Final Project

Sebastián Soriano Pérez [ss1072]

12/10/2019

Analyzing Video Game Sales

Summary

A few sentences describing the inferential question(s), the method used and the most important results.

Introduction

A more in-depth introduction to the inferential question(s) of interest.

The video game industry has been growing consistently during the last two decades, and in 2017 it was worth more than 78 billion USD worldwide. Video game software sales account for around 80% of total revenue. There are several factors that influence whether a video game will be successful or not, such as the developer studio, the critics rating, the user rating, among others. I will build a model to analyze what factors can help us determine how successful will a video game be in terms of global sales.

I analyzed a total of 4195 video game software releases across the world by 50 different developers. I built a hierarchical linear model on a logarithmic transformation of the global sales, with random slopes by developer, using a manual stepwise selection process using BIC and conditional R-squared as selection criteria. The variables on the final model were found to be significant in predicting the global sales for a video game.

The goal of this project is to find what are the most significant predictors for a video games success, measured as the number of sales around the world. I use a hierarchical linear regression model to explain the number of sales a particular videogame has. Considering that the developing studio plays a major role on a customer's decision to buy a new game, and due to the fact that accounting for every single studio included in the dataset (444 total) would make the interpretation too complicated, I used a random sample of 50 developers and built an appropriate hierarchical model with random intercepts effects for each one of them.

Data

You should describe the data in this section: how you obtained the data, the variables included, dealing with missing/erroneous values, exploratory data analysis etc.

The data was obtained from Kaggle (https://www.kaggle.com/rush4ratio/video-game-sales-with-ratings). It contains data from video games with 100,000 or more global sales from 1976 to 2016. The data contains 16719 rows with the following columns:

- *name* (categorical): Name of the video game
- *platform* (categorical): PLatform or console for which the video game was released
- *year_of_release* (categorical): Year of first release
- *genre* (categorical): Genre of the video game
- *publisher* (categorical): Publishing company
- *na_sales* (numerical): Units sold in North America
- *eu_sales* (numerical): Units sold in Europe
- *jp_sales* (numerical): Units sold in Japan
- *other_sales* (numerical): Units sold in the rest of the world
- *global_sales* (numerical): Total units sold worldwide

- *critic_score* (numerical): Average score (from 0 to 100) according to critics from other media aggregated by Metacritic
- *critic_count* (numerical): Number of critics taken into account for the Metacritic critic score
- *user_score* (numerical): Average score (from 0 to 100) according to Metacritic users
- *user_count* (numerical): Number of user scores on Metacritic
- *developer* (categorical): Video game developing company
- *rating* (categorical): Video game rating according to the ESRB that indicates the appropriate audience

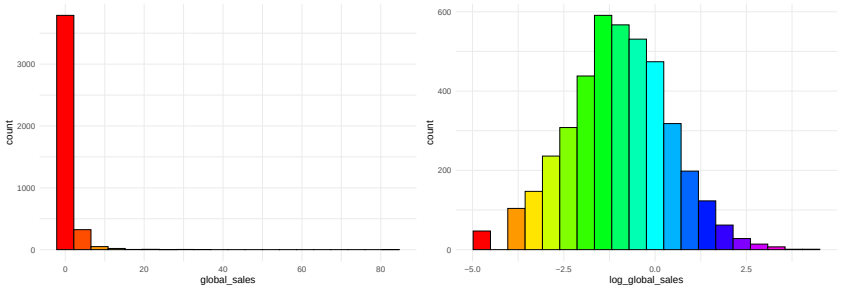
Around half of the rows in the dataset have missing values. First of all, I filtered out every video game release before 2000 since it seemed that for older releases there was more missing data and this report is only concerned with more recent releases. After observing the rest of the rows, the missing data seemed to be at random, so I decided to use only the 6951 rows with complete data.

Since the categorical variables have many levels (See Appendix 1.1 for a summary of the dataset *vgsales*) and in order to center the numerical variables to avoid multicollinearity issues, I created the following additional columns:

- *critic_score_c* (numerical): Values of *critic_score* minus the column's mean value
- *critic_count_c* (numerical): Values of *critic_count* minus the column's mean value
- *user_score_c* (numerical): Values of *user_score* minus the column's mean value
- *user_count_c* (numerical): Values of *user_count* minus the column's mean value
- *platform_company* (categorical): Company that manufactures the video game's platform
  - 'Nintendo' when *platform* is '3DS', 'DS', 'GB', 'GBA', 'GC', 'N64', 'Wii', or 'WiiU'
  - 'Sega' when *platform* is 'DC'
  - 'PC' when *platform* is 'PC'
  - 'Microsoft' when *platform* is 'X360', 'XB', or 'XOne'
  - 'Sony' when *platform* is 'PS', 'PS2', 'PS3', 'PS4', 'PSP', or 'PSV'
- *rating_everyone* (binary): Indicates if there is not an age restriction for the video game release
  - '1' when *rating* is 'E'
  - '0' when *rating* is not 'E'

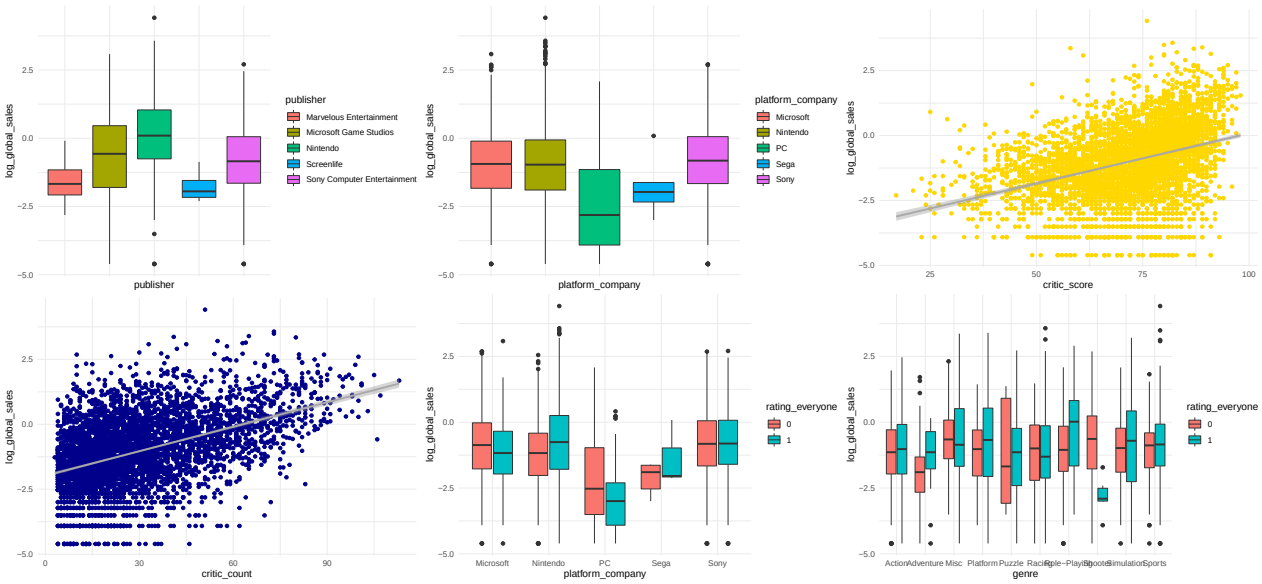
To consider the variable *publisher* into the model, I took a random sample of 50 different publishing companies from the dataset and built a sample dataset to work with. I observed at the distribution of the response variable *global_sales* and noticed that its distribution is not normal. After applying a logarithmic transformation, the distribution is normal, so I added the following column to the dataset:

- *log_global_sales*: Logarithmic transformation of *global_sales*



A summary of the data variables being analyzed can be found in Appendix 1.1.

