

John Schleider

DATA SCIENCE REFLECTIONS ON FIRM AGGREGATE STATISTICS

Overview

During my senior year, I examined aggregation of firm financial data for large public companies to analyze which macroeconomic variables best explain changes in company data. I show that, just as practitioners claim, nominal GDP growth best explains changes in firm aggregate revenue, but the situation is more complex for profitability.

Collect and analyze data in a reproducible and ethically responsible manner

My firm-level data comes from the Wharton Research database. I use quarterly data from 2012 to 2022, aggregating revenue (sales), capital expenditures, assets, market values, and different representations of profits. I gather quarterly macroeconomic data from the Federal Reserve database (FRED).

Obtain data through searching, scraping, mining or experimental methods

Most macroeconomic data comes with seasonal adjustments or unwanted aggregations. It took time to search FRED for time series without these modifications. I used Python tools and macros to efficiently search and download from the FRED database.

Parse, transform and generate wide-ranging data sets for analysis

It was tricky to aggregate firm-level statistics. I created two datasets: a naïve and conservative aggregation. The naïve aggregation simply summed the relevant variable for all firms, regardless as to whether they report annually or quarterly, whether they pop in and out of the data, report abnormal statistics, etc. The conservative aggregation only examined firms that reported quarterly data for all quarters for 2012-2022. This left large, stable firms with lots of available data, but might result in a sampling bias. I perform my analysis on both of these aggregations.

Statistically analyze data to summarize, draw inferences and make predictions

I used regression and correlation analysis to observe the relationships between the aggregation measures and macroeconomic variables. I find that nominal GDP growth is indeed the best predictor of firm revenue growth. Interestingly, it seems that firm revenue is more volatile than GDP—a one percent change in GDP results in a 1.4 percent change in aggregate firm revenue. This suggests that large firm activity is more risky than other parts of the economy.

Identify patterns and relationships in datasets using visualization and algorithms

To start my analysis, I created a simple correlogram that examined all the potential relationships to further narrow my analysis. The relationship between nominal GDP and firm

aggregate revenue was the strongest between any firm aggregations and macro variables. The correlogram illustrated all the relationships and their strengths well. For this project, I kept regression analysis to simple bivariate models, sometimes with quarterly controls. This is because macroeconomic models are difficult to make, and interactions or positive feedback loops between variables will make analysis useless without careful model creation. I only seek the relationship between variables, so bivariate models are appropriate.

Communicate data methods and conclusions to diverse audiences

The goal of this project was to identify which variables business managers, investors, and other practitioners should examine to understand the health of their fraction of the economy—public companies. I found through regressions and correlation analysis that macroeconomic variables do explain changes in aggregate firm data, but not on a one-to-one basis. Business leaders should note that they operate in a more volatile part of the economy, and their firms are more exposed to the business cycle. The data methods and analysis I used should be accessible to any practitioner with introductory knowledge of statistics.