STAT 204 Final Project: Predicting Video Game Sales

Gulzina Kuttubekova[1], Meltem Ozcan[1]
UC Santa Cruz[1]

Abstract

What properties of video games best predict global sales? Are there regional differences in the popularity of games with certain properties? In this article, we explore these questions and others using a dataset containing video game sales information globally. As part of our exploration, we conducted EDA to formulate specific hypotheses regarding interactions and built multiple linear regression models using the insights we gained through EDA. We found that there were significant interactions between region and genre, region and main developer company, region and main publisher company, and decade and genre. We built six MLR models and compared the models using various tools including AIC and BIC.

KEY WORDS: Multiple Linear Regression (MLR), Hypothesis Testing, Model Comparison

1. Introduction

1.1 The Data

The data we used for this investigation was obtained by a web scrape of Metacritic and VGChartz along with manually entered year of release values. The dataset initially contained 16719 observations with 6 categorical (Name, Platform, Genre, Publisher, Developer, Rating) and 10 numerical (Year_of_Release, NA_Sales, EU_Sales, JP_Sales, Other_Sales, Global_Sales, Critic_Score, Critic_Count, User_Score, User_Count) variables making up a total of 16 variables. Global_Sales is the sum of the number of units sold in each of the four regional sales variables for Japan, EU, North America, and the rest of the world (in millions of units), and is our key variable of interest.

1.2 Data Cleaning and Manipulation

1.2.1 Removal of NAs

We began cleaning our data by re-coding all empty cells to NA, and observing where the NAs were located. As the missing values were mostly concentrated in categorical variables such as Rating and Developer, we decided to remove all NAs. After removing observations with NAs and blank cells we were left with 6825 observations ($\sim$40% of the data).

To confirm that we did not fundamentally change variables with NAs by removing the NAs, we compared the distributions of the variables in the data set "before" and "after" this process. We found that the distributions for all numerical variables remained approximately the same after the removal of NAs. However, for most categorical variables, the proportions of each level within the variable distribution had changed. The most noticeable changes in proportion occurred in genre and developer variables.

Later, we tested some of the hypotheses on the dataset with NAs and ended up with results and conclusions similar to those from our analysis of the reduced dataset (without NAs). Moreover, in some cases, the $Adj - R^2$ of models doubled when fitted on the reduced data set in comparison to the model on the full data set.

1.2.2 Creation of new variables

Having removed the NAs, we created a melted data set that included Region as an explanatory variable (Levels: NA, Japan, EU, Other).

We then created 7 new variables reflecting information that we thought might be relevant to predicting game sales: Main_Publisher, Main_Developer, Developer_Country, Platform_Generation, Family_Platform, Platform_Company, and Decade.

To create the Main_Developer and Main_Publisher variables, we re-grouped observations for the Developer and Publisher variables by parent companies to reduce the number of levels to 11. These 11 levels correspond to the top 10 video game companies and an umbrella group for others (Levels: Activision Blizzard, Nintendo, Square Enix, EA, Tencent, Microsoft, Namco, Sony, SEGA, Ubisoft, Other).

We also re-coded EA, Microsoft, and Activision Blizzard to USA, Ubisoft to France, Tencent to China, and Nintendo, Namco, Square Enix, SEGA and Sony to Japan in order to create Developer_Country (Levels: USA, Japan, China, France, Other).

We researched which generation, family of systems, and company each game platform belonged to in order to create the Platform_Generation (Levels: 5th Generation, 6th Generation, 7th Generation, 8th Generation), Family_Platform (Xbox, DS, Wii, PS, Miscellaneous), and Platform_Company (Levels: Nintendo, SEGA, SONY, Microsoft) variables. Finally, we created Decade by re-coding the Year_of_Release variable (Levels: 80s, 90s, 00s, 10s).

2. Exploratory Data Analysis (EDA)

2.1 Exploration of the existing variables

2.1.1 Response variable: Sales

Globally, the mean number of units sold per game was 0.78 million. On average, a total of 0.39 million units were sold in

North America, 0.24 million units were sold in Europe, 0.06 million units were sold in Japan, and 0.08 million units were sold in other parts of the world. We identified "Wii Sports", "Mario Kart Wii", and "Wii Sports Resort" as the games with the highest number of units sold at 82.53 million, 35.52 million, and 32.77 million respectively.

