Individual Milestone 1

Team Abolishes

Dawei Wang

Video game is a huge market which provides a lot of job opportunities. However, how to produce a best-selling game is always a question that market researchers want to answer. Game sales could be determined by so many factors, the review from critics may boost or jeopardize game sale at any time. The reputation, genre of game, and publication platform may influence the sale in some degree. To provide an insightful perspective for marketing people in the game development industry, answering the question at the very beginning is essential.

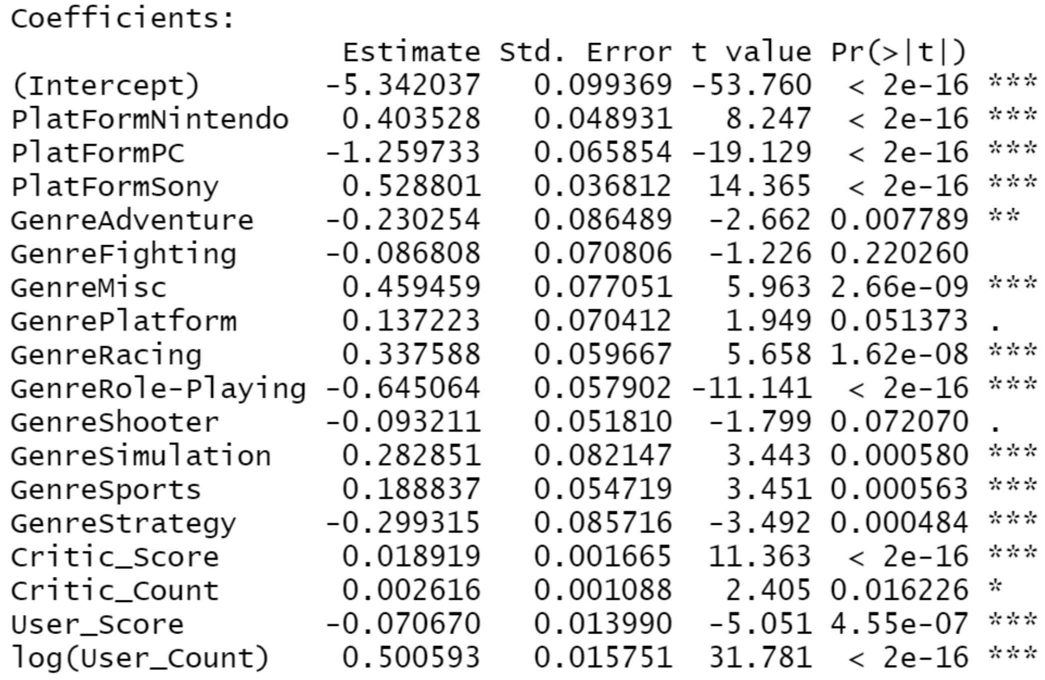
In this report, our group are trying to reveal the regression rule behind the global sales data of video game industry. The data covers most of the video games that published from 1980-2016. The European sales data will be analyzed by me, while all the other region sales and global sales data will be analyzed by my teammates. By focusing different regions, we hope the unique sales character could be evaluated and developer who wish to focus on certain region or global sales strategy could be enlighten by our work.

To have a direct-viewing impression of this European sales dataset, a linear regression model with all independent variables was made. After repeated adjustment, testing, and modeling, the final first order model is:

log(EU_Sales)

~PlatForm+Genre+Critic_Score+Critic_Count+User_Score+log(User_Count)

And the coefficients are:



```
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)       -5.342037   0.099369 -53.760  < 2e-16 ***
PlatFormNintendo   0.403528   0.048931   8.247  < 2e-16 ***
PlatFormPC        -1.259733   0.065854 -19.129  < 2e-16 ***
PlatFormSony       0.528801   0.036812  14.365  < 2e-16 ***
GenreAdventure    -0.230254   0.086489  -2.662 0.007789 ** 
GenreFighting     -0.086808   0.070806  -1.226 0.220260    
GenreMisc          0.459459   0.077051   5.963 2.66e-09 ***
GenrePlatform      0.137223   0.070412   1.949 0.051373 .  
GenreRacing        0.337588   0.059667   5.658 1.62e-08 ***
GenreRole-Playing -0.645064   0.057902 -11.141  < 2e-16 ***
GenreShooter      -0.093211   0.051810  -1.799 0.072070 .  
GenreSimulation    0.282851   0.082147   3.443 0.000580 ***
GenreSports        0.188837   0.054719   3.451 0.000563 ***
GenreStrategy     -0.299315   0.085716  -3.492 0.000484 ***
Critic_Score       0.018919   0.001665  11.363  < 2e-16 ***
Critic_Count       0.002616   0.001088   2.405 0.016226 *  
User_Score        -0.070670   0.013990  -5.051 4.55e-07 ***
log(User_Count)    0.500593   0.015751  31.781  < 2e-16 ***
```

The way I built this model was:
The first preview model shows that numeric independent variables show a strong linear relationship with euro sales, that includes critic score, critic count, user score and user count. In the meantime, part of the categorical variables has good p-value while a lot of them don't. The Adjusted R-squared is quite low at this time. The data transformation was tired at different variables. It shows when we log sales, it could improve our model by increasing adjusted R-squared from 0.16 to 0.34