For the Exploratory Data Analysis, I plotted every relevant variable versus the response variable *log_global_sales*. All of the categorical variables seemed to have different means for the response variable on at least two levels, and the numerical variables seemed to have a positive effect on the response variable. I observed at some interactions between categorical variables (*platform_company*:*rating_everyone*, *platform_company*:*genre*, and *genre*:*rating_everyone*) and since the distributions seemed to vary I took them into account during the model selection. (For a full EDA, see Appendix 1.2).



Model

A detailed description of the model used, how you selected the model, how you selected the variables, model assessment, model validation, and presentation of the model results. What are your overall conclusions in context of the inferential problem(s)?

In order to obtain a final model for the response variable *log_global_sales* various methods for model selection were tested and interactions between predictors were considered as part of the full model. Since the publishing companies in the sample dataset are a sample of the total publishers available, I included random intercept effects for *publisher* in the final model. The categorical and numerical variables (except for *na_sales*, *eu_sales*, *jp_sales*, and *other_sales*, since their sum is exactly equal to *global_sales*) were taken into account for the model selection as predictors. The interaction terms *platform_company*:*genre*, *platform_company*:*rating_everyone*, and *genre*:*rating_everyone* were considered too.

The final model was selected with a manual stepwise approach in R, using AIC and the conditional R_GLMM-squared (conditional R-squared) from the MuMIn package, which is described as "a variance explained by the entire model, including both fixed and random effects" in r.squaredGLMM function documentation. The final model's conditional R-squared is 0.4928372, meaning almost half of the data's variance is explained by the model. The variables *platform*, *genre*, *critic_score_c*, *critic_count_c*, *user_count_c*, as well as the interactions *platform*:*rating_everyone* and *genre*:*rating_everyone* were found to be significant. The variable *rating_everyone* was added to the model because its interactions had significant levels and improved the model considerably. With an AIC value of 12114, the final model has teh following formula:

$$log\_global\_sales = \beta_{(\text{Intercept})} + \gamma_{(\text{Intercept}),j} + \sum_{p \in P} \beta_p \left(platform\_company_p\right)_{i,j} + \sum_{g \in G} \beta_g \left(genre_g\right)_{i,j}$$

$$+\beta_{\text{rating\_everyone}} rating\_everyone_{i,j} + \beta_{\text{critic\_score\_c}} critic\_score\_c_{i,j} + \beta_{\text{critic\_count\_c}} critic\_count\_c_{i,j}$$

$$+\beta_{\text{user\_count\_c}} user\_count\_c_{i,j} + \sum_{p \in P} \beta_{p,\text{rating\_everyone}} \left(platform\_company_p\right)_{i,j} rating\_everyone_{i,j}$$

$$+ \sum_{g \in G} \beta_{g,\text{rating\_everyone}} \left(genre_g\right)_{i,j} rating\_everyone_{i,j} + \epsilon_{i,j};$$

Where:

$$P = \{\text{Nintendo}, \text{PC}, \text{Sega}, \text{Sony}\}$$

$$G = \{\text{Adventure}, \text{Fighting}, \text{Misc}, \text{Platform}, \text{Puzzle}, \text{Racing}, \text{Role-Playing}, \text{Shooter}, \text{Simulation}, \text{Sports}, \text{Strategy}\}$$

$$\epsilon_{i,j} \sim N\left(0, \sigma^2\right) \quad \text{and} \quad \gamma_{(\text{Intercept}),j} \sim N\left(0, \tau^2_{(\text{Intercept})}\right);$$

```
## Computing profile confidence intervals ...
```

