Video Game Sales Analysis

Boda Song, Bowen Gu, Yucheng Feng

Abstract

All the members of our group are interested in video games, so we decide to find out what factors might influence the global sales of video games. We created a linear model with sales in North America, sales in Japan, publisher of the game, the critic score of the game as well as the user score of the game. Then we performed several statistical methods including linear regression, model selection, ANOVA test and other methods to help our analysis. After getting the result from the statistical methods, we found out that the user score does not have a large influence on our final linear regression model, and the sales in North America, sales in Japan, publisher of the game and critic score have significant influence on the global sales of video games. Moreover, the publishers might have significantly different impacts in North America and in Japan, which is also a key finding from our analysis.

Introduction

In October 1958, the first video game was invented, which marks the start of an era. In the following years, video games have been constantly evolving and appear to be very popular especially in the 21th century. Now you can get millions of different kinds of video games on many platforms, and the players can always find a game that they definitely love. Given so many different kinds of video games, how can we determine which video games are popular all over the world? We decide to use the total global sales of video games as a way to determine the popularity of the game, and we want to find out what factors might influence the global sales of video games.

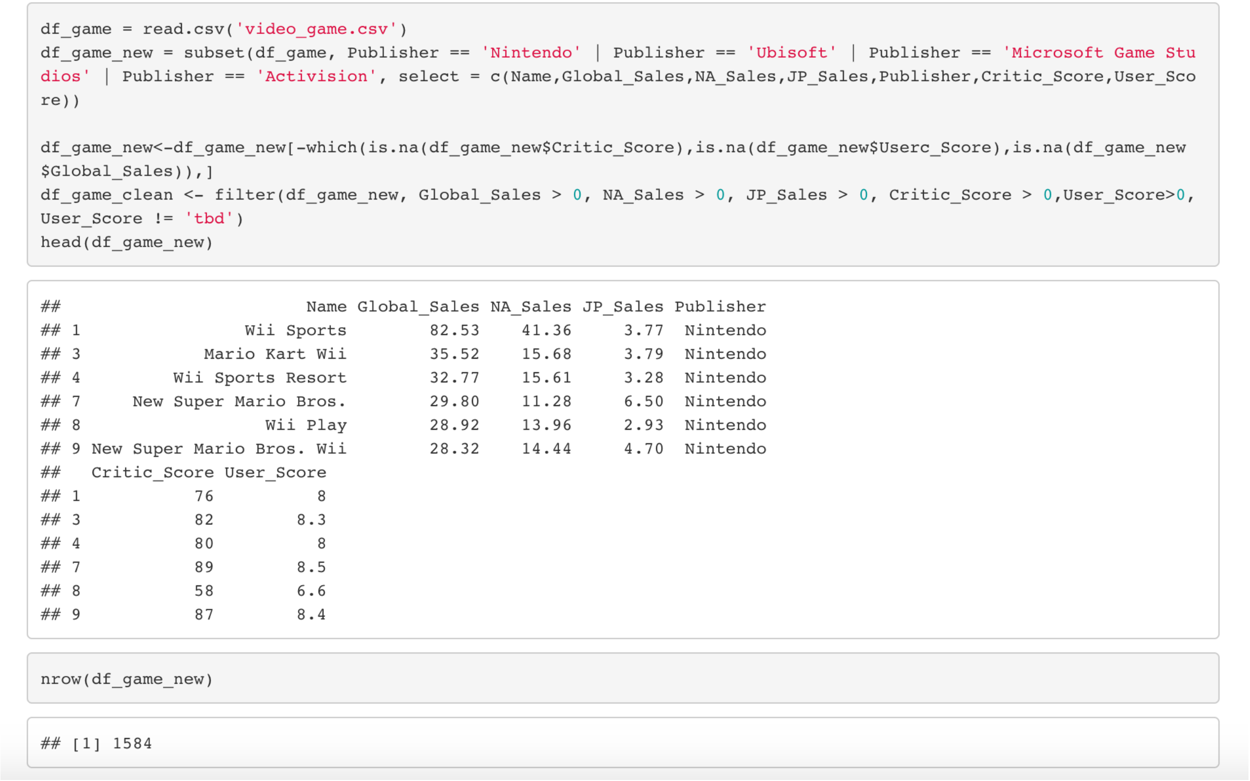
We downloaded our dataset from Kaggle.com and modified the dataset to fit our analysis. Our dataset contains 7 columns, which are video game name, global sales (in million), North America Sales (in million), Japan Sales(in million), publisher of the game, critic score of the game and user score of the game, and our dataset contains 462 rows, which means that there are 462 different video games in the dataset. We only included video games from the four largest publishers, which are Nintendo, Microsoft Game Studios, Activision, and Ubisoft, to ensure each company has enough video games. We want to find out how the North America Sales, Japan Sales, publisher, critic score, and user score influence the global sales of video games.

Results

The purpose of this project is to try to find a method to predict global video sales through a linear regression model. We can make this prediction if we have enough data about game sales from the past, along with other variables like local sales, publisher, user score, and so on. To achieve this goal, we carefully choose the first-order term. From the original data frame, there are many variables that could not help us evaluate the global sales, and also some categorical variables that will cause coefficients in other models to exceed the upper limit. In this case, we carefully select the predictor variables that seem to have a somewhat relationship with our response variables. Through our group discussion, we finally choose 5 first-order predictor variables, which include 4 quantitative variables and one categorical variable to predict our response variable, global sales. Four quantitative variables are game sales in North America (NA_Sales), game sales in Japan (JP_Sales), aggregate score compiled by Metacritic staff (Critic_Score), and score by Metacritic's subscribers (User_Score). One categorical variable is the publisher of the games (Publisher).

To prevent some outliers influence our model, we exclude rows that contain missing data from the original data frame. We also minimize the number of variables in our categorical variable to only contain four major companies, which are Nintendo, Ubisoft, Microsoft Game Studios, and Activision. The reason for us to choose those companies is that they are all well-known game publishers around the world, and they published many games in the past years, so we can have enough data to generate our linear model.

Below is a code block that shows the step to make data cleaning from the original data frame. As you see, it has 1584 rows after the cleaning, which is enough to give us a linear regression model.



```
df_game = read.csv('video_game.csv')
df_game_new = subset(df_game, Publisher == 'Nintendo' | Publisher == 'Ubisoft' | Publisher == 'Microsoft Game Stu
dios' | Publisher == 'Activision', select = c(Name,Global_Sales,NA_Sales,JP_Sales,Publisher,Critic_Score,User_Sco
re))

df_game_new<-df_game_new[-which(is.na(df_game_new$Critic_Score),is.na(df_game_new$Userc_Score),is.na(df_game_new
$Global_Sales)),]
df_game_clean <- filter(df_game_new, Global_Sales > 0, NA_Sales > 0, JP_Sales > 0, Critic_Score > 0,User_Score>0,
User_Score != 'tbd')
head(df_game_new)
```

```
##                         Name Global_Sales NA_Sales JP_Sales Publisher
## 1                 Wii Sports        82.53    41.36     3.77  Nintendo
## 3             Mario Kart Wii        35.52    15.68     3.79  Nintendo
## 4          Wii Sports Resort        32.77    15.61     3.28  Nintendo
## 7      New Super Mario Bros.        29.80    11.28     6.50  Nintendo
## 8                   Wii Play        28.92    13.96     2.93  Nintendo
## 9  New Super Mario Bros. Wii        28.32    14.44     4.70  Nintendo
##   Critic_Score User_Score
## 1           76          8
## 3           82        8.3
## 4           80          8
## 7           89        8.5
## 8           58        6.6
## 9           87        8.4
```

```
nrow(df_game_new)
```

