

Interpreting the impact of lead actor gender and age on movie gross

Hannah Lewis

Metis, Linear Regression & Web Scraping Module
August 6, 2021



Overview

Motivation

- Female protagonists—particularly, female protagonists portrayed by actors over the age of ~ 35 —are significantly underrepresented in movies

Objectives

- Determine what (if any) impact these differences in actor demographics (i.e., gender and age) have on the lifetime gross of a movie

Results

- Lifetime movie gross is strongly dependent on how it performs during its opening weekend, the budget, its reception by audiences, and its genre

Data & Methods

Box Office Mojo
by IMDbPro

IMDb

Rotten
Tomatoes

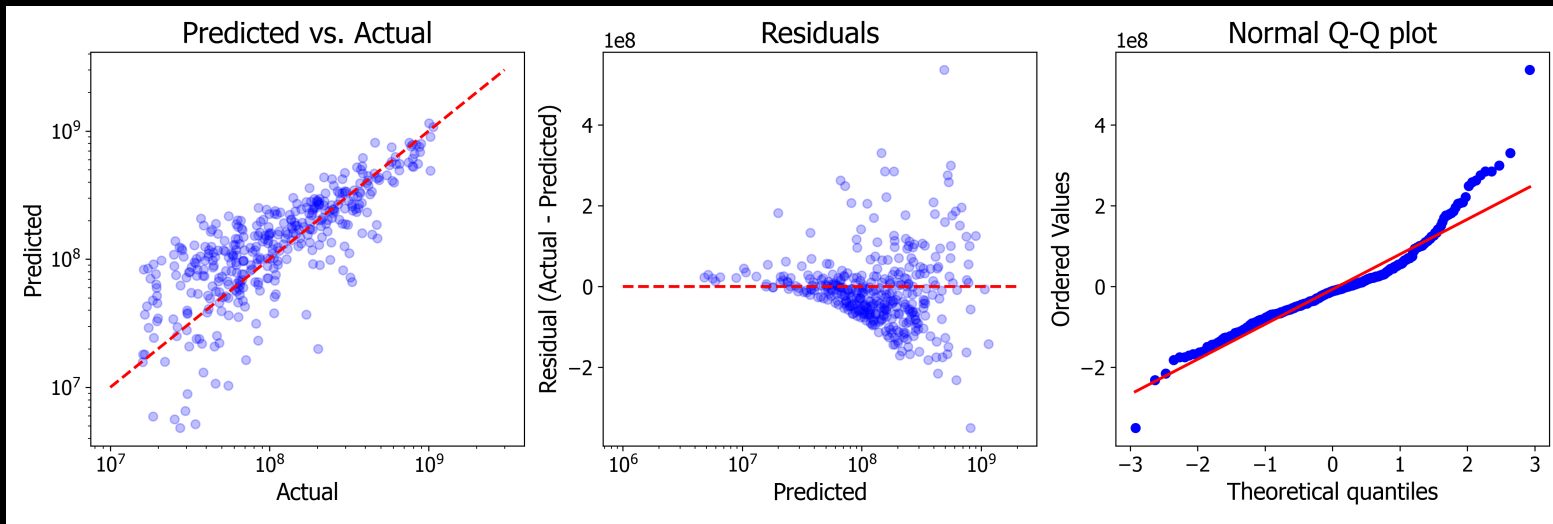
Target: worldwide lifetime gross

Features: domestic opening gross, budget, studio, MPAA rating, run time, genre, lead name, lead age, lead height, lead gender, Rotten Tomatoes audience and Tomatometer scores, international movie?

