Predicting Box Office Success

A Regression Analysis of TMDB Movie Data

Aitugan Shagyr • Makhabbat Batyrova
Aisultan Zhakupbayev • Raiymbek Arysbek

SDS 301: Modern Regression Analysis

December 2025

1 Introduction

The global film industry is a multi-billion dollar market characterized by high risk and extreme variance in financial outcomes. While a blockbuster can generate returns far exceeding production costs, a substantial proportion of films fail to recoup their budgets. In this environment, the ability to forecast box office performance prior to release is valuable for studios, investors, and distributors aiming to optimize budgeting and marketing strategies.

This project addresses the problem of predicting worldwide box office revenue based on pre-release movie characteristics. Using the *TMDB Box Office Prediction* dataset, we build and validate a multiple linear regression model that identifies the key drivers of financial success. Specifically, we investigate whether quantifiable factors such as production budget, cast/crew size, and genre are associated with a film's commercial performance, and whether temporal factors (release year) capture systematic trends over time.

1.1 Data Description and Processing

The dataset for this analysis was sourced from The Movie Database (TMDB) via Kaggle. The cleaned training data consists of $n = 3{,}000$ unique films with release years spanning 1921 to 2017.

The response variable is `revenue`, a continuous variable representing total worldwide box office earnings in US dollars. To predict this target, we selected explanatory variables that capture different dimensions of a film's production and appeal:

- **Budget (X_1):** `budget` in USD.
- **Popularity (X_2):** TMDB popularity metric.
- **Runtime (X_3):** Film duration (minutes).
- **Cast Count (X_4):** Number of credited cast members.
- **Crew Count (X_5):** Number of credited crew members.
- **Release Year (X_6):** Year of theatrical release.
- **Primary Genre (X_7):** Main genre label.

Data Engineering

The raw dataset contained nested JSON strings for `cast`, `crew`, and `genres`. We parsed these fields to obtain cast/crew counts and primary genre labels, imputed missing `runtime` with the sample mean, and standardized date formats (including correction of two-digit year encodings). The resulting dataset is analysis-ready for regression modeling and diagnostics.

2 Exploratory Data Analysis

Analysis of Figure 1: Most variables exhibit essentially 0% missing values. The variable `log_budget_reported` has approximately 27% missingness, which corresponds to films with zero or unknown budgets that were intentionally set to NA after constructing the reported budget field. The variable `genre_count` has a negligible amount of missingness, while all other variables are complete.

Action / Implication: Dropping observations with missing budget would remove roughly 27% of the data. Instead, the full sample is retained by introducing (i) a `budget_missing`

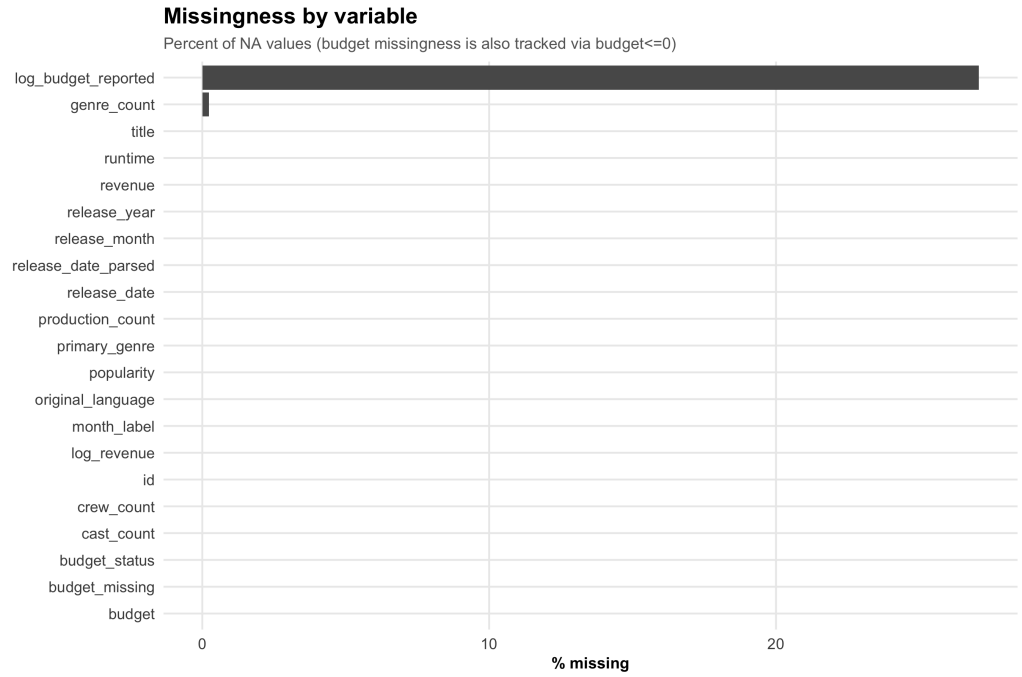


Figure 1: Missingness by variable

indicator and (ii) using `log_budget_reported` only when the budget is available (or via an imputation strategy), rather than treating zero as a real budget value.

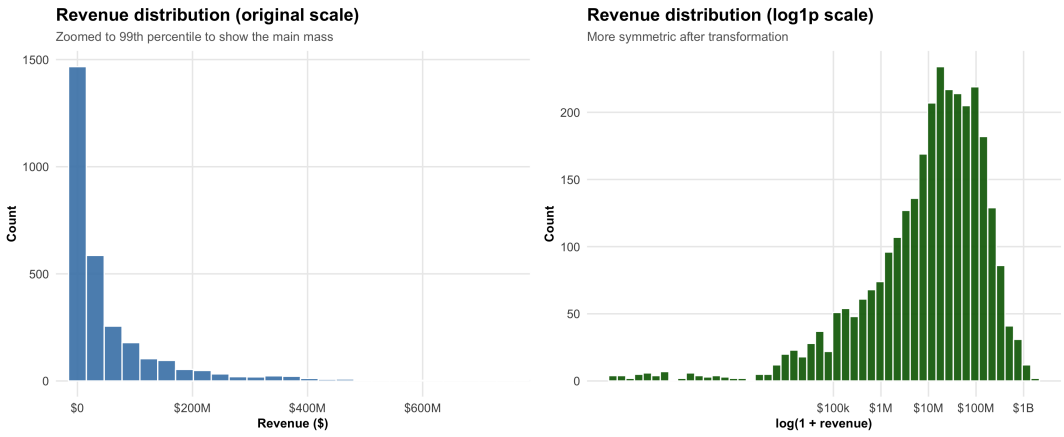


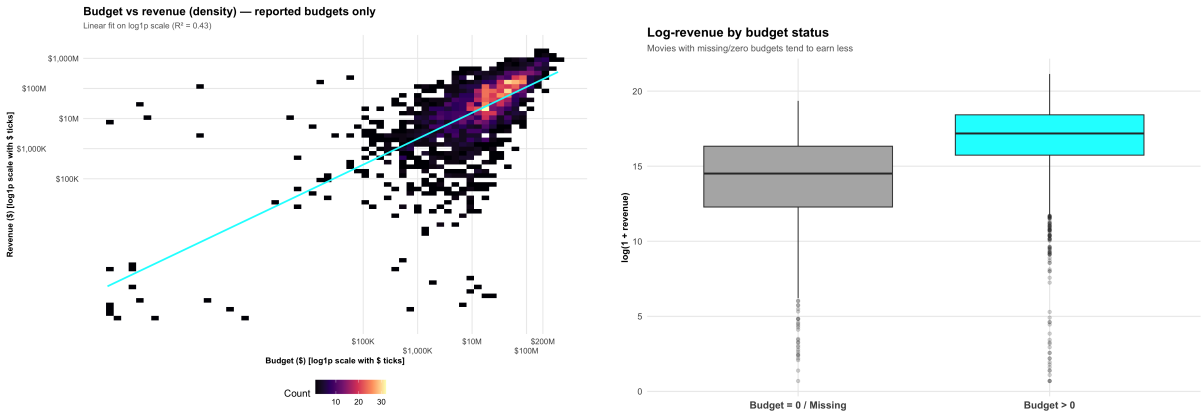
Figure 2: Revenue distribution on the original scale (top/left) and on the log1p scale (bottom/right)

