Final project

Henrique Magalhaes Rio

11/28/2020

Introduction

What makes a video games successful? As part of the entertainment industry video games are very subjective what is a good game for one person for another person it might not be. Like the movie industry, ratings are quite important, with mainly 2 types of ratings which are critics and user ratings, in a platform like STEAM, where small or large developers can freely publish their games ratings are extremely important since it could be the differential between the purchase decision or not, especially for unknown developers those ratings could be the difference between a successful or a unsuccessful game.

In this paper, I will analyze whether both types of ratings affect how successful a game is, in order to measure success the amount of units sold will be used, the differences in genre and platform(WII,Switch,PS4) will also be accounted as well as the amount of units sold in different region.

Model Formulation

In order to fully understand the relation between score(rating) and units, several variable will be included which leads to the full model being: $\log(GlobalSales) = \beta_0 + \beta_1 * UserScore + \beta_2 * CriticScore + \beta_3 * Genre + \beta_4 * Platform.$

The User Score and Critic Score variables, are the main variables since where are interested in how the change in those variables influences the amount of units sold. the Genre and Platform where introduced in order to account for any omitted variable bias that might happen since ratings might have different effects on different genres and platforms, where more popular genres such as Action ratings might have a lot more impact when compared to simulation that has a smaller more united community.

Data

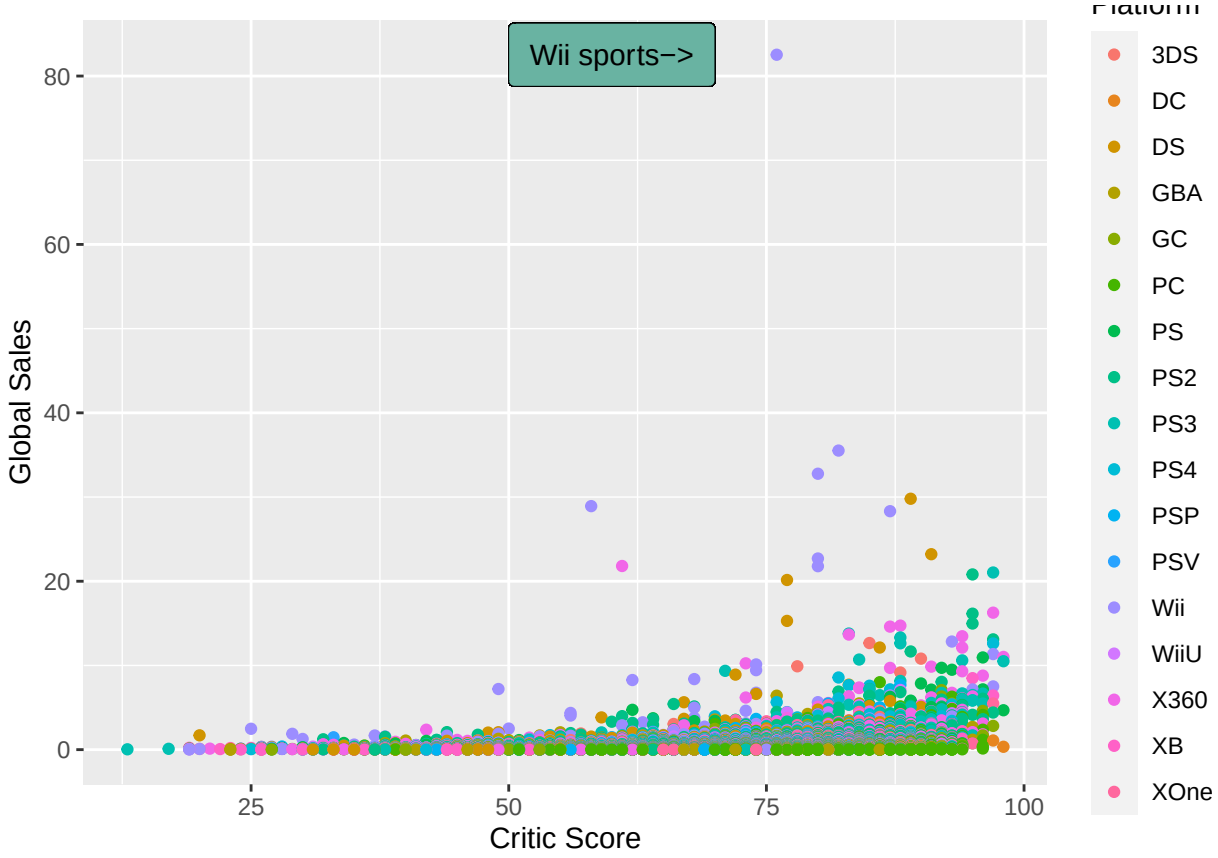
The data for this model was gathered from games that were released from 1998 to 2016, Some games where removed since they did not have all of the data necessary, which leaves about 8137 games, with 12 different genres and 17 different platforms.