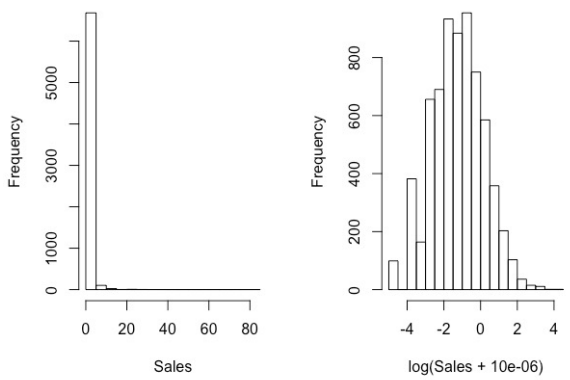

Figure 1: Epsilon and log transformation

We observed that the distribution of the Global Sales variable was highly right-skewed. Prior to transforming this variable, we shifted each data point by epsilon (0.000001) to avoid issues with 0s, and then explored a number of transformations including the reciprocal, square root, and log transformations. We found that re-expressing the sales data as log values were the most successful in making the distribution approximately symmetric. Figure 1 illustrates the distribution of the Sales variable before and after the transformation.

2.1.2 Genre

Globally, we found that shooter games, platform games, and sports games are the genres with highest numbers of units sold following the miscellaneous group. Strategy games had the least number of units of sold.

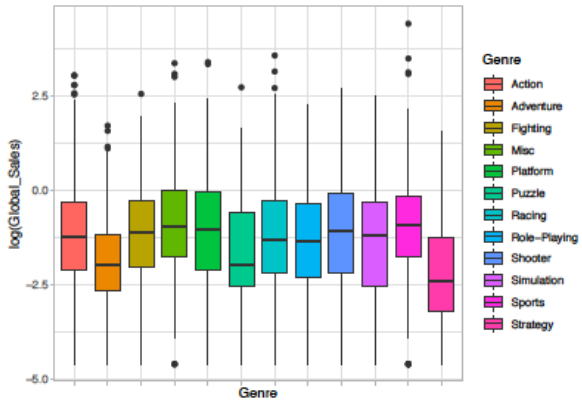

Figure 2: Log(sales) by genre globally.

When we explored sales in relation to genres and regions, we discovered a potential interaction between the region and genre variables. The miscellaneous level for Genre had the most units sold in all regions outside of Japan. In North America, the top three genres following miscellaneous are shooter games (11.92%), sports games (11.13%), and platform games (11.02%). We observed a similar pattern in the EU where the top three genres are shooter games (11.53%), racing games(10.81%), and platform games (10.27%), and in the rest of the world (Other_Sales) where genres with most units sold were shooter games (11.69%), racing games (11.53%), and sports games (11.46%). In contrast, the most popular games in Japan were in role-playing (20.04%), puzzle (14.79%), and Platform (12.46%) genres, followed by Simulation games which bring 12.46% of sales.

2.1.3 Critic and User Scores

We found that both User_Score and the Critic_Score were left-skewed, meaning that critics and users alike rated games more positively than negatively overall. User_Score and Critic_Score were positively correlated (Pearson r $= 0.58035$), indicating a close linear relationship between the two variables and potential multicollinearity.

2.1.4 Year of Release

The games in the dataset were released between 1985 and 2016. 2008, 2007 and 2005 were the years with the highest number of games released: 592, 590, and 562 respectively.

2.1.5 Publisher and Developer Companies

The games were created by 1289 unique companies (developers) and marketed and sold by 262 companies (publishers). The publisher and developer overlapped for 720 games.

2.1.6 Rating

The games in the dataset were rated Teen (34.83%), Everyone (30.5%), Mature(21%), and Everyone 10+ (13.63%). The levels Adult Only, Kids to Adults, and Rating Pending each had one game listed.

2.1.7 Platform

The dataset has information on games played on 17 different platforms. The majority of the games in the dataset run on the PS2 console (16.7%), the X360 console (12.57%), the PS3 console (11.27%), the PC console (9.54%), and the Xbox console (8.28%).