- Baseline models trained on domestic opening gross (and its square), budget, and Rotten Tomatoes scores as features
- Select ridge regression model (validation $R^2 \sim 0.720$, MAE $\sim \$68.4M$)

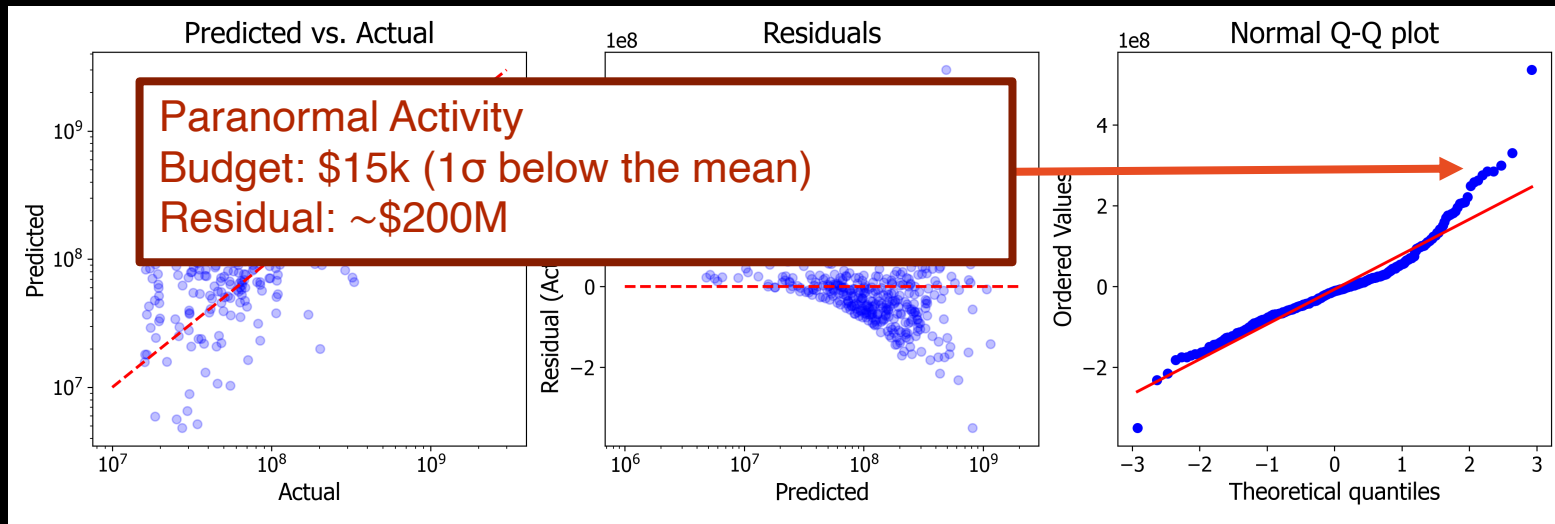
- Expanded model includes lead actor gender and age as features, as well as dummy variables for the studio that produced the movie and the genre of the movie

Results: model performance



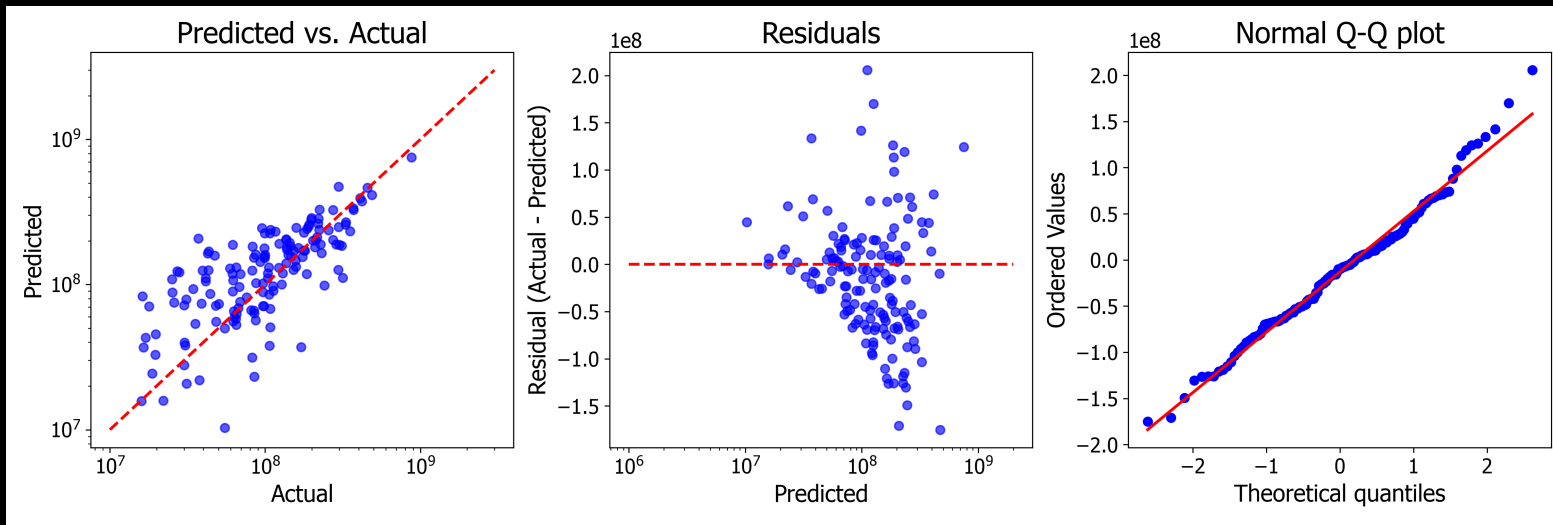
- Model does not generalize well for outliers in budget
- For any budget (\$15k up to 300M), MAE = \$63.4M

