MA678 Final Project

Chenxuan Xiong

2023-12-10

Abstract

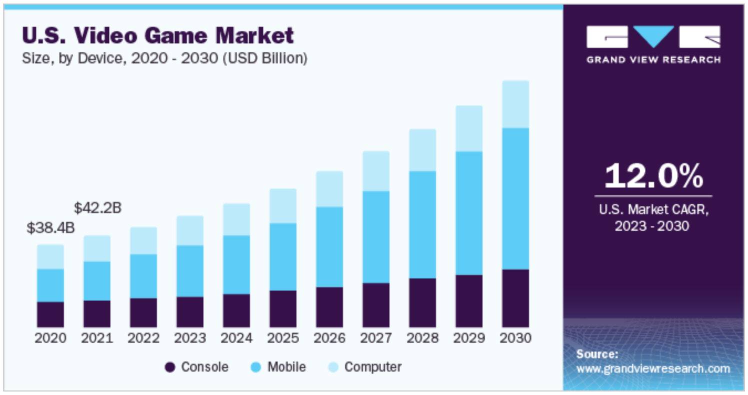
This study uses linear regression to build a video game pricing model. This model uses Game Features and device requirements as predictors to regress the Original price. Through this model, we can come up with a reasonable approximate game pricing for game companies to use as a reference.

1. Introduction

In the changing world of video games making the right pricing decisions is crucial, for a game's success. This study uses regression to create a concise pricing model focusing on game features and device requirements as factors that can predict pricing. The aim is to provide game companies with a guide for setting prices. By analyzing the relationships between these factors and the initial prices of games our model aims to simplify the dynamics of pricing into practical insights. This research offers data-driven approaches that empower industry players to navigate the market intricacies, with accuracy.

1.1. Literature Review

These days more and more people are willing to pay for digital content. Consumer’s willingness to pay is a key factor that drives the size of the game market to grow. The global video game market size was estimated at USD 217.06 billion in 2022 and is expected to grow at a compound annual growth rate (CAGR) of 13.4% from 2023 to 2030. (Grand View Research. 2021). As the industry is growing, the pricing of video games become more and more important. Reasonable game price functions not just as a signal of a game’s value but also as a sign of a game’s legitimacy or its lack in the eyes of developers and the larger game industry (Consalvo and Paul 2014).



1.2. Data Set

The data is downloaded on Kaggle: https://www.kaggle.com/datasets/nikatomashvili/steam-games-dataset/data.

The data was gathered by scraping the rolling page of the Steam search site in early September. The process began with scraping the Steam video game platform utilizing a rolling page, where games were continuously loaded through scrolling until completion. During the initial scraping phase, the team acquired game data, specifically the name, price, discounted price, release date, and link. Subsequently, leveraging the links, they expanded the dataset and extracted additional information for each game. The two datasets were then merged into a single file.

(Steam Search site:
https://store.steampowered.com/search/?category1=998&ndl=1&ignore_preferences=1.)

Due to memory constraints, we only use the first 5000 games to build the model and then use 5000 to 6000 as validation in this project. The model can be extended to the whole data set to get a more accurate result.

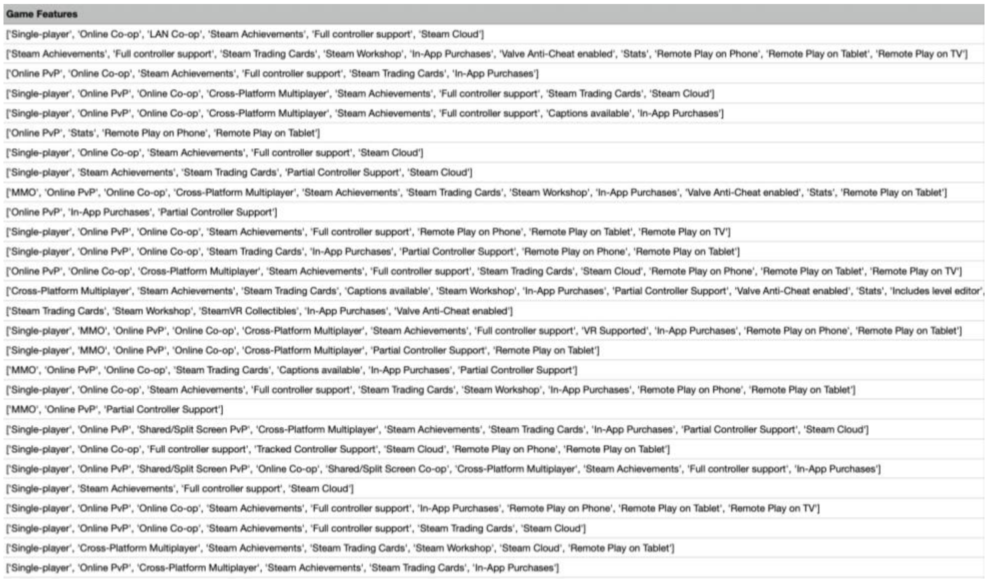
1.3. Research Objectives

1. Which features will have a significant impact on the price of the game?
2. What's the reasonable price for a typical type of game?

2. Method

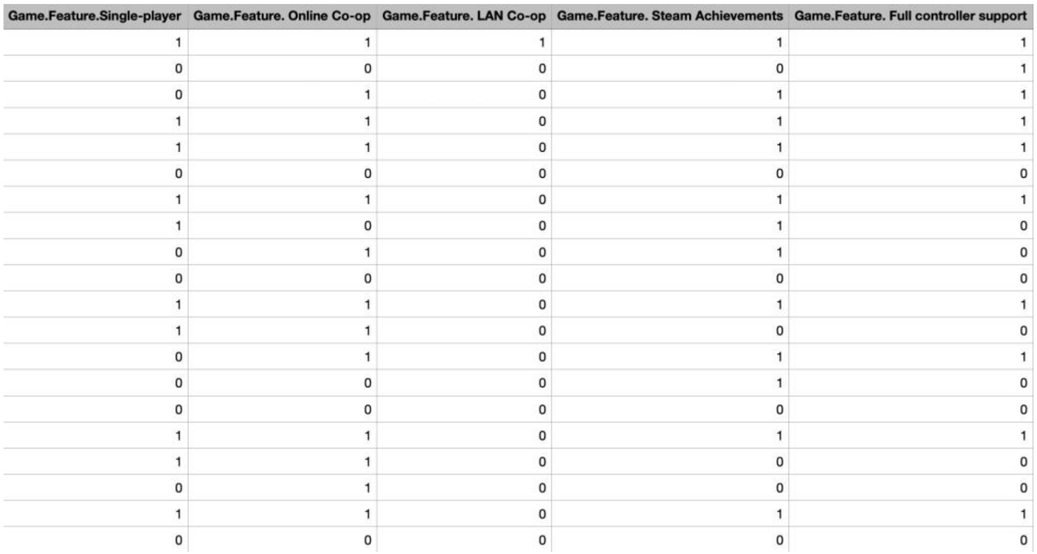
2.1. One-hot Encoder

One-hot encoding is a technique used to convert categorical variables into a binary matrix (1s and 0s). It is used in data cleaning. In the original data, the "Game feature" is a list of features that is not suitable for the mathematical algorithm to interpret the data.



