# Ph.D. Class in Empirical Corporate Finance -Professor Rüdiger Fahlenbrach

## Empirical Problem Set: Compustat and Its Perils

**Due Mar 31, 2023, via email to ruediger.fahlenbrach@epfl.ch**

The purpose of this problem set is to introduce you to the most commonly used database in empirical corporate finance research, Standard and Poor's Compustat database. Few universities in Europe can afford to subscribe to this database, so we are lucky to have it here in Lausanne. After this homework, you will know how to use Compustat, will be able to easily replicate the variables constructed in published research, know what some of the pitfalls using Compustat are, and know how to write some basic code to summarize and clean the data.

Everyone must hand in his or her own individually prepared responses to the questions. I consider it a violation of the honor code to copy computer code and present the output as your own work. To download the data, you will need a WRDS account (go to wrds.wharton.upenn.edu, and register for an account). If you encounter insurmountable difficulties in getting a WRDS account, let me know. I recommend that you use STATA or Python for this assignment, but you may use another statistical software package of your choice. Some people in the past have used Matlab but reported back that this was a very painful experience. Please submit a clean write-up of your answers, and also attach/email the code you used. **I expect this problem set to be a significant amount of work – you should budget at least 15 hours for completing it.**

### 1) Preliminaries

a) Using Compustat North America on WRDS, obtain the following document "Balancing Models – North American Company Data.xls" (available under support/Manuals and Overviews/Compustat, https://wrds-www.wharton.upenn.edu/pages/support/manuals-and-overviews/compustat/)) and keep the annual (A) worksheets. If you do not find the spreadsheet, let me know and I will send it to you.

b) Download annual data for the following firm variables for the universe of public companies for fiscal years 1970-2022. You should use the data under "Annual Updates", "Fundamentals Annual".

Name of Firm
GVKEY (the Compustat Identifier)
CIK Number
Month of Fiscal Year End
Fiscal year
Industry (NAICS and SIC)
IPO date
Number of Common Shares Outstanding at Close of Fiscal Year
Common Share Price at Close of Fiscal Year

Balance Sheet Variables: Cash and short-term investments, Total liabilities, Total Long-term debt, short-term debt, Total Assets, Total Shareholders' Equity at Book Value, Preferred Stock at Book Value, Net PP&E

Income Statement Variables: Sales, Net Income, Operating Income After Depreciation (EBIT), Operating Income Before Depreciation (EBITDA), Interest paid, Research and Development Expense

Cash Flow Variables: Capital expenditures, Cash Dividends, Purchase of Common and Preferred Stock ("Repurchases")

c) What is the CIK number? What is the gvkey? Outside compustat: What is the CUSIP? What is the permno? Keep these different identifiers in mind. A non-trivial task in any empirical project is to link different databases for your research in a consistent way.

d) Find a way to select U.S. headquartered firms only. Are there many non-US firms in the Compustat files? Why are there non-US firms in the Compustat North America files?

In the following parts 2) – 4), only use U.S. headquartered firms.

**2) Calculate the following ratios:**

Book Leverage 1 = Total Debt / Total Assets
Book Leverage 2 = Total Liabilities / Total Assets
Net Book Leverage 1 = (Total Debt – Cash and Short-term Investments) / Total Assets
Common Equity at Market Value = Number of Common Shares Outstanding * Common Share Price
Market Leverage = Total Debt / (Total Debt + Preferred Stock at Book Value + Common Equity at Market Value)
Asset Tangibility = Net PP&E / Total Assets
Cash&ST inv. ratio = Cash & short-term investments / Total Assets
Return on Equity (ROE) = Net Income / Total Shareholders' Equity at Book Value
Profit Margin = Net Income / Sales
Capex_ratio = Capital expenditures / Total Assets
R&D_ratio = Research and Development Expenditures / Total Assets
Dividend Yield = Dividends per Common Share / Lagged Price per Common Share
Dividend Payer = indicator variable equal to one if company pays dividends, zero otherwise
Total Payout Ratio = [Dividends + Repurchases] / Net Income
EBIT interest coverage = Operating Income After Depreciation / Interest Paid

a) Winsorize all of these variables at the 1st and 99th percentile year-by-year. In other words, for each variable, replace the values above the 99th percentile and below the 1st percentile in each year *t* with the 99th or 1st percentile value respectively. Keep an unwinsorized version of your data. You need it later. Here is a STATA winsorize routine: http://ideas.repec.org/c/boc/bocode/s361402.html

b) For each variable, tabulate means, medians, minimum, maximum, standard deviations and observations counts

ba) for the whole sample
bb) for companies with assets larger than the 75$^{th}$ percentile of assets in each year

How far off are means and medians? What does this suggest for empirical work? Is there a difference between the overall sample and the large company sample?

c) Draw a figure of the evolution of average asset tangibility, cash & short-term investment ratio, and the fraction of dividend-paying firms from 1970-2022. Compare your figures with the figures from Kahle and Stulz from lecture 1.

d) For the book and market leverage variables only, tabulate means, medians, standard deviations, min, max, and observation counts.
   a. Examine minimum and maximum. Are those numbers plausible? What do you suggest one should do in empirical research?
   b. Compare book leverage 1 with net book leverage 1. What do you conclude from this comparison?