Analysis of Figure 2 (Original Scale): Revenue is extremely right-skewed: most films generate relatively low revenue, while a small number of blockbuster films produce a long right tail, even when restricting the view to the 99th percentile.

Action / Implication: Modeling revenue on the original scale is likely to violate normality and homoscedasticity assumptions required for linear regression.

Analysis of Figure 2 (Log1p Scale): After applying the $\log(1 + \text{revenue})$ transformation, the distribution becomes much closer to symmetric, and the influence of extreme blockbusters is substantially reduced.

Action / Implication: Log-revenue is a more appropriate response variable for linear regression models.



(a) Budget vs revenue (density), reported budgets only

(b) Log-revenue by budget status

Figure 3: Budget and revenue relationships

Analysis of Figure 3a: In log–log space, there is a clear positive relationship between budget and revenue. The highest density of observations follows the fitted regression line, indicating that the relationship is not driven solely by outliers. The simple $\log(\text{revenue}) \sim \log(\text{budget})$ model yields $R^2 \approx 0.43$ when restricted to films with reported budgets.
Action / Implication: A log–log specification for budget is well justified. Budget should be treated as a core predictor, while missing budgets should be handled separately rather than implicitly coded as zero.

Analysis of Figure 3b: Films with missing or zero budgets exhibit substantially lower median log-revenue than films with reported budgets, and the entire distribution is shifted downward.
Action / Implication: Include a `budget_missing` indicator as a predictor and avoid interpreting zero budgets literally. This pattern suggests that budget missingness is not completely at random.

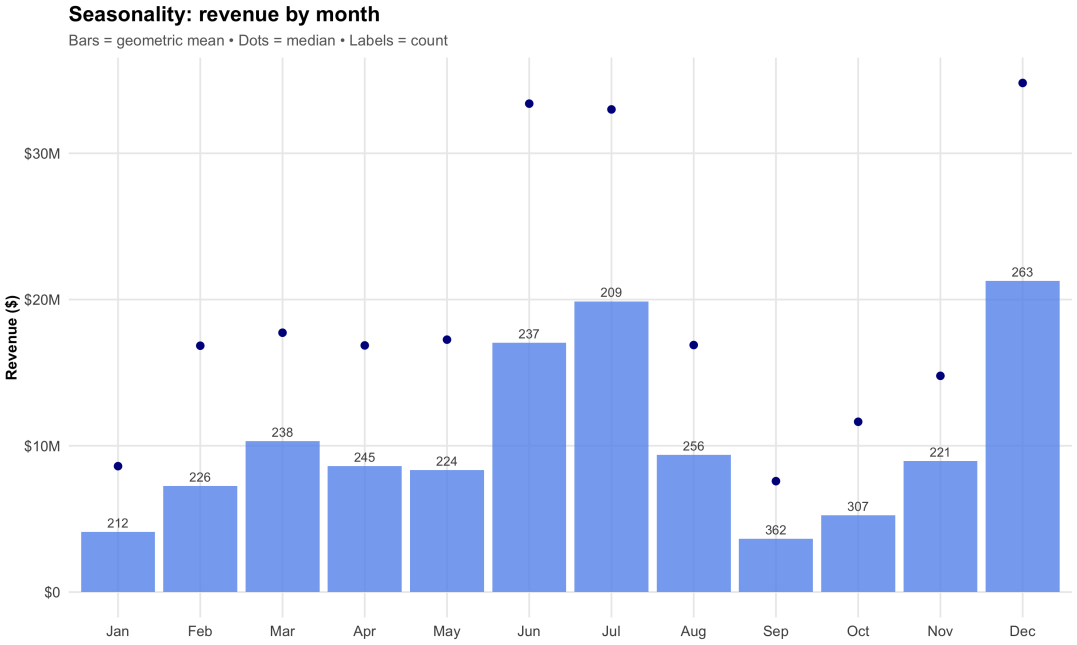
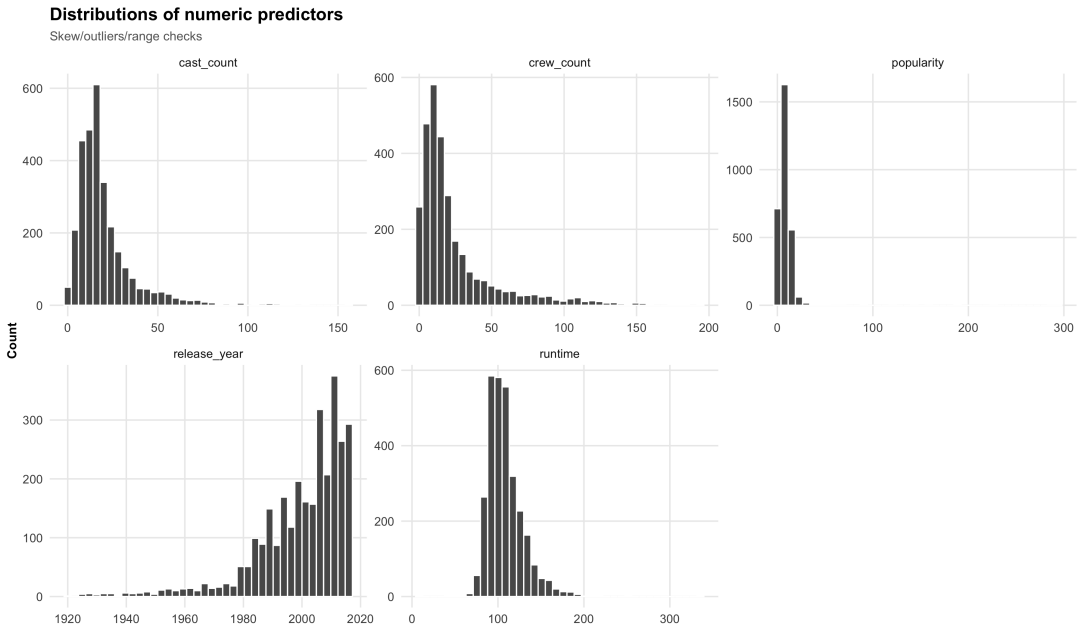