	Mean	Std. dev.	Max	Min
Global Sales	0.7	1.8	82.5	0.01
User Score	7.2	1.4	9.6	0.50
Critic Score	69.0	13.9	98.0	13.00

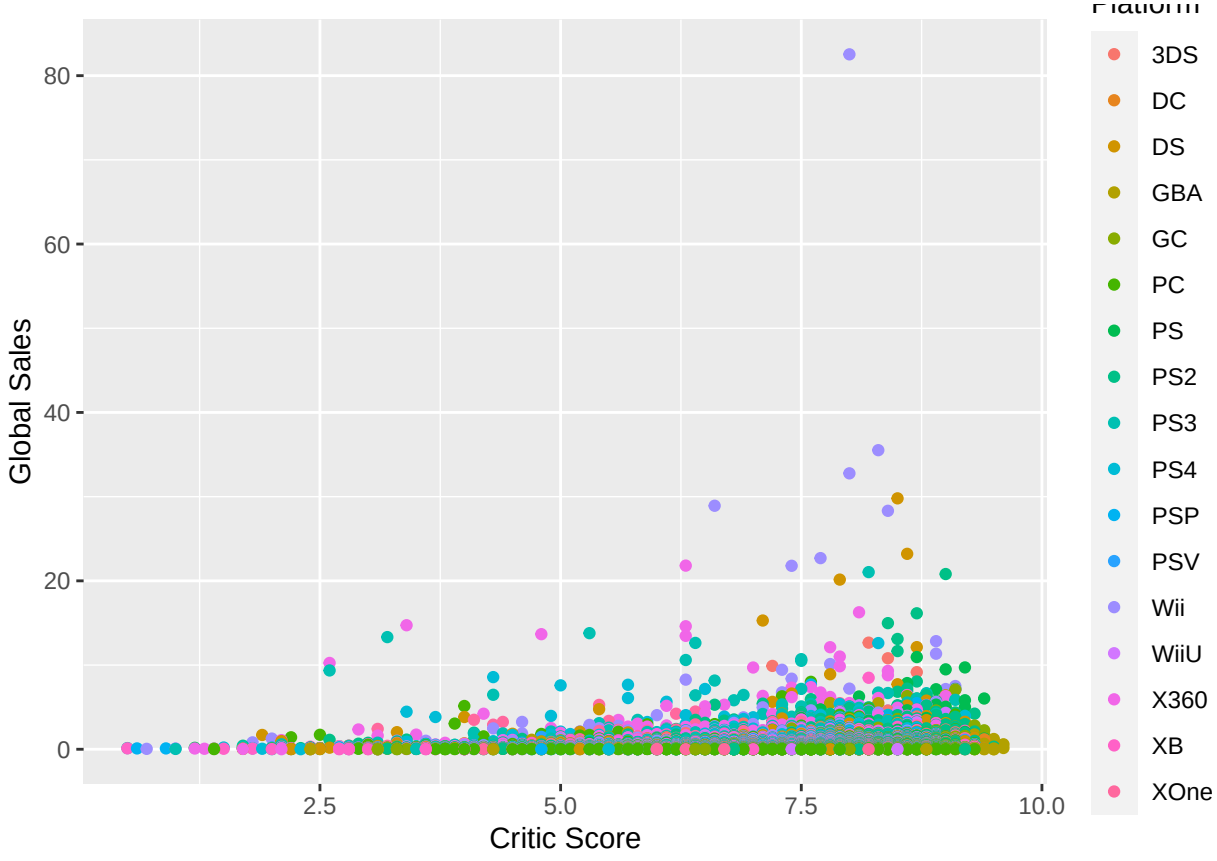
we can see that Global Sales varies by quite a lot with the mean being 0.7 millions of units, the max is 82.5 million units which belongs to Wii sports game and min is 0.01 and standard deviation 1.8, which makes sense since is not every game that gets to sell a lot of units. The User Score is done in a scaling from 0 to 10, with the highest score being 9.6 and the lowest 0.5 and the mean being 7.2 and standard deviation being 1.4.

