Subset Selection and Cross-Validation:

Comparisons and Verifications of Selected Video Games Sales Models

Sungwoo Kim

Teachers College, Columbia University

Abstract

This study reevaluated linear models proposed by Joe Cox in his study on finding variables that can contribute to creating a blockbuster game as well as created new models to suggest for better predictive performance. After re-conducting the study with ANOVA and multiple comparisons test, the models from the original study were replicated, and additional predictive models were formed through stepwise regression and random forest approach. The results showed that the models in the original studies were not valid for applicative use due to issues in multicollinearity and relatively lower predictive performance. As for predictability, the random forest model performed best, followed by stepwise regression models. However, the small difference in predictive performance suggests using stepwise regression models due to its relatively easy interpretability. There were contrasting results on influential variables compared to the original study, which is most likely due to difference in the study methods.

Keywords: multiple comparisons, subset selection, cross-validation, video games

Subset Selection and Cross-Validation:

Comparisons and Verifications of Selected Video Games Sales Models

Significance of Video Game Industry

The video game industry market has continuously grown over the past decade due to easy online accessibility and significant development of hardware and graphics. A study from video games market research company Newzoo shows that the total market revenue of video games is forecasted to be $137.9 billion, which is nearly double the amount from 2012 with $70.6 billion (Wijman, 2018). Newzoo also predicted that the compound annual growth rate (CAGR) will be around 11% until 2021, where "maintaining a double-digit growth rate for 10 years is truly remarkable" (Wijman, 2018). Though over 50% of such growth is derived from the mobile gaming market, PC and console market were also steadily growing as well.

In addition to market trend and forecasts, the increasing popularity of the video game industry also has been exhibited through recognitions and corresponding changes in mainstream entertainment systems. In 2016, leading sports channel ESPN recognized eSports as a type of sport and started broadcasting popular eSports game such as *League of* Legends, *Dota* 2, and *Overwatch* (Peckham, 2016). This was a sudden change of stance, as ESPN President John Skipper previously stated that eSports was "not a sport" in 2014 (Peckham, 2016). However, the decision was made as staff found that video games as a genre of sports was "[becoming] consistently more popular and [getting] a lot more coverage in the mainstream" based on the increasing scale of professional gaming competitions (Peckham, 2016).[1]

[1] This year's DOTA 2 world championship prize pool was announced to be around 25 million dollars, which was recorded as the largest prize pool in a single competition, in the history of eSports (E-Sports Earnings, 2018). This

As a result of the massive expansion of the video game industry, the definition of 'blockbuster' has expanded to video games as well. While *God of War* sold "over 5 million copies in its first month", another recently released *Red Dead Redemption 2* sold over 17 million copies in the first two weeks with $725 million in revenue in the first three days of release. From observing commercial success of these video games and the increasing video game market size. Clearly, analyzing and finding key qualities that influence commercial success would provide a significant guideline to game creators for future production.

Literature Review

Joe Cox conducted a study on console video games to find how game characteristics, platforms, genre, games' view perspectives, and developers influence sales revenue. He collected 1770 video games from 2004 to 2010 and performed two OLS linear regression analyses, where one being a full model with all variables and the other being a reduced model only with variables that were found significant in the initial analysis. He also performed a series of logistic regressions with different sales thresholds as dependent variables to find if the predictors are consistently effective at increasingly successful titles[2]. His research findings showed that the most consistently significant factors that are associated with high sales are publishers, home platforms, and review scores (Cox, 2013).

While Cox's study provides valuable insight and a general idea of primary factors, there are several notable criticisms of the outlined models. First, Cox's specified model is only based

record is estimated to be broken by Epic Games's *Fortnite*, as Epic Games announced a "$100 million prize pool over the course of 2017-2018 competitive season" (Thier, 2018).

[2] Three logistic regressions were performed, where each threshold was 2 million dollars, 1 million dollars, and 0.5 million dollars. The sales data would be transformed into a binary variable for each analysis, where 1 means that the title's sales revenue is over the threshold revenue. As a result, the analyses would find if some qualities are significantly different by the threshold. This was done due to the long-tailed nature of the dependent variable, where only less than about 150 games made over 2 million dollars revenue in the sample (Cox, 2013).

on factors that are found to be significant ($p<.05$) from the full model. Though the perception to infer from significant coefficients is common, such is not a valid action to create an appropriate predictive model as it may reduce the reliability of prediction. Second, following from the first argument, the full model may not be the best valid prediction model as well. While it may be used as an exploratory measure as an overview, having over 40 variables may result in overfitting the data, which risks reducing the predictive power of the model. Finally, the models do not test how each category of variables influences sales. These linear regression analyses measure how each variable contributes, but they do not test if variables in the same category are significantly different to one another. For example, Cox's study does not show if having Nintendo as a publisher has a significant difference in sales compared to other publishers such as Rockstar.

