MA678 Final Project

Chenxuan Xiong

2023-12-10

## Abstract

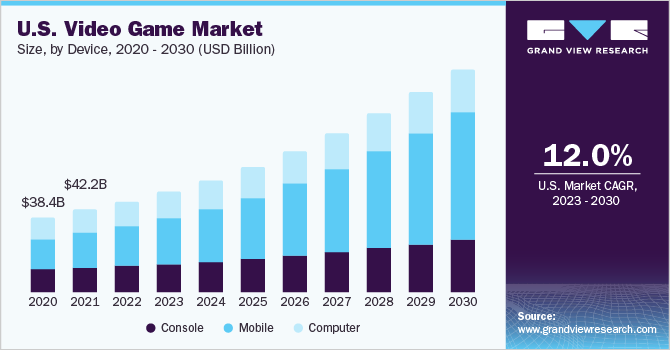
This study uses linear regression to build a video game pricing model. This model uses Game Features and device requirements as predictors to regress the Original price. Through this model, we can come up with a reasonable approximate game pricing for game companies to use as a reference.

## 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.

### 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).



### 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 extend to the whole data set to get a more accuracy result.

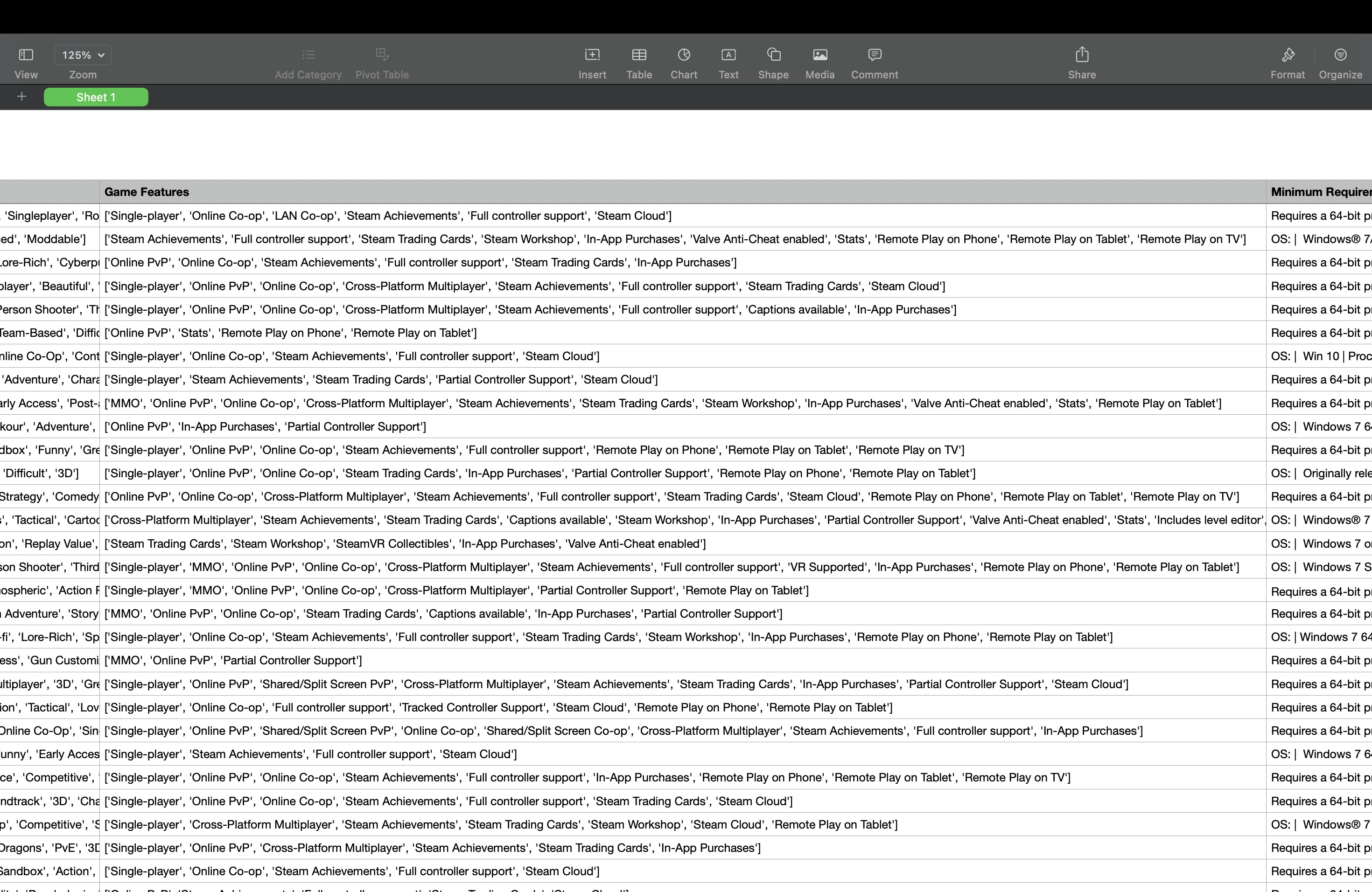
### Research Objectives

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?

