Exploring Factors Influencing Average Playtime of Video Games: A Regression Analysis*

Ziqi Zhu Raghav Sinha Christophe McWatt

December 4, 2024

Table of contents

*Code and data are available at: https://github.com/zzq20010617/sta302Paper.

1 Contributions

We found the dataset together, and seperate the works as follow: Ziqi: Introduction and data description Raghav: Ethics discussion and some plot in preliminary result Christophe: Most of Preliminary results section

2 Introduction

The gaming industry has experienced exponential growth in recent years, with millions of players engaging in various games on platforms like Steam. Understanding what influences average playtime in games is crucial for game developers, publishers, and marketers. Average playtime is a key indicator of player engagement and game success, and investigating factors that impact it can provide valuable insights into game design, pricing strategies, and marketing efforts. So our research aims to explore the factors that influence the average playtime of games available on Steam using a dataset we get from Kaggle. Specifically, we will investigate how variables such as price, user ratings, and other game characteristics contribute to player engagement, as measured by playtime. From some previous papers, we learned that both positive and negative reviews have a significant correlation with playtime, but the effect is more pronounced for positive reviews (Brodschneider and Pirker 2023). Another paper found out that game pricing significantly affects player engagement, with lower-priced games generally seeing higher playtime (Luisa et al. 2021). The third paper did an empirical study that also examined the influence of game reviews on playtime, focusing on how the number and sentiment of reviews can affect player engagement. It concludes that games with more positive reviews see increased playtime, especially if those reviews highlight the game's quality (Lin et al. 2019). We found this problem suitable for linear regression because it allows us to model the relationship between average playtime (the dependent variable) and various predictors (price, ratings, developer, etc.). This approach helps quantify how each factor influences playtime, enabling a clear understanding of the impact of different game attributes. By fitting a linear regression model, we can determine which factors have statistically significant effects on playtime, providing actionable insights for game developers and marketers.

3 Methods

To develop and refine the final model for predicting average playtime, a structured approach was employed that combined diagnostic assessments, transformations, and model selection techniques. The analysis began with data preparation and selection of relevant predictors based on domain knowledge and preliminary exploration. A multiple linear regression model was initially fitted using predictors including price, positive ratings, negative ratings, number of achievements as numeric variables, and ownership levels as categorical variable.

Preliminary model diagnostics, since we are using MLR model, we first need to check conditional mean condition with both response and predictors, including response vs. fitted plots and pairwise plots between predictors. Then residuals vs. fitted plots were used to assess the assumptions of linearity, constant-variance, and uncorrelated errors. While Normal Quantile-Quantile (QQ) plot is used to check for normality. To address violations of these assumptions, Box-Cox transformations were applied to the response variable and select predictors to make the most suited transform function based on their λ. Then pairwise scatterplots and the Variance Inflation Factor (VIF) were used to check for multicollinearity, and highly correlated variables were removed or adjusted to improve the model's stability.

Refinement of the model was carried out through stepwise selection using AIC (Akaike Information Criterion) to identify the best combination of predictors while balancing model complexity and fit. ANOVA tests were used to compare nested models and evaluate whether reductions in predictors led to significant changes in explanatory power. The final model was validated using diagnostic plots to confirm adherence to linear regression assumptions, ensuring that it was suitable for interpretation and prediction.

The methodology was implemented in programming language **R** (R Core Team 2023). Packages **readr** (Wickham, Hester, and Bryan 2024), **httr** (Wickham 2023), **jsonlite** (Ooms 2014) were used to download the data, **dplyr** (Wickham et al. 2023) and **lubridate** (Grolemund and Wickham 2011) were used to clean the data. The diagnostic process is done with help of **ggplot2** (Wickham 2016) for residual graphs, **GGally** (Schloerke et al. 2024) and **car** (Fox and Weisberg 2019) for multicollinearity checks, **MASS** (Venables and Ripley 2002) for stepwise selection, **caret** (Kuhn and Max 2008) is used to perform box-cox transform on variables. This systematic approach ensured that each decision was guided by statistical evidence and diagnostic tools, allowing for the development of a robust final model. A simplified version of the methodology flowchart is presented in Figure 1.

