1 Abstract

In recent years, the video game industry has been a stable segment of the entertainment industry and is on the rise globally. In this project, video games are analyzed from different aspects, such as user scores, critics, genre, and platforms that affect global video game sales. The project examines the relationship between users' scores and critic scores. It also examines the relationship between users' and critics' scores' influence on global sales. We also study other variables, such as game platforms.

2 Introduction

The project is tackling the video game industry based on the critics' and users' reviews. The game industry is a substantial market that is valued at 159 billion dollars. Thus, it is critical for companies to interpret the users' experiences with parameters such as users' scores so they can develop the games accordingly. This could be done by identifying many trends and analyzing them, such as game sales, reviews, and the most popular platforms. In this project, the focus is primarily on user reviews, critic reviews, and global sales to understand the relationship between them. Because critics have an impact on users' purchases, their reviews can impact the game's popularity and sales. So based on their reviews, sales can be affected dramatically if the review is harsh. On the other hand, if the review is positive, this can boost the game's popularity and, consequently, sales.

The dataset needed to be cleaned before using it in the project. There were many variables with NA values. By cleaning the dataset, the observations were reduced from approximately 17000 to 5000 observations. Additionally, the quantitative and qualitative variables became 15 variables. This range of information on video games includes Name, Genre, Platform, Publisher, Developer, Critic Score, User Score, Year of Release, Sales, and Sales by region

(North America, Japan, Europe, and Others). Note that sales are in millions, and scores range from 0 to 100. We are focusing on specific variables: Name, Genre, Platform, Critic Score, and User Score, that affect gaming sales. This dataset was published in Kaggle.

3 Platform Analysis

The goal of this model is to study the relationship between scores and platforms. In order to further analyze the effects of categorical variables on our models, we studied video game platforms' relationship with scores. the platform was picked as it was a shared predictor between our models.

Firstly, we check the scores' dispersion.

