Association Between Video Game Sales in North America and Video Game Critic Score, User Score, Platform and Genre.

Arthur Wu

Introduction

Playing video games is a hobby that has grown exponentially in popularity and technology within the past few decades. The years 2005 and 2006 mark a major advancement in home game consoles with the release of the 7th generation consoles Microsoft Xbox 360, Nintendo Wii and Sony PS3. These consoles brought forth new technology into gaming including motion capture and online gameplay, as well as a large selection of new video games of various genres and gameplay[1].

In this analysis, I have data on video games varying in platform, genre and publisher that have been released since 1980. This data includes critic scores and user scores drawn from Metacritic, game details such as genre and video game platform, and game sales by region. I will use this sample of video game data to represent the total population of released video games, including games that will be released in the future. Due to the dominance and competitiveness of the Xbox 360, Wii and PS3 consoles in the United States, I want to see how different aspects of 7th generation video games impact video game sales in North America. I also would like to consider the associative impact of other variables on video game sales in North America as well. Some additional variables outside of video game platform that I am interested in exploring include video game genre, critic score and user score. The specific genres of video games that I would like to study are the action, racing, role-playing, shooter and sports genres. The main question that I ask in this study is: Are video game platform, genre, critic score and user score significant predictors of video game sales in North America? Using data on video game critic scores and video game platforms, I will develop a model that will use critic and user score, platform and genre to predict North American video game sales.

[1] https://www.did.ie/content/blog/history-of-video-game-consoles

Data

The dataset used in this study is a video game sales with ratings dataset that was compiled by Rush Kirubi and made available on his Kaggle profile[2]. The columns that this dataset includes are basic game descriptions including game platform, genre and publisher, global and regional sales statistics, and critic and user ratings. Kirubi obtained data on global and regional video game sales from Gregory Smith's Video Games Sales dataset, which was compiled through a web scraping of VGChartz, a popular website used to track video game sales [3]. Kirubi extended Smith's video game sales data by adding critic and user score data from a web scraping of Metacritic.com, which is a review and rating website that gives ratings to video games.

To make the video game sales and ratings dataset more suitable for my study, I performed data cleaning to remove missing critic scores. A majority of video games released before 2001 had missing critic scores in the dataset. This is because Metacritic did not begin rating video games until it was founded in 2001[4]. Therefore, I needed to remove all games that did not have critic scores to prevent skewed results. I also filtered the Platform column so that the dataset would only contain the platforms that are important to the study which are the PS3, Wii and X360 platforms. Furthermore, I filtered the Genre column of the dataset to only include action, racing, role-playing, shooter and sports genres, which are the five most populated genres for the three consoles that I am interested in observing. I removed all games that had a critic score under 30.0 since there were only 7 entries within that interval and could skew the results of this study. I also addressed the outliers in the NA sales column by removing all games with a NA sales value greater than 0.5. Finally, I converted the user score variable from its original string format to a numeric variable format and removed all video games that were missing user scores. This allows my dataset to be complete and correctly formatted for my calculations. Before cleaning, the dataset contained 16,719 data entries. After cleaning, the dataset contained 1,156 entries.

---

[2] https://www.kaggle.com/rush4ratio/video-game-sales-with-ratings
[3] https://www.kaggle.com/gregorut/videogamesales
[4] https://www.metacritic.com/about-metacritic

Variables

To analyze if critic score, user score, platform and genre are predictors of video game NA sales, I will use critic and user score observations, the platform column containing only PS3, Wii and Xbox 360 video games, and the genre column containing only action, racing, role-playing, shooter and sports genres as my predictor variables. I will also use the NA sales column as my outcome variable.

NA Sales

NA sales is the total number of sales for each video game in North America measured in millions of units sold. The source of the NA sales data comes from the VGchartz website, where each released video game is tracked for the number of sales by region.

