Best Indicator of Video Game Sales in North America

Wai Kin Chan, Cindy Huang

BAIS:9100

December 2nd, 2020

Executive Summary:

The goal of our project is to predict which independent variable best predicts North America's video game sales. Because the goal is to figure out which independent variable is the best predictor of North American video game sales, this makes our dependent variable NA sales. The independent variables in question are: platform, genre, publisher, global sales, EU sales, or JP sales. To achieve this goal, we focused on the different video game trends throughout the years. By analyzing the historical data, we were able to use the information provided and found that JP sales and video game platforms were the best indicators to predict North America's video game sales.

To analyze which independent variable best predicts North America's video game sales, we have a dataset (https://www.kaggle.com/gregorut/videogamesales) that contains a list of video games with sales greater than 100,000 copies. This dataset was generated by a scrape of vgchatz.com and has a total of 16, 598 records. We took a sample of that and created a new dataset with 1200 records. The dataset has all the information about the video games, their sale numbers in various regions, and their ranking of overall sales in order for us to draw specific conclusions on our analysis. Variables included in the dataset include the ranking of overall sales, the platform of the games release (i.e. PC,PS4, etc.), the genre of the game, the publisher of the game, sales in North America (in millions), sales in Europe (in millions), sales in Japan (in millions), sales in the rest of the world (in millions), and total worldwide sales.

In order to answer our question and analyze the dataset, we first restructured the data by taking a sample of 1200 records out of the 16,598 total records. We then cleaned the data by taking out any null values and any irrelevant variables. Next, we coded the Platform, Genre, and Publisher to numerical values, Platform ID, Genre ID, and Publisher ID, respectively. After completing the maintenance, we then made sure the dataset satisfied all the assumptions for multilinear regressions by running an initial regression of the data and looking at the residual and normal probability plot. After running a regression, we created a correlation method to We then initiated a backwards stepwise regression methodology to remove variables one at a time based on highest p-value > .05. We ran a total of 4 regression models before coming up with the best model.

From our initial evaluation, the regression model satisfied all the regression assumptions. In the original dataset, we found correlation between EU sales, global sales, and North American sales. To fix this, we took out those variables and ran a new regression. After making sure all assumptions were satisfied, we proceeded to run the regression model. One limitation to our findings was that the dataset did not take into account user data, in order to find out user satisfaction, which can affect future sales. The final result determined that Japan sales and Platforms were the best indicators in predicting North American sales. The equation representing our findings is: North American Sales = 0.404 + 1.081(JP Sales) + 0.054(Platform).

Modeling Steps

1. Took a 1200 record sample set out of the 16,598 total records.
2. Restructured sample dataset to transform categorical data to numerical values for the Platform, Genre, and Publisher variables. To do this, we created a sorting table with the categorical variables on the left and the unique index number to represent each of them. We then used the VLOOKUP formula to match the variable to the index and created new columns called Platform ID, Genre ID, and Publisher ID, respectively. This was needed because multilinear regression needs numerical data, not categorical data.
3. Cleaned the data by removing any null values from the dataset.
4. Checked to make sure all assumptions were satisfied:
    a. The first assumption is to make sure the model is linear. To check that the model is linear, we ran a multiple linear regression (Figure 1) to see if the residual plot (Figure 1.1) was linear. We observed from the residual plot that the model is indeed linear.
    b. The second assumption to check the error has a population mean of 0. To check this, we concluded the y-intercept of the model, which is included in the model.
    c. The third assumption is to make sure independent variables are uncorrelated with the error. To check this, we observed the residual plot (Figure 1.1) from the initial multiple linear regression. We found that the residual plot was random. This shows that the independent variable is not dependent on the dependent variable.
    d. The fourth assumption is to make sure the error term is normally distributed. To check this, we generated a normal probability plot (Figure 1.2) to show that it follows a normal distribution. We found that the error term does follow a normal distribution.
    e. The fifth assumption is that there is no autocorrelation between variables. To check this, we ran a correlation matrix (Figure 1.3). We found that European Sales (EU sales) and Global Sales are highly correlated to North American sales (NA sales). Because they are correlated, we then took EU sales and Global sales out of the model.
    f. The sixth assumption is to make sure the error term has constant variance, or has no heteroscedasticity. After removing the two independent variables, we ran a second regression (Figure 2) and found that the residual plots produced were homoscedastic.
    g. The seventh assumption and final assumption is that the independent variable has no multicollinearity. To find this, we ran another correlation matrix (Figure 2.1), without European Sales and Global sales. The results showed that all the independent variables are not correlated with each other.
5. After satisfying all assumptions, we began the first iteration of the modeling phase by implementing a backward stepwise methodology. We looked at the p-value of the second

regression (Figure 2) model and found the publisher has the highest p-value at 0.4974. We removed this variable and ran another regression.