Game Features
['Single-player', 'Online Co-op', 'LAN Co-op', 'Steam Achievements', 'Full controller support', 'Steam Cloud']
['Steam Achievements', 'Full controller support', 'Steam Trading Cards', 'Steam Workshop', 'In-App Purchases', 'Valve Anti-Cheat enabled', 'Stats', 'Remote Play on Phone', 'Remote Play on Tablet', 'Remote Play on TV']
['Online PvP', 'Online Co-op', 'Steam Achievements', 'Full controller support', 'Steam Trading Cards', 'In-App Purchases']
['Single-player', 'Online PvP', 'Online Co-op', 'Cross-Platform Multiplayer', 'Steam Achievements', 'Full controller support', 'Steam Trading Cards', 'Steam Cloud']
['Single-player', 'Online PvP', 'Online Co-op', 'Cross-Platform Multiplayer', 'Steam Achievements', 'Full controller support', 'Captions available', 'In-App Purchases']
['Online PvP', 'Stats', 'Remote Play on Phone', 'Remote Play on Tablet']
['Single-player', 'Online Co-op', 'Steam Achievements', 'Full controller support', 'Steam Cloud']
['Single-player', 'Steam Achievements', 'Steam Trading Cards', 'Partial Controller Support', 'Steam Cloud']
['MMO', 'Online PvP', 'Online Co-op', 'Cross-Platform Multiplayer', 'Steam Achievements', 'Steam Trading Cards', 'Steam Workshop', 'In-App Purchases', 'Valve Anti-Cheat enabled', 'Stats', 'Remote Play on Tablet']
['Online PvP', 'In-App Purchases', 'Partial Controller Support']
['Single-player', 'Online PvP', 'Online Co-op', 'Steam Achievements', 'Full controller support', 'Remote Play on Phone', 'Remote Play on Tablet', 'Remote Play on TV']
['Single-player', 'Online PvP', 'Online Co-op', 'Steam Trading Cards', 'In-App Purchases', 'Partial Controller Support', 'Remote Play on Phone', 'Remote Play on Tablet']
['Online PvP', 'Online Co-op', 'Cross-Platform Multiplayer', 'Steam Achievements', 'Full controller support', 'Steam Trading Cards', 'Steam Cloud', 'Remote Play on Phone', 'Remote Play on Tablet', 'Remote Play on TV']
['Cross-Platform Multiplayer', 'Steam Achievements', 'Steam Trading Cards', 'Captions available', 'Steam Workshop', 'In-App Purchases', 'Partial Controller Support', 'Valve Anti-Cheat enabled', 'Stats', 'Includes level editor',
['Steam Trading Cards', 'Steam Workshop', 'SteamVR Collectibles', 'In-App Purchases', 'Valve Anti-Cheat enabled']
['Single-player', 'MMO', 'Online PvP', 'Online Co-op', 'Cross-Platform Multiplayer', 'Steam Achievements', 'Full controller support', 'VR Supported', 'In-App Purchases', 'Remote Play on Phone', 'Remote Play on Tablet']
['Single-player', 'MMO', 'Online PvP', 'Online Co-op', 'Cross-Platform Multiplayer', 'Partial Controller Support', 'Remote Play on Tablet']
['MMO', 'Online PvP', 'Online Co-op', 'Steam Trading Cards', 'Captions available', 'In-App Purchases', 'Partial Controller Support']
['Single-player', 'Online Co-op', 'Steam Achievements', 'Full controller support', 'Steam Trading Cards', 'Steam Workshop', 'In-App Purchases', 'Remote Play on Phone', 'Remote Play on Tablet']
['MMO', 'Online PvP', 'Partial Controller Support']
['Single-player', 'Online PvP', 'Shared/Split Screen PvP', 'Cross-Platform Multiplayer', 'Steam Achievements', 'Steam Trading Cards', 'In-App Purchases', 'Partial Controller Support', 'Steam Cloud']
['Single-player', 'Online Co-op', 'Full controller support', 'Tracked Controller Support', 'Steam Cloud', 'Remote Play on Phone', 'Remote Play on Tablet']
['Single-player', 'Online PvP', 'Shared/Split Screen PvP', 'Online Co-op', 'Shared/Split Screen Co-op', 'Cross-Platform Multiplayer', 'Steam Achievements', 'Full controller support', 'In-App Purchases']
['Single-player', 'Steam Achievements', 'Full controller support', 'Steam Cloud']
['Single-player', 'Online PvP', 'Online Co-op', 'Steam Achievements', 'Full controller support', 'In-App Purchases', 'Remote Play on Phone', 'Remote Play on Tablet', 'Remote Play on TV']
['Single-player', 'Online PvP', 'Online Co-op', 'Steam Achievements', 'Full controller support', 'Steam Trading Cards', 'Steam Cloud']
['Single-player', 'Cross-Platform Multiplayer', 'Steam Achievements', 'Steam Trading Cards', 'Steam Workshop', 'Steam Cloud', 'Remote Play on Tablet']
['Single-player', 'Online PvP', 'Cross-Platform Multiplayer', 'Steam Achievements', 'Steam Trading Cards', 'In-App Purchases']

After implementing one-hot encoding. The original game feature was melted to 59 columns. Games that have the feature will be marked 1 in that column. After encoding the game features data will be columns with binary categories.



Game.Feature.Single-player	Game.Feature. Online Co-op	Game.Feature. LAN Co-op	Game.Feature. Steam Achievements	Game.Feature. Full controller support
1	1	1	1	1
0	0	0	0	1
0	1	0	1	1
1	1	0	1	1
1	1	0	1	1
0	0	0	0	0
1	1	0	1	1
1	0	0	1	0
0	1	0	1	0
0	0	0	0	0
1	1	0	1	1
1	1	0	0	0
0	1	0	1	1
0	0	0	1	0
0	0	0	0	0
1	1	0	1	1
1	1	0	0	0
0	1	0	0	0
1	1	0	1	1
0	0	0	0	0

2.2. ANOVA Analysis

ANOVA analysis is used to analyze whether different hardware requirements will affect the price. For example, in the "Processor" column which gives the minimum processor requirement to run the game. There are about 1782 levels in this column. And the price is grouped by these levels to conduct an ANOVA analysis.

Source of Variations	Sum of Square*	Degree of Freedom**	Mean Square	F Statistics
Within Columns	$SSW = \sum_{all\ columns} \sum_{all\ i} (X_i - \bar{x}_{col})^2$	$df_w = (R-1) * C$	$MS_w = \frac{SSW}{df_w}$	$F = \frac{MS_b}{MS_w}$
Between Columns	$SSB_c = \sum_{all\ columns} (x_{col} - \bar{x}_o)^2$	$df_b = C - 1$	$MS_b = \frac{SSB_c}{df_b}$	
Total	$\sum_{all\ i} \left(x_i - \bar{x}_{grand}\right)^2$	$df_t = R * C - 1$		

$* \ X_i = \ individual\ data\ value, \bar{x}_{col} = \ mean\ of\ within\ each\ column, X_o = mean\ for\ all\ data$
$** R = \ number\ of\ rows,\ C = number\ of\ columns$

2.3. Forward Selection

After melting the game feature column to 59 new columns. Together with the hardware requirements, there are 63 columns we should consider in total.