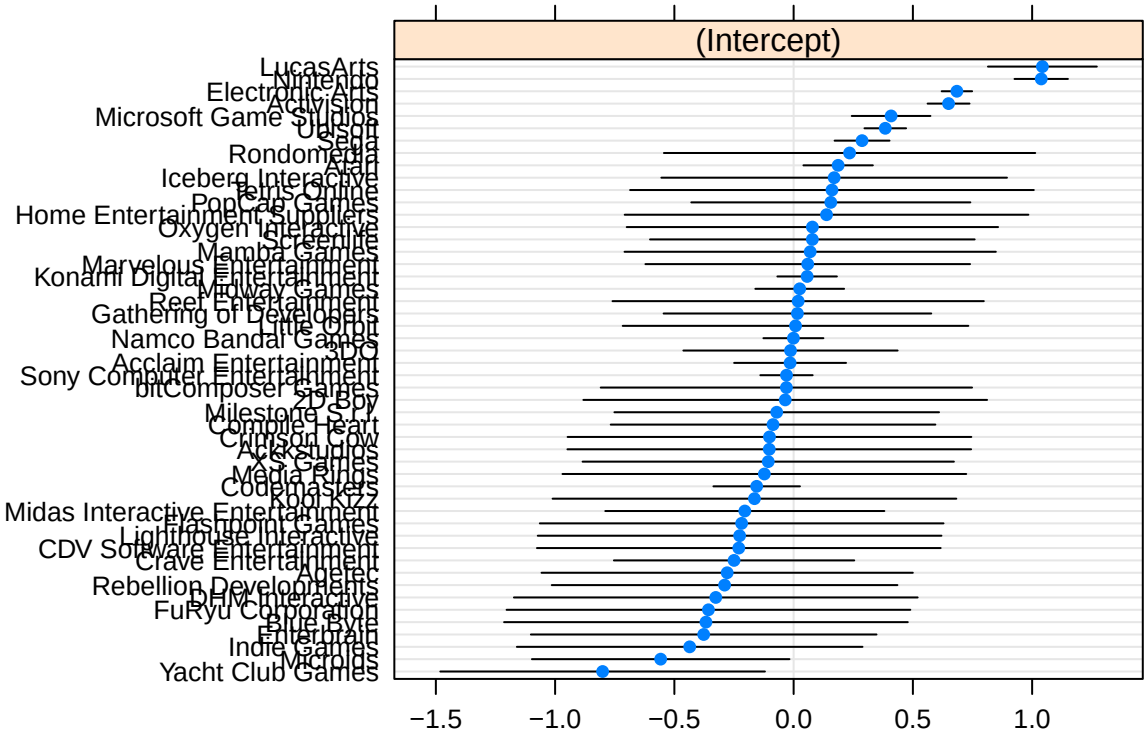
Table 1: Model's Fixed Coefficients

	Estimate	t.value	Pr...t..	2.5 %	97.5 %	Odds
(Intercept)	-1.71	-16.01	0.00	-1.93	-1.49	0.18
platform_companyNintendo	0.08	1.31	0.19	-0.04	0.21	1.09
platform_companyPC	-1.44	-18.27	0.00	-1.59	-1.28	0.24
platform_companySega	-0.66	-1.68	0.09	-1.42	0.11	0.52
platform_companySony	0.46	8.80	0.00	0.36	0.56	1.58
genreAdventure	-0.34	-2.99	0.00	-0.56	-0.12	0.71
genreFighting	0.31	3.61	0.00	0.14	0.47	1.36
genreMisc	0.53	6.00	0.00	0.36	0.71	1.71
genrePlatform	0.06	0.58	0.56	-0.14	0.26	1.06
genrePuzzle	-0.20	-0.74	0.46	-0.75	0.34	0.82
genreRacing	0.15	1.71	0.09	-0.02	0.33	1.16
genreRole-Playing	-0.19	-2.52	0.01	-0.33	-0.04	0.83
genreShooter	0.11	1.83	0.07	-0.01	0.22	1.11
genreSimulation	0.43	4.52	0.00	0.24	0.61	1.53
genreSports	0.14	1.45	0.15	-0.05	0.32	1.15
genreStrategy	-0.50	-4.48	0.00	-0.72	-0.28	0.61
rating_everyone1	0.16	1.37	0.17	-0.07	0.40	1.18
critic_score_c	0.02	17.08	0.00	0.02	0.03	1.02
critic_count_c	0.02	18.85	0.00	0.02	0.02	1.02
user_count_c	0.00	14.91	0.00	0.00	0.00	1.00
platform_companyNintendo:rating_everyone1	0.23	2.39	0.02	0.04	0.43	1.26
platform_companyPC:rating_everyone1	-0.04	-0.28	0.78	-0.35	0.26	0.96
platform_companySega:rating_everyone1	0.36	0.51	0.61	-1.01	1.74	1.43
platform_companySony:rating_everyone1	0.13	1.43	0.15	-0.05	0.30	1.13
genreAdventure:rating_everyone1	-0.13	-0.51	0.61	-0.61	0.36	0.88
genreFighting:rating_everyone1	-0.78	-1.48	0.14	-1.80	0.25	0.46
genreMisc:rating_everyone1	-0.23	-1.46	0.14	-0.54	0.08	0.79
genrePlatform:rating_everyone1	-0.15	-0.93	0.35	-0.46	0.16	0.86
genrePuzzle:rating_everyone1	-0.40	-1.23	0.22	-1.03	0.23	0.67
genreRacing:rating_everyone1	-0.26	-1.79	0.07	-0.54	0.02	0.77
genreRole-Playing:rating_everyone1	0.47	2.59	0.01	0.12	0.83	1.60
genreShooter:rating_everyone1	-1.68	-3.92	0.00	-2.52	-0.84	0.19
genreSimulation:rating_everyone1	-0.10	-0.53	0.60	-0.46	0.26	0.91
genreSports:rating_everyone1	-0.22	-1.58	0.11	-0.48	0.05	0.81
genreStrategy:rating_everyone1	0.07	0.27	0.78	-0.42	0.55	1.07

```
##                 platform_companyNintendo
##                                 2.375819
##                      platform_companyPC
##                                 1.901742
```

```
##                       platform_companySega
##                                   1.459059
##                       platform_companySony
##                                   2.135685
##                             genreAdventure
##                                   1.389655
##                              genreFighting
##                                   1.229708
##                                  genreMisc
##                                   1.934193
##                               genrePlatform
##                                   2.390652
##                                genrePuzzle
##                                   4.537605
##                                genreRacing
##                                   2.479311
##                           genreRole-Playing
##                                   1.521184
##                                genreShooter
##                                   1.513517
##                              genreSimulation
##                                   1.659963
##                                 genreSports
##                                   4.783893
##                                genreStrategy
##                                   1.435584
##                             rating_everyone1
##                                  11.753879
##                               critic_score_c
##                                   1.288739
##                               critic_count_c
##                                   1.587880
##                                 user_count_c
##                                   1.333754
## platform_companyNintendo:rating_everyone1
##                                   3.633627
##       platform_companyPC:rating_everyone1
##                                   1.644967
##     platform_companySega:rating_everyone1
##                                   1.436984
##     platform_companySony:rating_everyone1
##                                   3.286729
##           genreAdventure:rating_everyone1
##                                   1.477025
##            genreFighting:rating_everyone1
##                                   1.058247
##                genreMisc:rating_everyone1
##                                   2.599116
##            genrePlatform:rating_everyone1
##                                   3.330940
##              genrePuzzle:rating_everyone1
##                                   4.882269
##              genreRacing:rating_everyone1
##                                   3.918297
```

```
##         genreRole-Playing:rating_everyone1
##                                   1.603394
##              genreShooter:rating_everyone1
##                                   1.065803
##           genreSimulation:rating_everyone1
##                                   1.948509
##               genreSports:rating_everyone1
##                                   8.517019
##             genreStrategy:rating_everyone1
##                                   1.469225
```



Conclusions

Voting tendency of different demographic groups could be analyzed from the model we built. Baselines of our model are Age 18 - 25, Democratic Party, Female, Asian, and Hispanic/Latino and we could see the following interesting facts.

- Most of variables are significant.
- All county random intercept effects are significant (except for Jackson), which means the odds of voting differed by county in 2016.
- Except mixed race, most of races are less likely vote as compared to the baseline.
- People aged above 40 are approximately 150% more likely to vote than younger people.
- Libertarians are 75% more likely to vote than other parties although they only make a small fraction of the population
- Males belonging to the Republican Party, Libertarian Party or even if they are Unaffiliated are more likely to vote as compared to the baseline. But, overall males are 23% less likely to vote as compared to overall females.
- The Male supporting Democrats are the group that bring the odds down.

Through the multi-level logistic model from dataset from the North Carolina State Board of Elections (NCSBE), we could reveal the voting tendency of different demographic groups in North Carolina in 2016 and predict the turnout rate of each subgroup. From the analysis, we could find that age, party, race, sex,

ethnic, county, etc., are significant factors to decide their tendency of voting.

Although, we could predict turnout rate of voting and voter's tendency, categorical age and unidentified gender, race, ethnic would be clarified to get more precise results.