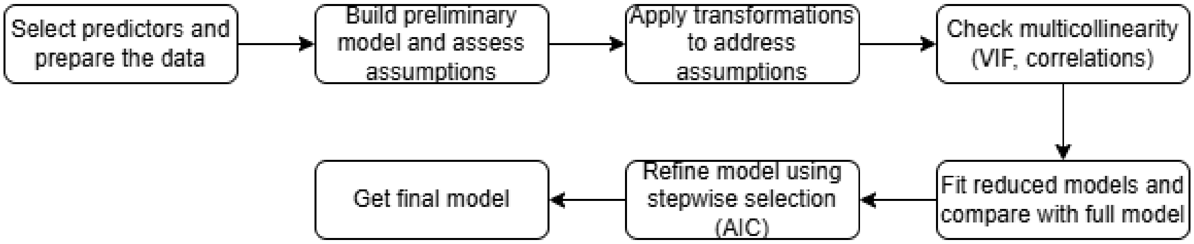
Select predictors and prepare the data
Build preliminary model and assess assumptions
Apply transformations to address assumptions
Check multicollinearity (VIF, correlations)
Fit reduced models and compare with full model
Refine model using stepwise selection (AIC)
Get final model

Figure 1: Methodology flowchart

4 Results

At the starting point, our baseline model uses `average_playtime` as the response variable, with `price`, `positive_ratings`, `negative_ratings`, `achievements`, and `owners` as predictors. The four key linear regression assumptions considered are linearity, uncorrelated errors, constant error variance, and normality. Additionally, since this is a multiple linear regression model, we also examine the conditional mean response and predictor assumptions.

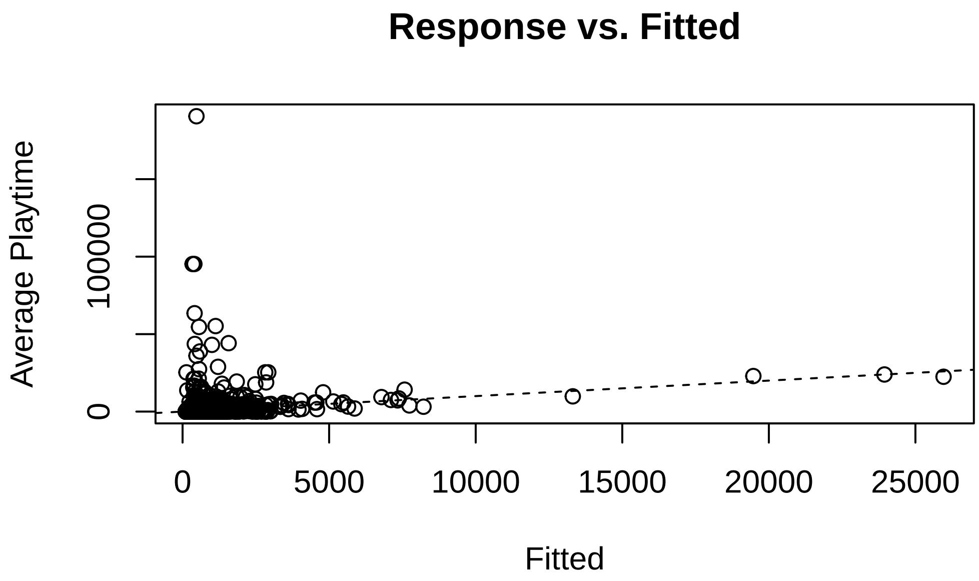

Figure 2: Response vs. Fitted plot to assess model fit.

To check the conditional mean response assumption, we analyze the Response vs. Fitted scatterplot Figure 2. The data points are tightly clustered in the bottom-left corner, which satisfies this assumption. The fitted line generally follows the line $y = \hat{y}$, but there are notable outliers along the Y-axis where the observed `average_playtime` deviates significantly from the predicted values. These extreme outliers suggest that the response variable may be highly skewed or contain problematic data points.