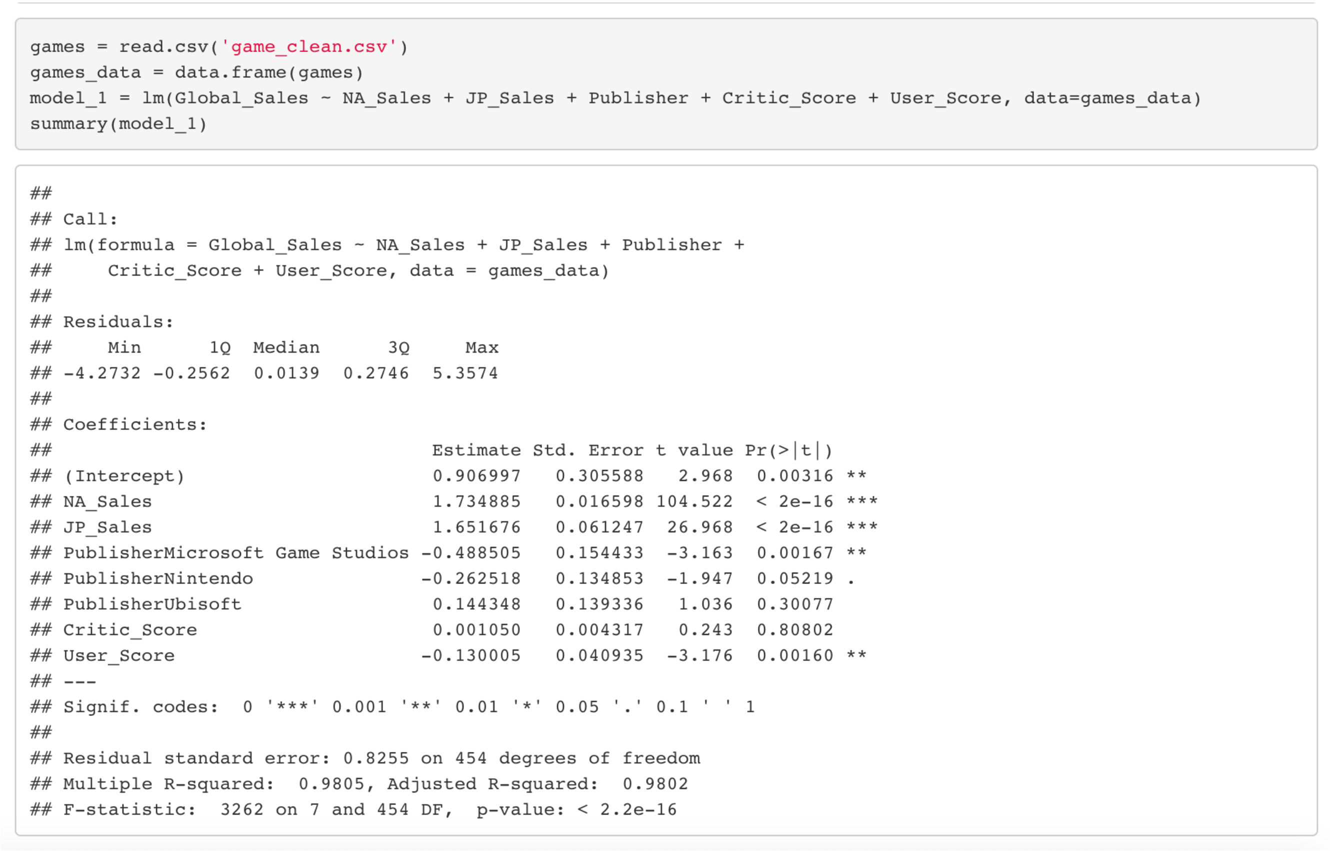
```
## [1] 1584
```

First Model

Then, we use this cleaned version of the data frame to make our linear model, the summary of this linear model is in the image below. This model only includes all the first-order terms.

We pick the quantitative variable NA_sales to make the interpretation. For every 1 million increase in NA_Sales, I expect the Global_Sales to increase by 1.7349 million on average, holding JP_Sales, Publisher, Critic_Score, and User_Score constant.

We pick the categorical variable Publisher and the baseline level is Activision. We can make an interpretation that if the publisher is Nintendo, I expect the intercept to decrease -0.262518 from the baseline Activision, which has an intercept of 0.906997, holding NA_sales, JP_sales, Critical_Score, and User_Score the same.

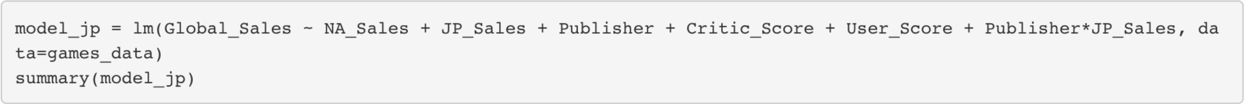


```
games = read.csv('game_clean.csv')
games_data = data.frame(games)
model_1 = lm(Global_Sales ~ NA_Sales + JP_Sales + Publisher + Critic_Score + User_Score, data=games_data)
summary(model_1)
```