Appendix

1.1 Dataset Summaries

```
## [1] "vgsales"
```

```
##                                          name         platform
##  Madden NFL 07                             :   9   PS2    :1161
##  LEGO Star Wars II: The Original Trilogy   :   8   X360   : 881
##  Need for Speed: Most Wanted               :   8   PS3    : 791
##  Harry Potter and the Order of the Phoenix :   7   PC     : 692
##  LEGO Batman: The Videogame                :   7   XB     : 581
##  LEGO Indiana Jones: The Original Adventures:   7   Wii    : 492
##  (Other)                                   :6905   (Other):2353
##  year_of_release          genre                             publisher
##  Min.   :2000    Action      :1666   Electronic Arts            : 961
##  1st Qu.:2004    Sports      : 972   Ubisoft                    : 512
##  Median :2007    Shooter     : 884   Activision                 : 505
##  Mean   :2008    Role-Playing: 708   Sony Computer Entertainment: 322
##  3rd Qu.:2011    Racing      : 591   Nintendo                   : 309
##  Max.   :2016    Platform    : 402   THQ                        : 309
##                  (Other)     :1728   (Other)                    :4033
##     na_sales         eu_sales          jp_sales        other_sales
##  Min.   : 0.000   Min.   : 0.0000   Min.   :0.00000   Min.   : 0.00000
##  1st Qu.: 0.060   1st Qu.: 0.0200   1st Qu.:0.00000   1st Qu.: 0.01000
##  Median : 0.150   Median : 0.0600   Median :0.00000   Median : 0.02000
##  Mean   : 0.395   Mean   : 0.2378   Mean   :0.06494   Mean   : 0.08351
##  3rd Qu.: 0.385   3rd Qu.: 0.2100   3rd Qu.:0.01000   3rd Qu.: 0.07000
##  Max.   :41.360   Max.   :28.9600   Max.   :6.50000   Max.   :10.57000
##
##   global_sales      critic_score    critic_count      user_score
##  Min.   : 0.0100   Min.   :13.00   Min.   :  3.00   Min.   : 5.00
##  1st Qu.: 0.1100   1st Qu.:62.00   1st Qu.: 14.00   1st Qu.:66.00
##  Median : 0.2900   Median :72.00   Median : 25.00   Median :76.00
##  Mean   : 0.7814   Mean   :70.14   Mean   : 29.02   Mean   :72.59
##  3rd Qu.: 0.7400   3rd Qu.:80.00   3rd Qu.: 40.00   3rd Qu.:83.00
##  Max.   :82.5300   Max.   :98.00   Max.   :113.00   Max.   :97.00
##
##    user_count                          developer        rating
##  Min.   :    4   Electronic Arts            : 614   T      :2380
##  1st Qu.:   11   Ubisoft                    : 305   E      :2106
##  Median :   27   Konami                     : 148   M      :1448
##  Mean   :  174   Capcom                     : 132   E10+   : 948
##  3rd Qu.:   89   Sony Computer Entertainment: 107          :  66
##  Max.   :10665   Nintendo                   :  85   RP     :   2
##                  (Other)                    :5560   (Other):   1
##  log_global_sales  critic_score_c    critic_count_c     user_score_c
##  Min.   :-4.6052   Min.   :-57.137   Min.   :-26.016   Min.   :-67.586
##  1st Qu.:-2.2073   1st Qu.: -8.137   1st Qu.:-15.016   1st Qu.: -6.586
##  Median :-1.2379   Median :  1.863   Median : -4.016   Median :  3.414
##  Mean   :-1.2509   Mean   :  0.000   Mean   :  0.000   Mean   :  0.000
##  3rd Qu.:-0.3011   3rd Qu.:  9.863   3rd Qu.: 10.984   3rd Qu.: 10.414
##  Max.   : 4.4132   Max.   : 27.863   Max.   : 83.984   Max.   : 24.414
##
##   user_count_c       platform_company rating_everyone
```

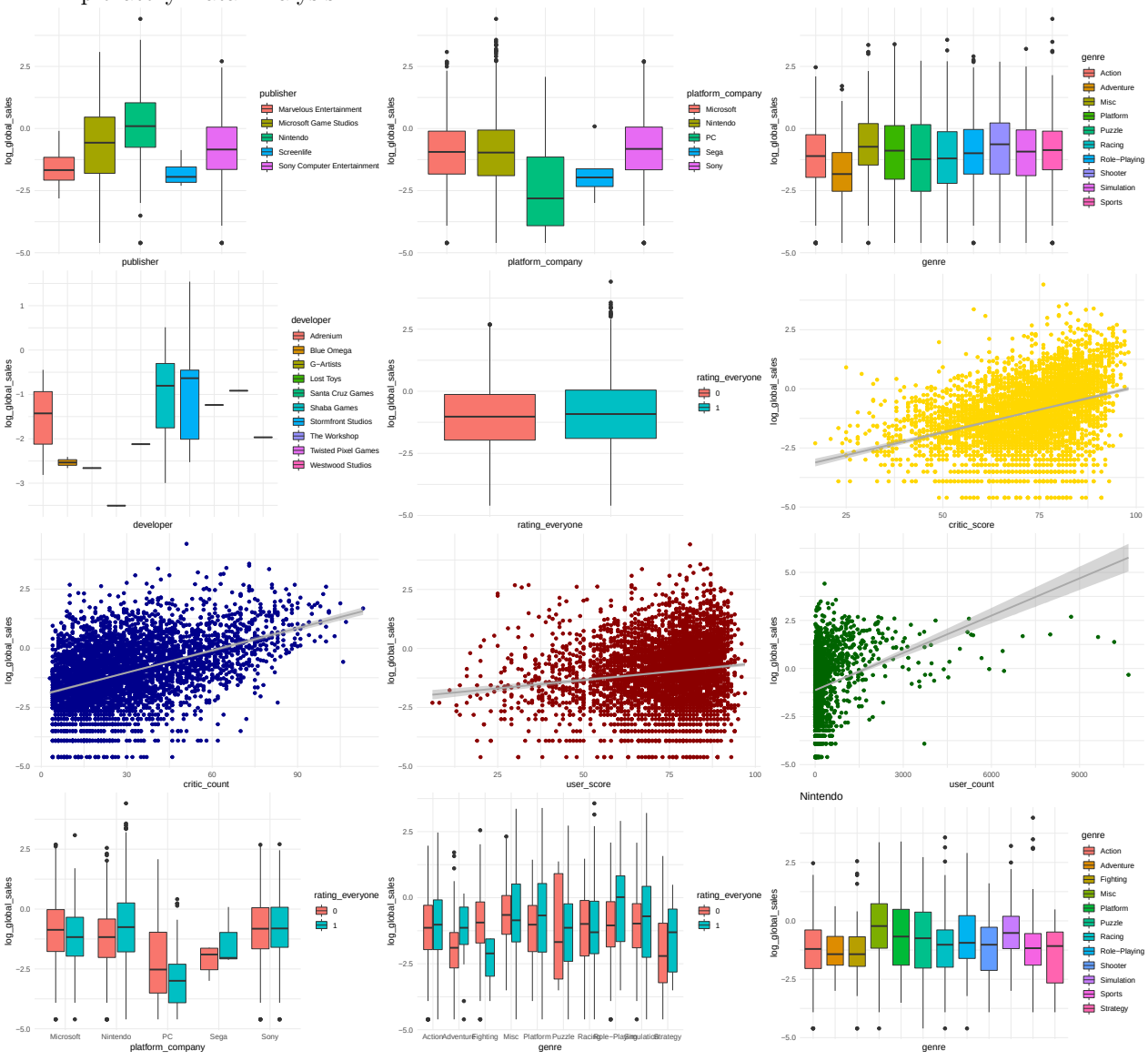
```
##  Min.   : -169.96   Microsoft:1628    0:4845
##  1st Qu.: -162.96   Nintendo :1818    1:2106
##  Median : -146.96   PC       : 692
##  Mean   :    0.00   Sega     :  11
##  3rd Qu.:  -84.96   Sony     :2802
##  Max.   :10491.04
##
```