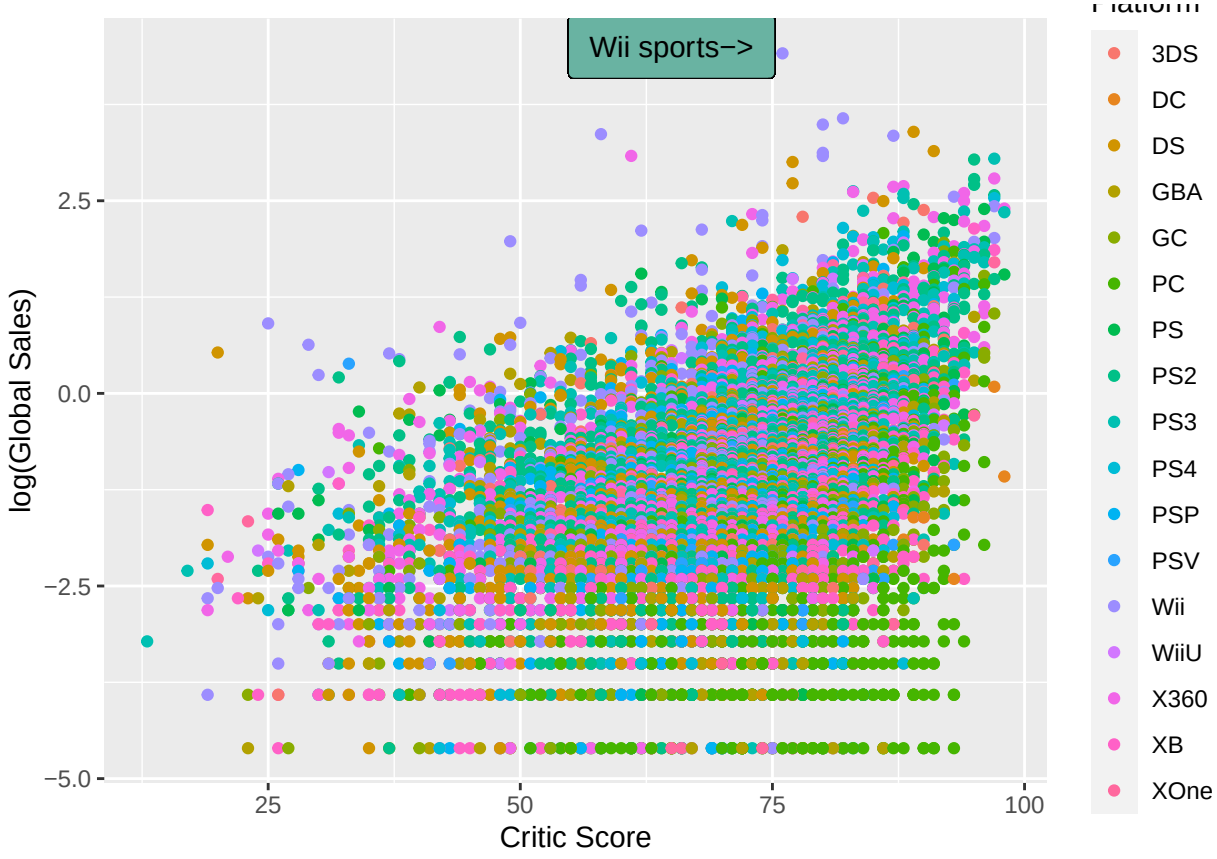
The critic score variable is scored from 0 to 100, with the highest score being 98.0 and the lowest being 13, and the mean is 69.0 which is lower than the User score mean, if the scales where the same they would have very close standard deviations.