Figure 4: Seasonality: revenue by month

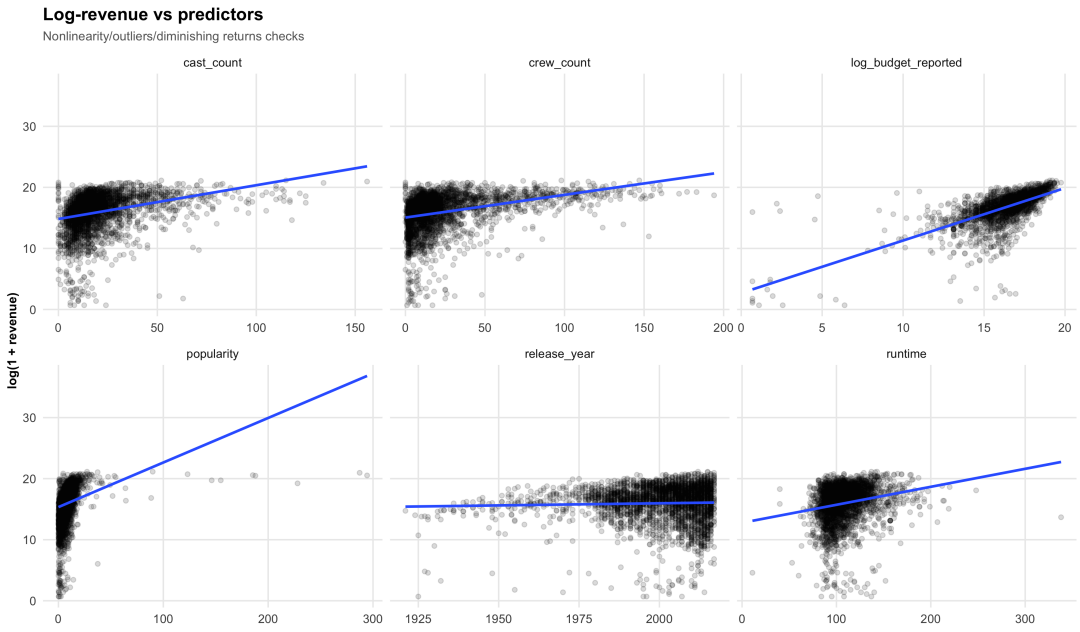
Analysis of Figure 4: Typical revenue varies by release month. June–July and December tend to have higher geometric means and medians, while September (and often January) exhibits

lower values. The number of releases also differs substantially across months.

Action / Implication: Include `release_month` (or a coarser seasonal factor) as a categorical predictor to capture systematic timing effects.



(a) Distributions of numeric predictors



(b) Log-revenue vs predictors

Figure 5: Predictor distributions and bivariate relationships

Analysis of Figure 5a: The variables `cast_count` and `crew_count` are strongly right-skewed, with long upper tails. `popularity` is extremely right-skewed, with most values near zero and a small number of extreme outliers. `release_year` is heavily concentrated in recent decades, while `runtime` is approximately unimodal around typical feature lengths, with some long-film outliers.

Action / Implication: Consider log-transformations (e.g. $\log(1 + x)$) for cast count, crew

count, and popularity. Runtime may enter linearly, but potential nonlinearity and outliers should be checked. Year effects may require centering or nonlinear modeling.

Analysis of Figure 5b: Log-revenue shows mild to moderate positive associations with cast count, crew count, popularity, and runtime, often with substantial dispersion and signs of diminishing returns. The strongest and cleanest relationship is observed with `log_budget_reported`. The association with `release_year` is weak, with wide variability.

Action / Implication: Retain these variables as predictors, applying transformations where appropriate, and assess leverage and influence diagnostics. Budget remains the most important explanatory variable.

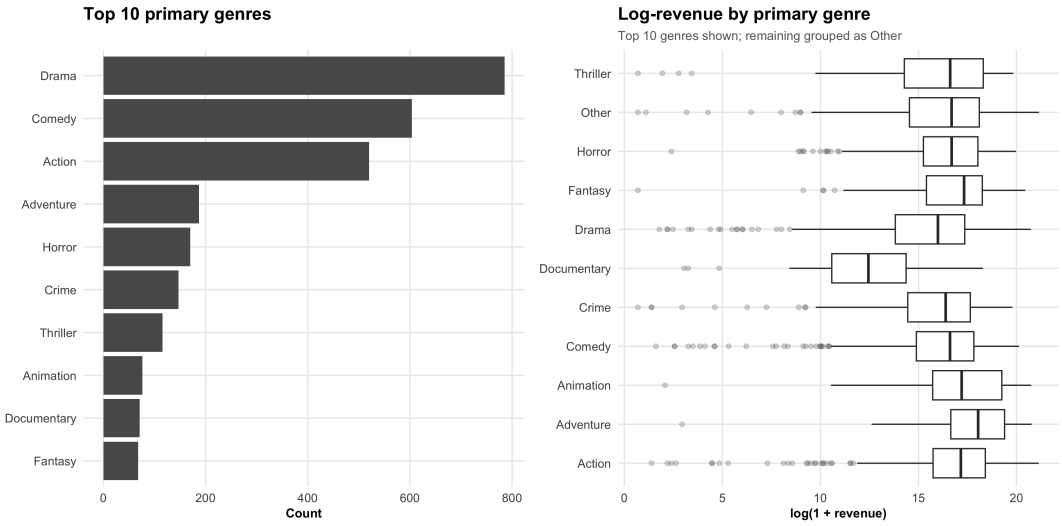


Figure 6: Top genres and log-revenue by primary genre

Analysis of Figure 6: The genre distribution is highly imbalanced: Drama, Comedy, and Action dominate the sample. Median log-revenue differs across genres. Documentary films exhibit notably lower typical revenues, while Action, Adventure, Animation, Fantasy, and Thriller tend to have higher medians.

Action / Implication: Include `primary_genre` as a categorical predictor, grouping rare genres into an "Other" category to ensure stable estimation. Genre effects should be interpreted cautiously due to potential confounding with budget and popularity.

Analysis of Figure 7: The strongest correlation with log-revenue is observed for `log_budget_reported` ($r \approx 0.66$). Other predictors show weaker but positive correlations. Moderate correlations exist among predictors, but none appear extreme.

Action / Implication: Budget should be central in the model. Multicollinearity should be assessed using variance inflation factors (VIF).

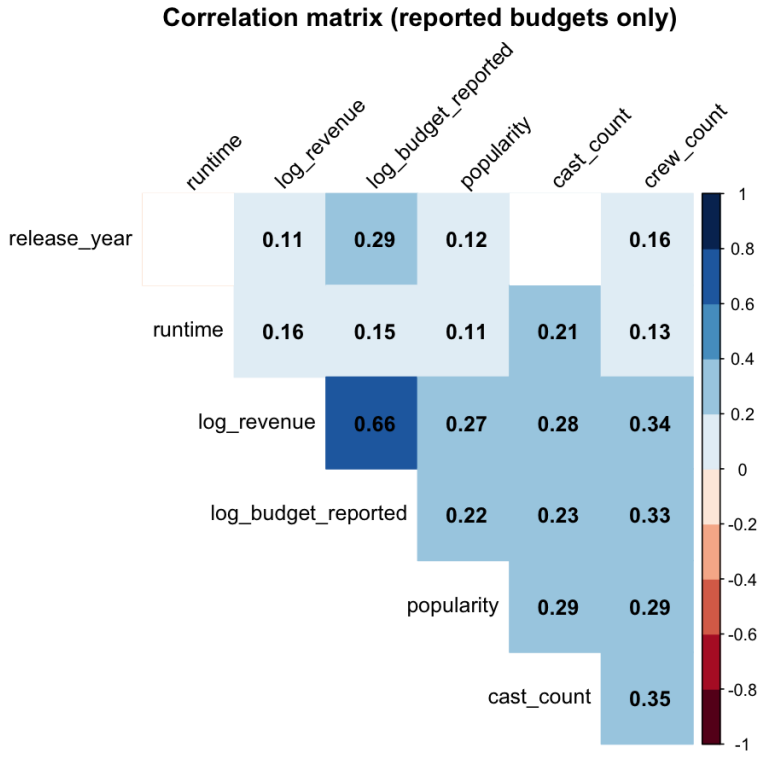


Figure 7: Correlation matrix (reported budgets only)