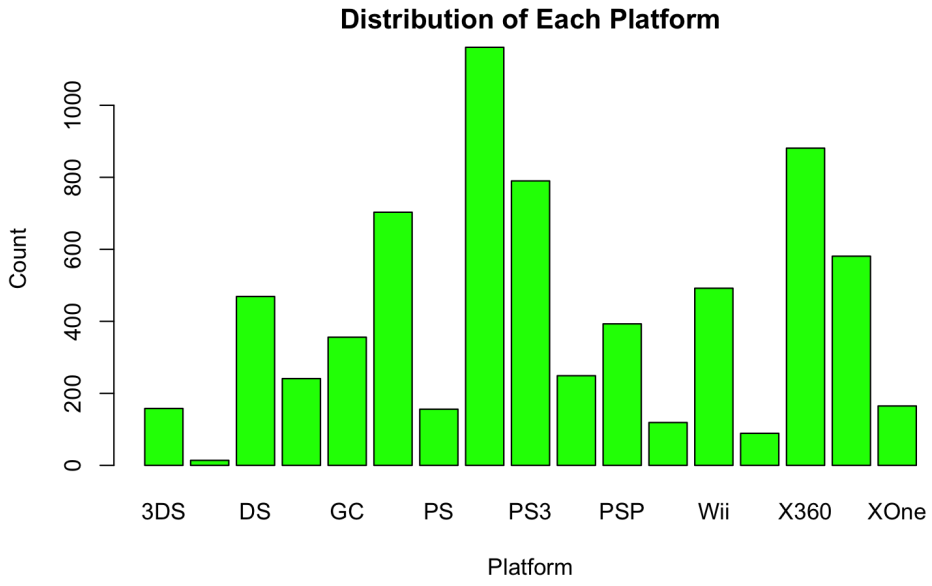

Figure 1: Distribution of Each Platform

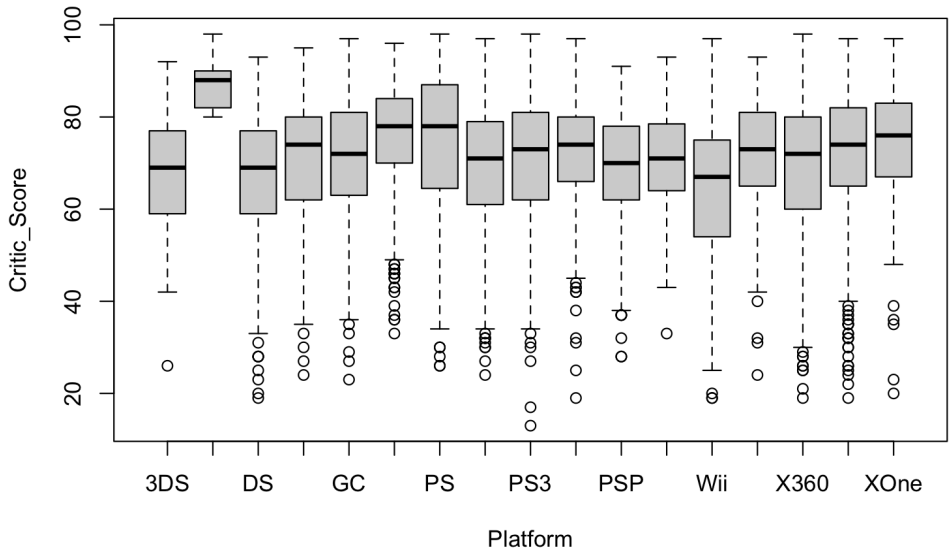

Figure 2: critic score vs. platform

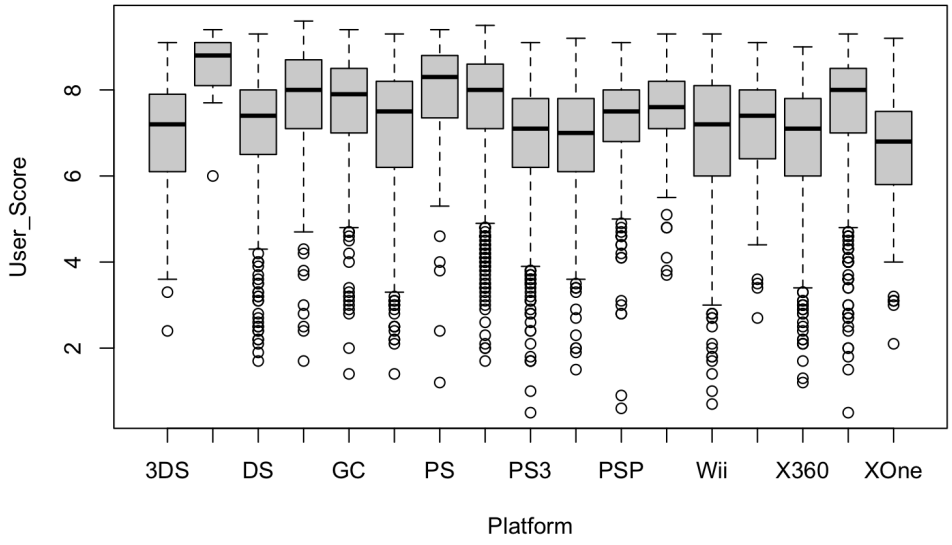

Figure 3: user score vs. platform

From Figure 2, the DC platform has the highest critic score, but since it only has 14 cases, it may be unconvincing. The Wii platform has the lowest critics scores. Most platforms

have symmetric critic scores, and their score variances are similar. It shows that the data is convincing and relatively fair.

From Figure 3, it can be shown that the DC platform has the highest user score too. No platform obviously has a low critic score. Most platforms have symmetric critic scores, which shows the data is convincing and relatively fair. We can notice that PS and PS2 games have obviously higher user-score than PS3.

Secondly, we fit the analysis of the variance model, the p-value is both smaller than 2e-16, which shows that the platform has a significant effect on the Critic Score and User Score at the 1% level.

We also check the assumptions of models.