To further understand the relationship of Critic/User Score with the global sales we can plot the data, if we plot without the log based Global_Sales we can see that the data is very concentrated at the bottom with few outliers. But if we add the log to the Global Sales, we can see that the data is much more spread out and has less evident outliers which makes the model more accurate.

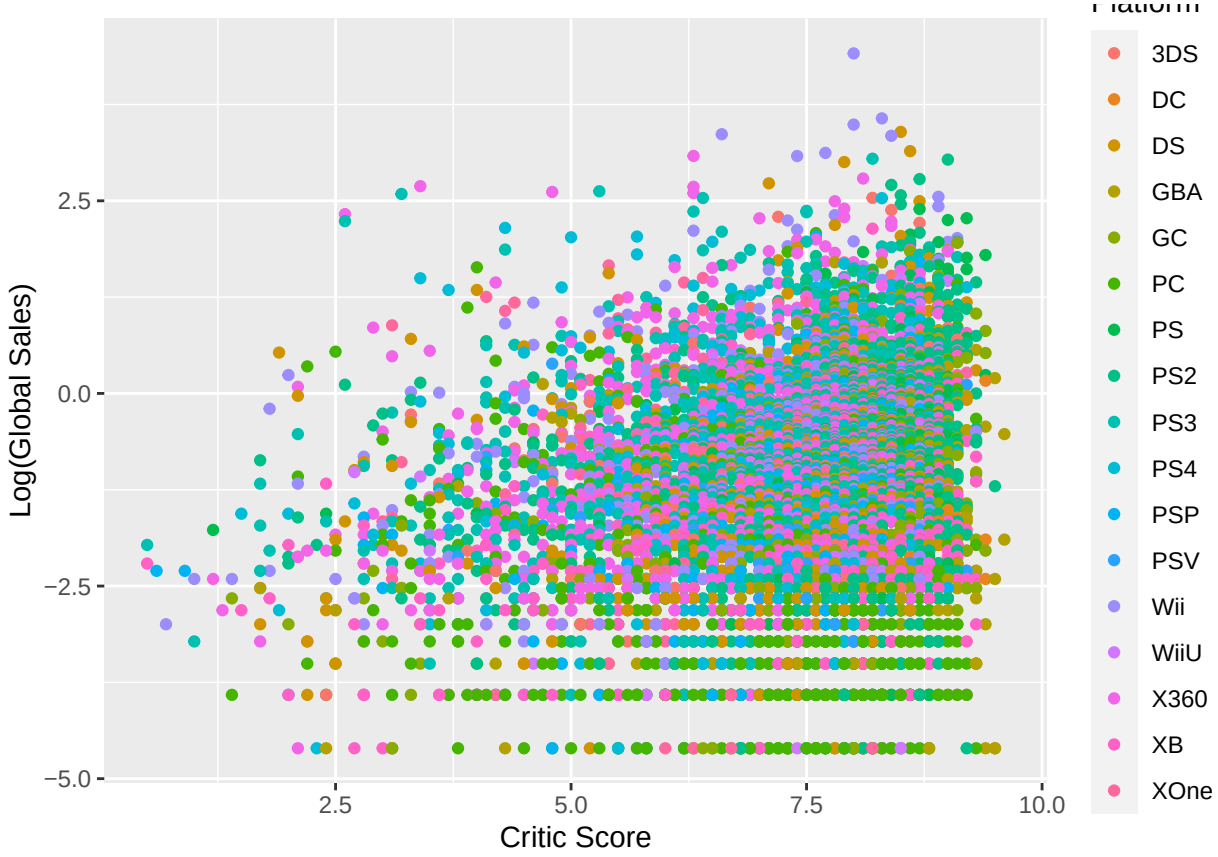



```
## Warning: Removed 1120 rows containing missing values (geom_point).
```


