Homework 1: Compustat and its Perils

Applied Corporate Finance – FINA6223A.H2026

Prof. Jakub Hajda

Homework Details	
Deadline:	January 31, 2025, 23h55 Montreal time
Max points:	100
% of final grade:	10%
Team composition:	Teams of 2 or 3
Submission form:	PDF file with results (tables, graphs) and discussion **and** commented code (for example, Interactive Python Notebook)
Submission method:	Submit by e-mail
Useful packages:	`pandas`, `linearmodels`, `statsmodels.api`, `collections`, `matplotlib.pyplot`

Introduction

In this problem set, you are going to investigate several empirical properties of the COMPUSTAT database, which is one of the most important tools available to empirical researchers in corporate finance and which we will be using throughout the course. The purpose of this homework is to (1) get acquainted with standard data manipulations in COMPUSTAT that will be used in the remaining homeworks and throughout the course and (2) improve your data analysis skills in Python (or another statistical software) and (3) get used to typical data operations as the ones in research papers that we analyze in this class.

General tips

- Read the instructions carefully.
- Make sure that your reported graphs and tables are clear, readable and contain all the information. When writing explanations be clear, concise and to the point. No points will be given for answers out of topic or for unnecessarily long and unclear arguments.
- You are free to discuss the homework with your classmates. However, each team should submit their own code and their own results.
- Please be reminded of the HEC Regulation regarding the intellectual integrity of students. Any suspicion of plagiarism will be treated accordingly.

- If you use Chat GPT, make sure to indicate it properly. Note that it is very easy to commit plagiarism when using it because it copies verbatim other sources.

Data preparation

Download the annual data on firms' fundamentals from Compustat North America, covering the fiscal years 1995–2020. In particular:

1. Click on "Compustat - Capital IQ", then click on "North America" under "Compustat", then click on "Fundamentals Annual".

2. In "Step 1", specify the start and the end of the sample period, i.e. January 1995 to December 2020.

3. In "Step 2", select "Search the entire database". Don't change anything in the section "Screening variables",.

4. In "Step 3", select all the variables that are necessary to compute the following financial ratios (see the column **Formula** in Table 1). Note that, for each firm, `lprice` denotes the lagged stock price `prcc_f` and `lat` denotes the lagged total assets `at`. You need to create these lags yourself. Apart from these variables, select also company name (`conm`), headquarters location (`loc`) and industry SIC code (`sic`).

5. In "Step 4", select the desired format, execute the query and download the data.

Financial ratio	Formula
Book leverage (1)	`(dlc+dltt)/at`
Book leverage (2)	`lt/at`
Market value of equity	`csho*prcc_f`
Market leverage	`(dlc+dltt)/(dlc+dltt+pstk+csho*prcc_f)`
Market-to-book	`(prcc_f*csho+dltt+dlc+pstkl-txditc)/at`
Asset growth	`at/lat-1`
Asset tangibility	`ppent/at`
Return on equity	`ni/ceq`
Profit margin	`ni/sale`
CAPEX ratio	`capx/at`
Dividend yield	`(dv/csho)/lprice`
Dividend payout ratio	`dv/ni`
Total payout ratio	`(dv+prstkc)/ni`
EBIT interest coverage	`ebit/xint`
Cash holdings	`che/at`
Profitability	`oibdp/at`

Tab. 1: Ratios used in the problem set and their formulas, computed using variables with original COMPUSTAT names.

1 Understanding Data Issues [20 points]

1. Check if you have any duplicates in your data. If so, remove them. Report the total number of observations in your data. Identify all firms which are not headquartered in the U.S. and drop them from the data. Report the new number of observations. Plot the evolution in the number of U.S.-based **firms** over the sample period.

2. Winsorize each ratio at 1st and 99th percentile in each fiscal year. Create a table with summary statistics that contains: the mean, the median, the minimum, the maximum, the standard deviation and the number of non-missing observations for each of the financial ratios. *Hint*: to winsorize means to set values below (or above) a certain quantile to the value of that quantile in the data; each ratio will typically have a different number of non-missing observations.

