Project Description

Foundations of Statistics and Econometrics

Dataset:

The panel dataset contains various metrics for a sample of cryptocurrencies (digital coins) in 2017 and 2018 at a monthly level, extracted from the CoinGecko website (https://www.coingecko.com/en).

Below is the description of the variables in the dataset.

- ✓ *marketcap*: the market capitalization (hereafter, market cap) of the coin in US dollars. The logged version of the variable is also provided (*log_marketcap*).
- ✓ *log_price*: natural logarithm of the price of the coin in US dollars.
- ✓ *log_twitter*: natural logarithm of the number of followers of the coin's Twitter account.
- ✓ *log_facebook*: natural logarithm of the number of likes the coin's Facebook account received.
- ✓ *log_star*: natural logarithm of the number of stars the coin received from developers in GitHub, which shows how much the technical aspects of the coin are appealing to the programmers and developers' community. GitHub is a website wherein software developers contribute to open projects.

- ✓ *log_subsc*: natural logarithm of the number of developers who subscribed to the coin's project account in GitHub.
- ✓ *alexa*: the ranking of the coin webpage based on Alexa website (https://www.alexa.com/topsites). Alexa is a website that ranks the top sites based on visibility and users traffic.
- ✓ *log_bing*: natural logarithm of the number of matched results in Bing search engine (https://www.bing.com) for the coin, which shows to what extent people are curious about the coin.
- ✓ *symbol*: the unique identifier of the coin.
- ✓ *period*: the time-period identifier (at a monthly level) of the panel data.

Note: The log transformation applied for some variable (such as *log_twitter*) is Ln(x+1), rather than Ln(x), to avoid losing observations with the non-logged version equals to zero; if the value is zero, the log-transformed version will be zero as well in this method—Ln(0+1)=0. For simplicity, you can interpret the effect size (if needed) as Ln(x). A similar note applies to *log_bing*, *log_facebook*, *log_star*, and *log_subsc*.

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).

Descriptive Analysis

- ➢ Provide a two-way table for summary statistics of the variables for the whole sample, 2017 and 2018 subsamples. Provide the correlation matrix of the variables. Beefily discuss the results.
- ➢ Apply a test to evaluate if there is any significant difference (at 0.05 significance level) between the years 2017 and 2018 regarding the market cap (logged).

Exploratory Analysis

- Inspect the data graphically, such as visual summary statistics, check the distribution/skewness of main variables (i.e., dependent and independent variable), pre-check the possibility of outliers, and pre-check the relationship between the dependent and independent variables, the longitudinal trend of the dependent variable, 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.
- Show the trend for Bitcoin market share across periods. The Bitcoin market share at period *t* is defined as the Bitcoin's market cap at period *t* divided by the sum of all coins' market cap at period *t*. Bitcoin's symbol in the dataset is ***btc***.

Main Regression Analysis:

- Conduct an OLS regression to estimate the effect of *the number of GitHub subscribers (logged)*, *the number of GitHub stars (logged)*, *Alexa ranking*, and *the number of matched results in Bing (logged)* on *the coin market cap (logged)*, while controlling for the coin price (logged), the number of Twitter followers (logged), the number of Facebook likes (logged), and monthly time-period. This will be the *baseline* model. Carefully interpret and discuss the results (e.g., R-squared, the statistical significance of coefficients, the effect size).
- Modify the baseline model to evaluate the differential effect of *the number of matched results in Bing* (logged) for Bitcoin vs non-Bitcoin coins. Based on the results, discuss the statistical significance and effect size of the difference. Run a margins plot and discuss how that graph supports the regression results for the differential effect of the number of GitHub stars (logged) for Bitcoin vs non-Bitcoin coins.

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.
- Run the baseline model with coins fixed effects with robust standard errors. Briefly compare the new results with the original results of the baseline model. Explain how the fixed effect model can mitigate the endogeneity problem in your baseline model.

Can you explain why the effect of so many of the variables becomes statistically non-significant in the fixed effect model?