# The Establishment of Multi-variable Linear Regression in Steam Sales

Name: Haocheng Zhang

*San Jose States University, Xiamen, China*
*Email: 0903006293asd@gmail.com*

**ABSTRACT**
Based on the development of the Internet, the game platform launched by the Internet companies led by Valve has changed the way consumers purchase games. Compared with the previous methods based through CDS, the emergence of game platforms makes game developers have more sales methods, such as progressive maintenance of games, the releases of subsequent content in the form of DLC and other forms, and the establishment of communities among players, release modules and participate in the development. Noted that promoting on gaming platforms has become the primary means of selling in process of game development, thus developers have moved the competition online. Steam is Valve's gaming platform that offers digital rights management, game publishing, streaming, social networking and more. This paper takes Valve's Steam platform as an example and uses multiple linear regression equation to explain the variability in game ownership.

***Keywords:*** *Steam platform, games, multiple linear regression*

## 1. BACKGROUND

Steam is currently the largest game publishing platform in the world. As of 2020, Steam has 1.2[1] monthly active users and 62.6 million daily active users. In terms of purchasing power, the Steam platform data in 2020 showed an increase of 2.6 million monthly users[2]. Steam's growing size has attracted a growing number of developers to choose Steam as their primary distribution channel, and even its competitors are putting their games on Steam. For example, Ubisoft is considering re-launching its games on Steam with its own games platform (UPlay) ([6]).

The emergence of Steam platform saves the weakening and withering distribution channel of PC video games. Before Steam's age, the marketing and release of video games highly rely on local game shop while this becomes a challenge for some rural area and developing countries. In the case of online distribution of goods, the sales situation of games is different from that of disc distribution in the early days of the game industry, and the channel for users to obtain game news has correspondingly shifted to online. Based on the Steam platform, developers can save a lot of marketing costs and have more opportunities to communicate with players. With the existence of the Steam community, improvements proposed by players can be quickly fed back to developers, and then improvements can be made in subsequent patches while it seems not possible for customers in pre-steam age to get patches of the games they purchased in local game shop. The presence of game reviews provides consumers a basic understanding of the game's information before they purchase, influencing their decisions. Some games also use so called "workshops" to allow players to create their own mods to make the game more playable or to make subsequent content updates without official support which significantly extent the life cycle of games. Therefore, the difference in sales between games is not limited to differences in gameplay, but is the result of multiple factors interacting with each other. These factors came together make it possible to build a model to explain the variability of sales among games. We collect explaining information from different platform including Is There Any Deal and Playtracker.

## 2. Methodology

In this section we discusses the data sources used for analysis, exploratory data analysis, to provide an overview of the games being sold on Steam.

### *2.1. Source of Data*

 

This study uses Steamworks store data and a random sample of 20,000 sold games from 140,000 sold games using an R language package. Collect game data on Steam Spy based on appID, an index for games sold in Steam platform, in store data including 1. Game developer, 2. Publisher, 3. Number of good reviews, 4. Number of negative reviews, 5. Average hours played since March 2009, 6. Average hours played in the last 2 weeks, 7. Median hours played since March 2009, 8. Median hours played in the last 2 weeks, 9. Release price (USD), 10. Number of supported languages, 11. Type of game. Actual game sales are not disclosed due to Valve's privacy policy.

Playtracker Insight uses the platform to randomly collect public data from players around the world, and is 90%certain that actual data to be within 10% of a given game[3]. Playtracker provides the number of games being sold as a response variable to this study. Steam promotions are so frequent that it's important to take Into account Steam discount data in the models that explain sales differences. IsThereAnyDeal records the average level of discount, the average duration of the discount, the number of discounts, and the average price since launch. These data provides valuable information so that we are able to focus on how change in game price interacting with game sales, unlike Playtracker, IsThereAnyDeal provides actual discount and price fluctuation data of Steam platform.