Goals of the Study

Due to the potential risks of weak predictability in the proposed models, this paper aims to propose alternative models with the same data with higher predictability. First, this study will re-conduct the study by performing an omnibus analysis of variance (ANOVA) test, where any significant results will be further explored with comparison tests. Following the initial analysis, this paper will assess the validity of the original models by analyzing its variance influence factor (VIF). Specifically, this paper will use forward and backward stepwise subset selection processes as well as a random forest approach to produce potential predictive models. Then, this paper will compare the predictive effectiveness of the model by conducting K-fold cross validation on all models except random forest model and comparing their root mean square error (RMSE). Lastly, this paper will attempt to review Cox's findings on most influential factors on sales by finding the most important variables used in random forest modeling.

Methods

The Data

The data is directly retrieved from Joe Cox's study, "What Makes a Blockbuster Video Game? An Empirical Analysis of US Sales Data". There are a total of 1770 observations, with no missing data. The dependent variable used is lifetime unit sales of games released in US from 2004 to 2010 in millions of dollars, which were collected from an online database by VGChartz (Cox, 2013). Because the dependent variable is heavily skewed and violates the Shapiro-test of normality, a logarithmic transformation is performed in order to adjust for normality as it is done in the original study. However, careful interpretation from analysis results is still needed as the transformed variable still violates normality. Refer to Figure 1 for Q-Q plots for both un-transformed and transformed sales variable.

Table 2 lists all independent variables used in Cox's study. There are five main categories of independent variables as defined by Cox: Game characteristics, published console platforms, predominant genre of the game, view perspective of the game, and publishers of the game. The general game characteristics categories are further divided into three categories: year published, ESRB rating, and others. A total of 49 independent variables is used in the study, though only two of them are non-binary variables, which are review scores and max number of players. Contrary to Cox's variable layout, this study will treat Year as a series of binary variables of each year from 2004 to 2010, which will be labeled as Y04 to Y10.

Analysis Plan

This paper first conducted an overall F-test on all categories to find if variables in each category are significantly different among each other. Each category will be treated as a factor

with relevant variables except the “others’ category, where each variable in the category will be a separate variable. Though results of scale variables in “others” category in ANOVA will not be as interpretable as other variables, they will be used to control multiple comparisons of other categories. Then, a follow-up all-pairwise comparison tests are performed to find specific differences among the variables while controlling for other variables. Since the multiple comparison analysis will be unplanned, Tukey’s HSD Procedure was used to adjust p-values for multiple comparisons in each category.

Following the initial study, this paper replicated the linear regression analyses with the dataset defined above. They included the full model with all 49 independent variables and the specified model suggested by Cox with only statistically significant independent variables. Although Cox’s specified model is previously argued as potentially invalid for predictions, it was included as a part of the analysis and used as a reference for comparison. Due to the adjustment of Year variable, Cox’s results were different to models in this replication, which were compared and noted. Because this paper’s scope is on linear regression models and assessing their predictability, logistic regressions in Cox’s study was not replicated in this study.

Next, both forward and backward stepwise subset selection was performed with the full model and null model (Sales on intercept) as the scope of selection. The Akaike and Bayesian Information Criteria (AIC and BIC) were found as not appropriate to be used in this study due to the high number of variables and thus significantly long time taken for these computations.

After finding above linear regression models, the (VIF) for each model was examined to check for multicollinearity. Since VIF of over 5 in any variable could mean that the model has risks of multicollinearity, VIF will effectively explain the validity of the model. If any variable is found to have high VIF in the full model, a new regression model without high-VIF variables

will be modeled and added to the analysis. Here, it is assumed that variables with high VIF in full model will also have high VIF in the specified model. As so, only full model will have its modified version.