Global Sales
80
60
40
20
0
2.5
5.0
7.5
10.0
Critic Score
Platform
3DS
DC
DS
GBA
GC
PC
PS
PS2
PS3
PS4
PSP
PSV
Wii
WiiU
X360
XB
XOne


Wii sports−>
log(Global Sales)
2.5
0.0
−2.5
−5.0
25
50
75
100
Critic Score
Platform
3DS
DC
DS
GBA
GC
PC
PS
PS2
PS3
PS4
PSP
PSV
Wii
WiiU
X360
XB
XOne

```
## Warning: Removed 1120 rows containing missing values (geom_point).
```



Empirical Results

Regressor	1	2	3	4	5
Intercept	-4.04(0.07)***	-2.37(0.08)***	-3.60(0.09)***	-3.55(0.09)***	-3.7(0.12)***
$\Delta CriticScore$	0.03(0.001)***		0.041(0.001)***	0.041(0.001)***	0.05(0.001)***
$\Delta UserScore$		0.15(0.01)***	-0.07(0.01)***	-0.06(0.013)***	-0.11(0.01)***
$\Delta(Genre)Adventure$				-0.75(0.084)***	-0.6(0.07)***
$\Delta(Genre)Fighting$				0.004(0.07)	-0.13(0.06)
$\Delta(Genre)Misc$				0.30(0.07)***	0.1(0.06)
$\Delta(Genre)Platformer$				0.047(0.070)	0.007(0.06)
$\Delta(Genre)Puzzle$				-0.52(0.12)***	0.6(0.1) ***
$\Delta(Genre)Racing$				-0.15(0.06)*	-0.14(0.05) **
$\Delta(Genre)Role-Playing$				-0.25(0.057)***	-0.198(0.05)***
$\Delta(Genre)Shooter$				-0.07(0.05)	0.03(0.05)
$\Delta(Genre)Simulation$				-0.25(0.08)**	0.03(0.07)
$\Delta(Genre)Sports$				-0.037(0.05)	-0.16(0.05)***
$\Delta(Genre)Strategy$				-1.26(0.08)***	-0.66(0.07)***

Regressor	1	2	3	4	5
$\Delta(Platform)DC$					-1.16(0.32)***
$\Delta(Platform)DS$					-0.03(0.10)
$\Delta(Platform)GBA$					-0.35(0.12)**
$\Delta(Platform)GC$					-0.48(0.11)***
$\Delta(Platform)PC$					-1.7(0.1)***
$\Delta(Platform)PS$					0.33(0.13)*
$\Delta(Platform)PS2$					0.17(0.09)
$\Delta(Platform)PS3$					0.25(0.099)*
$\Delta(Platform)PS4$					-0.32(0.12)**
$\Delta(Platform)PSP$					
$\Delta(Platform)PSV$					-0.73(0.14)***
$\Delta(Platform)Wii$					0.43(0.10)***
$\Delta(Platform)WiiU$					-0.25(0.15)
$\Delta(Platform)Xbox360$					0.19(0.09)
$\Delta(Platform)Xbox$					-0.67(0.1)***
$\Delta(Platform)XboxOne$					-0.4(0.13)**

=90% Significance ***=95% significance*** =99% significance

Summary Statistics	1	2	3	4	5
F-Statisic	1373	186.3	540.4	117.5	128.4
P-value(F-test)	$<$ 2.2e-16	$<$ 2.2e-16	$<$ 2.2e-16	$<$ 2.2e-16	$<$ 2.2e-16
R^2	0.1444	0.02587	0.1333	0.1775	0.3449
SER	1.307	1.386	1.307	1.273	1.136

**The Adjusted R^2 was used for regression 3/4/5.

Regression 1 and 2 where done in order to compare the effects of just the user scores and the critic scores on the percent change in Global Sales, and it was found that a one point change in Critic Score is associated with a 3% increase in Global Sales, Where, a 1 point change in User Score, is associated with 15% increase in Global Sales. It should be noted that User Score and Critic Score are in different scales, which is why there User Score appears to have a higher increase in Global Sales but in Reality User Score, has a Higher impact.