To select the most significant variables, the forward selection method was used. The process begins with an empty model and sequentially adds predictors that contribute the most to the model's performance until a stopping criterion is met. 22 variables are retained after the selection. The subsequent regression is based on these variables.

2.4. Multilevel Regression

There are 4 models in this research:

1) Null model
2) No-pooling model
3) Partial-pooling model
4) Complete-pooling model

Both no-pooling and partial-pooling models use “Memory” as the group factor.

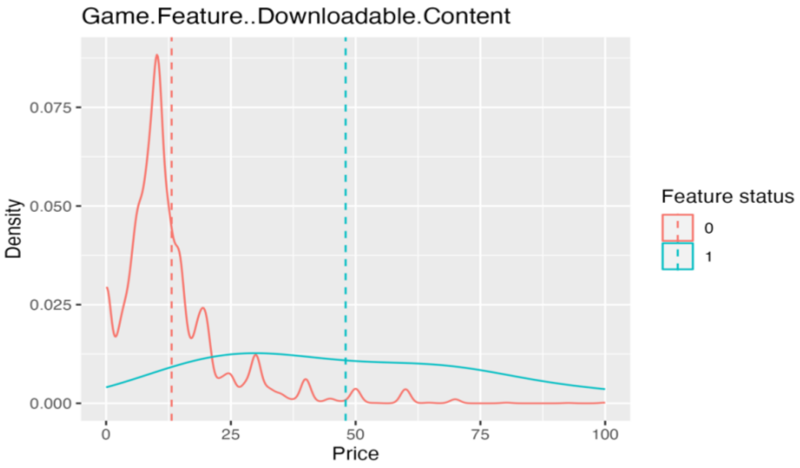
The complete-pooling model uses general linear regression with a log link function and Gamma distribution since the price is skewed and strictly non-negative.

3. Results

3.1. Game Feature Visualization

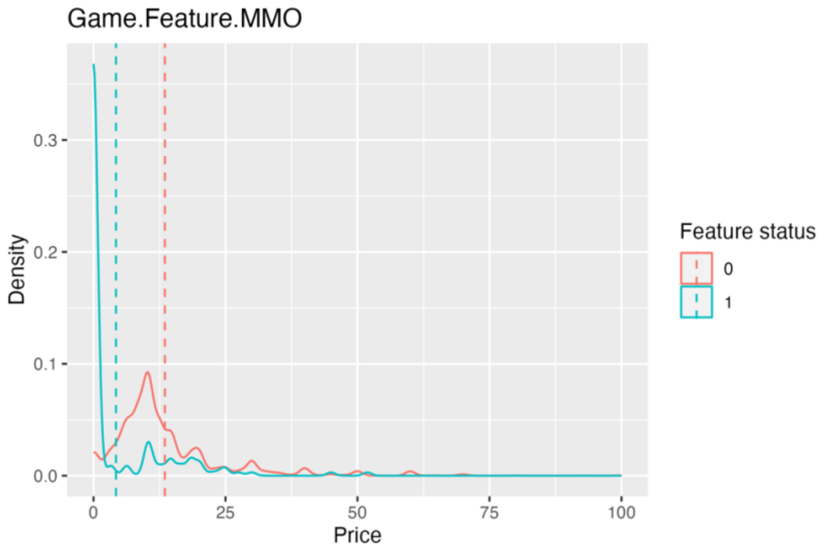
The visualization of game features against price. 59 plots are generated. Five plots that feature have the greatest impact on price. To see the full plots, please refer to the appendix.

Games with downloadable content have a higher price on average.

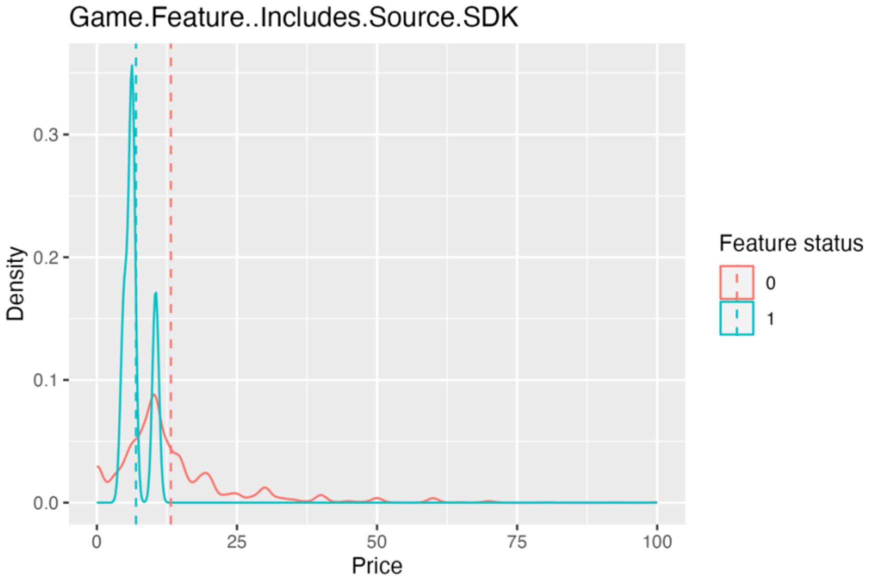


MMO (Massively Multiplayer Online): A type of video game that supports a large number of players interacting with each other within a persistent virtual world online.

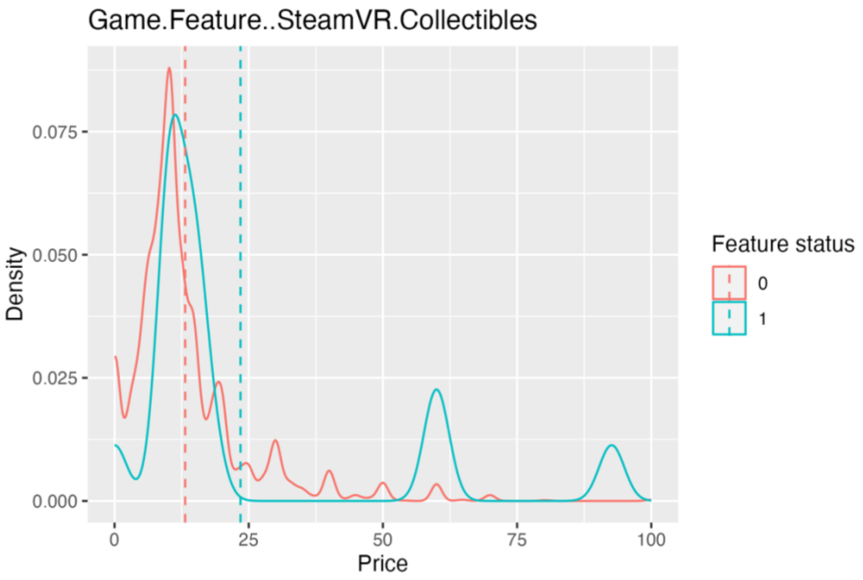
The vertical dash line represents the average price of games that are MMO type and not MMO. Games that are not MMO have higher prices on average.