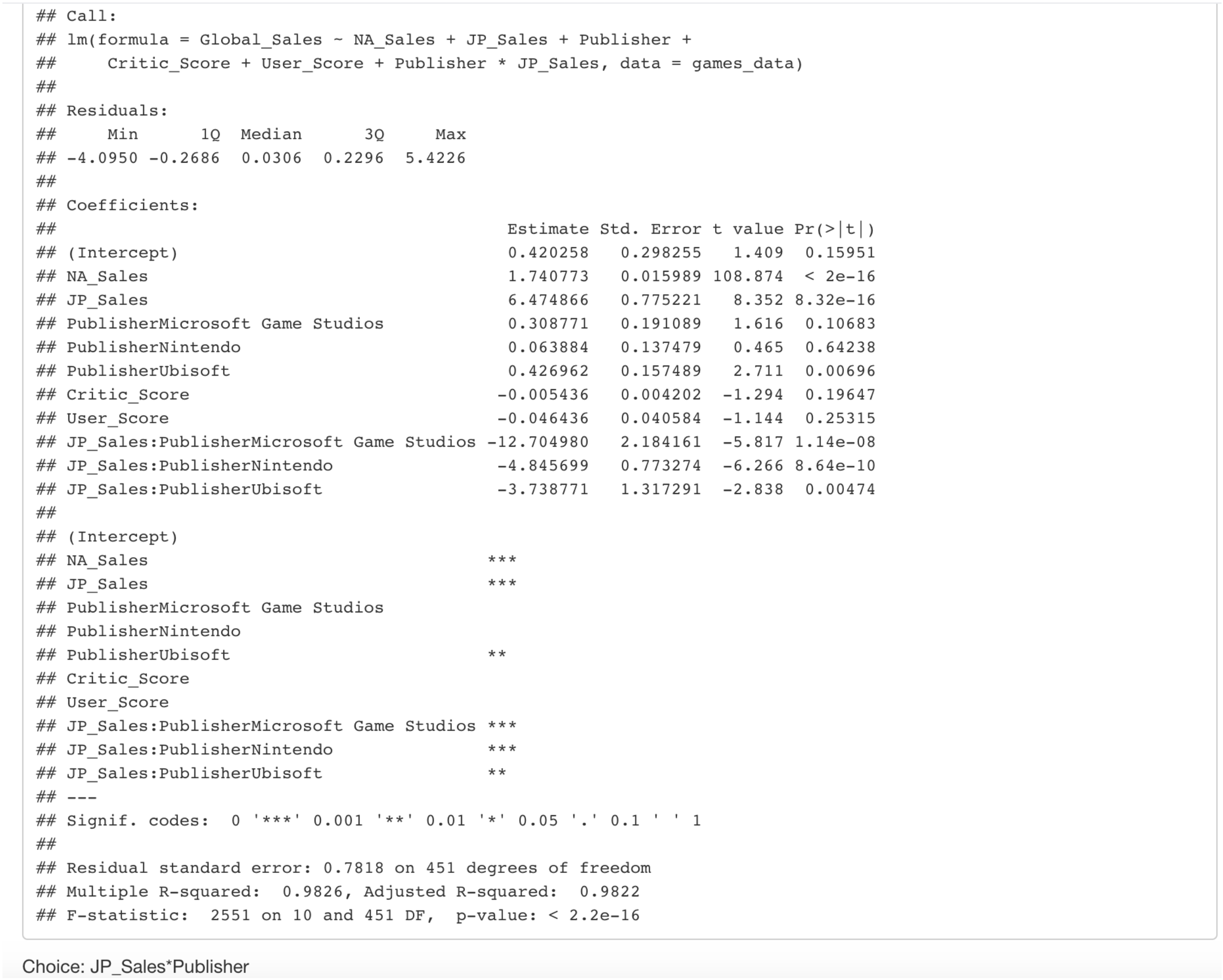
```
## 
## Call:
## lm(formula = Global_Sales ~ NA_Sales + JP_Sales + Publisher + 
##     Critic_Score + User_Score, data = games_data)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -4.2732 -0.2562  0.0139  0.2746  5.3574 
## 
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                      0.906997   0.305588   2.968  0.00316 ** 
## NA_Sales                         1.734885   0.016598 104.522  < 2e-16 ***
## JP_Sales                         1.651676   0.061247  26.968  < 2e-16 ***
## PublisherMicrosoft Game Studios -0.488505   0.154433  -3.163  0.00167 ** 
## PublisherNintendo               -0.262518   0.134853  -1.947  0.05219 .  
## PublisherUbisoft                 0.144348   0.139336   1.036  0.30077    
## Critic_Score                     0.001050   0.004317   0.243  0.80802    
## User_Score                      -0.130005   0.040935  -3.176  0.00160 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.8255 on 454 degrees of freedom
## Multiple R-squared:  0.9805, Adjusted R-squared:  0.9802 
## F-statistic:  3262 on 7 and 454 DF,  p-value: < 2.2e-16
```

Additional Variable

We choose to include an intersection term in our model, which is Publisher*JP_Sales. We choose this term because Japan is a small region compared to North American, and it would more likely to have some trends for buying its local publisher rather than companies far from other states. The interaction term could reflect this tendency. However, to our surprise, Nintendo doesn't incline in this interaction term heavily. Still, it does prove that company Microsoft Game Studio is not so welcomed compared with other companies in the Japanese market, because it has a negative slope.



```
model_jp = lm(Global_Sales ~ NA_Sales + JP_Sales + Publisher + Critic_Score + User_Score + Publisher*JP_Sales, da
ta=games_data)
summary(model_jp)
```

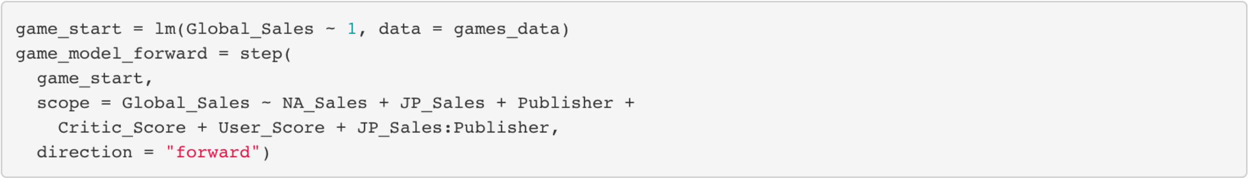


```
## Call:
## lm(formula = Global_Sales ~ NA_Sales + JP_Sales + Publisher + 
##     Critic_Score + User_Score + Publisher * JP_Sales, data = games_data)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -4.0950 -0.2686  0.0306  0.2296  5.4226 
## 
## Coefficients:
##                                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)                             0.420258   0.298255   1.409  0.15951
## NA_Sales                                1.740773   0.015989 108.874  < 2e-16
## JP_Sales                                6.474866   0.775221   8.352 8.32e-16
## PublisherMicrosoft Game Studios         0.308771   0.191089   1.616  0.10683
## PublisherNintendo                       0.063884   0.137479   0.465  0.64238
## PublisherUbisoft                        0.426962   0.157489   2.711  0.00696
## Critic_Score                           -0.005436   0.004202  -1.294  0.19647
## User_Score                             -0.046436   0.040584  -1.144  0.25315
## JP_Sales:PublisherMicrosoft Game Studios -12.704980   2.184161  -5.817 1.14e-08
## JP_Sales:PublisherNintendo             -4.845699   0.773274  -6.266 8.64e-10
## JP_Sales:PublisherUbisoft              -3.738771   1.317291  -2.838  0.00474
## 
## (Intercept)                              
## NA_Sales                                 ***
## JP_Sales                                 ***
## PublisherMicrosoft Game Studios          
## PublisherNintendo                        
## PublisherUbisoft                         ** 
## Critic_Score                             
## User_Score                               
## JP_Sales:PublisherMicrosoft Game Studios ***
## JP_Sales:PublisherNintendo               ***
## JP_Sales:PublisherUbisoft                ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.7818 on 451 degrees of freedom
## Multiple R-squared:  0.9826, Adjusted R-squared:  0.9822 
## F-statistic:  2551 on 10 and 451 DF,  p-value: < 2.2e-16
```