6. We repeated step 4 with the third regression model (Figure 3). We found that genre had a high p-value of 0.4055. We removed said variable and ran a fourth regression model.
7. Repeating step 4 again and running a fourth regression model (Figure 4), we found that JP sales and video game platforms were the best indicators for estimating North American video game sales.
8. The final equation is: North American Sales = 0.404 + 1.081(JP Sales) + 0.054(Platform)

Appendix:

Figure 1: Summary Output of First Regression Model

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.98499965
R Square	0.97022431
Adjusted R S	0.97007444
Standard Er	0.42152161
Observation	1199

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	6	6901.22702	1150.2045	6473.44373	0
Residual	1192	211.795116	0.17768047		
Total	1198	7113.02214			

	Coefficients	*Standard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-0.0219636	0.04591656	-0.4783375	0.63249772	-0.1120499	0.06812266	-0.1120499	0.06812266
Genre ID	0.00257196	0.00388977	0.66121053	0.50860509	-0.0050596	0.01020351	-0.0050596	0.01020351
EU_Sales	-1.0415091	0.01713175	-60.794084	0	-1.0751209	-1.0078974	-1.0751209	-1.0078974
JP_Sales	-0.828554	0.01723349	-48.078128	2.459E-281	-0.8623654	-0.7947427	-0.8623654	-0.7947427
Global_Sale	0.90316348	0.00654729	137.944653	0	0.89031798	0.91600897	0.89031798	0.91600897
Platform ID	-0.0003472	0.00164004	-0.2116981	0.83237878	-0.0035649	0.00287049	-0.0035649	0.00287049
Publisher ID	-0.0002876	0.0007033	-0.4088842	0.68269806	-0.0016674	0.00109228	-0.0016674	0.00109228

