Empirical Methods in Finance
Tutorial 1

Florian Peters, Xiao Xiao

Universiteit van Amsterdam
Finance Group

January 13 & 14, 2022

Exercise 1: CRSP/Compustat Merged Database (Accounting Data)

- Download from WRDS some key accounting variables (see below) for the period 1980-2020 for all firms in CRSP-Compustat Merged Database and annual frequency.
- Compute three key financial ratios: The market-to-book ratio, book leverage and market leverage
- Plot histograms of the raw and the winsorized ratios. What are reasonable thresholds to winsorize at in your opinion?
- Compute two measures of firm size: the market value and the book value of assets
- Present the following summary statistics for the raw values of all self-computed variables: quantiles 0% (=min), 1%, 5%, 10%, 25%, 50%, 75%, 90%, 95%, 99%, 100% (=max), mean and standard deviation.
- Plot the sample average (per year) of the two firm size measures against time into one single graph (y-axis: yearly mean; x-axis: year).
- Similarly, plot the sample average (per year) of the two leverage measures against time into a single graph.

Exercise 1: CRSP/Compustat Merged Database

- Run panel regressions of book leverage on its determinants
- The determinants (independent variables) to be used are 1. the log of book assets 2. tangibility (defined as PP&E divided by book assets) 3. market-to-book 4. profitability (defined as operating income before depreciation and amortization (oibdp) divided by book assets)
- Report four panel regressions: The first (colums 1) is a pooled OLS regression (without fixed effects) of book leverage on the four determinants. The second is a panel regression of the same variables but with year fixed effects added. The third is a panel regression that adds firm fixed effects instead of year fixed effects. The fourth is a panel regression that adds both firm and year fixed effects. You can use Stata's reghdfe function to run those regressions.

Exercise 1: Hints

- First, you need market and book value of assets:
  - Book value is ‘at’
  - Market value of assets is book value of assets minus book value of equity plus market value of equity
  - For book value of equity you may use the book value of shareholders’ equity, ‘seq’
  - The market value of equity is common shares outstanding (csho) times fiscal year closing price (prcc_f)
- Book leverage is debt in current liabilities (dlc) plus long term debt (dltt) divided by total assets (at)
- Market leverage is debt in current liabilities (dlc) plus long term debt (dltt) divided by market value of assets

Exercise 2: ExecuComp (Executive Compensation)

- Download key compensation variables for all CEOs over the period 1993-2020. Describe how you access the database, how you restrict the sample to CEOs, etc.
- Inflation-adjust total compensation (tdc1) to constant 1993-dollars using the time series of the US consumer price index.
- Compute winsorized total compensation, winsorize by year symmetrically at the 0.5% level.
- Plot two histograms of inflation-adjusted total compensation: raw and winsorized compensation. Do the most extreme outliers (left and right) represent data entry errors or correct entries in your opinion?
- Plot the mean and median (per year) of winsorized total compensation against time (y-axis: yearly mean/median; x-axis: year)
- Compute the fraction of options, stock, bonus and salary in total pay. Is there a complication regarding this data? Are these data available for the entire time period 1993-2020? Why (not)?
- Plot the mean fractions of options and stock compensation per year (two lines in the same graph with y-axis: yearly mean fraction of options / stock compensation, x-axis: year).

Exercise 2: Hints

- To inflation adjust, use Stata's ‘freduse’ function and select the CPIAUCSL series
- Collapse the data by year and normalize the CPI level to 1 in 1993
- Save the inflation index in a separate dta-file
- Then merge the inflation index into your main file by year

Exercise 3: CRSP (Stock Returns and Firm Size)

