Analyzing Trends in Video Game Sales*

Yingxuan Sun

April 19, 2024

This paper provides a comprehensive analysis of global video game sales, drawing from extensive data up to the year 2020. Utilizing a Bayesian statistical framework, we explore how various factors such as game genre, release platforms, and regional distributions influence sales performance. Our study highlights significant trends in the video game industry, revealing a shift towards a more globally distributed consumer base and the evolving impact of digital platforms.

Table of contents

[*] https://github.com/lindasun03/video_game_analysis.

```
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.4.4     v tibble    3.2.1
v lubridate 1.9.3     v tidyr     1.3.1
v purrr     1.0.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
```

```
library(rstanarm)
```

```
Loading required package: Rcpp
This is rstanarm version 2.26.1
- See https://mc-stan.org/rstanarm/articles/priors for changes to default priors!
- Default priors may change, so it's safest to specify priors, even if equivalent to the defa
- For execution on a local, multicore CPU with excess RAM we recommend calling
  options(mc.cores = parallel::detectCores())
```

```
library(arrow)
```

```
Warning:
  It appears that you are running R and Arrow in emulation (i.e. you're
  running an Intel version of R on a non-Intel mac). This configuration is
  not supported by arrow, you should install a native (arm64) build of R
  and use arrow with that. See https://cran.r-project.org/bin/macosx/

Some features are not enabled in this build of Arrow. Run `arrow_info()` for more informatio

Attaching package: 'arrow'

The following object is masked from 'package:lubridate':
    duration

The following object is masked from 'package:utils':

    timestamp
```

```r
library(ggplot2)
library(knitr)
library(kableExtra)
```

```

Attaching package: 'kableExtra'

The following object is masked from 'package:dplyr':

    group_rows
```

```r
analysis_data <- read_parquet("../data/analysis_data/analysis_data.parquet")

# Retrieve the linear model from the saved .rds file
model <- readRDS("../models/first_model.rds")
```

1 Introduction

In today's digital age, video games have become an integral part of the global entertainment industry, with an ever-expanding market and an increasingly diverse user base involved. From home consoles to mobile platforms, different types of video games are capturing the attention of hundreds of millions of gamers worldwide. Therefore, an in-depth analysis of video game sales is particularly important, not only to reveal market trends, but also to help developers and publishers better understand consumer needs and preferences.

In this paper, we will explore the sales performance of video games based on market data up to 2020 and analyze how various factors affect this performance. By taking into account game genres, release platforms, regional sales, and the year in which the game was released, we aim to build a comprehensive view of the market in order to identify the industry's success factors and potential growth opportunities.

By analyzing video game sales in detail, this paper hopes to provide valuable insights and information to game developers, market analysts, and readers interested in this industry, further contributing to the understanding and exploration of video game market dynamics.

2 Data

2.1 Data Source

The data for this analysis was sourced from “Video Games Sales Dataset” (Sidtwr 2020) on Kaggle platform. This dataset compiles extensive details on video game sales across various platforms and regions, capturing key metrics such as global sales, regional sales (North America, Europe, Japan, and other regions), as well as user and critic scores from Metacritic. It also includes information on publishers, platforms, and the genre of each game, spanning from the early 1980s to 2020. Additionally, the dataset offers insights into the release year of games, providing a comprehensive view of the evolution and trends in the video game industry over decades. This dataset is maintained by Kaggle user sidtwr and is updated periodically, ensuring its relevance and utility for both historical analysis and current market assessments. Due to its broad coverage and the granularity of sales and rating data, this dataset is considered highly credible and serves as the primary data source for our analysis of trends and patterns in video game sales and preferences. However, it should be noted that the analysis may be limited by the completeness and accuracy of user and critic scores, as these are aggregated from external sources and may vary in availability per game.