Table 1: Summary Statistics for NA Sales, Critic Score and User Score

Variable	Frequency	Min	Max	Mean	Median	Std. Dev.	Std. Error
NA_Sales	1156	0.0	0.5	0.1903	0.16	0.129	0.003793
Critic_Score	1156	30.0	92.0	65.33	67.00	13.23	0.389
User_Score	1156	0.7	9.3	6.708	7.00	1.347	0.03961

Looking at table 1, we can see that the sample mean for video games sales in our data set is 0.19, which can be interpreted as 190,000 units sold. The standard deviation of the NA sales data is 0.129, which is relatively small. This tells us that the spread of NA sales data is relatively concentrated near the mean of 0.19. The standard error of NA sales in our data is 0.0038, which is very small and indicates that our sample data accurately represents the general population of NA video game sales.

Critic and User Score

The critic score and user score data used in this study were taken from the "Metascore" on Metacritic's website. To determine the Metascore of a video game, Metacritic takes the weighted average of all critic scores and creates a final numeric score on a 100-point scale[5]. User scores are also compiled by Metacritic in a similar fashion as critic ratings but is instead rated on a 10-point scale. Following Metacritic's video game score systems, we can assume that a game is more enjoyable and fun if the critic and user scores are high, and a game is less enjoyable and unrewarding if the critic and user scores are low.[6]

Looking at the summary statistics for the critic score variable shown above in table 1, we can see that the sample mean of critic score is 65.33 and the median is 67. This shows us that on average, critics rate games around 65 points on the 100 point scale in our sample data, which is 15 points above the halfway mark of 50 on the variable's 100-point scale. We can also see that the standard deviation is 13.23, which tells us that the spread of data is relatively concentrated around the average. The standard error is 0.39 which is small and indicates that our sample of critic scores represents the general population well.

Looking at the summary statistics for the user score variable from the same table, we can see that the sample mean of user score is 6.708 and the median is 7.0. This is proportionally similar to the critic score sample mean, as both means are around 20% higher than the halfway mark of each respective numerical range. We can also see that the standard deviation is 1.347, which tells us that the spread of data is relatively concentrated around the average and is also proportionally similar to the standard deviation of critic score. The standard error is 0.040 which is small and indicates that our sample of critic scores represents the general population well. Overall, we can see that the summary statistics of user score is very similar in magnitude when compared with the statistics of critic score.

Platform and Genre

The two categorical variables that are used in this study are video game platform and genre. The platforms that I am interested in are PS3, Wii and Xbox 360 (X360). Examining the univariate summaries of the three platforms featured in the dataset shown in table 2, we can see that PS3 and Xbox 360 have very similar frequencies and percentages. PS3 has a frequency of 479 and a percentage of 41%, and Xbox 360 has a frequency of 456 and a percentage of 39.4%. However, we can also see that the frequency for the Wii platform is much lower compared to the

---

[5] https://www.metacritic.com/about-metascores
[6] https://www.metacritic.com/about-metacritic

frequencies of Xbox 360 and PS3 With a frequency of 221 and a percentage of 19.1%, This tells us that the number of Wii games is effectively half of the number of PS3 and Xbox 360 games, resulting in a underrepresented Wii platform category in our sample data. Overall, this difference in frequency and percentage reflects that our sample data is not as balanced is size as we would ideally like.

Table 2: Univariate Summaries for Video Game Platform and Genre

Category	Frequency	Percentage
Platform		
PS3	479	41.4 %
Wii	221	19.1 %
X360	456	39.4 %
Genre		
Action	452	39.1 %
Racing	133	11.5 %
Role-Playing	129	11.2 %
Shooter	208	18 %
Sports	234	20.2 %