2.2 Exploration of the new variables

2.2.1 Main_Publisher and Main_Developer

65.79% of the games in the dataset were developed by non top 10 companies. EA developed 10.81% of the games, Ubisoft 5.41% of the games and Sony developed 4.01% of the games in the dataset. Figure 3 illustrates a potential interaction between the developer company and region variables. The

majority of video games sold in Japan were developed by companies based in Japan, such as Square Enix, Nintendo, and Namco. American companies such as Microsoft, EA and Activision Blizzard appear to account for a larger proportion of the games sold in North America in comparison.

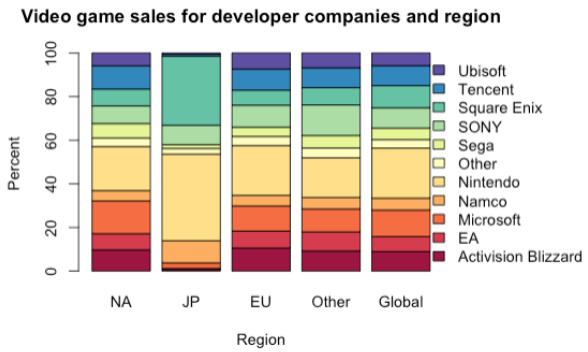

Figure 3: Video game sales for developer companies and region

41.51% of the games were published by non top 10 companies. EA published 13.98%, Ubisoft published 9.20%, and SEGA published 8.16% of the games.

2.2.2 *Platform_Generation*

Video games are characterized also by their generation, which is determined by the hardware and the graphical output of consoles. 43.37% of the games are in the 7th generation, 33.76% in the 6th generation, 11.74% in the 5th and 11.14% in the 8th generation. The most recent games in the dataset are from 2016, and there have been numerous game releases in the 8th generation (current video game generation) that are not accounted for in this dataset.

2.2.3 *Platform_Company*

41.11% of the games run on Sony platforms, 32.72% on Microsoft platforms, 25.96% on Nintendo platforms, and 0.21% on SEGA platforms. We checked whether there were multiples of games running on different platforms and did not see instances of this situation.

2.2.4 *Decade*

72.32% of the games were released in 2000s, and 25.03% in the 2010s, and 2.62% in the 1990s. We have only 8 games in our dataset from the 1980s, accounting for 0.03% of the games in our dataset.

3. Interactions

3.1 Hypothesis Testing: ANOVA

We built a least squares model as well as a weighted least squares model for each of the hypotheses regarding interactions. We tested our hypotheses in both the full dataset and also the subset of our data where the Sales variable was greater than 0.

3.1.1 *Region and genre*

We considered a two-way ANOVA model with interactions:

$$y_{ijk} = \mu + \alpha_i + \gamma_j + (\alpha\gamma)_{ij} + \epsilon_{ijk},\ \epsilon_{ijk} \overset{iid}{\sim} N(0, \sigma)$$

where α_i is the main effect of genre type i, for $i = 1, ..., 12$, γ_j is the main effect of region type j, for $j = 1, 2, 3, 4$ and y_{ijk} is the k^{th} sales such that $k = 1, ..., n_{ij}$ for genre i and region j.

We made the following assumptions for ANOVA:

1. Error terms are identically and independently distributed
2. Error terms are normally distributed with mean zero and fixed variance
3. Homogeneity of variance between treatments

Having built the model, we were interested in testing the following hypotheses:

$$H_0 : (\alpha\gamma)_{ij} = 0\ for\ all\ i, j$$
$$H_1 : (\alpha\gamma)_{ij} \neq 0\ for\ at\ least\ one\ pair\ of\ i, j$$

There was a significant interaction at $\alpha = 0.05$ significance level between genre and region when predicting log(sales) with an F value of 29.224. For instance, genres such as role-playing and simulation were significantly more popular in Japan, and genres such as shooter and sports games were more popular in North America and the EU.

We used the same two-way ANOVA model formulation for the remaining hypothesis tests by mainly changing the main effect variables. Model assumptions were held the same. Table 1 displays the ANOVA output for the four hypotheses tested.

