Do You Even Play Games?

A statistical analysis of video game sales

STAT 504 Applied Regression Final Project

Anh-Minh Nguyen, Kaelan Yu, Wenxian Fei

Background/Problem Statement

Video game sales are always influenced by many factors. As a huge market with great potential, identifying those factors can help people make appropriate market strategy and develop better video game for players. In 2018, Maksim Klimentyev fit multiple models (e.g. ridge regression, support vector machine) to predict global video game sales and compared their performance on the test data set using mean absolute error.

Goal

Find the most significant factors influencing video game sales in North America.

Dataset Description

This data set contains a list of video games with sales greater than 100,000 copies along with critic and user ratings. It is a combined web scrape from VGChartz and Metacritic along with manually entered year of release values for most games with a missing year of release. The original coding was created by Rush Kirubi and can be found here, but it limited the data to only include a subset of video game platforms. Not all of the listed video games have information on Metacritic, so this data set does have missing values.

Data Cleaning

Original Dataset

Step 1: Convert year to age: **age = 2020 - year**
Step 2: Remove variables with little practical significance (e.g. global sales)
Step 3: Decrease the levels of categorical variables by grouping appropriately

For example, for the variable **rating** we would group the following levels: **M, RP, T** into a new category called **"Mature"**

Step 4: Remove categorical variables with too many levels (e.g. name, platform)

Step 5: Remove User Count as a predictor:

- Severely right skewed

- The plot of log(NA_Sales) vs User_Count (below) is poorly defined

Step 6:
- Removed outlier "Wii Sports" since it was bundled with the Wii console release
- Removed Rare Platforms: Data is skewed left in terms of games release recently, older consoles
- Filtered out observations with null values in Critic_Score & User_Score: Continuous Values
- Name of row: Name + Platform + Year_of_Release – Games released amount + Competition within company

Initial Data
17416 Observations
14 Predictors

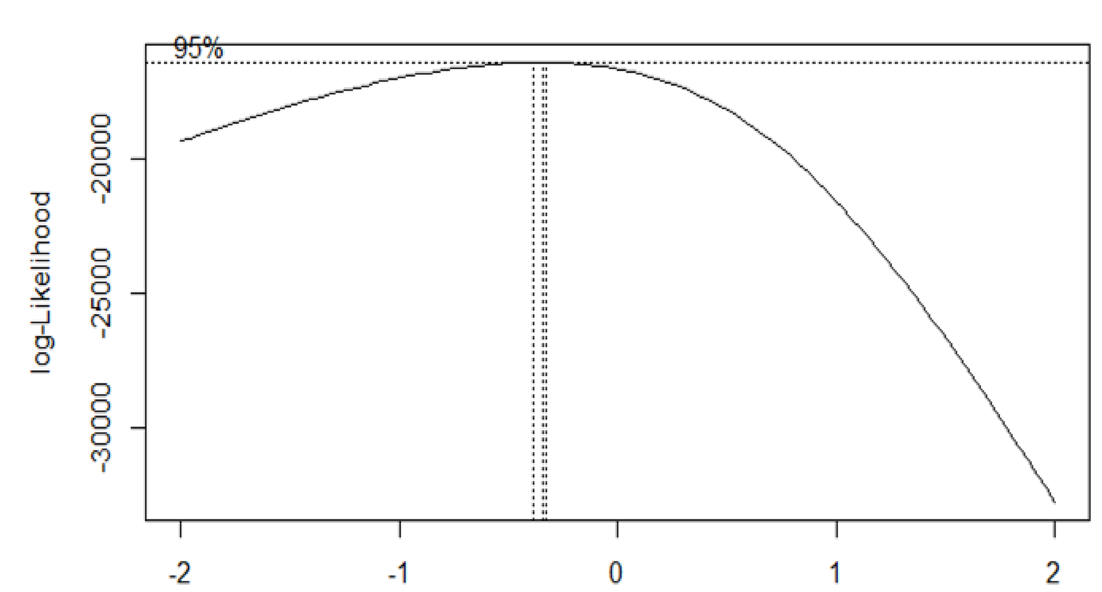
Cleaned Data
4244 Observations
7 Predictors

Multicollinearity

After removing User_Count as a predictor, there is only moderate positive correlation (at best) between User_Score and Critic_Score.

