Final Project

Sebastián Soriano Pérez [ss1072]

12/10/2019

Analyzing Video Game Sales

Summary

This report will try to find what factors can help determine whether a video game release will be successful in terms of the number of units it sells worldwide, and which of these factors have the strongest impact on sales. This analysis uses a public dataset that gathers data for video games that have sold more than 100,000 copies worldwide from 1976 to 2016, and includes the video game's development and release information, as well as information on the public reception of the video game.

By creating a hierarchical multiple linear regression model with random intercept effects, it was found that the most significant predictors are the platform manufacturer, the genre of the videogame, the number of users and critics that review the video game on Metacritic, as well as the aggregated critic score it receives.

Introduction

The video game industry has been growing consistently during the last two decades, and in 2017 it was worth more than 122 billion USD worldwide according to an article on Bloomberg. 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.

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 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 AIC and conditional R-squared as selection criteria.

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. The variables in the final model were found to be significant in predicting the global sales for a video game.

Data

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:

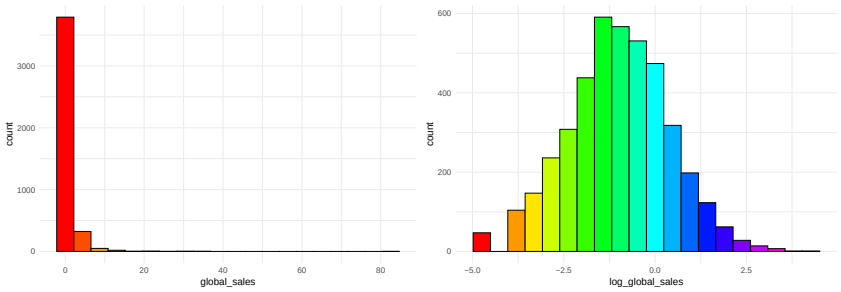
- *name* (categorical): Name of the video game
- *platform* (categorical): PLatform or console for which the video game was released
- *year_of_release* (categorical): Year of first release
- *genre* (categorical): Genre of the video game
- *publisher* (categorical): Publishing company
- *na_sales* (numerical): 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.

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*



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'

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

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:

$$
\begin{aligned}
\mathit{log\_global\_sales}_{i,j} &= \beta_{(\text{Intercept})} + \gamma_{(\text{Intercept}),j} + \sum_{p \in P} \beta_p \left(\mathit{platform\_company}_p\right)_{i,j} + \sum_{g \in G} \beta_g \left(\mathit{genre}_g\right)_{i,j} \\
&+ \beta_{\text{rating\_everyone}} \mathit{rating\_everyone}_{i,j} + \beta_{\text{critic\_score\_c}} \mathit{critic\_score\_c}_{i,j} + \beta_{\text{critic\_count\_c}} \mathit{critic\_count\_c}_{i,j} \\
&+ \beta_{\text{user\_count\_c}} \mathit{user\_count\_c}_{i,j} + \sum_{p \in P} \beta_{p,\text{rating\_everyone}} \left(\mathit{platform\_company}_p\right)_{i,j} \mathit{rating\_everyone}_{i,j} \\
&+ \sum_{g \in G} \beta_{g,\text{rating\_everyone}} \left(\mathit{genre}_g\right)_{i,j} \mathit{rating\_everyone}_{i,j} + \epsilon_{i,j};
\end{aligned}
$$

Where:

$j \in J$; (J is the set of 50 publishers, for the full set see Appendix 1.3.)

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

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

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

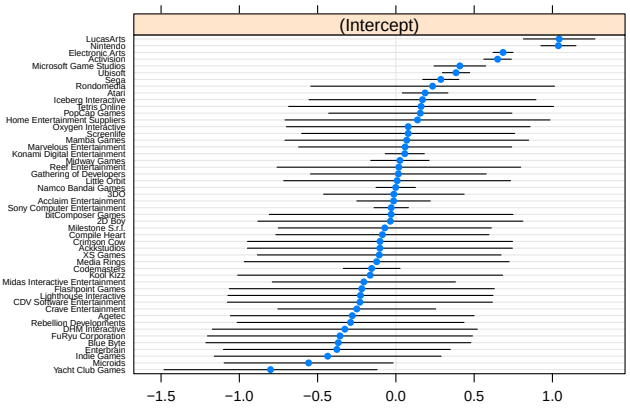
The final model's fixed coefficients have the following values for every β coefficient:

