Creating a Linear Model for Board Games with their Average Rating as the Response

By Adam Musa

Student Number 1005909073

Introduction

Board games have been one of the main forms of entertainment for nearly 7000 years. Presently, the board game industry is estimated to have a net worth of more than $10 billion worldwide. Moreover, the uses of board games aren't limited to business, research shows that there is a plethora of uses for board games. For example, multiple medical articles point to the positive effects board games can have on people. In addition, research shows that board games might be a useful aid to help with teaching and development of children.

Clearly, board games are a very important medium of entertainment. As such, I have decided to study a sample of board games to try creating a linear model with average rating as the response. These ratings are based on the website "boardgamegeeks.com", which is the online largest hub for board games. The average rating of a board game can be generally interpreted as how liked a board game is by people. This metric can be very useful as a higher rating could lead to better sales or a better brand image and being able to predict this value could have a multitude of uses.

Methods

There are a few steps to take before I begin creating the model. First, I need to adjust the dataset to only contain the data I need. This includes removing all the identifiers and adding the "Number of Mechanics" column, which is the number of mechanics a board game has. Then the data must be divided into a training and test dataset. The rest of the process is performed on the training dataset. Also note that there are no missing observations in the data.

To create a linear model, we must satisfy its assumptions first, which can be done using the residual plots. Since I have multiple predictors, I also need to satisfy the two conditions of the mean response being a function of the predictors (condition 1) and the predictors being linear functions of each other (condition 2). After attempting to satisfy the conditions by removing predictors if needed, I can create a qq-plot as well as residual plots of the fitted rating and all the other predictors.

If any of these residual plots show patterns/irregularities, then some of the predictors/response might require transformations. The transformations are found using the BoxCox method and then applied to the data. Next, all the plots (for both conditions and the residual plots) must be rechecked for any patterns. If nothing is wrong, I can move on, otherwise the model might require more adjustment.

If all the plots are adequate, then the first linear model can be made. After fitting this model, we can study its slopes and intercept. We can then choose the predictors that seem not to be

significant and we could conduct an F-test to check if we can remove them. After, we have performed enough tests and are satisfied, we have the first possible model.

Next, we can create a few more models using some automated selection procedures on the full transformed model. This involves using the All-Possible Subsets Selection Process as well as the Forward/Backward/Stepwise Selection Processes (with both AIC and BIC selection). Then we can choose a few promising models as possible candidates and check all their plots to make sure they satisfy all conditions and assumptions

Then we compare the first model and all the other chosen models with each other by comparing there AICs, BICs, adjusted R^2 values, their multicollinearity, and their influential points to see their effects on the model. These values allow us to eliminate some models if they seem insufficient.

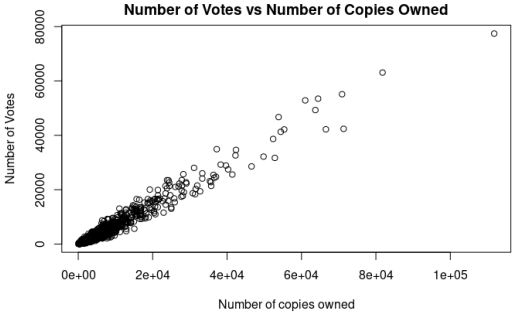
Finally, the few best models are selected and compared by validating them using the test dataset. This validation tries to see if the test models have relatively similar features and plots compared to the training models.

After doing these steps the findings can be compared and we can choose a model that fits the data well, while also being useful and practical.

Results

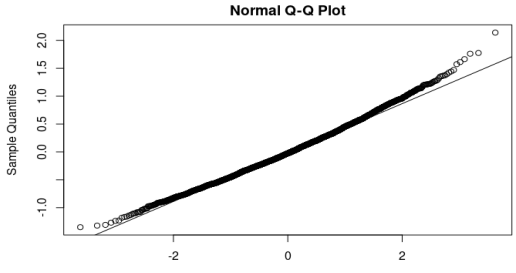
As discussed, the data was first modified then I split it into a 7:3 ratio. I choose this ratio as the sample has a lot of observations, so I felt that 30% of the data was enough for validation. The data is left with 11 variables; Minimum Time Needed, Maximum Time Needed, Minimum Players Needed, Maximum Players Needed, Year Made, Recommended Age, Average Rating, Weight, Copies Owned, Number of Votes, and Number of Mechanics.