Choice: JP_Sales*Publisher

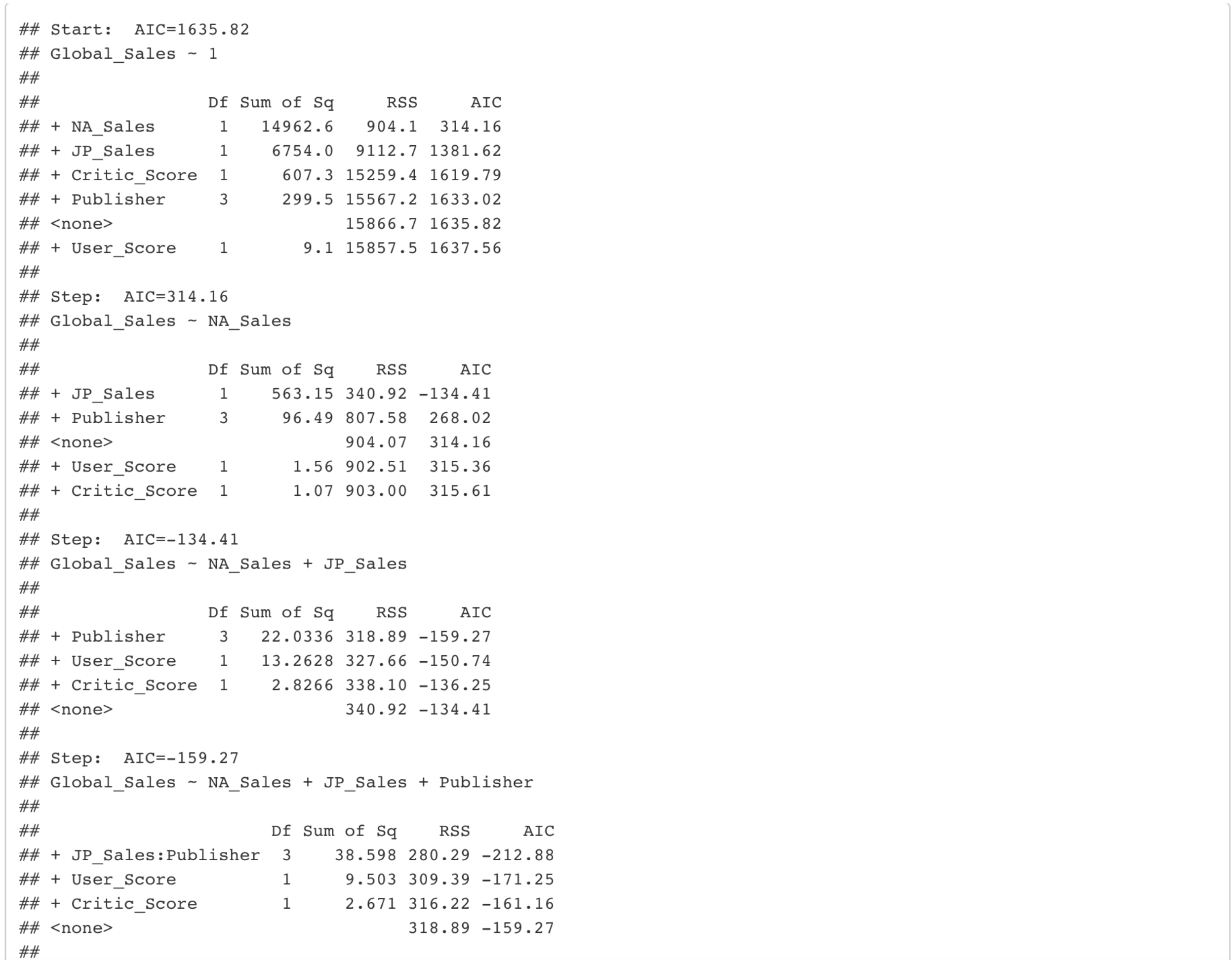
Model Selection

The method we choose is forward selection, we use the following coding block to construct the metric. This method successfully performs and helps to exclude the variable User_Score.



```
game_start = lm(Global_Sales ~ 1, data = games_data)
game_model_forward = step(
  game_start,
  scope = Global_Sales ~ NA_Sales + JP_Sales + Publisher +
    Critic_Score + User_Score + JP_Sales:Publisher,
  direction = "forward")
```

We are very surprised by the result of the selection first. In the normal sense, the user score is very important for buyin g the game. However, this makes sense, because the critic score is much more important in the real market. Many users would still buy the game if the media gives the game a good score, although it might have a few bad comments rated by players. You can see the AIC score details below.



```
## Start:  AIC=1635.82
## Global_Sales ~ 1
##
##                Df Sum of Sq     RSS     AIC
## + NA_Sales      1   14962.6   904.1  314.16
## + JP_Sales      1    6754.0  9112.7 1381.62
## + Critic_Score  1     607.3 15259.4 1619.79
## + Publisher     3     299.5 15567.2 1633.02
## <none>                      15866.7 1635.82
## + User_Score    1       9.1 15857.5 1637.56
##
## Step:  AIC=314.16
## Global_Sales ~ NA_Sales
##
##                Df Sum of Sq    RSS     AIC
## + JP_Sales      1    563.15 340.92 -134.41
## + Publisher     3     96.49 807.58  268.02
## <none>                      904.07  314.16
## + User_Score    1      1.56 902.51  315.36
## + Critic_Score  1      1.07 903.00  315.61
##
## Step:  AIC=-134.41
## Global_Sales ~ NA_Sales + JP_Sales
##
##                Df Sum of Sq    RSS     AIC
## + Publisher     3   22.0336 318.89 -159.27
## + User_Score    1   13.2628 327.66 -150.74
## + Critic_Score  1    2.8266 338.10 -136.25
## <none>                      340.92 -134.41
##
## Step:  AIC=-159.27
## Global_Sales ~ NA_Sales + JP_Sales + Publisher
##
##                      Df Sum of Sq    RSS     AIC
## + JP_Sales:Publisher  3    38.598 280.29 -212.88
## + User_Score          1     9.503 309.39 -171.25
## + Critic_Score        1     2.671 316.22 -161.16
## <none>                            318.89 -159.27
##
```

```
##
##                       Df Sum of Sq    RSS     AIC
## + JP_Sales:Publisher  3    38.598 280.29 -212.88
## + User_Score          1     9.503 309.39 -171.25
## + Critic_Score        1     2.671 316.22 -161.16
## <none>                            318.89 -159.27
##
## Step:  AIC=-212.88
## Global_Sales ~ NA_Sales + JP_Sales + Publisher + JP_Sales:Publisher
##
##                 Df Sum of Sq    RSS     AIC
## + Critic_Score  1    3.8330 276.46 -217.24
## + User_Score    1    3.6104 276.68 -216.87
## <none>                      280.29 -212.88
##
## Step:  AIC=-217.24
## Global_Sales ~ NA_Sales + JP_Sales + Publisher + Critic_Score +
##     JP_Sales:Publisher
##
##               Df Sum of Sq    RSS     AIC
## <none>                    276.46 -217.24
## + User_Score  1   0.80018 275.66 -216.58
```

After this procedure, our final model is **Global_Sales ~ NA_Sales + JP_Sales + Publisher + Critic_Score + JP_Sales:Publisher**

Our fitted line is below:

Global_Sales = 0.3061 + 1.7433*NA_Sales + 6.722*JP_Sales + 0.3235*PublisherMicrosoft + 0.03147*PublisherNintendo + 0.4168*PublisherUbisoft - 0.008357*Critic_Score - 13.1491*JP_Sales:PublisherMicrosoft - 5.0958*JP_Sales:PublisherNintendo - 4.0224*JP-Sales:PublisherUbsoft

Collinearity

We decide to use VIF (variance inflation factor) to measure the collinearity of coefficients in our model, and a large VIF indicates high collinearity.