3.1.2 *Region and developer company*

We found a significant interaction between the region and developer company variables such that a larger portion of video games sold in Japan were developed by companies based in Japan, while games developed by companies based in other regions such as the USA accounted for a smaller proportion of all sales. This interaction was significant at the 0.05 significance level with an F-value of 28.516.

3.1.3 *Region and main publisher*

We also found that the interaction between publisher company and the region variables was significant at the 0.05 significance level with an F-value of 43.339. For instance, people in Japan were more prone to buying video games published by Japanese companies such as SEGA, Namco, and Nintendo. EDA had revealed that the vast majority of the games sold in Japan were published by Japanese companies, and our analysis confirmed that there was a significant interaction between publisher company and region.

3.1.4 Genre and decade

Our fourth and final hypothesis regarding interactions was between genre and the decade. We found a significant interaction between these two variables at the 0.05 level with an F-value of 3.5808. For instance, action games became more popular over time, while strategy games lost popularity. As illustrated in Table 1, SS_interaction for the Genre:Decade term accounts for a small percentage of SSE.

When we repeated these four hypothesis tests using the dataset without 0s, we found that the interactions between region and genre, developer and region, publisher and region, and decade and genre were still significant.

3.2 Checking model assumptions for ANOVA

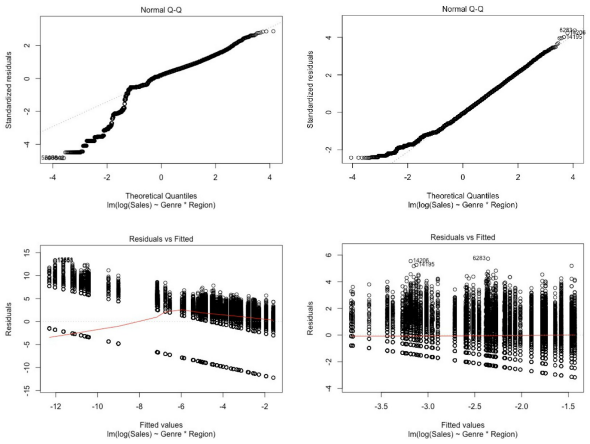

Figure 4: Checking model assumptions for ANOVA using datasets with 0s (left) and without 0s (right).

An examination of the residual plots shows that the normality and homogeneity of variance assumptions for the error term were violated in all four models. The Shapiro-Wilks test for normality and the Breusch-Pagan test for heteroscedasticity confirmed these violations for both the least squares and the weighted least squares approaches.

We then repeated this process using the dataset without 0s for both the least squares and the weighted least squares methods. While there was some visual improvement in the residual vs. fitted and Normal-QQ plots, the Shapiro-Wilks and Breusch-Pagan tests revealed p-values less than 0.05. Hence, the normality and homogeneity assumptions for ANOVA were violated for both LS and WLS methods in both datasets.

4. Multiple Linear Regression (MLR) Models

We split the dataset into training and test subsets with a 70-30 split. We used a stratified partitioning method to ensure that each level of each categorical variable would be represented in equal proportion in either set. We built 6 models using the training set, and then used these models in the test set for prediction.

4.1 Initial Variable selection

As the first step of variable selection, we eliminated a number of variables using our findings from the EDA and our intuitive biases. We excluded Platform, Publisher, and Developer from our analyses as these variables have too many levels to draw meaningful conclusions from, and instead used the consolidated versions of these variables.

We decided to eliminate the User_Score and User_Count variables to avoid multicollinearity issues as these variables had great overlap with the Critic_Score and Critic_Count. We eliminated Developer_Country after EDA as other variables such as Main_Developer provided the same information.

After this initial round of elimination, we retained the following variables: Global_Sales, Genre, Rating, Critic_Score, User_Score, Platform_Generation, Family_Platform, Platform_Company, Main_Developer, Decade, and Main_Publisher. Having narrowed down to these variables, we began building MLR models. We will first go through the models we explored and check for the assumptions of each model, and follow this exploration with a section on model comparison to select the best model(s) to predict video game sales globally.