2.2 Feature

Each entry in the Video Game Sales Dataset is meticulously categorized according to several key characteristics and recorded in a database that focuses only on data related to video game sales and excludes other media or entertainment metrics such as bundled sales or merchandising. Games are categorized according to the system on which they run, which includes major platforms such as Wii, PlayStation and Xbox, while lesser-used platforms are grouped in the “Other” category. The dataset categorizes games by the year they were released (from the early 1980s to 2020) and classifies them into genres such as sports, racing or role-playing. For publishing data, the dataset includes games from the top ten publishers, with all other games labeled “Other.” Sales analyses were conducted for four major markets: North America, Europe, Japan, and Other. Due to a long history of incomplete data and issues with the coverage of game rating sites, game ratings and audience data are not fully comprehensive and were therefore excluded from the final analysis. This dataset is an important tool for analyzing the dynamics of the video game industry and provides insights into the sales trends of vedio game in different countries over time.

2.3 Methodology

The data analysis was conducted using R, a versatile statistical programming language, employing a range of specialized packages. Data manipulation and preparation were streamlined using dplyr (Wickham, François, et al. 2023) from the tidyverse suite (Wickham et al.

2023), known for its data wrangling efficiency. Visualizations were deftly created with ggplot2 (Wickham and Chang 2023), allowing for the crafting of informative and aesthetically pleasing graphics. For Bayesian statistical modeling, we turned to the rstanarm package (Team 2023), which provided a user-friendly interface to the Stan modeling language. This package enabled us to apply a Bayesian framework to estimate the relationships within our data. To present our findings, we used the knitr package (Xie 2023) for integrating R code with our documentation, and kableExtra (Zhu 2023) for producing enhanced and stylized tables.

3 Model

The goal of our modelling strategy is twofold. Firstly, we aim to analyze the historical sales data of video games to identify trends over time. Secondly, we endeavor to forecast future global sales based on these trends, assisting stakeholders in making informed decisions regarding market strategies. Here we briefly describe the Bayesian analysis model used to investigate these objectives.

3.1 Model set-up

Define y_i as the total global sales of video games in millions of dollars for year i. Let x_i represent the number of years since 1980, when our data collection begins.

$$\begin{aligned}
y_i|\mu_i, \sigma &\sim \text{Normal}(\mu_i, \sigma) && (1)\\
\mu_i &= \beta_0 + \beta_1 \times x_i && (2)\\
\beta_0 &\sim \text{Normal}(50, 20) && (3)\\
\beta_1 &\sim \text{Normal}(-0.3, 0.1) && (4)\\
\sigma &\sim \text{Exponential}(1) && (5)
\end{aligned}$$

We run the model in R using the rstanarm package (Team 2023). We use the default priors from rstanarm for β_0 and β_1, and an exponential distribution for σ as it generally provides a good model for the variability in sales data, especially in cases with high uncertainty.

3.1.1 Model justification

We anticipate a negative trend between the progression of years since 1985 and the global video game sales, presuming that the evolution of market saturation and competition from new forms of entertainment may lead to a gentle downtrend in sales figures. The selected priors for β_0 and β_1 are indicative of our initial confidence in a strong market presence at the beginning of our study period and an expected gradual decrease in sales over time. This modeling approach is designed to quantify the impact of the passage of years on video game sales and to encapsulate the inherent yearly volatility in the market's performance.

