Determinants of Video Game Ownership

By Minsoo Lee

May 7, 2025

Abstract

This paper investigates the key factors influencing video game ownership using data from the popular digital distribution platform, Steam. A multivariate linear regression model is constructed based on datasets curated by expert game designers, encompassing data from over 30,000 games. Via this model, the analysis identifies significant predictors of game ownership, offering insights that can inform the strategies of developers and corporations aiming to better understand consumer behavior and improve product offerings.

1 Introduction

In today's digital world, video games have emerged as one of the leading forms of modern entertainment. Indeed, over the past twenty-five years, video games have evolved from large, expensive arcade machines with high barriers to entry into digital products that can be downloaded with the click of a button. With computers being more widely available and the spread of the internet, video games have become incredibly easy to access, with hundreds of millions of players gathering online to play their favorite games.

As such, corporations around the world have increasingly invested in understanding consumer behavior and developing video games that cater to evolving player preferences. According to Forbes, the global video game industry generated an estimated $184.4 billion in 2022-significantly surpassing the music industry's $26.2 billion and the movie industry's $26 billion. These figures highlight not only the financial dominance of video games in the entertainment sector but also their growing cultural relevance and influence. As the industry continues to expand, understanding the determinants of video game ownership becomes critical for developers, publishers, and large scale corporations alike.

This paper aims to support efforts to further develop the video game industry by identifying the key factors that influence estimated video game ownership. Through the use of a multivariate regression model, the analysis provides a data-driven understanding of consumer behavior within the gaming market. By applying the insights derived from this model, both corporations and independent developers can better understand their audiences' general preferences and tailor their products accordingly. This may thus help optimize marketing strategies, improve game design, and ultimately increase the likelihood of consumer engagement and purchase. The following sections will first review the relevant literature describing research conducted on consumer behavior in the video game industry, outline the methodology and data used for the analysis, and discuss the empirical results that present the significant predictors of game ownership.

2 Literature Review

Studies have been conducted to determine video game ownership rates, many of which echo a similar type of multivariate regression model that will be used in this paper. A study by Feray Adiguzel from LUISS University utilized data from 140 video games to determine whether reviews from online video sharing platform Youtube affected video game sales. The study used an ordinary least squares regression model with features including valence and volume from critics and consumers, Youtube's like to dislike ratio, interactivity, multiplatform availability, and genre against the dependent variable of log(Sales). The study discovered that although reviews on Youtube were no more significant than those by critics, consumer reviews were highly significant on determining video game sales, and had a positive impact on sales. I thus plan to use user reviews as a feature in my regression model as it has already proven to be significant through Adiguzel's model.

Another study by Joe Cox and Daniel Kaimann from the Portsmouth Business School sought to determine the interaction between critic reviews and other determinants of product quality in the video game industry. Like Adiguzel's study, the study uses an ordinary least squares regression to discover the effect of professional critic reviews on video games. The study found that critic reviews not only predict video game sales, but also influence them significantly. This suggests that from a consumer's point of perspective, reviews from professional

critics serve as a credible source of information and motivator to incentivize one to purchase and play a video game.

Finally, a study by Hoon S. Choi, Myung S. Ko, Dawn Medlin, and Charlie Chen models video game sales data by analyzing intrinsic and extrinsic cues. Intrinsic cues include company reputation, newness, and retro features, while extrinsic cues include review valence, product popularity, price, and user engagement. A multivariate linear regression is fitted utilizing these features, and video game sales are modelled accordingly. The study discovered that with the exception of company reputation, all factors were significant in determining video game sales. This implies that most consumers do not generally have a preference for a company's reputation when purchasing a game, and will instead select games based on qualities belonging to the game itself. The study notes that widely known video game corporations such as Microsoft and Electronic Arts are seen generally as publishers rather than developers, and thus consumers may not view them as having a large influence in creating the game. This study may shed light on the effects of indie versus corporate game development observed in the regression model in this paper, providing useful insights into this consumer behavior phenomenon.

3 Data

This paper will utilize datasets utilizing data collected from video game distributor Steam. Steam, created by Valve Corporation, is one of the most popular video game digital distribution services available online, with thousands of games available for download and purchase. Steam's globally renowned reputation in the video game industry, combined with its large amount of video game titles, will allow it to serve as an efficient platform to model this paper's regression analysis and generalize it to video game industry as a whole.

The first dataset is created by Alexander Barabanov (Lead Game Designer at Sad Cat Studios) and Lev Kobelev (Game Designer at Whalekit). The dataset obtains information directly from Steam's online store, and this paper will use the dataset's calculated review score, which is calculated as the percentage of positive reviews out of total reviews. Additionally, the dataset contains information regarding games' "tags", which are essentially labels for a video game's genre, development process, and multiplayer availability. This paper will convert these "tags" into dummy variables to determine whether these qualitative

variables will have a significant effect on video game sales. As there are too many tags on Steam's platform to create a universal analysis, only tags widely present across the platform have been selected, namely-Action, Shooter, Open World, RPG, Multiplayer, Strategy, Sandbox, Horror, Indie, Survival, Story Rich, Simulation, PvP, Difficult, Adventure, Puzzle, Moddable, Comedy, Casual, and 2D.