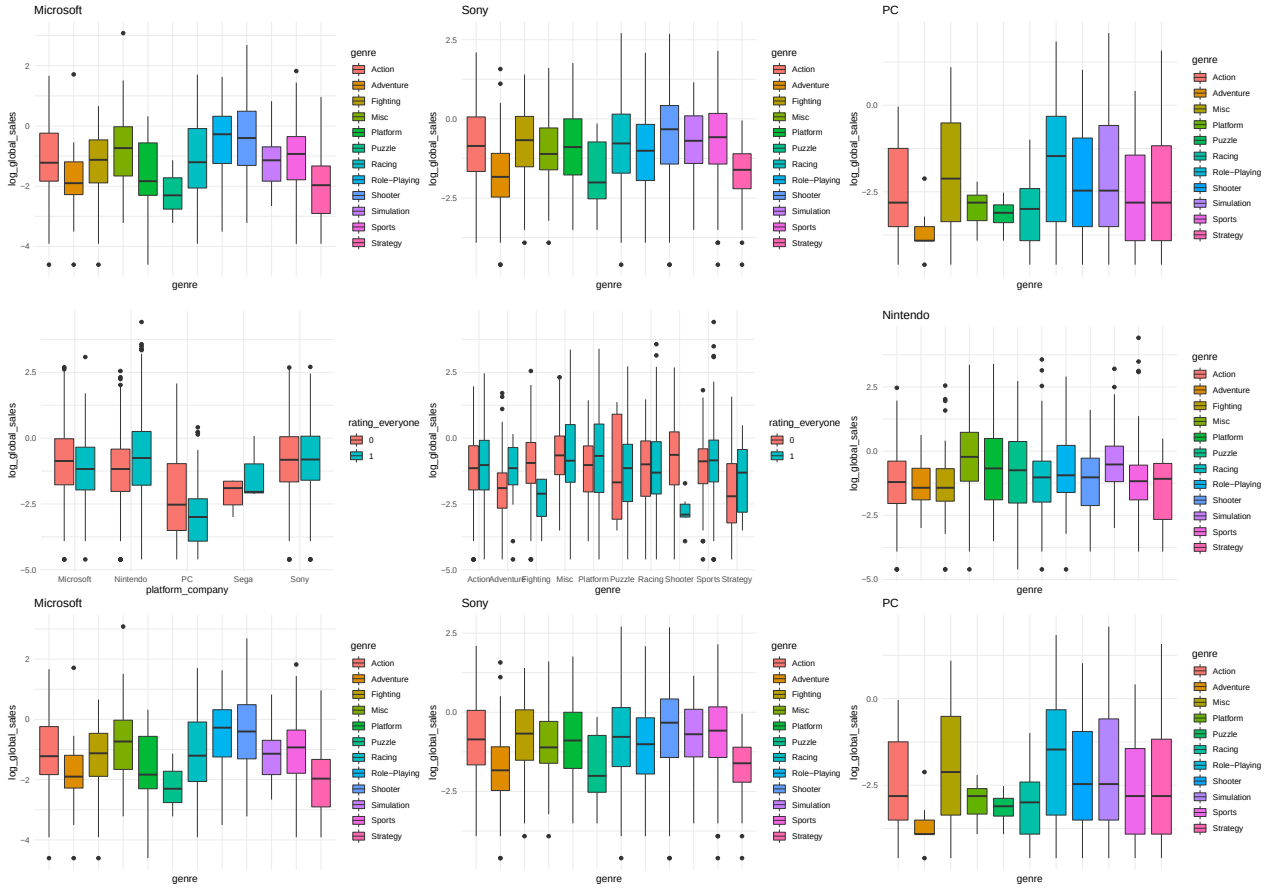
```
## [1] "sample_data"
```

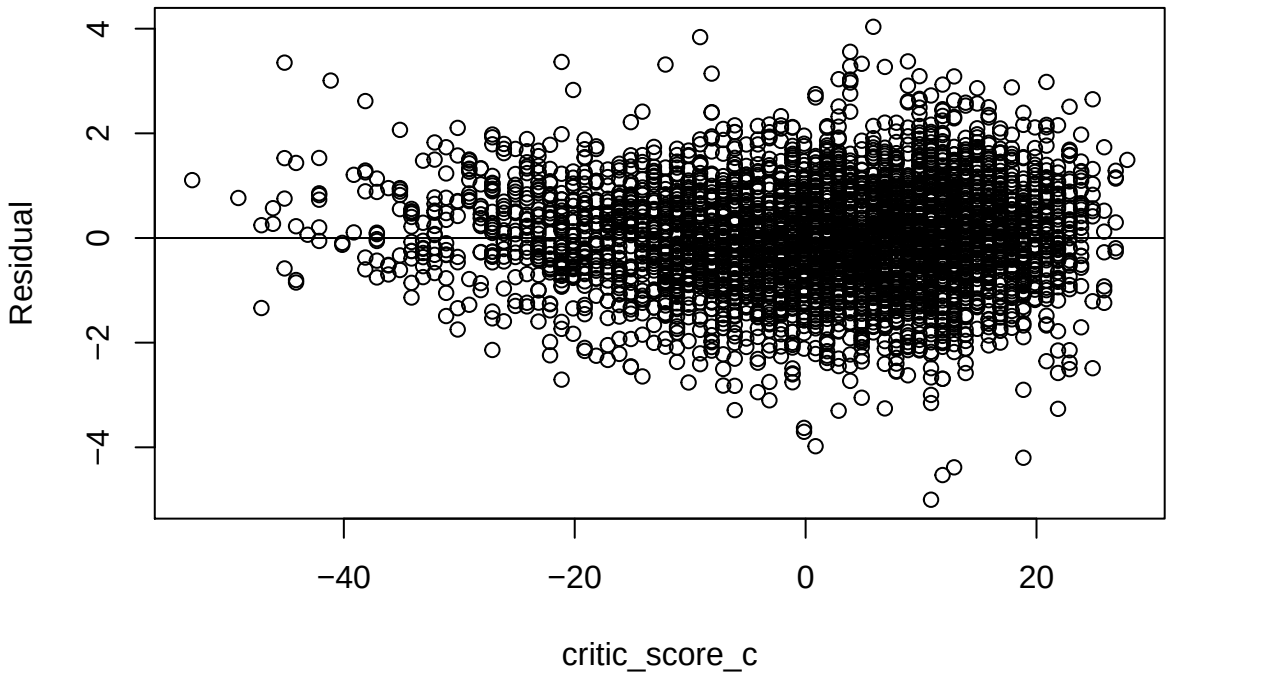
```
##                                        name         platform
##  Madden NFL 07                           :   9   PS2    : 717
##  LEGO Star Wars II: The Original Trilogy :   8   X360   : 516
##  Need for Speed: Most Wanted             :   8   PS3    : 478
##  Harry Potter and the Order of the Phoenix  :   7   XB     : 391
##  LEGO Indiana Jones: The Original Adventures:   7   PC     : 362
##  Madden NFL 08                           :   7   Wii    : 306
##  (Other)                                 :4149   (Other):1425
##  year_of_release          genre                          publisher
##  Min.   :2000    Action      :899   Electronic Arts            : 961
##  1st Qu.:2004    Sports      :793   Ubisoft                    : 512
##  Median :2007    Shooter     :529   Activision                 : 505
##  Mean   :2007    Racing      :396   Sony Computer Entertainment: 322
##  3rd Qu.:2010    Role-Playing:324   Nintendo                   : 309
##  Max.   :2016    Misc        :275   Sega                       : 292
##                  (Other)     :979   (Other)                    :1294
##     na_sales          eu_sales          jp_sales        other_sales
##  Min.   : 0.0000   Min.   : 0.0000   Min.   :0.00000   Min.   :0.0000
##  1st Qu.: 0.0800   1st Qu.: 0.0200   1st Qu.:0.00000   1st Qu.:0.0100
##  Median : 0.1900   Median : 0.0800   Median :0.00000   Median :0.0300
##  Mean   : 0.4842   Mean   : 0.2998   Mean   :0.08069   Mean   :0.1014
##  3rd Qu.: 0.4800   3rd Qu.: 0.2600   3rd Qu.:0.01000   3rd Qu.:0.0900
##  Max.   :41.3600   Max.   :28.9600   Max.   :6.50000   Max.   :8.4500
##
##   global_sales      critic_score    critic_count      user_score
##  Min.   : 0.0100   Min.   :17.00   Min.   :  3.00   Min.   : 6.00
##  1st Qu.: 0.1500   1st Qu.:64.00   1st Qu.: 16.00   1st Qu.:66.00
##  Median : 0.3700   Median :74.00   Median : 26.00   Median :77.00
##  Mean   : 0.9663   Mean   :71.58   Mean   : 30.77   Mean   :73.13
##  3rd Qu.: 0.9300   3rd Qu.:81.50   3rd Qu.: 42.00   3rd Qu.:83.00
##  Max.   :82.5300   Max.   :98.00   Max.   :113.00   Max.   :97.00
##
##    user_count                          developer        rating
##  Min.   :    4.0   Electronic Arts            : 612   E      :1493
##  1st Qu.:   12.0   Ubisoft                    : 305   T      :1396
##  Median :   29.0   Konami                     : 146   M      : 732
##  Mean   :  181.4   Sony Computer Entertainment: 107   E10+   : 548
##  3rd Qu.:   96.0   Nintendo                   :  85          :  26
##  Max.   :10665.0   Codemasters                :  66   AO     :   0
##                    (Other)                    :2874   (Other):   0
##  log_global_sales   critic_score_c    critic_count_c     user_score_c
##  Min.   :-4.60517   Min.   :-53.137   Min.   :-26.016   Min.   :-66.5857
##  1st Qu.:-1.89712   1st Qu.: -6.137   1st Qu.:-13.016   1st Qu.: -6.5857
##  Median :-0.99425   Median :  3.863   Median : -3.016   Median :  4.4143
##  Mean   :-1.01178   Mean   :  1.448   Mean   :  1.749   Mean   :  0.5435
```