You should present one table for each of (a), (b), and (d). Do not include statistics in your tables beyond the ones requested. These should be actual tables with one line per variable and no extra lines within. Each table should be clearly labeled with a title and intelligible variable names. You can format tables outside of your statistical package, but what I really want to see is that you can use your statistical package to create legible tables. If you have never done this, you will find this quite painful. However, developing some technology to do this now will help you in the future. If you write an empirical dissertation, chances are you will be glad that you got this out of the way early. If you use stata, you may want to consider the "estout" package. Just type "findit estout" in the command line.

**3) The importance of outliers and time-series dependence**

a. Estimate a regression of Book Leverage 1 on log (total assets), asset tangibility, R&D_ratio, dividend payer, and profit margin using the unwinsorized data from 1970-2022.
b. Estimate the same regression using the winsorized data. Compare the results.
c. In addition to b), restrict book leverage to be between zero and one. Re-estimate the regression and compare. What do you conclude?
d. Use the winsorized data. Now include, in this order, in the regression: 1) year-fixed effects, 2) year-fixed effects and industry-fixed effects, 3) year x industry-fixed effects. Compare the results and interpret differences. If you are motivated, look up "Fama-French 49 industry-fixed effects" and implement. Otherwise, it is fine to use industry-fixed effects based on two-digit SIC codes.
e. Estimate the regression of book leverage as in part a) and include year-fixed effects and industry-fixed effects. Now cluster the standard errors by firm. Compare t-statistics to those obtained in part c). Why do they change so much?

**4) Fiscal Year and calendar year in Compustat:**

Calculate the one-year change in the market value of equity, i.e. the percent change between the end of fiscal year *t* and the end of fiscal year *t-1*. Tabulate the mean and standard deviation of this quantity for observations with

fiscal year 2000, month of fiscal year end = 1 (Jan)
fiscal year 2000, month of fiscal year end = 2 (Feb)
[…]
fiscal year 2000, month of fiscal year end = 12 (Dec)
fiscal year 2001, month of fiscal year end = 1 (Jan)
fiscal year 2001, month of fiscal year end = 2 (Feb)
[…]
fiscal year 2001, month of fiscal year end = 12 (Dec)

so that you are presenting 24 means and standard deviations.

a) What is the calendar date range over which the one-year change in the market value of equity has been calculated for each of these rows?

*Hint: Compustat has a peculiar convention for how they treat fiscal years. The big difference is between fiscal year end month 5 and 6. Searching wrds for "compustat fiscal year" may greatly facilitate solving this exercise.*

b) Download the paper "Fama, Eugene F., and Kenneth R. French, 1992, The Cross-section of expected stock returns, Journal of Finance 67, 427-465, available here: https://onlinelibrary.wiley.com/doi/epdf/10.1111/j.1540-6261.1992.tb04398.x.

Read the first two paragraphs on page 430 (*Our use of December market equity […] with similar results.*). The paper describes the choices a researcher needs to make when combining accounting data with stock market data. In your own words, describe the two problems a researcher faces, **explicitly linking them to the exercise you just did for fiscal year 2000/2001**.

**5) Mergers and Acquisitions in Compustat:**

Compustat has a particular convention on how to treat companies after mergers and acquisitions, which is important for you to know. Take the merger of ExxonMobile. Look up when the merger happened. Locate all the observations in your data for the two pre-existing firms and the resulting combined firm from five years before to five years after the merger. Create an accurate list of book leverage and asset growth (i.e., (total assets year t / total assets year t- 1) – 1). Look at asset growth in the year of the merger and compare it to the other years. Under what circumstances could this become a problem?

Hint: You will mess this or similar searches up if you treat the Compustat company name field as definitively correct for any year other than 2022. The name item is backfilled in Compustat. Part of this exercise is identifying the Compustat observations that represent the acquirer and target, respectively. If you want a list of the historical names of companies, you will have to use the CRSP database, and their historical name field.

Now pick the AOL – Time Warner merger and figure out how Compustat treated the respective firms in this case pre- and post-merger. Why is this so much more complicated here than for the ExxonMobile case?