To check the conditional mean predictor assumption, we examine the pairwise scatterplots of predictors Figure 3. A lack of curves in all pairs of plots indicates that this assumption is satisfied. However, the Residuals vs. Fitted scatterplot Figure 4 shows a fanning pattern resembling a left-pointing arrow, which suggests a potential violation of the constant variance assumption. Additionally, the Residuals vs. Qualitative Predictor scatterplots Figure 7 indicate possible violations of linearity and uncorrelated errors due to clusters of points and curves in the `positive_ratings` and `negative_ratings` predictors.

Boxplots were used to assess linearity violations in the categorical predictor `owners`, and no issues were detected Section A. However, the Normal QQ Plot Figure 5 reveals significant deviations from the diagonal at the extremes, indicating a violation of the normality assumption.

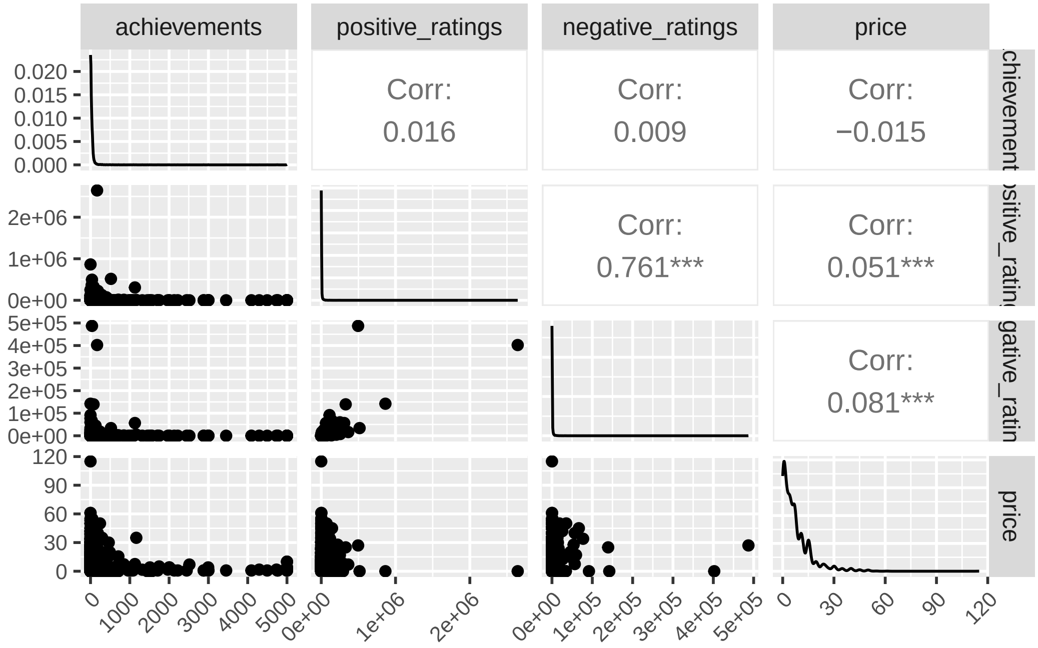

Figure 3: Pairwise plot between numeric predictors

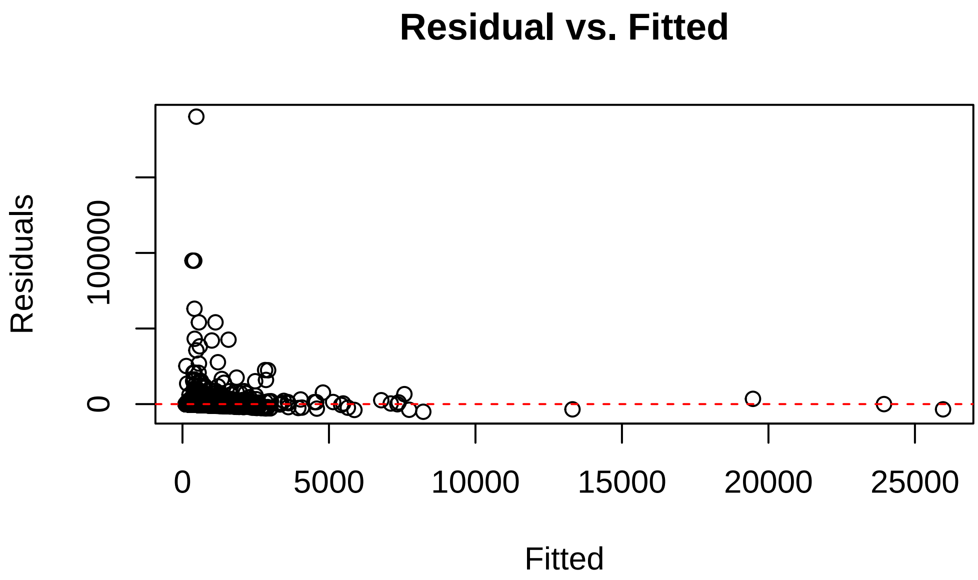

Figure 4: Residual vs fitted plot for model lm(average_playtime ~ price + positive_ratings + negative_ratings + achievements + owners, data=data)

The presence of these assumption violations aligns with the nature of the data, as confirmed by the outlier analysis in appendix Section A.4, which identifies several influential points even in the final model. For instance, the uncorrelated errors assumption may be violated because some predictors are inherently related (e.g., price