Explaining Box Office with Minimal Information

Can a Movie's Success Be Explained by Basic Metadata?

Isabelle Huang

December 15, 2025

Movie box office performance is an important topic in both industry and academic research, informing investment decisions, marketing strategy, and our understanding of audience behaviour. This study examines how well simple movie attributes can explain box office outcomes without relying on complex models. Using a dataset of 6,897 films released between 2006 and 2015, we fit linear regression and mixture-of-regressions models to explain box office performance using only genre, distributor, MPAA rating, release time, and a small set of title-based indicators. Our best model achieves an R^2 of 0.68, suggesting that basic metadata captures much of the variation in the data but is not sufficient for accurate explanation. Richer information such as budget, marketing, franchise status, and word-of-mouth effects is likely essential for more accurate analysis.

Table of contents

1 Introduction

Movie theatres are a common destination for families and friends to gather and briefly escape from daily life at a relatively low cost. Yet thousands of films are released worldwide each year, and only a small fraction are widely watched or remembered. What determines which movies audiences choose? Prior work has studied determinants such as budget, sequels, star power, and reviews, and has used these variables to model or forecast revenue [@{Hao2023FactorsFilmRevenue, Scott2019ReturnRegressions}]. Many studies also apply machine learning methods to improve predictive performance (Jange Zarate and Aragon Encarnacion 2025). This raises a natural question: how much of a movie's box office can be explained using only easily available information and simple models? If strong performance is possible with minimal predictors, decisions could be made at lower cost, both in data collection and in analysis. In this project, we investigate the explanatory power of basic movie metadata under simple regression models.

We first fit a linear regression model to explain ticket sales using only genre, distributor, MPAA rating, release year and month, and a small set of indicator variables capturing whether the title contains certain frequent words. We then fit a mixture-of-regressions model to test whether a more flexible specification can substantially improve fit using the same limited feature set. Our main estimand is the log expected number of tickets sold for a movie with given basic attributes. Let Y_i denote the log number of tickets sold for film i, and let X_i denote its features. The target regression function is

$$m(x) = \mathbb{E}[\, Y_i \mid X_i = x \,]$$

Overall, these simple variables explain box office outcomes to a meaningful extent, though not necessarily sufficient. After tuning, our best model achieves an R^2 of approximately 0.68, leaving roughly one-third of the variation in ticket sales unexplained. Among the predictors considered, genre has the strongest association with ticket sales, while title-based indicators contribute the least. Although the resulting models are not accurate enough for practical forecasting, they may serve as useful baselines and as intermediate proxies when richer covariates are unavailable. Future work could either extract more signal from limited data or incorporate additional predictors to capture more complex drivers of performance.

The remainder of this paper is organized as follows. Section 2 describes the dataset and cleaning procedures. Section 3 presents the modeling approach and empirical results. Section 4 discusses findings and limitations and outlines directions for future work.

2 Data

2.1 Overview

We use the statistical programming language Python (Python Software Foundation, n.d.) together with the data library Pandas (The pandas development team 2025) to clean and analyze our dataset. The data come from the Data and Story Library (Data and Story Library, n.d.), an open-source repository of real-world datasets for teaching and practice. Our dataset contains 6,897 major films released between 2006 and 2015, with variables including title, genre, distributor, MPAA rating, release date, and gross revenue.

2.2 Outcome variables

There are two variables in the dataset that measure box office performance: gross revenue and the number of tickets sold. Gross revenue is an aggregate outcome that depends on realized attendance, ticket prices, and the length of the theatrical run, reported in USD. The number of tickets sold is an approximate count based on information provided by theatres and production companies. Because gross revenue can be derived from ticket sales, we focus our analysis on the number of tickets sold. Since this response is numeric, linear regression is a natural starting point.

However, ticket sales are highly right-skewed: a small number of blockbusters sell millions of tickets, while many films sell only a few thousand or fewer. The mean number of tickets sold is about 1.77 million, but the median is only 22,962, indicating that the mean is heavily influenced by a small set of extreme values. Consistent with this, the standard deviation is large at 5.45 million, reflecting substantial dispersion in ticket sales across films.