Furthermore, VIFs are all now below 5 so multicollinearity does not appear to be an issue.

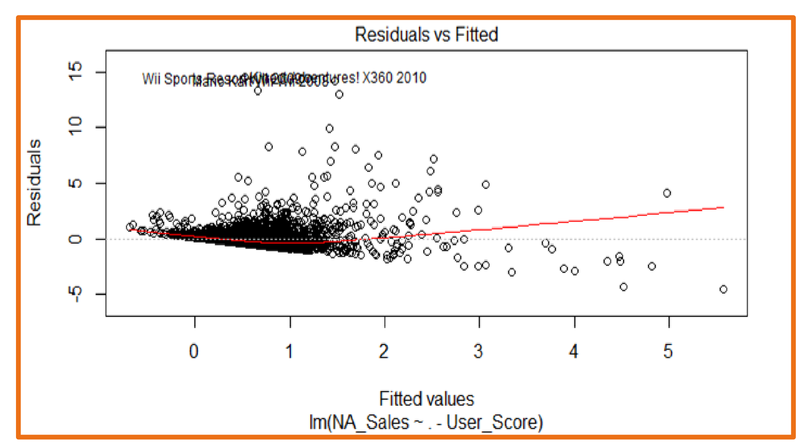
Box – Cox Transformation



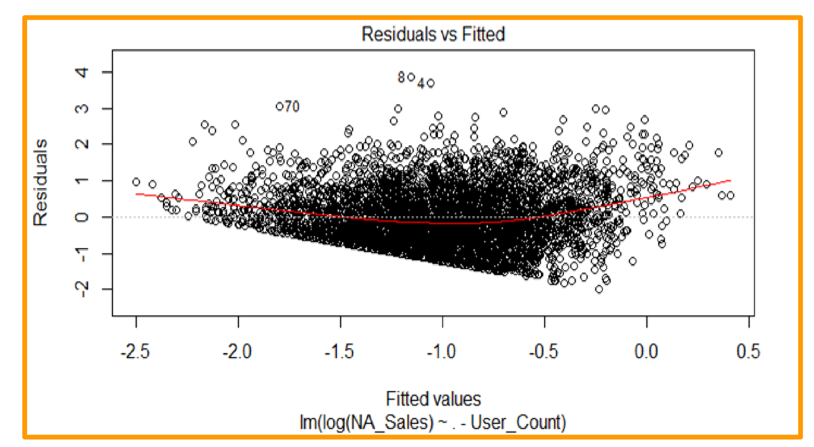
Lambda does not contain 0 but we will do a log transformation instead:

- E(Y|X) ~ V(Y|X)
0.374 ~ 0.998

- A log transformation is also more interpretable



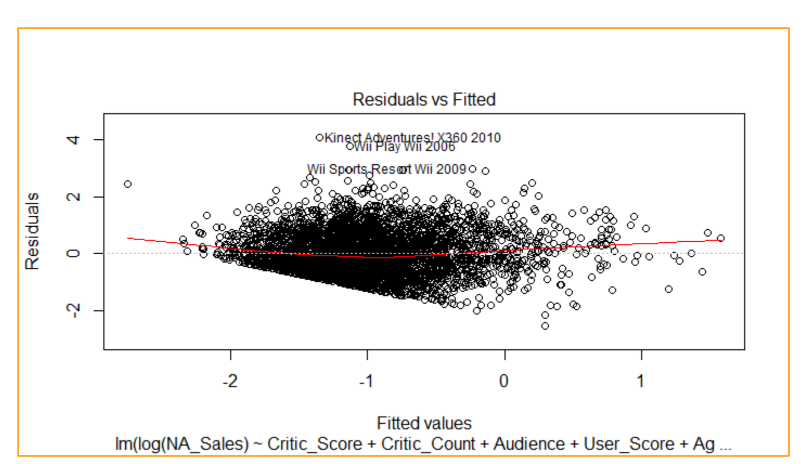
TA plot before Log Transformation



TA plot after Log Transformation

Model Diagnostics & Selection

- Threw in every single possible interaction term then only 3 interactions due to "singularities" or aliased coefficients
- BIC Selection: Forwards & Backwards
- Chose Forward Selection



