in driving box office performance.

From the model, the coefficient of IMDB rating is 229.39. This aligns with the intuition that audience perception significantly impacts a movie's commercial success. Genres such as Horror and Animation have a positive impact on revenue, whereas genres like Crime and Drama negatively affect movie’s box office performance. Finally, the negative impact of Release Year on Gross Revenue is somewhat surprising and could reflect the increasing competition in the industry.

Despite these findings, there exist lots of limitations. The model exhibited a low adjusted R-squared, suggesting that while the predictors are statistically significant, a big portion of the variability in gross revenue remains unexplained. This is unsurprising given the stochastic nature of the film industry, where unpredictable factors such as audience trends and external economic conditions through De Vany and Walls’s study (Pangarker, 2013). Furthermore, the quantile-quantile (QQ) plot does not fully validate even after applying Box-Cox transformations from right tailed skewed distribution. Additionally, the dataset contained numerous outliers and high-leverage points, which are likely to influence the model’s estimates.

While the proposed linear model provides useful insights, it can not fully capture the complex unexpectancy of movie revenue. We hope that future research could employ different modeling approaches, such as machine learning methods to better handle the problem and account for outliers. In future studies, expanding the dataset to include factors like marketing budgets, streaming performance, and social media engagement could potentially improve predictive accuracy. Despite the limitations, this analysis offers a meaningful exploration of the factors that influence movie gross revenue.

Bibliography

Original Datasets:

Sawhney, P. (n.d.). *IMDB Dataset*. Kaggle. Retrieved October 1, 2024, from https://www.kaggle.com/datasets/prishasawhney/imdb-dataset-top-2000-movies

Banik, R. (n.d.). *The Movie Dataset*. Kaggle. Retrieved October 1, 2024, from https://www.kaggle.com/datasets/rounakbanik/the-movies-dataset/data?select=movies_metadata.csv

Related academic paper:

Pangarker, N. A., & Smit, E. v. d. M. (2013). The determinants of box office performance in the Film Industry Revisited. *South African Journal of Business Management*, *44*(3), 47–58. https://doi.org/10.4102/sajbm.v44i3.162

Moon, S., Bergey, P. K., & Iacobucci, D. (2010). Dynamic Effects among Movie Ratings, Movie Revenues, and Viewer Satisfaction. *Journal of Marketing*, *74*(1), 108–121. http://www.jstor.org/stable/20619083

Alec, Kenedy. (2010). Predicting Box Office Success: Do Critical Reviews Really Matter? University of California at Berkeley. https://www.stat.berkeley.edu/~aldous/157/Old_Projects/kennedy.pdf