Table 1: Model's Fixed Coefficients

	Estimate	t.value	Pr...t..	2.5 %	97.5 %	Sales
(Intercept)	-1.63	-15.34	0.00	-1.86	-1.42	0.20
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

The baseline intercept indicates that a game of no particular publisher, released on a Microsoft Console, of the Action genre, with an age-restriction rating, with critic score of 71.58, reviewed by 30.77 critics and 181.4 users would sell 0.20 million (M) copies worldwide. The most significant variables is *critic_count_c*. A one-unit increase (31.77 critic reviews in total) would indicate a video game could be expected to sell 2% more units. The other significant variables are $platform\_company_{\text{PC}}$ (76% less sales), *critic_score_c* (2% more sales), *user_count_c* (0.05% more sales), $platform\_company_{\text{Sony}}$ (58% more sales), $genre_{\text{Misc}}$ (71% more sales), $genre_{\text{Simulation}}$ (53% more sales), $genre_{\text{Strategy}}$ (39% less sales), $genre_{\text{Shooter}}$:*rating_everyone* (additional 81% less sales), $genre_{\text{Fighting}}$ (36% more sales), $genre_{\text{Adventure}}$ (29% less sales), $genre_{\text{Role-Playing}}$:*rating_everyone* (additional 60% more sales), $platform\_company_{\text{Nintendo}}$:*rating_everyone* (additional 26% more sales), and $genre_{\text{Role-Playing}}$ (17% less sales) in order of significance.

The random intercept effects by plublisher or $\gamma_{(\text{Intercept})}$ random effects are shown in the following plot (See Appendix 1.3. for a table with the point estimates):



As for the fixed effects coefficients,

It can be observed that only 10 out of the 50 publishers have significant (with 95% confidence) random intercept effects: LucasArts (a baseline game would sell 0.55M copies), Nintendo (0.55M), Electronic Arts (0.39M), Activision (0.37M), Microsoft Game Studios (0.29M), Ubisoft (0.29M), Sega (0.26M), Atari (0.24M), Microids (0.11M), and Yacht Club Games (0.09M) (the latter two have negative random intercept effects). It is notewordthy that the first 8 publishers release the largest number of videogames and are among the best known video game companies in the world.

All of the VIF values of the model predictors are below 5 with the exception of *rating_everyone* (See Appendix 1.4. for each individual VIF value). I decided to keep the variable in the model because the interactions between *rating_everyone* and the categorical variables improve the model significantly. For the most part, the model does not show any serious multicollinearity issues. The model assumptions of linearity, independence, and equal variance seem to be met. The normality as- sumptions may be a reason to worry about since the Q-Q shows tails on both sides. (See Appendix 1.5.)

Conclusions

The final model indicates that the main factors that can help us determine how many units a video game will sell globally are the platform manufacturer, the genre of the videogame, the number of users and critics that review the video game on Metacritic, as well as the aggregated critic score it receives. The platforms in which video games sell the most units are made by Sony, belong to the Misc genre, and are rated E for everyone. The better the critic score a video game receives, and the more users and critics reviews it gets, the more units it will sell worldwide.

To improve the model I would start with the collection of data. Building a better web scrapping script I could get complete data for more video games, including recent years. A restriction on the publishing companies to only those that have sold more than a certain amount of units worldwide could also improve the analysis. To avoid the multicollinearity issue with *rating_everyone* I would use the variable *rating* instead, although the model interpretation and the interaction effects would become way more complicated to interpret, and the model wouldn't be practical or useful.

Although the conditional R-squared value is just a little bit below 0.5, I believe this hierarchical model does represent the reality well enough. The variables that were found to be significant confirm the information observed during the exploratory data analysis, and correspond to my knowledge and understanding of the video game industry. The most well-known publishing companies, best selling platform manufacturers, and most popular genres correspond to the interpretation of the model.