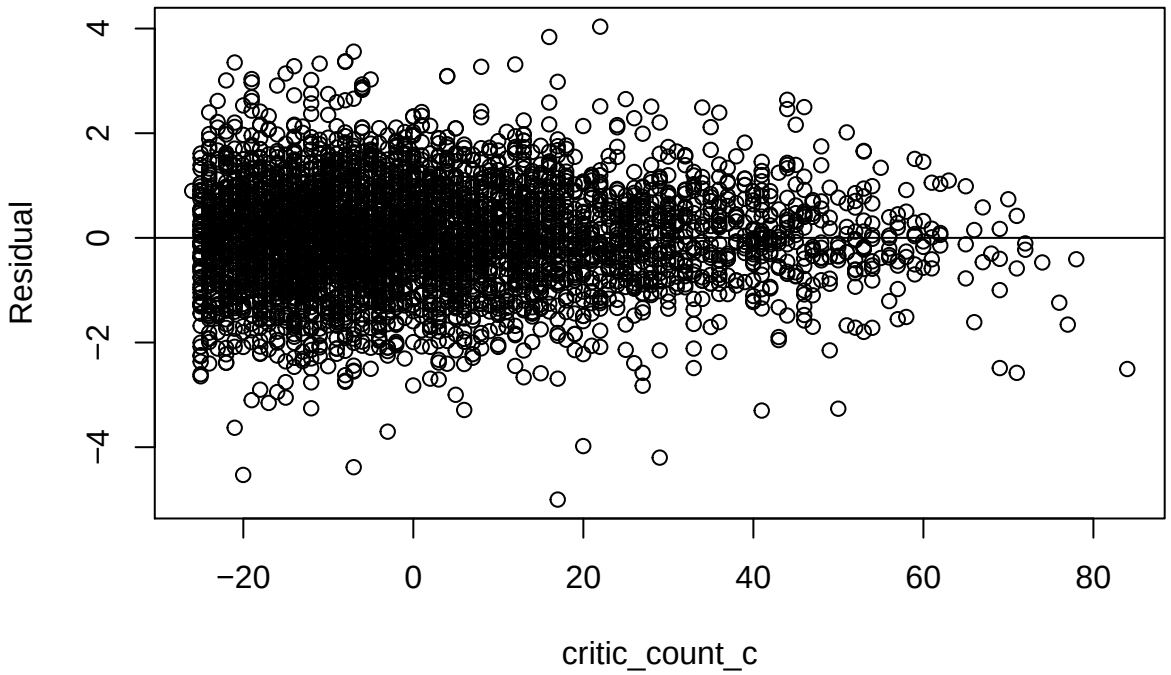
```
##  3rd Qu.:-0.07257   3rd Qu.: 11.363   3rd Qu.: 12.984   3rd Qu.: 10.4143
##  Max.   : 4.41316   Max.   : 27.863   Max.   : 83.984   Max.   : 24.4143
##
##   user_count_c         platform_company rating_everyone
##  Min.   : -169.960   Microsoft:1005    0:2702
##  1st Qu.: -161.960   Nintendo :1130    1:1493
##  Median : -144.960   PC       : 362
##  Mean   :    7.464   Sega     :  10
##  3rd Qu.:  -77.960   Sony     :1688
##  Max.   :10491.040
##
```