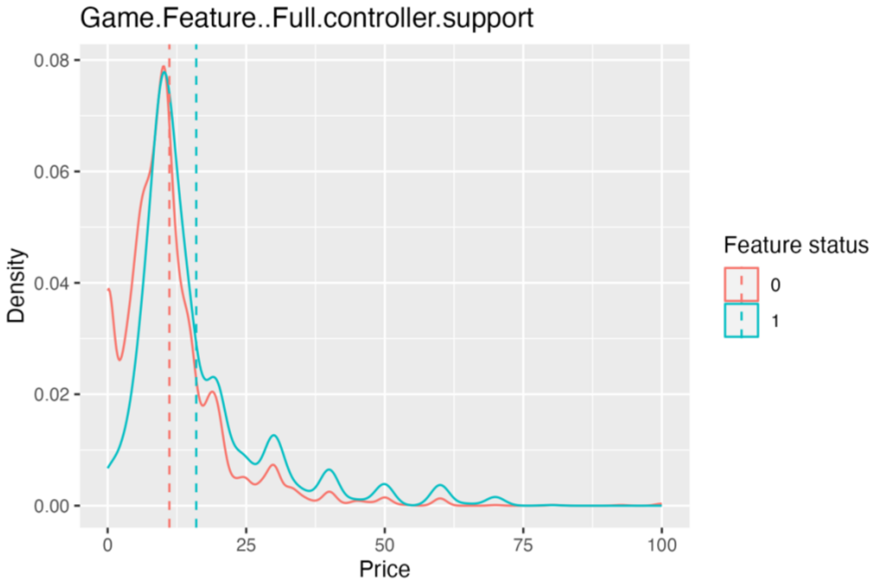
Games that do not include source SDK have a higher average price. (Source SDK, or Source Software Development Kit, is a set of tools and resources provided by Valve Corporation for game developers to create content for games built on the Source engine.)



Games with SteamVR Collectibles have a higher average price.

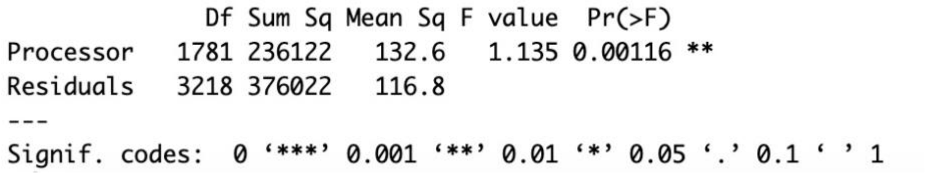


Games that support full controllers have a higher average price.



3.2. ANOVA table analysis

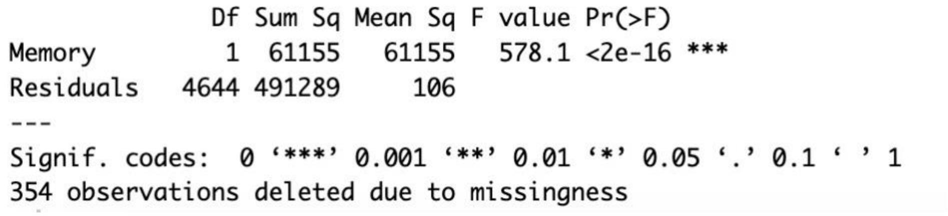
ANOVA Analysis for Processor



```
              Df Sum Sq Mean Sq F value  Pr(>F)   
Processor   1781 236122   132.6   1.135 0.00116 **
Residuals   3218 376022   116.8                   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The p-value is 0.00116, which is less than the commonly used significance level of 0.05. The low p-value suggests that there is a statistically significant difference among the levels of the "Processor" variable.

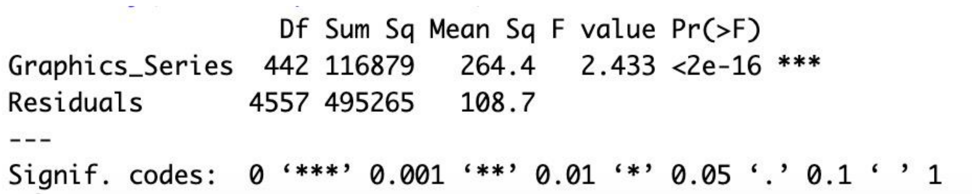
ANOVA Analysis for Memory



```
              Df Sum Sq Mean Sq F value Pr(>F)    
Memory         1  61155   61155   578.1 <2e-16 ***
Residuals   4644 491289     106                   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
354 observations deleted due to missingness
```

The p-value is less than 2e-16, which is essentially zero and much smaller than a typical significance level of 0.05. The very low p-value suggests that there is an extremely significant difference related to the "Memory" variable.

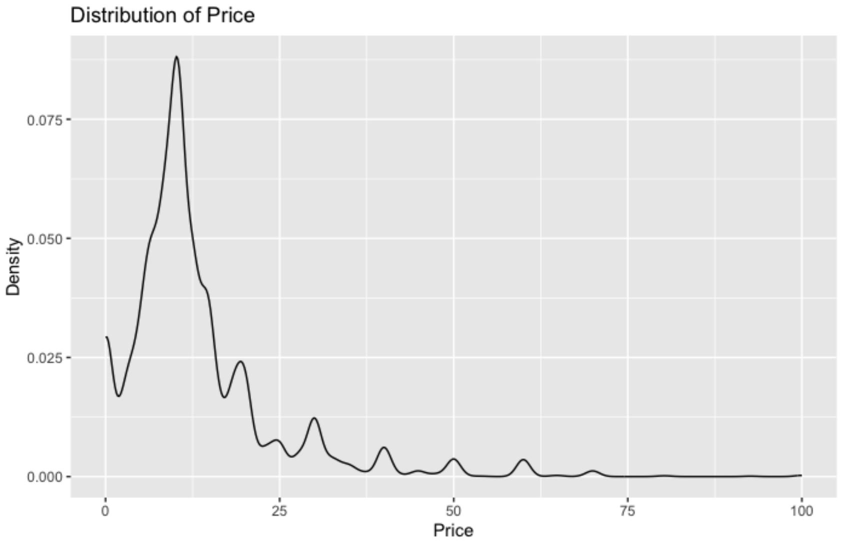
ANOVA Analysis for Graphics Series



```
                  Df Sum Sq Mean Sq F value Pr(>F)    
Graphics_Series  442 116879   264.4   2.433 <2e-16 ***
Residuals       4557 495265   108.7                   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