2.1 Summary

Table 1: Summary statistics for key variables ($n = 3{,}000$)

Variable	Mean	Median	Min	Max
Revenue (USD)	\$66.7M	\$16.8M	0	\$1.52B
Budget (USD)	\$22.5M	\$8.0M	0	\$380M
Budget missing indicator	0.271	0	0	1
Popularity	8.46	7.37	0	294.3
Runtime (minutes)	108.3	104.5	11	338
Cast count	–	16	0	156
Crew count	–	14	0	194
Release year	–	2004	1921	2017
Release month	–	7	1	12

Discussion of Table 1: Revenue is strongly right-skewed (mean \$66.7M vs median \$16.8M, max \$1.52B), supporting the use of $\log(1 + \text{revenue})$ as the response. Budgets are also right-skewed (mean \$22.5M vs median \$8.0M), and 27.1% of films have missing/zero budgets, implying that budget missingness should be handled explicitly. Popularity and cast/crew sizes exhibit heavy right skewness and potential outliers, motivating transformations and influence checks. Runtime is centered around typical feature lengths, while year and month variables justify controlling for time trends and seasonality.

2.2 Conclusion

Overall, the EDA supports modeling revenue on the log scale to address strong right-skewness and heteroscedasticity. Budget is the strongest single predictor in log–log space and should be central in the regression specification, while budget missingness (27.1%) must be treated explicitly via an indicator and by using `log_budget_reported` only for reported budgets. Additional predictors such as popularity, cast/crew size, runtime, release month, and release year show meaningful but weaker relationships with log-revenue and may require transformations or non-linear terms during model diagnostics. Genre effects should be included as a factor with rare levels grouped to ensure stable estimation.

3 Modeling

3.1 Goal and response transformation

Our goal is to predict worldwide box office revenue using pre-release movie characteristics. Because revenue is highly right-skewed and exhibits increasing variance with its mean, we model

$$y_i = \log(1 + \text{revenue}_i) = \log(1 + \text{revenue}_i)\,,$$

which yields a more symmetric response and improves the plausibility of linear-model assumptions.

3.2 Notation

Let $X \in \mathbb{R}^{n\times q}$ denote the design matrix (including an intercept) and let $\beta \in \mathbb{R}^q$ be the coefficient vector. The multiple linear regression model is

$$y = X\beta + \varepsilon, \qquad E[\varepsilon \mid X] = 0.$$

The ordinary least squares (OLS) estimator is

$$\hat{\beta} = (X^\top X)^{-1} X^\top y, \qquad \hat{y} = X\hat{\beta}, \qquad e = y - \hat{y}.$$

3.3 Candidate specifications

Guided by the EDA, we consider the following nested multiple linear regression specifications, all with response $y = \log(1 + \text{revenue})$.

Model 1 (baseline: budget with missingness handling). We treat reported budget as a core predictor and include a missing-budget indicator:

$$y_i = \beta_0 + \beta_1\, \texttt{log\_budget\_use}_i + \beta_2\, \texttt{budget\_missing}_i + \varepsilon_i.$$

Model 2 (adding popularity and production size). We add predictors capturing demand and production scale:

$$\begin{aligned} y_i = \beta_0 &+ \beta_1\, \texttt{log\_budget\_use}_i + \beta_2\, \texttt{budget\_missing}_i + \beta_3\, \texttt{log\_pop}_i + \beta_4\, \texttt{runtime}_i \\ &+ \beta_5\, \texttt{log\_cast}_i + \beta_6\, \texttt{log\_crew}_i + \varepsilon_i. \end{aligned}$$

Model 3 (full: time and categorical controls). We further control for release timing and genre using factor indicators:

$$\begin{aligned} y_i = \beta_0 &+ \beta_1\, \texttt{log\_budget\_use}_i + \beta_2\, \texttt{budget\_missing}_i + \beta_3\, \texttt{log\_pop}_i + \beta_4\, \texttt{runtime}_i \\ &+ \beta_5\, \texttt{log\_cast}_i + \beta_6\, \texttt{log\_crew}_i \\ &+ \sum_{m=2}^{12} \gamma_m\, \mathbf{1}(\texttt{release\_month} = m) \\ &+ \sum_{g} \delta_g\, \mathbf{1}(\texttt{primary\_genre} = g) \\ &+ \eta\, \texttt{release\_year\_c}_i + \varepsilon_i. \end{aligned}$$

Nonlinear refinements. To capture diminishing returns, we also consider quadratic terms for runtime and release year:

$$\begin{aligned} y_i = \text{Model 3 terms} &+ \text{poly}(\texttt{release\_year\_c}_i, 2) \\ &+ \text{poly}(\texttt{runtime}_i, 2) + \varepsilon_i. \end{aligned}$$

4 Diagnostics and Model Selection

4.1 Workflow

Guided by EDA, we built models in increasing complexity and checked assumptions at each stage using: (i) the standard 4-panel diagnostic plots (residuals vs fitted, Q–Q, scale-location, leverage), (ii) formal heteroskedasticity tests (Breusch–Pagan), and (iii) influence diagnostics (Cook's distance). When violations were detected, we applied targeted remedies (transformations, added terms, robust inference, and sensitivity refits) and retained changes only when supported by model-comparison evidence.

4.2 Model 1 (`m1`): budget-only baseline

Model 1 (`m1`) uses only budget information (including the missing-budget indicator) to predict $\log(1+\text{revenue})$. It provides a useful baseline (Adjusted $R^2 = 0.3929$), but diagnostics indicate assumption violations.

Heteroskedasticity. A Breusch–Pagan test strongly rejected constant variance ($BP = 189.39$, $df = 2$, $p < 2.2 \times 10^{-16}$), consistent with the fan shape in Figure 8. **Action:** We use heteroskedasticity-robust standard errors (HC3) for inference in subsequent models.

Budget missingness and collinearity. Large VIF values for `log_budget_use` and `budget_missing` are expected because these two terms are mechanically linked by the missingness-handling design. This inflates the uncertainty of *individual* coefficients, but retaining both terms avoids dropping $\approx 27\%$ of the data and prevents implicitly treating unknown budgets as literal zeros.

Action: Retain both terms and interpret them jointly.