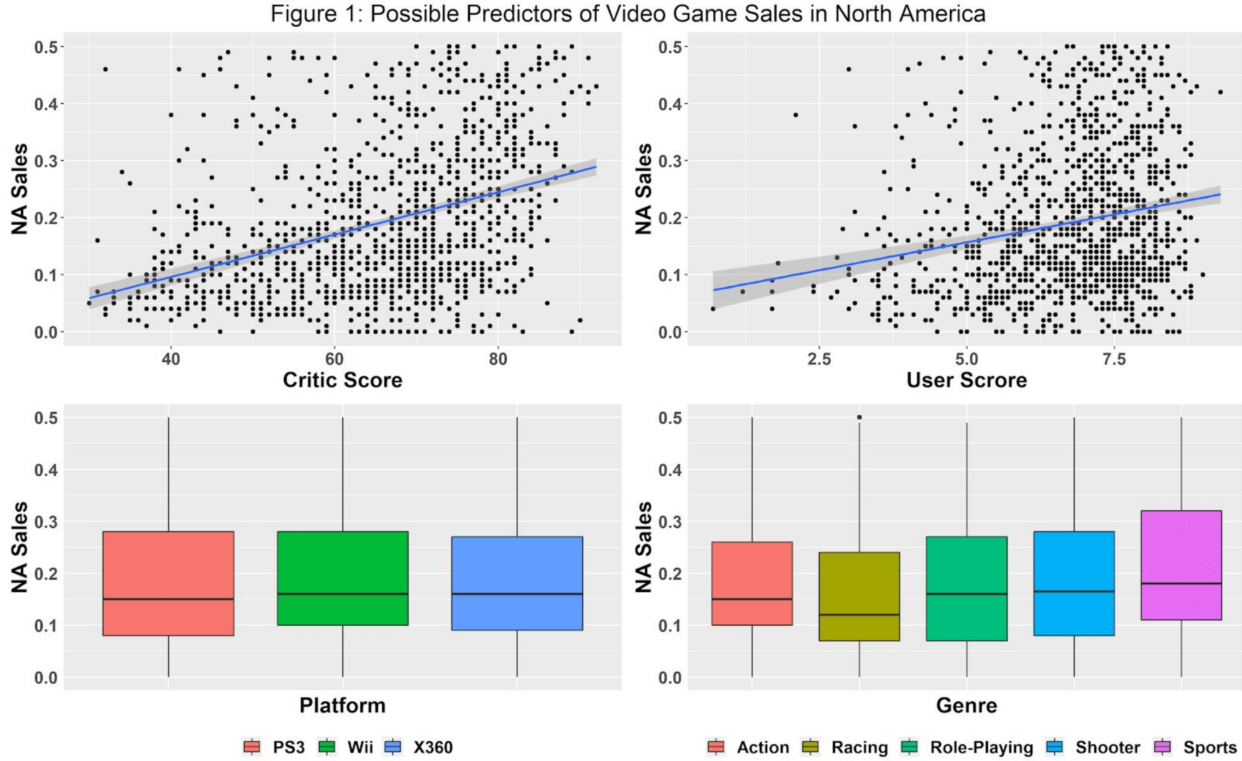
For the genre category, I specifically want to analyze the five most populated genres for PS3, Wii and Xbox 360 video games which include action, racing, role-playing, shooter and sports genres. Examining the summaries for genre in table 2, we can see that the frequencies and percentages of each genre is more varied across the categories compared to the platform category. Racing and role-playing genres have lower percentages of around 11%, shooter and sports genres have percentages around 20% and action genre has the highest percentage of 39.1%. The difference in frequency between the post populated genre (action) and least populated genre (role-playing) is 323 observations.

Analysis and Interpretation

To examine the relationship between platform, genre and critic score, I will use linear and multiple regression models. For my specific study, my independent predictor variables are critic score, user score, platform and genre. My dependent outcome variable will be NA sales. I will also use 0.05 as my alpha value.

Relationship Visualization

Figure 1: Possible Predictors of Video Game Sales in North America



Prior to performing linear and multiple regression, I used various plots to visualize the relationship between each predictor variable and our outcome variable NA sales as shown in figure 1. From our plots, we can see that both critic score and user score variables have a positive relationship with NA sales and both have weak correlations as shown from the upward angle of distribution and large amount of data spread from the line of best fit for both scatterplots. However, critic score seems to have a slightly higher correlation than user score as the plot points in the critic score plot seem to follow the angle of the reference line better than in the plot for user score. Overall, both critic and user score seem to have very weak, positive correlations with NA sales.

Looking at the platform and genre correlation plots, we can see that the NA sales distribution by platform are extremely similar in mean, range and interquartile range. This suggests that there is little to no significant difference in NA sales by platform. The genre correlation plot reveals slightly more variety in the NA sales distributions by genre. However, there does not seem to be any trends that would suggest correlation between NA sales and video game genre.

Linear Regression Model

To examine how our predictor variables influence NA sales of video games, I will first perform a simple linear regression model fitting using the critic score predictor variable to predict NA sales.