- Download, for all stocks with share code 10 or 11, for the period Jan 1990-Dec 2020 and monthly frequency, the following data:
  - the monthly stock returns including dividends (item "Holding Period Return"),
  - the stock price (item "Price"), and
  - the number of shares outstanding (item "Number of Shares Outstanding ")
- Plot a histogram of raw and winsorized (at the 0.5% level) returns for the whole sample. For the winsorized returns, let the x-axis go from -100% to +100%.
- Download the Fama-French and momentum factors (MKTRF, SMB, HML, UMD) and the risk-free rate (RF) from WRDS for the same frequency and period as the stock returns and merge them into the stock return file.
- Plot a histogram of the stock price. Why are there negative prices? What do you need to to about that?
- Compute the market capitalization (price times number of shares outstanding) for each observation, and create a variable that divides the sample, for each month separately, into ten deciles of market capitalization (i.e. create a variable taking the values 1,...,10 where 10 is the decile of the largest firms in a given year).

Exercise 3: CRSP (Stock Returns and Firm Size)

- Compute the equally-weighted and the value-weighted excess portfolio return of stocks for each size decile and month. (The excess return is the raw return minus the risk-free rate.) Collapse the data such that you retain only one observation per decile and month.
- Run regressions of the value-weighted excess stock return on the excess market return (MKTRF) separately for each of the ten deciles of market cap (i.e. report ten regressions in one table). Report t-stats (not standard errors or p-values) in parentheses underneath the regression coefficients using two decimals. Report coefficients with three decimals places.
- Run regressions of the excess stock return on the excess market return (MKTRF), the Fama-French factors and the momentum factor separately for each of the ten deciles of market cap (i.e. report ten regressions in one table).
- For which coefficient(s) (if any) in the two tables do you expect a particular pattern? Could these regressions reveal the so-called size effect? Is it present in your dataset?

Exercise 3: Hints

- Use the ‘xtileJ’ function (from Judson Caskey’s webpage) or the ‘astile’ function from fintechprofessor.com to create size deciles per month. To be able to use xtileJ or astile, you may need to install the function ‘egenmore’ first (type: ssc install egenmore). Stata’s own function for creating quantiles by subgroup, xtile2, may work but may be very slow.
- To create value-weighted returns, you may use the STATA’s ‘wtmean’ function (google it to see how to install and use it)
- Use the “outreg2” command to display the regression results. Column 1: decile 1, column 2: decile 2, etc. Type “help outreg2” so see the documentation of the command’s syntax
- it may be useful to download the entire CRSP monthly stock database (all variables) for the period 1960jan-2020dec to your hard drive. Then you can use Stata’s ‘use... using’ function to load only the variables you need for this exercise, e.g.
  use permno date ret prc shrout shrcd using .../crsp_monthly_19602020 if inrange(date,td(1jan1963),td(31dec2020)) & inlist(shrcd,10,11), clear

Exercise 4: CRSP (Stock Returns and Trading Volume)

- Download, for all stocks with share code 10 and 11, for the period Jan 1990-Dec 2020 and monthly frequency, the following data:
  - the monthly stock returns including dividends (item "Holding Period Return")
  - the stock price (item "Price")
  - the number of shares outstanding (item "Number of Shares Outstanding")
  - the trading volume (item "VOL")

  Winsorize stock returns at the 0.5% level symmetrically.
- Sum the dollar trading volume of all stocks for each month separately, and plot a graph showing the evolution of total trading volume in the U.S. over time from 1990 to 2020.
- For each stock and month, construct the variable *Turnover* as the ratio of the number of shares traded and the number of shares outstanding. Annualize the variable by multiplying by 12. Average this variable over all stocks by month and plot a graph showing the evolution of average share turnover in the U.S. over time. What happened in Aug 2019? What does the graph look like if you winsorize *Turnover* by month?