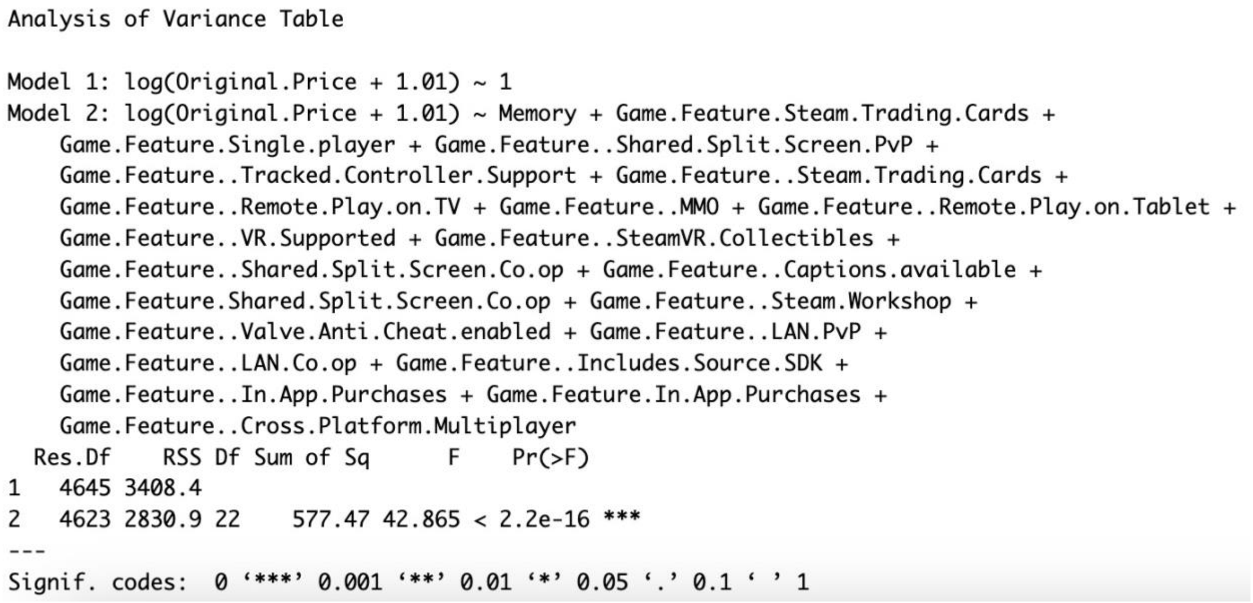
The p-value is less than 2e-16. The very low p-value suggests that there is a highly significant difference related to the "Graphics_Series" variable.

3.3. Comparison Between Null Model and Predictor Model After Forward Selection



Most games cost between $0-$25

The null model is fitted without adding any predictors. And the general linear regression model with log link and Gamma distribution is fitted on predictors selected by forward selection. Price is transformed to avoid a 0 value in price so that the Gamma distribution can be implemented. We use ANOVA analysis to get the significance of the model with predictors.



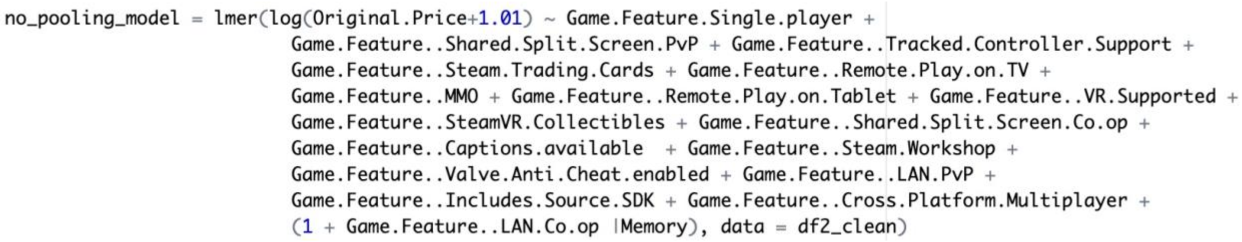
```
Analysis of Variance Table

Model 1: log(Original.Price + 1.01) ~ 1
Model 2: log(Original.Price + 1.01) ~ Memory + Game.Feature.Steam.Trading.Cards +
    Game.Feature.Single.player + Game.Feature..Shared.Split.Screen.PvP +
    Game.Feature..Tracked.Controller.Support + Game.Feature..Steam.Trading.Cards +
    Game.Feature..Remote.Play.on.TV + Game.Feature..MMO + Game.Feature..Remote.Play.on.Tablet +
    Game.Feature..VR.Supported + Game.Feature..SteamVR.Collectibles +
    Game.Feature..Shared.Split.Screen.Co.op + Game.Feature..Captions.available +
    Game.Feature.Shared.Split.Screen.Co.op + Game.Feature..Steam.Workshop +
    Game.Feature..Valve.Anti.Cheat.enabled + Game.Feature..LAN.PvP +
    Game.Feature..LAN.Co.op + Game.Feature..Includes.Source.SDK +
    Game.Feature..In.App.Purchases + Game.Feature.In.App.Purchases +
    Game.Feature..Cross.Platform.Multiplayer
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1   4645 3408.4
2   4623 2830.9 22    577.47 42.865 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The p-value, being practically zero, suggests that the improvement in the model fit by including these predictors is highly significant.

3.4. No-pooling, Partial-pooling, and Complete-pooling models

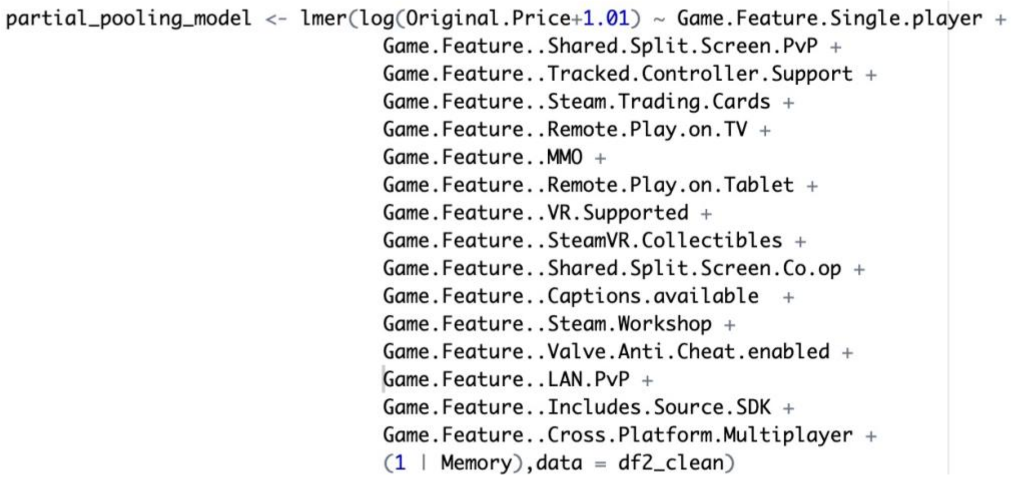
3.4.1. No-pooling model.