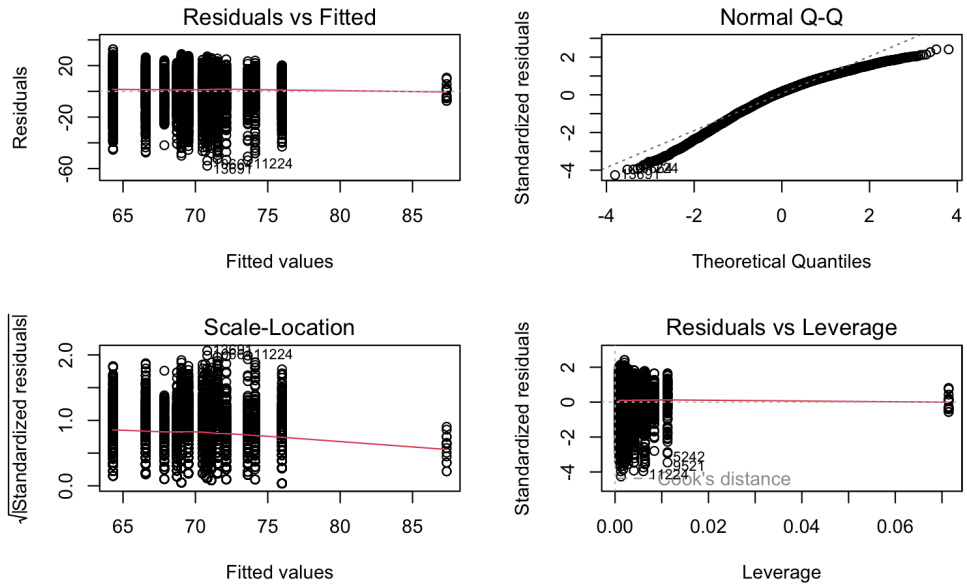

Figure 4: Critic-Score diagnostic plot

In the Critic-Score diagnostic plot, the residuals vs Fitted is good, the red line indicates that the residuals are centered at 0. In the Normal Q-Q plot, the points are close but do not precisely match the red line. In the Residuals vs Leverage plot, the red line is horizontal.

In the User-Score diagnostic plot. The quantile plot shows some departures. In the third plot, all points are below 1.0, and the red line is horizontal, indicating that the variances are similar.

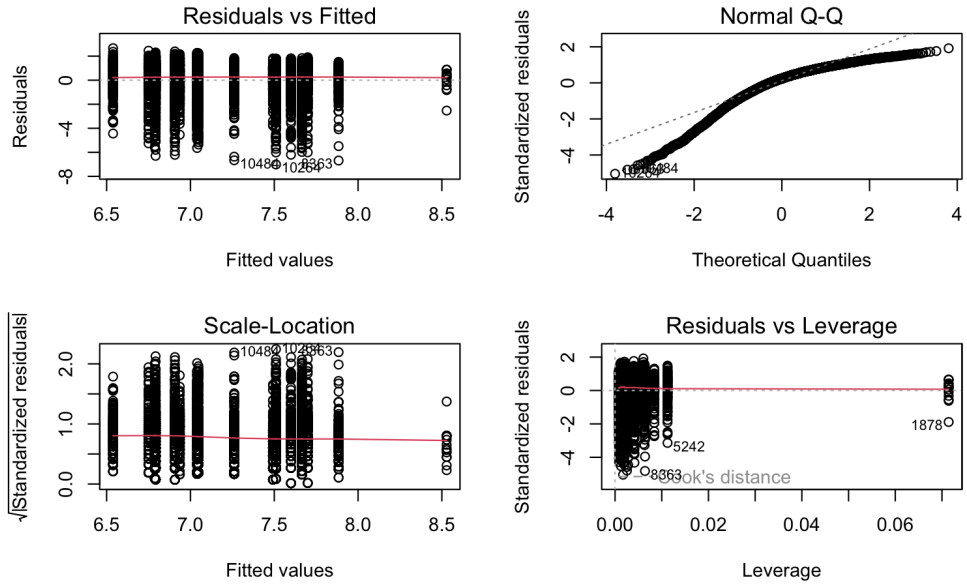

Figure 5: User-Score diagnostic plot

4 User Score Model

The goal of this model is to study the relationship between user scores and critic scores. The first model we built used only critic scores as its predictor. This model was tested using Tukey's test for nonadditivity; the P value for the test was significant and showed that a quadratic relationship exists between the scores. This was confirmed by the boxCox plot for the log-likelihoods of our predictors.

