**Abstract**

The surge in popularity of video games of all types has made this form of business a highly lucrative one – so long as the opinion of the game is strong enough to encourage sales. Using data from the Steam digital gaming platform, I evaluate whether certain general statistics – such as gameplay, ownership totals, and the like – can predict the likely public response of the game in total positive and negative reviews.

**Introduction**

Over the course of my lifetime, I have watched serious video game playing evolve from a “geeky” pursuit to a mainstream competitive event as culture and society has embraced the products of the information and data age. The variety of available themes, and styles of games seems to grow exponentially every year as new ideas and concepts develop. Even those that are avidly not gamers – such as my wife – often find something to capture their interest, from puzzles to simulations to games that do all the work for them. With such an expansive variety of options, the opinions about such games – both positive and negative – are even more so.

By using data provided by the SteamSpy API to acquire true values for games managed by Valve Corporation’s digital gaming service of Steam to learn how specific statistical factors – namely the overall median play time of the game by all owners, publisher, the standard cost of the game, current discounts, and approximate number of owners – appear to affect the games’ total positive and negative reviews, this project has come to the conclusion that while specific values in each of these factors have a great deal of influence for the positive scores, the negative scores are far less predictable.

**Methods**

The data for this project was pulled from the SteamSpy API on seven different days through the use of Python with the urllib and requests modules. The first day pulled was for all data collected by the Steam platform for July 2nd of 2020, which was subsequently used as the training data. From this data, ten linear multiple regression models were performed through the use of the lm function in R, five each for predicting the total positive and total negative reviews for each game. Additionally, five Poisson regression models were also used through the glm function in R – four for the positive review totals and one for the negative review totals – for additional analysis to help minimize bias; while Poisson regression models corresponding to all of the linear regression models were attempted, the remaining five combinations all resulted in models that failed to converge and thus failed to produce any predictions.

The five predictors used in these models were the overall median play time of the game by all owners, the game’s publisher, the standard price of the game, any discounts applied to the price of the game at the time that the data was pulled, and number of owners of the game as broken up into specific range groups. One of the primary individual statistics for every game on the Steam platform which is listed for every owner is the total amount of time that individual has spent playing that game – consequently, the median play time is for that value across all owners of the game from March 2009 to the present as a value in minutes, as indicated by the SteamSpy website.

While publisher is rather self-explanatory, more than 19000 publishers are represented in the data, which is beyond the scope of the processing ability available for this project; therefore, all publishers with less than 20 games represented were relabeled as “Other” in order to accommodate this issue. This was implemented through the addition of a publisher\_counts column which matched each factor found in the publisher variable to the total number of times that a game in the data showed that publisher, after which all publisher values which had a total value in publisher\_counts that was less than 20 were changed to “Other” in the data.

The standard price of the game is again self-explanatory, while the discount, which is listed as the percentage value (50 is 50%, 20 is 20%, etc.), will be used to see if the types of reviews received are correlated with the amount of discount showing on the game. Finally, the number of owners of the game per the data pulled from the SteamSpy API places the exact totals into the following ranges both to account for margin of error and for privacy purposes as dictated by the Steam platform’s regulations: 0 to 20,000 owners; 20,000 to 50,000 owners; 50,000 to 100,000; 100,000 to 200,000; 200,000 to 500,000; 500,000 to 1,000,000; 1,000,000 to 2,000,000; 2,000,000 to 5,000,000; 5,000,000 to 10,000,000; 10,000,000 to 20,000,000; 20,000,000 to 50,000,000; 50,000,000 to 100,000,000; and 100,000,000 to 200,000,000.