Appendix

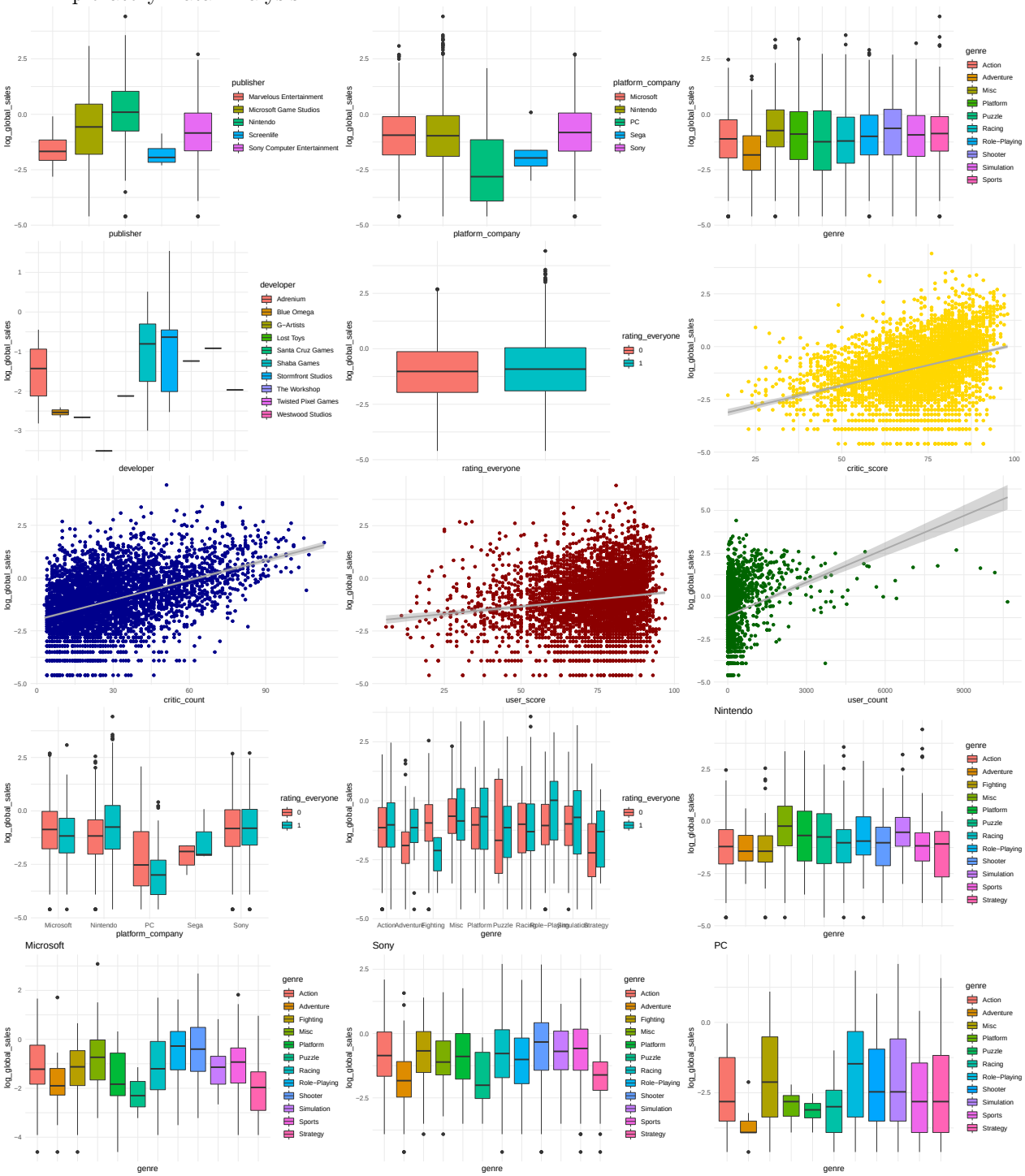
1.1 Dataset Summary

```
## [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    platform_company rating_everyone critic_score_c
##  Min.   :-4.60517   Microsoft:1005    0:2702          Min.   :-54.585
##  1st Qu.:-1.89712   Nintendo :1130    1:1493          1st Qu.: -7.585
##  Median :-0.99425   PC       : 362                    Median :  2.415
##  Mean   :-1.01178   Sega     :  10                    Mean   :  0.000
##  3rd Qu.:-0.07257   Sony     :1688                    3rd Qu.:  9.915
##  Max.   : 4.41316                                    Max.   : 26.415
##
##  critic_count_c     user_score_c      user_count_c
```

```
##  Min.   :-27.766   Min.   :-67.129   Min.   : -177.42
##  1st Qu.:-14.766   1st Qu.: -7.129   1st Qu.: -169.42
##  Median : -4.766   Median :  3.871   Median : -152.42
##  Mean   :  0.000   Mean   :  0.000   Mean   :    0.00
##  3rd Qu.: 11.234   3rd Qu.:  9.871   3rd Qu.:  -85.42
##  Max.   : 82.234   Max.   : 23.871   Max.   :10483.58
##
```