```
no_pooling_model = lmer(log(Original.Price+1.01) ~ Game.Feature.Single.player +
                       Game.Feature..Shared.Split.Screen.PvP + Game.Feature..Tracked.Controller.Support +
                       Game.Feature..Steam.Trading.Cards + Game.Feature..Remote.Play.on.TV +
                       Game.Feature..MMO + Game.Feature..Remote.Play.on.Tablet + Game.Feature..VR.Supported +
                       Game.Feature..SteamVR.Collectibles + Game.Feature..Shared.Split.Screen.Co.op +
                       Game.Feature..Captions.available  + Game.Feature..Steam.Workshop +
                       Game.Feature..Valve.Anti.Cheat.enabled + Game.Feature..LAN.PvP +
                       Game.Feature..Includes.Source.SDK + Game.Feature..Cross.Platform.Multiplayer +
                       (1 + Game.Feature..LAN.Co.op |Memory), data = df2_clean)
```

	Estimate	Std. Error	t value
(Intercept)	1.01053	0.08856	11.41
Game.Feature.Single.player	1.19917	0.0452	26.528
Game.Feature..Shared.Split.Screen.PvP	0.21193	0.04406	4.811
Game.Feature..Tracked.Controller.Support	-0.02825	0.06385	-0.442
Game.Feature..Steam.Trading.Cards	0.16054	0.02249	7.138
Game.Feature..Remote.Play.on.TV	0.14946	0.03514	4.253
Game.Feature..MMO	-1.13031	0.10193	-11.089
Game.Feature..Remote.Play.on.Tablet	0.08316	0.0371	2.242
Game.Feature..VR.Supported	0.13059	0.07759	1.683
Game.Feature..SteamVR.Collectibles	0.25399	0.1931	1.315
Game.Feature..Shared.Split.Screen.Co.op	0.09795	0.04334	2.26
Game.Feature..Captions.available	0.01371	0.05982	0.229
Game.Feature..Steam.Workshop	0.07463	0.03492	2.137
Game.Feature..Valve.Anti.Cheat.enabled	0.48407	0.1152	4.202
Game.Feature..LAN.PvP	-0.01634	0.09675	-0.169
Game.Feature..Includes.Source.SDK	-0.41918	0.29843	-1.405
Game.Feature..Cross.Platform.Multiplayer	-0.39584	0.04116	-9.616

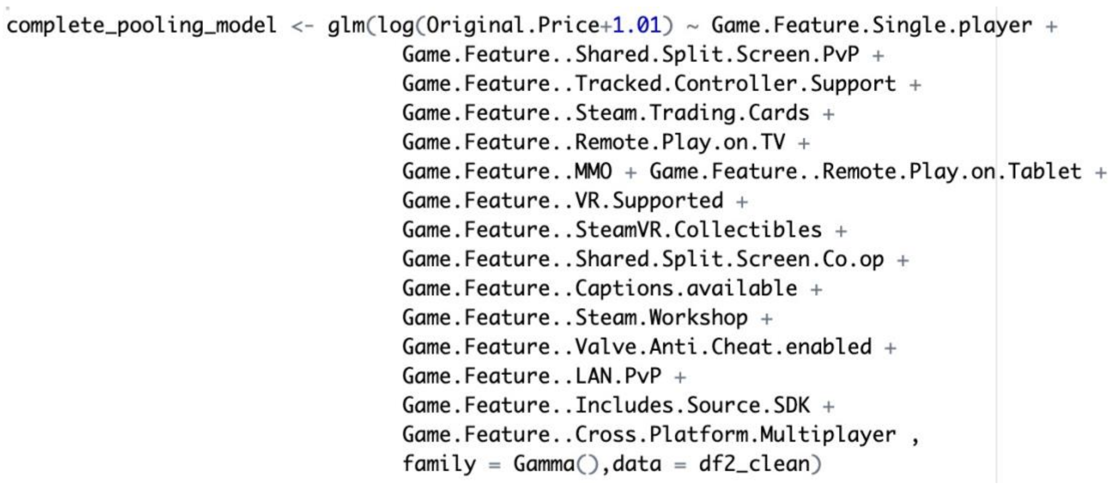
3.4.2. Partial-pooling model.



```
partial_pooling_model <- lmer(log(Original.Price+1.01) ~ Game.Feature.Single.player +
                              Game.Feature..Shared.Split.Screen.PvP +
                              Game.Feature..Tracked.Controller.Support +
                              Game.Feature..Steam.Trading.Cards +
                              Game.Feature..Remote.Play.on.TV +
                              Game.Feature..MMO +
                              Game.Feature..Remote.Play.on.Tablet +
                              Game.Feature..VR.Supported +
                              Game.Feature..SteamVR.Collectibles +
                              Game.Feature..Shared.Split.Screen.Co.op +
                              Game.Feature..Captions.available  +
                              Game.Feature..Steam.Workshop +
                              Game.Feature..Valve.Anti.Cheat.enabled +
                              Game.Feature..LAN.PvP +
                              Game.Feature..Includes.Source.SDK +
                              Game.Feature..Cross.Platform.Multiplayer +
                              (1 | Memory),data = df2_clean)
```

	Estimate	Std. Error	t value
(Intercept)	0.98042	0.09166	10.697
Game.Feature.Single.player	1.19983	0.04523	26.525
Game.Feature..Shared.Split.Screen.PvP	0.20926	0.04406	4.75
Game.Feature..Tracked.Controller.Support	-0.01405	0.06384	-0.22
Game.Feature..Steam.Trading.Cards	0.16183	0.02253	7.183
Game.Feature..Remote.Play.on.TV	0.14413	0.03517	4.099
Game.Feature..MMO	-1.11997	0.10197	-10.983
Game.Feature..Remote.Play.on.Tablet	0.08286	0.03714	2.231
Game.Feature..VR.Supported	0.10973	0.07748	1.416
Game.Feature..SteamVR.Collectibles	0.24747	0.19344	1.279
Game.Feature..Shared.Split.Screen.Co.op	0.10292	0.0433	2.377
Game.Feature..Captions.available	0.01107	0.05986	0.185
Game.Feature..Steam.Workshop	0.07571	0.03492	2.168
Game.Feature..Valve.Anti.Cheat.enabled	0.48681	0.11519	4.226
Game.Feature..LAN.PvP	0.01805	0.08043	0.224
Game.Feature..Includes.Source.SDK	-0.42029	0.29897	-1.406
Game.Feature..Cross.Platform.Multiplayer	-0.38947	0.04105	-9.487

3.4.3. *Complete-pooling model.*

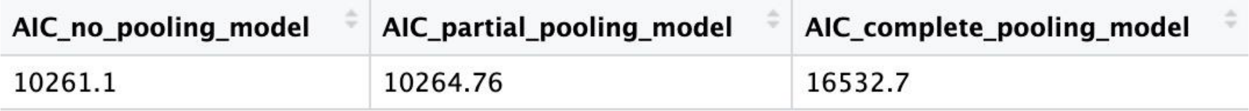