4.2 Models

4.2.1 Model 1

$$\begin{aligned}\log(\text{Sales}) = \quad & \text{Genre} \\ & + \text{Rating} \\ & + \text{Critic\_Score} \\ & + \text{User\_Score} \\ & + \text{Platform\_Generation} \\ & + \text{Family\_Platform} \\ & + \text{Platform\_Company} \\ & + \text{Main\_Developer} \\ & + \text{Decade} \\ & + \text{Main\_Publisher}\end{aligned}$$

Implicitly, the MLR1 model is given as:

$$\mathbf{y} = \mathbf{X}^\top \beta + \epsilon$$

where $\mathbf{y} = [y_1, ..., y_{N_{train}}]^\top$ is the log(Sales), $\mathbf{X}$ is model matrix, β is a vector of parameter coefficients and ϵ is the error vector s.t. $\epsilon_i \sim N(0, \sigma)$.

We consider the following assumptions for MLR model:

1. There is a linear relationship between response variable *log(Sales)* and numerical explanatory variables *Critic_Count, Critic_Score*

2. Error terms are identically and independently distributed
3. Error terms are normally distributed with mean zero and fixed variance
4. Homoscedasticity: variance of error terms is similar across the independent variables

In MLR1 all added variables were significant at 0.05 significance level. Table 2 contains different measurements on model assessment. For instance, adjusted R-squared value ($AdjR^2$) for MLR1 was 0.4625, meaning that only 46.25% of variation in log(Sales) is explained by the explanatory variables in model MLR1. The train root mean squared error (RMSE) value was 6.774657, which decreased to 6.494338 when we applied the MLR1 predictive model on the test data set.

4.2.2 Model 2

In the second model, we added the interaction terms we had determined to be significant in our ANOVA section, to MLR1. This model has the same assumptions as MLR1. Log(Sales) are *iid* and error terms are normally distributed with zero mean and a fixed variance. MLR2 structure:

$$\begin{aligned}\log(\text{Sales}) = \quad & \text{Genre} \\ & + \text{Rating} \\ & + \text{Critic\_Score} \\ & + \text{User\_Score} \\ & + \text{Platform\_Generation} \\ & + \text{Family\_Platform} \\ & + \text{Platform\_Company} \\ & + \text{Main\_Developer} \\ & + \text{Decade} \\ & + \text{Main}_{\text{P}}\text{ublisher} \\ & + \text{Genre} : \text{Region} \\ & + \text{Main\_Developer} : \text{Region} \\ & + \text{Main\_Publisher} : \text{Region} \\ & + \text{Genre} : \text{Decade}\end{aligned}$$

All variables including the interactions were significant in this model. We observed that the AIC score went down by 1975.9 with the inclusion of the interaction terms, and BIC dropped by 1045.6, indicating that the model with the interactions is better despite having a larger number of parameters. The second model captures an additional 4%, hence 50.72% of the variation in the response variable compared to the first model without interactions.

Both train and test RMSE values have increased compared to the RMSE values in MLR1. We also observed a similar pattern between train and test error values for MLR2, i.e. test error = 6.650431 was smaller than train error = 6.862823, while in practice we would have expected the opposite. Normally, we could interpret this pattern as a good performance of the fitting model. However this interpretation is likely misleading as both of the error values have relatively large magnitude. This unexpected finding can be explained with the famous the variance-bias trade-off we encounter in the model. Since each categorical variable, on average, has 7 levels, we have relatively small variance within each group. This high bias created by small variance in the groups leads to "underfitting", which in turn causes the model to have a higher train and smaller test error. This odd relationship between the train and test errors can also be explained by the "curse of dimensionality" such that we have a relatively small number of observations in the training data set to estimate the relatively large number of coefficient parameters.

We then ran these models using the data set without 0s in the response variable, and found that all the variables in MLR1 and MLR2 were still significant. Table 2 displays the measurements for the performance of each of the models when we used the datasets with and without 0s. When the data set without 0s were used, AIC went from 53396.81 to 52267.82 when the interaction terms were added to MLR1 to get MLR2. BIC went from 53831.76 to 53588.21 with the inclusion of the 4 interaction terms. $AdjR^2$ values also decreased from 0.4625 to 0.4014 for MLR1 and from 0.5072 to 0.4425 for MLR2. RMSE for MLR1 was 2.900504 on the train data without 0s and 2.8137 on the test data without 0s. Similarly, RMSE for MLR2 was 2.908342 on the train data without 0s and 2.806784 on the test data without 0s.