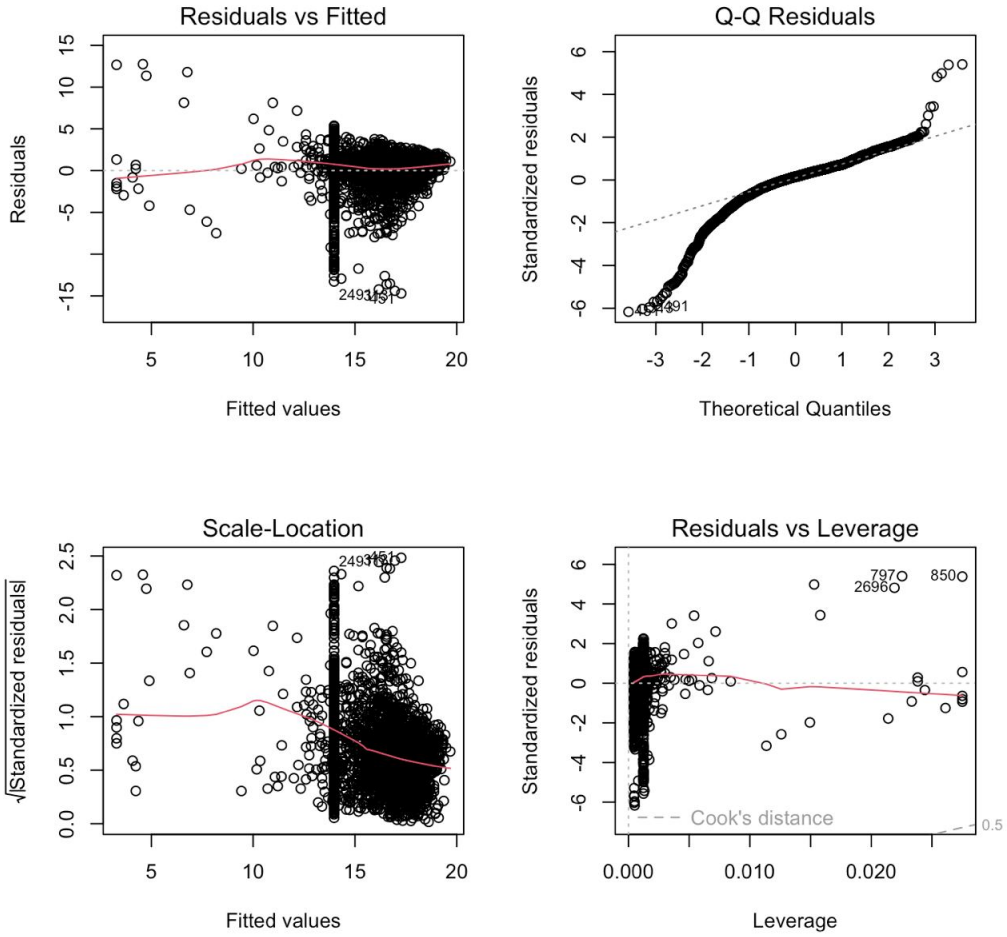


Figure 8: Diagnostic plots for Model 1 (`m1`). The residual spread increases with fitted values and the Q–Q plot shows heavy tails, indicating heteroskedasticity and non-normality.

Table 2: Variance Inflation Factors (VIF) for Model 1

Predictor	VIF
`log_budget_use`	18.47
`budget_missing`	18.47

Note: High VIFs are due to the structural construction of the missing budget indicator.

Motivation to expand the model. To verify that non-budget features add explanatory power, we compared `m1` to a version adding runtime and a quadratic runtime term. A partial F-test indicated a statistically significant improvement:

Table 3: Partial F-test showing improvement when adding runtime (with a quadratic term) to the budget-only baseline.

Model	Res.Df	RSS	F	Pr(>F)
`m1` (baseline)	2997	17050	–	–
`m1` + runtime + runtime^2	2995	16803	22.06	3.08×10^{-10}

Decision: Because `m1` is under-specified and assumption violations are severe, we proceed to

a richer specification.

4.3 Model 2 (m2): adding popularity and production size

Model 2 (m2) adds predictors capturing audience interest and production scope: log_pop, runtime, log_cast, and log_crew. Fit improves over m1, but heteroskedasticity and tail departures remain.

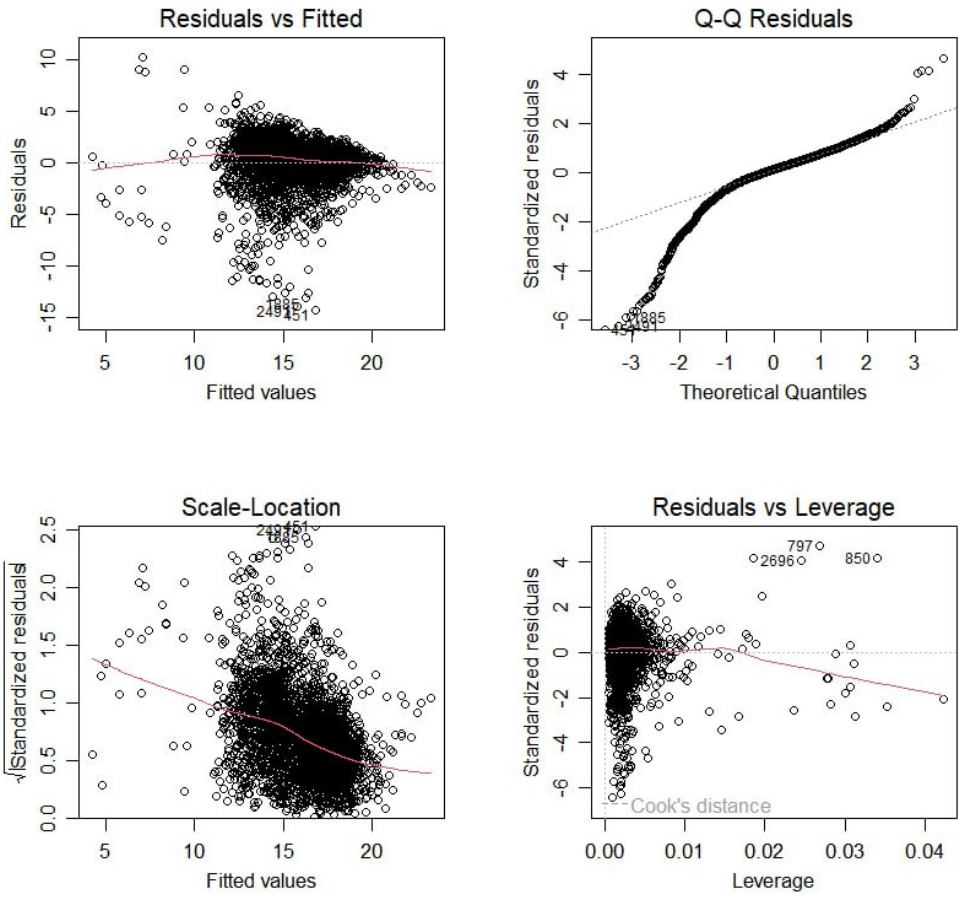


Figure 9: Diagnostic plots for Model 2 (m2). Fit improves, but heteroskedasticity and tail departures remain, motivating robust inference and further refinements.

Heteroskedasticity (formal test). Breusch–Pagan test:

$$BP = 191.14, \quad df = 6, \quad p < 2.2 \times 10^{-16}.$$

Action: Inference for m2 is based on HC3 robust standard errors.

Robust coefficient summary (HC3).

Multicollinearity (VIF). Motivation for further refinement: EDA indicated seasonality by release month and differences across genres, motivating additional categorical controls and a time trend.