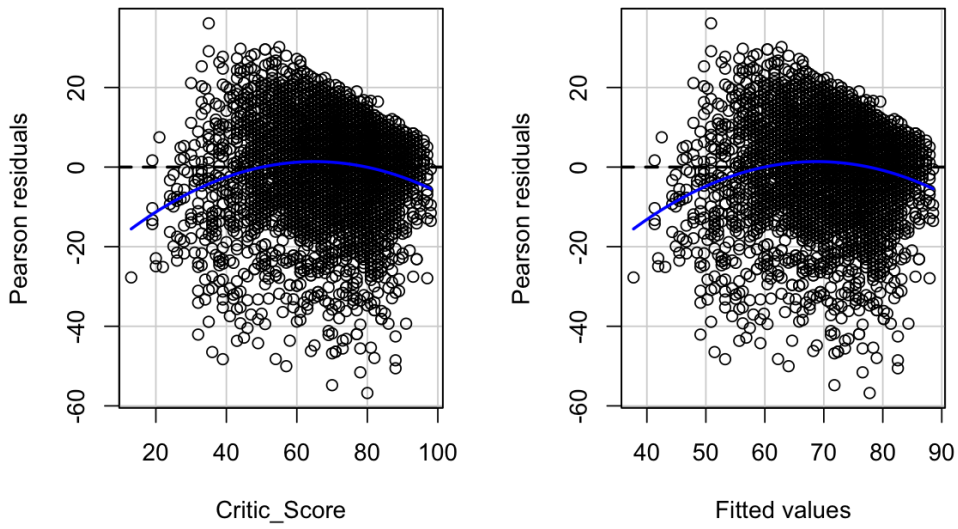

Figure 6: Pearson residuals plot for critic score

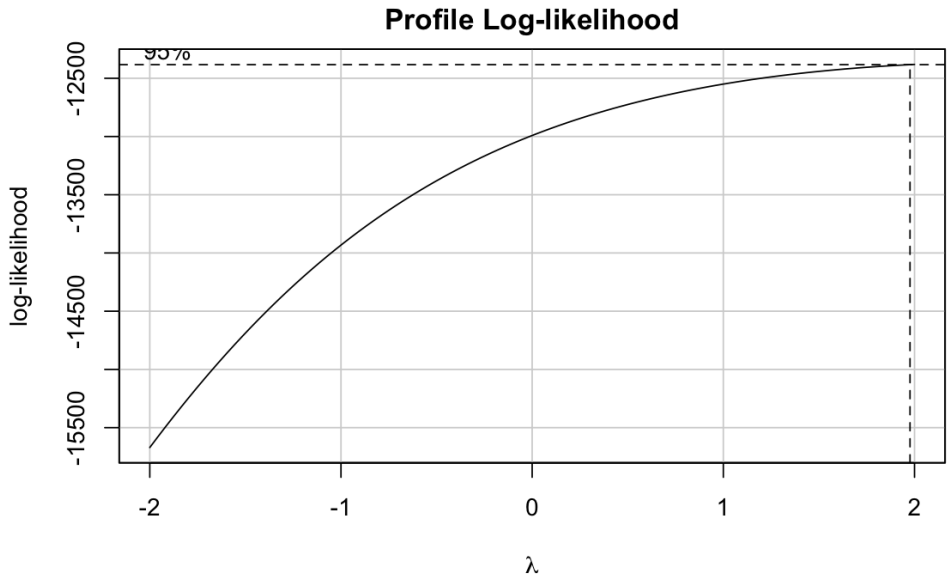

Figure 7: box-cox plot for user score vs critic score

4.1 Model Selection

After establishing this relationship, we built the full model with the following predictors: Critic score, Critic score squared, Rating, Genre, and Platform. We also included interactions between the Critic score and categorical models. Global sales were excluded from this model. This full model was then simplified using two methods, BIC and backward selection. The results from both methods were very similar, and in the end, BIC's model was chosen for its simplicity. Thus, our final model is:

```
lm(formula = User_Score ~ Critic_Score + I(Critic_Score^2) +
    Rating + Platform , data = sales_cleaner)
```

4.2 Model Assumptions

There are three assumptions for our model, linearity, Homoscedasticity, and Normality. These assumptions are checked through plots and tests.