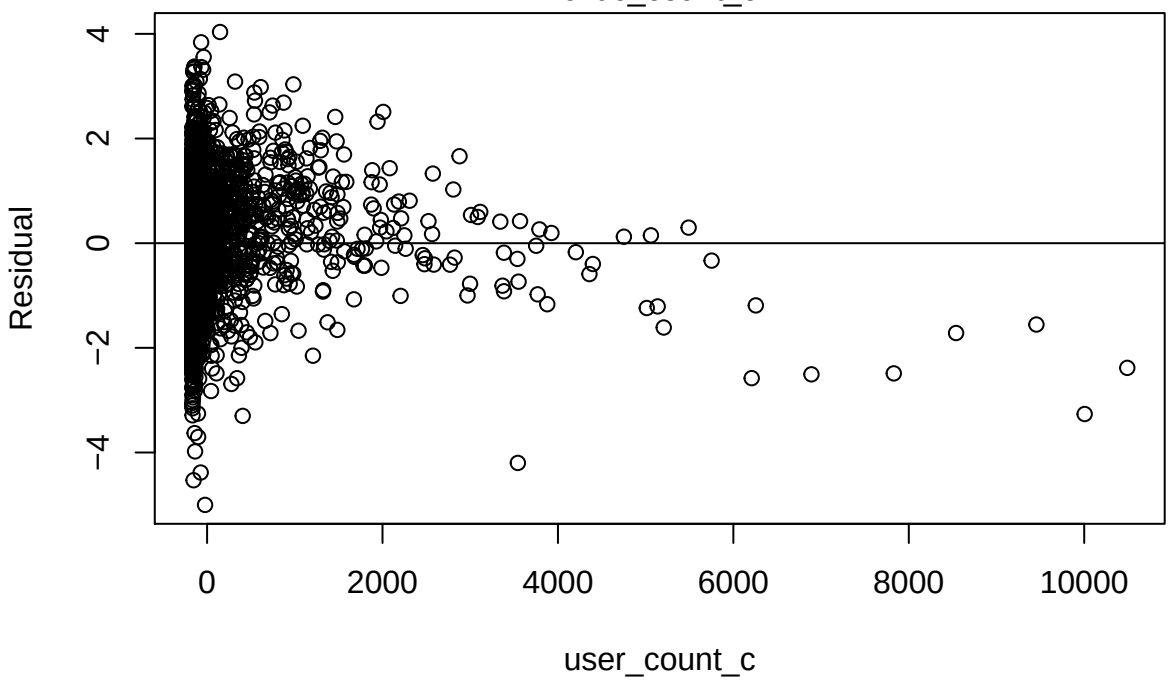
1.2. Exploratory Data Analysis

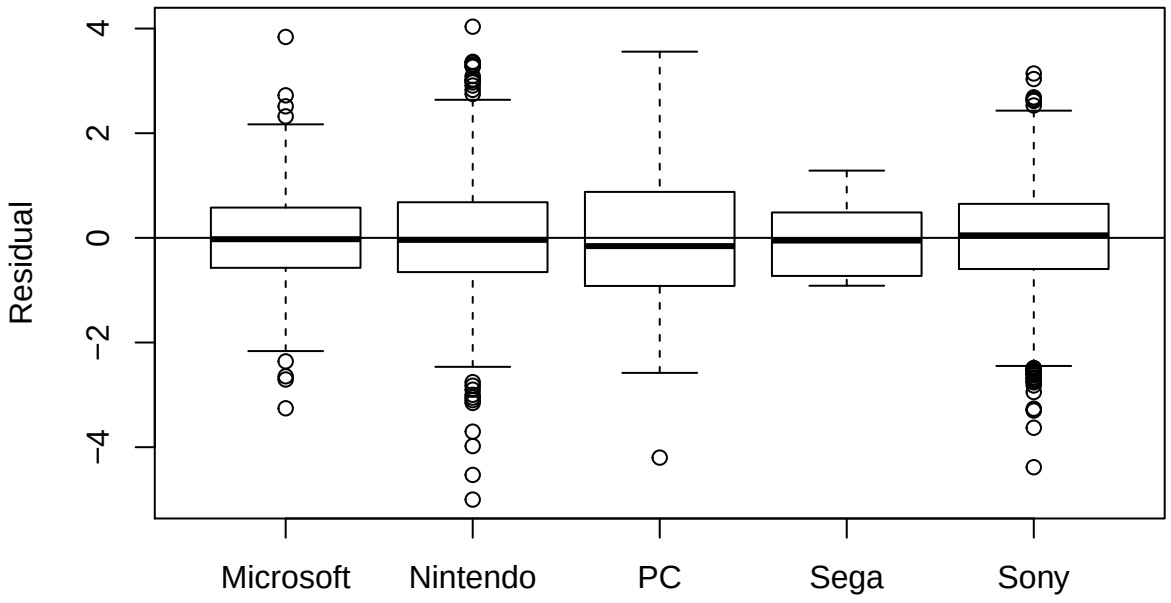


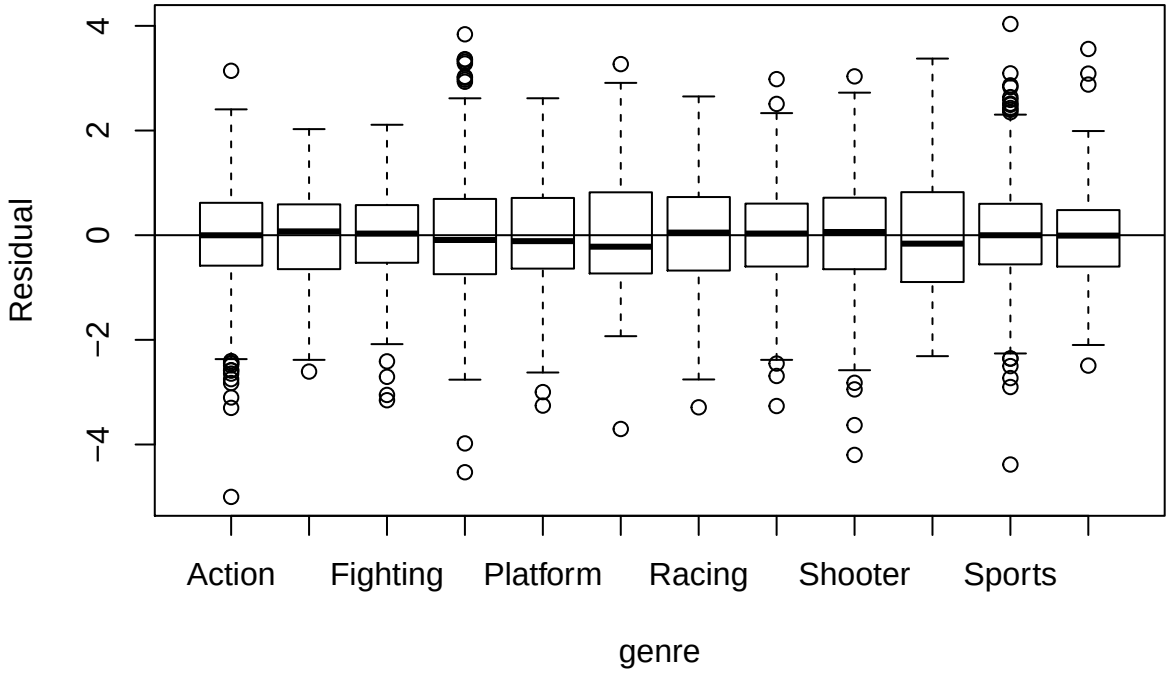

Microsoft
Sony
PC
Nintendo
log_global_sales
genre
Action
Adventure
Fighting
Misc
Platform
Puzzle
Racing
Role−Playing
Shooter
Simulation
Sports
Strategy
rating_everyone
0
1
platform_company
Microsoft
Nintendo
PC
Sega
Sony