### *2.2. Explanatory variable*

According to previous studies, there is a phenomenon on Steam that during the promotion period of games, the sales of games will be greatly increased. The lower the average price of games, the higher the sales of games [4]. The strength of the game's discount, the length of each discount, the game's release time, and user reviews were all selected as explanatory variables for variation in game ownership. Reviews are an important factor affecting consumers' decisions. Before consumers make decisions, reviews of related products determine their main driving force[5]. Steam offers community features that allow players to comment on games after purchasing them and in which there are player's information available by viewing their profile such as gaming time and whether they actually buy the game, people can decide the value of the comment. Since Steam is a global consumer, the amount of language support for the game we believe will affect sales. Based on the above, we can make the following assumptions. 1. There is a positive correlation between game sales and reviews. 2. There is a positive correlation between sales and reviews. 3. There is a negative correlation between game sales and average price. 4.There is a positive correlation between game sales and release time. 5. Game sales are positively correlated with language support. 6. Game correlation is positively correlated with discounts (number of discounts, intensity, duration, average duration)

## 3. EXPLORATORY DATA ANALYSIS

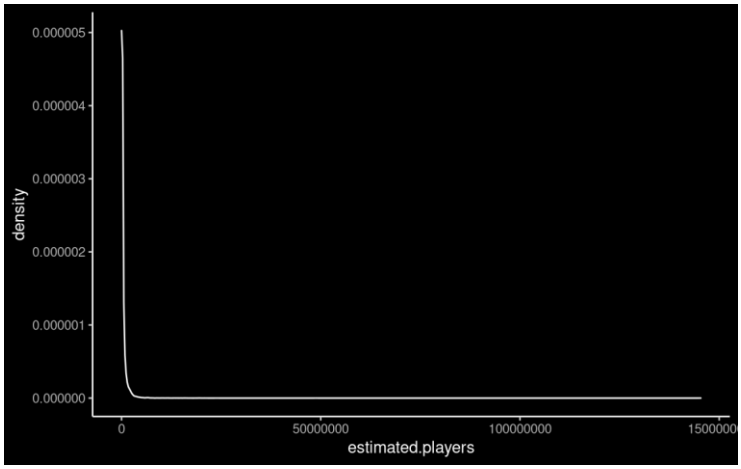


**Figure 1.** Game sales distribution

The R package's GGplot2 is used to make a group graph to give a quick view of distribution of estimated players. Input the game sales data of Playtracker Insight (n=24358), and the data shows an extreme right skew distribution which means that only small amount of game have a dominant sales while and large amount of games Cut up a small portion of the remaining market shares in Steam platform reflecting highly disparity. The mean value was 356,223 and the median was 96,000.

Correlation analysis Pearson's correlation coefficient in R language reflects the relationship between sales volume and explanatory variables, at a significant level of 0.05. The correlation test including correlation coefficient and correlation test using R package is shown in the following table.

**Table 1.** Correlation Test

| Variable | Correlation Coefficient | T | P |
|---|---|---|---|
| Sales and reviews | 0.78(CI=(0.78,0.79)) | 198.96 | 2.2*10^-16 |
| Sales and positve reviews | 0.04 CI=(0.03,0.05) | 7.0458 | 1.893*10^-12 |
| Sales and average price | 0.00 CI=(-0.01,0.01) | -0.00039919 | 0.9997 |
| Sales and release time | 0.26 CI=(0.25,0.27) | 43.026 | 2.2*10^-16 |
| Sales and language support | 0.09 CI=(0.08,0.10) | 15.164 | 2.2*10^-16 |
| Sales and discounts | 0.03 CI=(0.01,0.04) | 4.8774 | 1.082*10^-6 |
| Sales and average discount duration | -0.01 CI=(-0.02,0.00) | -1.6538 | 0.09818 |
| Sales and average discount strength | 0.09 CI=(0.08,0.1) | 15.586 | 2.2*10^-16 |

The results showed that the average discount duration and average price don't have enought evidence to support that there is a correlation with sales in which the correlation test under the significant level of 0.05, and there was a strong positive correlation between the number of user reviews and game sales (P <0.05), and the correlation coefficient r = 0.78CI = (0.78, 0.79). The average number of discounts showed a weak positive correlation (P <0.05).

## 4. ESTABLISHMENT OF MULTIPLE REGRESSION

Based on the previous relationship hypothesis, establish a multiple regression equation between game sales and the selected explanatory variables. The estimated players from Playertracker estimation that is from various word-wide social media platforms is assigned to be the response variable in model , while we select some variables as explanatory variables. They are positive review rate from players on Steam platform, average price of a specific game, days from its original releases, the number of languages supported by publisher, the number of discount which we denoted as cut_times in formula, and average extent of discount during the game's sales period.