We visualize films with the highest and lowest ticket sales in the following graphs.

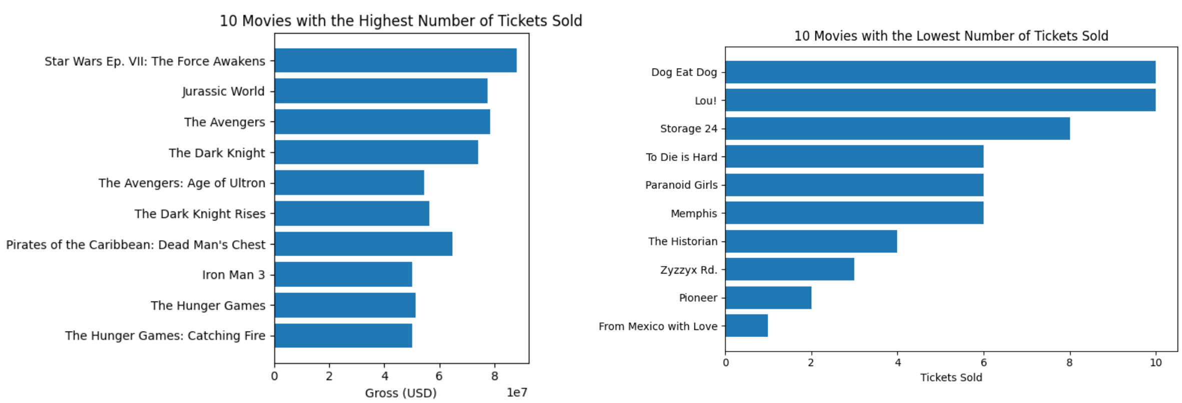

Figure 1: Ticket sales for the ten best- and worst-selling movies.

Because of the strong clustering near zero, we work with a base-10 log transformation of tickets sold instead. This compresses the scale and reduces the influence of extreme values, producing a distribution that is more symmetric and closer to normal. The mean of log tickets sold is about 7.38, the median is about 10.0, and the standard deviation is about 2.87. The transformed distribution is still slightly bimodal, with one peak around 4 and another around 7.

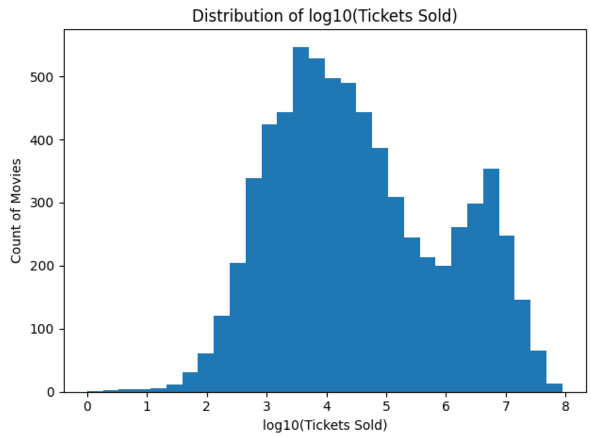

Figure 2: Number of tickets sold on the log base 10 scale

2.3 Predictor variables

We treat all remaining variables in the dataset (excluding gross revenue) as potential predictors of ticket sales. Our primary predictors of interest are Genre and Distributor. Genre is a categorical variable describing the film's primary genre such as Action and Comedy. In our sample, Comedy and Adventure are among the most popular genres in recent years, while Drama is the most frequently produced genre (Figure 3). Distributor is also categorical and identifies the company responsible for releasing the film to theatres. Between 2006 and 2015, IFC and Warner Bros appear most often in the dataset, with roughly 300 films each.

The remaining variables are expected to have weaker effects on box office performance, but we include them for a more complete picture. MPAA rating captures the film's content rating (e.g., G, PG, PG-13, R). Release year and release month record when the film was released and may capture timing effects such as holiday-season boosts.