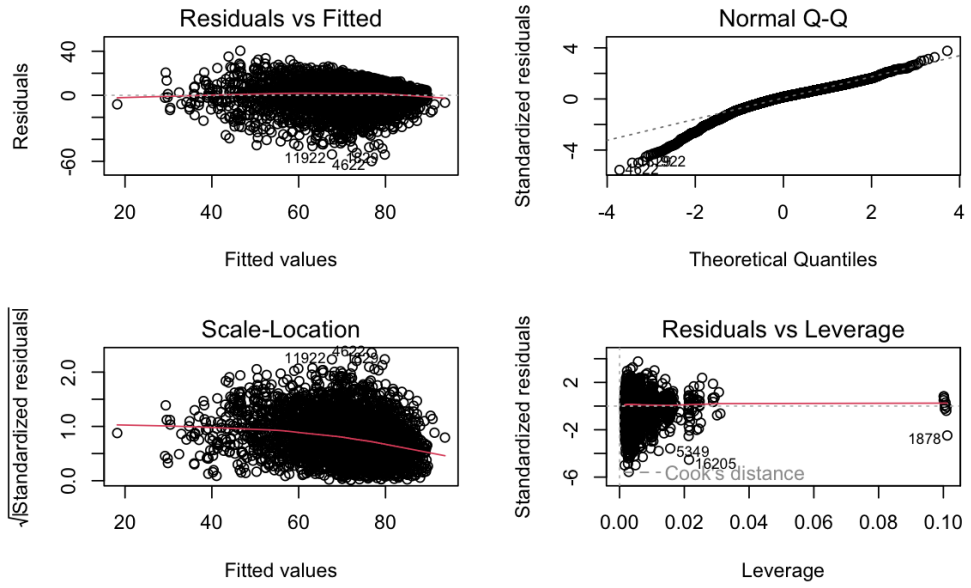

Figure 8: The Four plots for the selected model

Linearity: From the Residuals vs Fitted plot, we can see that there is a slight curve to our model relationship, even after incorporating a quadratic term into our model. Thus we tested our linearity assumption using the Tukey test, and the P value this time was high enough so that we do not reject the linearity assumption.

Homoscedasticity: From the Scale-location plot, we can see that there is a clear pattern of decrease for the line. After testing with the Non-constant Variance Score Test, the p-value is extremely small, which means we reject the homoscedasticity assumption.

Normality: The normal Q-Q plot does not show a normal distribution at all. after testing with Shapiro-Wilk normality test, we reject we get an extremely low P value and normality assumption.