After acquiring the data, Python was used to clean and prepare it for modeling. This preparation consisted of altering the publisher predictor as mentioned above as well as dropping all instances where there were any missing values in the data – this project relied heavily on specific details from each game, and since these details varied widely between games and those incidents with missing values tended to be situations where insufficient data was available for the game, there was not sufficient support any application of dummy variables or imputing values that could be done without significantly skewing the results. The final step for the preparation was to remove the extra columns acquired with the data through the API pull, namely the appid (unique identification number supplied through Steam to each game which was duplicated as the dataframe index by the data acquiring process), userscore (supposedly a score rank of the game based on user reviews, but the details of how the score was determined was unclear, so the variable was dropped in favor of the positive and negative review totals), developer (typically included the publisher and often a collection of multiple different publishers, which would confuse the results), average\_forever (the mean value of the total play time as opposed to the median value – the median value was chosen as it is more robust against the influence of outliers), average\_2weeks and median\_2weeks (indicated values in minutes for the data of the last 2 weeks for each game, and dropped due to the fact that this project aimed to take a much wider view of time), and price (current price of the game after the removal of the discount, therefore basically duplicate data). The final result of the data for each of the seven dates consisted of the columns for the index (as mentioned previously, the same value as the game’s identification number in the Steam platform), the game’s name, the game’s publisher, the total number of positive reviews, the total number of negative reviews, the total number of owners as a range, the median play time in minutes for all owners since March of 2009, the base price of the game, the current discount of the game, and the count of the total number of games for each publisher as previously mentioned for the creation of the “Other” category.

The prepared data for July 2nd, 2020, was then imported into R for the training process of the models. As mentioned previously, five linear multiple regression models for each of the positive and negative review totals were implemented, consisting of one model including each of the predictor terms individually with no analysis of interactions between any of them; one including the comparison of the interaction between the publishers and the number of owners; one comparing the interaction between the base price and the discount; one comparing the interaction between the number of owners and the median play time; and one comparing the interactions between the publishers, the number of owners, and the median play time, in each possible pair and all together.

Since the data does resemble a Poisson distribution to some degree, multiple Poisson regression models were also attempted using the same parameters as listed above; however, due to most of the various models failing to converge, only five of the models produced viable formulae for use as a predictive model: the separate combinations of all predictors individually, the interaction between the publishers and the number of owners, the interaction between the base price and the discount, and the interaction between the number of owners and the median play time for the number of positive reviews; and the interaction between the number of owners and the median play time for the number of negative reviews.

From these results for the initial training models, the summaries were reviewed for the resulting coefficient values and their significance. For each of the variables that contained factors, the following were used as the baseline values: 1C Entertainment for the “publisher” variable and 0 to 20,000 owners for the “owners” variable. The following figures are portions of plots which display the coefficients for each of the positive review linear regression models through by using a central point for the estimated value and lines extending out to the sides for its associated standard error confidence interval – the order of these plots follows the same order in which they were introduced in the earlier paragraph. In each plot portion, the light gray dotted line which passes through nearly all of the points is for the coefficient value of 0.A picture containing sitting, water, room, people

Description automatically generated

A screenshot of a cell phone

Description automatically generatedA close up of a logo

Description automatically generatedA screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

As can be seen from these plots, in most cases there is little deviation for the majority of coefficients in each model comparatively, but it should be noted that some interaction variables in both the positive reviews and negative reviews models measuring the interactions between the publishers, the number of owners, and the median play time have very large standard error confidence intervals. When comparing the various statistical results for each of the models, the statistical significance of many individual variables and factors is not very strong – only the ownership groups containing 100,000 or more owners and the initial price show consistent significance at 95% or higher confidence intervals for all models, and certain publishers are statistically significant in either the positive reviews or negative reviews models, but the only publisher which shows strong statistical significance in both groups of models in Valve (which, ironically, is the company that also maintains the Steam gaming platform).

All models, however, show very strong values for the goodness of fit, which may indicate overfitting due to the high number of degrees of freedom in each model and the extremely low p-values (all are noted with the result – p-value : < 2.2e-16). On the other hand, the adjusted *R2*values for all of the linear regression models do indicate a middling to strong fit for all models, especially the positive ones and as additional interactions are applied. The linear regression models that have the best fit for both the positive reviews and the negative reviews are the ones which test the interactions between the publishers, the number of owners, and the median play time, with the model for the positive reviews showing an adjusted *R2*value of 0.9705 and the model for the negative reviews showing an adjusted *R2*value of 0.9231.