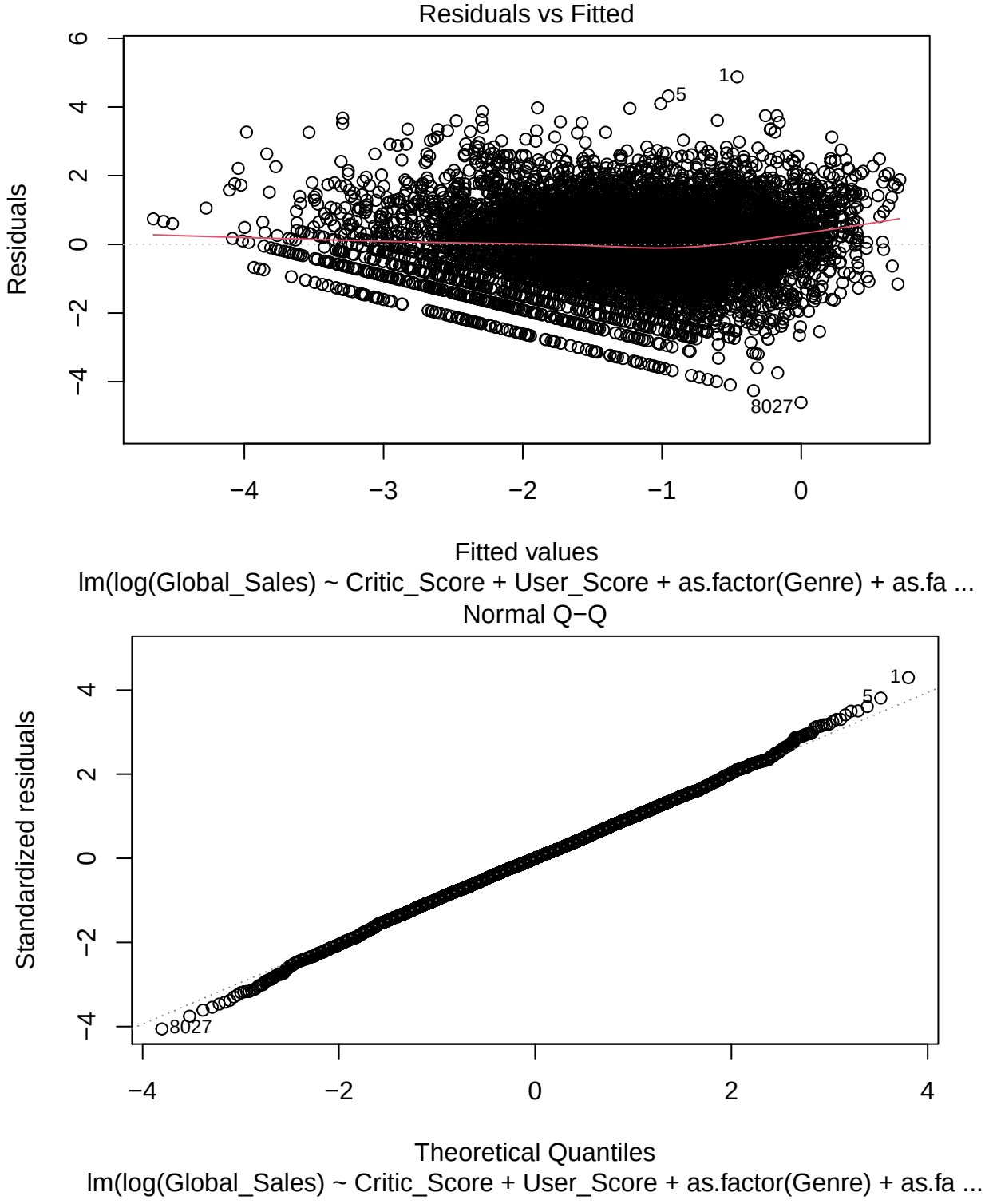
In Regression 3 we confirm what was previously stated that Critic Score has a higher impact in the increase in global sales, since for a one increase in User Score we now see a 7% percent decrease in Global Sales(p-value<0.0001), as for a one score increase in Critic Score we see a 4.1% increase in Global Sales(p-value<0.0001) which is higher the previously seem.