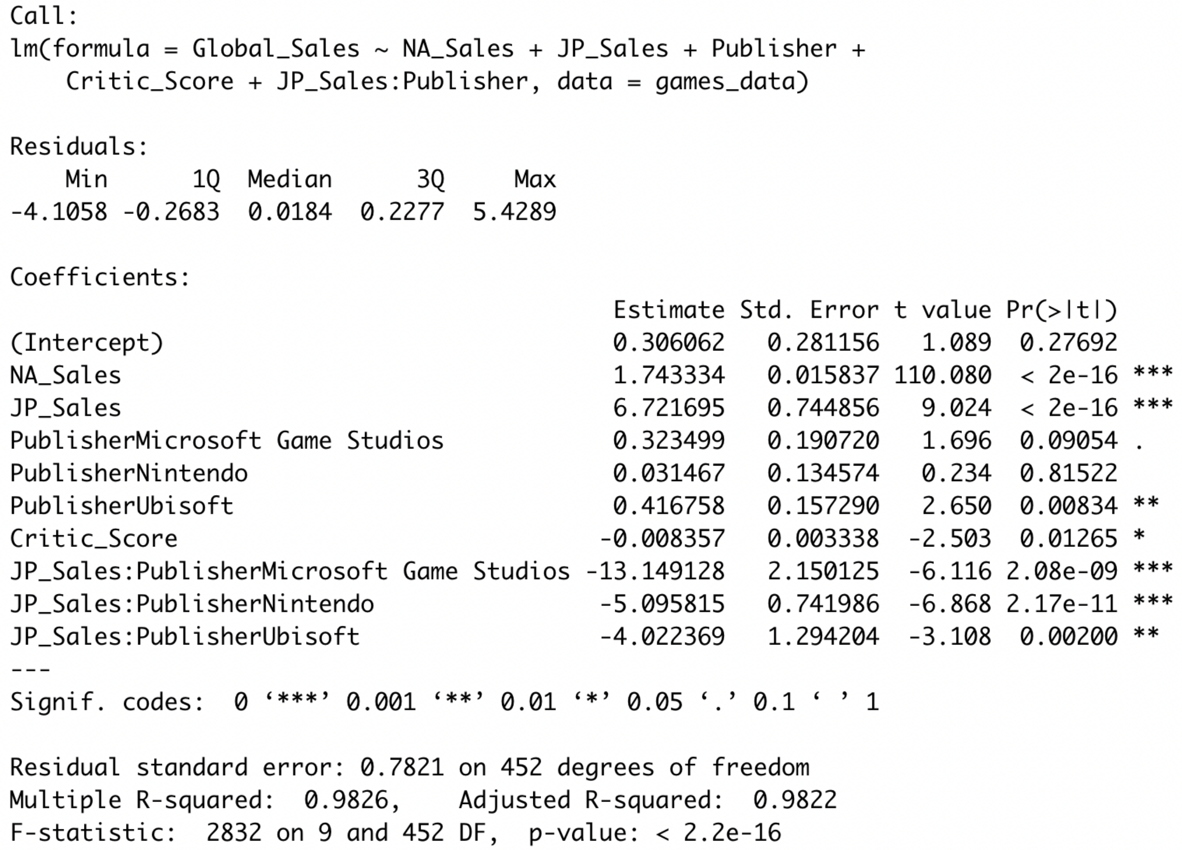
After removing all the categorical predictors in our model, we included NA_Sales, JP_Sales, Critic_Score, and User_Score to be the predictors in our first model. After calling the VIF function on the first model, we get the VIF of NA_Sales is 1.4783, the VIF of JP_Sales is 1.4121, the VIF of Critic_Score is 1.4720, the VIF of User_Score is 1.4684. All the four predictors in the first model have a relatively low VIF, which means that all of them are not highly correlated to other predictors, and there is no cause of concern. In our final model, we decided to remove User_Score after using the method of forward selection, and the final model includes NA_Sales, JP_Sales, and Critic_Score. After calling VIF function, we get the VIF of NA_Sales is 1.3806, the VIF of JP_Sales is 1.3421, the VIF of Critic_Score is 1.0468. We can see that all the VIFs of the predictors in the final model are also relatively low, which means they are not highly correlated to each other and there is no cause of concern.

Additionally, we can see that for all the predictors included in the final model, their VIFs are lower than the corresponding VIFs in the first model, which means our model selection is effective and lowered the collinearity of predictors even more.

R^2

We use the summary function to generate a summary for our first and final model. Our first model has six predictors, which are NA_Sales, JP_Sales, Publisher, Critic_Score, User_Score, and interaction between Publisher and JP_Sales. From the output of R, we can see that multiple R^2 (coefficient of determination) is 0.9826, which means that about 98.26% of the variability in the global sales of video games can be explained by their linear relationship with their sales in North America, their sales in Japan, their publisher, the critic score, the user score and the interaction term between the publisher and sales in Japan. 0.9826 is definitely a high R^2 value, much higher than most R^2 values we worked in exercises, which means that the prediction from our linear regression model is really close to the actual results. One reason we think our R^2 is very large is that we included the sales from North America and sales from Japan to be our predictors, which are actually a part of global sales. Our final model includes all the predictors in the first model except the user score. After looking at the output from R, we found that the multiple R^2 is also 0.9826, and the adjusted R^2 is also 0.9822, which is exactly the same as our first model. Bothe the multiple R^2 and adjusted R^2 stay the same. Since we did not expect two coefficients of determination to be exactly the same, this result is quite surprising for us. The reason for this situation might be that the user scores for video games actually have a weaker influence on global sales than we think, at least weaker than any other predictors in our model.

```
Call:
lm(formula = Global_Sales ~ NA_Sales + JP_Sales + Publisher + 
    Critic_Score + User_Score + Publisher * JP_Sales, data = games_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.0950 -0.2686  0.0306  0.2296  5.4226 

Coefficients:
                                           Estimate Std. Error t value Pr(>|t|)    
(Intercept)                                0.420258   0.298255   1.409  0.15951    
NA_Sales                                   1.740773   0.015989 108.874  < 2e-16 ***
JP_Sales                                   6.474866   0.775221   8.352 8.32e-16 ***
PublisherMicrosoft Game Studios            0.308771   0.191089   1.616  0.10683    
PublisherNintendo                          0.063884   0.137479   0.465  0.64238    
PublisherUbisoft                           0.426962   0.157489   2.711  0.00696 ** 
Critic_Score                              -0.005436   0.004202  -1.294  0.19647    
User_Score                                -0.046436   0.040584  -1.144  0.25315    
JP_Sales:PublisherMicrosoft Game Studios -12.704980   2.184161  -5.817 1.14e-08 ***
JP_Sales:PublisherNintendo                -4.845699   0.773274  -6.266 8.64e-10 ***
JP_Sales:PublisherUbisoft                 -3.738771   1.317291  -2.838  0.00474 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7818 on 451 degrees of freedom
Multiple R-squared:  0.9826,	Adjusted R-squared:  0.9822 
F-statistic:  2551 on 10 and 451 DF,  p-value: < 2.2e-16
```