and developer). These findings are consistent with those in the literature, where similar violations and errors have been reported. Addressing these violations through transformations is critical to ensure the stability of the model.

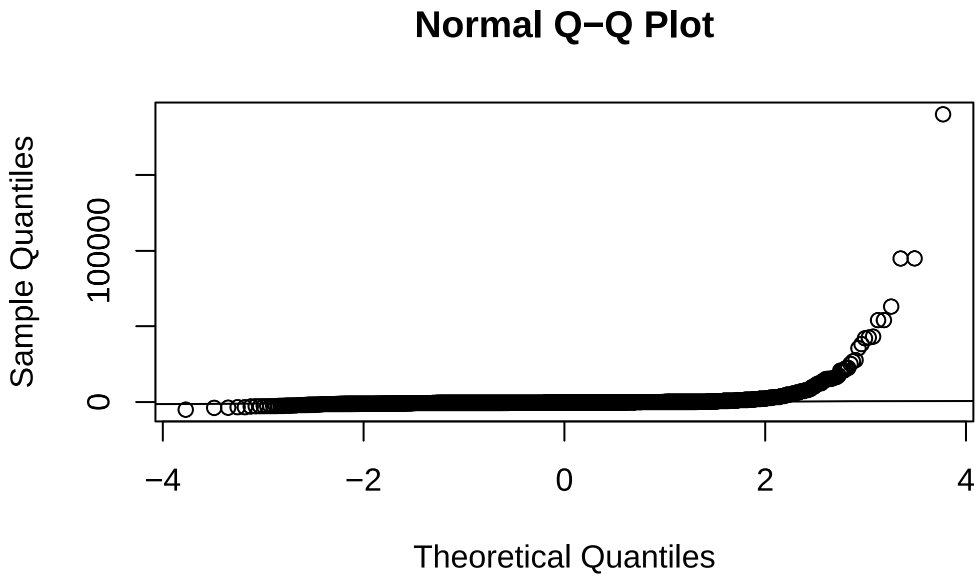

Figure 5: Residual vs fitted plot for model

To mitigate these violations, Box-Cox transformations were applied to all numeric predictors to improve normality and address linearity issues. Different transformations were applied to the response and predictors. For instance, the response variable `average_playtime` was log-transformed after calculating a Box-Cox λ 0.1, as an example, an original average_playtime value of 17,612 was transformed to 9.78. Note that a constant of 1 was added to all predictors before applying the Box-Cox transformation, as it requires strictly positive values.

After transforming the model, the Variance Inflation Factor (VIF) was calculated to check for multicollinearity. Notably, `bc_positive_ratings` (log-transformed `positive_ratings`) and `bc_negative_ratings` (log-transformed `negative_ratings`) exhibited moderately high VIF values of approximately 5 and 4.5 respectively, indicating multicollinearity. This is reasonable, as a higher number of positive reviews typically corresponds to greater attention and a larger player base of a game, which can also result in more negative reviews. The pairwise scatterplots Figure 3 confirm this, showing a positive linear correlation between these two predictors.

