Social Media Data Analysis Final Project

01/1358810 Kuon Ito Github repository

Can YouTube Increase Sales of Game Hardware?

- An Analysis Through Nowcasting -

---

1. Motivation

Apart from my personal aspiration to work as a marketer for Nintendo in the future, the motivation behind this research is twofold: marketing strategy and supply chain management. Understanding how YouTube impacts game hardware sales can provide valuable insights for marketing efforts, helping companies optimize their promotional strategies. By using time series analysis to predict demand, we can better understand the dynamics of game hardware sales, which is crucial for effective inventory management.

Accurate demand forecasting is essential, especially in situations where shortages can lead to issues such as the emergence of resale markets that exploit supply constraints. This was evident during the release of the PlayStation 5 in Japan, where groups took advantage of limited stock to sell the console at inflated prices. By applying forecasting from this perspective, the research aims to offer solutions that could mitigate such problems in the future.

I tested the following hypotheses:

1. The number of views on official channels’ videos is positively correlated with the sales of game hardware.
2. Forecast models incorporating YouTube information are better than the ARIMA model.

To test these hypotheses, I constructed the following models:

1. **Cross-Sectional Model**:
    - Multiple Linear Regression
2. **Time Series Models** for each hardware:
    1. ARIMA Model
    2. PCA Regression
    3. Ridge Model
    4. Lasso Model

The details of each model are explained in Chapter 4: Analysis.

---

2. Data Retrieval

Dependent Variable:

I examined the sales of three major video game consoles: Nintendo Switch, PlayStation, and Xbox. I used data from VGChartz.

Independent Variables of the Cross-Sectional Model:

The independent variables I considered were the number of days from the release date to June 30, 2024, and various metrics obtained from YouTube. Specifically, I collected the number of views and likes from the 50 most recent videos on each official channel, resulting in 150 observations. This approach allowed me to capture a comprehensive snapshot of the engagement each console's content receives.

Additionally, I included the total number of videos and the number of subscribers for each official channel. However, to avoid multicollinearity, I excluded the total number of views for each channel, as this could be linearly related to the views of individual videos. If the coefficient for the number of views is not significantly different from 0, it would indicate a correlation between views and sales.

Independent Variables of the Time Series Model:

For this project, I collected weekly data on view counts and subscriber changes for the official channels of various gaming consoles, as well as for 20 similar channels. The similar channels were defined as those in the "games" genre on Social Blade's "Similar Channels" list. If there were fewer than 20 such channels, I included channels from the "Featured Box" and their similar channels. This data collection strategy was one of the key innovations of this project.

Typically, YouTube only provides access to the data at fixed times. However, by consistently gathering data at the same time each week, I was able to create a time series dataset independently. Specifically, I ran a script every Saturday at 10 PM from June 1 to July 6 to capture this information.

To calculate weekly changes, I used the actual difference between the cumulative data from one week to the next. This approach ensured that the weekly fluctuations reflected real data, based on the differences in the cumulative figures obtained each week.

3. Data Processing

3.1 Cross-Sectional Analysis

In the "data processing - analysis" section of the cross-sectional folder on GitHub, I merged the channel information with the respective video information to create a single DataFrame called `video_df`, where each row represents a video. Since some variables were not integers, I converted them to numeric values. Additionally, some newly uploaded videos had 0 views. Given that this project involves nowcasting, I wanted to examine the immediate impact of videos, so I did not drop videos simply because they were newly uploaded. However, due to the significant variance in view counts, I applied a logarithmic transformation. During this process, any videos with 0 views would result in missing values, so I dropped those videos from the dataset.

According to the graph 3.1 (heatmap), there is a combination of variables with strong correlations between explanatory variables. Focusing on the relationship between video view counts and sales, I selected variables from YouTube that had a weak correlation with views but included as many relevant variables as possible. This was done to investigate the impact of YouTube on game console sales. Variables with a correlation coefficient of 0.7 or higher were dropped, leaving the following variables: views, likes, videos, and dates_passed. The "videos" variable represents the number of videos on the official channel, while "dates_passed" indicates the number of days since the game was released.

3.2 Time Series Analysis

In the "data processing - analysis" section of the time series folder on GitHub, I adjusted the data to units of 1,000 for subscriber numbers. This was done because it is unlikely that changes of just a few hundred subscribers in a week would significantly impact global game console sales. The reason for focusing on absolute values is that there were instances where video view counts were negative, likely due to certain videos being deleted during that week. These negative values were retained in the analysis. Additionally, since there was a significant disparity in view counts, I applied a logarithmic transformation.

4. Analysis

4.1 Cross-Sectional Analysis

When graphing video view counts and game console sales (graph 4.1), a positive correlation was observed. Will this result hold even when controlling for other variables?

Table 4.1 presents the results of regressing sales on views, likes, videos, and dates_passed. Since there are no zero values in any of the explanatory variables in the DataFrame, the constant term is not meaningful. However, the coefficients of the other variables are somewhat counterintuitive. The coefficient for video view counts is not significant at the 5% level, while the other variables are significant at the 1% level. An increase in the number of likes is associated with higher sales, which makes sense. Although one might expect the number of videos on the official channel to be positively correlated with game console sales, the results show a negative coefficient. Additionally, the results indicate that the more days have passed since the game console's release, the higher the sales tend to be.