The second dataset is compiled by Anton Kozyriev (PhD, Igor Sikorsky Kyiv Polytechnic Institute) and will be primarily utilized for its detailed information on platform availability, specifically for Apple MacBooks. Given that a significant portion of modern video game consumers use MacBooks as their primary gaming device, incorporating a qualitative variable that captures a game's availability on macOS is essential, particularly as most recent larger AAA-games are available solely on Windows due to the Macbook's lack of dedicated graphics support and limited compatibility with many game engines and APIs such as DirectX. This variable will help determine whether platform compatibility has a measurable impact on video game ownership. Including this factor allows for a more nuanced understanding of how hardware accessibility influences consumer purchasing behavior in the gaming industry.

Finally, the third dataset is compiled by Martin Bustos and published on Kaggle under license by the Massachusetts Institute of Technology (MIT). This dataset provides the model's dependent variable-estimated ownership. Since Steam does not release official data on game ownership, the dataset utilizes estimates generated by SteamSpy, a reputable video game sales tracking platform created by Sergey Galyonkin. SteamSpy collects data by polling user profiles through the Steam API, and has become widely recognized for producing accurate and reliable estimates of game ownership. Additional features included from this dataset-such as years since release, adult-only playability, price (USD), number of achievements, number of user recommendations, and average playtime-will serve as independent variables in the regression model. These variables offer a robust framework for identifying the key drivers behind estimated game ownership and help ensure the model captures both economic and behavioral dimensions of consumer decision-making.

The three datasets have been merged together to fit this paper's research using an inner join on the game's universal app identification number. The data has been cleaned to remove video games that do not have sufficient information to be used in the paper's regression model, and has been examined closely to remove any possibilities of errors in data analysis. In total, 34,017 titles

were obtained from the datasets after cleaning and merging. Data cleaning was completed using Excel, and the datasets have been merged using Python and Pandas tools. The code used to complete the merge is included in the appendix of this paper.

To summarize, the model will use estimated ownership as a dependent variable, and the independent variables will include years since release, 18+, Price (USD), Mac, Achievements, Recommendations, Average playtime forever, Reviews Score, Action, Shooter, Open World, RPG, Multiplayer, Strategy, Sandbox, Horror, Indie, Survival, Story Rich, Simulation, PvP, Difficult, Adventure, Puzzle, Moddable, Comedy, Casual, and 2D. A full statistical model including interaction terms will be described in the next section.

4 Results

The model below was estimated using a Weighted Least Squares (WLS) regression framework to address heteroscedasticity present in the initial model (provided in appendix). Weights were calculated using the Price (USD) variable. The model incorporates key interaction terms to account for joint effects.

$$\begin{aligned}\text{estimated\_ownership} = {} & \beta_0 + \beta_1 \cdot \text{years\_since\_release} + \beta_2 \cdot \text{age\_18plus} + \beta_3 \cdot \text{price\_usd} + \\ & \beta_4 \cdot \text{mac} + \beta_5 \cdot \text{achievements} + \beta_6 \cdot \text{recommendations} + \\ & \beta_7 \cdot \text{avg\_playtime} + \beta_8 \cdot \text{review\_score} + \beta_9 \cdot \text{action} + \beta_{10} \cdot \text{shooter} + \\ & \beta_{11} \cdot \text{open\_world} + \beta_{12} \cdot \text{rpg} + \beta_{13} \cdot \text{multiplayer} + \beta_{14} \cdot \text{strategy} + \\ & \beta_{15} \cdot \text{sandbox} + \beta_{16} \cdot \text{horror} + \beta_{17} \cdot \text{indie} + \beta_{18} \cdot \text{survival} + \\ & \beta_{19} \cdot \text{story\_rich} + \beta_{20} \cdot \text{simulation} + \beta_{21} \cdot \text{pvp} + \beta_{22} \cdot \text{difficult} + \\ & \beta_{23} \cdot \text{adventure} + \beta_{24} \cdot \text{puzzle} + \beta_{25} \cdot \text{moddable} + \beta_{26} \cdot \text{comedy} + \\ & \beta_{27} \cdot \text{casual} + \beta_{28} \cdot \text{2D} + \beta_{29} \cdot \text{action\_multiplayer} + \\ & \beta_{30} \cdot \text{price\_rpg} + \beta_{31} \cdot \text{multiplayer\_pvp} + \\ & \beta_{32} \cdot \text{playtime\_storyrich} + \beta_{33} \cdot \text{sandbox\_openworld} + \\ & \beta_{34} \cdot \text{action\_shooter} + \beta_{35} \cdot \text{rpg\_storyrich} + \beta_{36} \cdot \text{comedy\_casual} + \epsilon\end{aligned}$$