Interpreting the results of the simple linear regression model fitting displayed in model 1, we can see that the intercept coefficient of critic score and NA sales is -0.053. This tells us that when the critic score is 0, NA sales is predicted to be -0.053 million copies sold. This intercept does not give any valuable information for our model, since it is impossible for the number of NA video game sales to be negative. The independent variable (critic score) coefficient is 0.004, which tells us that the relationship between critic score and NA sales is a positive relationship due to the positive slope. Furthermore, it tells us that a 1 point increase in critic score is associated with a 0.004 (4,000 copies) increase in NA sales. The null hypothesis that we are testing is if our predictor (critic score) is not significantly different from 0. The p-value of the lstat coefficient is less than 0.001, which is less than our alpha value of 0.05. This allows us to reject null which means that the coefficient for critic score is significantly different from 0. The r-squared value of our model is 0.14, which tells us that critic score accounts for 14% of the variance in NA sales. The p-value of the F-statistic is less than 0.001, which is smaller than our alpha value of 0.05. This allows us to reject null and tells us that our model is an acceptable model for predicting NA sales and that our r-squared value is significantly different from 0. From these results, the complete linear model formula for the relationship between critic score and NA sales is:

Model 1: Linear Model between Critic Score and NA Sales

Predictor	Estimate	Std. Error	t-statistic	p-value
Intercept	-0.053	0.018	-2.97	0.003
Critic_Score	0.004	0.000	14.01	< 0.001

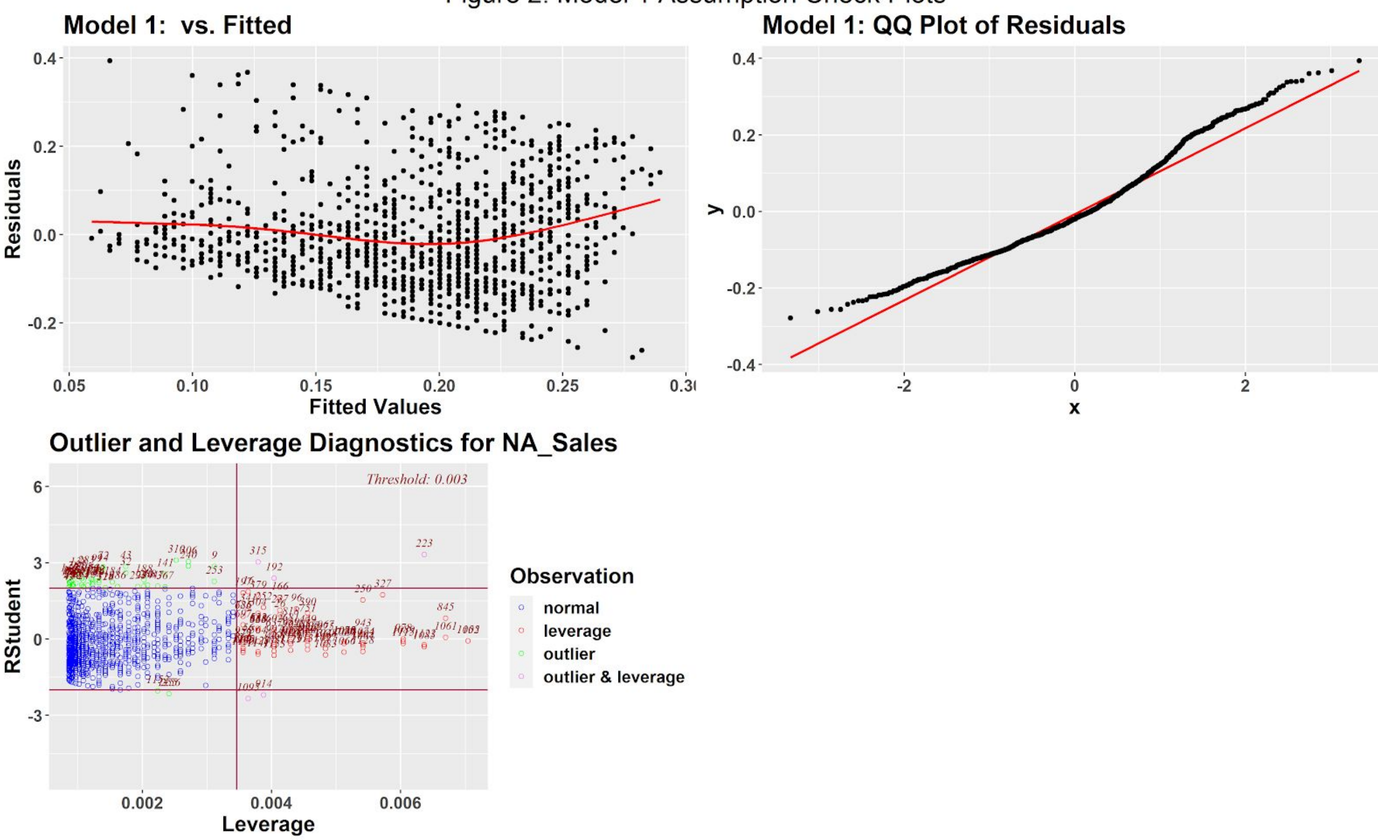
Note:
Reference level is Critic Score
n = 1156. r-squared = 0.14, F(1,1154) = 196.35.
F-Statistic p-value: < 0.001