As an additional alternative to stepwise regression in terms of creating a predictive model, a random forest model was created. Unlike linear regression, random forest is modeled by fitting multiple decision trees, which would induce higher predictability. After modeling, the study measured the number of trees with the least mean square error (MSE), which was used to find out-of-bag RMSE of the model. To find major contributing variables to the model, a variance importance plot of mean decrease in node impurity was also produced. However, since random forest model is a series of piecewise functions, interpreting the model in detail may be difficult compared to linear regression models.

Finally, the performance of the models was compared through cross-validation method. With the four linear regression models specified above, a K-fold cross validation was performed to assess the accuracy of measurement of each model, where each fold is an equal portion of the data to test the model's performance. 5 folds are used for cross validation, and RMSE was computed by averaging and rooting mean square error from each fold. As for random forest model, out-of-bag RMSE was used to compare with other RMSEs as the random forest method cross-validates itself during the modeling process. Since RMSE is defined as the average variance of the difference between predicted value from the model and actual value from the data, lower RMSE would mean that the model is more likely to have a lower error when predicting.

The study hypothesized that the random forest method will have the best predictability. This hypothesis is based on random forest's nature of fitting multiple randomized low-bias

models, which would be an advantage over linear regression models with a single decision tree. The study also hypothesized that the stepwise regression models will have the second-best predictability. However, it is unknown of which stepwise regression (forward or backward) will have a better performance. As for influential factors, this study hypothesized that all categories will have a significant difference within each category, though specific differences are not known. In addition, the study hypothesized that significant variables in stepwise-selected models and Random Forest model will be mostly different from those from Cox's models.

Results

ANOVA and Multiple Comparisons

The result of omnibus ANOVA on all variables and categories can be found in Table 2. While Console ($F(4, 440)=4.765$, $p<.001$), Publisher ($F(17,440)=4.034$, $p<.001$), and Genre ($F(6, 440)=3.521$, $p=.002$) categories are found to have a significant difference within as hypothesized, Rating, Year, and Perspective of video games did not have a significant difference within each category. Another notable result is the significant F test of the Accessory variable ($F(1, 440)=8.963$, $p=.003$), which was found not to be significant in Cox's study.

Table 3 displays significant multiple comparison results within the Console, Publisher, and Genre categories. The most conspicuous aspect is the continuous presence of Nintendo in comparisons of publishers, where it is significantly higher than 7 other publishers. This follows Cox's finding that Nintendo is significantly related to higher sales revenue in the United States. Namco, on the other hand, is found to most significantly underperform among other publishers. However, Namco was not found to be significant to sales in Cox's study. In the Console category, two comparisons were found to be significant, which were both related to higher sales

in comparison with PSP, t(440)=4.149, p<.001, and t(440)=-2.841, p=.038. Lastly, only one significant comparison was found in the Genre category, which is between simulation and strategy, t(440)=3.496, p=.009.

Linear Regression

Table 4 lists the coefficient estimates for each linear regression model performed. The relative base variables for years, rating, platforms, perception, and genre are Y04, RatingE, X360, FirstPerson, and Action, respectively. As for developer dummy variables, the observations that are not marked by any developers listed in the variables are used as relative base observations, since several video games were made from non-major developers. While the full model measured all 50 variables, the specified model measured 31 variables, and the stepwise regression model selected 32 variables. Interestingly, both forward and backward stepwise regression processes selected the same model as a result.

In comparison with the original study, the overall F values of both full model and specified model decreased, with F(50, 1719)=28.2 compared to F(45, 1724)=30.99 and F(31,1738)=44.7 compared to F(27, 1743)=53.012. While the cause of the decrease is most likely due to increase in degrees of freedom, this also reduced the adjusted power of the study, with the replication having R^2=.435 for full model (compared to .450) and R^2=.434 for specified model (compared to .443).

The examination of variance inflation factor shows that the full model and the specified model possess some risk of multicollinearity. While full model possessed 6 variables with VIF of over 10, the specified model had 5 variables with the same criteria. Most of these variables were found to be years released. However, there was no significantly high VIF (>1) in the stepwise

regression model, even when the model included years as 3 of the variables. As this analysis suggests that the full and specified models may not be valid models to be used, a new model without high-VIF variables is made with 45 variables, which is also found at Table 3. The new model has higher F but slightly lower adjusted power, $F(45, 1724)=31$, $R^2=.433$.