1.2. Exploratory Data Analysis



1.3. Random Effects for the Intercept by Publisher

Table 2: Random Effects for the Intercept by Publisher

	gamma	log_global_sales Scale	global_sales Scale
2D Boy	-0.04	-1.67	0.19
3DO	-0.01	-1.65	0.19
Acclaim Entertainment	-0.01	-1.65	0.19
Ackkstudios	-0.10	-1.74	0.18
Activision	0.65	-0.98	0.37
Agetec	-0.28	-1.91	0.15
Atari	0.19	-1.45	0.24
bitComposer Games	-0.03	-1.66	0.19
Blue Byte	-0.37	-2.00	0.14
CDV Software Entertainment	-0.23	-1.86	0.16
Codemasters	-0.15	-1.79	0.17
Compile Heart	-0.09	-1.72	0.18
Crave Entertainment	-0.25	-1.88	0.15
Crimson Cow	-0.10	-1.73	0.18
DHM Interactive	-0.33	-1.96	0.14
Electronic Arts	0.68	-0.95	0.39
Enterbrain	-0.38	-2.01	0.13
Flashpoint Games	-0.22	-1.85	0.16
FuRyu Corporation	-0.36	-1.99	0.14
Gathering of Developers	0.02	-1.62	0.20
Home Entertainment Suppliers	0.14	-1.50	0.22
Iceberg Interactive	0.17	-1.46	0.23
Indie Games	-0.44	-2.07	0.13
Konami Digital Entertainment	0.06	-1.58	0.21
Kool Kizz	-0.16	-1.80	0.17
Lighthouse Interactive	-0.23	-1.86	0.16
Little Orbit	0.01	-1.63	0.20
LucasArts	1.04	-0.59	0.55
Mamba Games	0.07	-1.56	0.21
Marvelous Entertainment	0.06	-1.57	0.21
Media Rings	-0.12	-1.76	0.17
Microids	-0.56	-2.19	0.11
Microsoft Game Studios	0.41	-1.22	0.29
Midas Interactive Entertainment	-0.20	-1.84	0.16
Midway Games	0.03	-1.61	0.20
Milestone S.r.l.	-0.07	-1.70	0.18
Namco Bandai Games	0.00	-1.63	0.20
Nintendo	1.04	-0.60	0.55
Oxygen Interactive	0.08	-1.55	0.21
PopCap Games	0.16	-1.48	0.23
Rebellion Developments	-0.29	-1.92	0.15
Reef Entertainment	0.02	-1.61	0.20
Rondomedia	0.23	-1.40	0.25
Screenlife	0.08	-1.55	0.21
Sega	0.29	-1.35	0.26

Table 3: Random Effects for the Intercept by Publisher (continued)