$$NA\ Sales\ =\ 0.004(Critic\ Score)\ -\ 0.053$$

Checking Assumptions

After developing our simple linear model, we must check if our newly fitted model violates any assumptions that simple linear regression models require. We will check these assumptions by performing calculations and analyzing summary graphs shown in Figure 2. The first assumption that we need to check is the assumption of normally distributed residuals. Examining the model 1 QQ plot of residuals, we can see that the distribution of residuals seems

Figure 2: Model 1 Assumption Check Plots



to closely follow the reference line with slight upward deviation at the tails of the graph. Although there is deviation, the deviation is minor enough to claim that the residuals distribution is normal. The second assumption to check is for the model to have no influential outliers. Model 1: Residuals vs Leverage shows that there are many plots that have leverage and many plots that are outliers, but there are approximately 5 data points that are considered both outliers and have leverage, which suggest the presence of influential outliers. Therefore, we violate the assumption that the model is free of influential outliers. Next, we will check the assumption of homoscedasticity, or the constant variance in errors. To test this assumption, we will look at Model 1: Residuals vs Fitted. From the plot, we can see that the data points seem to maintain a consistent distance as the fitted values increase, but there is also a slope to the graph which may suggest inconsistent variance. To further analyze homoscedasticity, I performed the Breusch Pagan Test for Heteroscedasticity and obtained a p-value of 2.89e-5, which is less than our alpha value. However, the Breusch Pagan Test for Heteroscedasticity is very sensitive and could be misleading. Therefore, it is difficult to determine if our model violates homoscedasticity or not. For now, I will assume the worst and reject the null hypothesis that our variance is constant, which means we have heteroscedasticity and ultimately violate the assumption of homoscedasticity. Lastly, we will evaluate the assumption of independent errors. Examining Model 1: Residuals vs Fitted, we can see that the curvature may suggest a curvilinear relationship between residuals and fitted values.With a curvilinear shape, our model violates the assumptions of independent errors.

Overall, we can see that our linear model between NA sales and critic score suffers from many potential limitations due to violating multiple assumptions of linear regression. By violating the assumption of no influential outliers, our model may have bias estimates and standard error bias. Violating the assumption of homoscedasticity may lead to less precise estimates and biased significance tests and violating the assumption that errors are independent may lead to biased standard errors and poor confidence intervals and hypothesis tests.

Multiple Linear Regression Model

In an attempt to improve upon my first model, I will develop a full model with all predictors (critic score, user score, platform and genre) and perform a multiple linear regression. For this new model, the reference groups for the platform and genre categorical variables default to the PS3 platform and the Action genre.