Random Forest

A random forest model was performed with 500 total number of trees, with 19 variables tried at each split. The number of trees with minimum mean squared error of the model was 182, with RMSE of .885. The variance importance plot is illustrated in Figure 2, which shows that review scores, publisher (Nintendo), and being a sequel most influenced the formation of the prediction model, though sequel variable was not as influential as the first two. Although the significance of review scores and publisher supports Cox's findings, the variance importance plot mostly disagrees to Cox's findings as 5 out of top 10 influential variables are in the "game characteristics" category. However, it is worth noting that the specific coefficient is not defined, signifying that the variance importance plot would not be interpretable in terms of finding the direction (positive or negative) of each variable's influence.

Cross-Validation and RMSE

A follow-up calculation of RMSE through 5-fold cross validation analysis for linear regressions and split cross validation for random tree is indicated in Table 5. Out of linear regression models, stepwise regression models produced the least amount of RMSE as expected. However, its difference with the specified model is relatively minor compared to the full model. Interestingly, the predictive ability of the adjusted full model without high VIF did not have a

significant difference with that of full model as well. Out of all models, the random forest tree model had the best predictive accuracy with significantly lower RMSE of .885.

Discussion

The evaluation of replicated models through VIF and RMSE showed that both full and specified models had high risks of multicollinearity and relatively low predictability. Due to the question on the specified model's validity from its criterion of choosing only significant variables, only the full model would be at least interpretable in the original study. However, the modified model without significantly high VIF variables would be preferred than full model in terms of interpretation, though its predictability is on a similar level to other basic linear regression models.

In terms of predictability, the random forest model was found to have the lowest RMSE followed by stepwise regression model as hypothesized. However, the difference in prediction performance among the models is debatable, as the random tree model has around 7% lower than the stepwise regression model. As a result, stepwise regression model may be preferred by some considering that it is similarly effective but more easily interpretable than the random forest model.

In terms of verifying Cox's findings on influential variables, Cox's argument was only partially confirmed in a way that only few variables were found to be highly influential: review scores and Nintendo. Though it was argued that the coefficients are unknown and thus the random forest model would be difficult to interpret, the strong positive coefficients of such variables in linear models and their existence in both stepwise regressions suggest that the two variables are very likely to be crucial positive criteria of producing a successful video game title.

This argument is further supported by the multiple comparisons test, as games published by Nintendo had significantly higher sales compared to several video game publishers. However, the significant F-test result of accessories and multiple comparisons of Namco provided a contrasting argument to Cox's findings. These differences can be explained by the different analysis approach; while Cox's study treated all variables as an independent variable, this study conducted the F test by putting variables into several categories. In other words, Cox's study tested each variable's influence on US sales, while this study tested for significant difference within the category.

Limitations and Future Directions

There are several key limitations to the study, which could also be significantly improved. First, the models might have had a better fit with several additional variables that would be significantly related to game's financial success in theory. Such variables include original sales price and binary variable of award achievement. Though these variables were not able to be analyzed as they were not included in the dataset, such variables should be collected and investigated for their potential influence.

The most major limitation of this study is that the data is likely to be outdated. Although earlier report shows that the console market is still expanding, more than a decade has passed since 2004, and numerous consoles and games were developed and published. While some variables such as review scores, Nintendo, and sequel may still have some degree of influence on current model, the data needs to be updated with new observations and variables. Some of the potential new variables for consideration could be existence of downloadable content (DLC), accessibility of the game (acquirable strictly through hard copies, or also through downloads), and new consoles and platforms such as Nintendo Switch and Steam.

Another potential direction of the study could be analyzing mobile games, since reports indicated that mobile video games currently take over 50% of the video game market. Considering the possibility of difference in model structure compared to console markets, such study will both be more relevant to current industrial situations and produce more productive, applicative outcomes. The findings of this study would still be relevant, as some of the key influential variables may be similarly effective in the mobile games market.

