Motion Picture Revenue Prediction Model

By Ryan Chou

Dec, 15th, 2025

Introduction

Cinema has been a major industry since the initial growth from 19th century. Modern films often involved multimillion dollars investments. Studio executives must allocate production budgets efficiently, selecting projects with the strongest potential for financial success. Therefore, it is crucial and valuable to identify factors that reliably predict a film's earnings.

The purpose of this report is to develop a statistical model capable of predicting a film's worldwide revenue using a variety of observable characteristics. A central question of the report is to forecast whether higher production budgets lead to higher future revenue. Additionally, I will be discussing whether such spending is justified once other film factors, such as genre, runtime, viewer rating, and director attributes. Particular emphasis is placed on interpretability so that results can be communicated to non-statistical decision-makers.

The dataset contained around 100 films. Every movie is described using nine variables related to film characteristics, production information, and financial outcomes. Prior to modeling section, we will conduct Exploratory Data Analysis (EDA) to better understand the structure of the data.

Exploratory Data Analysis (EDA)

The variables included are as follows: The variable **movie_title** served only as an identifier and was excluded from modeling. Film characteristics include **production_date**, **runtime_minutes**, **movie_averageRating**, and **genres**. Director information is captured by **director_name** and **birth_year**. **Budget** and **gross**, which represent production costs and global revenue in nominal dollars, are used to measure financial outcomes of interest.

The distribution of global gross revenue is shown in Figure 1(A). A few movies make disproportionately large sums of money, and the distribution is heavily skewed to the right. The

use of logarithmic transformations for both gross revenue and budget in later modeling is motivated by this noticeable skewness and rising variance at higher revenue levels. The distribution of log(gross) is shown in Figure 1(B). The log-transformed values are more evenly distributed than the original revenue figures, giving the data a more balanced appearance. This implies that the log transformation is a suitable option for regression modeling.

Figures 1(C) and 1(D) illustrate the relationship between revenue and budget. In Figure 1(C), we see that movies with larger budgets typically generate more revenue, but the pattern isn't a straight line, and there is significant variation. When we use the log transformation, as shown in Figure 1(D), the relationship between log(gross) and log(budget) becomes much clearer and almost forms a straight line.

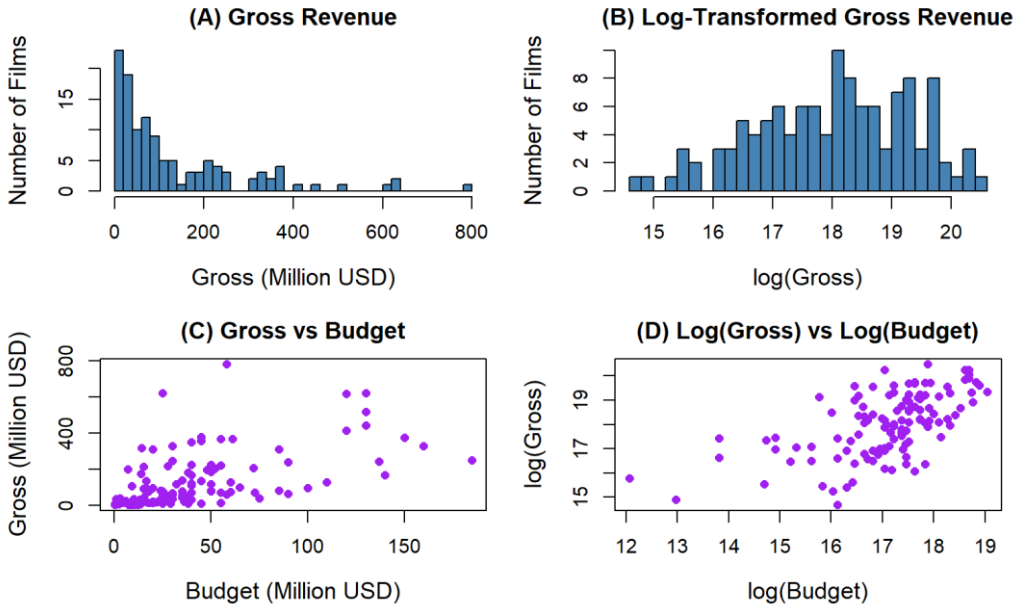

Figure 1.

Each movie has up to three genre labels in the genres variable, which are arranged alphabetically. To enable films to belong to more than one genre simultaneously, I assigned them with binary indicator variables (e.g. is_drama, is_comedy). The dataset contains 18 different genres with wildly different representations. I applied a threshold of 10 observations per genre, retaining 10

major genres: Drama (77 films), Comedy (41), Action (32), Romance (30), Adventure (19), Crime (19), Thriller (19), Mystery (12), Biography (10), and Family (10). Eight genres with insufficient observations were excluded to prevent overfitting and unreliable coefficient estimates (see Table 2 and 3 for complete distribution). These 10 genres account for 96% of the dataset.

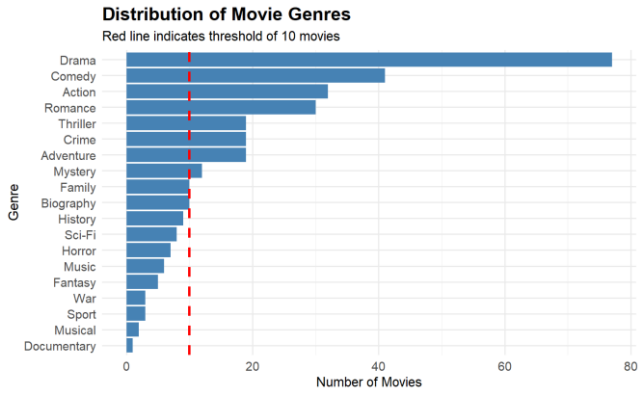

Table 2.

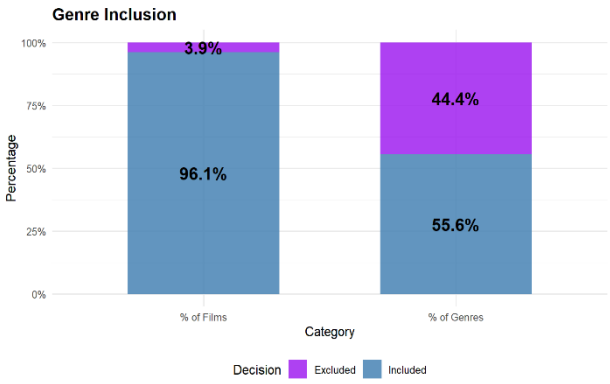

Table 3.

Model Selection

I started with the full linear regression model for log scale of worldwide gross revenue on log(budget), viewer rating, runtime, and production year, along with dummy variables for the top genre categories. Since the structure of the genre categories is listed alphabetically, they were treated more like unordered tags and expressed by dummy variables. It enables films to have more than one genre included. This ensured the exclusion of categories with fewer than 10 instances to prevent less accurate estimation. To create the optimal model, I applied backward stepwise selection using AIC. The method removes predictors that didn't contribute significantly to the fit. The resulting model contained log(budget), movie_averageRating, runtime_minutes, as well as genre variables is_drama, is_romance, is_thriller, and is_family. The final selection(Figure 4):

Let $Y_i = log_(gross_i)$,

$$Y_i = \beta_0 + \beta_1 log_(budget_i) + \beta_2(rating_i) + \beta_3(runtime_i) + \beta_4((isdrama)_i) + \beta_5(isromance)_i) + \beta_5(isthriller)_i) + \beta_7(isfamily)_i) + \epsilon_i$$

Figure 4.

Each genre indicator is equal to 1 if the movie exhibits that genre characteristic, and 0 otherwise. Note that these coefficients reflect the effect of each genre characteristic, conditional on the inclusion of all other genre characteristics. The final model shows strong explanatory power, with an adjusted R-squared value of around 0.58. It indicates a significant portion of the variation in log-transformed revenue is accounted for by the predictors used.

Results & Interpretation

Since the response variable is log-transformed, coefficients are interpreted as multiplicative effects on expected revenue.

- **Budget:** The value of the coefficient for 'log(budget)' is approximately 0.63, signifying a unit increase in 'budget' will translate to a 0.63% increase in expected worldwide revenue. This indicates diminishing returns.
- **Viewer Rating:** movie_averageRating's coefficient is 0.71. It means that a one-unit increase in the rating increases box office earnings by a factor of nearly twice the magnitude of the rating.
- **Runtime:** The fact that the coefficient for runtime is negative indicates that longer movies tend to make lower profits.
- **Genres:** Differences by genre appear: Drama Movies are less profitable, Family Movies are more profitable, while Romance has a positive impact and Thriller has limited statistical significance. This may indicate the available data does not provide sufficient evidence to confidently distinguish thriller films from the baseline genre. Potentially due to greater variability in financial performance or a smaller number of observations in this category.

Model Fit and Diagnostics

Figure 5 shows diagnostic plots of the final model. It seems residuals are equally spaced in normal and Q-Q plot. Therefore, residuals are normally and equally distributed. In scales and locations, variance is similar at all levels, so whether variance is equal is met. In residuals and leverage charts, there are some influential observations, but none of them exceed threshold levels of Cook's D statistics, and some of them are more influential compared with others.

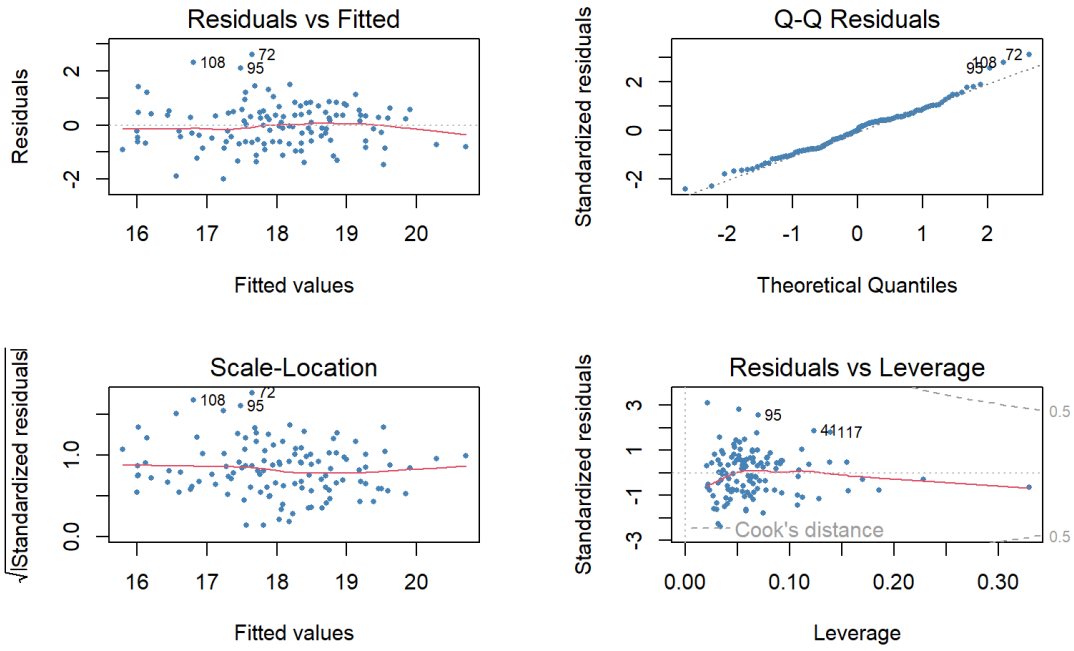

Figure 5.

Budget Justification

The model points out that increased production budgets result in increased expected worldwide box office revenues. Specifically, the estimated budget elasticity is approximately 0.63, implying that a 1% increase in budget is associated with a 0.63% increase in expected revenue. Furthermore, it is implied that incrementally higher spending would incrementally increase box office revenues.

However, the studios normally give a bigger budget to movies that involve franchises, talent, or huge marketing. These variables are not directly observed in a dataset. This might lead to overestimation of the coefficient of the budget. Therefore, including additional data on marketing expenditures, franchise status, or star power would help produce a more accurate estimate of the true effect of budget on revenue.

Prediction for New Film

By using the final model, I predicted the revenue for the film following the characteristics shown in Figure 7. Expected global gross revenue is projected by the model at around **$65 million USD**. The corresponding 95% prediction interval is quite broad, range from **$12 million USD to $358 million USD.** It incorporated the natural uncertainty and volatility associated with box office revenues, despite observable film attributes.

Production date	Run time	Rating	Genre	Budget
June 15, 2013	2hrs	7.0	Comedy	$10 million

Figure 7.

Conclusion

In conclusion, the analysis demonstrated the production budget, ratings, and genre proportions to be key variables correlated with worldwide box office gross revenue. While increasing budgets are related to higher profits, the possibility of diminishing returns and other factors being present makes it clear that simply relying on increased spending is not an effective method for the budget and profits. In the same manner, other unobserved variables, including the franchise nature of the particular film and the number of star performances could also be taking a role in budget expenditure and revenue generation to the extent that budget expenditure alone would not be the singular factor to optimize to ensure the success of any particular film release.

Appendix

```r
library(readxl)
library(tidyverse)
library(ggplot2)
final_data <- read_excel(
  "C:/Users/rchou/OneDrive/Ryan/Ohio State/Courses/5th Semester Fall 2025/Sta
t 3301/Final/dataset10-1.xlsx"
)

final_data$log_gross <- log(final_data$gross)
final_data$log_budget <- log(final_data$budget)

# Figure 1a
par(mfrow = c(2, 2), mar = c(4.5, 4.5, 2.5, 1))
hist(
  final_data$gross / 1e6,
  breaks = 30,
  main = "(A) Gross Revenue",
  xlab = "Gross (Million USD)",
  ylab = "Number of Films",
  col = "steelblue",
  border = "black",
  cex.main = 1.2,
  cex.lab = 1.2
)

# Figure 1b
hist(
  final_data$log_gross,
  breaks = 30,
  col = "steelblue",
  main = "(B) Log-Transformed Gross Revenue",
  xlab = "log(Gross)",
  ylab = "Number of Films",
  border = "black",
  cex.main = 1.2,
  cex.lab = 1.2
)

# Figure 1c
plot(
  final_data$budget / 1e6,
  final_data$gross / 1e6,
  main = "(C) Gross vs Budget",
  xlab = "Budget (Million USD)",
  ylab = "Gross (Million USD)",
  col = "purple",
  pch = 16,
```

```r
  cex.main = 1.2,
  cex.lab = 1.2
)

# Figure 1d
plot(
  final_data$log_budget,
  final_data$log_gross,
  main = "(D) Log(Gross) vs Log(Budget)",
  xlab = "log(Budget)",
  ylab = "log(Gross)",
  col = "purple",
  pch = 16,
  cex.main = 1.2,
  cex.lab = 1.2
)
par(mfrow = c(1, 1))

par(mfrow = c(2, 1))
# Figure 2
all_genres <- final_data$genres %>%
  str_split(",") %>%
  unlist() %>%
  str_trim() %>%
  unique() %>%
  sort()

for (genres in all_genres) {
  col_name <- paste0("is_", genres)
  final_data[[col_name]] <- as.integer(str_detect(final_data$genres, fixed(ge
nres)))
}

# Figure 3
genre_cols <- names(final_data)[str_detect(names(final_data), "^is_")]
genre_counts <- final_data %>%
  select(all_of(genre_cols)) %>%
  summarise(across(everything(), sum)) %>%
  pivot_longer(everything(), names_to = "Genre", values_to = "Count") %>%
  mutate(Genre = str_remove(Genre, "^is_")) %>%
  arrange(desc(Count))

ggplot(genre_counts, aes(x = reorder(Genre, Count), y = Count)) +
  geom_col(fill = "steelblue") +
  geom_hline(
    yintercept = 10,
    color = "red",
    linetype = "dashed",
    linewidth = 1
  ) +
```

```r
  coord_flip() +
  labs(
    title = "Distribution of Movie Genres",
    subtitle = "Red line indicates threshold of 10 movies",
    x = "Genre",
    y = "Number of Movies"
  ) +
  theme_minimal() +
  theme(plot.title = element_text(face = "bold", size = 15),
        axis.text = element_text(size = 10))

summary_data <- data.frame(
  Decision = c("Included", "Excluded"),
  Genres = c(10, 8),
  Films = c(269, 11)
)

plot_data <- summary_data %>%
  mutate(Genre_Pct = Genres / sum(Genres) * 100,
         Film_Pct = Films / sum(Films) * 100) %>%
  select(Decision, Genre_Pct, Film_Pct) %>%
  pivot_longer(
    cols = c(Genre_Pct, Film_Pct),
    names_to = "Category",
    values_to = "Percentage"
  ) %>%
  mutate(Category = recode(Category, "Genre_Pct" = "% of Genres", "Film_Pct"
= "% of Films"))

ggplot(plot_data, aes(x = Category, y = Percentage, fill = Decision)) +
  geom_col(position = "fill",
           alpha = 0.85,
           width = 0.6) +
  geom_text(
    aes(label = sprintf("%.1f%%", Percentage)),
    position = position_fill(vjust = 0.5),
    color = "black",
    fontface = "bold",
    size = 5
  ) +
  scale_y_continuous(labels = scales::percent_format(scale = 100)) +
  scale_fill_manual(values = c(
    "Included" = "steelblue",
    "Excluded" = "purple"
  )) +
  labs(title = "Genre Inclusion", y = "Percentage", ) +
  theme_minimal() +
  theme(plot.title = element_text(face = "bold", size = 14),
        legend.position = "bottom")
par(mfrow = c(1, 1))
```

```r
# Modeling
library(lubridate)

final_data$production_date <- mdy(final_data$production_date)
final_data$production_year <- year(final_data$production_date)

all_genres <- trimws(unlist(strsplit(final_data$genres, ",")))
genre_counts <- sort(table(all_genres), decreasing = TRUE)

major_genres <- names(genre_counts[genre_counts >= 10])
major_genres
for (g in major_genres) {
  final_data[[paste0("is_", tolower(g))]] <-
    as.integer(grepl(paste0("\\b", g, "\\b"), final_data$genres))
}

genre_terms <- paste0("is_", tolower(major_genres))

full_model <- lm(
  log_gross ~ log_budget + movie_averageRating + runtime_minutes +
    production_year +
    is_drama + is_comedy + is_action + is_romance +
    is_adventure + is_crime + is_thriller +
    is_mystery + is_biography + is_family,
  data = final_data
)

summary(full_model)

reduced_model <- step(full_model, direction = "backward", trace = FALSE)
summary(reduced_model)

# Figure 5
final_model <- reduced_model
par(mfrow = c(2, 2), mar = c(5, 5, 2.5, 2.5))
plot(final_model,
     col = "steelblue",
     pch = 16,
     cex = 0.6)
par(mfrow = c(1, 1))

new_film <- data.frame(
  log_budget = log(10000000),
  movie_averageRating = 7.0,
  runtime_minutes = 120,
  is_drama = 0,
  is_romance = 0,
  is_thriller = 0,
  is_family = 0
)
```

```
pred <- predict(final_model, newdata = new_film, interval = "prediction")
pred
exp(pred)
```

Citation

Beautiful plotting in R: A ggplot2 cheatsheet. (2014, August 4). Technical Tidbits from Spatial Analysis & Data Science. https://www.zevross.com/blog/2014/08/04/beautiful-plotting-in-r-a-ggplot2-cheatsheet-3/

Creating Graphs with ggplot2. (2025). Grinnell.edu. https://www.stat2labs.sites.grinnell.edu/Handouts/rtutorials/IntroToGgplot0725.html

Posit Software, PBC. (2023, July 26). Date import problem in R. Posit Community. https://forum.posit.co/t/date-import-problem-in-r/170481/17