From the results of the model shown in Model 2, we can see that the intercept coefficient is -0.057, which is the mean of NA sales when all of the numerical variables are 0 and all the categorical variables are at their reference level. However, this intercept is not meaningful because NA video game sales cannot be a negative value. The coefficient for critic score is

0.004, which tells us that critic score has n positive relationship with NA sales as shown by the positive sign. It also tells us that as critic score of a video game increases by 1 point, NA sales increases by 0.004 (4,000 copies sold). The coefficient for user score is -0.003, which tells us that user score has n negative relationship with NA sales as shown by the negative sign. It also tells us that as the user score of a video game increases by 1 point, NA sales decreases by 0.003 (3,000 copies sold). The coefficient for the Wii platform is 0.028, which tells us that compared to the PS3 platform, the Wii platform on average has 28,000 higher NA sales holding all else constant. For the Xbox 360 platform, the coefficient is 0.009, which tells us that on average, the Xbox 360 has 9,000 higher NA sales holding all else constant. The coefficient for the racing genre is -0.040, which tells us that compared to the action genre, the racing genre has on average 40,000 fewer NA sales holding all else constant. For the role playing genre, the -0.019 coefficient tells us that on average, the role playing genre has 19,000 fewer NA sales when compared to the action genre holding all else constant. For the shooter genre, the 0.005 coefficient tells us that on average, the shooter genre has 5,000 more NA sales when compared to the action genre holding all else constant. Lastly, the 0.009 coefficient for the sports tells us that on average, the sports genre has 9,000 more NA sales when compared to the action genre holding all else constant. Out of the many reference groups, there are only four reference groups that have p-values less than our alpha value which suggests that they are

Model 2: Characteristics Associated with NA Video Game Sales

Predictor	Estimate	Std. Error	t-statistic	p-value
Intercept	-0.057	0.021	-2.76	0.006
Critic_Score	0.004	0.000	11.96	< 0.001
User_Score	-0.003	0.003	-1.06	0.290
Platform_Wii	0.028	0.010	2.78	0.006
Platform_X360	0.009	0.008	1.14	0.254
Genre_Racing	-0.040	0.012	-3.39	< 0.001
Genre_Role-Playing	-0.019	0.012	-1.63	0.104
Genre_Shooter	0.005	0.010	0.52	0.603
Genre_Sports	0.009	0.010	0.90	0.368

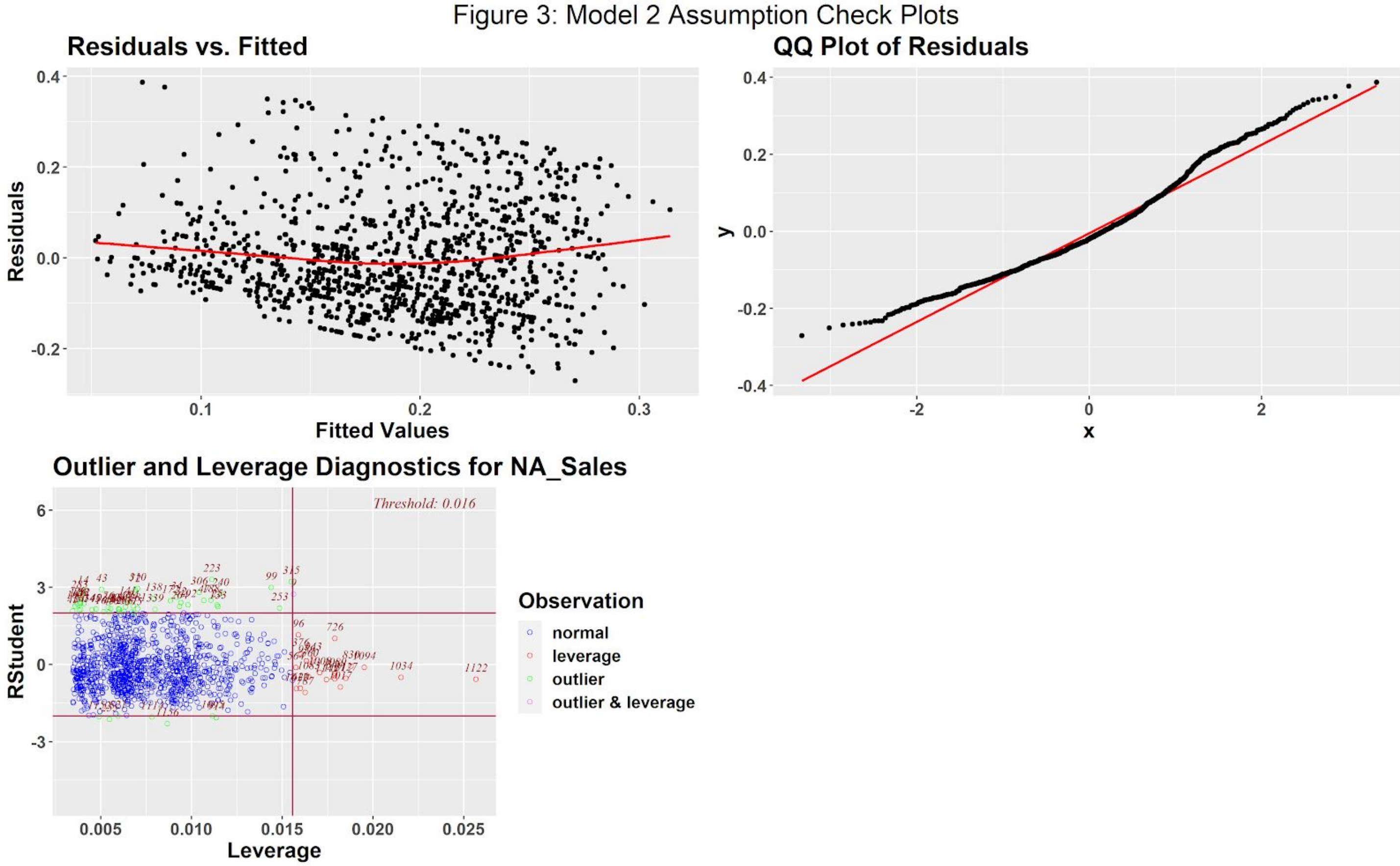
Note:
Reference level is Critic Score
n = 1156. r-squared = 0.16, F(8,1147) = 28.39.
F-Statistic p-value: < 0.001