The linear regression model formula is shown as follow.

$$\begin{aligned} \text{Estimated.players} \sim\ & \text{review\_num} \\ & + \text{positive\_rate} \\ & + \text{average\_price} \\ & + \text{days\_from\_releases} \\ & + \text{languages\_supports} \\ & + \text{cut\_times} + \text{duration} \\ & + \text{average\_cut} \end{aligned}$$

**Table 2.** Regression report

| Variable | Coefficients | SE | T | P |
|---|---|---|---|---|
| Intercept | -3.245*10^5 | 1.974*10^4 | -16.441 | 2*10^-16 |
| Review num | 7.69*10^1 | 3.419*10^-1 | 209.703 | 2*10^-16 |
| Positive rate | 4.720*10^4 | 1.975*10^4 | 2.390 | 0.0169 |
| Average price | -2.580*10^2 | 1.303*10^2 | -1.980 | 0.0477 |
| Release time | 3.158*10^2 | 5.747 | 54.951 | 2*10^-16 |
| Languages supports | 7.152*10^3 | 1.058*10^3 | 6.763 | 1.39*10^-11 |
| Discount times | -6.985*10^2 | 4.372*10^2 | -1.598 | 0.1101 |
| Average discount duration | -1.636*10^4 | 1.584*10^3 | -10.330 | 2*10^-16 |
| Average discount strength | 3.646*10^5 | 2.739*10^4 | 13.310 | 2*10^-16 |

In the regression equation, at the significance level of 0.05, all variables except the number of discounts (P =0.1101) passed t test, indicating that there is not enough evidence to support the correlation between the discount times and the sales of games under the current sample size (n = 24358). The regression equation's coefficient R^2 = 67.56%, indicating that the regression model can explain 67.56% of the variation in Steam sales. The remaining 32.44% of sales volume variation could not be explained by the model, which had a high

goodness of fit. In Regression model, reviews numbers, positive rate, release time, language support, average discounts strength and sales were positively correlated, consistent with previous assumptions, while the average price, discount times, the average discount duration is negative correlation with sales, discount time and average discounts duration is not consistent with the previous assumption. The model passed the F-test (p = 2.2*10^-16), and the variation of the corresponding variable explained by the selected variables was better than that explained by the random variables.

## 5. CONCLUSION

1. There is a strong correlation between the number of reviews and sales. The choice of number of reviews as an explanatory variable does not mean that high number of reviews leads to high sales, but rather that high-selling games tend to get more reviews.

2. Among the four variables reflecting game discounts (average price, average number of discounts, average discounts strength, and average duration of discounts), only the average discounts strength is significant, and there is a significant weak correlation between them. This shows that among different promotional strategies, stronger discounts are better than high-frequency, long-term ones.

3. Obviously, earlier released games have more sales than late released games.

4. More language support can significantly boost sales.

## REFERENCES

[1] Clement, J. (2021, June 8). _Number of steam mau 2020_. Statista. Retrieved October 21, 2021, from https://www.statista.com/statistics/733277/number-stream-dau-mau/.

[2] Valve. (2021, January 19). _Steamworks development - steam - 2020 year in Review - Steam News_. Welcome to Steam. Retrieved October 21, 2021, from https://store.steampowered.com/news/group/4145017/view/2961646623386540826.

[3] Mikolić, M. (2021). _About playtracker insight_. About PlayTracker Insight. Retrieved October 21, 2021, from https://playtracker.net/insight/about/.

[4] An empirical study of the impact of product factors on game-related consumption based on STEAM platform（2016) Retrieved October 21, 2021, from https://wap.cnki.net/touch/web/Conference/Article/JSGA201610001052.html.

[5] Cui, G. Lui, H.K. and Guo, X. (2012), “The Effect of Online Consumer Reviews on New Product Sales,” in: International Journal of Electronic Commerce, 17(1), 39-58.

[6] Jeffrey, C. (2021, July 20). Ubisoft titles may come back to steam if valve's handheld takes off, says CEO_. TechSpot. Retrieved October 23, 2021,from https://www.techspot.com/news/90497-ubisoft-titles-may-come-back-steam-if-valve.html.