Table 4: Model 2 (m2) coefficient tests using HC3 robust standard errors.

Variable	Estimate	Std. Error	t value	p-value
Intercept	1.995	0.718	2.780	0.0055
log_budget_use	0.638	0.050	12.754	$< 2.2 \times 10^{-16}$
budget_missing	8.498	0.807	10.525	$< 2.2 \times 10^{-16}$
log_pop	1.162	0.083	13.956	$< 2.2 \times 10^{-16}$
runtime	0.00738	0.00212	3.474	0.00052
log_cast	0.291	0.069	4.218	2.536×10^{-5}
log_crew	0.062	0.051	1.213	0.225

Table 5: Variance inflation factors (VIF) for Model 2 (m2).

Variable	VIF
log_budget_use	22.30
budget_missing	20.63
log_pop	1.57
runtime	1.07
log_cast	1.49
log_crew	1.61

4.4 Model 3 baseline (m3) and nonlinear refinement

Model 3 (m3) includes month and genre factors and a centered year term:

$$\begin{aligned}\log(1+\text{revenue}) \sim\ & \texttt{log\_budget\_use} + \texttt{budget\_missing} + \texttt{log\_pop} \\ & + \texttt{runtime} + \texttt{log\_cast} + \texttt{log\_crew} \\ & + \texttt{release\_year\_c} + \texttt{release\_month\_f} + \texttt{primary\_genre\_top}.\end{aligned}$$

Testing nonlinearity (quadratic terms). We tested quadratic terms via nested partial F-tests:

$$\begin{aligned}\texttt{m3\_y2\_clean} &: \texttt{m3} + \text{poly}(\texttt{release\_year\_c}, 2), \\ \texttt{m3\_run2\_clean} &: \texttt{m3\_y2\_clean} + \text{poly}(\texttt{runtime}, 2).\end{aligned}$$

Table 6: Nested partial F-tests for quadratic terms in year and runtime.

Model	Res.Df	RSS	Df	Sum Sq	p-value
m3	2971	13747	–	–	–
m3_y2_clean	2970	13472	1	275.38	8.215×10^{-15}
m3_run2_clean	2969	13423	1	48.14	0.001115

Table 7: Model selection criteria for m3 vs m3_run2_clean.

Model	AIC	BIC	Adj. R^2
m3	13140.23	13320.42	0.5063
m3_run2_clean	13072.78	13264.99	0.5176

Decision: We selected m3_run2_clean because it improves fit (lower AIC/BIC; higher adjusted R^2) while remaining interpretable.

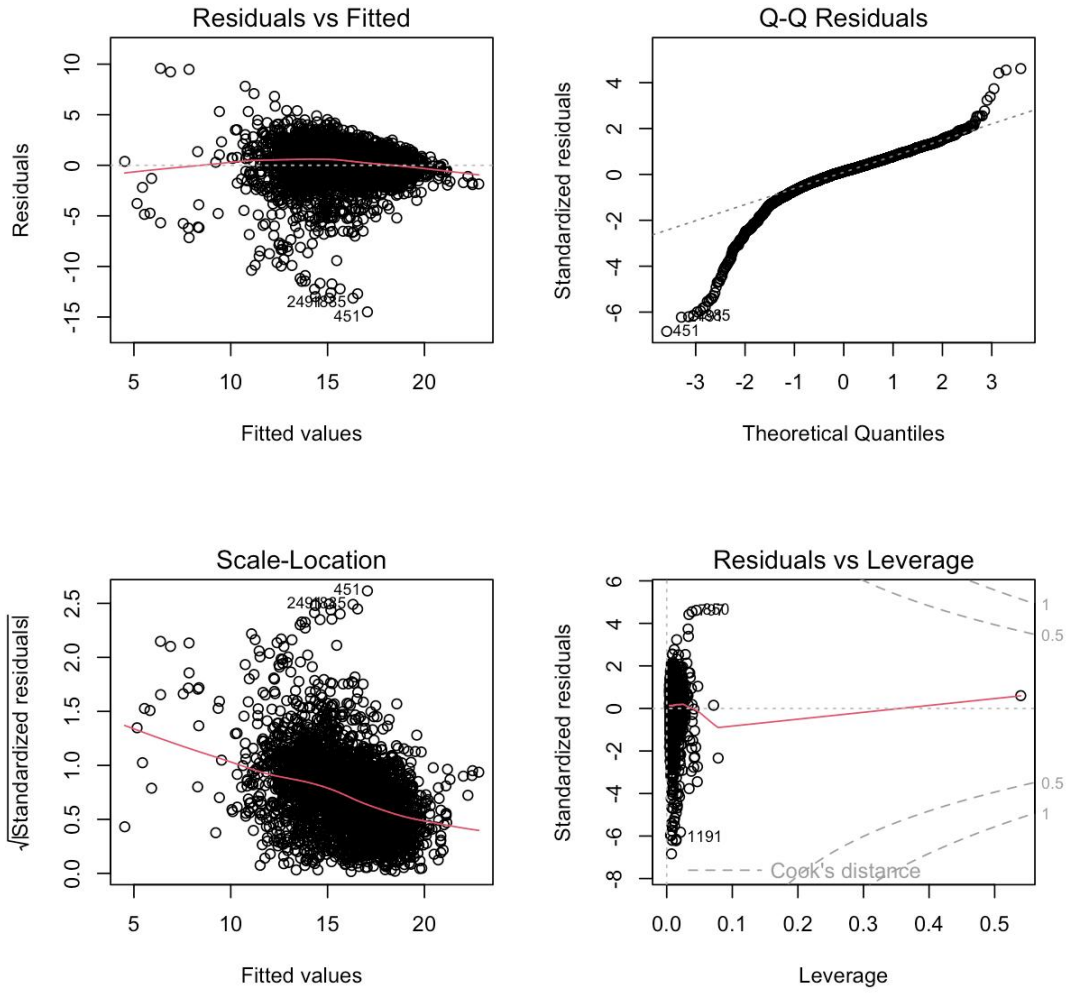


Figure 10: Diagnostic plots for `m3_run2_clean`. Remaining heteroskedasticity motivates robust inference.

Residual assumptions for the selected model (`m3_run2_clean`)

Heteroskedasticity (formal test). A studentized Breusch–Pagan test strongly rejected homoskedasticity:

$$BP = 222.33, \quad df = 30, \quad p < 2.2 \times 10^{-16}.$$

Action: We report inference for the selected model using **HC3 robust standard errors**.

Normality. The Q–Q plot indicates heavy-tailed residuals. With $n = 3000$, the primary concern is valid inference under heteroskedasticity rather than exact normality. **Action:** Robust (HC3) standard errors are used; conclusions are checked via influence sensitivity analysis.

Influence diagnostics and sensitivity analysis We evaluated influence using Cook's distance. The largest Cook's distances were modest (maximum ≈ 0.032). As a robustness

check, we refit the selected model excluding the top 10 observations by Cook's distance:

$$850,\ 797,\ 1191,\ 2696,\ 1918,\ 2384,\ 2491,\ 1162,\ 470,\ 2578.$$

Comparing HC3-robust coefficient tests before and after removal showed that key predictors (budget, popularity, and the quadratic year/runtime terms) retained the same direction and statistical significance. Some secondary categorical effects (specific months/genres) were more sensitive and are interpreted cautiously.