## Method

### 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.



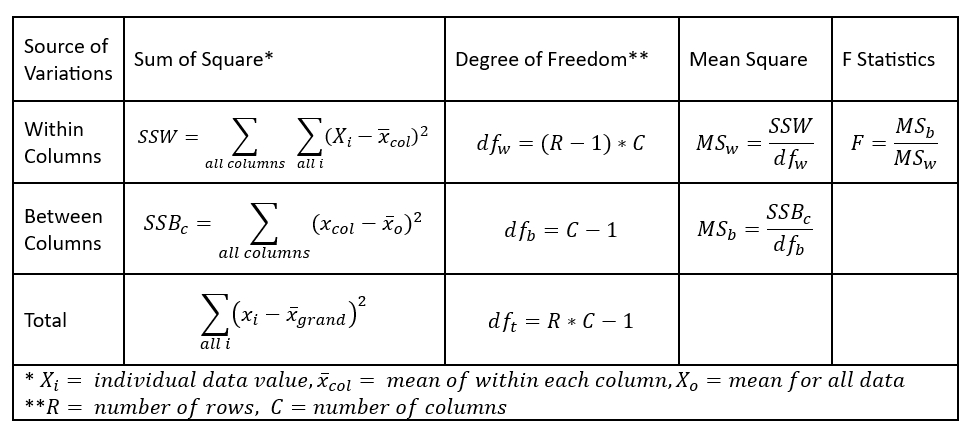
After implementing one-hot encoding. The original game feature was melted to 59 columns. Games that have the feature will be marked 1 in that column. After encoding the game features data will be columns with binary categories.

A screenshot of a computer

Description automatically generated

### 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.



### 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.

### 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.

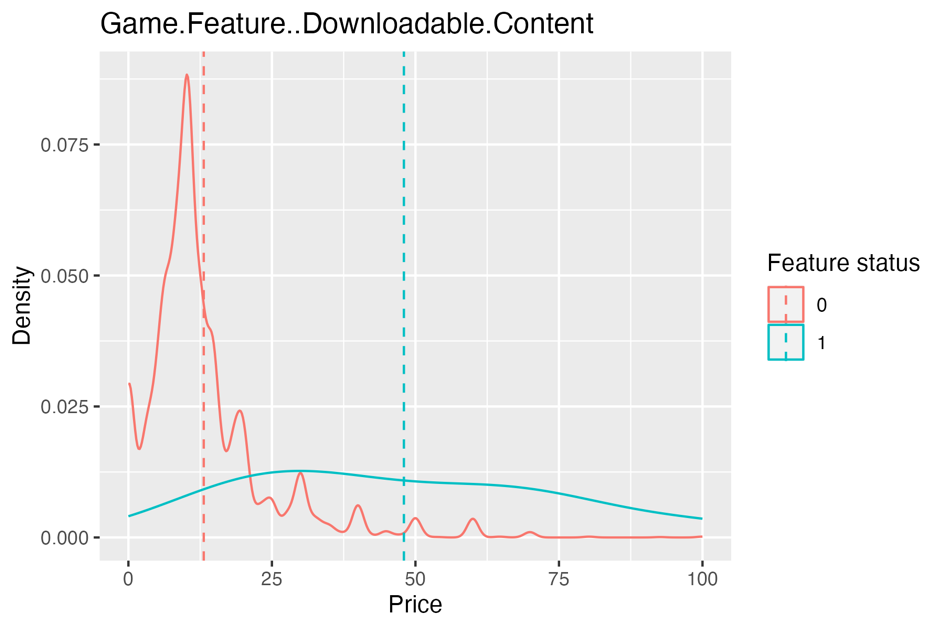
The complete-pooling model uses general linear regression with a log link function and Gamma distribution since the price is skewed and strictly non-negative.