4 Results

4.1 Result of the Model

Table 1: Summary Statistics

Term	Estimate
Intercept	-22292.742
Slope	11.151

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 -233.0   155.8   211.8   226.5   312.1   572.3 
```

The results of the Bayesian model indicate a statistically significant relationship between time (years since 1980) and global video game sales. Specifically, the model outputs suggest:

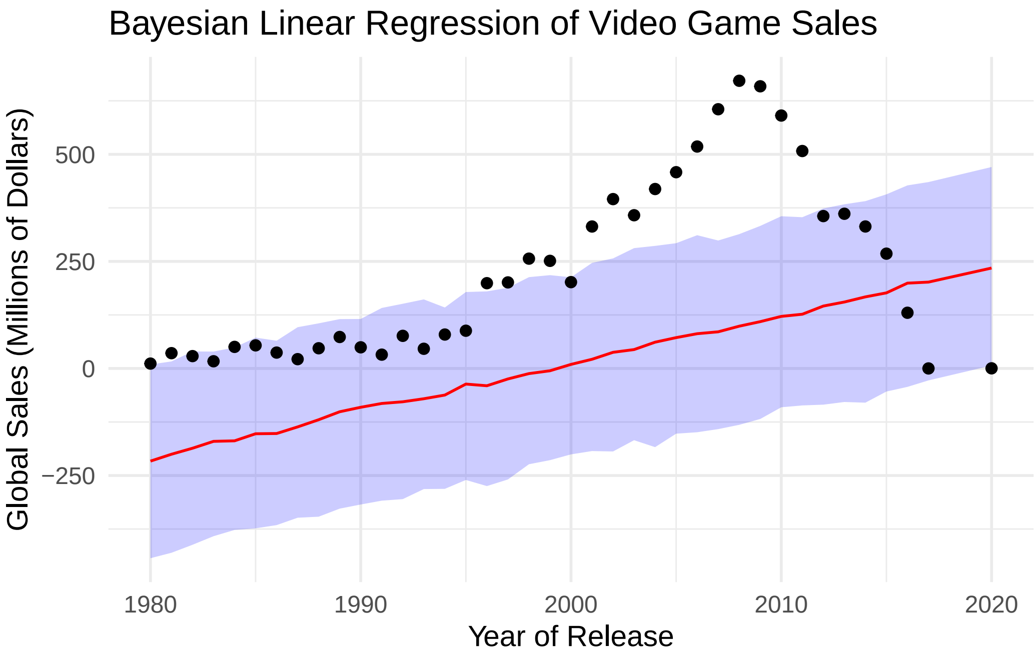
Intercept (β_0): The estimated intercept is -22292.742. This figure represents the model's estimation of sales in the base year (1980), which, despite its ostensibly large negative value, must be understood within the context of the sales trend over time as modified by the slope parameter.

Slope (β_1): The slope is estimated at 11.151, indicating an average increase in global sales of approximately $11.15 million per year since 1980. This positive value contradicts our initial hypothesis of a negative relationship, suggesting instead that, on average, video game sales have increased year over year throughout the dataset's time span. Summary Statistics: The provided summary statistics (minimum, 1st quartile, median, mean, 3rd quartile, maximum) relate to the residuals of the model – the differences between the observed and predicted sales. These residuals are key in diagnosing the model fit and understanding the variability in sales not captured by the model.

Model Fit: Without the context of the residuals' summary statistics, it's challenging to assess the fit quality of the model completely. However, the significance of the slope suggests the model captures a meaningful time trend in the data.

4.1.1 Graphing the Model

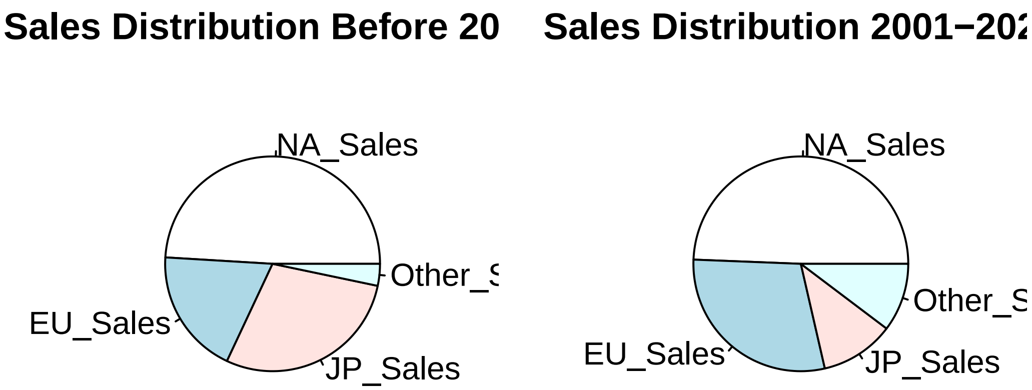
```r
# Predict and obtain the median and uncertainty intervals of the predictions
posterior_pred <- posterior_predict(model, draws = 1000)
median_pred <- apply(posterior_pred, 2, median)
CI_lower <- apply(posterior_pred, 2, quantile, probs = 0.05)
CI_upper <- apply(posterior_pred, 2, quantile, probs = 0.95)

