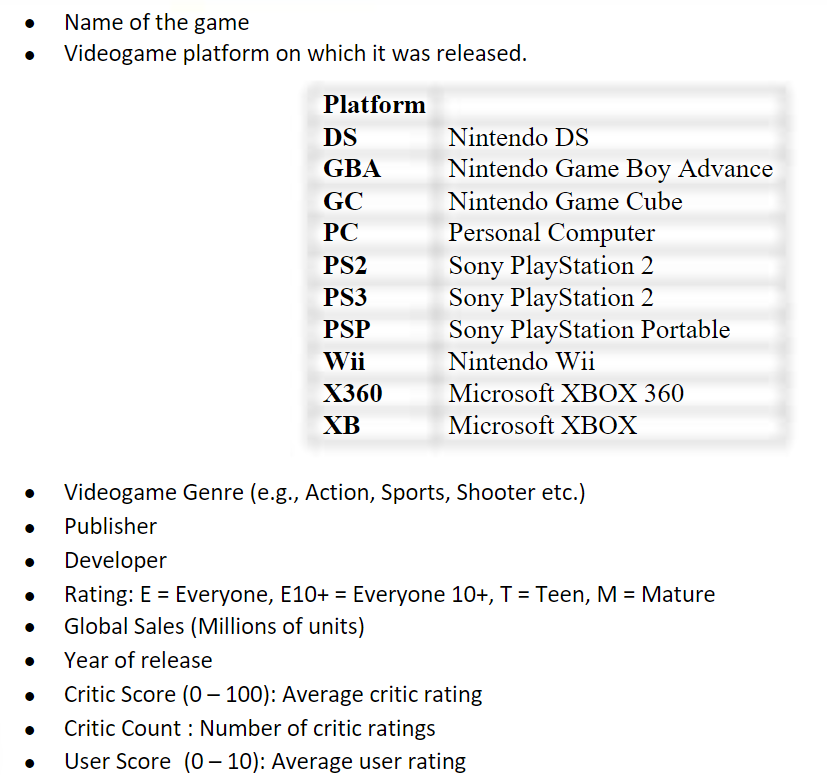
**BUAN 6337 Case Study**

The dataset for this exercise is available in VideoGamesSales\_Main.7bdat. This dataset contains information on the global sales and critic and user review ratings for videogames launched between 2001 and 2012 (from).

The variables are:



**Question 1**

Develop a regression model that links global sales to video game reviews. Explore ways in which

the model fit could be improved through suitable changes to the model specification and variables.

1. Present the final model and results.
2. Explain how you developed your model (what was your initial model, what were the key variations you tried and how did you arrive at the final model – and the thought process behind these steps).
3. Interpret the model results.

**Answer**

**a)**

Our final model used user and critic scores, user and critic counts, and platforms as the independent variables and global sales as the dependent variable. We saw that this model passed the P-value test and most variables were significant. However, it still struggled with the R-square and we believe that was due to regression assumptions being violated. We did observe R-square increasing as we made changes to the model (from .05 to .16). We have further improvised the model fit by addressing the skewness of the dependent variable in Q2 which has increased the Adj R-square to 0.4237

**Final Regression Model:**

Table

Description automatically generated

Table

Description automatically generated

**b)**

Our first model was a basic regression analysis that considered Critic\_Score and User\_Score (the independent variables) on Global Sales (the dependent variable). The resulting model passed the P-value test which rejects the null that the variances are 0 and can conclude that the model is significant, but R-square looked to be weak (we want to see a number closer to 1) which implies it does not explain the variation in the data very well. The P-Values of the parameter estimates also look good stating that the variables are statistically significant with each other and do have an impact on global sales. However, we can see that the User Score variable causes a decrease in global sales for every 1-point increase in user score, and we wanted to explore this further by adding more variables as this does not make sense.

**Regression Model with Critic Score and User Score as Independent Variables:**

Table

Description automatically generated

The second variation included User\_Count and Critic\_Count as additional independent variables. It seems this strengthened the original outcomes of the prior model, but R-Square was still relatively weak and User\_Score was now not showing no relationship based on the P-value and still negative, so we wanted to try a variation of other variables in attempt 3 to see if these significantly contributed to revenue and would smooth out the negative impact of user score. It is interesting to note that this model says User\_Score does not have an impact on global sales and User\_Count has minor impact on global sales. It is appearing Critic\_Score and Critic\_Count are influence global sales where as User\_Score and User\_Count do not or have much less of an impact.

**Regression Model Including Critic\_Count & User\_Count:**

Table

Description automatically generated

We then compared just Critic\_Score and Critic\_Count and see both variables pass the P-value test but R-square is still pretty low.

**Regression model with just Critic\_Score and Critic\_Count:**

Table

Description automatically generated

For the third variation we thought the platform could have an effect on sales and wanted to compare these different variables. Xbox (XB) is the variable that was left off (similar to Q4 in our lecture). R-square has improved, and we can see all models had higher sales than Xbox besides PC. The P-value for Game Cube, PSP, and Xbox 360 were above the rejection region of .05 which means sales from these platforms are not significantly different than Xbox and all other platform sales are. We also concluded that this would be our final model as we observed outliers and many residuals and wanted to focus on resolving issues in order to build a more accurate model supporting global sales. We anticipate that R-square will improve once changes are made.