## Results

### 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.

****

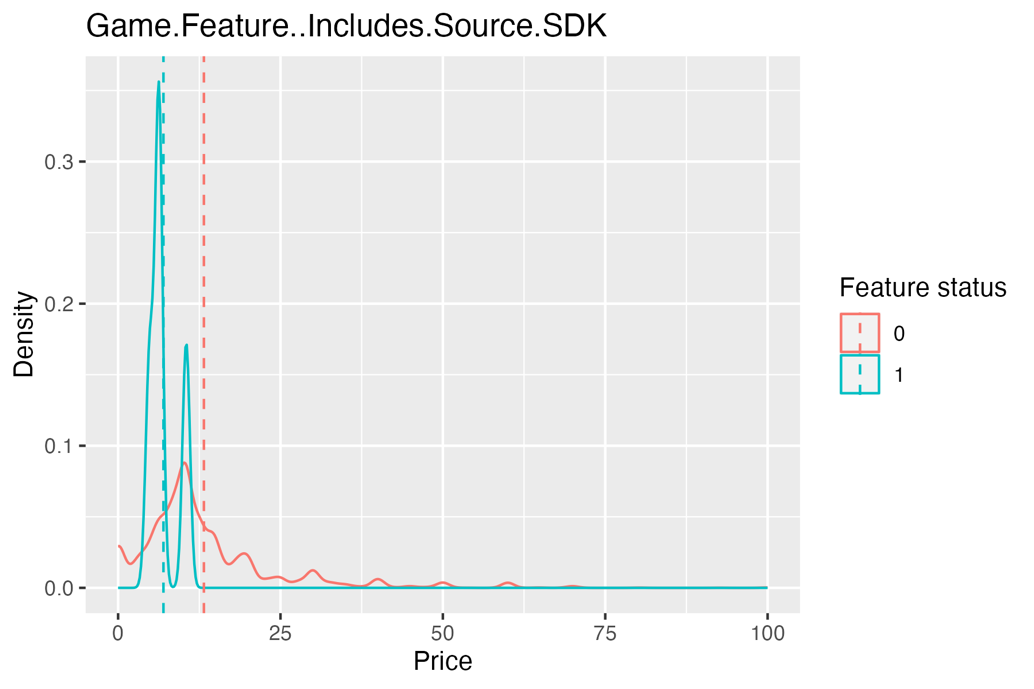
MMO (Massively Multiplayer Online): A type of video game that supports a large number of players interacting with each other within a persistent virtual world online.

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.

**A graph with red and blue lines

Description automatically generated**

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.)