Results: model performance



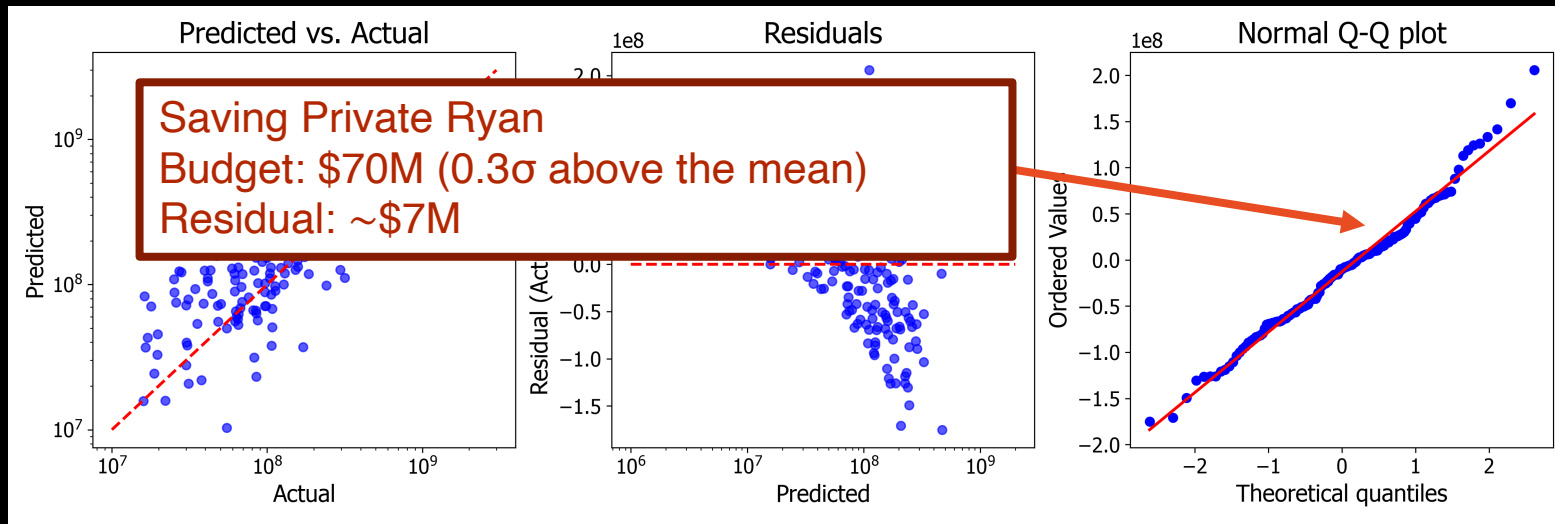
- Model does not generalize well for outliers in budget
- For any budget (\$15k up to 300M), MAE = \$63.4M

