STATS 551 Final Report Group 23

Abner Heredia Bustos, Chittaranjan Velambur Rajan, Brendan Matthys

12/14/2022

Contributions

Abner Heredia Bustos was in charge of specifying the model and its hierarchical structure, deriving the posterior distributions, writing the code in R to sample from the posterior distributions, and writing the code to run the simulations with Gibbs sampling. He was also in charge of writing the section "Method" in the final report.

Brendan Matthys: was in charge of data cleaning, model diagnostics, exploratory data analysis, calculation, and the presentation and interpretation of results.

Chittaranjan Velambur Rajan was in charge of data cleaning, data scraping, derivation of posterior distributions, writing code for sampling functions, and the presentation and interpretation of results.

Introduction:

People expect different things from different genres of movies. We expect plenty of CGI and big explosions in superhero movies but not in Westerns; and we expect intricate plots in spy thrillers but not in slasher movies. So, it is possible that people value the features of movies from different genres.

Our main goal is to understand what features are important in choosing whether to watch a movie on its opening weekend, and if that importance depends on the genre of the film. To estimate the importances, we use a hierarchical linear regression where revenue in the opening weekend is the dependent variable and movie characteristics are the covariates. This model is described in detail in the methods section.

Data Description

Our data was scraped from both IMDB and Box Office Mojo, and the original columns scraped for each movie were: imdb_id (a unique identifier on imdb's website for each movie), the title of the movie, the genre or genres of the movie, the release date, the country of origin (original data had multiple countries at some points), budget, revenue, duration.

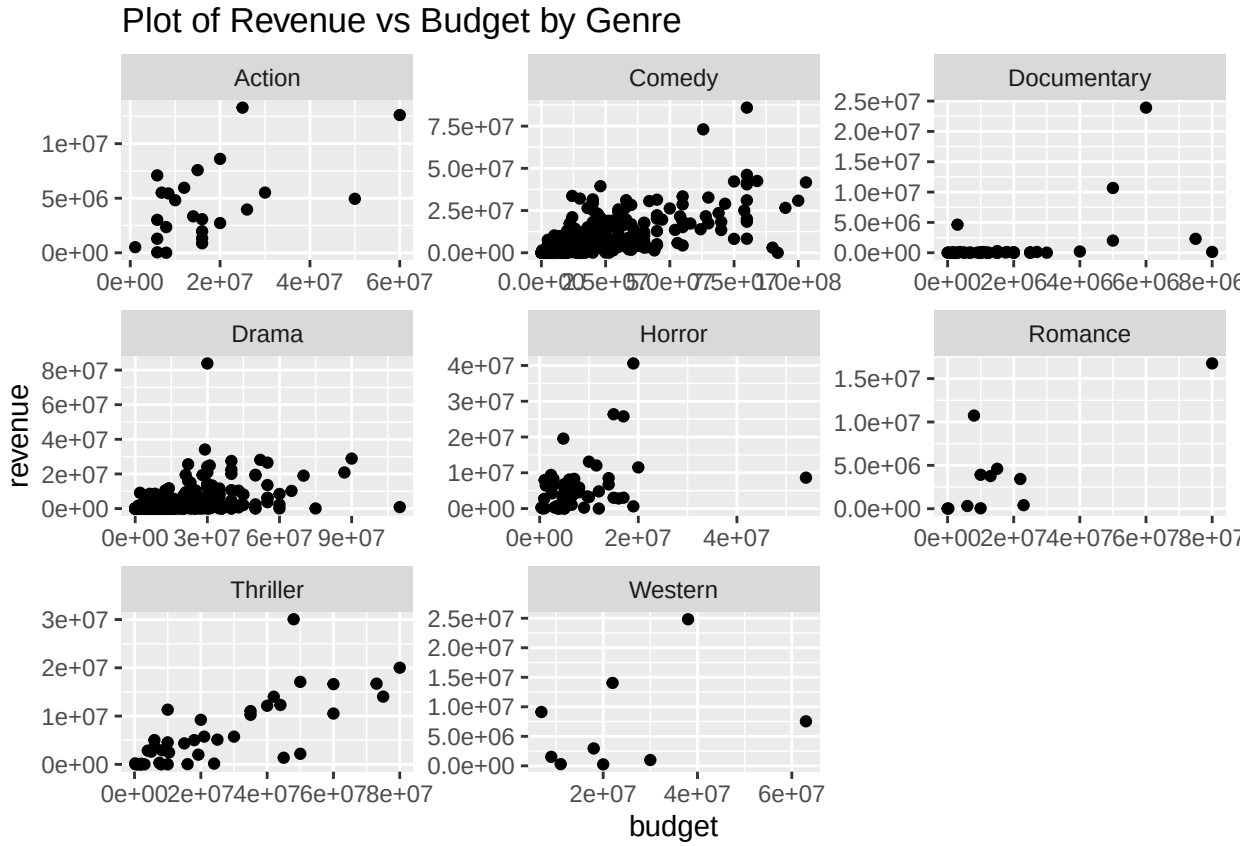
There was cleaning required before being able to use this in our model. First of all, we only wanted to use movies with a single genre. Including movies with multiple genres in our analysis would be too complicated for the model, so we decided to filter out the observations just to include single genre films.