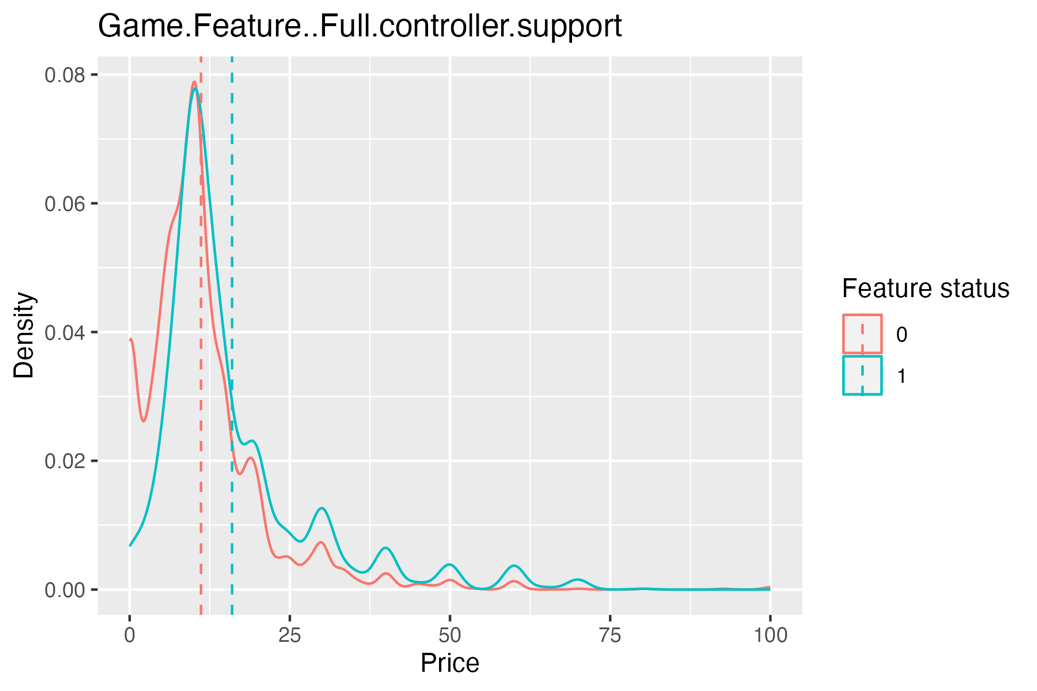
****

Games with SteamVR Collectibles have a higher average price.

**A graph of a graph

Description automatically generated with medium confidence**

Games that support full controllers have a higher average price.

****

### ANOVA table analysis

ANOVA Analysis for Processor

A close-up of numbers

Description automatically generated

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

A close-up of numbers

Description automatically generated

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

A close up of numbers

Description automatically generated

The p-value is less than 2e-16 (essentially zero), which is much smaller than a typical significance level of 0.05. The very low p-value suggests that there is a highly significant difference related to the "Graphics\_Series" variable.

### Comparison Between Null Model and Predictor Model After Forward Selection

A graph showing a price

Description automatically generated

Most games cost between $0-$25

A null model is fitted without adding any predictors. A 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.

**A screenshot of a computer program

Description automatically generated**

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

### No-pooling, Partial-pooling, and Complete-pooling models

#### No-pooling model.

A screenshot of a computer

Description automatically generated



#### Partial-pooling model.

A screenshot of a computer

Description automatically generated



#### Complete-pooling model.

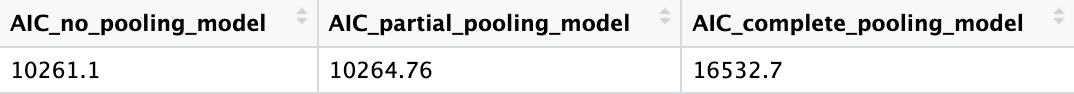
A screenshot of a computer program

Description automatically generated



#### AIC results

We use AIC to compare the fitness of these models.



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

### Prediction Results Residual Plot

Prediction results for a validation set of 1000 games.

A graph with orange and purple lines

Description automatically generated

The residual plot:

A graph with black dots

Description automatically generated

The model gives a relatively conservative estimate of the price.

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. 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.

4.2. Try to use PCA to analyze since the mean square error of GLM is still large (approximately 40)

## 

## 5. Appendix

[Plots Appendix](https://github.com/SwaineXiong/MA678-midterm-proj/blob/main/Appendix.pdf)

## 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#](https://www.grandviewresearch.com/industry-analysis/video-game-market).