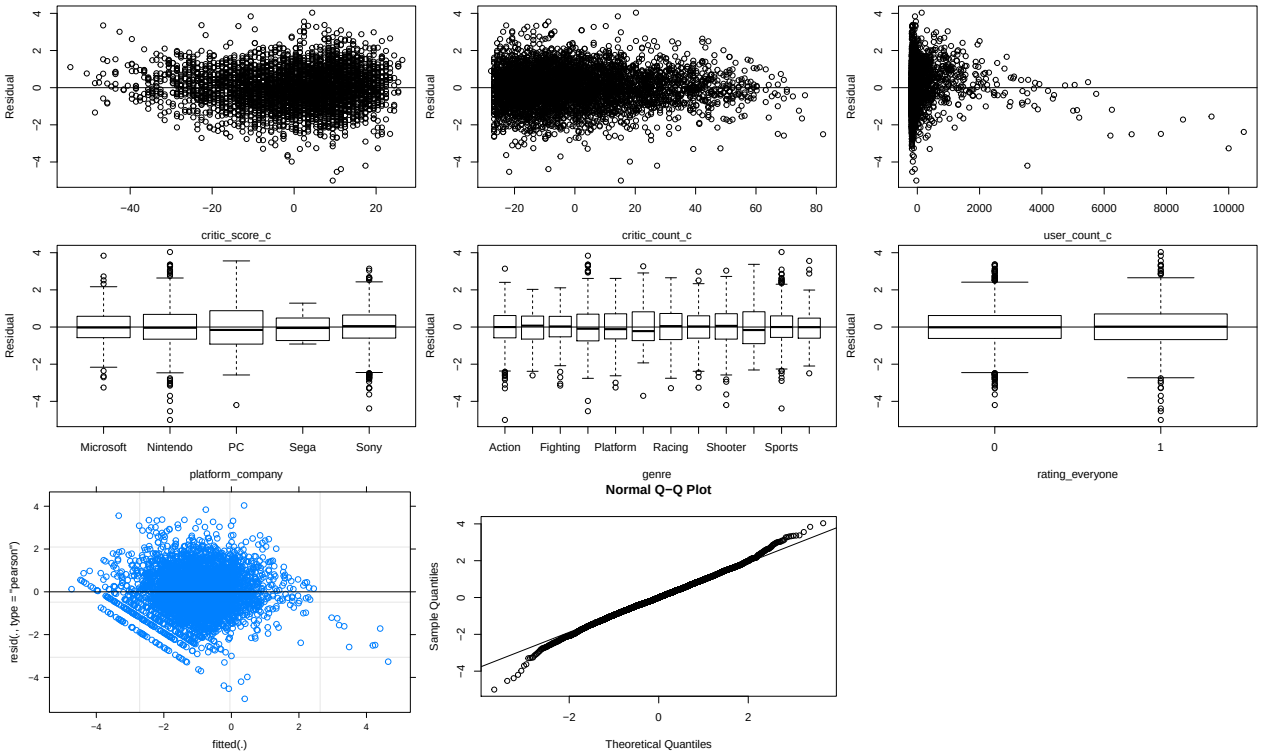
	gamma	log_global_sales	global_sales
Sony Computer Entertainment	-0.03	-1.66	0.19
Tetris Online	0.16	-1.47	0.23
Ubisoft	0.38	-1.25	0.29
XS Games	-0.11	-1.74	0.18
Yacht Club Games	-0.80	-2.43	0.09

1.4. Model's VIFs for the Fixed Effects

Table 4: Model's VIFs for the Fixed Effects

	x
platform_companyNintendo	2.38
platform_companyPC	1.90
platform_companySega	1.46
platform_companySony	2.14
genreAdventure	1.39
genreFighting	1.23
genreMisc	1.93
genrePlatform	2.39
genrePuzzle	4.54
genreRacing	2.48
genreRole-Playing	1.52
genreShooter	1.51
genreSimulation	1.66
genreSports	4.78
genreStrategy	1.44
rating_everyone1	11.75
critic_score_c	1.29
critic_count_c	1.59
user_count_c	1.33
platform_companyNintendo:rating_everyone1	3.63
platform_companyPC:rating_everyone1	1.64
platform_companySega:rating_everyone1	1.44
platform_companySony:rating_everyone1	3.29
genreAdventure:rating_everyone1	1.48
genreFighting:rating_everyone1	1.06
genreMisc:rating_everyone1	2.60
genrePlatform:rating_everyone1	3.33
genrePuzzle:rating_everyone1	4.88
genreRacing:rating_everyone1	3.92
genreRole-Playing:rating_everyone1	1.60
genreShooter:rating_everyone1	1.07
genreSimulation:rating_everyone1	1.95
genreSports:rating_everyone1	8.52
genreStrategy:rating_everyone1	1.47

1.5. Model Assumptions and Validation


Residual
critic_score_c
critic_count_c
user_count_c
Microsoft
Nintendo
PC
Sega
Sony
Action
Fighting
Platform
Racing
Shooter
Sports
genre
rating_everyone
platform_company
resid(., type = "pearson")
fitted(.)
Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles