EC 8862:Empirical exercise 2, Fall 2019

Learning How to Use Compustat and to Estimate Investment Models by GMM

Part 1 is taken from a S. Terry Problem set 1 for the Ec741 course at BU.

Due: Jan 13 by 5.00pm

Part 1: Firm-Level Data from Compustat

At anytime up to 9pm on Nov 17, email me the response to Part 1 as a written up set of solutions and answers. Use Tex to write up your answers, and use any statistical language or package you'd like for the analysis. STATA is cheap for students, and R is free. With your formatted answers, also include your code. The code doesn't have to be pretty, but it does have to be your own work.

1. ## WRDS + NBER Preliminaries

(a) Your first task is to get access to Compustat's North America Fundamentals Annual database. Compustat annual includes standardized information from US-listed public firms' financial statements at fiscal year end, and it's an important source of US firm- level data. You will access Compustat through the Wharton Research Data Services (WRDS) site. WRDS is a business data warehouse to which BC has a subscription. Go to the website here and register for an account using your BC email. Approval can take a day or two, so sign up immediately. Do not wait until some future date before the problem set is due. Once you're set up, Compustat documentation is readily accessible via Google or the WRDS website.

(b) Your next task is to download the NBER-CES manufacturing database here. This database, compiled in a consistently measured fashion, contains data on cost shares, prices deflators, output, etc. for individual manufacturing industries in the US. Down- load the 1987 SIC version. You can use whatever file format you'd like, e.g., STATA or CSV. For each SIC-4 digit industry and year, compute and/or save the following information:

- `labshare`: the labor share in value added, using a 10-year moving average of payroll costs relative to value added
- `capshare`: the capital share, as the residual of the implied labor share
- `vaddfrac`: a 10-year moving average of the ratio of value added to gross output (shipments)
- `piship`: the shipments price deflator
- `piinv`: the investment price deflator

The result should be a panel dataset by `sic` x `year` with each of the variables above.

2. Download, Clean, & Merge Compustat

In this step, you'll download the Compustat dataset itself. To select the right dataset, first login to WRDS. Then navigate to "Compustat-Capital IQ," then "Compustat Monthly Updates - North America," then "Fundamentals Annual." At that point, make the following selections:

(a) Select 1980-01 to 2011-12 for "Step 1. Choose your date range."

(b) Select "Search the entire database" to download the full dataset under "Step 2. Apply your company codes."

(c) Under "Screening Variables," uncheck "FS" from "Industry Format," which will avoid duplication of certain observations

(d) Under "Screening Variables," uncheck "CAD" from "Currency," which will avoid duplication of certain observations

(e) Select "All" under "Step 3. Choose variable types" in order to download all available data. This will result in a large file, but you'll have more information to play with later in Part 2.

(f) Choose your output format under "Step 4. Select query output." I prefer to download files in STATA format with .zip compression, although you can use whatever approach you like.

(g) WRDS will construct your dataset and notify you when it's available for download.

Now that you have the Compustat database in hand, read the data in and perform the following basic cleaning exercises:

(h) Keep only US-based firms, i.e., keep if `fic`="USA".

(i) Keep only final versions of statements, i.e., keep if `final`="Y".

(j) To avoid firms with strange production functions, drop regulated utilities and financial companies, i.e., drop if the 4-digit `sic` code is in the range [4900,5000) or [6000,7000).

(k) Get rid of years with extremely large values for acquisitions to avoid the influence of large mergers, i.e., drop if acquisitions `aqc` is greater than 5% of assets `at`.

(l) Since Compustat records end-of-year capital values, shift the reported book value `ppent` forwards one year, i.e. set $\texttt{ppent}_{it} = \texttt{ppent}_{it-1}$.

(m) Keep only if the book value of assets `at`, book value of capital `ppent`, number of employees `emp`, capital investment `capxv`, and revenues `sale` are nonmissing and positive.

(n) Keep only if the firm exists in the data for greater than or equal to two years.

(o) Merge in the NBER-CES manufacturing data by `sic` · `year`, where you can take the Compustat fiscal year `fyear` = NBER-CES `year`. Keep only the matched observations.

The result of this exercise should be an unbalanced panel of manufacturing firm-years, with tens of thousands of observations. The exact number of observations may differ based on when you download the Compustat data, which is continually updated.

Compute Some Simple Series of Interest

3. Now, you will compute some crude versions of economically interesting series.

(a) `investment`: Set the investment series equal to capital expenditures `capxv` deflated by the investment price `piinv`.

(b) `capital`: Set the capital series equal to the book value of capital `ppent` deflated by the investment price `piinv`. How does this approach for computing capital differ from, and when is it likely to be incorrect relative to, the perpetual inventory method?

(c) `irate`: Set the investment rate equal to the ratio of investment and capital, times 100.

(d) `output`: Set the value added output series equal to `sale` times `vaddfrac`, deflated by the shipments deflator `piship`. Under what assumptions on the production function does this approach make sense?

(e) `outputgrowth`, `empgrowth`: Compute output and employment growth using the Davis-Haltiwanger formulas, converted to percent.

(f) `tfpr`: At this point you have value added, capital, and labor. Use the industry-year specific labor and capital cost shares from the NBER-CES database to compute the log of revenue TFP based on these two inputs. You should not need to estimate any production functions here. Under what assumptions does such a cost share-based TFP measure make sense?

(g) `tfprgrowth`: Compute the growth rate, i.e., 100 times the log difference, of TFP.

At this point, if you compute a histogram or moments of any of these variables, you'll imme- diately see that plenty of outliers dominate mean calculations. You'll also see that there are big, persistent, differences in scaling across firms of different sizes. Your next step will deal with both of these challenges to inference:

(h) Remove firm effects from `tfpr`, `irate`, `empgrowth`, `outputgrowth`, `tfprgrowth`. Hint: This is easiest to do by extracting residuals from a firm effects panel regression.

(i) Drop any observations with residual values less than the 1.5%-ile or greater than the 98.5%-ile for any of these series.

(j) Plot histograms of each of the cleaned series from subpart (h), pooled over all years. Label your figures and all axes.

(k) Report the mean and standard deviation of each of the cleaned series from subpart (h), pooled over all years. Interpret units in all cases.

4. Putting the Dataset to Work

At this point, you have some roughly computed but economically interesting series in hand, and a reasonably cleaned dataset. Let's put this data to work!

(a) **Size Distributions** For the years 2003-2006, plot the distribution of `output`, censoring above at the 90%ile. Label your axes. Is this distribution symmetric? What family of distributions might naturally fit this data?

(b) **Macro TFP Fluctuations and Volatility** For each year in the dataset, compute the average levels of `tfpr` and the average growth rate `tfprgrowth`. Call these macro TFP and macro TFP growth. Report the interquartile range (IQR), i.e., the difference between the 25-th and 75-th percentiles, of macro TFP growth. Interpret this number. In what cases would the IQR be more appropriate to compute than the standard devi- ation? Now, plot the macro level of TFP by year, together with a linear trend, labelling

your axes in all cases. In which set of years is the macro TFP level below its trend? What broad events do these dates correspond to in US macroeconomic history?

(c) **Micro TFP Fluctuations and Moments** Remove firm effects from micro TFP growth `tfprgrowth`. The resulting series is an estimate of micro TFP growth relative to a firm-specific trend. Compute and interpret the IQR of this residualized series. Is the value from this subpart higher or lower than the value from subpart (b)? What can we infer about the relative magnitude of micro and macro volatility?

(d) **Firm-level TFP and Inputs** Compute the correlation matrix of micro TFP levels, investment rates, and employment growth rates, after removing firm effects from each variable. Do the signs of the correlation matrix seem qualitatively consistent with benchmark models of firm investment? Why or why not?

(e) **Distributional Shifts over the Cycle** Remove firm effects from `tfprgrowth` and `output growth`, i.e., compute TFP growth and output growth at the firm level accounting for firm-specific trends. Consider two separate periods of time. Let the first period, "Pre-Recession," include all firms years in [2005,2006]. Let the second period, "Reces- sion," include all firm years in [2008,2009]. On the same figure, plot two histograms or two kernel-smoothed densities (whichever you prefer) of output growth in each period. Label your axes. Also, report the IQR and non-parametric Kelly Skewness of output growth for the two separate periods. How did the first and second moments of output growth at the firm level shift over the business cycle? Interpret these results in statements such as "The size of the output innovation for a typical firm during the Great Recession..." and "Large negative output innovations for the typical firm during the Great Recession became..." Repeat this exercise, i.e., the graph, moments, and interpretation, for TFP growth. Are the output and TFP growth distributional dynamics qualitatively similar?

Part 2: Test Q Models-theory

Part 2 requires you to estimate a couple of investment models

1) the Q-model of investment.
2) The Euler equations for capital in the parametrization of the lecture notes

For simplicity, define Q as the sum of the market value of shares plus debt divided by the value of by the fixed capital stock If you want to be fancy subtract from the numerator the value of inventories and financial assets. Be careful that the timing is right. For simplicity use the beginning of period Q as the regressor (not Q at time t+1, as we have in the lecture handout)

For both 1) and 2) divide the sample into two types of firms based on the total value of assets in 1974. Divide into those who belong to the first quartile and the rest, with the quartile calculated on the basis of the distribution in 1974. Omit firms that enter after 1974. Think about what instruments you may want to use.

Estimate by GMM using xtabond2 in Stata. Comments on the your results.