The release date needed to be cleaned as well, but there weren't any major transformations that needed to be executed, just string manipulation.

Budget matters because having a bigger budget means that a movie can include more popular features. Famous and/or talented actors charge more for their participation than relatively unknown ones. It is expensive to produce flashy special effects and to travel to exotic locations. The budget is related to these appealing features, so more people may want to watch movies with larger budgets. Thus, budget may be seen as a predictor of movie revenue

The revenue was our response variable in this instance. In this model, revenue represents the opening weekend revenue earned by the movie, in US Dollars.

The excess duration represents the length of the movie past a standard 120 minutes. From our initial analysis, we think that duration could have some sort of effect on the revenue in our hierarchical model, and that it is worth including. Specifically, we were interested in determining the effect of how much longer the movie is than two hours on the revenue. In the opening weekend where there is less information about the movie, people may choose to watch longer movies only after they know more about them, or perhaps not watch them at all. Thus, we expect excess duration to have a negative effect on the revenue during the opening weekend.



This is a preliminary look at the relationship between revenue and budget. From these scatterplots, there seems to exist a postive relationship for every genre. Thus, we found it appropriate to include in our model.

Methods

We use four variables from our data set: revenue, genre, budget, and excess duration. Revenue is the amount of money, measured in dollars, that the movie earned during its opening weekend. Genre is the genre that imdb gives to the movie. Budget is the estimated amount of money spent to make the movie, also measured in dollars. Excess duration is the number of minutes above 120 (two hours) of the movie duration. So, for example, a movie with a duration of 2 hours and 15 minutes would have an excess duration of 15. If a movie lasts 120 minutes or less, we set the excess duration to 0.

For each genre, we want to estimate how much budget and excess duration matter to people when choosing whether to watch a movie on its opening weekend. To determine this, we choose all the movies from our sample that imdb classifies as belonging to only one genre. People may look at movies from "hybrid" genres (e.g., romantic comedy) as combinations of "pure" genres. If so, the importance they give to budget and excess duration may also be a combination of the importance they would give in each separate genre. Thus, by focusing on movies from pure genres, we hope to find the most marked differences between the importance of movie characteristics.

We write a linear model for each group. Let $\mathbf{y}_g$ be the vector[1] of length n_g revenues for movies from genre g; let $\mathbf{x}_1^{(g)}$ be the vector of budgets; let $\mathbf{x}_2^{(g)}$ be the vector of excess durations; and let $\boldsymbol{\epsilon}_g$ be a vector if iid Normal$(0, \sigma_g^2)$. Then

$$\begin{aligned}\mathbf{y}_g =& \beta_0^{(g)} + \beta_1^{(g)}\mathbf{x}_1^{(g)} + \beta_2^{(g)}\mathbf{x}_2^{(g)} + \boldsymbol{\epsilon}_g \\ =& X_g\boldsymbol{\beta}_g + \boldsymbol{\epsilon}_g\end{aligned}$$

where $\boldsymbol{\beta}_g^T = (\beta_0^{(g)}, \beta_1^{(g)}, \beta_2^{(g)})$ and X_g is a matrix with columns given by $(\mathbf{1}, \mathbf{x}_1^{(g)}, \mathbf{x}_2^{(g)})$.

Thus $(\mathbf{y}_g \mid X_g, \boldsymbol{\beta}_g) \sim \text{Normal}(X_g\boldsymbol{\beta}_g, \sigma_g^2\mathbf{I}_g)$, where $\mathbf{I}_g$ is the $n_g \times n_g$ identity matrix.