References

Cox, J (2013, Apr 2013). What Makes a Blockbuster Video Game? An Empirical Analysis of US Sales Data. *Managerial and Decision Economics* (35), 189-198. https://doi.org/10.1002/mde.2608

E-Sports Earnings (2018). *Largest Overall Prize Pools in eSports*. Retrieved from https://www.esportsearnings.com/tournaments

Peckham, M. (2016, Mar 1). *Why ESPN Is So Serious About Covering Esports*. Retrieved from http://time.com/4241977/espn-esports/

Wijman, T. (2018, Apr 30). *Mobile Revenues Account for More than 50% of the Global Games Market as It Reaches $137.9 Billion in 2018*. Retrieved from https://newzoo.com/insights/articles/global-games-market-reaches-137-9-billion-in-2018-mobile-games-take-half/

Thier, D. (2018, Mar 21). *Fortnite's Massive Prize Pool Just Made It The Biggest Game In Esports*. Retrieved from https://www.forbes.com/sites/davidthier/2018/05/21/fortnites-massive-prize-pool-just-made-it-the-biggest-game-in-esports/#67ec6fea1063

Figure 1. Q-Q Plots of Sales Before and After Logarithmic Transformation

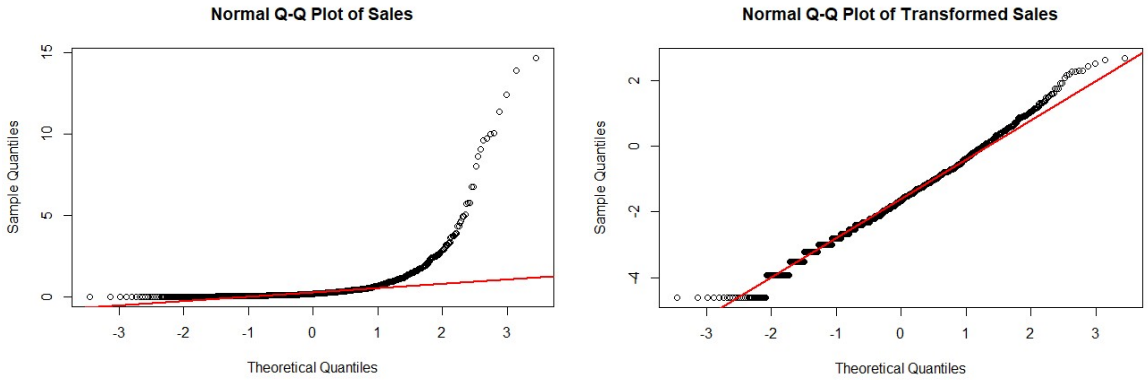

Figure 2. Variance Importance Plot of Random Forest Model by Mean Decrease in Node Impurity