**Final Regression Model:**

Table

Description automatically generated

Table

Description automatically generated

We did perform testing for non-linear effects of independent variables ie., user/critic score and count but did not perceive significant changes in the r-square or to the model overall. We have further improvised the model fit by addressing the skewness of the dependent variable in Q2 which has increased the Adj R-square to 0.4237

**C)**

**Interpreting the results:** Our model fell within the rejection region of the P-Value so we could state it as statistically significant. It did have a low R-Square value (.16 on a range of 0 to 1. 1 Being a strong representation of the variances) which we noted as an issue. We tested several variations, and this model had the strongest R-Square value. We also tested for non-linear effects among the key parameters and observed little change in R-Square. The key parameters Critic\_Score, User\_Score, Critic\_Count, and User\_Count all fell within the rejection region of the P-Value which meant the variables were statistically significant in changing global sales. It is notable that the Critic\_Score and Critic\_Count appeared to carry much more weight based on coefficient change than User\_Score and User\_Count. We further compared platforms which showed that the Wii sold the most and only the PC sold less than Xbox. Most of the platforms were within the rejection region confirming that they were a significantly different sales number than Xbox. Overall, we see the model to be significant, but it did have a high number of residuals which we believe the testing in part 2 will identify the source of issue and fixes will lead to a higher R-Square and a stronger model overall.

**Question 2**

For the final model you constructed in question 1, verify whether the various regression

assumptions discussed in the lecture are satisfied. If an assumption is violated, discuss how it can

be handled, and implement the same. Discuss whether this change had a practically significant

impact on your model results.

**Answer**

**A) Outliers**

We did observe numerous outliers (almost 10%!!!) and after applying the fix, the variable were all considered significant.

It can be seen that most observations fall within the predicted range, but several residuals fall way beyond the predicted range (graphs 1 and 2). Looking at Cooks D plot it can also be observed that many outliers exist. Outliers can either be handled by dropping observations that are beyond the CookD threshold or by running robust regression. We went for robust regression since it doesn’t involve any loss of information.

**Graphic Interpretation of Model:**

Graphical user interface, diagram, application

Description automatically generated

After running robust regression, we see that all variables are now significant, but R-Square is still considerably low and went down a few points.

**Robust Regression Image:**

Table

Description automatically generated

**B) Multi-collinearity**

In running a correlation matrix we can see primary variables (Critic\_Score, Critic\_Count, User\_Score, and User\_Count) are moderately correlated, some have low correlation. This moderate to low correlation hints that we do not have a multicollinearity problem.

**Correlation Matrix:**

Table

Description automatically generated

Furthermore, our variance inflation factors are all 1 or slightly higher which is very low (well below the threshold of 10), and we do not need to be worried about multicollinearity affecting our model. Additionally, our condition index is well below 10 which is a beginning indicator of multicollinearity with 100 or more being strong multicollinearity.

Table

Description automatically generated

**C)** **Heteroscedasticity**

In checking for heteroscedasticity we do observe that the parameter estimates are mostly significant with User Score being the only insignificant parameter so we may have an issue.

Table

Description automatically generated

Overall the results of the Whites test show that the regression assumptions are not satisfied by being able to reject the Null Hypothesis. This shows that we do have a Heteroscedasticity problem.

**Whites Test:**

Graphical user interface, text, application

Description automatically generated

In reviewing the first two graphs we can see that it does appear that the beginning residuals are expanding as you move further down the x-axis.

**Visual Check for Heteroscedasticity:**

Chart, scatter chart

Description automatically generated

Finally, the heteroscedasticity Consistent T-Values in all instances are quite different than the parameter T-Values, another indicator of heteroscedasticity.

**Parameter Estimates Chart:**

Table

Description automatically generated

To fix heteroscedasticity we can do so by transforming the dependent variable Global\_Sales using the log transform of y variable process. We then see that the T-Values are much closer though not exact, and the P-values are much better. Additionally, observing the first few residual graphs shows a wider distribution that does not resemble the triangles observed earlier. However, it does still appear to slightly shape from left to right in the opposite direction. the Whites test is still within the rejection region which indicates we are still being affected by heteroscedasticity. It is nice to note that our R-Square is improving with a score of .4254 which is a much healthier model than what we started with.

**Results of Log Transform of Y Variable:**

Graphical user interface, application, table

Description automatically generated

Graphical user interface, diagram

Description automatically generated

**D) Normality of error term**

In testing for normality of error term we see that in the Goodness of Fit Test in all of the tests the P-Value are significant. We reject the null hypothesis that the residuals follow the normal distribution and thus we do have a normality of error problem. The histogram also shows that the distribution is effected by outliers.

**Reviewing Residuals Information with Proc Univariate:**

Table

Description automatically generated

Graphical user interface

Description automatically generated

We run a Cooks D Test to remove the outliers and see a much more normal distribution. However, we still reject the null hypothesis due to P=Values falling in the rejection region for all 3 tests. This normality of error problem can also be commonly solved by A Log Transformation of the Y Variable which we performed earlier when checking for heteroscedasticity and saw a positive result.

**Reviewing Residuals Information with Proc Univariate after Removing High Cooks D Values:**

Table

Description automatically generated

Chart, histogram

Description automatically generated