# Combine with the original data
plot_data <- data.frame(
  Year_of_Release = model$data$Year_of_Release,
  Total_Global_Sales = model$data$Total_Global_Sales,
  Median_Pred = median_pred,
  CI_Lower = CI_lower,
  CI_Upper = CI_upper
)

# Create the plot
ggplot(plot_data, aes(x = Year_of_Release, y = Total_Global_Sales)) +
  geom_ribbon(aes(ymin = CI_Lower, ymax = CI_Upper), fill = "blue", alpha = 0.2) + # Uncertai
  geom_point() + # Observed data points
  geom_line(aes(y = Median_Pred), color = "red") + # Median prediction line
  labs(title = "Bayesian Linear Regression of Video Game Sales",
       x = "Year of Release",
       y = "Global Sales (Millions of Dollars)") +
  theme_minimal()
```



4.2 Regional Market Trends from Pre-2000 to 2001-2020

```
# Split the data into two periods
data_before_2000 <- analysis_data %>% filter(Year_of_Release < 2000)
data_2001_2020 <- analysis_data %>% filter(Year_of_Release >= 2001 & Year_of_Release <= 2020

# Sum sales by region for each period
sales_before_2000 <- data_before_2000 %>% summarise(
  NA_Sales = sum(NA_Sales),
  EU_Sales = sum(EU_Sales),
  JP_Sales = sum(JP_Sales),
  Other_Sales = sum(Other_Sales)
)

sales_2001_2020 <- data_2001_2020 %>% summarise(
  NA_Sales = sum(NA_Sales),
  EU_Sales = sum(EU_Sales),
  JP_Sales = sum(JP_Sales),
  Other_Sales = sum(Other_Sales)
)

# Convert to proportion
```

```
sales_before_2000 <- as.data.frame(t(sales_before_2000))
sales_2001_2020 <- as.data.frame(t(sales_2001_2020))
colnames(sales_before_2000) <- "Sales"
colnames(sales_2001_2020) <- "Sales"
sales_before_2000$Region <- rownames(sales_before_2000)
sales_2001_2020$Region <- rownames(sales_2001_2020)