We believe that there is a dependency between the parameters $\boldsymbol{\beta}_g$ and σ_g^2 of all genres. If we knew that increasing the budget also increases expected revenue for action, comedy, and horror movies, we would expect that more budget also increases revenue for other types of movies.

Our key assumption is that, conditional on the genres, observations $y_i \mid \mathbf{x}_i$ are exchangeable. Then, by de Finetti's theorem we can model the dependency between the genres by setting population distributions for $\boldsymbol{\beta}_g$ and for σ_g^2. This way we can share information from each genre to improve our estimates of all the other genres.

Consider

- $\boldsymbol{\beta}_g \sim \text{Normal}(\boldsymbol{\theta}, T)$ for all g.
- $\sigma_g^2 \sim \text{Inverse-gamma}(\text{shape} = v/2, \text{scale} = vu^2/2)$ for all g.

Now we assume that the genres of movies are also exchangeable, so we apply de Finetti's theorem again and set

- $\boldsymbol{\theta} \sim \text{Normal}(\boldsymbol{\mu}_0, W_0)$.
- $T \sim \text{Inverse Wishart}(t_0, S_0^{-1})$.
- $v \sim \text{Gamma}(\text{shape} = \delta, \text{rate} = \lambda)$.
- $u^2 \sim \text{Gamma}(\text{shape} = a, \text{rate} = b)$.

See figure below for a graphical representation of our model.

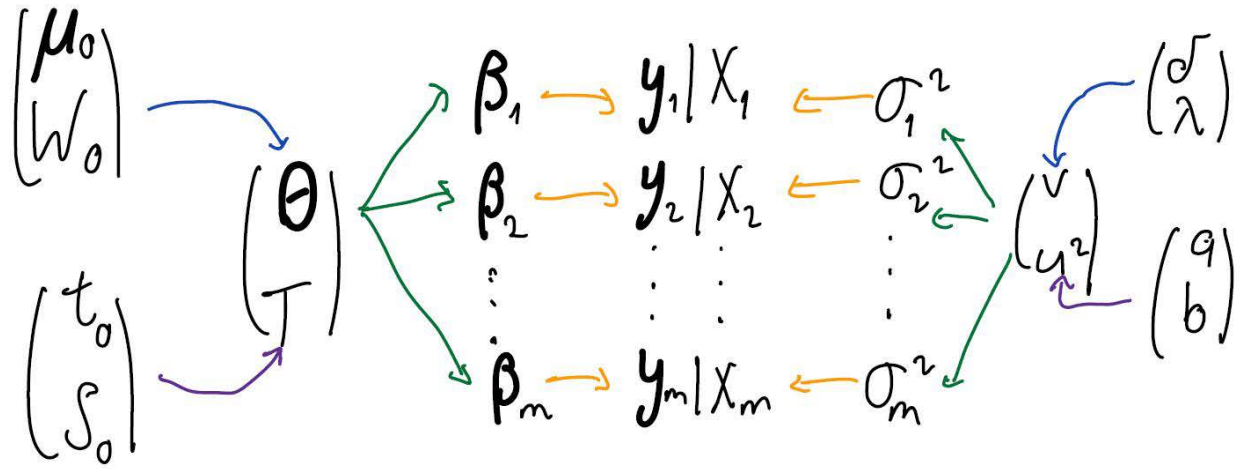


Figure 1: Hierarchical Structure

[1]All vectors are in column form.

It can be shown that, with these prior specifications and using conditional independence, the full posterior distributions for the parameters are:

- $\boldsymbol{\beta}_g \mid \mathbf{y}_g, X_g, \boldsymbol{\theta}, T$ is a Multivariate Normal distribution with parameters
  - Mean $\left(T^{-1} + \frac{1}{\sigma_g^2} X_g^T X_g\right)^{-1} \left(T^{-1}\boldsymbol{\theta} + \frac{1}{\sigma_g^2} X_g^T \mathbf{y}_g\right)$
  - Covariance $\left(T^{-1} + \frac{1}{\sigma_g^2} X_g^T X_g\right)^{-1}$
- $(1/\sigma_g^2) \mid \mathbf{y}_g, X_g, \boldsymbol{\beta}_g, v, u^2$ has a gamma distribution with shape equal to $(v + n_g)/2$ and rate equal to $\frac{1}{2}(vu^2 + \text{SSR}_g)$, where $\text{SSR}_g = (\mathbf{y}_g - X_g\boldsymbol{\beta}_g)^T(\mathbf{y}_g - X_g\boldsymbol{\beta}_g)$.
- $\boldsymbol{\theta} \mid \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m, T$ is a Multivariate Normal distribution with parameters
  - Mean $\left(W_0^{-1} + mT^{-1}\right)^{-1} \times (W_0^{-1}\boldsymbol{\mu}_0 + T^{-1}\sum_{g=1}^m \boldsymbol{\beta}_g)$
  - Covariance $\left(W_0^{-1} + mT^{-1}\right)^{-1}$

  where m is the total number of genres.
- $T \mid \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m, \boldsymbol{\theta}$ is Inverse-Wishart$(t_0 + m, [S_0 + S_{\boldsymbol{\theta}}]^{-1})$, where $S_{\boldsymbol{\theta}} = \sum_{g=1}^m (\boldsymbol{\beta}_g - \boldsymbol{\theta})(\boldsymbol{\beta}_g - \boldsymbol{\theta})^T$.
- $u^2 \mid \sigma_1, \ldots, \sigma_m, v$ has a gamma distribution with shape equal to $a + mv/2$ and rate equal to $b + (v/2)\sum_{g=1}^m (1/\sigma_g^2)$.
- $v \mid \sigma_1, \ldots, \sigma_m, u^2$ has a density proportional to

$$\text{dgamma}(v, \text{shape} = \delta, \text{rate} = \lambda) \times \prod_{g=1}^m \text{dgamma}((1/\sigma_g^2), \text{shape} = v/2, \text{rate} = vu^2/2).$$

We choose these priors and hyperpriors so as to make our model simple to compute and to interpret. Except for the distribution of v, all prior and posterior distributions are conjugate. Still, the multivariate normal and the inverse Wishart distributions can model a wide variety of parameter combinations. So, our model does not seem overly simplistic.

Computing posterior estimates

We use a Gibbs sampling procedure to estimate the posterior distributions of all of our parameters. Our algorithm is

1. Set values for the hyperparameters at the "top" (third) level: δ, λ, which inform v; a, b inform u^2; $\boldsymbol{\mu}_0$ and W_0 inform $\boldsymbol{\theta}$; and t_0 and S_0 inform T.
2. Based on values from step 1, sample one item from each of the hyperprior distributions of the variables at the second level: $\boldsymbol{\theta}$, T, v and u^2.
3. With values from step 2, sample values of $\boldsymbol{\beta}_g$ and σ_g^2 for each of the genres g.
4. In each simulation s, sample one value from each parameter from their corresponding full posterior distributions as detailed above. v does not have a conjugate distribution, so we use the Metropolis-Hastings algorithm to sample from its full posterior distribution. We set the proposal distribution for v to be a Normal distribution with mean "old v" and variance 2.

These two paragraphs below are for interpreting the value of the betas. Budget matters because having a bigger budget means that a movie can include more popular features. Famous and/or talented actors charge more for their participation than relatively unknown ones. It is expensive to produce flashy special effects and to travel to exotic locations. The budget is related to these appealing features, so more people may want to watch movies with larger budgets. Thus, budget may be seen as a predictor of movie revenue.

Excess duration is Thus, excess duration is meant to capture whether the movie is "getting too long". If the movie is long, then going to watch it entails a "bigger" commitment. In the opening weekend where there is less information about the movie, people may choose to watch longer movies only after they know more about them, or perhaps not watch them at all. Thus, we expect excess duration to have a negative effect on the revenue during the opening weekend.

We attempt to use a hierarchical model with linear regression to model our data and interpret the parameter coefficients. The model is described in Figure 1. There are three layers in this model, and they are as follows.

Results and Interpretation:

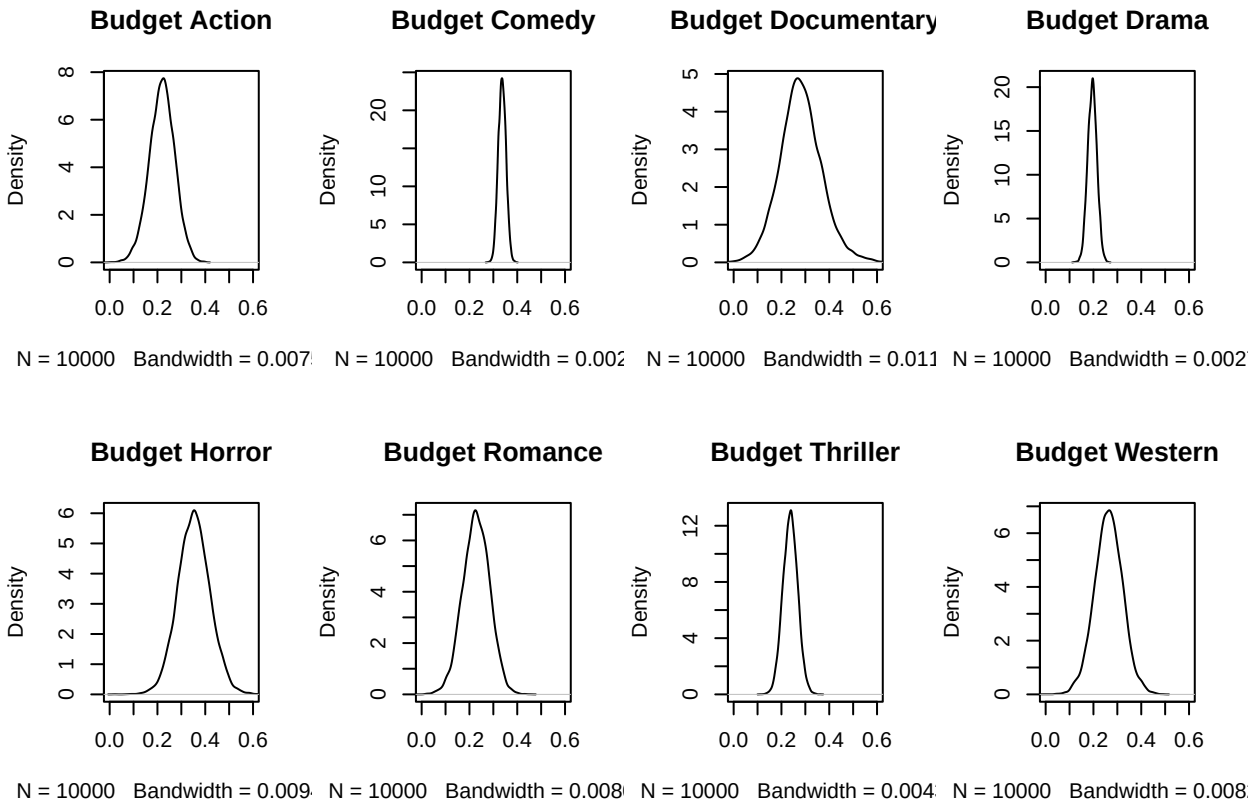
We conducted 10,000 iterations of the Gibbs sampler with our selected parameters. Below is a table representing the means of the betas for each genre.

Genre	Intercept	Budget	Excess Duration
Action	793131.2	0.219	-50663.57
Comedy	1096980	0.337	-56677.14
Documentary	754919.6	0.281	-50318.71
Drama	131700.1	0.195	-45363.6
Horror	1885349	0.353	-61435.8
Romance	748651.7	0.228	-51392.54
Thriller	723077.3	0.237	-50317.77
Western	904973.9	0.263	-56154.94

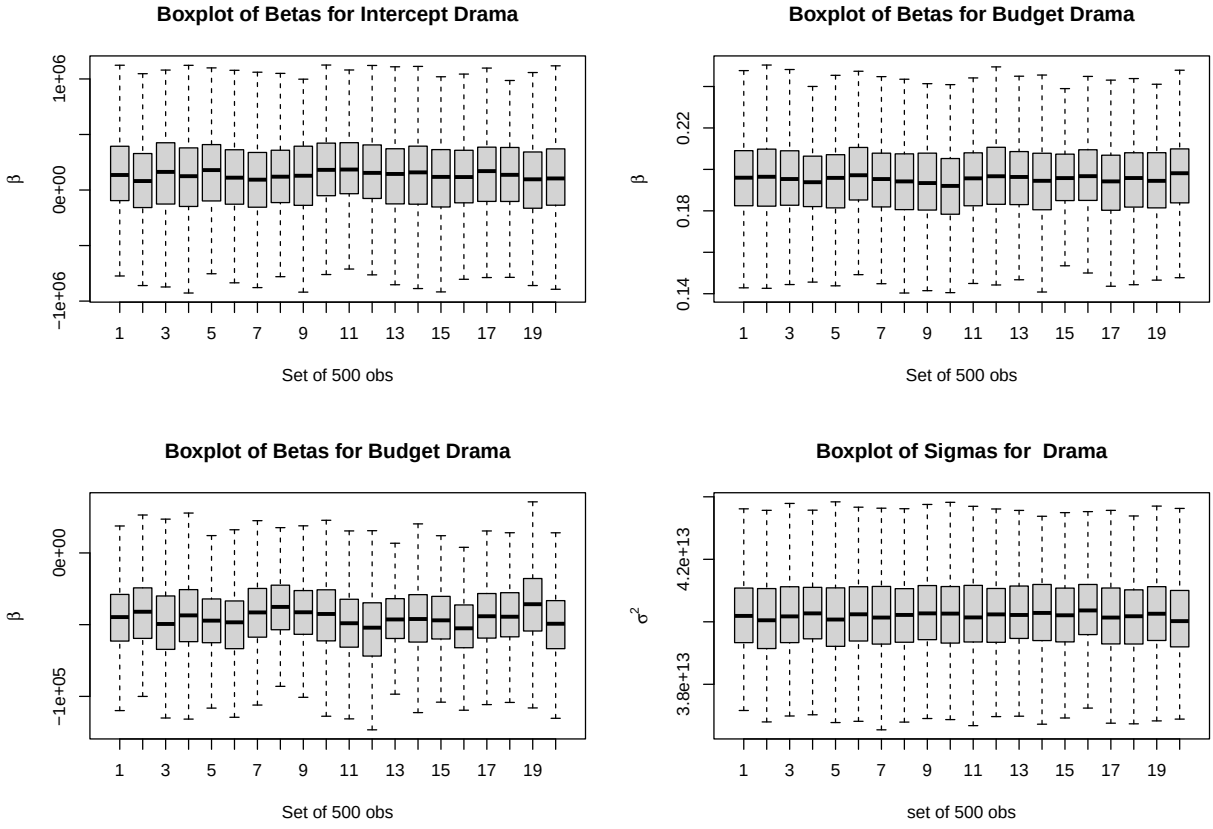
From this table, we can deduce that, at the most basic level, movies with higher budgets tend to generate higher revenues. Inherently this makes sense because movies with larger budgets tend to have stronger effects and bigger names associated with them. This would lead more people to come in and watch them. The negative coefficient for excess duration makes sense as well. Producing long movies costs more in terms of budget, and making a movie marginally longer shouldn't cause a substantial increase in viewership.

A really interesting takeaway from this table is that the genre with the highest beta for budget is horror, and the genre with the lowest beta for budget is drama. In dramas, one doesn't need too many special effects to create a successful film. The success of the movie is generally driven by the intrigue of the story. With regards to horror movies, many viewers watch because of the adrenaline rush generated by more realistic special effects. A higher budgeted movie will advertise stronger special effects, and thus more people will come to see it, leading to an increased revenue. Out of the genres above, horror is the genre which has a revenue that is most dependent on the quality of the special effects.

A distribution of the betas regarding the budget's effect on the revenue per genre can be found below. As confirmed by our table of betas, the horror genre seems to have the highest mean beta and drama looks to have the lowest mean beta. However, it looks as if the spread of these density plots is the highest for documentary and the lowest for comedy.



The above density plots resemble the densities regarding the beta parameter for budget for each genre. From this plot, we see that there is a lot of uncertainty for genres such as horror and documentary, whereas there is a lot more certaininty in the parameter for the comedy and drama genres.



The above boxplots represent the beta values for the intercept, budget, and excess duration, as well as the sigma-squared values for drama movies. Over many iterations, some of our parameters seem to converge, but some others seem unstable. We cannot attribute this to pure convergence issues, but maybe that the true space where the parameter lives being too hard to find in the allotted iterations.

Genre	**Mean of σ**
Action	6462204
Comedy	6882927
Documentary	6428063
Drama	6343471
Horror	6531584
Romance	6484759
Thriller	6456177
Western	6511200

With regards to the standard deviation σ by genres, drama has the lowest and comedy has the highest. This intuitively makes sense in that it's easier for a comedy to flop, and dramas are normally more consistent.

The below density plots follow closely to the means of the sigmas above.

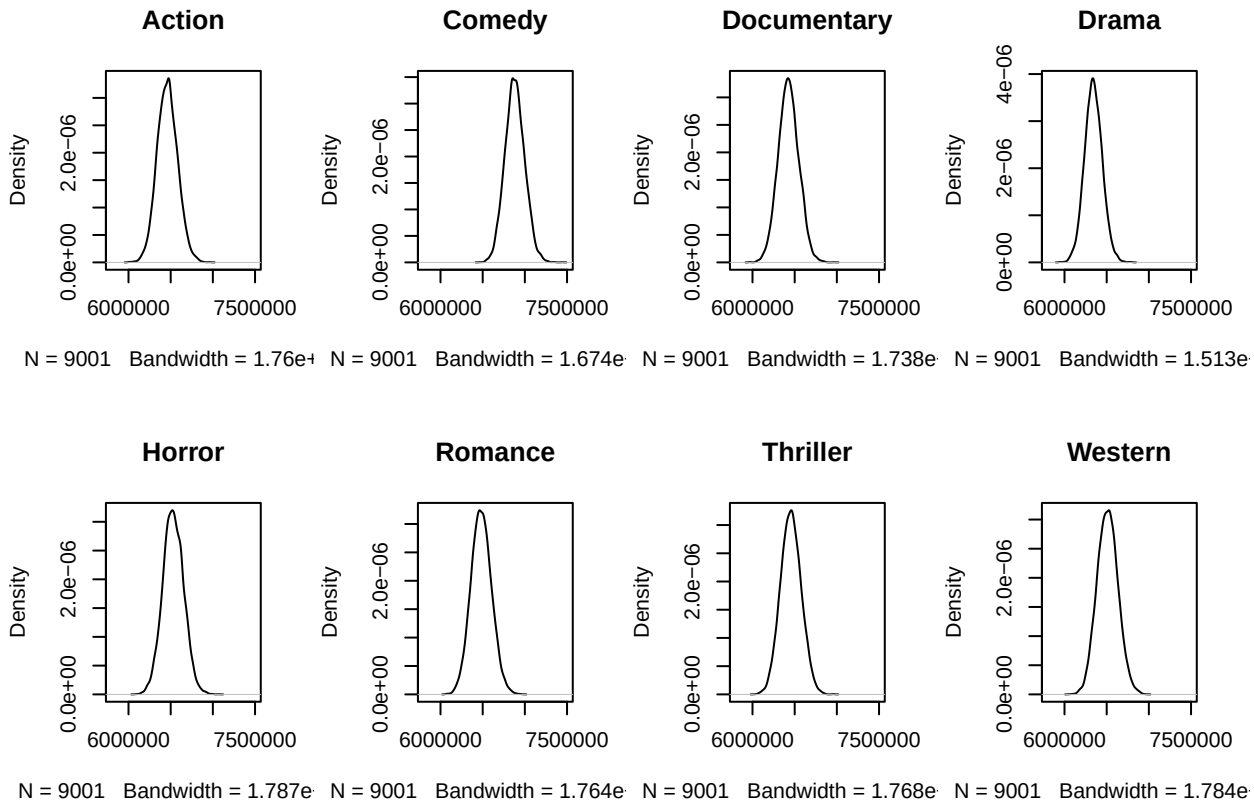


Table 1: RMSE per Genre

Genre	RMSE
Action	3591727
Comedy	6894122
Documentary	3141729
Drama	5502408
Horror	12669594
Romance	1712798
Thriller	4722269
Western	1781315

From creating predictions of our test data based on the model fit with the training data, we were able to find a root mean squared error for each genre. We found these to be reasonable values, and these RMSE's validate that the model parameters are legitimate.

Limitations

One of the major limitations our group faced while trying to scrape the data for this project was that after one trial of scraping the data, IMDB blocked the scraper we had. This meant that we were limited to the variables we had initially thought to scrape, and that we could not gather any more data past our initial scrape. We originally intended for the data set to be three times bigger than it was, but the web scraper getting blocked proved to be a significant issue. This proved to be an issue when some of our genres had very small sample sizes.

Conclusion

Our main goal was to understand what features are important in choosing whether to watch a movie on its opening weekend, and if that importance depends on the genre of the film.

In conclusion, we found that the importance of each feature is dependent on the genre of the film. Instead of having posterior values that were consistent with each other, we found that both the beta and the sigma values differed from genre to genre. This supports our initial hypothesis that the features of movies have an effect on how much money a movie makes during its opening weekend.