

Final Project Description- *Resit*

Foundations of Statistics and Econometrics, MANM467

Semester 1, 2020-21

Essential Information:

- Students need to prepare the final project individually and submit it before **17th August 2021, 4:00 pm** via Surrey Learn.
- The project consists of applying econometric analyses based on a real-world dataset as described below, using statistical software (Stata). The expected level of econometric analyses is based on the lectures and lab sessions.
- Stata software is available on All FASS labs and ALL central labs computers. The network version is also accessible via Surrey virtual desktop.
- All documents and files related to the final project are available from SurreyLearn: Course Materials → Assessment Information → Resit Assessment.

Dataset:

The panel dataset contains various metrics for a sample of cryptocurrencies (digital coins) for 21 months (in 2018 and 2019) 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.
- ✓ 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*

Other variables (calendar year, calendar month, and coin name) are also included.

Overall, the objective is to estimate the effect of the number of Twitter followers (logged), the number of GitHub subscribers (logged), Alexa ranking, and the number of matched results in Bing on the coin market cap (logged), 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.

Descriptive Analysis

- Provide a two-way table for summary statistics of the variables for the whole sample, 2018 and 2019 subsamples. Provide the correlation matrix of the variables. Briefly discuss the results.

- Apply a t-test and evaluate if there is any significant difference (at 0.05 significance level) between the years 2018 and 2019 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 .

Main Regression Analysis:

- Conduct an OLS regression to estimate the effect of *the number of Twitter followers (logged)*, *the number of GitHub subscribers (logged)*, *Alexa ranking*, and *the number of matched results in Bing* on *the coin market cap (logged)*, while controlling for the coin price (logged), the number of stars (logged), the number of Facebook likes (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, the effect size). Conceptually, why is the effect of Alexa ranking on the market cap negative? Is this an impactful variable on the coin market cap?
- Modify the baseline model to evaluate the differential effect of the number of GitHub subscribers (logged) for Bitcoin vs non-Bitcoin coins. Based on the results, discuss the statistical significance and effect size of the difference. You may use 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 Twitter followers (logged) on the market cap (logged), and clearly discuss the result. You may use graphical illustration to enhance your discussion.
- 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?

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 length of the project should be no more than 3500 words (excluding tables, graphs, and appendix). You should report the word count along with 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.

Guideline and Tips:

- Apply the analyses required as explained *orderly*, section by section (from Introduction to Diagnostics and Robustness Analysis).
- The report —the writing, explanations, tables, and graphs— should be **clear and informative as a self-sufficient and stand-alone document** for readers who do not have access to this Final Project Description.
- In the **introduction**, concisely explain the aim of the empirical report, sample and data, definition of all final variables which are incorporated in your regression models. Some of this information (such as sample and variables definition) has been already provided, but you need to give a concise summary of them in your report.
- All **tables and graphs** should be numbered and titled (and with captions if an additional explanation is required) and should be referred to in the report accordingly. The label of the variables in tables and graphs should be informative.

- Graphs should be visually clear (axis title, colour, legend, axis scale, etc.). You may use image format for your graphs. Do not populate the report with lots of graphs; be selective and use the most informative ones for your purpose.
- Tables should be exported from the statistical software to a proper and readable Word format. You can report various models in one or two tables (each model in one column). Yet, you need to clearly number your models and refer to them in the discussions accordingly. Moreover, you don't need to report the time-periods coefficients in your tables. Still, you should clearly indicate whether they are included in the models (see the Sample Report). All other coefficients should be reported in the tables.
- In the regression tables, standard errors should be reported below each coefficient (in the parenthesis), the significance level of the coefficient should be determined by asterisks. The R-squared and number of observations for each model should be reported (see the Sample Report).
- The programming codes used for preparing the tables, graphs, and regressions should be provided in a clear, easy to trace, and readable format in the appendix or as a separate do-file.
- You don't need to cite any reference, but if you intend to do so, use a proper citation style and provide the reference list in the appendix.
- Overall, the quality (i.e., clarity, rigour, precision, and depth) of the project is more important than the length.