A regression summary using the above model is presented below:

Table 1: WLS Regression Results: Estimated Owners

Variable	**Coefficient**	**Std. Error**	**t-Stat**	**P-value**	**[0.025, 0.975]**
const	55150.00	28700	1.922	0.055	[-600.35, 110000]
Years since release	43.12	59.02	0.731	0.465	[-72.57, 159.81]
18+	133300	22200	5.995	0.000	[90000, 177000]
Price (USD)	-1766.05	403.09	-4.381	0.000	[-2556.10, -976.00]
Mac	20400	12000	1.693	0.090	[-3000.00, 43800]
Achievements	39.13	27.00	1.450	0.147	[-13.77, 92.04]
Recommendations	48.70	0.24	201.21	0.000	[48.23, 49.17]
Average playtime forever	30.79	3.85	7.994	0.000	[23.25, 38.34]
Reviews Score Fancy	612.53	293.37	2.088	0.037	[38.52, 1186.54]
Action	7325.63	12400	0.589	0.556	[-16800, 31400]
Shooter	80330	65500	1.227	0.220	[-48000, 209000]
Open World	12150	18200	0.667	0.505	[-23400, 47700]
RPG	31500	19600	1.608	0.108	[-6700, 69700]
Multiplayer	1196.80	21900	0.055	0.956	[-41400, 43800]
Strategy	37060	12300	3.007	0.003	[13000, 61100]
Sandbox	47650	24400	1.951	0.051	[-180.00, 95500]
Horror	-37610	16900	-2.224	0.026	[-70800, -4460]
Indie	-63330	10800	-5.849	0.000	[-84500, -42100]
Survival	-49290	18500	-2.668	0.008	[-85400, -13200]
Story Rich	53230	15800	3.360	0.001	[22300, 84200]
Simulation	-16790	12500	-1.345	0.179	[-41300, 7700]
PvP	-71110	43700	-1.628	0.104	[-157000, 14700]
Difficult	50190	16400	3.058	0.002	[18000, 82400]
Adventure	43320	11200	3.861	0.000	[21300, 65300]
Puzzle	46690	14600	3.202	0.001	[18000, 75400]
Moddable	176500	30200	5.843	0.000	[117000, 236000]
Comedy	36050	23000	1.565	0.118	[-9000, 81100]
Casual	-4828.70	12100	-0.399	0.690	[-28500, 18800]
2D	-11070	12300	-0.904	0.366	[-35100, 13000]
Action_Multiplayer	195900	25500	7.683	0.000	[146000, 246000]
Price_RPG	3585.21	659.31	5.438	0.000	[2300.00, 4870.00]
Multiplayer_PvP	202000	49900	4.048	0.000	[104000, 300000]
Playtime_StoryRich	-53.93	9.10	-5.923	0.000	[-71.77, -36.09]
Sandbox_OpenWorld	23880	36100	0.661	0.509	[-46800, 94600]
Action_Shooter	-29720	67600	-0.440	0.660	[-163000, 103000]
RPG_StoryRich	-121000	25400	-4.767	0.000	[-170000, -71700]
Comedy_Casual	-15560	36100	-0.431	0.666	[-86400, 55300]
Model Summary					
R-squared	0.613				
Adj. R-squared	0.612				
F-statistic	1459.0				
Prob (F-statistic)	0.000				
Observations	34,017				
AIC	inf				
BIC	inf				

Standard errors are not robust to heteroscedasticity. All variables are weighted using inverse variance weights.