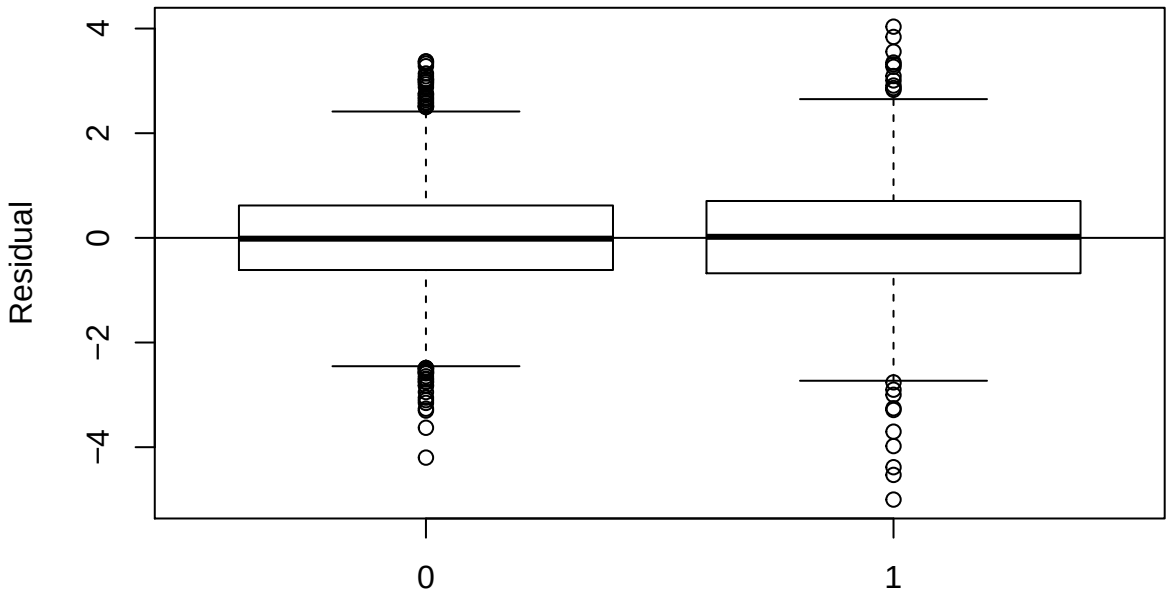

Residual
critic_score_c
−40
−20
0
20
−4
−2
0
2
4

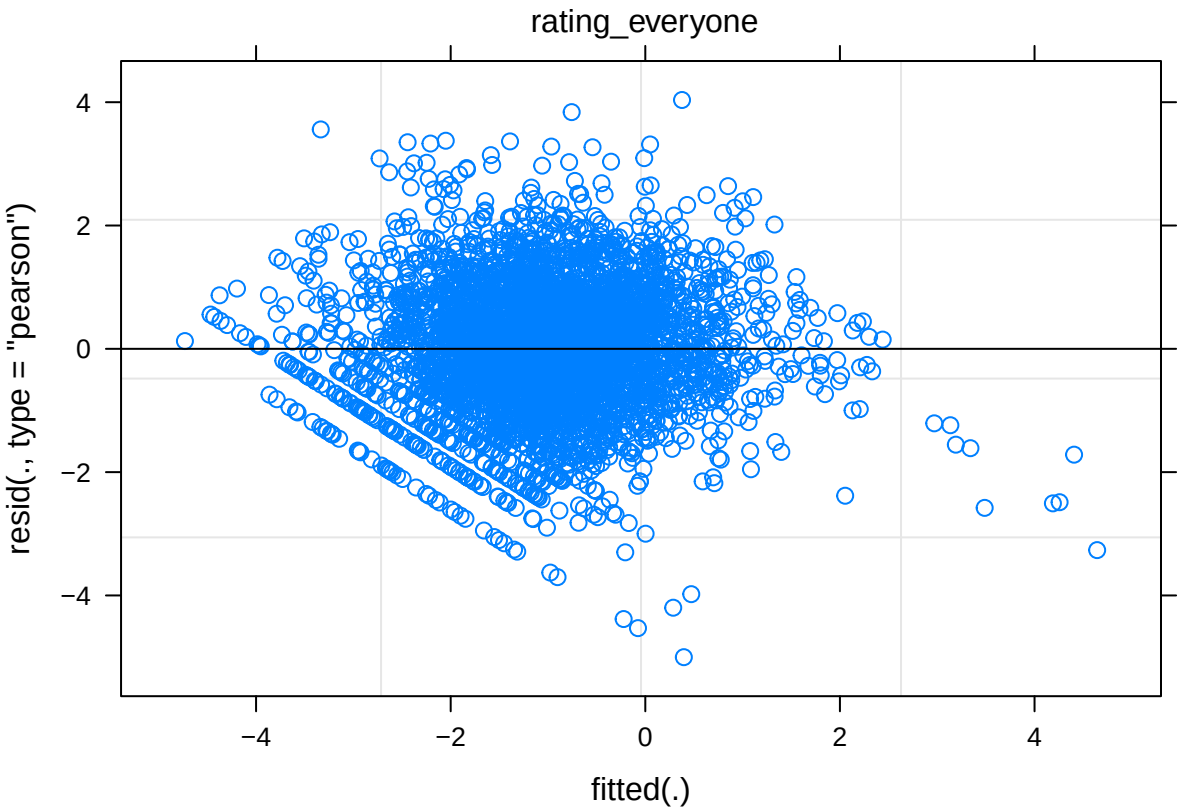

Residual
4
2
0
−2
−4
−20
0
20
40
60
80
critic_count_c

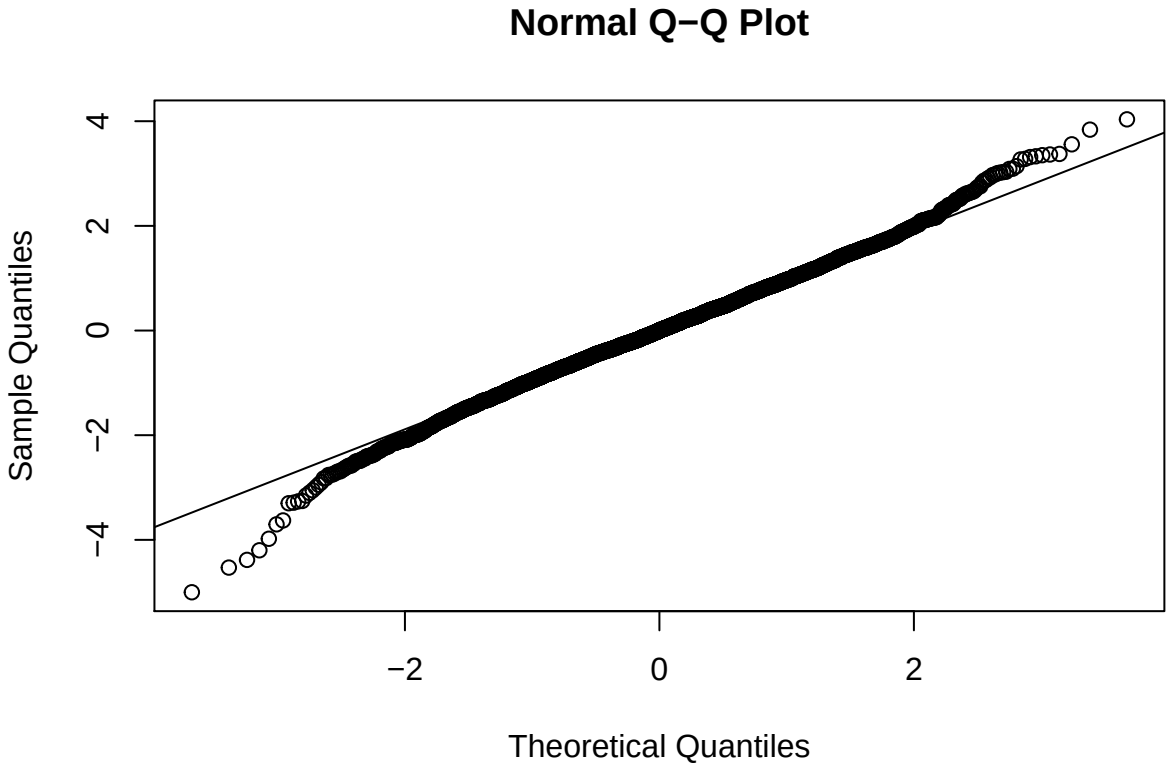

Residual
4
2
0
−2
−4
0
2000
4000
6000
8000
10000
user_count_c


Residual
4
2
0
−2
−4
Microsoft
Nintendo
PC
Sega
Sony
platform_company


Residual
4
2
0
−2
−4
Action
Fighting
Platform
Racing
Shooter
Sports
genre


Residual
4
2
0
−2
−4
0
1


rating_everyone
resid(., type = "pearson")
4
2
0
−2
−4
−4
−2
0
2
4
fitted(.)


Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles
4
2
0
−2
−4
−2
0
2