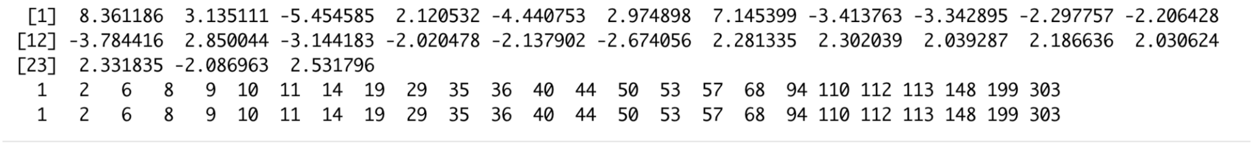
```
Call:
lm(formula = Global_Sales ~ NA_Sales + JP_Sales + Publisher + 
    Critic_Score + JP_Sales:Publisher, data = games_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.1058 -0.2683  0.0184  0.2277  5.4289 

Coefficients:
                                        Estimate Std. Error t value Pr(>|t|)    
(Intercept)                             0.306062   0.281156   1.089  0.27692    
NA_Sales                                1.743334   0.015837 110.080  < 2e-16 ***
JP_Sales                                6.721695   0.744856   9.024  < 2e-16 ***
PublisherMicrosoft Game Studios         0.323499   0.190720   1.696  0.09054 .  
PublisherNintendo                       0.031467   0.134574   0.234  0.81522    
PublisherUbisoft                        0.416758   0.157290   2.650  0.00834 ** 
Critic_Score                           -0.008357   0.003338  -2.503  0.01265 *  
JP_Sales:PublisherMicrosoft Game Studios -13.149128   2.150125  -6.116 2.08e-09 ***
JP_Sales:PublisherNintendo             -5.095815   0.741986  -6.868 2.17e-11 ***
JP_Sales:PublisherUbisoft              -4.022369   1.294204  -3.108  0.00200 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7821 on 452 degrees of freedom
Multiple R-squared:  0.9826,	Adjusted R-squared:  0.9822 
F-statistic:  2832 on 9 and 452 DF,  p-value: < 2.2e-16
```

Unusual Observations

We decide to use leverage, standardized residuals, and Cook's distance to differentiate unusual observations. We first take the final model after selection and call the hat values function to find the leverage, and then we find the observations which have large residuals. We have 10 coefficients in the final model (p = 10), and we have 462 observations in our dataset (n = 462), which means we have 462 individual games. From the output of R, we get that the limit of leverage is 2p/n which is equal to 0.0433, so observations have high leverages if their leverage values are greater than 0.0433. These are observations that have high leverages, and we can see that there are 33 observations that have high leverage.

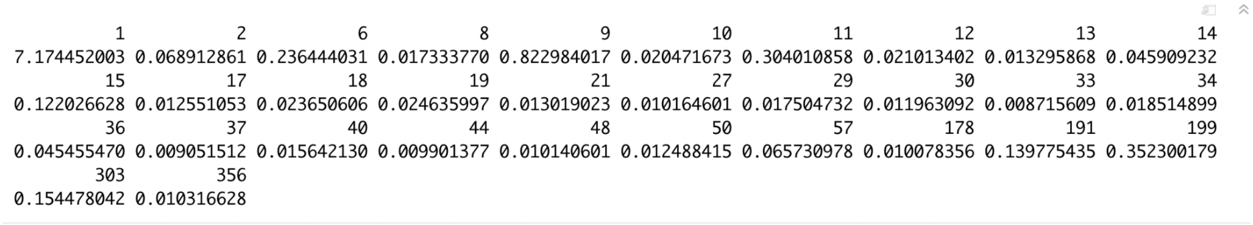
```
         1          2          3          4          5          6          7          9
11         12         15
0.50647736 0.06551866 0.06150859 0.13359049 0.06210401 0.07361972 0.04994818 0.29444734
0.05619753 0.10915600 0.29850118
        18         21         22         23         26         32         34         46
57         59        178
0.16030150 0.15352595 0.11542842 0.06615791 0.09396425 0.09610388 0.07707797 0.05188896
0.08418540 0.05368880 0.04955746
       191        199        226        262        303        306        317        349
356        391        437
0.26144525 0.44717142 0.05365355 0.04373027 0.19419543 0.04406540 0.06179288 0.10543266
0.07587370 0.06463392 0.05933022
```

Then we call the r-standard function to find the standardized residuals of observations, and observations have large standardized residuals if their standardized residuals have a value greater than 2. These are the observations with large standardized residuals, and we can see that there are 25 observations with large standardized residuals.



```
 [1]  8.361186  3.135111 -5.454585  2.120532 -4.440753  2.974898  7.145399 -3.413763 -3.342895 -2.297757 -2.206428
[12] -3.784416  2.850044 -3.144183 -2.020478 -2.137902 -2.674056  2.281335  2.302039  2.039287  2.186636  2.030624
[23]  2.331835 -2.086963  2.531796
  1   2   6   8   9  10  11  14  19  29  35  36  40  44  50  53  57  68  94 110 112 113 148 199 303
  1   2   6   8   9  10  11  14  19  29  35  36  40  44  50  53  57  68  94 110 112 113 148 199 303
```

Then we call the Cook's Distance function to find the Cook's distance for all the observations. We have 462 observations, so observations have a large Cook's Distance if they have Cook's Distance greater than 0.0087. These are observations with values of large Cook's Distance, and we can see that there are 32 observations that have high Cook's Distance.

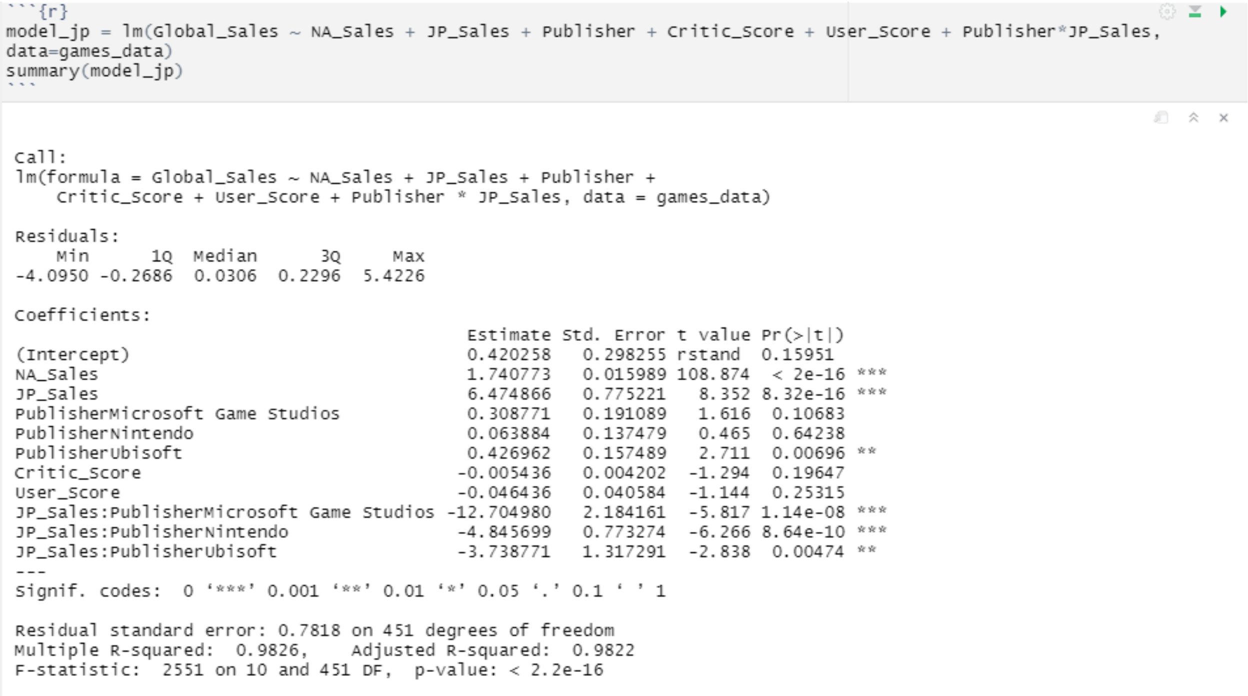


```
          1           2           6           8           9          10          11          12          13          14
7.174452003 0.068912861 0.236444031 0.017333770 0.822984017 0.020471673 0.304010858 0.021013402 0.013295868 0.045909232
         15          17          18          19          21          27          29          30          33          34
0.122026628 0.012551053 0.023650606 0.024635997 0.013019023 0.010164601 0.017504732 0.011963092 0.008715609 0.018514899
         36          37          40          44          48          50          57         178         191         199
0.045455470 0.009051512 0.015642130 0.009901377 0.010140601 0.012488415 0.065730978 0.010078356 0.139775435 0.352300179
        303         356
0.154478042 0.010316628
```

Overall we can see that the number of observations that are especially influential to our model is generally the same with three different methods (leverage, standardized residuals, and Cook's Distance), so we can say that there are about 25-32 unusual observations in our dataset. Since we have a total of 462 observations, and they only take about 6% to 7% percent of our dataset, we can conclude that the number of unusual observations in our model is relatively small and does not create a large influence on our final model.