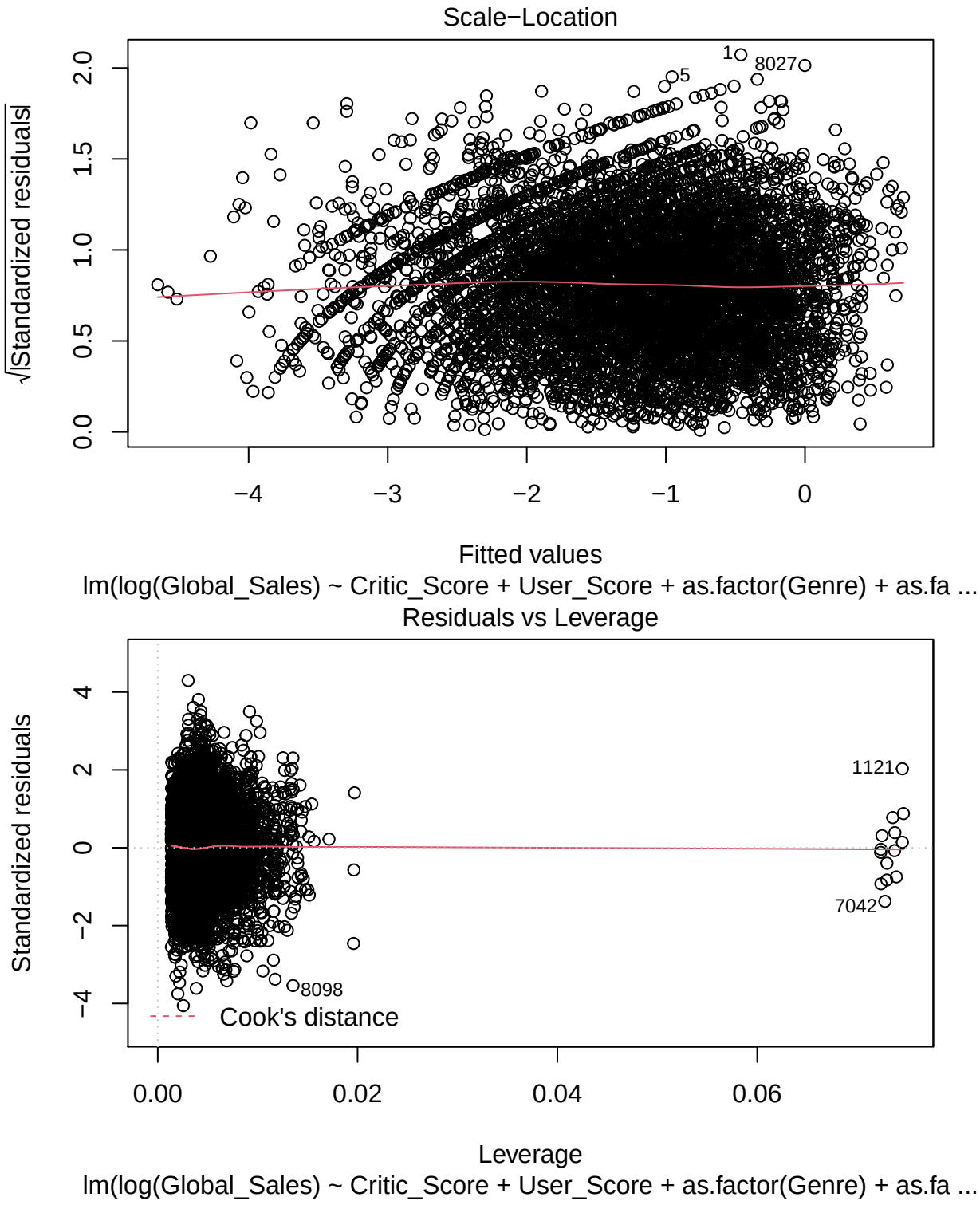
For Regression 3 and 4 we add all of the genres and platforms as categorical data, which further increases the effect of Critic Score, and decreases the effect of User Scores. Not all genres and platforms are statistically significant which means, that some games are more popular and therefore sell more while other games aren't as popular therefore, sell less, while other genre do not influence the amount of sales at all.