To address this issue, `bc_negative_ratings` was removed from the model, leaving the final predictors as `bc_price` (log-transformed of `price`), `bc_positive_ratings`(log-transformed of `positive_ratings`), `bc_achievements`(log-transformed of `achievements`) and `owners`. The refined model was re-evaluated with summary statistics and residual plots. The summary of coefficients as shown in Table 1 will be discussed in Section Section 5. The residual diagnostics Figure 6 show significant improvement, with a more random distribution around zero in the Residuals vs. Fitted plot and a QQ plot with better alignment to the reference line.

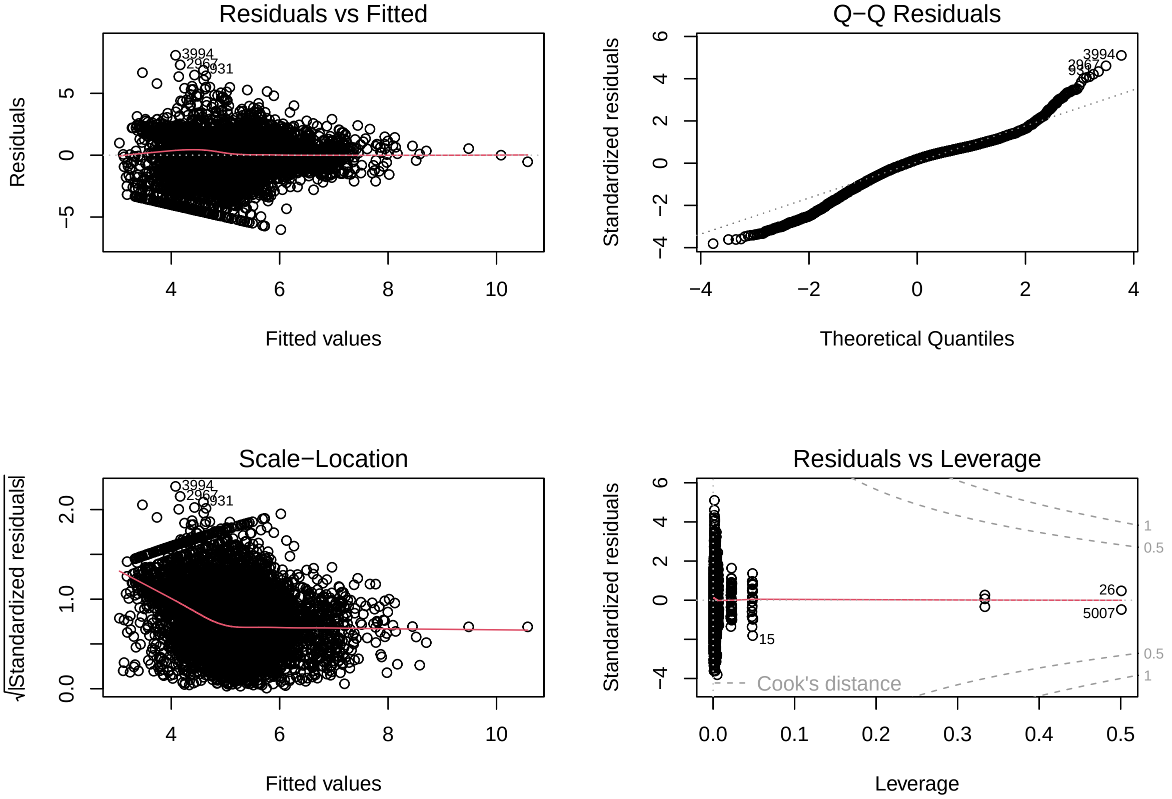

Figure 6: Diagnostic plots for lm(bc_average_playtime ~ bc_price + bc_positive_ratings + bc_achievements + owners, data=data)

To further simplify the model, automated selection procedures using AIC in both forward and backward directions were conducted. The procedure resulted in the same final model as described above, confirming its robustness.