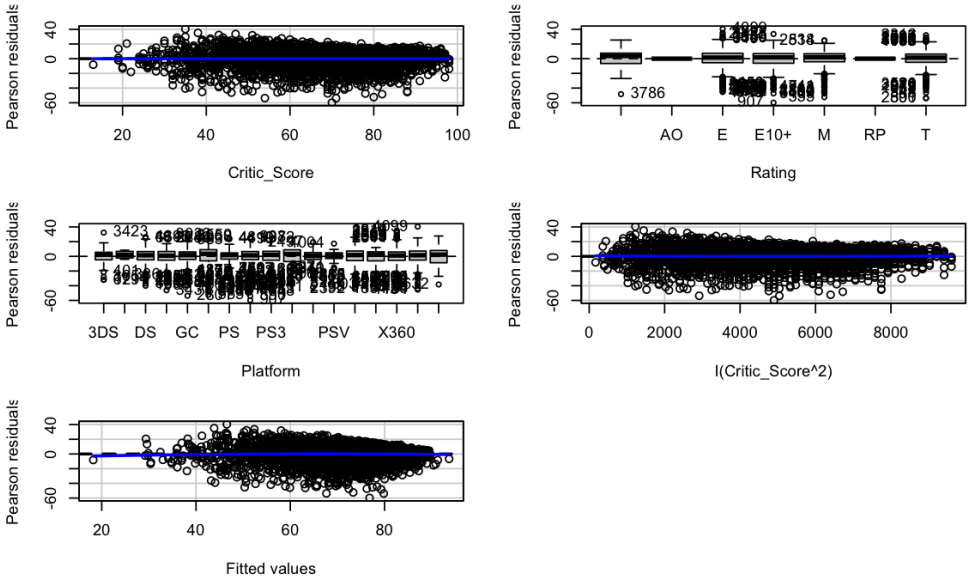

Figure 9: The Four plots for the selected model

4.3 Model Results

the adjusted R-squared of this model is 0.43, which is not great. it has an RSE of 10.78, and since our scores are between 1 and 100, it is a very significant error rate.

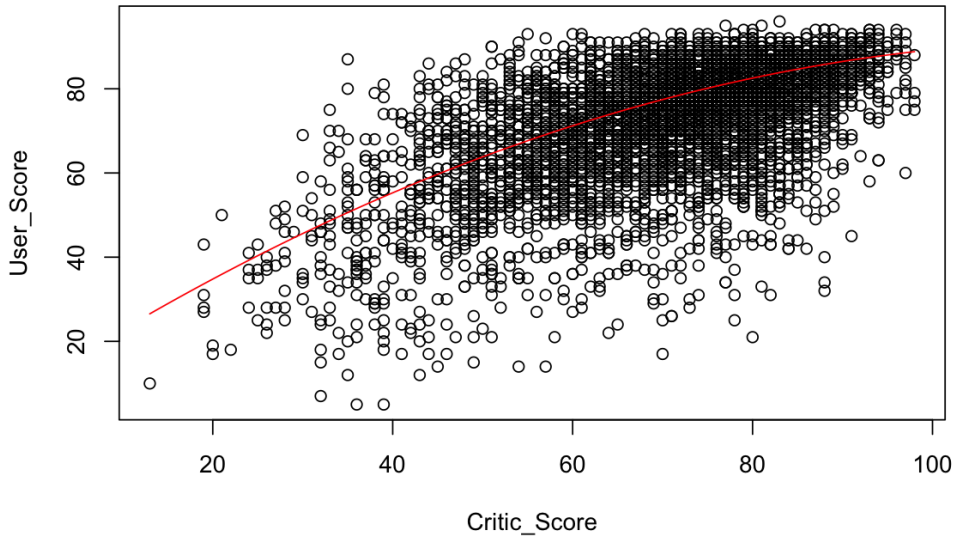

Figure 10: scatter plot for user score vs critic score with a curve for the model (categorical variables were averaged for this plot)

5 Global Sales Model

If we look at the correlation of all numeric variables with respect to global sales, we observe the highest correlation to be between global sales and critic and user counts. This is to be expected since the higher the number of copies sold, the more it is rated. To get a more practical model with which we can predict the number of sales, we should omit at least the user count since the data would not be available yet. Critic scores, on the other hand, are primarily published within a short time window of the release date and can therefore be used to make a sales prediction.

We create a multiple linear regression model using all available data as regressors, then use BIC to reach a minimally adequate model. The resulting model has an adjusted R2 value of only 0.1455, using the platform, year of release, critic score, critic count, and user score as the regressors. This indicates that the age rating is not significant to global sales. However, as seen in the diagnostic plots below, the model fails both normality and homoscedasticity.