For the rest of this study, we will be reporting the AIC, BIC, and RMSE values for the models when tested on the data set with 0s on the 2 only as the pattern we have outlined here persisted for each of our models.

Since we ran into the "curse of dimensionality" in the first two models with 11 explanatory variables, and 182 coefficient parameters in MLR2, we decided to drop some of the least significant variables among the 11. We tried a stepwise selection method by AIC values, which is the combination of forward and backward elimination of explanatory variables. This approach did not eliminate any of the variables, so in the next steps we tried to reduce the number of variables by half by eliminating the explanatory variables intuitively. We will refer to MLR1 and MLR2 as "full" models and MLR3 through MLR6 as "reduced" models in the rest of this text.

For models MLR3, MLR4, and MLR5, we decided to fix our number of explanatory variables at 5 (plus relevant interactions) in order to reduce the number of coefficient parameters to be estimated. If we were to run and compare each of the 5-variable models we could build from the 11 variables, we would need to compare 462 models. We instead generated our models with the following procedure:

- We initially concentrated on keeping the variables that can be controlled by the company developing, publishing or investing in the game. Hence, 3 main explanatory variables are fixed: Region, Genre, and Rating.

- We excluded Decade as it is a variable that cannot be controlled by stakeholders and we also removed Critic_Count.
- We made sure at least one of the models contains Critic_Score as a predictor.
- To select the remaining variables, we treated Main_Developer and Main_Publisher as one family and Family_Platform and Platform_Company and Platform_Generation as a second family of variables as these variables are interconnected, and selected the remaining variables for the models from the two families.

In MLR6, we only included the three variables {Genre, Region, Rating} we deemed essential, and one interaction term.

4.2.3 Model 3

MLR3 structure:

$$\begin{aligned}\log(\text{Sales}) = \quad & \text{Region} \\ & + \text{Genre} \\ & + \text{Rating} \\ & + \text{Critic\_Score} \\ & + \text{Platform\_Generation} \\ & + \text{Region} : \text{Genre}\end{aligned}$$

All predictors in MLR3 were significant at the 0.05 significance level. Even though the number of variables in MLR3 was less than about half of the full models, the adjusted R-squared value indicates that about 41.57% of the variation in the response variable was explained by explanatory variables in the model.

4.2.4 Model 4

MLR4 structure:

$$\begin{aligned}\log(\text{Sales}) = \quad & \text{Region} \\ & + \text{Genre} \\ & + \text{Rating} \\ & + \text{Main\_Publisher} \\ & + \text{Platform\_Generation} \\ & + \text{Region} : \text{Genre} \\ & + \text{Region} : \text{Main\_Publisher}\end{aligned}$$

All predictors in MLR4 were significant at the 0.05 significance level. Explanatory variables in MLR4 explained 42.97% of the variation in the response variable.

4.2.5 Model 5

MLR5 structure:

$$\begin{aligned}\log(\text{Sales}) = \quad & \text{Region} \\ & + \text{Genre} \\ & + \text{Rating} \\ & + \text{Main\_Developer} \\ & + \text{Platform\_Company} \\ & + \text{Region} : \text{Genre} \\ & + \text{Region} : \text{Main\_Developer}\end{aligned}$$

All predictors in MLR5 were significant at the 0.05 significance level. The combination of explanatory variables used in MLR5 explained 40.94% of the variation in log(Sales).

4.2.6 Model 6

MLR6 structure:

$$\begin{aligned}\log(\text{Sales}) = \quad & \text{Region} \\ & + \text{Genre} \\ & + \text{Rating} \\ & + \text{Region} : \text{Genre}\end{aligned}$$

All predictors in MLR6 were significant at the 0.05 significance level. The adjusted r-squared was 0.3556, which is to be expected as this is the model with the fewest variables and the number of variables tends to inflate the adjusted r-squared metric.

4.3 Checking Model Assumptions

Our assumptions for the normality and homogeneity of variance for the error terms were also violated for the multiple linear regression models, checked visually and confirmed though the the Shapiro-Wilks test for normality and the Breusch-Pagan test for homogeneity of variance. The left-hand side of Figure 5 displays the Normal-QQ plot and the residuals vs fitted plot for MLR1. The pattern of violations observed here were replicated across all 6 models.

