Foundations of Statistics and Econometrics

Context and Dataset:

The panel dataset contains various metrics for videogame consoles at a monthly level (from 1995 to 2008). In this context, videogame consoles (such as Xbox owned by Microsoft or PlayStation owned Sony) compete in two ways. First, a given console competes with other consoles to attract customers (i.e., gamers) to adopt (i.e., buy) the console and play the available games on it. Second, it competes with other consoles to incentivise third-party game developers to release their games on that console. On the one hand, the more gamers adopt the console (i.e., the hardware); the game developers are more incentivised to publish and sell their games on that console. On the other hand, the more variety of game titles are available on the console; the more gamers are willing to adopt that console.

- *Please note, as per the confidentiality of the data, the consoles' names have been changed, and some numbers have been amended; therefore, this dataset is not suitable for research purposes.*

Below is the description of the variables in the dataset.

- ✓ *log_unit_sale*: natural logarithm of the number of units sold of the console in the given period (*unit_sale* is the non-logged version).
- ✓ *log-games*: natural logarithm of the number of *new* games published on the console (*games* is the non-logged version)
- ✓ *age*: age of the console (in months) since its launch.
- ✓ *price*: the price of the console (i.e., the hardware) in USD.
- ✓ *active_n*: the total number of active consoles in the given period.
- ✓ *ram*: RAM capacity of the console in megabytes.
- ✓ *mhz*: The CPU speed of the console in megahertz.
- ✓ *tech*: the console's technology generation (for example, https://bit.ly/3036N57)
- ✓ *leader*: a variable that indicates if the console's market share is above the median, compared to other consoles, in the given period—i.e., if the console is among the leaders of the market or not.
- ✓ *console*: the console's name (as said above, they are pseudonyms)
- ✓ *period*: the time-period identifier (at a monthly level) of the panel data

Note: The log transformation applied for the number of new games (*log_game*) is Ln(x+1), rather than Ln(x), to avoid losing observations with the value of zero; hence, if the number of new games is zero, the log-transformed

version will be zero as well—Ln(0+1)=0. For simplicity, you can interpret the effect size as Ln(x). A similar note applies to *log_unit_sale*.

Other variables (*calendar year, calendar month)* are also included. Overall, the objective is to estimate the effect of the number of new games, the console's age, and the number of active consoles in the market on the console's unit sale (while controlling for some relevant factors, as explained below).

Content and Structure:

Introduction

- Provide a brief explanation for the methodology, such as data, the definition of dependent, independent, and control variables, the objective of the analyses, and the baseline model (see the description of the baseline model below).

Descriptive Analysis

- Provide a two-way table for summary statistics of the variables for the entire sample and different technology generations subsamples. Provide the correlation matrix of the variables (for the entire sample). Beefily discuss the results.
- Apply a statistical test and evaluate if there is any significant difference (at 0.05 significance level) across the years regarding the unit sale (logged). You may use a relevant graphical illustration to enhance your discussion.

Exploratory Analysis

- Inspect the data graphically, such as visual summary statistics across technology generations or years, checking the distribution/skewness of *main* variables (i.e., dependent and independent variable), pre-checking the possibility of outliers, pre-checking the relationship between the dependent and independent variables, the longitudinal trend of the main variables, etc. The details and types of graphs are your decision—the objective is to provide a concise yet informative inspection of the data before running the regression. You may select a few graphs from the list mentioned above (or other graphs), which efficiently describe various aspects of the data.

- Show the trend for console's market share across periods; each console's trend with a line; all in one graph. The console's market share at period *t* is defined as the console's unit sale at period *t* divided by the sum of all console's unit sale at period *t*.

Main Regression Analysis:

- Conduct an OLS regression to estimate the effect of the number of new games, console's age, and the number of active consoles in the market on the console's unit sale while controlling for console's price, console's technology generation, console's RAM capacity, console's CPU speed, and time-period. This will be the ***baseline*** model. *The natural logarithm version of unit sale, number of new games, price, RAM capacity, and CPU speed should be used in all models*. *Other variables should be used as not-logged.*
- Carefully interpret and discuss the results (e.g., R-squared, the statistical significance of coefficients and the effect size of independent variables).
    - Note: as we do not need to interpret periods' effects, you do not need to report the coefficients of the periods in your tables. Still, you should explicitly indicate whether they are included in the models (see the Sample Report for an example). All other coefficients should be reported in the tables.
- Looking at the baseline model's results, evaluate whether there is any differential effect of technology generation on the unit sale. You may use a margins plot to enhance your discussion.
- Modify the baseline model to evaluate the differential effect of the number of new games for leader vs non-leader consoles. Based on the results, discuss the statistical significance and effect size of the difference. You may use a relevant graphical illustration to enhance your discussion.

Diagnostics and Robustness Analysis:

- Apply diagnostic analyses on the baseline model to check the potential heteroskedasticity and apply an appropriate remedy if needed. Briefly compare the new results with the original results of the baseline.
- Investigate the possibility of a quadratic effect of the number of new games (logged) on the unit sale (logged) and clearly discuss the result. You may use graphical illustration to enhance your discussion.

- Run the baseline model with console fixed effects. Briefly compare the new results with the original results of the baseline. Why are some variables dropped from the model with fixed effects? Explain how the fixed effect model can mitigate the endogeneity problem in your baseline model.
- Even after implementing the fixed effect model, what other endogeneity or omitted variable bias (particularly regarding the number of new games and console's price) may exist in the model? Shortly explain.

Appendix

- Copy the programming codes in the appendix in Word format. Do not copy the codes as a screenshot. Alternatively, you can upload the Stata do-file along with your report on SurreyLearn as a separate file.

Format:

- The project file should be in Microsoft word format in Times New Roman 12-point font double spaced. The word count of the project should be no more than 3500 words.
- The word count includes everything from the first word of the introduction to the last word of the conclusion. The word count does not include tables, figures or images, and appendices. It does not include abstract, table of contents, abbreviation pages, or references (though these are not mandatory in this project). You should report the word count, your name and student number at the beginning of your project. According to the university policy, exceeding the word count limit is subject to a 10-point penalty.