Then I conducted a short EDA of the data where, most of the histograms seemed skewed, but the histogram for Average Rating was approximately normal.



Then I checked the conditions required to use the residual plots. Condition 1 seemed okay, however, the number of votes and number of copies owned seemed to be non-linear functions of each other. To solve this problem, I removed number of votes as a predictor, since I felt like the number of copies owned is a more general and useful metric.

After the condition are satisfied, I plot all the residual plots, as well as the qq-plot. Of these, the residual plot for minimum number of players, number of copies owned seemed to show some small irregularities. Meanwhile, the qq-plot also had a large violation.



QQ-Plot of full model before transformation

To fix these problems, I used the BoxCox function to transform the variable. The result was using the Inverse of the Average Rating, Square Root the Minimum Number of

Players, and Logarithm the Number of copies owned. This resulted in residual plots that satisfied model assumptions. (The new qq-plot is shown in the appendix)

Next, I conducted an F-test to try and remove the predictors Year Made, Max Players, Min Time, and Age Recommendation. The F-test had a P-value of 0.17, so I decided to remove the predictors. This gives us the first model with 5 predictors, which are the Weight, Number of Mechanics, Root of the Minimum Players, Logarithm of the Copies, and Maximum Time Needed.

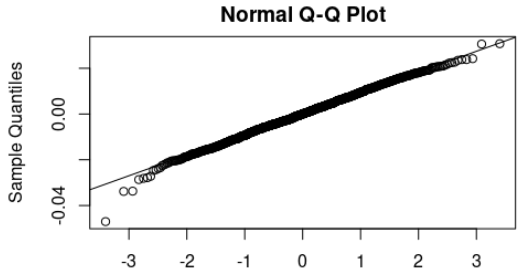
Then I used the automated selection methods to derive two more models. These models are ones with 4 predictors (derived using BIC selection) and 6 predictors (using AIC selection). The BIC model's predictors are the same as those of the 5-predictor model with Max Time Needed removed. The AIC model has the same 5 predictors in addition to Year Made. To make sure the models were adequate I rechecked the two conditions as well as the residual plots and everything seemed fine.

Model	Adjusted R^2	AIC	BIC	Influential Points
6 predictor model	0.3207	-22762	-22719	636
5 predictor model	0.3203	-22762	-22713	628
4 predictor model	0.3199	-22761	-22724	614

All the models had similar AIC, BIC, and adjusted R^2 values, and none suffered from multicollinearity. Moreover, the models had a similarly large number of influential points.

Model	6 predictors (AIC model)		5 predictors (F-test model)		4 predictors (BIC model)	
Data	Training	Test	Training	Test	Training	Test
Intercept	0.16	0.16	0.16	0.16	0.16	0.16
Weight	-6.9×10^{-3}	-6.6×10^{-3}	-6.9×10^{-3}	-6.6×10^{-3}	-7×10^{-3}	-6.9×10^{-3}
Number of Mechanics	-5×10^{-4}	-6×10^{-4}	-5.1×10^{-4}	-6×10^{-4}	-5.1×10^{-4}	-5.8×10^{-4}
Root of Min Players	5.7×10^{-3}	7.3×10^{-3}	5.6×10^{-3}	7.3×10^{-3}	5.6×10^{-3}	7.2×10^{-3}
Log of Copies Owned	-5.7×10^{-4}	-5.1×10^{-4}	-5.7×10^{-4}	-5.1×10^{-4}	-5.6×10^{-4}	-4.8×10^{-4}
Max Time Needed	-5×10^{-7}	-2.5×10^{-6}	-5×10^{-7}	-2.5×10^{-6}		
Year Made	-1.8×10^{-6}	-5.5×10^{-8}				
Adjusted R^2	0.3207	0.3328	0.3203	0.3332	0.3199	0.3287

Finally, I attempted to validate the model using the test data. All the plots pointed to the conditions and assumptions being followed except for the qq-plot of the 4-predictor model (BIC model), that had some deviation. Next, I compared the actual models, and looking at the results we can see a few values that are too far apart from each other, which are highlighted in yellow above.