Exercise 4: CRSP (Stock Returns and Trading Volume)

- For each stock and month, construct the following variables: 1. a variable called *Log(volume)* as the natural logarithm of (1 + the **previous month's** trading volume); 2. the variable *SIZE* as the natural logarithm of (the previous month's market capitalization in billions of dollars); and 3. the variable *BM* as the ratio of book value of equity (item CEQ in the *CRSP-Compustat Merged* database, from the previous fiscal year-end) to market capitalization (price times number of shares outstanding from *CRSP Stock/Security files*). Winsorize all the above variables at the 0.5% level symmetrically.
- Report six Fama-MacBeth regressions in one table: the first three regressions use the raw return as the dependent variable, regressions four to six use the excess return (raw return minus risk-free rate) as the dependent variable. Regressions 1 and 4 are univariate, they use only *Log(volume)* as the independent variable. Regressions 2 and 5 add *SIZE* as a control variable. Regressions 3 and 6 add *B/M* as a second control variable. If you interpret *Log(volume)* as a proxy for a stock's liquidity, do these regressions show a liquidity premium in stock returns?

Exercise 4: Hints

- Use the ‘xtileJ’ function (from Judson Caskey’s webpage) or the ‘astile’ function from fintechprofessor.com to create size deciles per month. To be able to use xtileJ or astile, you may need to install the function ‘egenmore’ first (type: ssc install egenmore). Stata’s own function for creating quantiles by subgroup, xtile2, may work but may be very slow.
- To construct BM, you need to merge the book value of equity from the CRSP-Compustat Merged (CCM) into the stock file. To do this, download the CCM database and use Stata’s rangejoin function. The code will look something like:

  ```
  g datadate=date
  rangejoin datadate -365 -1 using /.../.../ccm, by(permno) keepusing(ceq)
  ```
- Use the function asreg from fintechprofessor.com to run Fama-MacBeth regressions
- it may be useful to download the entire monthly stock database (all variables) for the period 1960jan-2020dec to your hard drive. Then you can use Stata’s ‘use... using’ function to load only the variables you need for this exercise, e.g.

  ```
  use permno date ret prc shrout shrcd using .../crsp_monthly_19602020 if
  inrange(date,td(1jan1963),td(31dec2020)) & inlist(shrcd,10,11), clear
  ```

Exercise 5: CRSP (The Value Premium)

- From CRSP, download, for all stocks with share code 10 and 11, for the period Jan 1963-Dec 2020 and monthly frequency, the following data:
  - the monthly stock returns including dividends (item "Holding Period Return"),
  - the stock price (item "Price"), and
  - the number of shares outstanding (item "Number of Shares Outstanding ")

  Winsorize stock returns at the 0.5% level symmetrically.
- Download the Fama-French factors and the risk free rate for the same time period and frequency, and merge them into the stock file.
- From the *CRSP-Compustat Merged (CCM)* database, download the following annual accounting data for all companies and the same time period:
  - the historical CRSP-PERMNO link
  - the variable *datadate* (this is selected by default)
  - stockholders' equity - total (seq)
  - deferred taxes and investment tax credit (txditc)
  - preferred stock redemption value (pstkrv)
  - the fiscal year end's share price (prcc_f)
  - the number of common shares outstanding at fiscal year end (csho)

Exercise 5: CRSP (The Value Premium)

- Compute, for each firm and year, the book-to-market value (BM) as Book Equity divided by Market Equity (ME).
- Book Equity (BE) is defined as stockholders' equity plus deferred taxes and investment tax credit (if available), minus the book value of preferred stock redemption value (if available).
- Market equity is defined as price times number of shares outstanding.
- Save the file containing the annual BM ratio keeping the variables PERMNO, datadate and BM.
- Merge the annual book-to-market data into the stock return file on PERMNO and time such that for each observation in the stock return dataset, the BM ratio of the firm from the most recent fiscal year is matched, i.e. 'datadate' from CCM need to be earlier than 'date' in CRSP. You can do this using STATA's *rangejoin* function. Example:
  ```
  g datadate = date
  rangejoin datadate -365 -1 using BMData, by(permno)
  ```

Exercise 5: CRSP (The Value Premium)

- For each month, create 10 book-to-market deciles using the astile function from fintechprofessor.com or the xtileJ function from Judson Caskey's webpage. Compute (a new variable as) the equally weighted portfolio return for each month and BM decile, collapse the data to one observation per month and decile.
- Run 10 regressions – one for each BM decile – of the excess return on the excess market return (MKTRF), and report them in one table, decile 1 in column 1, decile 2 in column 2, etc. Report t-stats in parentheses below the coefficients. Use three decimal places for coefficients and two decimals for t-stats.
- Run the same 10 regressions on subperiods 1963-1982, 1983-2002, and 2003-2019, i.e produce three more tables with 10 columns each.
- Do the regressions reveal the *value premium*? Is there a pattern in the value premium over time?