Size of the Model

Next, we will discuss the size of the model. In our first model, the number of observations is 462 (n=462). The model consists of all the first-order coefficients and the number of coefficients (including the intercept) is 11 (p=11). According to the rule of thumb, the number of observations should be greater than 10 times the number of the coefficient (n>=10*p), so that we have enough data to fit the model. The first model satisfies this condition (462>=10*11), which means the number of coefficients is reasonable for this size of the dataset.

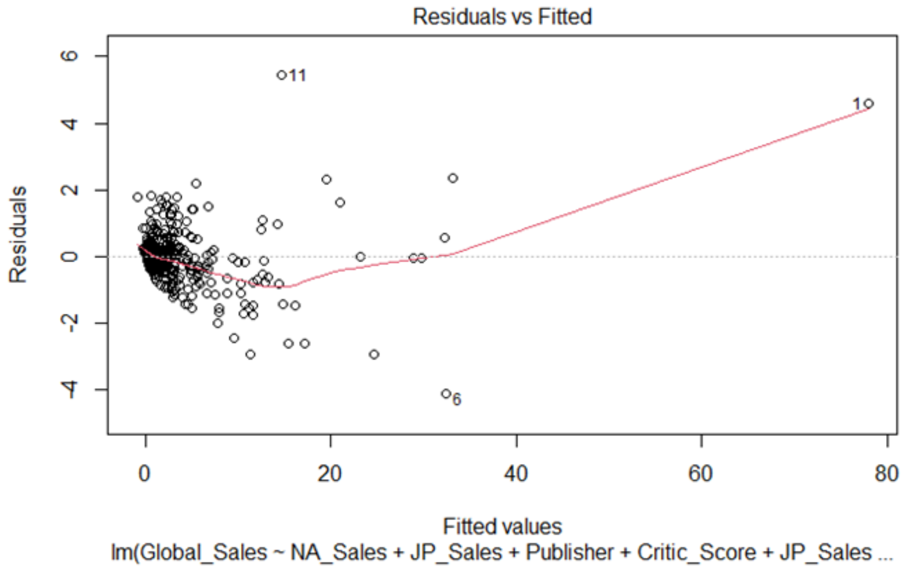


~~~
```{r}
model_jp = lm(Global_Sales ~ NA_Sales + JP_Sales + Publisher + Critic_Score + User_Score + Publisher*JP_Sales,
data=games_data)
summary(model_jp)
```
~~~

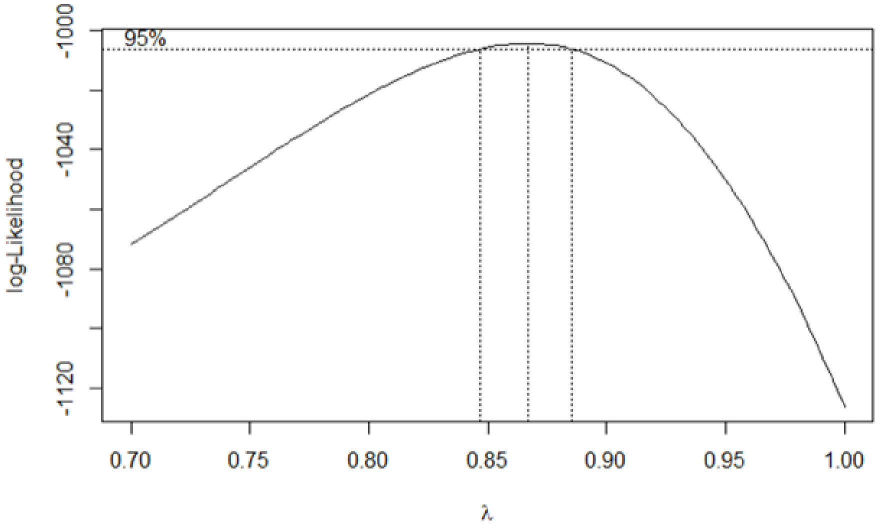
```
Call:
lm(formula = Global_Sales ~ NA_Sales + JP_Sales + Publisher +
    Critic_Score + User_Score + Publisher * JP_Sales, data = games_data)

Residuals:
    Min      1Q  Median      3Q     Max
-4.0950 -0.2686  0.0306  0.2296  5.4226

Coefficients:
                                      Estimate Std. Error t value Pr(>|t|)
(Intercept)                           0.420258   0.298255 rstand  0.15951
NA_Sales                              1.740773   0.015989 108.874  < 2e-16 ***
JP_Sales                              6.474866   0.775221   8.352 8.32e-16 ***
PublisherMicrosoft Game Studios       0.308771   0.191089   1.616  0.10683
PublisherNintendo                     0.063884   0.137479   0.465  0.64238
PublisherUbisoft                      0.426962   0.157489   2.711  0.00696 **
Critic_Score                         -0.005436   0.004202  -1.294  0.19647
User_Score                           -0.046436   0.040584  -1.144  0.25315
JP_Sales:PublisherMicrosoft Game Studios -12.704980   2.184161  -5.817 1.14e-08 ***
JP_Sales:PublisherNintendo           -4.845699   0.773274  -6.266 8.64e-10 ***
JP_Sales:PublisherUbisoft            -3.738771   1.317291  -2.838  0.00474 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7818 on 451 degrees of freedom
Multiple R-squared:  0.9826,    Adjusted R-squared:  0.9822
F-statistic:  2551 on 10 and 451 DF,  p-value: < 2.2e-16
```

Boxcox

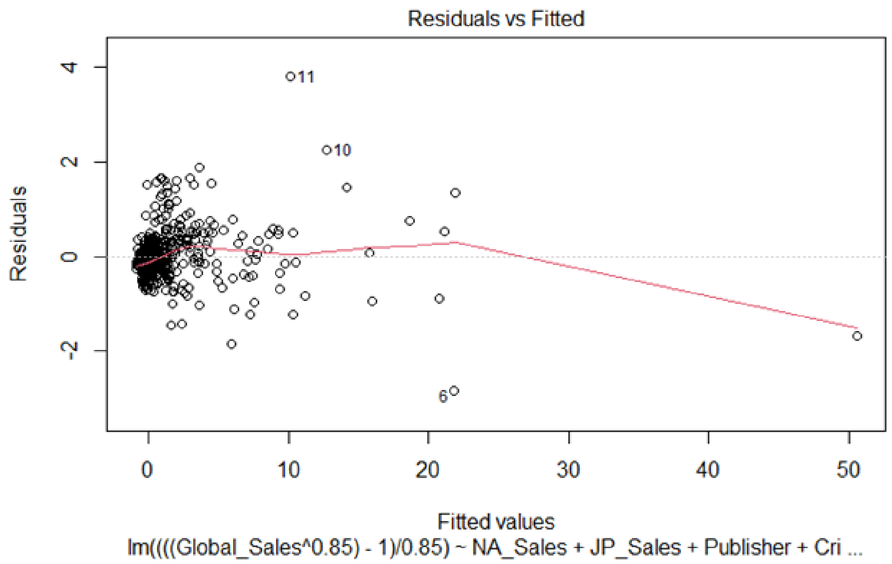
By plotting the data on the original scale, we can see from the plot below, it shows that the residual is not fixed around 0. In this case, we can do the box-cox transformation to fix this problem.