In addition, we incorporate information from movie titles by extracting simple text-based features. We test whether including certain keywords in a title is associated with higher ticket sales, under the idea that recognizable or appealing words may attract audience attention. We identify the most frequent words appearing in titles—"man", "love", and "life"—and create three indicator variables for whether each word appears in a given title. These variables are then merged into the main dataset.

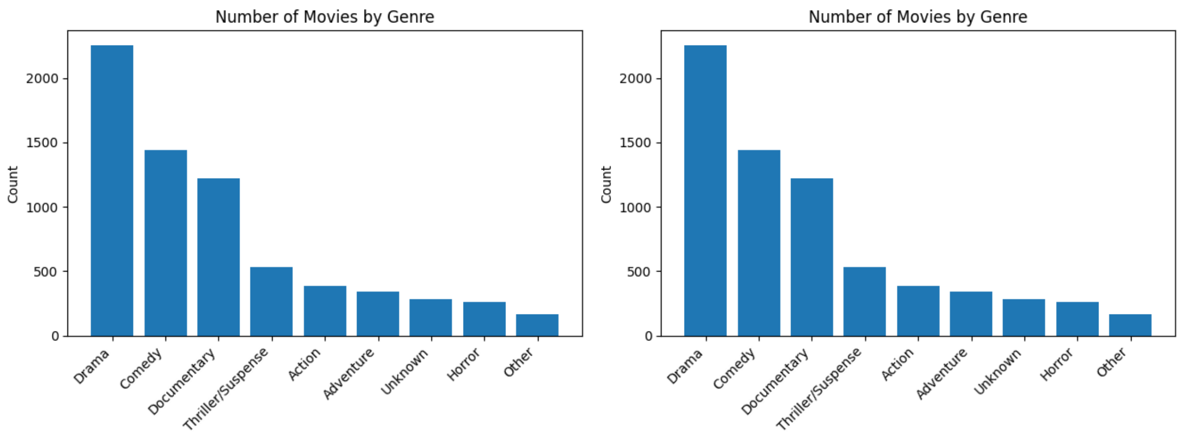

Figure 3: The best selling genres and most produced genres of movies.

Rows with missing values were removed from the dataset prior to analysis, leaving 6,337 observations. The dataset is relatively well curated, so no further cleaning is required. That said, because this is an observational dataset of commercially released films, some measurement error may be present, for example, in approximate ticket counts or genre assignments. We treat the recorded values as given throughout.

3 Model

Based on the exploratory analysis, it appears reasonable to model the base-10 log of tickets sold as a linear function of the predictors. In addition, motivated by the bimodal shape of the transformed distribution, we also fit a mixture-of-regressions model to test whether allowing two latent groups can better capture the data and improve model fit.

3.1 Model set-up

3.2 Linear regression model

To estimate the regression function $m(x)$ we first fit Linear Regression model for the logged tickets sold using Ordinary Least Square Estimators. For each movie $i \in \{1, \dots, n\}$, let

- Y_i be logged number of tickets sold,
- $x_{i,\text{year}}$ and $x_{i,\text{month}}$ be release year and month of the movie,
- G_i, D_i, and M_i denote the categorical variables Genre, Distributor, and MPAA rating,
- $z_{i,\text{man}}, z_{i,\text{love}}, z_{i,\text{life}} \in \{0, 1\}$ be the three title indicators.

Then the full linear model can then be written as

$$
\begin{aligned}
Y_i &= \beta_0 + \beta_1 x_{i,\text{year}} + \beta_1 x_{i,\text{month}} \\
&\quad + \sum_g \gamma_g \, \mathbb{1}\{G_i = g\} + \sum_d \delta_d \, \mathbb{1}\{D_i = d\} + \sum_m \eta_m \, \mathbb{1}\{M_i = m\} \\
&\quad + \alpha_{\text{man}} z_{i,\text{man}} + \alpha_{\text{love}} z_{i,\text{love}} + \alpha_{\text{life}} z_{i,\text{life}} + \varepsilon_i
\end{aligned}
$$

We implement this model in Python using Seabold and Perktold (2010) Linear Rregression makes several assumptions, including linearity, independence, homoscedasticity, and normality of errors and these assumptions are roughly satisfied in our case.