```
complete_pooling_model <- glm(log(Original.Price+1.01) ~ Game.Feature.Single.player +
                                Game.Feature..Shared.Split.Screen.PvP +
                                Game.Feature..Tracked.Controller.Support +
                                Game.Feature..Steam.Trading.Cards +
                                Game.Feature..Remote.Play.on.TV +
                                Game.Feature..MMO + Game.Feature..Remote.Play.on.Tablet +
                                Game.Feature..VR.Supported +
                                Game.Feature..SteamVR.Collectibles +
                                Game.Feature..Shared.Split.Screen.Co.op +
                                Game.Feature..Captions.available +
                                Game.Feature..Steam.Workshop +
                                Game.Feature..Valve.Anti.Cheat.enabled +
                                Game.Feature..LAN.PvP +
                                Game.Feature..Includes.Source.SDK +
                                Game.Feature..Cross.Platform.Multiplayer ,
                                family = Gamma(),data = df2_clean)
```

	Estimate	Std. Error	t value
(Intercept)	0.831848	0.020452	40.674
Game.Feature.Single.player	-0.409181	0.020501	-19.959
Game.Feature..Shared.Split.Screen.PvP	-0.030375	0.009754	-3.114
Game.Feature..Tracked.Controller.Support	-0.03978	0.014397	-2.763
Game.Feature..Steam.Trading.Cards	-0.020675	0.005307	-3.896
Game.Feature..Remote.Play.on.TV	-0.018188	0.007865	-2.312
Game.Feature..MMO	0.428271	0.049487	8.654
Game.Feature..Remote.Play.on.Tablet	-0.012126	0.008438	-1.437
Game.Feature..VR.Supported	-0.018358	0.017169	-1.069
Game.Feature..SteamVR.Collectibles	-0.038639	0.03988	-0.969
Game.Feature..Shared.Split.Screen.Co.op	-0.016219	0.009648	-1.681
Game.Feature..Captions.available	-0.004789	0.013666	-0.35
Game.Feature..Steam.Workshop	-0.013178	0.007991	-1.649
Game.Feature..Valve.Anti.Cheat.enabled	-0.103099	0.024213	-4.258
Game.Feature..LAN.PvP	-0.01067	0.018621	-0.573
Game.Feature..Includes.Source.SDK	0.116185	0.080757	1.439
Game.Feature..Cross.Platform.Multiplayer	0.094802	0.012177	7.785

3.4.4. AIC results

We use AIC to compare the fitness of these models.

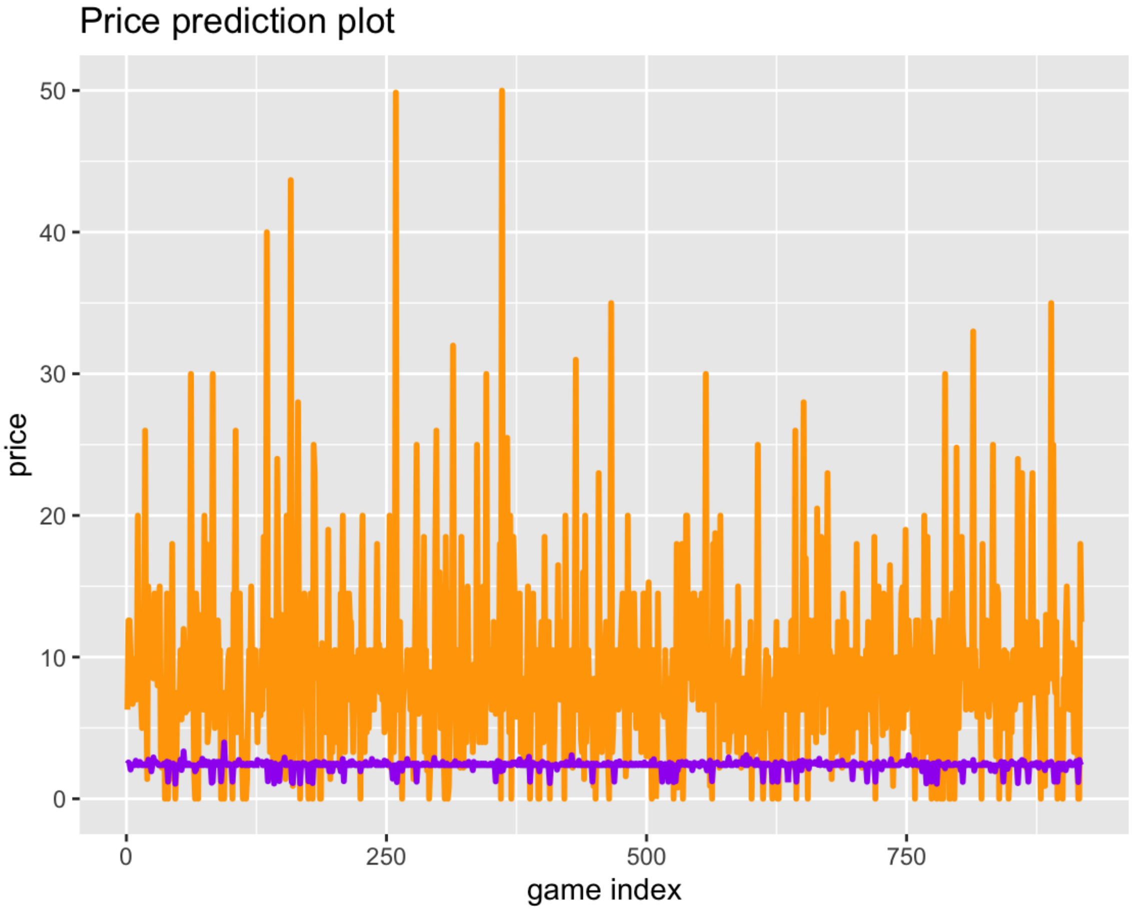


AIC_no_pooling_model	AIC_partial_pooling_model	AIC_complete_pooling_model
10261.1	10264.76	16532.7

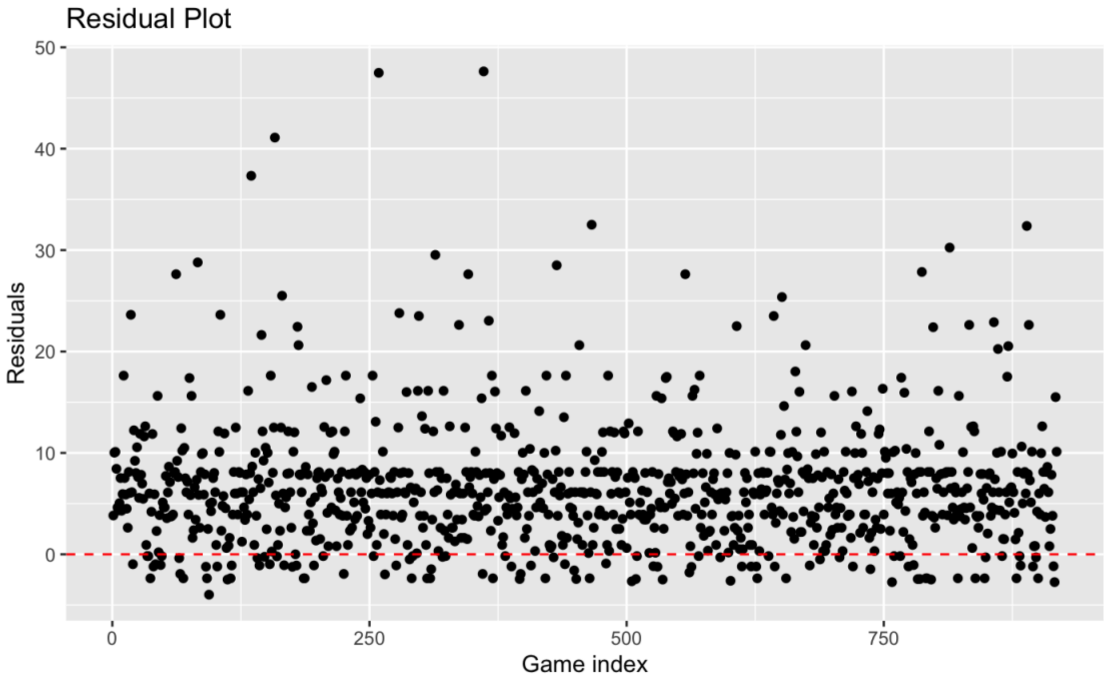
