Homework 2: Determinants of Capital Structure

Applied Corporate Finance – FINA60223A.H2026
Prof. Jakub Hajda

Homework Details	
Deadline:	21 February 2026, 23h55 Montreal time
Max points:	10
Team composition:	Teams of 4 (or 3 if odd # of students)
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 Python packages:	`pandas`, `linearmodels`, `statsmodels.api`, `collections`, `matplotlib.pyplot`

Introduction

In this problem set, you are going to replicate several results from Leary, Roberts, and Zender (2008)[1] and expand on their analysis. This homework will test your skills across three dimensions: *(a)* understanding panel data models *(b)* getting from the data to the model and *(c)* understanding how we can study the determinants firms' financing decisions.

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.

[1] Lemmon, Michael L., Michael R. Roberts and Jaime F. Zender, 2008, **Back to the Beginning: Persistence and the Cross-Section of Corporate Capital Structure**,'*Journal of Finance* 63 (4), 1575-1608.

- 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 period beginning with the first year used in the paper and extending it to the **most recently available period**. 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.
3. In "Step 2", select "Search the entire database" (do not touch the section "Screening Variables").
4. In "Step 3", select all the variables that are necessary to compute the variables you will need (see below). Apart from these variables, select also industry SIC code (`sic`) which will help you create industry fixed effects.
5. In "Step 4", select the desired format, execute the query and download the data.

You will also be asked to download several items from the *quarterly* Compustat. To do so, follow the steps similar the ones above, except choose "Fundamentals **Quarterly**" and select all the items that the exercise will ask. Make sure that the data period corresponds to the one of the annual Compustat (to the extent possible; it could be that the quarterly data starts later – in this case, your sample will start later as well, so just replace the starting year with the first year available).

1 Exploratory data analysis [50 points]

1. Based on the definitions in the appendix of the paper, create all the variables needed to replicate the summary statistics table and the regression tables (see below exactly which ones), except for cash flow volatility (you will calculate it differently below). Apply any data filters used by the authors. Note that the authors trim (and not winsorize) the variables: that is, they simply remove the values from the tail. Use the translation between the data item number and the variable name in Compustat from this spreadsheet (look at first column): `https://wrds-www.wharton.upenn.edu/documents/894/Compann_Variable_Translation.xls`. You can use the GDP deflator from FRED (if you need it): `https://fred.stlouisfed.org/series/GDPDEF`. Note that while not specified in the paper, use the 4-digit SIC code to identify industries. You may also not find the definitions of some variables in the paper, in this case either *i.* use common sense (e.g., dividend payer is a dummy variable when the firm pays cash dividends, 0 otherwise; industry median leverage

is the leverage median in each industry-year, etc.) or *ii.* use the definitions from the first homework.

2. For firms in your sample, obtain **quarterly** data on the variables necessary to calculate profitability. The variables' names are `oibdptq` (operating income before depreciation) and `atq` (total assets). Calculate profitability as defined in the paper. For each firm and each fiscal year, calculate the volatility (i.e., the standard deviation) of profitability and call it `cf_volatility`. To calculate it, use the observation from the given and the preceding year (so that you have eight observations). For example, to calculate `cf_volatility` for Apple in the fiscal year 2001, use the observations of profitability in the 8 quarters in fiscal years 2001 and 2000. Replace all calculated cash flow volatilities with missing values if there were less than 4 observations used to compute them (i.e., if there were at least 5 observations missing in the 8-quarter window). Merge the cash flow volatilities to your main, **annual**, data. You should now have one observation of `cf_volatility` per firm per fiscal year. Remember to lag the volatility data by one year for each firm.

3. Replicate the "All Firms" panel of Table I (you need the IPO date for survivors, which you don't have). Add two more panels to your table: that is, replicate the "All Firms" panel for two periods, 1965–2003 and 2004–most recent year (based on what you downloaded). Are there any differences between the period used by the authors and the most recent period? If so, speculate briefly as to why.

4. Replicate Panel C of Figure 1. Notice the difference between event year and calendar year. Simulate two sets of random numbers (call them x and y) from the standard normal distribution of 1,000 observations each. Create 10 quantiles based on the realization of x. Calculate the mean of x within each quantile range, starting from the values between quantile 1 and 2 and ending between quantile 9 and 10 – $\bar{x}_{q(x)}$. Calculate the mean of y based on the same quantiles of x — $\bar{y}_{q(x)}$. Plot $\bar{y}_{q(x)} - \bar{x}_{q(x)}$ (y-axis) vs. $\bar{x}_{q(x)}$ (x-axis). Explain how this demonstrates that the results in Figure 1 might be spurious.

5. Note that in the paper, values of leverage in year t are regressed on leverage determinants in year $t-1$. Apply the same transformation to your data, i.e., for each firm, lag the explanatory variables by one period. After this transformation, you will have many rows with missing data. Remove all of them (including those that were missing earlier).

2 Leverage models [50 points]

1. Replicate Panel A of Table II (using all the data; this applies to all the models below as well).

2. Estimate the following regression model using all the variables in Table II except for initial leverage

$$Leverage_{it} = \alpha + \beta X_{it-1} + \varepsilon_{it}$$

(a) using pooled OLS,

(b) using fixed effects with year fixed effects,

(c) using fixed effects with firm and year fixed effects,

(d) using fixed effects with firm and industry × year fixed effects.

For now, you do not need to cluster the standard errors. Report the estimation results together in a single table, with one panel for each measure of leverage. That is, each panel will have estimates from 4 different regressions, depending on the use of fixed effects, which you should indicate at the bottom of the table. Does adding different fixed effects substantially change the estimated values of the parameters?

3. Comment on the economic importance of the results (you can use the summary statistics table you created earlier). Explain why the coefficients change between specifications and elaborate on the kind of unobserved heterogeneity these different fixed effects help control for.

4. Is it a good idea to add industry fixed effects instead of firm fixed effects? Why (not)?

5. Re-estimate the model which replicates Table II with standard errors clustered by *i.* firm and *ii.* industry. Focus on book leverage only. What happens to the t-statistics when you cluster? Do any conclusions about the significance of the factors change? How should you cluster the standard errors?

6. Based on the analysis in Frank and Goyal (2009) or your own knowledge, propose and execute a robustness test that would provide further evidence regarding the sign and/or magnitude of the result for a selected variable from the model which replicates Table II. Remember to cluster the standard errors in an appropriate way.

7. **Bonus question – extra credit (only if you want to try)**
Do a variance decomposition of the following model, similar to the one in the paper (only report the % of explained variation of each variable):

$$Leverage_{ijt} = \alpha + \beta X_{ijt-1} + \eta_i + \nu_{jt} + \varepsilon_{ijt},$$

where i corresponds to firm, j corresponds to industry and t corresponds to time.