3.3 Mixture Regression Models

We also consider a finite mixture and linear regression for Y_i with the same predictors that we previously considered. Let $Z_i \in \{1, 2\}$ be an unobserved component indicator such that

$$
\Pr(Z_i = k) = \pi_k, \qquad k = 1, 2,
$$

with $\pi_k > 0$ and $\pi_1 + \pi_2 = 1$. Conditional on $Z_i = k$ we assume a standard linear regression model

$$
Y_i \mid (Z_i = k, \mathbf{X}_i) \sim \mathcal{N}(\mathbf{X}_i^\top \beta_k, \, \sigma_k^2), \qquad k = 1, 2,
$$

where β_k is the vector of regression coefficients and σ_k^2 is the error variance for component k. So this allows us to model the relationship specific to the two groups of movies that we suspect exist in the data.

Marginally, the conditional density of Y_i given $\mathbf{X}_i$ is

$$
f(y_i \mid \mathbf{X}_i, \theta) = \sum_{k=1}^{2} \pi_k \, \phi(y_i \mid \mathbf{X}_i^\top \beta_k, \sigma_k^2),
$$

where $\theta = (\pi_1, \pi_2, \beta_1, \beta_2, \sigma_1^2, \sigma_2^2)$ and $\phi(\cdot \mid \mu, \sigma^2)$ denotes the univariate Normal distribution. The observed-data likelihood for θ is therefore

$$
L(\theta \mid \{y_i, \mathbf{X}_i\}_{i=1}^{n}) = \prod_{i=1}^{n} \left\{ \sum_{k=1}^{2} \pi_k \, \phi(y_i \mid \mathbf{X}_i^\top \beta_k, \sigma_k^2) \right\},
$$

and can be maximised using an expectation–maximization (EM) algorithm which is built in to the gmm model in statsmodels.

3.4 Model justification

As we have discussed earlier, our response variable is numerical in nature, making linear regression a natural choice. Linear regression comes with four assumptions, linearity, uncorrelated errors, homoscedasticity, and normality assumptions. In our exploratory data analysis, we saw that after the log transformation, our response variable is roughly normal, satisfying the normality assumption. And since we are not in particular interested in the effect of each predictor, violation of normality assumption would not be too problematic in our case. There were no clear patterns in the residuals plots and the span of residuals remains similar throughout, suggesting that the linearity and homoscedasticity assumption are also roughly satisfied. However, there might be correlation issues among the predictors as we used many indicator predictors for categorical variables with many levels, such as Genre and Distributor. This might cause some instability in our coefficient estimates, but since we are more interested in overall model performance rather than individual coefficients, this is not a major concern.

On the other hand, the bi-modal distribution of the response variable after log transformation suggests that the population may consist of two groups, leading to the use of mixture models. Therefore, it was also reasonable to fit a separate linear regression model after separating each group with GMM.

4 Results

The linear regression on the log scale achieves an in-sample R^2 of about 0.68 and a root mean squared error (RMSE) of approximately 0.82 in log10 units. An RMSE of 0.82 on the $\log_{10}$ scale corresponds to prediction errors on the order of a factor of about 6 in the original ticket counts, so while the model explains substantially more variation than the unlogged version (which had $R^2 \approx 0.35$), there is still considerable unexplained heterogeneity in box office performance.