Exercise 6: I/B/E/S (Analyst Forecasts)

- Download, for each US firm over the period Jan 1 2000 to Dec 31 2020, the mean, the median, and the standard deviation of analysts' quarterly, one-quarter-ahead, EPS forecasts as well as the number of analysts following the firm. Also, download the forecast period end date (the date of the fiscal period end to which the forecast pertains). Use the so-called "unadjusted summary data". What does "unadjusted" mean in this context? Which database do you use for this?
- Which are the stock identifiers available in I/B/E/S, and does one of them uniquely identify a firm?
- download WRDS's IBES-CRSP link table and merge it into the IBES dataset (you find the table on WRDS⇒Linking Suite by WRDS⇒IBES CRSP Link)
- Define a new variable indicating the quarter of the year of the IBES statistical period (e.g. 1990q1 if the IBES statistical period is 28feb1990). You can use Stata's qofd() function to do this. Plot the mean and median number of analysts following a firm over time into one single graph (x-axis: calendar quarter, left y-axis: mean number of analysts, right y-axis: median number of analysts)

Exercise 6: I/B/E/S (Analyst Forecasts)

- Download actual unadjusted quarterly earnings per share over the same period as above, and merge them into the IBES forecast dataset.
- Compute the forecast error as the difference between the median forecast and actual (realized) earnings, divided by the CRSP stock price one week prior to the earnings announcement, and multiplied by 100. Why does it make sense to scale the forecast error by price? Why not scale by actual earnings? (If you're not able to scale the forecast error by the stock price, scale it by the actual earnings.)
- Plot a histogram of the forecast error using 100 bins. If there are outliers, plot the graph for percentiles 2-99 or 5-95 only. (Use the option *bin(100)* of STATA's histogram function)
- Plot a histogram of the forecast error using bins with a fixed width of 0.1, and showing only forecast errors in a range of -1 to $+1$. (Use the option *width(0.1)* of STATA's histogram function)

Exercise 7: ISS (Corporate Governance)

- In WRDS you find two different governance databases provided by ISS (formerly RiskMetrics). What is the difference between the two? Why are they separate?
- Download the variables classified board (CB), and blank check preferred (BCP) and poison pill (PP) for all firms, separately from the governance and governance legacy databases for the maximum time period possible. What do these variables mean?
- Append the two files.
- Compute the percentage of firms having a CB, BCP, and PP for each year and plot these percentages against time.
- Download the variables Director ID, Director full name, Female, and Board affiliation from the directors database of ISS for the years 1996-2015.
- Compute the percentage of female board members and the percentage of independent directors for each firm. Then collapse the data so that you retain only one observation per firm and year.
- Compute the average percentage of female board members and the average percentage of independent directors for each year and plot both against time in the same graph using two different y-axes. (You can use STATA's option "yaxis(1)", "yaxis(2)" to do this.)

Exercise 8: Zephyr (M&A)

- From the Zephyr database, download all M&A deals that satisfy the following criteria:
    - the acquiror is a publicly listed US firm
    - the deal is announced between 1 Jan 1995 and 31 Dec 2021
    - the deal was successfully completed
    - the deal value is above $10m
    - the acquirer controls less than 50% of the shares of the target at the announcement date and obtains 100% of the target shares in the acquisition
- For the above sample of deals, download the following data items:
    - acquiror and target name, acquiror and target ISIN, announcement and completion date of the merger, transaction value, and measures of acquiror and target firm size (eg sales, total assets).
- Zephyr can be accessed by selecting 'Zephyr' on https://databases.uba.uva.nl/. Once you are on the Zephyr web page, select *Zephyr Advanced*, and continue with the sample and variable selection.