significant coefficients. These reference groups are the critic score, user score, wii platform and racing genre . All other references groups have p-values greater than our alpha value which suggests that their coefficients are not significantly different from 0. The r-squared value of our model is 0.16, which tells us that the model as a whole explains 16% of the variance in NA sales. Lastly, the p-value of the F-statistic is less than 0.001 which is less than our alpha value. This allows us to reject the null hypothesis and reveals that our new model is significantly better than a null model.

Assumptions

We will now examine the assumptions for the new model using various tests and plots from Figure 3. Examining the qq plot of residuals, we can see that the distribution of residuals is relatively normal except for slight deviation at the tails. The distribution of residuals looks very similar to the residuals distribution in our first model. I would say that due to how minor and gradual the deviations are, the residuals for model 2 are normally distributed. Looking at the graph of Outlier and Leverage Diagnostics for NA sales, we can see that there is one plot point (9) that is both an outlier and has leverage, making it an influential outlier. We violate the

Figure 3: Model 2 Assumption Check Plots



assumption of no influential outliers, but tHis is a large improvement from model 1 where there were multiple influential outliers. Examining the Residuals vs fitted plot for homoscedasticity, there is no funnel shape in the distribution that suggests unequal variance in residuals. The reference line also is relatively horizontal and straight which further supports successfully fulfilling the assumption. Looking at the Residuals vs Fitted plot for the assumption of independent errors, we can see that the reference line is more horizontal and straight compared to the plot from model 1, but there is still a very slight curve to the reference line which may suggest that the errors are not independent. At the end, I will conclude that the model still violates the assumption of independent errors. Lastly, I will test the assumption of multicollinearity using a VIF test. Examining the results, we can see that all independent variables have VIF values between 1 and 2. This tells us that we are under the threshold (4.0) for VIF and that we are safe from multicollinearity issues. Overall, There seems to be an improvement in fulfilling assumptions in model 2 compared to model 1. However there are still violations in the assumption of no influential outliers and the assumption of independent errors.

We will now use ANOVA to determine if our fuller model 2 is significantly better at predicting NA sales than nested model 1. From our ANOVA results, model 2 has 1 degree of freedom which makes it the full model and makes model 1 the nested model. We can also see that the p-value is 0.0003, which is smaller than our alpha value. Therefore, we reject null and conclude that our second full model is significantly better at predicting NA sales than our first nested model.