TA plot: Forward BIC selection

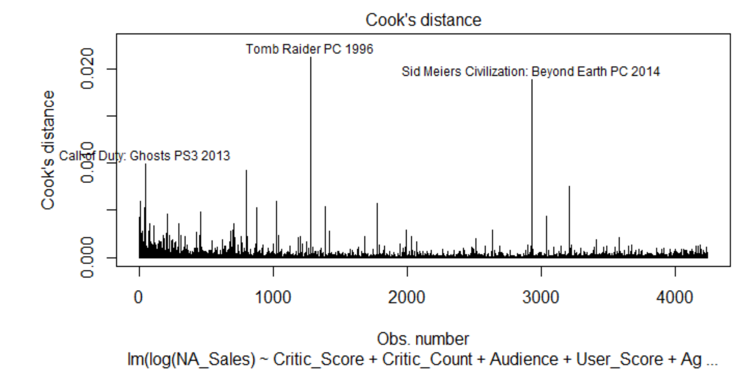
	log(NA_Sales)		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	0.02	-0.55 – 0.59	0.948
Critic_Score	-0.00	-0.01 – 0.00	0.391
Critic_Count	-0.03	-0.04 – -0.03	**<0.001**
Audience [Mature]	-0.04	-0.34 – 0.25	0.764
User_Score	-0.16	-0.23 – -0.09	**<0.001**
Age	-0.08	-0.12 – -0.04	**<0.001**
Brand [PC]	-2.76	-4.34 – -1.18	**0.001**
Brand [Sony]	-1.60	-1.97 – -1.24	**<0.001**
Brand [Xbox]	-0.39	-0.81 – 0.03	0.067
Critic_Score * Critic_Count	0.00	0.00 – 0.00	**<0.001**
User_Score * Age	0.01	0.01 – 0.02	**<0.001**
Age * Brand [PC]	0.17	0.13 – 0.20	**<0.001**
Age * Brand [Sony]	0.03	0.02 – 0.05	**<0.001**
Age * Brand [Xbox]	-0.04	-0.06 – -0.03	**<0.001**
Audience [Mature] * Brand [PC]	0.74	0.16 – 1.32	**0.013**
Audience [Mature] * Brand [Sony]	0.43	0.31 – 0.54	**<0.001**
Audience [Mature] * Brand [Xbox]	0.39	0.26 – 0.53	**<0.001**
Critic_Score * Brand [PC]	0.00	-0.01 – 0.02	0.609
Critic_Score * Brand [Sony]	0.01	0.01 – 0.02	**<0.001**
Critic_Score * Brand [Xbox]	0.01	0.00 – 0.01	**0.001**
Critic_Count * User_Score	-0.00	-0.00 – -0.00	**<0.001**
Critic_Count * Age	0.00	0.00 – 0.00	**0.003**
Audience [Mature] * User_Score	-0.05	-0.09 – -0.02	**0.004**
Observations	4244		
R^2 / R^2 adjusted	0.324 / 0.321		

