

Problem Set: Linear Panel Data Models

Empirical Methods in Corporate Finance

Michael R. Roberts

August 16, 2010

I. Two-Period Panel Leverage Regressions

Using data from fiscal year-ends 1990 and 1991, consider the following model of corporate leverage:

$$\begin{aligned} \text{Leverage}_{it} &= \beta_0 + \beta_1 \text{Profitability}_{it} + \beta_2 \text{Tangibility}_{it} + \beta_3 \text{FirmSize}_{it} \\ &+ \beta_4 \text{MB}_{it} + a_i + v_t + \varepsilon_{it} \end{aligned} \quad (1)$$

where a_i and v_t are firm and year fixed effects and all other variables are as defined in the attachment to this homework. Exclude all financial firms and utilities, as well as foreign governments, international affairs non-operating establishments and companies not headquartered in the United States (see attachment).

1. Do you have a balanced panel? If not, why not? Describe the distribution of observation counts per firm. (I.e., how many firms have complete data for both 1990 and 1991, just 1990, just 1991?).
2. Compute summary statistics for the level and first difference of each variable in the regression. Then recompute these summary statistics after Winsorizing the level of each ratio at the lower and upper one percentiles of their distribution. Make sure to distinguish between the three types of variation: total, within, and between. Are the observation counts the same for each variable? If not, why not? Should we be concerned?
3. Estimate a pooled OLS regression (i.e., restricting $a_i = 0$) and a firm fixed effect regression using both within and first difference estimators (three separate regressions). Discuss and compare the results (i.e., direction and magnitude of estimated coefficients, statistical and economic significance, model fit, residual variance decomposition, etc.)

II. T-Period Panel Leverage Regressions

Using data from fiscal year-ends 1965 through 2005, inclusive, reconsider the model of corporate leverage in equation (1) above. As before, exclude all financial firms and utilities, as well as foreign governments, international affairs non-operating establishments and companies not headquartered in the United States (see attachment). Winsorize all ratios at the upper and lower one percentiles before answering the following questions.

1. What are the mean and median number of time series observations per firm?
2. Compute summary statistics for all of your regression variables in both levels and first differences. Make sure to distinguish between the three types of variation: total, within, and between.
3. Run a pooled OLS regression (i.e., restricting $a_i = 0$) and estimate four different sets of standard errors: OLS, heteroskedasticity robust, within firm dependence robust, and a combination of heteroskedasticity and within firm dependence robust.
4. Does including firm size as $\log(\text{assets})$ make sense here?
5. Estimate the leverage regression with the following estimators: (1) pooled OLS, (2) first difference, (3) fixed effects, and (4) random effects. All standard errors should be heteroskedasticity-consistent. For (1), be sure to correct the standard errors for clustering at the firm level.
6. Estimate an AR(1) model for the level of variable (y and x's) via pooled ols. For example,

$$Profitability_{it} = \alpha Profitability_{it-1} + u_{it} \quad (2)$$

Relate these results to those in the previous problem. That is, how is the persistence of each variable relevant for each panel estimator?

7. Run a Hausman test to determine whether random effects are appropriate.

III. Static LPDM Simulations

Consider the following model:

$$y_{it} = \beta_0 + \delta_2 d2_t + \dots + \delta_T dT_t + \beta_1 x_{it} + a_i + u_{it} \quad (3)$$

where $i = 1, \dots, N$ and $t = 1, \dots, T$ index cross-sectional (firms) and time-series (years) units.

We're going to simulate this model under a variety of assumptions to see what happens to the parameter estimates under different estimation strategies.

The following will remain constant throughout the exercise:

- u is i.i.d. standard normal across firm-years.
- a is i.i.d. standard normal across firms (remember this is constant within firms).
- $x_{it} = \gamma a_i + e_{it}$ where e is i.i.d. normal with mean 0 and SD = 0.25 across firm-years.
- u , a and e are mutually independent.
- $\beta_0 = 4.0, \beta_1 = 5.0$
- Each experiment will require 100 simulations. The results to be reported should be averages across these simulations. The point of the simulations is to ensure our results aren't an artifact of one "weird" draw.

1. Assume $\gamma = 0$, $\delta_2 = 1.5$, $N = 100$, and $T = 2$. Simulate data according to the parameters given above. Estimate the model in levels via pooled OLS. Repeat this process 100 times to get a distribution of parameter estimates. Construct a 95% confidence interval from the distribution of estimates. Does the true parameter values fall inside this interval?
2. Using the same assumptions on γ , δ_2 , N , and T as in (1), repeat the exercise only replacing pooled OLS estimates with first difference estimates. How do your results here compare with those from (1)?
3. Now assume $\gamma = 0.5$ (δ_2 , N , and T unchanged from above). Repeat exercises (1) and (2) and compare the pooled OLS results with the first difference estimators. How do they compare now? If there is a difference, explain?

