Team:Manipogo

Group Members:	Richard Lee	rlee106@gmail.com
	Deannia Lucas	deannialucas80@gmail.com
	Caglar Cinar	cag.cin@gmail.com
	Hershel Don	hdon@mail.depaul.edu
	Hima Spandana Barla	spandanabarla26@gmail.com

Topic: Video Games

Dataset : Video Games Sales

We set out to identify factors that could predict video game sales. Our data set included several variables related to the type of game and its critical & user receptions, as well as the sales data broken down by regions, which we used as the dependent variables in our analysis.

We created several multiple regression models to try and predict sales, with different models for each respective sales region and globally. Our sales regions were broken down into North America, Europe, Japan, other regions, and overall globally. Our models included numeric quantities as well as categorical dummy variables, and transformed variables, as we created them by considering several kinds of terms introduced in the course. Models were constructed with consideration to avoid multicollinearity.

What we discovered is that the transformation used to log our dependent variable greatly increased our r-squared values. Interaction terms did not make any relevant change or difference to our models. The most effective interaction term was between the genre and platform which contributed to a value of over 0.01 to the adjusted r-squared value. Second order terms were also attempted but none were successful enough to make a difference on our models, there were no changes in our r-squared values. Overall this was the sum of our highlights and finds that made an impact on our models.

In the NA sales, a log transformation greatly boosted the accuracy of our models adjusted r-square value. It more than doubled the value in our first order model. Some adjustments made that were attempted were adding second order models and interaction models. Unfortunately these variables did not make much of a difference of any at all. The most relevant in the NA sales section happened to be the second order model of the squaring the critic score and critic user value. By potentially increasing the number of critics, it was hypothesized that the model would increase in correlation to the volume of critics and scores but unfortunately it didn't impact the model the way it was hypothesized, it made a slight difference in the r-squared value by boosting it by a mere value of 0.02. Other than these highlights, the log transformation made the biggest difference. With a heteroscedastic plot in the model, the log transformation was able to show a better pattern by being transformed, but ultimately our data set was not preferable for being used in a linear regression model. Either additions of variables or using a different model would give better results for our particular data set.

For Japan Sales, we didn't find a very concerning multicollinearity in our models. User_Score and Credit_Score was somewhat correlated but it wasn't too concerning with the 0.58 correlation. This probably wasn't significant enough to remove one of the variables from the model but the p-value for Critic_Score became bad after doing a log transformation on the User_Count so we removed it. Log transformation made sense because the number grew very large for some games and the count difference between a game that's rated by 500 users and a game that's rated by 1000 users does not have the same impact when we are comparing it against 8500-9000. The Scale Location plot showed that our data was very heteroscedastic with a very distinguishable "V" pattern. Doing a log transformation on the response variable (JP_Sales) helped to alleviate the degree of it and made the graph look a little bit more spread. Doing a log transformation on User_Count and the response variable helped a lot, almost doubling the adjusted R-squared score. The interaction terms didn't add much. Two biggest contributors were Platform:UserCount and Platform:Genre which added about 0.035 each. There wasn't any second order term that had a significant effect on the model. One major concern I had with our data and models is that it looks like there is a serial correlation in our data and no matter what variables I chose, I couldn't get rid of it. I've tried a couple methods and transformations to resolve it but none was successful.

While performing the regression model on EU_Sales,we used critic score, critic count, user score and user count as the dependent variables, it passed the t-test and f-test,but the adjusted R-square was very low 0.1215. Created two second order terms and one interaction term using the variables and tried to predict the response variables,but that could not contribute much to the model.After applying the log on EU_Sales the adjusted R-square value slightly increased to 0.3805, but did not make any significant impact on the model. Performed Pseudo Random Number Generators for the model, we set the seed and generated a random number, this is to reproduce our experiments. Built a model for test and train dataset, model is built entirely on trained dataset, but is evaluated on the test data. Removed a few variables from the dataset like X,name and platform columns. Approximately 80% of these are 1's and aprroximately 20% of these are 2's. Performed multicollinearity by pruning some of the variables and obtained VIF values of critic count and user count and critic values under 10.

For the other unspecified sales regions, we constructed a first order model that predicts the log of the millions of titles sold. The log transformation of the response variable was used in response to a pattern identified in the residual plot for the plain first-order multilinear model. Creating the model that predicted the log of the sales increased the adjusted R-squared of the model from 0.0935 to 0.1805, while maintaining significance for the t-test on each beta as well as the F-test for the model. However, there is still a notable pattern in the residual plot, indicating a violation of the assumption of homoscedasticity. The independent variables in the model are critic scores, critic count, user scores, and user count. The model was trained on a randomly selected 20% of the data and tested on the other 80%. The validation on the testing data indicates that the model is good in terms of generalization and does not overfit, despite it being a very weak model, that is not seemingly useful.

My final selected explanatory variables are critic_score, critic_count, user_score and user_count used to predict Global Sales. Before my log transformation I was getting 12% for the adjusted R square with excellent P-values. After the log transformation my adjusted r square went up to 22%. There was no sign of multicollinearity after using the VIF to check the model for multicollinearity shows that there is no indication of multicollinearity because all the values are below 10. The sum of the residuals of -1.836101e-15 on the model looks good because it can be rounded up to zero. The histogram of the residuals shows a normal distribution curve; the peak and tails are showing which depicts a normal distribution of the data. There are no outliers, all data seems to be in the data frame and 95% of the residuals are within 2 standard deviations of the mean. During the feature selection none of the four selected variables was removed as the variables were important to the model prediction.

In conclusion, it is difficult to predict with much certainty how well a video game will sell based on the variables we had available to consider, with the best models still having very low predictive value. All of the models we created were similarly limited in their value, indicating a need for different streams of data or possibly a more sophisticated modelling approach