In the analyses for the Poisson regression models the project found more difficulty in specifying their success. In all cases both the residual deviance and the Akaike’s information criterion (AIC) for the models were very high – over twenty million in all cases except one – despite the high degrees of freedom (over thirty thousand). The interesting exception to this fact was the model for the negative reviews based on the interaction between the number of owners and the median play time – in this case, these values were just over five million. Additionally, the differences between the AIC values were large enough that the difference between the positive review models in how probable they were to minimize information loss compared to one another is exceptionally minute. Otherwise, the coefficient plots for each of the Poisson regression models were very similar in the amount of standard error confidence interval to the linear regression models, but the variety in the actual values of the coefficients varied widely in comparison.

At this point, the project applied the fifteen different models to the test data, which consists of the SteamSpy data for the before mentioned days between July 13th and July 18th, 2020, inclusive, after they had been cleaned and prepared using the same methods as the training data. However, in the intervening week and a half, some of the publishers added additional games to the Steam platform, resulting in some factors that appeared in the testing data which did not appear in the training data (since they were originally included in the “Other” value) – in order to maintain the consistency supplied by the models, these publishers were added to the “Other” grouping despite having twenty or more games available in the Steam platform.

**Results**

After applying the fifteen different models to each of the various sets of testing data, the following results were observed in the analyses – while the various test models suggested a strong goodness of fit for the data, the resulting predictions produced very high residual scales, indicating large residual standard deviations. This is further supported by the following plots demonstrating the relationship of the standardized residuals to the theoretical quantiles of the distribution.

A picture containing photo, sitting, table, different

Description automatically generatedA picture containing photo, sitting, table, bird

Description automatically generatedA picture containing sitting, photo, table, field

Description automatically generatedA picture containing sitting, photo, room

Description automatically generatedA close up of a person

Description automatically generatedA picture containing photo, bird, sitting, flying

Description automatically generatedA picture containing photo, bird, sitting, table

Description automatically generatedA close up of a logo

Description automatically generatedA close up of a logo

Description automatically generatedA close up of a logo

Description automatically generatedA picture containing photo, bird, sitting, table

Description automatically generatedA picture containing bird, flying, air, person

Description automatically generated

Linear Regression – Negative Reviews – Publishers, Owners, & Median Play Time

Poisson Regression – Positive Reviews – No Interactions

Poisson Regression – Positive Reviews – Price & Discount

Poisson Regression – Positive Reviews – Owners & Median Play Time

Poisson Regression – Negative Reviews – Owners & Median Play Time

Poisson Regression – Positive Reviews – Publishers & Owners

Linear Regression – Positive Reviews – Owners & Median Play Time

Linear Regression – Negative Reviews – Owners & Median Play Time

Linear Regression – Positive Reviews – Publishers, Owners, & Median Play Time

Linear Regression – Negative Reviews – Publishers & Owners

Linear Regression – Positive Reviews – Price & Discount

Linear Regression – Negative Reviews – Price & Discount

A picture containing sitting, bird, flying, air

Description automatically generatedA picture containing photo, sitting

Description automatically generatedA picture containing sitting, bird, flying, standing

Description automatically generated

Linear Regression – Positive Reviews – No Interactions

Linear Regression – Positive Reviews – Publishers & Owners

Linear Regression – Negative Reviews – No Interactions

**Conclusion**

Note that in each of the reviews of the standardized residuals the tails drift away from the theoretical line, which indicates that the models are not the best fit despite what the initial model statistics indicated. In particular, all of these indicators contain extreme outliers which present additional questions. These results call for additional review of the data and consideration of additional models. Subsequent reviews will be considered hereafter.

**Acknowledgements**

The project’s analyst would like to acknowledge with gratitude Sergey Galyonkin for his creation and continual maintenance of the SteamSpy website as a generalized source of statistical data for the Steam digital gaming platform, Prof. Amirfarrokh Iranitalab for his guidance and encouragement in this project, and his wife for her support through the stresses and anxieties experienced while this project was performed