QQ-plot of 4 predictor model with test data

Of these models, I discarded the 4-predictor model as a violation of normality can have a lot of impact on the usefulness of the model. I then choose to use the 5-predictor model as both the remaining models were very similar, but the 5-predictor model was simpler. The final model is,

$$\begin{aligned}(\text{Average Rate})^{-1} &= 0.16 - 6.9 \text{ x } 10^{-3}(\text{Weight}) - 5.1 \text{ x } 10^{-4}(\text{Number of Mechanics}) \\ &+ (5.6 \text{ x } 10^{-3})(\text{Min Players Needed})^{0.5} \\ &- (5.7 \text{ x } 10^{-4}) \log(\text{Number of Copies Owned}) - 5 \text{ x } 10^{-7}(\text{Max Time Needed})\end{aligned}$$

Discussion

To start I will interpret the intercept, the slopes, and R^2 value.

- If all the predictors have a value of 0, then $(\text{Average Rate})^{-1}$ will be 0.16.
- If Weight increases by 1 unit, then $(\text{Average Rate})^{-1}$ will decrease by $6.9 \text{ x } 10^{-3}$, given all other predictors are constant.
- If the Number of Mechanics increases by 1, then $(\text{Average Rate})^{-1}$ will decrease by $5.1 \text{ x } 10^{-4}$, given all other predictors are constant.
- If the Root of the Min Players Needed increases by 1 unit, then $(\text{Average Rate})^{-1}$ will increase by $5.6 \text{ x } 10^{-3}$, given all other predictors are constant.
- If $\log(\text{Number of Copies Owned})$ increases by 1 unit, then $(\text{Average Rate})^{-1}$ will decrease by $5.7 \text{ x } 10^{-4}$, given all other predictors are constant.
- If the Max Time Needed increases by 1 unit, then $(\text{Average Rate})^{-1}$ will decrease by $5 \text{ x } 10^{-7}$, given all other predictors are constant.
- 0.3218 of the total variation in $(\text{Average Rate})^{-1}$ can be explained by the model.

The model found answers the research question as it follows all the linear model assumptions, which lets us use this model safely to predict/approximate a board games average rating given its features. As mentioned, this metric is important as it help judge how favorable a game is to the public. This can be used in many areas, like trying to boost sales (as higher rated board games look more appealing), trying to find what the public prefers to see in a board game, etc.

However, I would advise against using this model blindly as this model has a few flaws as shown in the results section. Of these, I think the R^2 value of this model is very low, which means we don't explain much of the response using this model. There could be a lot of reasons for this and one such reason is the presence of confounding variables. For example, a board games advertisement budget can have a large impact on its ratings. These confounding variables are generally not as readily available as the rest of the data.

In conclusion, I think the model is adequate, however, further research and data collection needs to be done to improve the model, so it can give better results and be used safely.

Sources

- Vatvani, Dinesh. “An Analysis of Board Games.” A Python and Data Analysis Blog, Dinesh Vatvani, 5 Mar. 2018, https://dvatvani.github.io/BGG-Analysis-Part-1.html

- Beardsley, Sam. “What Makes a Board Game Good: Exploratory Data Analysis of Games on BoardGameGeek.com.” LinkedIn, 23 Aug. 2020, https://www.linkedin.com/pulse/what-makesboard-game-good-exploratory-data-analysis-games-beardsley

- Nakao, M. Special series on “effects of board games on health education and promotion” board games as a promising tool for health promotion: a review of recent literature. BioPsychoSocial Med 13, 5 (2019). https://doi.org/10.1186/s13030-019-0146-3

- Siegler, R. S., & Ramani, G. B. (2009). Playing linear number board games—but not circular ones—improves low-income preschoolers’ numerical understanding. Journal of Educational Psychology, 101(3), 545-560. http://dx.doi.org/10.1037/a0014239

- Ramani, & Siegler, R. S. (2008). Promoting Broad and Stable Improvements in Low-Income Childrens Numerical Knowledge Through Playing Number Board Games. Child Development, 79(2), 375–394. https://doi.org/10.1111/j.1467-8624.2007.01131.x