The right-hand side of Figure 5 illustrates testing MLR1 using the data set without 0s affected the residuals. Visually, there is great improvement in the left tail of the Normal QQ-plot, and the red reference line in the residuals vs. fitted plot is closer to the horizontal 0 line. However, we confirmed via the Shapiro-Wilks and Breusch-Pagan tests that these assumptions were indeed violated.

5. Model Comparison

Having built the six multiple linear regression models, checked for the assumptions for the error term, and explored how the performance of the models change for the train and

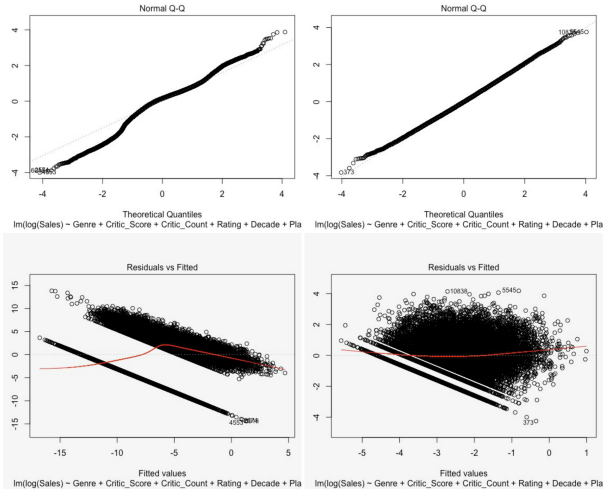

Figure 5: Checking model assumptions for MLR1 using dataset with (left) and without (right) 0s.

test datasets with and without 0s, we then compared them using various model assessment criteria. Table 2 displays the relevant metrics.

MLR1 and MLR2 explained the highest percentage of variance in the data, with adjusted R-squared values of 0.4625 and 0.5072 respectively. MLR2 had the lowest AIC and BIC values across the six models. However, MLR1 and MLR2 were the first two models we eliminated due to the high number of variables and the issues we encountered with the variance-bias trade-off and the "curse of dimensionality".

We then considered comparisons between the remaining four models.

Looking at the train RMSE and test RMSE shift, we found that the models where we observe an increase from the train RMSE to the test RMSE were models MLR4, MLR5, and MLR6 while RMSE decreased from the train test to the test set in MLR3. The increase in the RMSE value from the train to the test sets can be an indication that these models are behaving as they should and the underfitting we had observed due to the curse of dimensionality might have been solved for these there models, but not for MLR3.

We found that reducing the number of variables actually led to an increase in the AIC and BIC values in MLR6, and this model explained the smallest proportion of the variance in the data at 35.56%. Hence, it was the next model we eliminated.

MLR4 had the lowest AIC across the remaining models at 134653 and the lowest BIC at 135445.7. BIC more strongly penalizes the number of terms included in the model, and while MLR3 and MLR6 had fewer terms in comparison to MLR4, the BIC value was lower for MLR4. The fact that MLR4 is the preferred model with both AIC and BIC is a good indication that MLR4 is the best model.

MLR4 also explained the largest proportion of the variance in the data at 42.97% (adjusted R-squared value of 0.4297). We found that removing the additional variables from MLR1 to get MLR4 only led to a small decrease in the percentage of variance explained in the data, and not a lot of information was lost in the process.

When we repeated this process for the dataset without 0s, we observed the lowest AIC and BIC values in MLR3 closely followed by MLR4 which had the advantage of having an increase in RMSE for the train to test shift.

The adjusted R-squared values were much lower for the dataset without 0s compared to the full dataset across all 6 models. The average train RMSE was 6.893953 for the dataset with 0s, compared to an average of 2.891143 for the dataset without 0s. This pattern indicates while the models explain smaller proportion of the variance in the data for the dataset without 0s, removing the 0s allowed for higher accuracy of the models.

Overall, we conclude that model MLR4 which predicts log(sales) with region, genre, rating, main publisher, and platform generation variables and the region-genre and region-main publisher interactions is the best model for the purposes we have outlined previously.

