Board Game Insights Derived From User Reviews

Introduction

Dataset Description

The dataset is a subset of board games user ratings from the website BoardGameGeek. Originally posted to Kaggle, this subset was sourced from the TidyTuesday R for Data Science community. There are two tables—ratings and details. The ratings file contains a record for each board game and includes its average rating, rank on BoardGameGeek, number of reviews, and a few other miscellaneous fields such as year published and url. The details table contains information about each board game such as minimum/maximum number of players, estimated play times, publisher, minimum suggested age, etc.

Research Question

The goal is to analyze the relationship between game characteristics and the ratings to determine if certain types of games tend to receive higher ratings.

Methodology

The analysis applies Bayesian Linear Regression to model the relationship between game characteristics (such as min/max play time and min/max number of players) and the average user rating. This approach allows the use of prior information about the regression coefficients and updates it with the data to obtain a posterior distribution for each coefficient.

Bayesian linear regression allows the quantification of uncertainty in the regression coefficients while incorporating prior beliefs. The ratings are assumed to follow a normal distribution (normal likelihood), and the regression coefficients are assigned weakly informative normal priors.

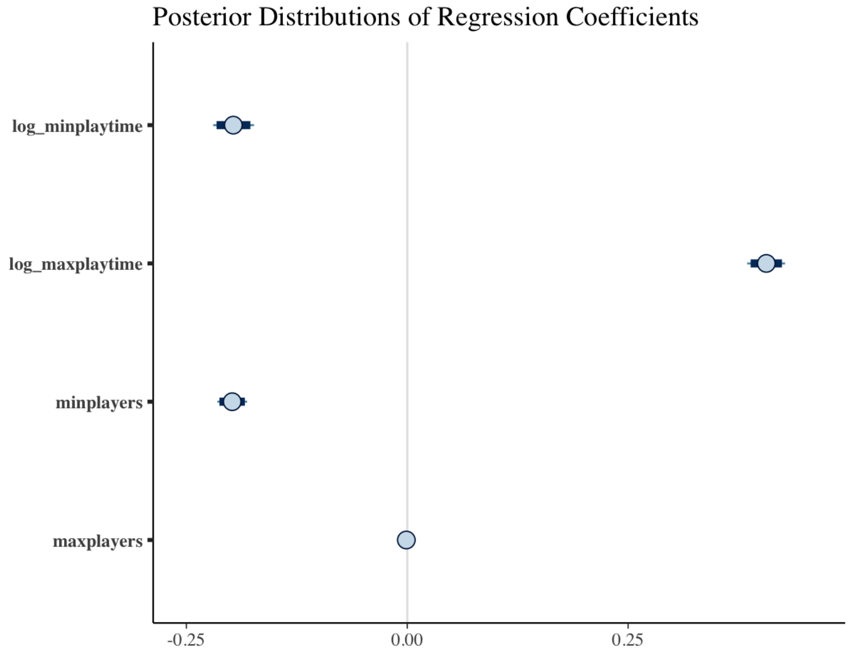
Model Application:

- **Data Preprocessing**: The two tables (ratings and details) are merged based on the BoardGameGeek game ID. The play time fields (minimum, maximum, and estimated play time) were log transformed to address skewness in these variables.
- **Posterior Approximation**: I fit a Bayesian linear regression model where the dependent variable was the average rating, and the independent variables were log-transformed min, max, and estimated play time and the number of players (min/max).
- **Priors:** I placed weakly informative normal priors on the regression coefficients, which allow for flexible estimation without strongly influencing the results.

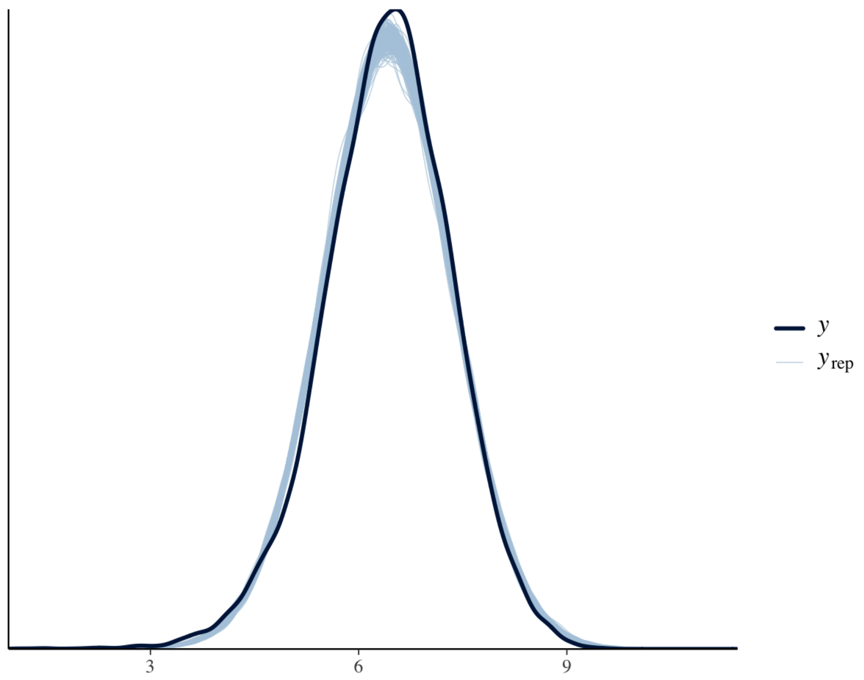
Results

Posterior Distributions

- **Intercept**: The average rating across all board games, holding the other variables constant, is approximately 6.0. The 95% credible interval for the intercept is [5.94, 6.04].
- **Log Minimum Play Time**: The coefficient for log minimum play time is approximately -0.2, with a 95% credible interval of [-0.22, -0.17], indicating that games with longer minimum play times tend to receive lower ratings.
- **Log Maximum Play Time**: The coefficient for log maximum play time is approximately 0.4, with a 95% credible interval of [0.38, 0.43], suggesting that games with longer maximum play times are generally rated higher.
- **Minimum Number of Players**: The coefficient for the minimum number of players is approximately -0.2, with a 95% credible interval of [-0.22, -0.18], indicating that games with a higher minimum number of players tend to have lower ratings.
- **Maximum Number of Players**: The coefficient for the maximum number of players is close to zero, with a 95% credible interval of [-0.002, 0.00], indicating that this variable has minimal influence on ratings.



Posterior Predictive Check (PPC): To assess the model’s goodness of fit, I conducted a posterior predictive check by comparing the predicted ratings from the model with the observed ratings. The following density overlay plot shows the distribution of observed ratings (solid dark line) versus the posterior predictive distribution (lighter lines).



The PPC plot reveals that the predicted distribution aligns very closely with the observed data, especially around the peak, indicating that the model captures the central tendency of the ratings very well. There is a slight deviation in the tails, where the model slightly underestimates the variability in ratings at the lower ends and overestimates the higher ends. However, the overall fit suggests that the model provides a good approximation for most of the data, reinforcing the validity of the posterior predictions.

Model Fit

The posterior predictive distribution's mean is around 6.4, which closely aligns with the average rating observed in the dataset. Diagnostics such as Rhat (all values close to 1.0) indicate that the model has converged well, and the effective sample sizes (n_eff) are sufficiently large, ensuring that the posterior estimates are reliable.

Prediction

- **Short Play Time Game**: For a game with a short play time (10 minutes minimum, 20 minutes maximum) and accommodating 2 to 6 players, the model predicts an average rating of approximately **6.34**.
- **Long Play Time Game**: For a game with a long play time (180 minutes minimum, 3000 minutes maximum) and the same player range (2 to 6), the model predicts an average rating of approximately **7.82**.

Discussion & Conclusions

- **Implications of Findings**: The analysis reveals that certain game characteristics can significantly affect board game ratings. Games with shorter minimum play times tend to have lower ratings, while games with longer maximum play times tend to have higher ratings. This suggests that players may appreciate games that can accommodate a wide range of session lengths, and specifically, the option for long games to extend into multiple sessions. The minimum number of players also negatively affects ratings, indicating that games requiring more players to start might be less favored by some users.
- **Suitability of Bayesian Linear Regression**: Bayesian linear regression was quite suitable for this dataset, as it provided precise estimates of the relationship between game characteristics and ratings, along with credible intervals for the coefficients. The use of weakly informative priors allowed the data to guide the results while accounting for uncertainty.
- **Concerns and Reliability**: One limitation of the dataset is the presence of outliers, such as games with extremely long play times, which could skew results. I attempted to mitigate this by log-transforming the play time variables, but further investigation into these extreme values or removal of outliers might improve the model. Additionally, some interactions between variables (e.g., how the number of players affects the relationship between play time and rating) could be explored in future analyses.
- **Future Directions**: In future analyses, I could consider including more variables (such as game publisher and genre/category) and exploring interaction effects between variables to provide a deeper understanding of what drives board game ratings. Other considerations could be made such as only including games with more than a set minimum number of ratings since the games with a small sample size of ratings may not reflect the community's true sentiment towards those games.

Bibliography

Dataset: TidyTuesday subset of BoardGameGeek data, originally posted to Kaggle.
https://github.com/rfordatascience/tidytuesday/blob/master/data/2022/2022-01-25/readme.md