Figure 1.1: Residual Plot from First Regression Model

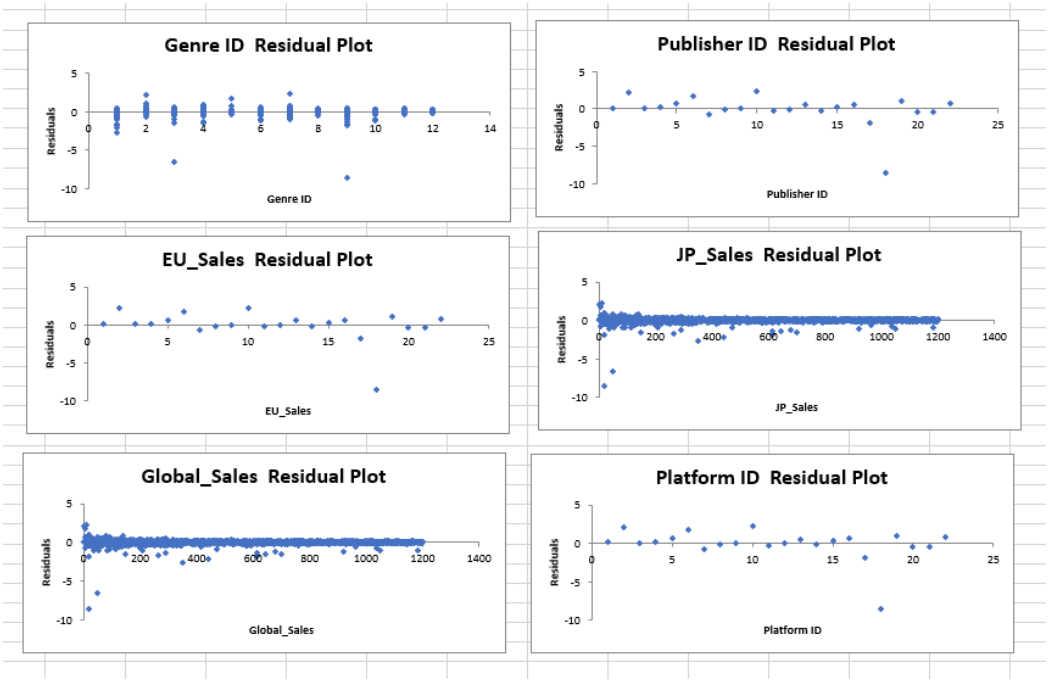

Figure 1.2: Normal Probability Plot of First Regression Model

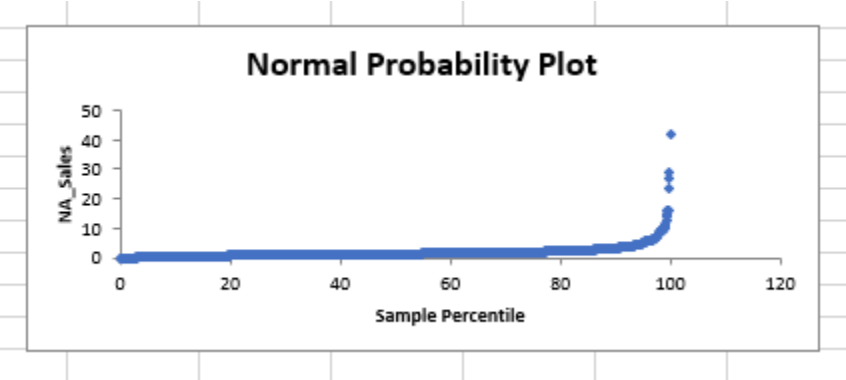

Figure 1.3: Correlation Matrix with All Variables

1st correlation matrix

	Genre ID	EU_Sales	JP_Sales	Global_Sales	Platform ID	Publisher ID	NA_Sales
Genre ID	1						
EU_Sales	-0.0269206	1					
JP_Sales	-0.1368969	0.3568517	1				
Global_Sales	-0.0607626	0.8685111	0.5627443	1			
Platform ID	0.0079916	0.043	-0.2909264	-0.0098818	1		
Publisher ID	0.0823938	0.0127261	0.1149413	0.0424409	-0.0618403	1	
NA_Sales	-0.0357565	0.6847922	0.3787771	0.9215635	0.0504899	0.023043	1

Figure 2: Summary Output of Second Regression Model (No EU_Sales and Global_Sales)

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.4153056
R Square	0.17247874
Adjusted R Sq	0.16970647
Standard Erro	2.22031451
Observations	1199

ANOVA

	df	SS	MS	F	Significance F
Regression	4	1226.84509	306.711271	62.2158074	8.5323E-48
Residual	1194	5886.17705	4.92979653		
Total	1198	7113.02214			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	0.37727166	0.24147806	1.5623434	0.1184721	-0.0964969	0.85104022	-0.0964969	0.85104022
Genre ID	0.01832846	0.02047807	0.8950288	0.37095208	-0.02184855	0.05850547	-0.02184855	0.05850547
JP_Sales	1.09511644	0.07038833	15.5582114	7.729E-50	0.95701787	1.23321501	0.95701787	1.23321501
Platform ID	0.05408875	0.00847311	6.38357467	2.4712E-10	0.0374649	0.0707126	0.0374649	0.0707126
Publisher ID	-0.00251371	0.00370301	-0.67883009	0.49737708	-0.00977884	0.00475141	-0.00977884	0.00475141

Figure 2.1:Correlation Matrix without EU_Sales and Global_Sales

2nd correlation matrix

	Genre ID	JP_Sales	Platform ID	Publisher ID	NA_Sales
Genre ID	1				
JP_Sales	-0.1368969	1			
Platform ID	0.0079916	-0.2909264	1		
Publisher ID	0.0823938	0.1149413	-0.0618403	1	
NA_Sales	-0.0357565	0.3787771	0.0504899	0.023043	1

Figure 3: Summary Output of Third Regression Model (Publisher variable taken out)

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.41492092
R Square	0.17215937
Adjusted R Sc	0.17008111
Standard Erro	2.21981355
Observations	1199

ANOVA

	df	SS	MS	F	Significance F
Regression	3	1224.57338	408.191128	82.8381834	1.0809E-48
Residual	1195	5888.44875	4.92757218		
Total	1198	7113.02214			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	0.30067713	0.21344735	1.40867116	0.15919256	-0.1180961	0.71945041	-0.1180961	0.71945041
Genre ID	0.01695477	0.02037324	0.83220791	0.40545787	-0.0230165	0.05692608	-0.0230165	0.05692608
JP_Sales	1.08963383	0.06990765	15.586761	5.2899E-50	0.95247844	1.22678923	0.95247844	1.22678923
Platform ID	0.05424221	0.00846819	6.40541167	2.1517E-10	0.03762805	0.07085638	0.03762805	0.07085638

Figure 4: Summary Output of Fourth and Best Regression Model

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.4143424
R Square	0.1716796
Adjusted R S	0.1702944
Standard Err	2.2195282
Observation	1199

ANOVA

	df	SS	MS	F	Significance F
Regression	2	1221.1607	610.58035	123.94285	1.21E-49
Residual	1196	5891.8614	4.9263056		
Total	1198	7113.0221			

	Coefficients	tandard Erro	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	0.4041123	0.1735051	2.3291091	0.0200192	0.0637041	0.7445205	0.0637041	0.7445205
JP_Sales	1.0814505	0.0692037	15.627057	3.105E-50	0.9456763	1.2172247	0.9456763	1.2172247
Platform ID	0.0540055	0.0084623	6.3818784	2.496E-10	0.0374028	0.0706081	0.0374028	0.0706081