Improving the Model

Examining the results of the information and assumption checks of our previous two models, I will now refine my model by addressing existing assumption violations. First, I will also address violation of independent errors caused by the curvilinear shape of the residuals vs fitted plot in model 2. To fix this issue, I will add critic score squared to the new model. I will remove data entry numbers 9 from the dataset as it acted as an influential outlier in our second model. I also removed data entry numbers 223, 914 and 1095 from the dataset as they were newly emerging influential outliers in our new model. Removing these influential outliers will allow our new model to successfully fulfill the assumption check for no influential outliers.

Examining the model 3 summary, we can see that the r-squared value is 0.19 compared to the model 2's r-squared value of 0.16. This means that the new model explains 19% of the variance in NA sales compared to the previous model explaining 16% of the variance in NA sales. Also, the p-value of the F-statistic for the new model is less than 0.001, which is equivalent to the previous model's F-statistic p-value. Therefore, our new model also rejects the null hypothesis and maintains the property of being significantly better than a null model.

Model 3: Characteristics Associated with NA Video Game Sales

Predictor	Estimate	Std. Error	t-statistic	p-value
Intercept	0.284	0.068	4.16	< 0.001
Critic_Score	-0.008	0.002	-3.44	< 0.001
Critic_Score_sqr	0.000	0.000	5.36	< 0.001
User_Score	-0.002	0.003	-0.60	0.549
Platform_Wii	0.026	0.010	2.66	0.008
Platform_X360	0.008	0.008	0.99	0.324
Genre_Racing	-0.041	0.012	-3.49	< 0.001
Genre_Role-Playing	-0.017	0.012	-1.43	0.152
Genre_Shooter	-0.002	0.010	-0.17	0.865
Genre_Sports	0.006	0.010	0.58	0.564

Note:
Reference level is Critic Score
n = 1156. r-squared = 0.19, F(9,1143) = 30.49.
F-Statistic p-value: < 0.001

Assumptions

Looking at the assumption checks, we can see that the new model has improved independent errors and a lack of influential outliers. Examining the assumption check plots shown in Figure 4, The residuals distribution of the new model looks almost identical to the distributions of both model 1 and model 2, which maintains its assumption of normal residuals. The residuals vs fitted values plot also does not have a particular shape to the plot and the reference line is completely horizontal, which verifies that we have removed the curvilinear shape that caused us to violate the assumption of independent errors in our previous model. In the Outlier and Leverage Diagnostics for NA Sales plot of our new model, we see that there are no plot points that are both an outlier and leverage, which allows us to fulfill the assumption of no influential outliers. This is fulfilled due to removing the influential outliers for the new model's dataset.. The multicollinearity of our new model is worse than in the previous model, but that is expected due to adding the squared critic score variable to our model

Figure 4: Model 3 Assumption Check Plots

to fix the curvilinear shape in independent errors as overlap between critic score and squared critic score is inevitable. Lastly, the assumption of homoscedasticity is improved by the lack a cone shaped spread in the Residuals vs Fitted distribution.

Finally, we will conduct an ANOVA test comparing our final model with model 2, making sure to apply the same altered dataset to both models. From our ANOVA results, we see that model 3 has a degree of freedom of 1, which makes model 3 the full model and makes model 2 the nested model. Also, the p-value is less than 0.001, which is smaller than our alpha value. Therefore, we reject null and conclude that our third full model is significantly better at predicting NA sales than our second nested model.

Conclusion

By removing influential outliers and addressing the assumption of independent errors by adding a squared critic score variable to the model, we were able to create an effective, violation-free model that accurately and significantly predicted North American video game sales using video game critic score, user score, platform and genre. Not only did we develop a model to predict NA video game sales, but we also determined that critic score was a significant predictor of NA sales, that the NA sales of the Wii platform is significantly higher than the PS3 platform reference group, and that the NA sales of the racing genre is significantly lower than the action genre reference group. Also, by cleaning and filtering the dataset of this study to only include video games released on the Wii, PS3 and Xbox 360, and removing outliers in the critic score and NA sales categories, my dataset was very comprehensive and allows me to apply the results of this study to the larger population.