The regression coefficients confirm that distributor and genre are among the strongest predictors of log ticket sales. Large studio distributors such as Walt Disney, Universal, 20th Century Fox, Sony Pictures and Warner Bros. are associated with substantially higher log ticket sales than the baseline distributors, while many smaller or niche distributors are associated with lower expected ticket sales. This matches with the expectation that audience would tend to select movies distirbuted by famous componnies. Among genres, adventure and action movies have positive effects on log ticket sales relative to the reference genre, whereas documentary and some drama categories are associated with lower ticket sales. MPAA ratings also show a clear pattern: PG-13 and PG movies tend to have higher log ticket sales than the baseline rating, while Not Rated and G titles are associated with lower sales. Though since most of the movies in the dataset were Not Rated, this could be a result of unequal representation. The simple title indicators (whether the title includes “man”, “love”, or “life”) have small coefficients and do not materially change model fit. Overall, after log transformation, the model captures broad patterns in how basic attributes relate to demand, but the remaining residual variation indicates that important drivers such as budget, marketing, franchise status and word-of-mouth are missing from this specification.

Predictor	Estimate	Std Error	p-value
Genre = Adventure	0.097	0.072	0.178
Distributor = 20th Century Fox	2.835	0.862	5.9e-09
MPAA = T.PG-13	0.395	0.092	1.7e-05
Genre = Documentary	-0.233	0.061	0.000
T.Universal	3.062	0.862	0.000

In terms of the mixture regression, we fit a two-component Gaussian mixture model to the log-transformed ticket sales to split the data into two groups, as suggested by the bimodality of response. yielding a "low–box-office" group (component 1, $n = 4{,}484$) and a "high–box-office" group (component 0, $n = 2{,}196$). We then fit separate linear regression models for each group, using the same set of predictors as the model above.

However, when we then fit two separate regression models on the two components (the high and low box office groups) identified by the mixture model, the overall R^2 for each component dropped to 0.38 and 0.52 respectively. This might be a result of the drop in data size after splitting into two components, leading to less stable estimates. Overall, the mixture model did not substantially improve our ability to explain variation in ticket sales using the limited feature set. Nonetheless, we discovered some similar patterns like action and adventure–type genres and major studio distributors are associated with higher log ticket sales, while more niche genres and non-major distributors tend to be associated with lower sales, conditional on the other covariates.

High Box Office Group (Component 0):

Predictor	Estimate	Std Error	p value
Distributor = Sony Pictures Classics	-0.859	0.062	0.000
Distributor = Roadside Attractions	-0.8956	0.094	0.000
Distributor = Paramount Pictures	0.0453	0.053	0.393
MPAA = PG	0.154	0.065	0.018
Genre = Drama	-0.3106	0.040	0.000

Low Box Office Group Coefficients(Component 1):

Predictor	Estimate	Std Error	p value
Distributor = 20th Century Fox	0.6007	0.707	0.396
Genre = Adventure	0.1209	0.101	0.232

Predictor	Estimate	Std Error	p value
MPAA = Not Rated	-0.3786	0.120	- 0.002
Distributor = Walt Disney	0.5205	0.705	0.460
Genre = Drama	0.1609	0.066	0.015

5 Discussion

5.1 Summary of analysis

This report answers the question of how much of a movie's box office performance can be explained using simplistic model and easily obtained data. We explored the dataset containing 6,897 major releases from 2006–2015, and based on the exploratory data analysis, we decided that a linear regression model would be be an appropriate simple model for modelling number of tickets sold. We modeled log-transformed ticket sales as a function of release year and month, distributor, genre, MPAA rating, and a small set of title indicators. After the first simple model, based on the observation that our response variable was bi-modal, we fit a two-component Gaussian mixture model to the log-tickets to identify the two potential underlying groups. Regression model was then fit to each compoenent to see whether relationships between predictors and outcomes differ in the low- and high-grossing segments and whether explanatory power of the model can be improved by grouping.

5.2 key Findings

We discovered that we are able to explain a moderate amount of variation in box office performance using only basic metadata. The linear regression on the log scale attains an R^2 of 0.68, which was non-trivial. It can be used as a proxy for understanding broad patterns in what movie attributes contribute to the its box office performance, when computation power is limited or precise outcomes were not required. Indeed, we were able to learn from our simple models that films released by large studio distributors and in commericial genres like advanture movies systematically sell more tickets than those from smaller distributors and narrower genres. People tend to favour movies that are relatively shallow, compared to genres like documentary and drama, which are associated with lower sales. This confirms the intuitive idea that institutional backing and genre positioning matter for commercial success, and the effect is visible even when we restrict attention to a small set of coarse variables. Although we initially believe that audience could be attracted to movies with certain key words in the title, our results show that these title indicators have minimal effects on ticket sales. So we are able to make some useful inferences about movie box office performance using only minimal information and simple models.