5 Conclusion and Limitations

In conclusion, the model demonstrates that the average playtime of a game is positively influenced by several factors, including the game's price, the number of achievements, the number of positive reviews, and the estimated number of owners. Notably, the number of owners has the most substantial effect, as shown in the model summary Table 1. For instance, games with an estimated ownership level of 100,000–200,000 have a coefficient of approximately 0.81; as ownership grows by a factor of 10, the effect increases to ~1.70, and for another 10-fold increase, the effect grows to ~3.04. This indicates that games with a larger player base tend to have higher average playtime, likely due to increased community engagement and multiplayer opportunities.

The positive relationship between price and average playtime is also significant. This is consistent with the notion that higher-priced games often provide more content, greater replayability, and better quality, which encourages players to spend more time playing. Additionally, players who invest more money in a game may feel a stronger obligation to "get their money's worth," resulting in longer playtime. Similarly, a greater number of positive reviews likely attracts more players and increases playtime, as positive feedback signals quality and encourages engagement.

Overall, the model aligns well with practical expectations. It suggests that developers and publishers aiming to increase average playtime should focus on improving game quality, fostering positive user reviews, and building a strong player base.

This study faces limitations primarily due to assumption violations and data issues. The Residuals vs. Fitted plot revealed a fanning pattern, indicating heteroscedasticity, while the Normal QQ Plot showed significant non-normality. Outliers in the response variable `average_playtime` suggest skewed data, despite transformations like Box-Cox improving stability. Multicollinearity between `bc_positive_ratings` and `bc_negative_ratings` required removing the latter, though some interdependence remains.

The dataset, limited to Steam games, introduces platform-specific biases, restricting generalizability. Observational data prevents causal inference, and omitted variables, such as marketing strategies, likely impact playtime. Future studies should refine transformations, address multicollinearity more rigorously, and expand variables and platforms to enhance the robustness and applicability of the findings.

6 Ethics discussion

The Steam Games dataset (Davis 2023) used in this research contains metadata on over 97,000 games published on the Steam platform. The dataset is sourced transparently from Steam store pages and the Steam API, hosted on Kaggle, and widely used in existing literature, ensuring its credibility.

In our analysis, we chose automated selection methods due to the dataset's size and complexity. Manual selection methods were not used as they would have been impractical and prone to human error, reducing consistency and reproducibility. While both methods are ethically equivalent when executed carefully, automated methods better align with the ethical principles outlined in the second ethics module by minimizing foreseeable risks and ensuring reliability.

Avoiding blameworthy practices requires addressing foreseeability, avoidability, and potential errors. Automated methods allowed us to mitigate these risks efficiently, ensuring accountability in our approach. Manual methods, while suitable for smaller datasets, would have introduced avoidable errors and diminished reliability, which could be considered ethically negligent.

By prioritizing virtues such as precision and accountability, we ensured an ethically sound methodology. Our choice was guided by practical considerations without compromising ethical standards, fostering robust and reproducible results.

Appendix

A Model details

A.1 Residuals vs each Quantitative Predictor scatterplot to check Linearity, Uncorrelated Errors, and Constant Variance

