An Analysis on what Affects North American Video Game Sales

By Linh Hoang

1006961343

Introduction

Video games have become accessible across multiple formats. Far past the days where these games were only accessible in arcades, consumers now have options on where they want to play their games. These range from consoles designed specifically for gaming, such as the Nintendo Switch or PS5, to common devices like a computer or mobile phone. Along with general mainstream popularity, it comes to no surprise that the video industry itself is large. Revenue for the industry is expected to be 196.8 billion USD by the end of 2022 with more market growth expected in the future (Newzoo, 2022).

Given this large and growing market, understanding factors can impact sales of a game is a point of interest. Consumers can understand why certain trends appear in the types of video games that are sold. Meanwhile, video game companies can use this knowledge in determining if a potential game is likely to turn a profit.

In *The effect of intrinsic and extrinsic quality cues of digital video games on sales: An empirical investigation* by H.S. Choi and Co., research was specifically aimed at how different factors change the digital sales of a video game. The research finds multiple factors, such as developer reputation and price, that do impact the digital sales. The paper, however, does not investigate the physical copy sales and if the same factors would also impact them similarly. Another paper looked at, *What Makes a Blockbuster Video Game? An Empirical Analysis of US Sales Data* by Joe Cox also investigates factors that may relate to the sales of video games. While this paper also finds factors that can change the sales performance, Cox looks only at the market in the United States, rather than the region of North America as a whole. Other articles are like the ones mentioned above, with the analyses being done on either a specific type of sales, sales not in North America as a whole or both. Our report aims to add ore to the current research that's been done of video game sales by looking at all types of sales for a video game in the North American region.

The purpose of this report is to analyze and understand what factors affect video game sales in North America. The genre; average review rating scores; platform; earliest publishing year; publisher; developer; and sales in other regions are hypothesized as factors that majorly impact sales in North America.

Methods

The data used for this analysis was sourced via Kaggle from user Rush Kurubi and provides information on a large number of video games released up to December 22nd 2016. It must be noted that only games with at least 100 000 global sales have been included. The variables; Genre; Platform; Year_of_Release; Publisher; Developer; EU_Sales; JP_Sales; Other_Sales; Critic_Score; and User_Score were the initial predictors.

The data had been cleaned. Incomplete cases were removed to ensure that all data points remaining can used for the entire analysis, rather than some being present in one part of the analysis and not present for another part. Any 0 values in numeric predictors have 1e-5 added to its original value so any needed box cox transformations can be applied. Some variables, such as

User_Score, were converted from the character data type to the numeric data type based on context of what the variable represented. Lastly, under variable Platform, any values associated with “2600” was changed to “Atari 2600” to combat any potential confusion on what the original value meant.

The data was then divided in half, with one half being the training data and the other half being the testing data. Both sets were checked to ensure they were similar in distribution for each variable. The exploratory data analysis was then conducted on the training data. Left skews were noted to appear in variable histograms involving sales and right skews were noted to appear in variable histograms involving ratings. It was also shown from the bar graphs that certain values in the categorical variables occur more often, such as ‘Action’ in the Genre. The scatterplots were observed for any linear relations.

A linear model was fitted using all the predictors mentioned. The residuals plots were then generated to check to see if the two additional conditions for multiple linear regression have been met. Afterwards, the linear model assumptions were then checked by comparing residuals to the actual values between each numeric predictor and comparing the data quantiles to the theoretical quantiles. Since the variance and normality assumptions were not met, Box-Cox had been used to transform the data and model. Conditions and assumptions were checked once more, with any violations noted.

The ANOVA-F test was then run on the model to check for a significant linear relationship. After confirming there was a significant linear relationship, the T-test was run to find any insignificant predictors.

A new model was built without the insignificant predictors included. The new model was checked for condition and assumption violations. A partial F test was conducted comparing the new model to the old model, and the new model was deemed better to keep.

95% confidence intervals for the predictors were be found. New models were built with various combinations of the predictors having 0 within its confidence interval removed. Conditions and assumptions are checked once more, with the violations noted. A partial F test was conducted with the new models and the previous one. Based on the results, the new model with JP_Sales removed was kept..

Once the T-test the second partial F test had been conducted, the Variance Inflation Factor was then used to assess for any severe correlation between predictors. As there were not any, the current model was kept. The adjusted R^2; Akaike’s Information Criterion (AIC); Corrected AIC; and Bayesian Information Criterion methods were not used as a result of this.

Leverage, outlier and influential points were found using Cook’s distance, DFFITS and DFBETAS.

The model was then validated by fitting the model in the test dataset. The result was then compared to when the model was fitted in the training dataset.

Results

Based on the final model determined, the factors that impact the North American sales of a video game are the average user rating; the average critic rating; the year of release; the platform/console the game is playable in; the sales performance in Europe; and the sales performance in regions outside of North America, Japan and Europe.

The F-statistic of the final model is found to be 506, with a p-value of less than 2.2e-16, indicating a strong linear relation that is likely to be true between the response and final predictors.