Conclusion

In order to analyze the effect of ratings on I would choose model 5 since not only it has the highest Adjusted R^2 at 0.3, which means 30% of the variation of Global Sales is explained by the model, it also has the lowest standard error. Model 5 also includes the categorical variables Genre and Platform which account for omitted

variable bias hence the increase in the R^2. By looking at the diagnostics plots Residual vs Fitted and QQ plot, we can see that the constant variance and the normality assumption are both respected, and the model also has a very high F-statistic.





Data Source

https://www.kaggle.com/gregorut/videogamesales

Appendix

```r
# Insert packages you need here
library(knitr)

library(tidyverse)

library(expss)
# define the markup language we are working in.
# options(qwraps2_markup = "latex") is also supported.

data <- read.csv('videogamesales.csv')

data$User_Score <- as.numeric(data$User_Score)


data <- data %>% filter(data$Critic_Score>0)
data <- data %>% filter(data$Global_Sales>0)

data %>%
    tab_cells(Global_Sales, User_Score, Critic_Score) %>%

    tab_stat_fun(Mean = w_mean, "Std. dev." = w_sd, "Max" = w_max, "Min"= w_min ,method = list) %>%
    tab_pivot()









ggplot(data,aes(Critic_Score,Global_Sales,color=Platform))+geom_point()+geom_label(
    label="Wii sports->",
    x=60,
    y=82.53,
    label.padding = unit(0.55, "lines"), # Rectangle size around label
    label.size = 0.20,
    color = "black",
    fill="#69b3a2"
  )+xlab("Critic Score")+ylab("Global Sales")
ggplot(data,aes(User_Score,Global_Sales,color=Platform))+geom_point()+xlab("Critic Score")+ylab("Global


ggplot(data,aes(Critic_Score,log(Global_Sales),color=Platform))+geom_point()+xlab("Critic Score")+ylab(
    label="Wii sports->",
    x=65,
    y=4.5,
```

```
    label.padding = unit(0.55, "lines"), # Rectangle size around label
    label.size = 0.20,
    color = "black",
    fill="#69b3a2"
  )
ggplot(data,aes(User_Score,log(Global_Sales),color=Platform))+geom_point()+xlab("Critic Score")+ylab("L
```

```
# geom_abline(slope = coef(lm)[[2]], intercept = coef(lm)[[1]])
```

```
lm1 <- lm(log(Global_Sales)~Critic_Score,data=data)
lm2 <- lm(log(Global_Sales)~User_Score,data=data)
lm3 <- lm(log(Global_Sales)~Critic_Score+User_Score,data=data)
lm4 <- lm(log(Global_Sales)~Critic_Score+User_Score+as.factor(Genre),data=data)
lm5 <- lm(log(Global_Sales)~Critic_Score+User_Score+as.factor(Genre)+as.factor(Platform),data=data)
```

```
summary(lm1)
summary(lm2)
summary(lm3)
summary(lm4)
summary(lm5)
```

```
plot(lm5)
```