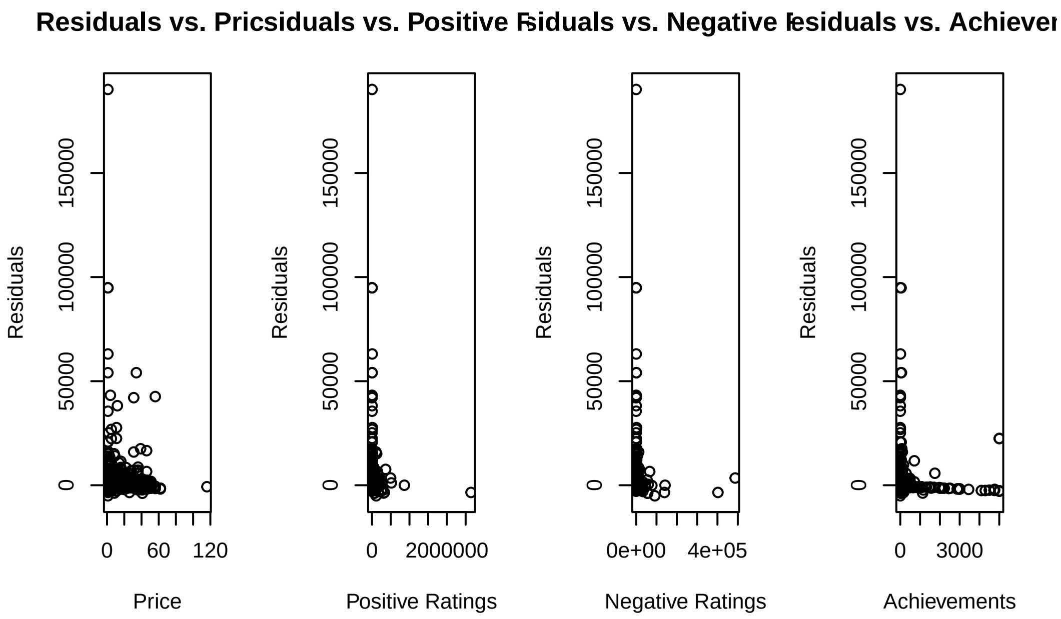

Figure 7: Residuals vs each Quantitative Predictor

A.2 BoxPlot to check assumptions for Categorical Predictors

See figure Figure 8

A.3 Final model results

Table 1: Summary of the Reduced Model

term	estimate	std.error	statistic	p.value
(Intercept)	2.9659147	0.0861182	34.440030	0.0000000
bc_price	0.3895455	0.0224633	17.341425	0.0000000

Table 1: Summary of the Reduced Model

term	estimate	std.error	statistic	p.value
bc_positive_ratings	0.0714378	0.0186039	3.839941	0.0001243
bc_achievements	0.0608850	0.0121352	5.017220	0.0000005
owners100000-200000	0.8144914	0.0853337	9.544775	0.0000000
owners1000000-2000000	1.7079929	0.1425771	11.979430	0.0000000
owners10000000-20000000	3.0419084	0.3772361	8.063672	0.0000000
owners100000000-200000000	6.1410922	1.5988890	3.840850	0.0001238
owners20000-50000	0.6760547	0.0766939	8.814970	0.0000000
owners200000-500000	0.9454732	0.0915972	10.322072	0.0000000
owners2000000-5000000	2.0279996	0.1665449	12.176894	0.0000000
owners20000000-50000000	4.4182711	0.9337416	4.731792	0.0000023
owners50000-100000	0.7583126	0.0816802	9.283927	0.0000000
owners500000-1000000	1.2317147	0.1178649	10.450222	0.0000000
owners5000000-10000000	2.6046296	0.2679580	9.720289	0.0000000
owners50000000-100000000	5.1525334	1.1366458	4.533104	0.0000059

A.4 Problematic points detection

Checking for problematic points-leverage cutoff: ~0.006, cook's distance cutoff: ~0.961, dffits cutoff: ~0.105, dfbeta cutoff: ~0.025. Number of influential points by each case-leverage points: 216, regression outliers: 7, Cook's distance: 0, DFFITS: 211, by Beta from 0 to 7: 431, 476, 377, 394, 464, 321, 49, 0.