Moreover, the mixture analysis shows that the movie market can be effectively partitioned into at least two groups: a large group of low- to moderate-grossing films and a smaller group of high-grossing films. The two-component Gaussian mixture on fits substantially better than a single Normal distribution, with one component centered around a few thousand tickets and another around millions of tickets. When we fit separate regressions within these groups, we see that basic metadata are more informative among films that have high box office, compared to those with lower ticket sales. The high group regression has higher R^2 and cleaner, more stable effects for distributors and genres, while the low group shows weaker fits and more residual variability. This suggests that once a film reaches the commercial "upper tier", its performance is more systematically tied to structural choices like genre and distribution, whereas performance in the lower tier is more idiosyncratic and less predictable from simple attributes. This also suggests that future work could build on mixture frameworks to better capture the characteristics of different movies, thereby better explaining and predicting box office outcomes.

5.3 Limitations and future work

However, it is important to recognize that there are several important limitations in our analysis. Substantively, the dataset is observational and restricted to major movies released from 2006 to 2015. This is ten years from now and trends in audience preference could change, the market could become more volatile and harder to explain with simple features and models. Being restricted to major movies would impact the generalization to smaller independent releases. Measurement could be imperfect in the dataset as assignments of genre and MPAA could be coarse and merging heterogeneous content. Methodologically, our models had issues in mutlicollinearity due to the large number of dummy variables for categorical predictors with many levels, which could lead to unstable coefficient estimates. Linearity assumption does not hold perfectly despite the transformation of the response variable, potentially leading to biased coefficients. We also did not consider interaction effects between predictors, due to the already existing multi-collinearity which we did not want to exacerbate. Finally, the mixture regressions rely on hard assignment of movies to latent components and do not carry mixture uncertainty into the regression stage, which could lead to unwanted bias.

The analysis here is a first step rather than a complete model of box office performance. If we wish to continue down this path of using minimal information and simple models, we would need to develop methods that can better extract signals from limited data. This is an active researsch area and there are many potential directions. But if we just want to improve box office explanation and prediction more generally, richer data and more sophisticated models would be helpful. Future work can extend the feature set by incorporating production budgets, sequel and franchise indicators, star metrics, and measures of pre- and post-release attention such as critic scores, online ratings and social media activity. With richer covariates, it would be natural to compare regularised linear models, tree-based methods and hierarchical models that allow distributor and genre effects to vary across time or markets. On the mixture side, a more formal finite mixture of regressions—where both component membership and regression parameters are estimated jointly—would provide a principled framework for studying regime-specific relationships, at the cost of more complex computation. Finally, a deeper treatment of selection and measurement would clarify how much of the remaining unexplained variation is due to genuinely unobserved drivers versus noise in the observed data.

Appendix

A Additional data details

Here are a few additional plots to further illutrate the data.

