Project Description
Foundations of Statistics and Econometrics

Dataset:

The panel dataset contains some variables for app developers in the Google Play app store. The data has been captured for eight periods during the two years of 2020 and 2021. The app store is an intensively competitive market where app developers release their apps to gain users' attention and engagement. Many of the apps, even after being installed by users, fail to attract users' engagement and eventually will be either uninstalled or abandoned from further usage. Therefore, the number of *active* users is a critical success factor in this market. The objective of this project is to explore some of the determinants of the total number of active users of app developers (as a measure of their success), as explained further.

Below is the description of the variables in the dataset.

- ✓ *apps*: Total number of apps released to the app store by a given app developer until a given period.
- ✓ *free_percent:* Percentage of free apps to all apps released by a given app developer until a given period. For example, '*free* equals one' means all developer's apps are *free* apps (i.e., no *paid* apps).

    - *Free* apps are those that can be downloaded and installed free of charge from the app store. *Paid* apps are those that users should pay the app price before being able to download and install them.

- ✓ *users*: Total number of active users for all apps released by a given app developer until a given period.
- ✓ *installs*: Total number of installs by app users across all apps released by a given app developer until a given period.
- ✓ *log_price*: Natural logarithm of the average price for all apps released by a given app developer until a given period. App price is the price that users are required to pay before downloading the *paid* apps.
- ✓ *hhi_category*: Herfindahl–Hirschman Index (HHI) for measuring the diversity of app categories across which a given app developer released its apps until a given period. It is a ratio between zero and one and simply means that the app developer is focused on only few app categories or diversified across many app categories. For example, if a developer publishes only Entertainment apps, HHI will be one, but if a developer has a portfolio of apps across many categories such as Games, Business, Finance, Entertainment, Music & Audio, etc. HHI will be very small. In other words, a small HHI (closer to zero) implies that the app developer is a *generalist*, while a large HHI (closer to one) implies that the app developer is a *specialist*.
- ✓ *size*: a categorical variable which indicates if the app developer is a small, medium, or large firm.
- ✓ *dev_id* is the unique identifier of the app developer.
- ✓ *period*: the time-period identifier of the panel data. Consider it as a *categorical* variable.

Note: The log transformation applied for the price variable (*log_price*) is Ln(x+1), rather than Ln(x), to avoid losing observations with price=0; hence, if the price is zero, the log-transformed version will be zero as well—Ln(0+1)=0. For simplicity, you can interpret the effect size (if needed) as Ln(x).

Note: If during your analysis you face this error: "*matsize too small*", which may or may not happen depending on your working memory, run the below code and then continue your analysis:

- ○ *set matsize 1000*

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 (as explained in the Main Regression Analysis section).

- The total number of installs, the total number of active users, and the average price should be used in the natural-log-transformed version in all models. Other variables should be used as not-logged.
    - *Hint:* In the dataset, the price variable is already logged, but you need to generate the natural log version of the two abovementioned variables for your analysis.

Descriptive Analysis

- Provide a two-way table showing the summary statistics of the variables for subsamples of small, medium, and large developers, as well as the full sample. Briefly discuss the results.
- Apply an appropriate test to evaluate if there is any statistically significant difference (at 0.05 significance level) between small, medium, and large developers regarding the total number of active users (logged). Briefly discuss the results.
- Provide the correlation matrix of the variables. Briefly discuss the results.

Exploratory Analysis

- Inspect the data graphically, such as visual summary statistics, check the distribution/skewness of variables, pre-check the possibility of outliers, and pre-check the relationship between the dependent and independent variables, the longitudinal trend of 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 pick up a few of the above-mentioned list of potential graphs (or other graphs), which describe various aspects of the data efficiently. *Hint: more than six graphs would be too much!*

Main Regression Analysis:

- Conduct an OLS regression to estimate the effect of the total number of installs (logged), the total number of apps, and category HHI on the total number of active users (logged), while controlling the average app price (logged) and time period. This will be the ***baseline*** model. Carefully interpret and discuss the results (e.g., R-squared, the statistical significance of coefficients, and the effect size).
- Briefly justify the positive or negative effects of the regressors conceptually. Particularly, based on the results, does it seem to be better to be a specialist or a generalist app developer to gain user engagement?
- Looking at the *free_percent* variable (i.e., the percentage of free apps to total apps), you can see in this data that some of the app developers release only free apps, and some release both free and paid apps. Generate a categorical variable which distinguishes these two types of developers—"only free" or "free & paid". Modify the baseline model to estimate the differential effect of *the total number of apps* on the total number of active users (logged) *for*

"only free" *vs* *"free & paid"* developers. Based on the results, discuss the statistical significance and effect size of the difference. Run a margins plot and discuss how this graph supports the regression results. Can you explain what this means conceptually?

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 total number of apps on the total number of active users (logged) and discuss the result. You can use graphical illustrations to enhance your discussion.
- Run the baseline model with app developers fixed effects with robust standard errors. Briefly compare the new results with the original results of the baseline model. Given the context of the data and variables, explain how the fixed effect model can mitigate the endogeneity problem in your baseline model.