6. Conclusions, Challenges, and Future Directions

From the analyses we did above, we have determined that the genre and rating of video games and the companies that develop and publish the game are some of the most important variables in predicting sales. The region where the video games are developed and published also plays an important role, as we saw from the interactions related hypothesis tests. These conclusions can have a big impact on the game industry. For instance, giant Japanese companies like Nintendo, Namco or SONY can focus on designing and publishing more role-playing games and targeting Japan as their major market for these games as we have determined that role-playing is a highly popular genre in Japan and these companies have the advantage of originating from that region.

Similarly, games developed or published by American companies account for a small proportion of games sold in Japan as discussed previously. If American companies such as Microsoft wanted to increase their market share in Japan, they would also benefit from targeting genres not usually preferred by the NA audience for this particular region.

We had previously presented potential challenges that we might face in our project proposal. We attempted different types of possible remedies for each of these challenges during the final phase of the project. While we tried a number of different transformations on the response variable {log, sqrt, etc.}, the major model assumptions of the normality and homogeneity of error terms were violated in both ANOVA and MLR models. We used the initial EDA to detect multicollinearity and dependence between the explanatory variables and were able to eliminate some variables through

these findings. However, this approach did not fully fix the violations of our assumptions. Similarly, we used both LS and WLS methods in the estimation of coefficient parameters in both types of models, but we did not see any improvement in the violations of the assumptions.

We also encountered new challenges such as the variance-bias trade-off and the "curse of dimensionality". Reducing the number of terms in the models was helpful in remedying this problem to a certain extent.

Taking into account the challenges we faced, we can make the following improvements if we were to conduct a second study on the video games industry:

1. Collect more data
2. Reduce the number of coefficient parameters even further by selecting only few explanatory variables
3. Use different explanatory variables
4. Attempt transformations of the coefficient parameters
5. Try different coefficient parameters estimation methods
6. Investigate other KPIs
7. Conduct Bayesian analyses

REFERENCES

Rush Kirubi, Dec 2016. Video Game Sales with Ratings. Retrieved October 2019 from https://www.kaggle.com/rush4ratio/video-game-sales-with-ratings.

VGChartz. Platform Totals. Retrieved October 2019 from https://www.vgchartz.com/analysis/platform_totals/.

Wikipedia. Video Game Console. Retrieved November 2019 from https://en.wikipedia.org/wiki/Video_game_console/.

Wikipedia. Video Game Consoles by Generation. https://en.wikipedia.org/wiki/Category:Video_game_consoles_by_generation

Table 1: Output for the interaction terms: ANOVA.

Interaction term	*DF*	*F-value*	*p-value*	$SS_{interaction}$	*SSE*
Genre:Region	33	29.224	$<$ 2.2e-16 ***	16967	479456
Main_Developer:Region	30	28.516	$<$ 2.2e-16 ***	14842	472869
Main_Publisher: Region	30	43.339	$<$ 2.2e-16 ***	21888	458846
Genre:Decade	22	3.5808	$<$ 2.655e-08 ***	2122	734337

Table 2: Model Comparison.

Model	*Dataset*	*AIC*	*BIC*	$Adj - R^2$	*Train RMSE*	*Test RMSE*
MLR 1	with 0s	133184.3	133637.3	0.4625	6.774657	6.494338
	without 0s	53396.81	53831.76	0.4014	2.900504	2.813700
MLR 2	with 0s	131208.4	132591.7	0.5072	6.862823	6.650431
	without 0s	52267.82	53588.21	0.4425	2.908342	2.806784
MLR 3	with 0s	135200.5	135677.7	0.4157	6.700798	6.534950
	without 0s	55886.36	56344.61	0.3097	2.890188	2.837222
MLR 4	with 0s	134653.0	135445.7	0.4297	6.731926	6.807955
	without 0s	56472.16	57225.55	0.2877	2.886228	2.951015
MLR 5	with 0s	135495.3	136288.0	0.4094	6.694827	6.713769
	without 0s	56671.03	57432.19	0.2795	2.886106	2.929614
MLR 6	with 0s	137552.4	137997.3	0.3556	6.598686	6.718925
	without 0s	58565.85	58993.03	0.1949	2.875492	2.986754