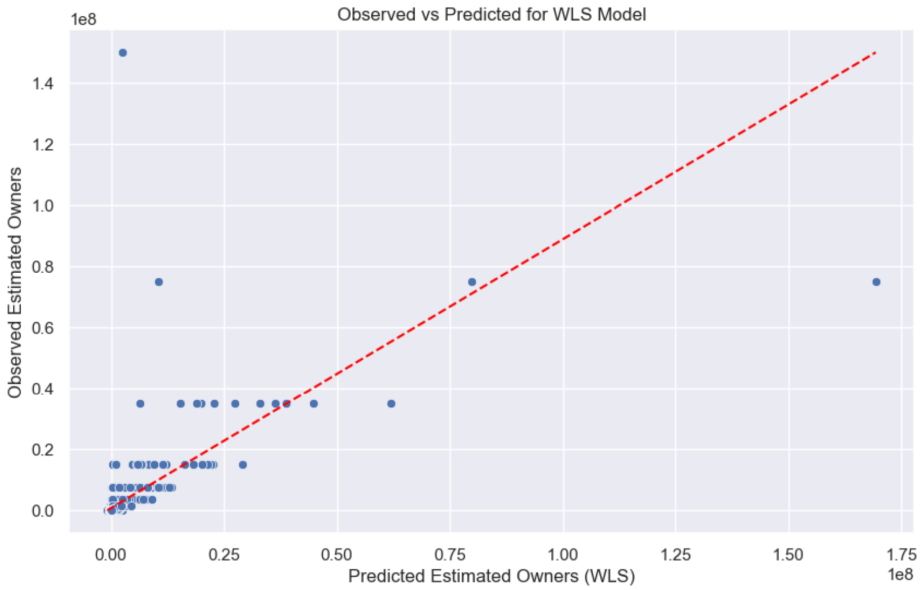

Figure 1: Observed vs Predicted with y = x

5 Discussion

Model Performance The model indicates a moderate R-squared value at 0.613, indicating that 61.3% of the variation in the data was captured by the model. Converting the model into a weighted least squares model and including interaction terms improved the model significantly, as the initial model without either had a relatively low R-squared value at 0.460 (results in appendix). Figure 1 presents a plot of observed and predicted estimated owners, and it is clear that the model moderately fits estimated ownership at most levels, excluding a select few games that have an extremely high of estimated owners.

Quantitative variables Assuming the test was conducted under $\alpha = 0.05$, significant positively correlated quantitative variables included 18+, recommendations, average playtime, and reviews score. These variables, particularly review scores and recommendations-were expected to be positively correlated, as noted in the literature review. Previous studies have shown that both user and critic reviews play a significant role in incentivizing game purchases. Price, as expected, is negatively correlated with video game sales, as consumers tend to budget their purchases. The number of achievements and years since release

were insignificant variables, and we can thus infer that these are variables that consumers do not consider when purchasing video games.

Qualitative variables Of the qualitative variables, Mac availability was insignificant at the 0.05 level, yet the p-value of 0.09 is still considerably low. Game developers should thus still consider Mac availability as an option to boost sales, as it still had a positive correlation with estimated ownership. Genre variables varied highly, but most notably, indie games were significant and negatively correlated with estimated ownership. This indicates that, on average, Indie titles tend to have lower ownership levels compared to mainstream games, possibly due to smaller marketing budgets, limited reach, or niche appeal. Additionally, multiplayer was insignificant as a single variable, but as an interaction with action and PvP, the variable was significant and positively correlated with ownership with a very large coefficient at 195,900 and 202,000 respectively. Many popular games enjoyed today involve battling other players in large online lobbies, which could explain this phenomenon.

6 Conclusion

We have thus created a system for video game developers and corporations to create video games that would suit modern consumers' tastes. To increase consumer sales, developers should take action to collaborate with video game critics to share their game with a larger audience, and be open to receiving feedback via user reviews. Creators may also be encouraged to collaborate with larger corporations and mainstream developers to enhance marketing and expand their reach online. However, this may not be a permanent solution, as many people specifically seek games developed by independent creators for the unique indie feeling. For some, this independent feel is a major factor in their purchasing decisions, meaning that the appeal of indie games cannot be replaced entirely by mainstream collaborations. Future research should be conducted to discover factors which may incentivize developers to work independently, and influence consumer behavior to give indie games a chance. A key limitation to this study would be the calculation of estimated ownership, as it has been obtained from a third-party source. However, without Valve Corporation officially publishing statistics regarding their player base data, this limitation would be extremely difficult to address. Ultimately, understanding the factors that influence game sales can help developers and companies navigate the ever-evolving

gaming landscape and connect with a broader audience.

7 Appendix & References

File containing code and datasets can be downloaded in the GitHub link below:
`https://github.com/Elee1602/Steam_Game_Analysis`

References

[1] Arora, Krishan. "The Gaming Industry: A Behemoth with Unprecedented Global Reach." Forbes, Forbes Magazine, 13 Aug. 2024 `www.forbes.com/councils/forbesagencycouncil/2023/11/17/the-gaming-industry-a-behemoth-with-unprecedented-global-reach/`

[2] Feray Adiguzel et al. "The Effect of YouTube Reviews on Video Game Sales." Journal of Business Research-Turk, 13 (2021): 2096-2109 `https://doi.org/10.20491/ISARDER.2021.1249`

[3] H. S. Choi et al. "The effect of intrinsic and extrinsic quality cues of digital video games on sales: An empirical investigation." Decis. Support Syst., 106 (2018): 86-96

`https://doi.org/10.1016/j.dss.2017.12.005`

[4] J. Cox et al. "How do reviews from professional critics interact with other signals of product quality? Evidence from the video game industry†." Journal of Consumer Behaviour, 14 (2015): 366-377 `https://doi.org/10.1002/CB.1553`