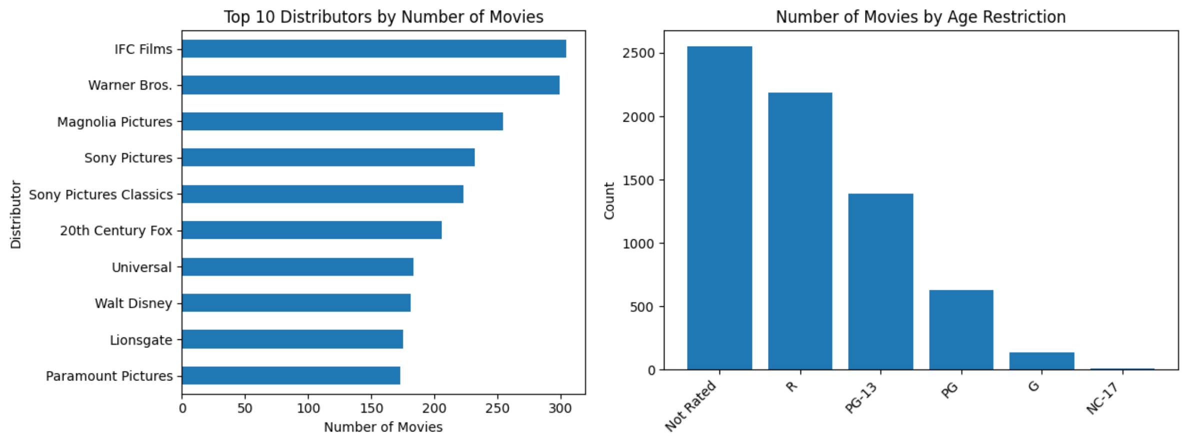

Figure 4: Distributors with the most movies and MPAA age restrictions overview of movies contained in the dataset.

B Model details

On top of the two models considered in the main text, we also consider the following two specifications as part of exploration.

Ordinary linear regression on raw ticket sales

Let $i = 1, \dots, n$ index movies and define the outcome as the number of tickets sold,

$$Y_i = \text{Tickets.Sold}_i.$$

For each movie i, we observe:

- Year_i: release year,
- Month_i: release month $(1, \dots, 12)$,
- G_i: genre (categorical),
- D_i: distributor (categorical),

- M_i: MPAA rating (categorical),
- $z_{i,\mathrm{man}}, z_{i,\mathrm{love}}, z_{i,\mathrm{life}} \in \{0,1\}$: indicator variables for whether the title contains the words "man", "love", or "life".

We treat G_i, D_i, and M_i as categorical variables and represent them with dummy variables. The linear regression model on the original ticket scale is

$$\begin{aligned} Y_i = \beta_0 &+ \beta_{\mathrm{year}}\,\mathrm{Year}_i + \beta_{\mathrm{month}}\,\mathrm{Month}_i \\ &+ \sum_g \gamma_g\,\mathbb{1}\{G_i = g\} + \sum_d \delta_d\,\mathbb{1}\{D_i = d\} + \sum_m \eta_m\,\mathbb{1}\{M_i = m\} \\ &+ \alpha_{\mathrm{man}} z_{i,\mathrm{man}} + \alpha_{\mathrm{love}} z_{i,\mathrm{love}} + \alpha_{\mathrm{life}} z_{i,\mathrm{life}} + \varepsilon_i, \end{aligned}$$

where one level of each factor (genre, distributor, MPAA) is taken as the reference category, and the error terms satisfy

$$\varepsilon_i \sim \mathcal{N}(0, \sigma^2), \qquad i = 1, \dots, n.$$

The parameter vector $(\beta_0, \beta_{\mathrm{year}}, \beta_{\mathrm{month}}, \{\gamma_g\}_g, \{\delta_d\}_d, \{\eta_m\}_m, \alpha_{\mathrm{man}}, \alpha_{\mathrm{love}}, \alpha_{\mathrm{life}})$ is estimated by ordinary least squares, i.e. by minimizing the sum of squared residuals

$$\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2.$$

Lasso regression on log-transformed ticket sales

For the penalised model, we work with a log-transformed outcome

$$Y_i = \log_{10}(\mathrm{Tickets.Sold}_i)$$

using the same underlying predictors (release year, release month, distributor, genre, MPAA rating, and title indicators).

Let $\mathbf{x}_i \in \mathbb{R}^p$ denote the design vector for movie i after preprocessing, where:

- continuous predictors (such as year and month, and the binary title indicators) have been centered and scaled, and
- categorical predictors (distributor, genre, MPAA) have been expanded into dummy variables via one-hot encoding, with one reference level dropped for each factor.