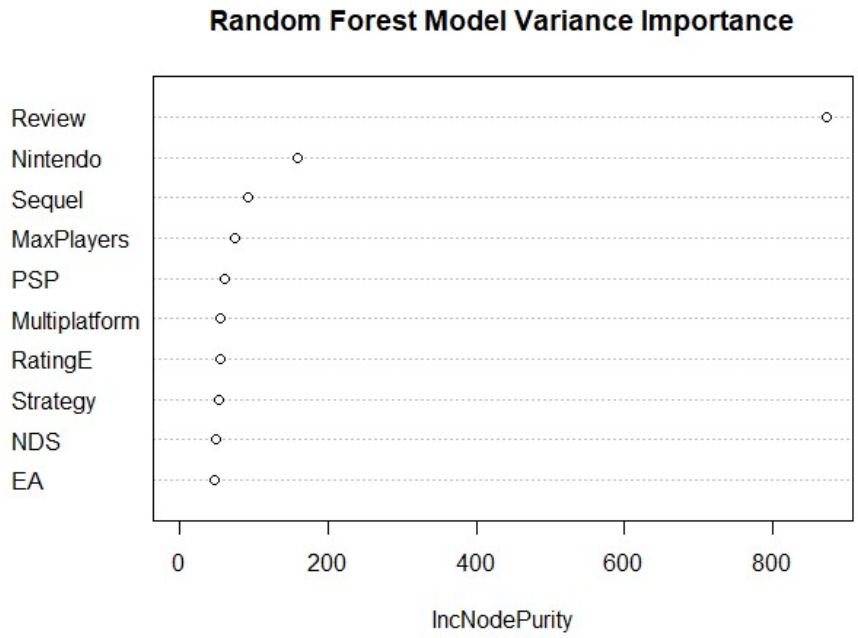

Table 1. Descriptives of Variables

Variable	Definition	Mean	Std. Deviation	Range
Dependent Variable				
Sales	US unit sales (in millions)	0.4798	1.06491	.01-14.6
logSales	US unit sales (in millions) with logarithm applied	-1.6	1.25	-4.61-2.69
Independent Variables				
Game Characteristics				
Y04	Game is released in 2004	0.01		0-1
Y05	Game is released in 2005	0.08		0-1
Y06	Game is released in 2006	0.17		0-1
Y07	Game is released in 2007	0.26		0-1
Y08	Game is released in 2008	0.25		0-1
Y09	Game is released in 2009	0.18		0-1
Y10	Game is released in 2010	0.04		0-1
Sequel	Game is a sequel	0.53		0-1
Rerelease	Game is a re-release	0.06		0-1
Review	Metacritic Review Score	68.43	13.895	12-98
RatingE	Game is rated as “E”	0.33		0-1
RatingT	Game is rated as “T”	0.29		0-1
RatingM	Game is rated as “M”	0.15		0-1
MaxPlayers	Maximum number of players	1.66	1.236	1-8
Online	Game has online functionality	0.35		0-1
Licensed	Game is officially licensed	0.35		0-1
Accessory	Game comes with an accessory	0.01		0-1
Multiplatform	Game released for multiple platforms	0.51		0-1
Platform				
NDS	Game released for Nintendo DS	0.24		0-1
Wii	Game released for Nintendo Wii	0.17		0-1
PS3	Game released for PlayStation 3	0.18		0-1
PSP	Game released for PlayStation Portable	0.17		0-1
X360	Game released for Xbox 360	0.25		0-1
Genre				
Action	Dominant genre is action	0.61		0-1
Adventure	Dominant genre is adventure	0.08		0-1
Educational	Dominant genre is education	0.01		0-1
Racing	Dominant genre is racing	0.12		0-1
RPG	Dominant genre is role playing	0.12		0-1
Simulation	Dominant genre is simulation	0.11		0-1
Sports	Dominant genre is sports	0.19		0-1
Strategy	Dominant genre is strategy	0.13		0-1
Publisher				

X2K	Game is published by 2K Games	0.04	0-1
Activision	Game is published by Activision	0.08	0-1
Atari	Game is published by Atari	0.02	0-1
Capcom	Game is published by Capcom	0.03	0-1
Disney	Game is published by Disney	0.01	0-1
Eidos	Game is published by Eidos	0.02	0-1
EA	Game is published by EA	0.13	0-1
Konami	Game is published by Konami	0.04	0-1
Microsoft	Game is published by Microsoft	0.02	0-1
Midway	Game is published by Midway	0.02	0-1
Namco	Game is published by Namco	0.03	0-1
Nintendo	Game is published by Nintendo	0.06	0-1
Rockstar	Game is published by Rockstar	0.01	0-1
Sony	Game is published by Sony	0.05	0-1
Sega	Game is published by Sega	0.06	0-1
THQ	Game is published by THQ	0.06	0-1
SquareEnix	Game is published by Square Enix	0.03	0-1
Ubisoft	Game is published by Ubisoft	0.07	0-1
Pespective			
FirstPerson	Game is played from a first-person perspective	0.24	0-1
Platform	Game is a platformer	0.09	0-1
Isometric	Game is played from an isometric perspective	0.05	0-1
SideScrolling	Game is played from a side-scrolling perspective	0.08	0-1
TopDown	Game is played from a top-down perspective	0.12	0-1
ThirdPerson	Game is played from a third-person perspective	0.75	0-1

Table 2. Result of Type 3 Analysis of Variance

Variable	Sum of Squares	DF	F	p value	
(Intercept)	50.53	1	65.972	<.001	***
Sequel	0.51	2	0.331	0.719	
Re.release	0	1	0.003	0.954	
Review	74.54	1	97.321	<.001	***
MaxPlayers	3.66	1	4.781	0.029	*
Online	0.23	1	0.300	0.584	
Licensed	0.1	1	0.131	0.717	
Accessory	6.86	1	8.963	0.003	**
Multiplatform	1.59	1	2.076	0.150	
Rating	4.43	2	2.893	0.056	
Console	14.6	4	4.765	<.001	***
Publisher	52.53	17	4.034	<.001	***

Perspective	1.9	5	0.495	0.780	
Genre	16.18	6	3.521	0.002	**
Year	3.25	6	0.708	0.643	
Residuals	337.01	440			