Fitting the model into the test data and comparing the results, a similar outcome in the regression line and residuals can be seen. It can be concluded that the model is valid and not overfitted. Referring to the graphs below, there are many similarities in the resulting model and residuals between the training and testing data using the same model.

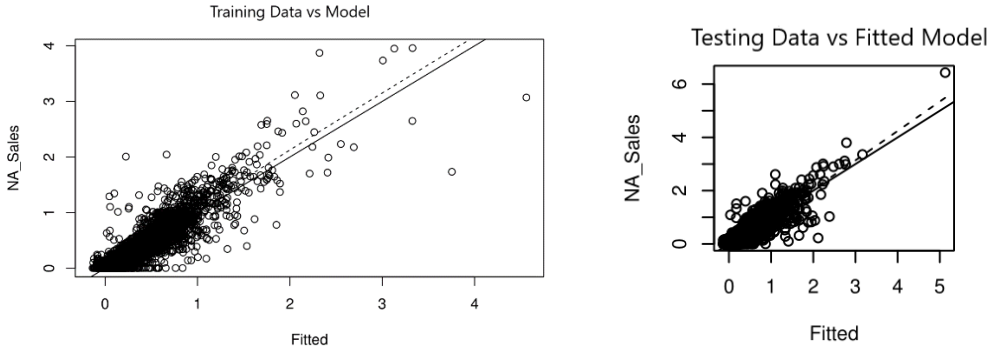

Figure 1: Graphs of the testing and training data compared to the model that was fitted. Notably, data points cluster similarly and resulting model share similar slopes

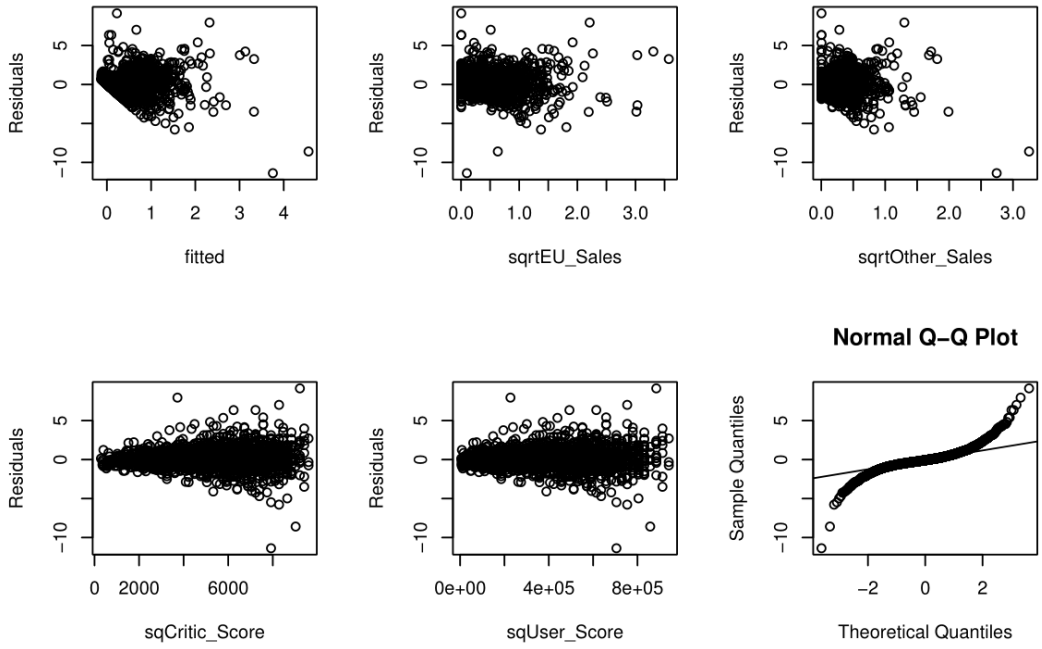

Figure 2: Residuals in comparison to actual values along with QQ plot for the training data and model. Note how normality is shown to be violated but other assumptions seem to hold