Results: model performance



- Model works best for movies with budgets within ± 0.5 standard deviations of the mean (\$30M up to \$80M), MAE = \$51.0M

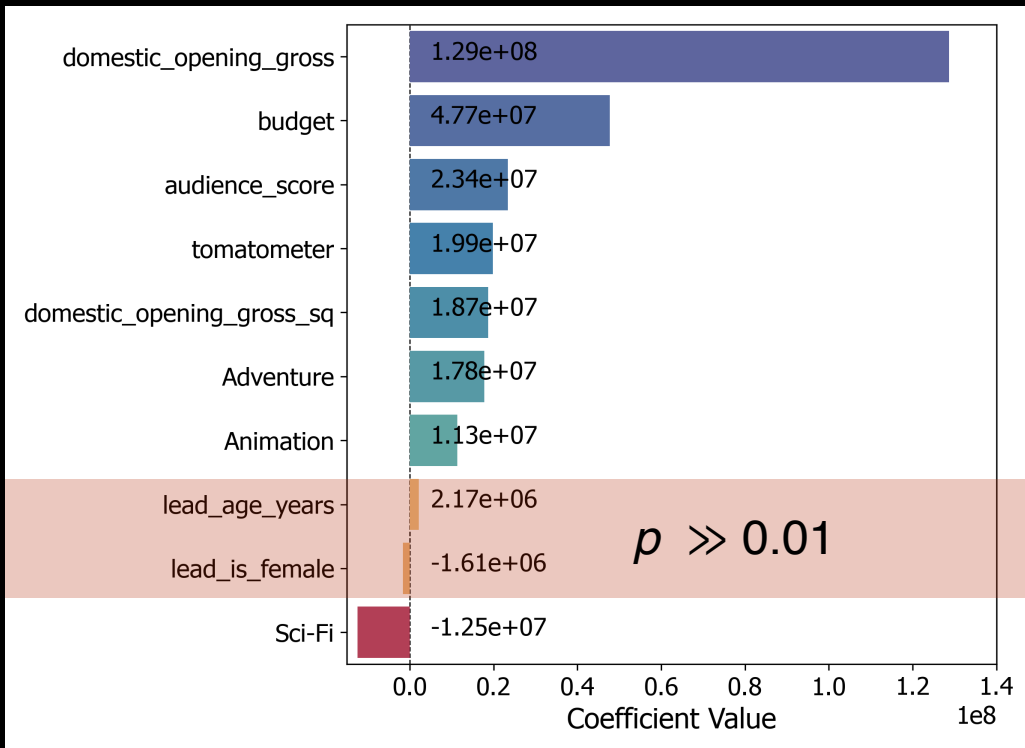
Results: model performance



- Model works best for movies with budgets within ± 0.5 standard deviations of the mean (\$30M up to \$80M), MAE = \$51.0M

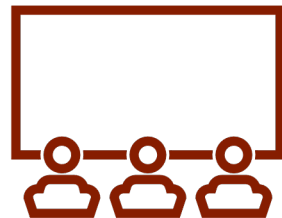
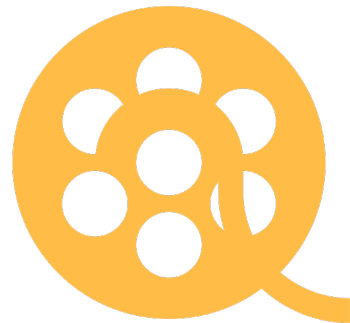
Results: significant features

- Significant coefficients
 - $p < 0.01$
 - Reject the null hypothesis that $\beta_i = 0$
- In the final model, lead actor age and gender **do not** have significant coefficients



Conclusions & Recommendations

- Lead actor gender and age are not among the most important features that determine movie gross
- Studios should produce a larger fraction of movies that are led by a more gender and age-diverse range of actors





Future Work

- Scrape movies with lower lifetime gross and broader range of budgets
- Subset data by genre (e.g., action vs. romance) to see in gender does become a significant feature

Questions?