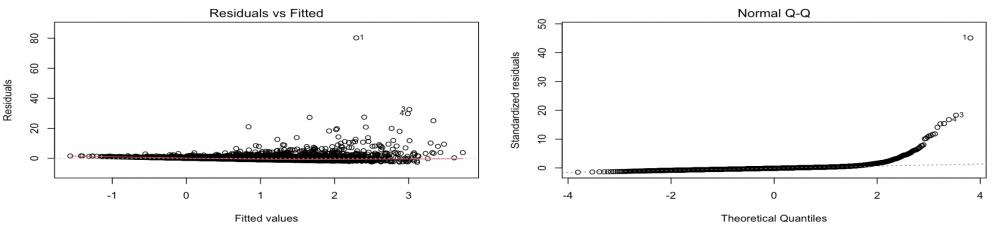

Figure 11: diagnostic plots of linear model

We can try adding quadratic numeric terms (critic score, critic count, user score) to the model and use the BIC process to find a minimally adequate problem, which increases the adjusted R2 value to 0.1735. Still, yields similarly terrible diagnostic plots, failing our normality and homoscedasticity assumptions.

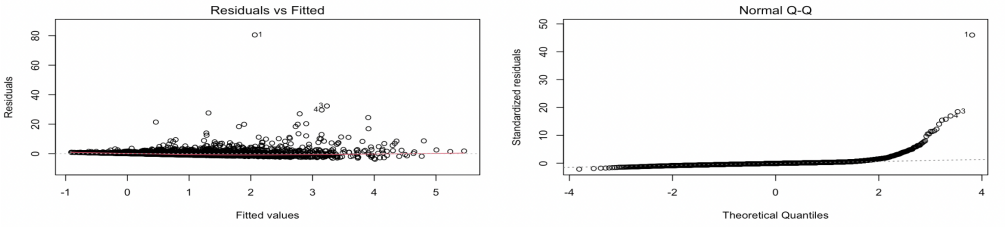

Figure 12: diagnostic plots of quadratic model

To improve the model, we look back to the earlier correlation plot and observe that, counter to intuition, global sales are not correlated with the year of release. It might be a fair assumption to think that the older the game, the more time it has had to sell or that the newer the game, the higher the sales due to the growing number of video game players. We explore the possibility of changing the year of release from a numeric term to a factor, which

could address this possible correlation discrepancy. If we look at the adjusted R2 of the model with a numeric release date before the BIC process, we get an R2 value of 0.1663. If we change the year of release to a factor, the R2 value increases slightly to 0.17 but still gets eliminated with BIC. The year of release term is ultimately not very significant. Even if we choose to include it in the model (which would not account for much more), it would be better to treat it as categorical data. There is no clear analytical expression to account for the relationship between global sales and the year of release. Based on the diagnostic plots and the very poor R2 value, we conclude that we cannot reliably predict global sales based on the data available. We could try adding user count, which is another variable with a high correlation to global sales, but it would no longer be practical, seeing, as discussed before, we would not have access to such data. However, for the sake of exploration and completion, we now include user count in the model with both linear and quadratic terms. The adjusted R2 increases to 0.1858, but we once again get terrible diagnostic plots, not satisfying either of the previous assumptions, as shown in the figure below.

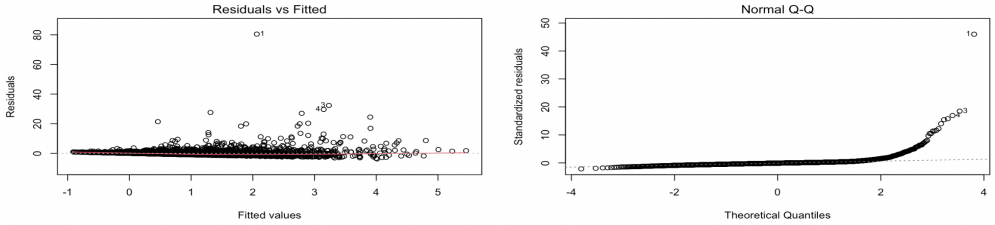

Figure 13: diagnostic plots with user count

Even with user count taken into account, which we would usually assume to be somewhat proportional to global sales, we cannot reliably predict the number of sales. Making any meaningful prediction would require many more variables unavailable in this dataset.

6 Conclusion

As our first model shows, there does exist a relationship between user scores and critic scores. However, this relationship is not very well covered by linear (or quadratic) models. Thus, while critic scores do seem to have an effect on user scores, it is perhaps weak and requires a more complex model to be captured. On the other hand, the results for the second model for global sales show that none of our other variables (and any combination of them) can be significant predictors for video game sales.