FIN9013 Assignment 2

Ang Zhang

March 1, 2025

Introduction

This report examines unobserved heterogeneity and correlations in a corporate finance setting. The research question is the capital structure measured by market debt ratio. The methodologies used follows Petersen (2009) and Gormley and Matsa (2014).

Part A. Data management and variable construction

Data are primarily from Compustat, with the *linkdt* variable from the CRPS/Compustat link table being used to determine the firm age. As both Compustate and CRSP have an indicator for exchange code (exchg in Compustat, exchd in CRSP) and they yield different results, this study uses exchg in Compustat to match the original study. Independent variables are lagged a year. Summary statistics are presented in Table 1 in appendix.

Part B. Standard errors

The results of this part are shown in Table 2 in appendix. Comparing the first four columns we can draw conclusion regarding the firm effect and time effects. Although the coefficients in the first four regressions are the same, their standard errors are vastly different. From how biased the estimation is, we can infer for each variable time effect is stronger or firm effect is stronger. In general, the standard errors in II is pretty close to that in IV, indicating that for this study correlation within firm dominates. This is intuitive, as we would effect strong time-series correlation in these variables for a given firm. Firms are not likely to change their fundamentals drastically year over year, so firm effect can be very persistent. For example, the standard errors of ln_firm_age and ln_market_value_of_assets are much larger in the OLS with White standard errors than in the OLS with clustered standard errors by firm, time, and both firm and time. This suggests that the time effect is stronger than the firm effect for these two variables. The standard errors of market_to_book_assets and profits_to_sales are much larger in the OLS with clustered standard errors by firm, time, and both firm and time than in the OLS with White standard errors. This suggests that the firm effect is stronger than the time effect for these two variables.
Fama-MacBeth is a different story, as it uses a two step approach which is different from the first four estimation. As can be seen from the results, Fama-MacBeth produced a different result. Given the background under which Fama-MacBeth is developed, it might not be a good fit here. It works well when only time effect exists but doesn't work when firm effect exists and can produce biased results. It was invented to tackle asset pricing problems where time effect is strong. But in this example, as we expect economically that firm effects is strong while time effects is weak, Fama-MacBeth would not be a good fit.
Since the standard errors are biased, the statistical significance is not reliable either.

Part C. Unobserved heterogeneity

In this setting, the only consistent estimator is the GFE. The GFE estimator is equivalent to demeaning both the independent variable and dependent variable. Consider a panel data model with entity fixed effect:

$$y_{it} = \alpha_i + \beta x_{it} + \epsilon_{it} \tag{1}$$

Demeaning all independent variables with respect to group (firm) : $\tilde{x}_{it} = x_{it} - \bar{x}_i$, and demeaning the dependent variable:$\tilde{y}_{it} = y_{it} - \bar{y}_i$. Plugging in this OLS, we have:

$$\tilde{y}_{it} = \alpha + \tilde{x}_{it} + \epsilon_{it} \tag{2}$$
$$y_{it} - \bar{y}_i = \alpha + \beta(x_{it} - \bar{x}_i) + \epsilon_{it} \tag{3}$$
$$y_{it} = (\alpha - \beta\bar{x}_i + \bar{y}_i) + \beta x_{it} + \epsilon_{it} \tag{4}$$

This is equivalent to the GFE model mathematically.
All other estimations, including OLS, AvgE and AdjY, are not consistent. And their estimated coefficents are all over the place: some are biased upward, some are biased downward.
The difference in each model represents the different cause of heterogeneity. The difference in OLS estimats suggests in general the presence of heterogeneity issue possibly caused by unobserved firm characteristics that correlates with both the dependent variable and independent variables. The difference between AdjY / AvgE and FE may imply the presence of correlation between independent variable and average which is not removed by simply demeaning the dependent variable or controlling the firm mean. This is pervasive in corporate finance applications because firm fundamentals are, intuively, correlated with firm (and it's mean).

Part D. Write-up and exposition

Standard errors: Having the correct standard error is sometimes vital to the validity of an empirical study. When i.i.d assumption is violated, OLS can produce biased standard errors. And the method to correct standard error would largely depend on the economic nature of the research question. A method that works well in one setting (like FM in asset pricing) might not work in another (like in CF).
Unobserved heterogeneity: Unobserved heterogeneity is a common problem in panel data analysis. The GFE model is a powerful tool to tackle this problem. It is consistent and unbiased. In practice, I found several ways to implement the GFE model. While both SAS and Python, the two programming tools that I'm familiar with, both have packages and procedures to implement the GFE model, it is possible to do it by hand and having the same results. For instance, demeaning the dependent variable and independent variable by group (firm) and then running OLS on the demeaned data would produce the same results as including a entity fixed effect in the regression. Also, I found a two step approach, where I ran a ordinary OLS regression and then demean the dependent variable with the mean of the residual in that group, and then ran the OLS regression again, would produce the same results as the FE model.

Part E. Extra credit: replication

Results are shown in Table 4 and Table 5 in appendix. The data managment process is a little bit different from Part A. As both linkdt from the CCM link table and namedt from CRSP produced unreliable results for firm age computation (they are later than fyear, possibly because Compustat included data before a firm went public), for these observations I used the first year the firm appears in Compustat to compute their firm age. Secondly, since there's a huge amount of missing data in advertising expense (77% are missing), instead of dropping all missing values I replaced them with 0 to avoid lossing too much data.

Appendix

Table 1: Summary statistics table

Variable	Observations	Mean	Std. Dev.	Min	Median	Max
market_debt_ratio	10094	0.176866	0.142859	0.000000	0.146809	0.790811
ln_market_value_of_assets	10094	8.510383	1.881171	4.110472	8.461485	13.179756
ln_firm_age	10094	2.835833	0.903136	0.000000	3.091042	4.007333
profits_to_sales	10094	0.204212	0.130631	-0.045369	0.168042	0.825143
tangible_assets	10094	0.254928	0.212209	0.000000	0.213651	0.908860
market_to_book_assets	10094	1.762658	0.962787	0.771887	1.434866	7.282147
advertising_to_sales	10094	0.027397	0.031033	0.000000	0.016362	0.185723
rd_to_sales	10094	0.011700	0.023733	0.000000	0.000000	0.135685
rd_positive	10094	0.381117	0.485685	0.000000	0.000000	1.000000

Table 2: Regression results

	I	II	III	IV	V
ln_firm_age	-0.0227**	-0.0227**	-0.0227**	-0.0227**	-0.0024
	(0.0016)	(0.0035)	(0.0025)	(0.0035)	(0.0030)
ln_market_value_of_assets	0.0143**	0.0143**	0.0143**	0.0143**	0.0303**
	(0.0008)	(0.0025)	(0.0015)	(0.0025)	(0.0011)
market_to_book_assets	-0.0543**	-0.0543**	-0.0543**	-0.0543**	-0.0545**
	(0.0013)	(0.0034)	(0.0052)	(0.0033)	(0.0033)
profits_to_sales	-0.1077**	-0.1077**	-0.1077	-0.1077**	-0.0031
	(0.0138)	(0.0395)	(0.0704)	(0.0376)	(0.0287)
tangible_assets	0.0692**	0.0692**	0.0692**	0.0692**	0.0886**
	(0.0068)	(0.0189)	(0.0169)	(0.0185)	(0.0102)
advertising_to_sales	-0.1557**	-0.1557	-0.1557*	-0.1557	-0.0521
	(0.0382)	(0.0917)	(0.0605)	(0.0911)	(0.0559)
rd_to_sales	-0.3880**	-0.3880**	-0.3880**	-0.3880**	-0.7087**
	(0.0554)	(0.1325)	(0.0564)	(0.1315)	(0.0830)
rd_positive	0.0048	0.0048	0.0048	0.0048	0.0188**
	(0.0031)	(0.0083)	(0.0080)	(0.0082)	(0.0036)
R-squared	0.2028	0.2028	0.2028	0.2028	0.6370
Coefficient estimates	OLS	OLS	OLS	OLS	FM
Standard errors	White	CL - F	CL - T	CL - F&T	FM

Note: This table presents the regression results for different model specifications. Column I shows the results using OLS with White standard errors. Columns II, III, and IV show the results using OLS with clustered standard errors by firm, time, and both firm and time, respectively. Column V presents the results using the Fama-MacBeth method. The dependent variable is the market debt ratio. The independent variables include log(firm_age), log(market value of assets), market_to_book_assets, profits_to_sales, tangible_assets, advertising_to_sales, rd_to_sales, and rd_positive. Standard errors are reported in parentheses. ** indicates significance at the 1% level, and * indicates significance at the 5% level.

Table 3: Unobserved heterogeneity

	OLS	AdjY	AvgE	GFE
ln_firm_age	-0.0230***	0.0003	-0.0006	-0.0063
	(0.0035)	(0.0010)	(0.0010)	(0.0047)
ln_market_value_of_assets	0.0128***	0.0013***	0.0017***	0.0134***
	(0.0024)	(0.0005)	(0.0005)	(0.0041)
market_to_book_assets	-0.0547***	-0.0117***	-0.0132***	-0.0352***
	(0.0033)	(0.0012)	(0.0013)	(0.0031)
profits_to_sales	-0.1050***	-0.0207**	-0.0237***	-0.0992**
	(0.0394)	(0.0083)	(0.0083)	(0.0432)
tangible_assets	0.0764***	-0.0041	-0.0012	-0.0278
	(0.0184)	(0.0034)	(0.0034)	(0.0346)
advertising_to_sales	-0.1296	0.0728***	0.0655***	-0.1394
	(0.0908)	(0.0220)	(0.0223)	(0.1168)
rd_to_sales	-0.3548***	0.0237	0.0101	-0.1503
	(0.1255)	(0.0347)	(0.0350)	(0.2464)
rd_positive	0.0053	0.0026	0.0027	0.0118
	(0.0083)	(0.0019)	(0.0019)	(0.0126)
Observations	10094	10094	10094	10094
R-squared	0.1827	0.0224	0.7301	0.0721

Note: This table presents different model specifications examining unobserved heterogeneity. Columns represent OLS, Adjusted Y (AdjY), Average Effects (AvgE), and Group Fixed Effects (GFE). The dependent variable is the market debt ratio. Independent variables include log(firm age), log(market value of assets), market-to-book assets, profits-to-sales, tangible assets, advertising-to-sales, R&D-to-sales, and an indicator for positive R&D. Standard errors are in parentheses. *** indicates significance at the 1% level, ** at 5%.

Table 4: Replication of Table 7 of Petersen (2009)

	I	II	III	IV	V
ln_market_value_of_assets	0.0004	0.0004	0.0004	0.0004	0.0197**
	(0.0004)	(0.0014)	(0.0020)	(0.0014)	(0.0010)
ln_firm_age	0.0006	0.0006	0.0006	0.0006	0.0189**
	(0.0008)	(0.0022)	(0.0046)	(0.0022)	(0.0018)
profits_to_sales	0.1725**	0.1725**	0.1725**	0.1725**	0.2288**
	(0.0083)	(0.0253)	(0.0221)	(0.0248)	(0.0176)
tangible_assets	0.1616**	0.1616**	0.1616**	0.1616**	0.1994**
	(0.0033)	(0.0109)	(0.0231)	(0.0106)	(0.0080)
market_to_book_assets	-0.0776**	-0.0776**	-0.0776**	-0.0776**	-0.0627**
	(0.0009)	(0.0024)	(0.0020)	(0.0024)	(0.0022)
advertising_to_sales	-0.0560	-0.0560	-0.0560	-0.0560	-0.0165
	(0.0315)	(0.0879)	(0.0731)	(0.0878)	(0.0314)
rd_to_sales	-0.1566**	-0.1566	-0.1566	-0.1566	-0.5242**
	(0.0380)	(0.0938)	(0.0951)	(0.0932)	(0.0682)
rd_positive	-0.0172**	-0.0172**	-0.0172**	-0.0172**	-0.0035
	(0.0018)	(0.0051)	(0.0027)	(0.0050)	(0.0043)
R-squared	0.3090	0.3090	0.3090	0.3090	0.6988
Coefficient estimates	OLS	OLS	OLS	OLS	FM
Standard errors	White	CL - F	CL - T	CL - F&T	FM

Note: This table presents the regression results for different model specifications. Columns I–IV show OLS with White standard errors, clustered by firm, time, and both, respectively. Column V is Fama-MacBeth. The dependent variable is the market debt ratio. The independent variables include log(firm_age), log(market value of assets), market_to_book_assets, profits_to_sales, tangible_assets, advertising_to_sales, rd_to_sales, and rd_positive. Standard errors are in parentheses. ** indicates significance at the 1% level.

Table 5: Replication of Table 2 of Gormley and Matsa (2014)

	OLS	AdjY	AvgE	FE
Fixed Assets/Total Assets	0.264***	0.119***	0.204***	0.137***
	(0.037)	(0.025)	(0.022)	(0.043)
Ln(Sales)	0.009***	0.009***	0.009***	0.008***
	(0.001)	(0.000)	(0.000)	(0.001)
Return on Assets	-0.047***	-0.034***	-0.063***	-0.071***
	(0.005)	(0.004)	(0.004)	(0.005)
Z-score	-0.010***	-0.005***	-0.009***	-0.008***
	(0.000)	(0.000)	(0.000)	(0.000)
Market-to-book Ratio	0.013***	0.005***	0.015***	0.015***
	(0.001)	(0.001)	(0.000)	(0.001)
Observations	208262	208262	208262	208262
R2	0.30	0.13	0.37	0.26

Note: This table presents regression results for different model specifications. Columns show OLS, Adjusted Y (AdjY), Average Effects (AvgE), and Fixed Effects (FE). Standard errors are in parentheses. *** indicates significance at the 1% level.