We then use the **boxcox()** function to find the best transformation of the form considered by the Box-Cox method. And we can find a specific interval for λ, here we can see λ=0.85 is one of the values in the confidence interval. After the box cox transformation, the Global_Sales is transformed into (Global_Sales^0.85-1)/0.85



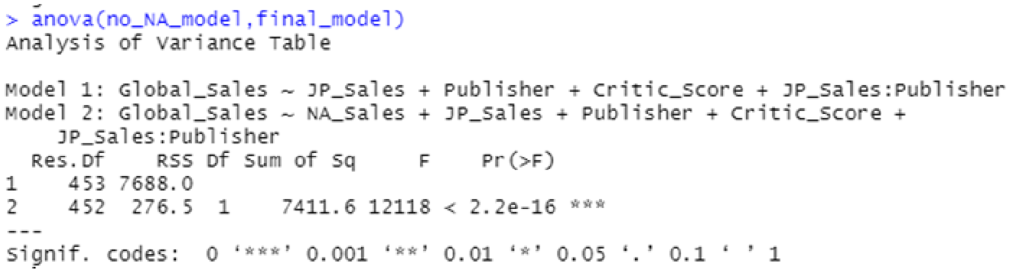
And as we can see in the plot below, the residual is more fixed around 0 compared to the residual before the transformation.



Statistical Test

We will now use the Anova test to help decide whether the sales in North America and the critic scores are important in our final model.

First, we can have an Anova test for the final model and the final model without NA_Sales.



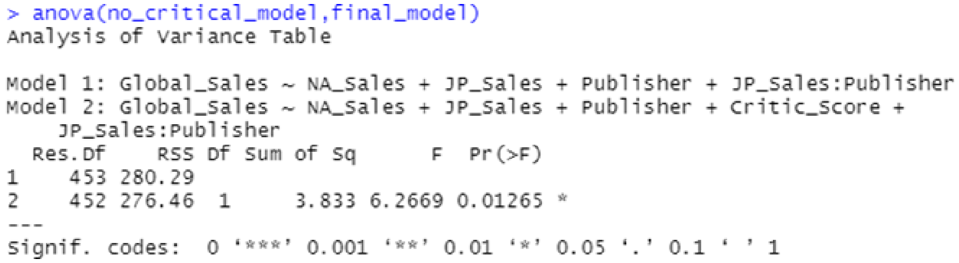
```
> anova(no_NA_model,final_model)
Analysis of Variance Table

Model 1: Global_Sales ~ JP_Sales + Publisher + Critic_Score + JP_Sales:Publisher
Model 2: Global_Sales ~ NA_Sales + JP_Sales + Publisher + Critic_Score +
    JP_Sales:Publisher
  Res.Df    RSS Df Sum of Sq     F    Pr(>F)    
1    453 7688.0                                  
2    452  276.5  1    7411.6 12118 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

H0: Global_Sales = β0 + β1*JP_Sales + β2*PublisherMicrosoft + β3*PublisherNintendo + β4*PublisherUbisoft+β5*Critic_Score +β6*JP_Sales:PublisherMicrosoft +β7*JP_Sales:PublisherNintendo+β8*JP-Sales:PublisherUbsoft+ε

H1: Global_Sales = β0 + β1*JP_Sales + β2*PublisherMicrosoft + β3*PublisherNintendo + β4*PublisherUbisoft+β5*Critic_Score +β6*JP_Sales:PublisherMicrosoft +β7*JP_Sales:PublisherNintendo+β8*JP-Sales:PublisherUbsoft+β9*NA_Sales +ε

The p-value is 2.2e-16, which means that we will reject H0. It shows that NA_Sales are important in our final model to predict the Globle_Sales, which means that the Sales in North America is an important indicator of global sales.

Then we can have an Anova test for the final model and the final model without Critic_Score.



```
> anova(no_critical_model,final_model)
Analysis of Variance Table

Model 1: Global_Sales ~ NA_Sales + JP_Sales + Publisher + JP_Sales:Publisher
Model 2: Global_Sales ~ NA_Sales + JP_Sales + Publisher + Critic_Score + 
    JP_Sales:Publisher
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)  
1    453 280.29                              
2    452 276.46  1     3.833 6.2669 0.01265 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

H0: Global_Sales = β0 + β1*JP_Sales + β2*PublisherMicrosoft + β3*PublisherNintendo + β4*PublisherUbisoft+β5*NA_Sales +β6*JP_Sales:PublisherMicrosoft +β7*JP_Sales:PublisherNintendo+β8*JP-Sales:PublisherUbsoft+ε

H1: Global_Sales = β0 + β1*JP_Sales + β2*PublisherMicrosoft + β3*PublisherNintendo + β4*PublisherUbisoft+β5*NA_Sales +β6*JP_Sales:PublisherMicrosoft +β7*JP_Sales:PublisherNintendo+β8*JP-Sales:PublisherUbsoft+β9* Critic_Score +ε

The p-value is 0.01265, which means that we will reject H0. It shows that Critic_Score is important in our final model to predict the Globle_Sales, which means that the critic score of the game is an important indicator of global sales.

Discussion

In our model, we use linear regression to predict future global sales using game sales in Japan, North America sales, game publishing companies, critic scores, and user scores. In the model design, we found that the performance of different companies in Japan is quite different, which will have an impact on the model’s predictions, so we added the intersection of sales in the Japanese market and the game publishing companies. This effectively improves the adjusted R square of the model. And, in our analysis, we found that for Microsoft's games, sales in the Japanese market are negatively correlated with global sales, while for Activision, Japanese sales are positively correlated with global sales. After introducing intersection terms, we chose the variables. We chose to abandon user_score, which may mean that user scores cannot predict global sales well after known issuers, sales in Japan and North America, and critic scores. Through the Anova test, we found that the critic score and North American sales are both important to the model.

In addition, I think this data has many limitations, such as the lack of data from other publishing companies. Since the performance of different companies is quite different in various variables, this model may not perform well to predict the performance of other companies. Besides, if we want to analyze the performance of the model more deeply, we need to introduce more intersection variables. For example, we can introduce the intersection between critic score and Company, we can analyze how critic score may help us predict the global sales for different companies.

Reference

Kaggle Dataset. (2019). Video game sales from Vgchartz and corresponding ratings from Metacritic. https://www.kaggle.com/mohalim/video-games-sales