From now on, use the winsorized data unless asked explicitly not to.

3. Split the firms into 4 groups (*quartiles*) each year depending on the market value of equity. Create a table with summary statistics of ALL variables in Table 1 for the firms in the smallest and largest quartiles that contains the mean, the median and the standard deviation for each of the two subsamples. Comment on the main differences between the two samples and provide a reason why this may be the case.

4. Split the firms into **financial** and **non-financial** ones. To do so, use the industry indicator `sic`: you can identify financial firms as those whose first two digits of the SIC code are between 60 and 67 (inclusive). How many financial and non-financial firms are there, *on average*, in the sample every fiscal year? Create also an indicator for **utility/regulated** firms, whose first two digits of the SIC code are between 40 and 49 (inclusive).

5. Using book leverage (1) and (2) as well as market leverage, create a table that contains the mean, the median, the standard deviation and the number of observations for each:

   (a) the sample of financial firms,

   (b) the sample of utility firms,

   (c) the sample of non-financial and non-utility firms,

   (d) the sample of non-financial and non-utility/regulated firms with **non-missing** value of total assets throughout the **whole** sample period.

6. Comment on the following (provide a reason why we observe what we observe):

   (a) The differences in financial leverage between 1. financial, 2. utility/regulated and 3. non-financial, non-utility/regulated firms.

   (b) The differences in financial leverage between non-financial firms and non-financial with non-missing value of total assets throughout the whole sample period.

   Based on the results, is it justified that financial and utility/regulated firms are often excluded from empirical analysis? Why (not)?

2 Exploratory Data Analysis [40 points]

We are now going to focus on the winsorized sample of non-financial and non-utility firms. Focus on six financial ratios: **book leverage (1)**, **EBIT interest coverage**, **cash**, **profitability**, **total payout ratio**, **market-to-book**.

1. Report the following

   a. a histogram-scatter matrix (*Hint*: check `pandas.plotting.scatter_matrix`),

   b. for each ratio, a time-series graph that contains the *average* and the *aggregate* value of the corresponding ratio over the sample period. *Hint*: to aggregate, compute the sum of each of the components of the financial ratio in a given fiscal year and then compute the ratio itself.

2. Create a correlation matrix of the six variables. Then, transform each of the six variables by removing the **firm-specific** mean and create another correlation matrix for these variables.[1] Which differences do you observe between the two matrices?

3. Let us consider the following model:

$$y_{it} = \alpha + \beta x_{it} + \varepsilon_{it},$$

where y is book leverage and x is profitability.

   a. Estimate the model using OLS. Report the results. Use appropriate standard errors. Motivate your choice.

   b. Estimate the following model using OLS

$$\tilde{y}_{it} = \bar{\alpha} + \bar{\beta}\tilde{x}_{it} + \bar{\varepsilon}_{it},$$

   where $\tilde{x}_{it}$ correspond to transforming the data in the same way as in question 2 where you transformed it before reporting correlations. What do you notice about $\bar{\beta}$ as compared to the estimates that you found in parts a?

   c. Estimate the following model using OLS

$$y_{it} - y_{it-1} = \tilde{\alpha} + \tilde{\beta}(x_{it} - x_{it-1}) + \tilde{\varepsilon}_{it}.$$

   What do you notice about $\tilde{\beta}$ as compared to the estimates that you found in parts a and b?

   d. Estimate the model **firm-by-firm** (i.e., you are going to get an estimate of β_i for each firm i). Only use data of firms with at least 10 non-missing observations. Plot the histogram of the resulting estimates of β_i and create a table with the mean, the median, the minimum, the maximum for the full sample.

   e. Divide the firms into 4 groups depending on their market value. Create a similar table that contains the same summary statistics for each subsample obtained by dividing the sample of firms into these groups. Comment your results. *Hint*: make sure to divide the sample using **entire** firms and not particular firm observations, given that β_i is firm-specific.