4.2 Time Series Analysis

The following four models were used in the analysis:

- **ARIMA**: This model was used as a benchmark and did not incorporate any information obtained from YouTube.
- **PCA Regression**: Given the large number of variables (40), I anticipated that their variances could become excessive. Therefore, I performed dimensionality reduction using PCA. The optimal number of factors was determined based on the value that minimized the information criterion (IC). The extracted factors were then used for regression analysis.
- **Ridge**: I collected data on both view counts and subscriber fluctuations for each channel, which are likely to be correlated. To mitigate multicollinearity, I used Ridge regression.
- **Lasso**: As mentioned in the context of PCA, reducing the number of variables is beneficial due to the large dataset. Lasso was effective in further reducing variables and alleviating correlation issues, making it a suitable model for this analysis.

Five weeks of data were obtained, and the two most recent periods were used as test data.

The performance of the models was evaluated using the following metrics:

- RMSE (Root Mean Squared Error)
- MAE (Mean Absolute Error)
- MAPE (Mean Absolute Percentage Error)

For each game console, I created four models to predict sales. To address Hypothesis 2 and investigate the influence of YouTube, I focused more on the three evaluation metrics (RMSE, MAE, MAPE) rather than simply whether the predicted sales were increasing or decreasing. By examining how well the models fit the actual sales data, I aimed to assess the impact of YouTube on game console sales.

4.2.1 Switch

For the Nintendo Switch, comparing the ARIMA model, which does not incorporate YouTube data, with the other models, we can see that the Ridge and Lasso models have lower MAPE values. This suggests that using YouTube data results in a better fit to the sales data.

SWITCH	RMSE	MAE	MAPE
ARIMA	14288	12967	0.08
PCA	158556647	112119424	680
Ridge	0.0074	0.0057	3.5e-8
Lasso	11786	101234	0.064

4.2.2 PlayStation

For the PlayStation, the Ridge model has a lower MAPE compared to the ARIMA model, indicating that it provides a better fit to the data.

PlayStation	RMSE	MAE	MAPE
ARIMA	3596	3517	0.018
PCA	275080255	194514233	975
Ridge	0.024	0.023	1.1e-7
Lasso	9174	8974	0.044

4.2.3 Xbox

For the Xbox, similar to the Switch, both the Ridge and Lasso models have lower MAPE values compared to the ARIMA model. This suggests that these two models provide a better fit to the data.

Xbox	RMSE	MAE	MAPE
ARIMA	5918	5533	0.080
PCA	46998322	33234076	469
Ridge	0.039	0.037	5.38e-7
Lasso	2939	2363	0.034

5. Conclusion

5.1 Cross-Sectional Analysis

Since the variable for view counts was not significant, we cannot conclude that view counts have an impact on game console sales through YouTube. However, other information obtained from YouTube, such as the number of likes and the number of videos on the official channel, was significant. This suggests that the number of likes has a positive impact, while the number of videos on the official channel has a negative impact on game console sales.

5.2 Time Series Analysis

Only the MAPE values from section 4.2 are summarized here.

MAPE	Switch	PlayStation	Xbox
ARIMA	0.08	0.018	0.080
PCA	680	975	469
Ridge	3.5e-8	1.1e-7	5.38e-7
Lasso	0.064	0.044	0.034

The Ridge model demonstrates higher accuracy compared to the ARIMA model. Given that the Ridge regression model shows greater accuracy than the ARIMA model, which does not include YouTube data, it suggests that incorporating YouTube information enhances the explanatory power of the model for making predictions.

To summarize the results of the cross-sectional and time series analyses for testing two hypotheses:

In the cross-sectional analysis, no correlation was found between YouTube video views and game console sales, but likes on videos can increase the sales of game hardware. However, by conducting a time series analysis, I aimed to capture the immediate effects of videos by examining the trends in game console sales—something that the cross-sectional analysis could not achieve. As a result, the Ridge model demonstrated higher accuracy than ARIMA model. The strong fit of the Ridge model suggests that YouTube information may indeed be linked to game console sales.

6. Critique

The cross-sectional regression analysis yielded results that were contrary to intuition. For example, it suggested that an increase in the number of videos on the official channel would lead to a decrease in game console sales. This raises the possibility of omitted variable bias. There may be other variables that explain game console sales which I have overlooked, and this omission could be causing bias in the coefficients, leading to incorrect signs.

The Ridge model showed a better fit compared to the other models, but this may simply indicate the inherent strength of the Ridge model itself. However, even if YouTube data were added as exogenous variables to the ARIMA model, multicollinearity could lead to an increase in the standard errors of the variables.

Examining a longer time period might have provided a clearer picture of how well the models actually fit the data.

Among the selected channels, there were some with hundreds of thousands of subscribers. When selecting channels, I chose those that seemed likely to contribute to game sales based on their subscriber count. However, if I had selected channels with even greater influence, such as those with millions of views or subscribers, the results might have been different.