According to the AIC value, the no-pooling model fits the best. Unfortunately, there are 39 different levels in the group in the training set with 5000 games but for the validation set of 1000 games, there are three more levels than the training set. So, the no-pooling and partial-pooling models cannot be used to predict the price for the validation set. And due to the memory constraint, we cannot increase the training set size. Therefore, we use the complete model to do the prediction.

3.5. Prediction Results Residual Plot

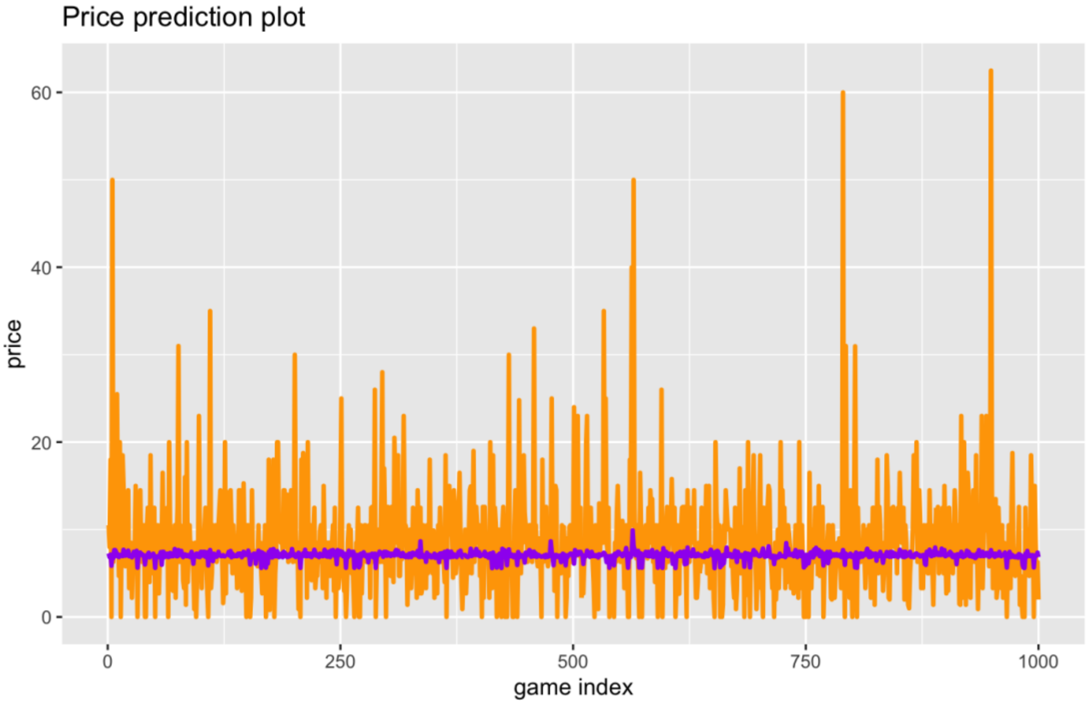
Prediction results for a validation set of 1000 games.



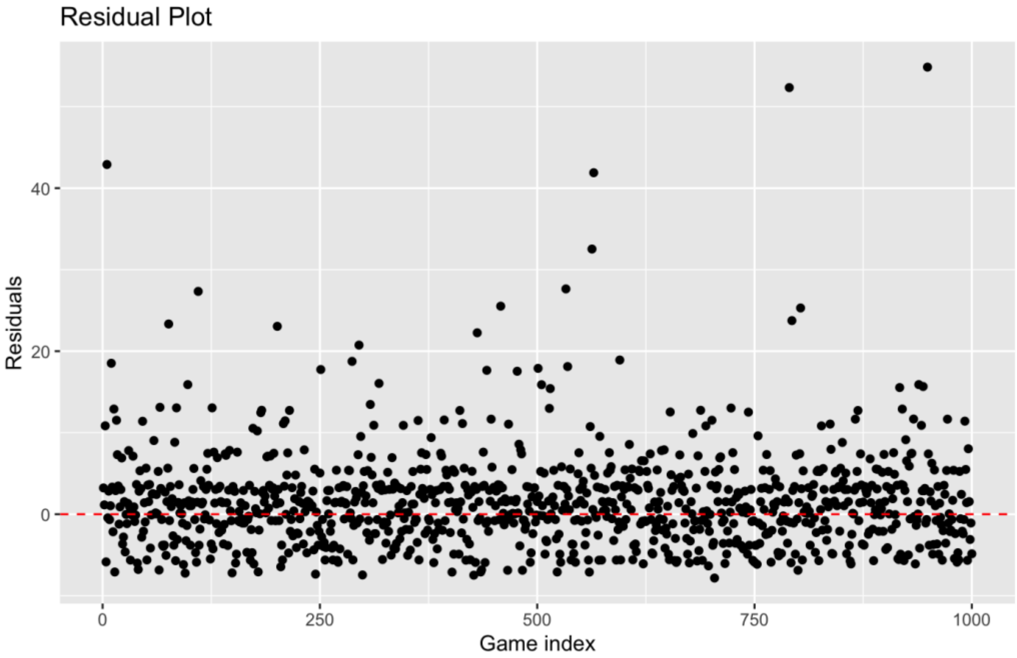
And the residual plot:



Both plots show that the model underestimates the price. However, it was found that if I added the null model's intercept to the complete-pooling model, the result will be much better:



The residual plot:



The model gives a relatively conservative estimate of the price. It cannot deal with the variability in the real data.

The residual plot shows there are still some outliers in the data. But most of the residuals are symmetric along 0.

4. Discussion

4.1. Null model intercept

The first thing to do in the further work is to find out why the null model's intercept can improve the complete-pooling model's performance and the problems with the model. Is it too simple? Or maybe non-linear model should be tried in the future.

4.2. Outliers

From the residual plot, we can find there are some points that with high residuals. The model should be adjusted to deal with these points or consider dropping these points as outliers.

4.3. PCA Method

The mean square error of GLM is still large (approximately 40). PCA might be a choice to fit the model. It might generate a more accurate result.

4.4. Inflation in Video Market

The model should be robust enough to consider the inflation in the today's world

5. Appendix

Plots Appendix

Code Appendix

6. References

Consalvo, Mia, and Christopher A. Paul. 2015. Paying to play: the evolving structure of game pricing and industry legitimacy. https://spir.aoir.org/ojs/index.php/spir/article/view/8780/6991.

Grand View Research. 2021. Video Game Market Size, Share & Trends Analysis Report By Device (Console, Mobile, Computer), By Type (Online, Offline), By Region (Asia Pacific, North America, Europe), And Segment Forecasts, 2023 – 2030. https://www.grandviewresearch.com/industry-analysis/video-game-market#.