The Lasso model assumes the linear relationship

$$Y_i = \mathbf{x}_i^\top \beta + \varepsilon_i, \qquad i = 1, \dots, n,$$

with mean-zero errors ε_i, but estimates $\beta \in \mathbb{R}^p$ by solving the penalised least squares problem

$$\widehat{\beta}_\lambda = \arg\min_\beta \left\{ \frac{1}{2n} \sum_{i=1}^{n} (Y_i - \mathbf{x}_i^\top \beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\},$$

where $\lambda \geq 0$ is a tuning parameter controlling the strength of the ℓ_1 penalty. The penalty is applied to the slope coefficients β_j (the intercept is left unpenalised or absorbed by centering). In practice, we select λ by K-fold cross-validation and use $\widehat{\beta}_{\widehat{\lambda}}$ as the final model. The ℓ_1 penalty both shrinks coefficients toward zero and sets many exactly to zero, performing variable selection and mitigating the impact of multicollinearity among the large set of dummy variables.

B.1 Diagnostics

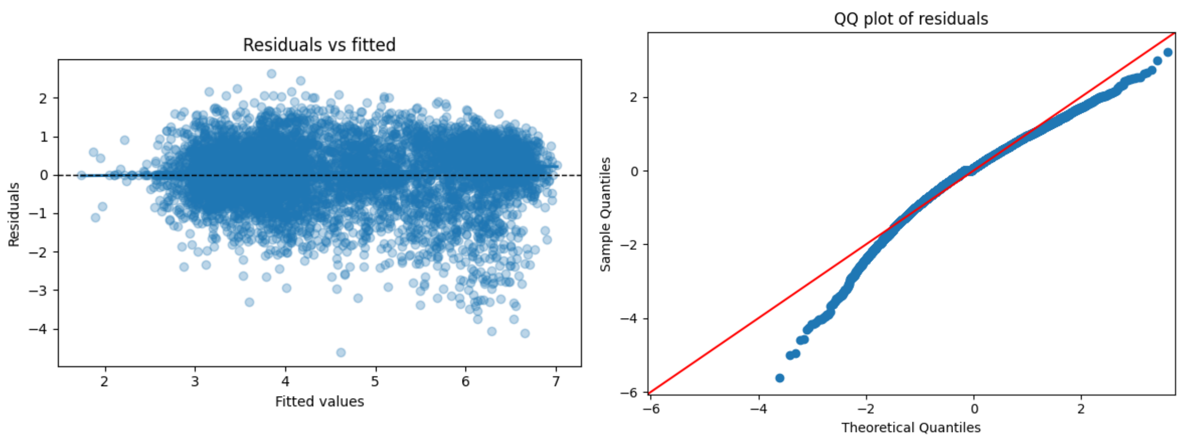

Figure 5: Diagnostic plots for linear regression model residuals. As mentioned in main text, there seems to be

C References

We use Python and several open-source packages for our analysis (**python?**; **harris2020numpy?**; **mckinney2010pandas?**; **pedregosa2011scikit?**; **seabold2010statsmodels?**)

Data and Story Library. n.d. "Movies Box Office Dataset." https://dasl.datadescription.com/.

Jange Zarate, Karla Jemima, and Michael Moises Aragon Encarnacion. 2025. "A Systematic Review on Forecasting for Box Office Success of a Movie Using Machine Learning Prediction Methodologies." In *Research Perspectives on Software Engineering and Systems Design: Proceedings of 8th Computational Methods in Systems and Software 2024, Volume 3*, edited by Radek Silhavy and Petr Silhavy, 1491:378–92. Lecture Notes in Networks and Systems. Cham: Springer Cham. https://doi.org/10.1007/978-3-031-96380-3_33.

Python Software Foundation. n.d. "Python Language Reference." https://www.python.org/.

Seabold, Skipper, and Josef Perktold. 2010. "Statsmodels: Econometric and Statistical Modeling with Python." In *Proceedings of the 9th Python in Science Conference (SciPy 2010)*.

The pandas development team. 2025. "Pandas-Dev/Pandas: Pandas (V2.3.3)." Zenodo. https://doi.org/10.5281/zenodo.17229934.