*** p value is lower than .001
** p value is lower than .01
* p value is lower than .05

Table 3. Multiple Comparisons with Tukey's Post-Hoc Type-I Error Adjustment

Contrast	Estimate	Standard Error	DF	t value	p value	
Platform						
Wii-PSP	0.713	0.172	440.000	4.149	<.001	***
PSP-X360	-0.427	0.150	440.000	-2.841	0.038	*
Genre						
Simulation-Strategy	0.972	0.278	440.000	3.496	0.009	**
Publisher						
X2K-Nintendo	-1.069	0.286	440.000	-3.739	0.024	*
Activision-Namco	0.908	0.242	440.000	3.747	0.023	*
Activision-Sega	0.702	0.194	440.000	3.610	0.036	*
Atari-Nintendo	-1.305	0.353	440.000	-3.696	0.027	*
Eidos-Nintendo	-1.274	0.328	440.000	-3.883	0.014	*
Konami-Nintendo	-1.320	0.284	440.000	-4.640	0.001	***
Microsoft-Namco	1.206	0.325	440.000	3.712	0.026	*
Namco-Nintendo	-1.405	0.281	440.000	-4.995	0.000	***
Nintendo-Sega	1.199	0.243	440.000	4.935	0.000	***
Nintendo-Ubisoft	1.049	0.264	440.000	3.972	0.010	**

*** p value is lower than .001
** p value is lower than .01
* p value is lower than .05

Table 4. Coefficient Estimates of Linear Regression Models

Variable	Full Model	Specified Model	Full Model without VIF > 5	Stepwise Model
(Intercept)	-4.016	-3.862	-4.309	-4.175
Y05	-0.146	-0.154		
Y06	-0.361	-0.358		-0.131
Y07	-0.26	-0.258		
Y08	-0.264	-0.281		
Y09	-0.395	-0.427		-0.154
Y10	-0.602	-0.659	-0.29	-0.368

Sequel	0.144	0.149	0.142	0.146
ReRelease	0.042		0.037	
Review	0.033	0.034	0.033	0.034
RatingT	0.037		0.04	
RatingM	0.258	0.253	0.253	0.223
MaxPlayers	0.014		0.015	
Online	0.064		0.073	
Licensed	0.017		0.012	
Accessory	-0.077		-0.074	
Multiplatform	0.177	0.192	0.173	0.161
NDS	-0.279	-0.415	-0.226	-0.36
Wii	0.141		0.15	
PS3	-0.121	-0.157	-0.125	-0.167
PSP	-0.617	-0.717	-0.584	-0.655
Adventure	-0.239	-0.254	-0.247	-0.256
Educational	0.847	0.861	0.831	0.817
Racing	-0.014		-0.004	
RPG	0.05		0.03	
Simulation	-0.017		-0.024	
Sports	-0.259	-0.237	-0.246	-0.262
Strategy	-0.35	-0.354	-0.355	-0.356
Platform	0.246	0.164	0.261	0.241
Isometric	-0.077		-0.073	
SideScrolling	-0.172		-0.179	-0.165
TopDown	-0.195	-0.192	-0.197	-0.177
ThirdPerson	0.118	0.122	0.107	0.126
X2K	0.239		0.254	0.243
Activision	0.595	0.546	0.598	0.627
Atari	-0.076		-0.074	
Capcom	0.356	0.315	0.361	0.385
Disney	0.754	0.747	0.769	0.767
Eidos	-0.275	-0.345	-0.279	-0.286
EA	0.62	0.556	0.642	0.645
Konami	-0.155		-0.142	
Microsoft	0.855	0.773	0.88	0.841
Midway	0.325		0.338	0.362
Namco	-0.019		0.005	
Nintendo	1.5	1.49	1.512	1.548
Rockstar	0.955	0.901	0.983	0.972
Sony	0.754	0.716	0.774	0.773
Sega	0.323	0.259	0.342	0.337
THQ	0.45	0.385	0.452	0.462
SquareEnix	0.148		0.16	
Ubisoft	0.198		0.213	0.225

Table 5. RMSE of 5-Fold Cross Validation in Each Linear Regression Models

	Regression Models				**Random Tree Model**
	Full Model	Specified Model	Full Model Without VIF > 5	Stepwise Regression Model	Random Tree (# of tree=182)
RMSE	0.964	0.958	.962	0.955	.885