Decision: We retained all observations and documented sensitivity results as a robustness check rather than a data-cleaning step.

4.5 Weighted least squares (WLS) sensitivity check (not selected)

Because heteroskedasticity remained in Model 3, we tested WLS as a *sensitivity* approach. In WLS, the estimator is most efficient only when the weights are proportional to the inverse conditional error variance:

$$w_i \ \propto\ \frac{1}{\mathrm{Var}(\varepsilon_i \mid X_i)}.$$

In our check, we used a heuristic weight based on fitted values from `m3`:

$$w_i \ =\ \frac{1}{\max(\hat{y}_i^2,\ \varepsilon)}, \qquad \varepsilon = 10^{-6},$$

which implicitly assumes

$$\mathrm{Var}(\varepsilon_i \mid X_i) \ \approx\ \hat{y}_i^2.$$

This variance model was not derived or validated from the data, so the WLS fit could mis-weight observations if the assumption is incorrect.

WLS fit summary. The WLS sensitivity fit (`m3_wls`) achieved

$$R^2 = 0.5148, \qquad \bar{R}^2 = 0.5101,$$

with weighted residual standard error

$$\widehat{\sigma}_w = 2.248 \quad \text{on } 2972 \text{ degrees of freedom},$$

and an overall model test

$$F(27,\ 2972) = 109.4, \qquad p < 2.2 \times 10^{-16}.$$

It solves a different problem than the one required. Given the significant Breusch–Pagan result for the selected model, the minimum requirement for valid inference is heteroskedasticity-robust standard errors (HC3), which deliver valid standard errors and p-values under heteroskedasticity *without specifying* $\mathrm{Var}(\varepsilon_i \mid X_i)$. By contrast, WLS changes the objective function and can alter coefficient estimates (the mean fit) depending on the chosen weights, which is risky when the weights are not well-justified.

It did not fix the diagnostics. The WLS diagnostic plot in Figure 11 still shows substantial tail deviations in the Q–Q plot and a remaining trend in the scale–location panel, indicating that heavy tails and non-constant variance patterns persist. Thus, WLS did not provide a clean "assumptions now OK" outcome.

Decision. We therefore did not adopt WLS as the primary remedy. Instead, we kept the OLS mean specification for interpretability and reported inference using HC3 heteroskedasticity-robust standard errors to obtain reliable standard errors, confidence intervals, and p-values under heteroskedasticity.

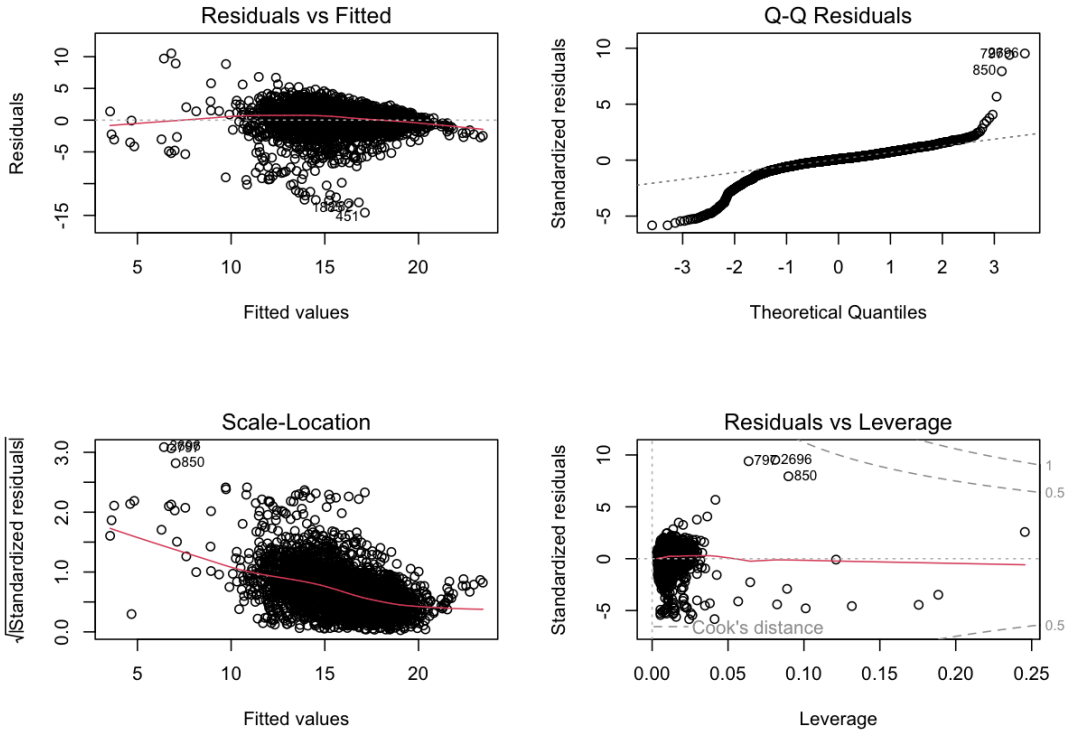


Figure 11: WLS diagnostic plots for Model 3 (`m3_wls`). Tail deviations in the Q–Q plot and a trend in the scale–location panel remain visible.

4.6 Alternative remedies considered

We explored spline terms as an alternative to quadratic effects. Although splines improved fit relative to `m3`, the quadratic model `m3_run2_clean` provided strong improvement with lower complexity and clearer interpretation. Weighted least squares (WLS) was also considered; however, WLS requires a correctly specified variance model (weights). Instead, HC3 robust inference provides valid standard errors under heteroskedasticity without imposing a parametric variance form.

5 Final Models

5.1 Model comparison and selection

Three nested specifications were compared: the baseline full model (`m3`), the quadratic extension (`m3_run2_clean`), and an alternative spline model (`m3_spl2`). Table 8 reports the nested F-tests for nonlinear terms, while Table 9 summarizes information criteria and 10-fold cross-validation prediction error.

Table 8: Nested partial F-tests for nonlinear extensions.

Model	Res.Df	RSS	Df	Sum Sq	F	p-value
`m3`	2971	13747.36				
`m3_y2_clean`	2970	13471.98	1	275.38	60.91	8.22e-15
`m3_run2_clean`	2969	13423.85	1	48.14	10.65	0.0011

Table 9: Model comparison by information criteria and 10-fold cross-validation.

Model	df	AIC	BIC	Adj. R^2	RMSE	RMSE SD	MAE	MAE SD
`m3`	30	13138.26	13318.45	0.5063	1.811	0.079	1.356	0.061
`m3_run2_clean`	32	13072.78	13265.00	0.5176	1.803	0.079	1.345	0.060
`m3_spl2`	40	13114.60	13354.84	0.5135	1.818	0.081	1.352	0.060

Decision: The quadratic specification `m3_run2_clean` achieved the lowest AIC/BIC, the highest adjusted R^2, and the best cross-validated RMSE and MAE. It was therefore chosen as the final model.