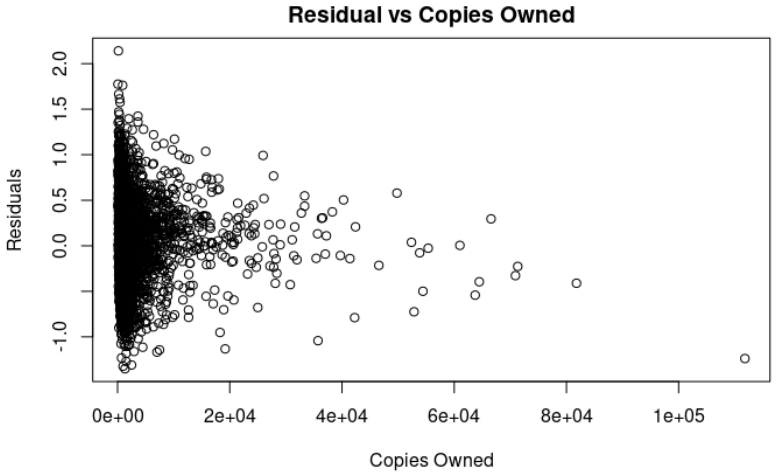
- Attia, P. (2016, January 21). The Full History of Board Games. Medium. Retrieved December 17, 2021, from https://medium.com/@peterattia/the-full-history-of-board-games-5e622811ce89

- Seetharaman, S. (2021, May 20). “Digital fatigue” is fuelling board game sales among adults - report. Dicebreaker. Retrieved December 17, 2021, from https://www.dicebreaker.com/categories/board-game/news/board-game-sales-2020-digital-fatigue

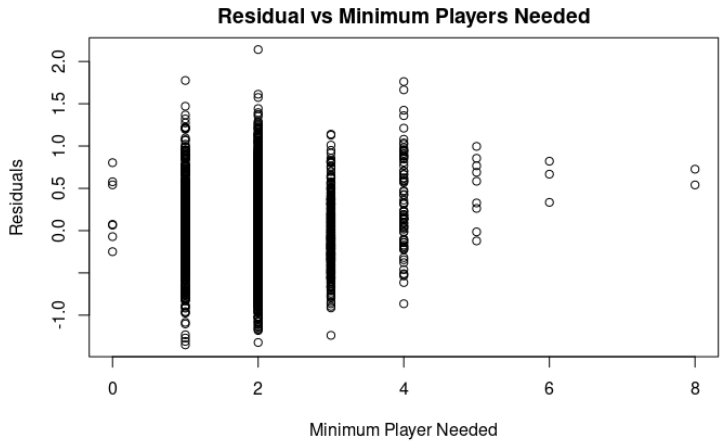
Dataset

- Mrpantherson, 2017-04-07, Board Game Data, bgg_db_1806.csv, Retrieved 18th October 2021 from https://www.kaggle.com/mrpantherson/board-game-data?select=bgg_db_1806.csv

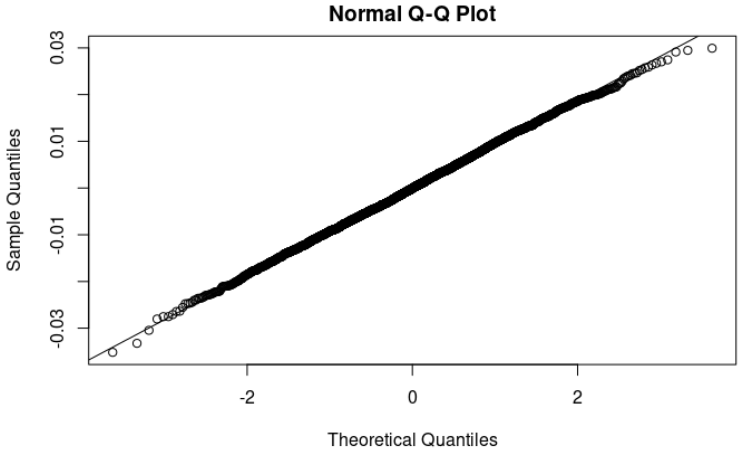
Appendix



Residual Plot for Copies Owned. Nonconstant Variance seems to be present.



Residual Plot for Minimum Players Needed. Nonconstant Variance seems to be present.



QQ-Plot of full model after transformation (training data)