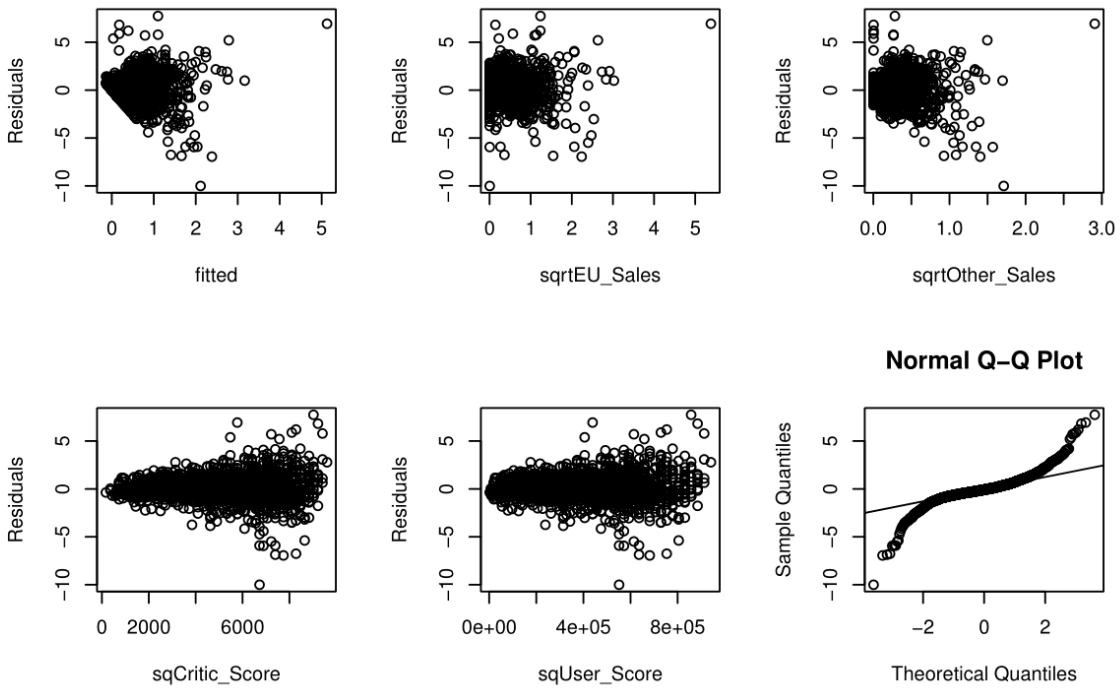

Figure 3: Residual plots and QQ plot for testing data and its model. Note the similarity in clustering and shape for between these plots and the plots in figure 2

Looking into details of the model, we see that the coefficients associated with each predictor is very small, with the biggest factor that increases sales performance of a game in North America being the performance of the game in other regions, with a coefficient of 1.293e+00. Meanwhile type of platform tends to negatively impact the sales performance, with over half the potential platforms providing a negative coefficient in the model equation. This negative impact could be due to how the associated platforms are older and have thus been replaced with newer version. Thus, releasing a game in an older platform could turn users away if they only have the newer platforms. The dataset itself is shown to have many outlier and leverage points, which may have strongly influenced the resulting model.

It is notable that the final model did not see developer and publisher as big factors in sales performance. Furthermore, while the other predictors for sales in multiple regions were seen as significant, sales performance in Japan was not.

Discussion

To summarize, the factors that were found to impact North American Sales were ratings, release year, sales performance in other regions and the platform the game is played in. Strong linear relations do exist, thus showing the sales of a game is based heavily on a linear equation.

The results of this analysis are like results shown in the research papers mentioned previously, namely the impact of publishers and developers towards sales.

It does need to be noted that the QQ plots generated throughout (see below) have been shown to have the Normality assumption of linear regression violated. Confidence intervals were used to cut insignificant predictors out and p-values have been used as evidence to draw conclusions in this analysis. Given their reliance on normality, the results from these methods could be bias and inaccurate. Further transformation or methods that do not rely on normality should be investigated to check for the validity of the conclusions drawn.

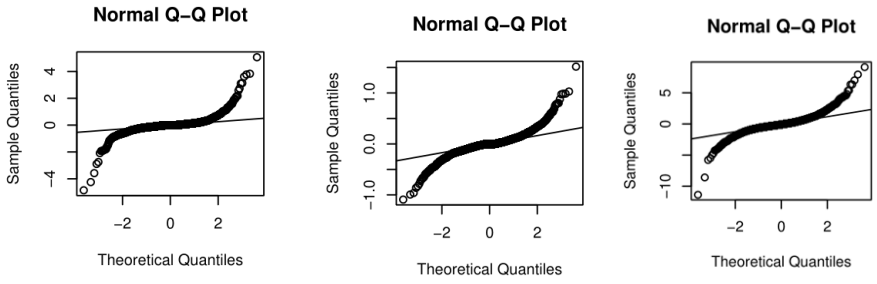

Figure 4: Various QQ plots generated throughout analysis between different models. From left to right the QQ plots are for: First model with all predictors, second model with transformed predictors and the final model

Works Cited

Choi, H. S., Ko, M. S., Medlin, D., & Chen, C. (2018). *The effect of intrinsic and extrinsic quality cues of digital video games on sales: An empirical investigation*. Decision Support Systems, 106, 86–96. https://doi.org/10.1016/j.dss.2017.12.005

Cox, J. (2014). *What Makes a Blockbuster Video Game? An Empirical Analysis of US Sales Data. Managerial and Decision Economics*, 35(3), 189–198. https://doi.org/10.1002/mde.2608

Newzoo. (2022, December 5). *Newzoo Global Games Market Report 2022: Free version*. Retrieved December 20, 2022, from https://newzoo.com/insights/trend-reports/newzoo-global-games-market-report-2022-free-version?utm_campaign=GGMR2022&utm_source=press