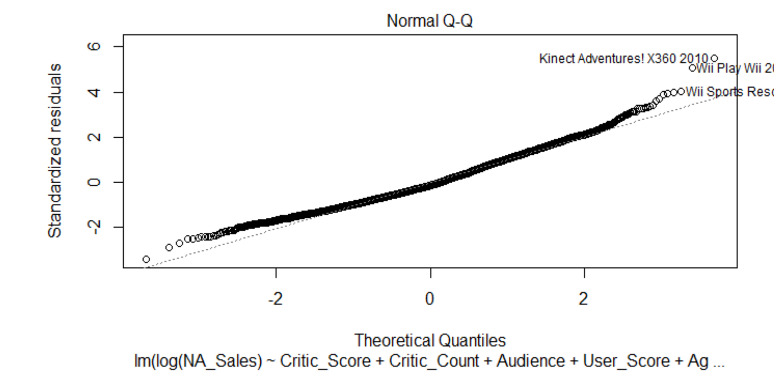
Modeling

Linear Regression

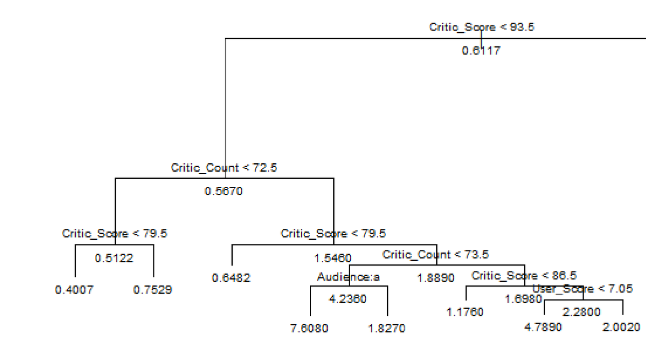
MSE: 0.5498

- Normality and constant variance assumption holds
- No influential points





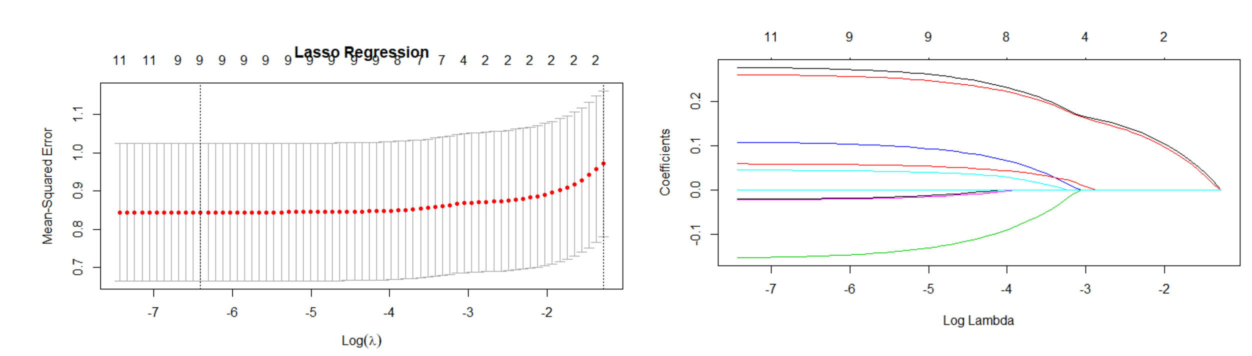
Regression Tree



- Cost-complexity pruning of tree to 10 nodes
- Decision nodes mostly involve continuous predictors

Training MSE: 0.7208
Testing MSE: 0.7122

Lasso Regression



(Intercept) .
Critic_Score 1.095094e-16
Critic_Count .
User_Score .
Age .
Brand_Nintendo .
Brand_PC .
Brand_Sony .
Brand_Xbox .
Audience_Family .
Audience_Mature .
Category_Action .
Category_Other.

Optimal Lambda: 0.2767 **Training MSE:** 1.3377 **Testing MSE:** 1.4522

Conclusions

- According to our multiple linear regression model, the interactions between the rating M and the Xbox, Sony, and PC group of consoles were significant. (Games like Halo, Uncharted, etcx)
- The regression tree had the smaller test MSE among the prediction models.
- Critic score appears to be the most significant predictor of video game sales in each prediction model, but it is not significant in our inference model.

Future Work

Some possible next steps to develop this project further include:

- Creating new variables (as transformations of predictor variables) and refitting our current models
- Modifying existing models (e.g. using random forests/other ensemble methods in place of the regression tree model) to improve predictive performance of video game sales
- Building regression models to predict video game sales in other regions (e.g. Japan)
- Using statistical methods to fill in missing data, as we had to remove many top hits published by Nintendo (Be careful of removing important observations)

References

Dataset
https://www.kaggle.com/kendallgillies/video-game-sales-and-ratings

Background Information
https://www.kaggle.com/maxklimentyev/video-games-predicting-global-sales