To have a better understanding of data attributes, I check this sub-dataset again with the impression of p-value of different platforms. There are two things I've noticed. One is some of the sales equal to zero, and the other is there are too many platforms, and only contains limited observations. The sales data were present in two decimal places million dollars. In this case, the game sales with 0 value are either too small to display, or equal to 0. This may lead by didn't releases in Europe or only a few people know about this game. It makes these observations has no practical meaning for us to make the prediction. Thus, the observations that sales equal to 0 were removed. Considering the small quantity for some game platform. A new PlatForm variable was made, and the platforms were grouped by its' manufacturing enterprise. The platform was transferred in this way:
(Manufacture: Original Platforms)
Nintendo: 3DS DS GBA Wii WiiU
Sega: DC
Sony: PS PS2 PS3 PS4 PSP PSV
PC: PC
Microsoft: X360 XB XOne

For genres, the Puzzle games are only 75 observations in this dataset while other genre has hundreds of observations. The p-value of Puzzle is also quite high. Thus, this genre was removed from dataset. After these modifications, the size of dataset drops from 7017 obs to 5987 obs.

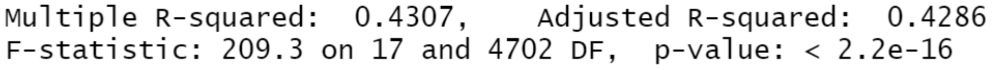
The dataset was divided to two a training set and a testing set with a ratio of 4:1. A model with log(EU_Sales) with PlatForm+ Genre+ Critic_Score + Critic_Count+ User_Score + log(User_Count) was built.

To remove outliers, I define a cook's distance greater than 5 times the mean as influential. There were 128 outliers and they were removed right after.

Build the model with the same ratio of training and testing data. Then use backward stepwise selection, the final model was made by now.

The final variables are: Numeric variables - critic score, critic count, user score, user count; categorical variable genre without Puzzle, and newly made categorical variable PlatForm.

In the coefficients table we can see, all p-values for these coefficients list are good except 3 genres are above 0.05. Considering the rest of them are great, we kept this variable. Platforms are all great. In numeric variables, only critic count is a little bit higher. But deletion of this variable only lead to adjusted R^2 reduce a bit, so, we are keeping this one.

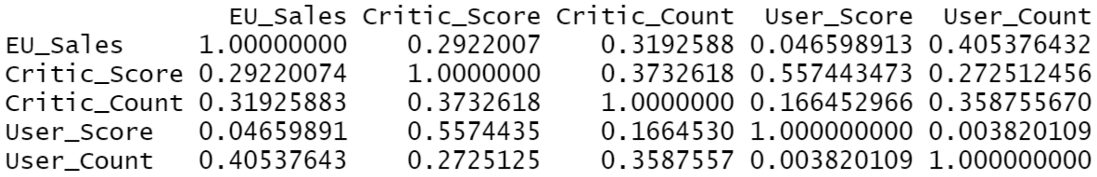


```
Multiple R-squared:  0.4307,    Adjusted R-squared:  0.4286
F-statistic: 209.3 on 17 and 4702 DF,  p-value: < 2.2e-16
```

The F-test of this model is great. This means we can reject H_0 and accept H_1 which means at least one $\beta \neq 0$. The adjusted R^2 means there are 42.86% could be explained by this model. RMSE of this mode is 0.94. This means about 95% of our observations or our predictions going to be within a 1.88 of the true value.

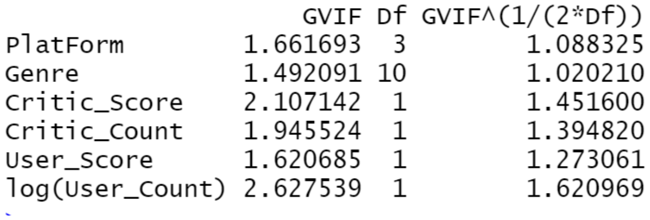
In the process of model selection, backward and forward were both executed. However, the forward one shows a very different result with a low adjusted R^2. It might because of the data transformation. I did it manually and select the best one while the computer uses a different approach which cannot find what I found.

Correlation Table:



```
               EU_Sales Critic_Score Critic_Count  User_Score  User_Count
EU_Sales     1.00000000    0.2922007    0.3192588 0.046598913 0.405376432
Critic_Score 0.29220074    1.0000000    0.3732618 0.557443473 0.272512456
Critic_Count 0.31925883    0.3732618    1.0000000 0.166452966 0.358755670
User_Score   0.04659891    0.5574435    0.1664530 1.000000000 0.003820109
User_Count   0.40537643    0.2725125    0.3587557 0.003820109 1.000000000
```

VIF Table



```
                     GVIF Df GVIF^(1/(2*Df))
PlatForm         1.661693  3        1.088325
Genre            1.492091 10        1.020210
Critic_Score     2.107142  1        1.451600
Critic_Count     1.945524  1        1.394820
User_Score       1.620685  1        1.273061
log(User_Count)  2.627539  1        1.620969
```

Correlation and VIF both show there’s no concern of multicollinearity.

In sum, this model shows that, in European video game market, sales related to user and critic appraise and the number of these people. Besides, the publication platform and genre also affect sales.

For next milestone, I will discover more about the relationship between independent variables and refine this model by their interaction or other possible methods.