5.2 Final model specification

The final fitted equation is:

$$\begin{aligned}\log(1+\text{revenue}) = \beta_0 &+ \beta_1\,\texttt{log\_budget\_use} + \beta_2\,\texttt{budget\_missing} + \beta_3\,\texttt{log\_pop} \\ &+ \beta_4\,\texttt{log\_cast} + \beta_5\,\texttt{log\_crew} + \sum_{m=2}^{12}\gamma_m\,\mathbf{1}(\texttt{month}=m) \\ &+ \sum_{g}\delta_g\,\mathbf{1}(\texttt{genre}=g) + \text{poly}(\texttt{release\_year\_c},2) + \text{poly}(\texttt{runtime},2) + \varepsilon.\end{aligned}$$

5.3 Parameter estimates and robust inference

The Breusch–Pagan test ($BP = 222.22$, $df = 29$, $p < 2.2\times10^{-16}$) confirmed heteroskedasticity, so inference is based on HC3 heteroskedasticity-robust standard errors. Table 10 reports robust estimates, 95% confidence intervals, and significance levels.

5.4 Interpretation in context

Because the response is $\log(1+\text{revenue})$, coefficients on log-transformed predictors represent approximate **elasticities**:

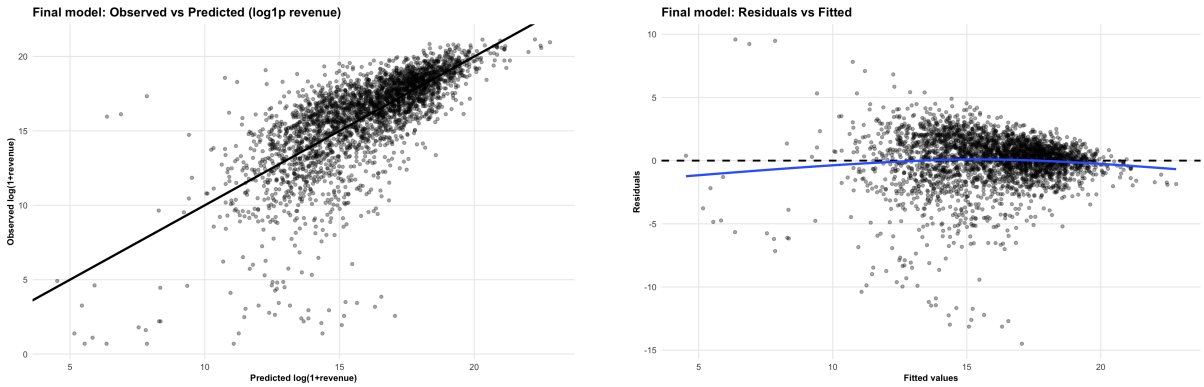
Table 10: Final model (`m3_run2_clean`) with HC3 robust standard errors and 95% confidence intervals.

Term	Estimate	Robust SE	t	p	95% CI
(Intercept)	0.399	0.951	0.419	0.6751	[-1.465, 2.263]
log_budget_use	0.616	0.055	11.115	$<2.2\times10^{-16}$ ***	[0.508, 0.725]
budget_missing	8.201	0.889	9.227	$<2.2\times10^{-16}$ ***	[6.460, 9.943]
log_pop	1.168	0.082	14.244	$<2.2\times10^{-16}$ ***	[1.007, 1.329]
log_cast	0.209	0.069	3.008	0.0027 **	[0.073, 0.345]
log_crew	0.115	0.050	2.296	0.0217 *	[0.017, 0.213]
poly(release_year_c, 2)1	-0.039	0.004	-9.561	$<2.2\times10^{-16}$ ***	[-0.047, -0.031]
poly(release_year_c, 2)2	-0.001	0.000	-5.653	1.72e-08 ***	[-0.001, -0.000]
poly(runtime, 2)1	0.038	0.011	3.590	0.0003 ***	[0.018, 0.059]
poly(runtime, 2)2	-0.000	0.000	-2.946	0.0032 **	[-0.000, -0.000]

Notes: HC3-robust p-values; significance codes: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

- **Budget elasticity:** a 1% increase in production budget is associated with a 0.62% increase in expected revenue, holding all else constant.
- **Popularity elasticity:** a 1% increase in TMDB popularity corresponds to a 1.17% rise in expected revenue, the strongest marginal effect in the model.
- **Cast / crew size:** both have positive but smaller elasticities, suggesting diminishing returns to scale in production team size.
- **Time and runtime:** significant quadratic terms imply nonlinear trends, capturing diminishing or accelerating effects over time and with longer runtimes.

5.5 Model fit and predictive performance



(a) Observed vs predicted (log1p revenue) **(b)** Residuals vs fitted

Figure 12: Model fit diagnostics for the final specification `m3_run2_clean`.

Observed-predicted values lie near the 45° line, indicating good log-scale calibration. Residuals show no major curvature but mild heteroscedasticity, consistent with the Breusch–Pagan test.

5.6 Comparison and conclusion

Across all evaluation criteria— (i) partial F-tests (Table 8), (ii) residual and influence diagnostics (Figure 12), and (iii) 10-fold cross-validation (Table 9)— the quadratic specification `m3_run 2_clean` achieves the best balance between interpretability, in-sample fit, and out-of-sample

predictive accuracy. It is therefore retained as the **final model** for explaining and forecasting log-transformed movie revenue.

6 Discussion

6.1 Final Conclusions

This analysis established a predictive model for worldwide box office revenue, explaining approximately 52% of the variance in the log-transformed response (Table 9). Our findings highlight three primary drivers of financial success:

- **The primacy of popularity and budget:** While production budget is a critical prerequisite for high revenue, our final model indicates that `popularity` (a proxy for pre-release buzz and audience engagement) has the highest elasticity (about 1.17) compared to budget (about 0.62). This suggests that while spending money is necessary, generating audience interest yields a proportionally higher return on investment.
- **Diminishing returns to scale:** The significance of log-transformed predictors and the negative quadratic term for runtime suggest that "more" is not always better. Increasing cast size or runtime contributes to revenue only up to a point, after which the marginal benefit declines.
- **Seasonal and temporal dynamics:** We confirmed that release timing is strategic. Significant coefficients for specific months indicate that aligning releases with peak windows (summer and December) is associated with systematically higher revenues, independent of budget.

6.2 Limitations

Despite the robustness of the final model (verified via influence diagnostics and cross-validation), several limitations must be acknowledged:

- **Data quality and missingness:** Approximately 27% of films in the dataset had missing or zero reported budgets. While we employed a missingness indicator (`budget_missing`) to retain these observations, this approach assumes the missingness mechanism is captured by the indicator and other covariates. If budget reporting is correlated with unobserved factors (e.g., independent studios with systematically different reporting practices), estimates may still be biased.
- **Omitted variable bias:** Our model explains roughly half of the variance ($R^2 \approx 0.52$). The remaining unexplained variance suggests the influence of unobserved variables, such as marketing spend (distinct from production budget), intellectual property strength (sequels vs. originals), distribution intensity, and critic reviews.
- **Heteroskedasticity:** Despite the log transformation, diagnostic plots (Figure 12) and the Breusch–Pagan test confirmed persistent heteroskedasticity. While we mitigated inference risk using HC3 robust standard errors, the residual "fan shape" implies predictive precision varies with movie scale; blockbusters remain harder to predict precisely than smaller films.
- **Causality vs. association:** As with all observational regression analyses, these findings represent associations, not causal mechanisms. Increasing a budget does not guarantee higher revenue; rather, high-budget films tend to co-occur with other characteristics (marketing, stars, wide release) that also drive revenue.