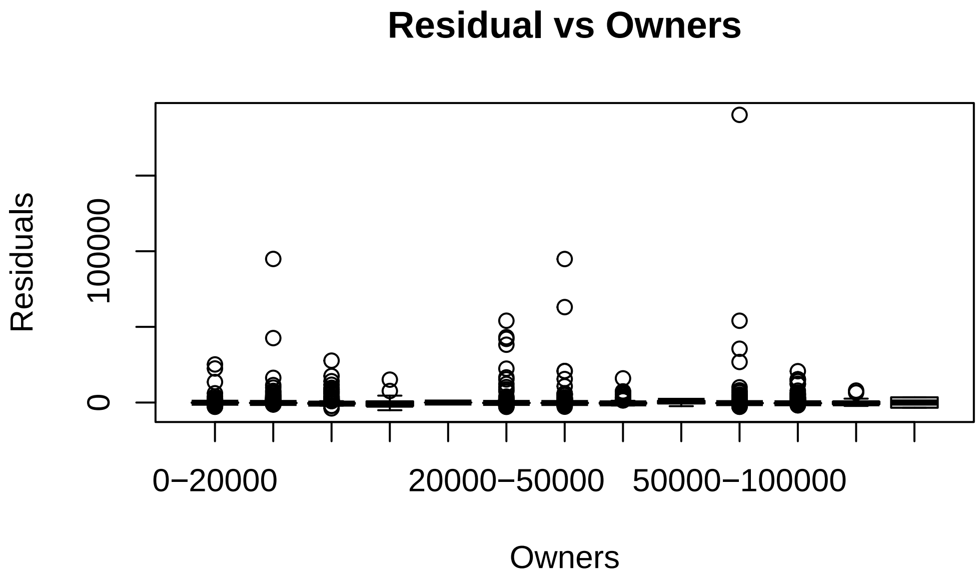

Figure 8: Boxplot residual vs. Owner levels suggest no linear assumption violation regarding the categorical predictors

References

Brodschneider, Vinzenz, and Johanna Pirker. 2023. "On the Influence of Reviews on Play Activity on Steam - a Statistical Approach." *ResearchGate*. https://www.researchgate.net/publication/376227235_On_the_Influence_of_Reviews_on_Play_Activity_on_Steam_-_A_Statistical_Approach.

Davis, Nik. 2023. "Steam Store Games (Clean Dataset)." https://www.kaggle.com/datasets/nikdavis/steam-store-games/data.

Fox, John, and Sanford Weisberg. 2019. *An R Companion to Applied Regression*. Third. Thousand Oaks CA: Sage. https://www.john-fox.ca/Companion/.

Grolemund, Garrett, and Hadley Wickham. 2011. "Dates and Times Made Easy with lubridate." *Journal of Statistical Software* 40 (3): 1–25. https://www.jstatsoft.org/v40/i03/.

Kuhn, and Max. 2008. "Building Predictive Models in r Using the Caret Package." *Journal of Statistical Software* 28 (5): 1–26. https://doi.org/10.18637/jss.v028.i05.

Lin, Dayi, Cor-Paul Bezemer, Ahmed E. Hassan, and Ying Zou. 2019. "An Empirical Study of Game Reviews on the Steam Platform." In *Empir Software Eng*, 24. https://doi-org.myaccess.library.utoronto.ca/10.1007/s10664-018-9627-4.

Luisa, Andraž De, Jan Hartman, David Nabergoj, and Samo Pahor. 2021. "Predicting the Popularity of Games on Steam." *ResearchGate*. https://www.researchgate.net/

publication/355110719_Predicting_the_Popularity_of_Games_on_Steam.

Ooms, Jeroen. 2014. "The Jsonlite Package: A Practical and Consistent Mapping Between JSON Data and r Objects." *arXiv:1403.2805 [Stat.CO]*. https://arxiv.org/abs/1403.2805.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Schloerke, Barret, Di Cook, Joseph Larmarange, Francois Briatte, Moritz Marbach, Edwin Thoen, Amos Elberg, and Jason Crowley. 2024. *GGally: Extension to 'Ggplot2'*. https://CRAN.R-project.org/package=GGally.

Venables, W. N., and B. D. Ripley. 2002. *Modern Applied Statistics with s*. Fourth. New York: Springer. https://www.stats.ox.ac.uk/pub/MASS4/.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. 2nd ed. Use r! Cham, Switzerland: Springer International Publishing.

———. 2023. *Httr: Tools for Working with URLs and HTTP*. https://CRAN.R-project.org/package=httr.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Wickham, Hadley, Jim Hester, and Jennifer Bryan. 2024. *Readr: Read Rectangular Text Data*. https://CRAN.R-project.org/package=readr.