Exercise 8: Zephyr (M&A)

- Import the data from Excel into Stata. Do you encounter any problems?
- Compute descriptive statistics (min, 5th and 95 percentile, max, mean, median, standard deviation) for the following variables: acquiror size, target size, deal value, and the time from announcement to completion.
- Define a new variable called "quarter" which indicates the calendar quarter of the acquisition announcement date. (You can use Stata's qofd() function for this). Compute the number and the total deal value of mergers completed in each calendar quarter from 1 Jan 1995 (or from the 1st quarter that has data on M&A deals) to 31 Dec 2021. Collapse the data to calendar quarter level, i.e. make sure that there remains only one observation per calendar quarter. Save the resulting file.
- Download the monthly return of the CRSP index including dividends (VWRETD) for the period Jan 1995 (or from the 1st quarter that has data on M&A deals) to Dec 2021 (or as long as data is available).
- Import the data into Stata. Create a new variable containing the cumulative monthly level of the index, setting the value for the earliest month to one. (Cumulative returns are computed as $R_t = R_{t-1} \cdot (1 + r_t)$, , where R_t is the cumulative return and r_t is the monthly return in period t.)

Exercise 8: Zephyr (M&A)

- Define a new variable called "quarter" which holds the calendar quarter of each month. Compute the average index level for each quarter, then collapse the data to calendar quarter level, i.e. make sure that there remains only one observation per calendar quarter. (When you plot the index level against time, it should look similar to the S&P500 index over the same period)
- Merge the M&A and CRSP index data files on the variable "quarter".
- Plot the number of mergers (y-axis) and the index level (y-axis) against calendar quarter (x-axis) using two separate y-axes (left axis for the number of mergers, right axis for the index level). You can use STATA's option "yaxis(1)", "yaxis(2)" to do this.
- Plot the total deal value of mergers (y-axis) and the index level (y-axis) against calendar quarter (x-axis) using two separate y-axes (left for the deal value, right for the index level).
- Generate a scatter plot of the total deal value against the index level and add a linear fitted line
- Do these plots suggest any correlation between market valuations and merger activity?

Exercise 1 for Real Estate Finance: Micro Data

This exercise is based on anonymized transaction data from the municipality of Ede. Unfortunately, this type of micro data is usually not freely available on the web. The file Prices_Ede contains transaction prices and dates, the file House_characteristics_Ede contains house characteristics. The data are posted on the Blackboard website in the folder "Tutorial".

- Import the data into Stata
- Merge the house characteristics data to the price data (the price data is the master dataset, use transaction ID as the variable to merge on). Do you get a perfect match?
- Describe your dataset (sample period? number of houses?)
- Provide some descriptive statistics of the data. What is the average, median, and standard deviation of the price per year? Are there outliers?

Exercise 1 for Real Estate Finance: Micro Data

- Generate the log of prices, house type dummies (use the command tab housetype, gen(housetype_dum)), and month dummies.
- A waste incinerator was placed in neighborhood 4001 in 2008. Create a dummy for neighborhood 4001, a dummy that is 1 starting in 2008 (and zero before), and an interaction term between the two (i.e. multiply the dummies with each other.)
- Regress the log of prices on the neighborhood 4001 dummy (is the incinerator placed in a low price neighborhood?), the post-2008 time dummy (are prices trending?), and the interaction term (the treatment effect). Include house type (housetype_dum*) and month dummies.
- Interpret the treatment effect.
- If you added housid dummies as additional controls to run a fixed effects regression (instead of creating a zillion dummies, use areg y x, absorb(housid)), would all of the variables remain in the regression? What happened?