4. Now assume $\gamma = -0.5$ (δ_2 , N , and T unchanged from above). Repeat exercises (1) and (2) and compare the pooled OLS results with the first difference estimators. How do they compare now? If there is a difference, explain?
5. Repeat the pooled OLS portion of the previous exercise for $N = 1000$, $N = 10,000$, and $N = 100,000$. What happens to the pooled OLS estimate as we ramp up the sample size in N ? Explain your results?
6. Repeat the pooled OLS portion of the previous exercise for $T = 20$, $T = 100$, and $T = 10,000$. What happens to the pooled OLS estimate as we ramp up the sample size in T ? Explain your results?
7. In empirical corporate (or even in much of asset pricing), which asymptotics make more sense: large N and small T , or vice versa?

IV. Dynamic LPDM

Using data from fiscal year-ends 1965 through 2005, inclusive, consider the following model of corporate leverage:

$$Leverage_{it} = \beta_0 + \beta_1 Leverage_{it-1} + a_i + v_t + \varepsilon_{it} \quad (4)$$

As before, exclude all financial firms and utilities, as well as foreign governments, international affairs non-operating establishments and companies not headquartered in the United States (see attachment). Winsorize all ratios at the upper and lower one percentiles before answering the following questions.

1. Estimate equation (4) via: pooled OLS, within estimation, first-difference estimation, GMM estimation (Arellano and Bond (1991), and System GMM estimation (Bond and Blundell (1998)). Discuss and compare your results. Also, be clear to state your assumptions regarding the GMM estimators (e.g., number of lags).
2. Include the four covariates (size, profitability, tangibility, and market-to-book) into the specification as strictly exogenous variables. Now repeat the estimation exercises from above. Discuss and compare your results. Also, be clear to state your assumptions regarding the GMM estimators (e.g., number of lags).
3. Focusing on the system GMM estimator, what happens to the estimates when you treat the firm characteristics as predetermined? Endogenous?
4. Experiment with the number of instruments (i.e., lags) for the lagged dependent variable and report the impact on your estimates.

I. Data

The data is from the annual Compustat database, FUNDA, and is located on WRDS at `"/wrds/comp/sasdata/na"`. In the investment and capital structure literatures, a variety of screens are frequently used to eliminate certain observations from the sample. A few of these screens include the following.

1. (`indfmt == "INDL" & datafmt == "STD" & popsrc == "D" & consol == "C"`): These conditions ensure that `gvkey-datadate` uniquely identify each observation. (`gvkey-fyear` is "almost" the unique identifier, but for 48 obs.)
2. (`year ≥ 1965`): Observations with year-ends greater than or equal to 1965. Prior to this year, selection issues become particularly severe in Compustat.
3. (`sic ≥ 0000 & sic ≤ 999`): Agriculture, Fishing & Hunting.
4. (`sic ≥ 4900 & sic ≤ 4999`): Utilities.
5. (`sic ≥ 6000 & sic ≤ 6999`): Financial Firms.
6. (`sic = 8888`): Foreign Governments.
7. (`sic ≥ 9000 & sic ≤ 9999`): International affairs & non-operating establishments.
8. (`gvkey != ""`): No Missing company indicators.
9. (`fyear != .`): No Missing fiscal year indicators.
10. (`fic == "USA"`): Only firms headquartered in the United States.
11. (`prcc.f = . & csho = .`): Nonmissing stock market data (price & shares outstanding).

The investment variable definitions come from Hennessy, Levy, & Whited 2007 (JFE). The capital structure variable definitions come from Lemmon, Roberts, and Zender (2008) (JF). (In the construction of the market-to-book ratio here, you must first set all missing observations for `pstkl` and `txditc` to zero.)

Investment Variables			
g	mb	= (at + (prcc.f * csho) - ceq - txdb) / at;	(Market-to-Book)
g	cf	= (ib + dp);	(Cash flow)
g	cf_k	= cf / ppent[_n-1];	(Cash flow / capital(t-1))
g	inv	= (capxv - sppe);	(Net Investment)
g	inv_k	= inv / ppent[_n-1];	(Investment / capital(t-1))

Capital Structure Variables			
g	td	= dlc + dltd;	(Total Debt)
g	bl	= td / at;	(Book leverage)
g	ml	= td / (td + (prcc.f * csho));	(Market leverage)
g	prof_a	= oibdp / at;	(Profitability)
g	tang_a	= ppent / at;	(Tangibility)
g	me	= prcc.f * csho;	(Market equity)
g	mk2bk	= (me + dlc + dltd + pstkl + txditc) / at;	(Market-to-book)
g	loga	= log(at);	(Log(assets))
g	logsale	= log(sale);	(Log(sales))
g	zscore	= (3.3 * pi + sale + 1.4 * re + 1.2 * (act - lct)) / at;	(Altman's unlevered Z-score)