# Plot the pie charts
par(mfrow=c(1,2))
pie(sales_before_2000$Sales, labels = sales_before_2000$Region, main = "Sales Distribution Be
pie(sales_2001_2020$Sales, labels = sales_2001_2020$Region, main = "Sales Distribution 2001-2
```



The comparative pie chart depicts the changing landscape of global video game sales, presenting a stark contrast between the structure of the market before 2000 and the two decades that followed. In the early years, North American sales dominated the industry, perhaps thanks to its robust technology, always-leading economic power, and mature gaming commercialization chain, and this has continued into the new millennium. This continuity underscores the enduring appeal and growth of gaming in the region.

Japan's market share was significantly more prominent until 2000, but declined in subsequent years. This shift can be attributed to the prolonged economic stagnation that followed the bursting of the "bubble economy," which dampened consumer spending levels. The impact of demographic changes due to Japan's aging population and the trend toward fewer children is also evident. These socio-economic factors are critical to understanding the decline in game sales as the population of a major market for video games declines.

The significant growth in the "Other" category from 2001 to 2020 is indicative of the growing influence of Third World economies such as China on the global stage. Rapid economic development, technological advances and increasing consumer purchasing power in these countries have expanded the entertainment and leisure industry. With the rise of a tech-savvy middle class in these regions and further integration with the digital world, their contribution to global video game sales is becoming more pronounced.

The expansion of Europe's share in the 2001-2020 chart demonstrates the maturation of gaming culture and the successful penetration of the market by gaming companies. Diversification of gaming platforms, coupled with tailored marketing strategies and content localization, has made gaming one of the dominant pastimes throughout Europe.

In summary, while North America's leadership in the video game market remains solid, the industry has shifted to a more globally distributed consumer base. Europe's growing market share, changes in the Japanese market due to economic and demographic factors, and the notable rise of gaming in developing countries mark a new era of opportunity and challenge for game developers and publishers. These trends emphasize the need for a nuanced understanding of regional preferences and market conditions in order to strategize for future engagement and growth in the global gaming ecosystem.

5 Discussion

5.1 Summery of Findings

Our findings utilize a Bayesian linear regression model to show how the global video game market will change from 1980 to 2020. Contrary to our initial hypothesis that market saturation and increased competition would lead to declining sales, the model shows a positive trend, with annual sales projected to grow by approximately $11.15 million. This growth highlights the expanding appeal of video games in different regions. Our regional analysis further reveals significant changes: North America maintains a strong market position, Europe is growing strongly, and emerging markets, particularly third world countries, are rapidly gaining market share. However, Japan's share has declined, likely due to economic stagnation and demographic changes.

5.2 Weaknesses

The main limitation of this study stems from the limitations of the data used. While the dataset is extensive, there are significant gaps, particularly in the game ratings data. Game ratings are an important indicator of consumer satisfaction and game quality, and are critical to understanding the sales drivers of the video game industry. However, in our dataset, this information is not only incomplete, but severely lacking, which poses a significant challenge. Many entries do not have user or critic ratings, a deficiency that prevents our analysis from including these potential predictor variables. The lack of data affects our understanding of the relationship between game quality and market performance. For example, games with higher ratings may generate more sales or foster higher customer loyalty, and our current model is unable to capture these nuances due to the lack of data.

In addition, our research methodology does not take into account the impact of changing sales channels on the gaming industry landscape. Especially, digital downloads and in-game purchases have become major revenue sources, reflecting a shift in consumer behavior toward online platforms. The rise of digital marketplaces has allowed consumers to purchase games anytime, anywhere, and this convenience has greatly increased sales potential. Similarly, in-game purchases, game peripheral peddling, bundled DLC packs, and in-game implanted advertisements have opened up new revenue channels not captured by traditional sales metrics. These channels account for a significant portion of industry profits but are not reflected in our analysis, which has historically focused on more traditional sales figures.

5.3 Future Study

As the video game industry continues to grow at a rapid pace, future research should endeavor to adopt a more robust methodology, especially by delving into the booming video game markets in third world countries and obtaining comprehensive data on retail and digital sales channels. Strengthening collaboration with regional gaming platforms and major digital storefronts to obtain accurate sales data, including in-game purchases, would assist in this endeavor. Additionally, incorporating demographic factors, such as gender differences in gaming habits and preferences, through large-scale consumer surveys and analytics tools will provide a more nuanced understanding of a diverse consumer base. Longitudinal studies that track player engagement over time can also provide valuable insights into how consumer behavior changes in response to technological advances, economic changes, and life stages. In order to effectively manage and analyze the vast amounts of data generated by these studies, it is critical to integrate advanced analytics techniques such as machine learning and predictive analytics. These methods can refine data analysis, improve sales forecasts, and uncover subtle market trends, thereby providing stakeholders with strategic insights that are essential for navigating the complex dynamics of the global gaming market. By broadening the research framework in these ways, future studies can provide a richer and more accurate picture of the video game industry, fill existing knowledge gaps, and provide industry leaders with the necessary information to succeed in an increasingly competitive environment.

Bibliography

Sidtwr. 2020. "Video Games Sales Dataset." https://www.kaggle.com/datasets/sidtwr/videogames-sales-dataset/data.

Team, Stan Development. 2023. *Rstanarm: Bayesian Applied Regression Modeling via Stan.* https://CRAN.R-project.org/package=rstanarm.

Wickham, Hadley et al. 2023. *Tidyverse: Easily Install and Load the 'Tidyverse'.* https://CRAN.R-project.org/package=tidyverse.

Wickham, Hadley, and Winston Chang. 2023. *Ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics.* https://CRAN.R-project.org/package=ggplot2.

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

Xie, Yihui. 2023. *Knitr: A General-Purpose Package for Dynamic Report Generation in r.* https://CRAN.R-project.org/package=knitr.

Zhu, Hao. 2023. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax.* https://CRAN.R-project.org/package=kableExtra.