Exercise 2 for Real Estate Finance: Regional Data

This exercise is based on Dutch regional house price (assessed value) data from the CBS. These data are available in the excel files Data_CBS_prices and Data_CBS_other. The data are posted on the Blackboard website in the folder "Tutorial".

- Import the data into Stata
- Merge the other municipality data to the price data (the price data is the master dataset, use municipality and year as IDs to merge on). Do you get a perfect match?
- Describe your dataset (sample period? number of municipalities?)
- Generate the variables log of price average, log of living space, population density (1000 persons per square km), percentage married, and a market tension indicator (households - residences) / residences
- Provide some descriptive statistics of the data. What is the average, median, and standard deviation of the price per year? Are there outliers?

Exercise 2 for Real Estate Finance: Regional Data

- Run a regression of log price on the log of living space, market tension indicator, percentage married, population density, percentage higher educated, and a time trend by including the calendar year in the regression. Also add fixed effects by including municipality dummies (tab municipality, gen(mun_dum) to create these dummies; add mun_dum* in the regression). Interpret your findings (coefficients).
- If you forget important variables in a model you can get omitted variable bias. Can you think of any key (omitted) variables that you would like to include in the model?

Exercise 3 for Real Estate Finance: International Data

This exercise is based on international house price data from Eurostat/Hypostat. These data are available in the excel files Data_eurostat_prices and Data_eurostat_other. You can freely download this data from the Eurostat website.

- Import the data into Stata
- Merge the other eurostat data to the price data (the price data is the master dataset, use country_id and year as id to merge). Do you get a perfect match?
- Describe your dataset (years? countries?)
- Generate the variables log of prices, log of turnover, log of gdp, and log of inflation (interest rate is already a percentage, so no need to create the log variable, inflation is actually the level of CPI).

Exercise 3 for Real Estate Finance: International Data

- Provide some descriptive statistics of the data (do not forget to do tsset country_ID year). What is the average percentage price increase per country (hint: you can use d.log_price as approximate percentage)? Are there outliers?
- Do a first-differenced regression of log price on the log of turnover, log of gdp, interest rate, log inflation and year dummies (tab year, gen(yeardum) to create these dummies, add yeardum* in the regression, you can use the d. operator to do the regression in first differences. Interpret your findings (coefficients).
- Do you have any concerns with adding turnover in the price equation?

Exercise 4 for Real Estate Finance: Time Series Data

This exercise is on time series data. These data are available in the excel files Data_timeseries_prices and Data_timeseries_other. You can freely download this data from the CBS/DNB.

- Import the data into STATA
- Merge the other time series data to the price data (the price data is the master dataset, use time as id to merge). Do you get a perfect match?
- Describe your dataset (years? missings?)
- Generate the variables log of prices, log of gdp (interest rate is already a percentage, so no need to create the log variable), also create season dummies (hint: tab quarter, gen(quarterdum), you should have 4 dummies)
- Provide some descriptive statistics of the data (do not forget to do tsset time). What is the average percentage price increase (hint: you can use d.log_price as approximate percentage)? Are there outliers?

Exercise 4 for Real Estate Finance: Time Series Data

- Run a regression of log prices on log gdp, interest rates, and the season dummies (leave 1 out). Create a variable with the residuals.
- Run a first-differenced (error correction model) regression of log price on log price t-1, log of gdp, interest rate, the season dummies (leave 1 out) and the residuals (you can use the d. operator to do the regression in first-differences, the l. operator to create lags). Note that the residuals are not differenced. They are the error correction term. Interpret your findings (coefficients).

[Note: Prices are typically non-stationary, so you estimate the model in first differences, by including a constant in the regression you automatically detrend the data (trend in levels)]
[Note: Before you can add the residuals in your main equation you typically have to test for cointegration first using a dickey fuller unit root test. If the residuals are stable, you should reject the unit root null hypothesis]
[Note: These types of time series models you can also use to forecast!]
[Note: in time series the number of observations can drop quickly. Try to use high frequency data]