3 Bankruptcies in Compustat [20 points]

We are going to analyze a mini "case study" of General Motors Corp., which went bankrupt in June 2009, and of Enron, which went bankrupt in December 2001.

[1] This means you should have as many different means as there are firms and then remove this mean from each observation of interest: $\tilde{x}_{it} = x_{it} - \bar{x}_i$, where i is firm and t is fiscal year.

1. Isolate the companies in the raw, **non-winsorized**, data (*Hint*: search by company name `conm`). Create a table with the yearly evolution (over the sample period) of book leverage (1) and market leverage, market value of equity, CAPEX ratio, asset growth, ROE, EBIT interest coverage, profitability, total payout ratio, dividend yield, and profit margin for each company.

2. What happens before, during, and after the fiscal year in which the companies went bankrupt? How did the financial ratios evolve?

3. Are they indicative of financial distress? Why (not)? Do you notice any concerns for empirical analysis? *Hint*: think of the following (1) why did Enron go bankrupt? (2) do you see any issues why re-listing bankrupt firms may be bad for data analysis?

4. Using the **winsorized** data for *all* firms, compute the *standard deviation* of book leverage (1), market leverage, dividend yield, dividend payout and total payout for **each firm** with at least 5 years of non-missing observations. Using the industry classification from Table 2 (the remaining firms should be discarded), create a table that shows the mean and the standard deviation of the firm-level standard deviations across the 9 industries.

5. Which industries exhibit the most variation in the financial ratios? What is special about these industries? Taking into account what we learned from the case GM's bankruptcy, what could it imply for analyzing financial (and other) policies of firms in these industries?

SIC code	**Industry**
0100-0999	Agriculture, Forestry and Fishing
1000-1499	Mining
1500-1799	Construction
2000-3999	Manufacturing
4000-4999	Transportation & Public Utilities
5000-5199	Wholesale Trade
5200-5999	Retail Trade
6000-6799	Finance, Insurance and Real Estate
7000-8999	Services

Tab. 2: Industry classification.

4 Properties of OLS Estimator [20 points]

We are going to analyze several properties of the OLS estimator that we discussed in class using simulation. Suppose that you have a linear regression of the form

$$y_i = \alpha + \beta x_i + \varepsilon_i,$$

where $x_i \overset{\text{iid}}{\sim} \text{Uniform}[0, 50]$ and $\varepsilon_i \overset{\text{iid}}{\sim} \text{Normal}(0, 1)$. Assume that the true values of α and β are $\alpha = 10$ and $\beta = 0.5$. Note: *i.* if you write the code as functions, you can reuse them in each point, *ii* set the seed for generating your random numbers. This is important so that you always work with the same simulations.

1. Simulate $n = 100$ values of x_i and ε_i. Use these values to calculate y_i. Estimate the linear regression using OLS and report the resulting estimates.

2. Repeat step 1 $N = 10,000$ times, that is, simulate $N = 10,000$ datasets and for each dataset estimate the model using OLS. Save the resulting estimates $\widehat{\alpha}_i$ and $\widehat{\beta}_i$.

3. Plot the histogram of the resulting estimates of α_i and β_i. Create a table with the mean, the standard deviation, the median, the minimum, and the maximum of each estimate for α and β. What do you observe? What does this tell us about the OLS estimator?

4. Repeat points 1 and 2 by using $n = \{25, 100, 1000, 10000\}$. Plot the four histograms in a single graph. What do you observe? What does this tell us about the OLS estimator?

5. Repeat points 1, 2 and 4 by by assuming that:

   (a) $\varepsilon_i \overset{\text{iid}}{\sim} \text{Uniform}[-5, 5]$

   (b) $\varepsilon_i \overset{\text{iid}}{\sim} \text{Poisson}(1) - 1$

   What do you observe? What does this tell us about the OLS estimator?