

Estimating Standard Errors in Finance Panel Data Sets: Comparing Approaches

Mitchell A. Petersen

The Review of Financial Studies , 2009
Presented by Jonghyeok Cho

02/03/2025

Contents

- 1 Introduction
- 2 Problem of SE in finance
- 3 Estimating Standard Errors in Finance Panel Data Sets
- 4 Number of Clusters
- 5 Double Clustering

1. Introduction

- What is the standard error?
 - ▶ The standard deviation of the sampling distribution of an estimated regression coefficient, representing the accuracy of the estimate.
 - ▶ In $Y = X\beta + \varepsilon$, $SE(\hat{\beta}) = \sqrt{\text{var}(\hat{\beta})}$
- It is important to get accurate standard errors to avoid faulty estimation.
 - Since $t = \frac{\hat{\beta}}{SE(\hat{\beta})}$, we will reject null hypotheses too often if we underestimate standard errors. It means that it makes us mistakenly believe there is 'statistical significance'.

2. Problem of SE in finance

Homoskedasticity

- First, we have to determine homoskedasticity versus heteroskedasticity.
 - ▶ Homoskedasticity assumes that variance of the residuals does not depend on the X ; $\text{var}(\hat{\varepsilon}_i) = \hat{\sigma}^2$
 - ▶ Heteroskedasticity assumes that variance of the residuals depends on the X ; $\text{var}(\hat{\varepsilon}_i) = \hat{\sigma}_i^2$
 - ▶ In $ROA = \alpha + \beta \text{debt} + \gamma \text{size} + \varepsilon$, residuals can depend on the size of the firm.
- We can use 'robust standard errors' in stata, which allows for heteroskedasticity assumption. ; reg y x, vce(robust) or reg y x, r
- Since the robust SE can be smaller than the classical SE in small samples, compare the robust SE to the classical SE and take maximum of the two. Angrist-Pischke argue that this will tend to be closer to the true SE in samples that exhibit heteroskedasticity.

2. Problem of SE in finance

Independence

- Secondly, we have to recognize that 'Classical' or 'robust' SE calculation in OLS are based on 'independence' assumption.
 $\rightarrow Cov(\varepsilon_{it}, \varepsilon_{it-k}) \neq 0$ and $Cov(\varepsilon_{jt}, \varepsilon_{it}) \neq 0$
- OLS standard errors are unbiased (converge to true) when the errors are independent and identically distributed. In practice, independence is unrealistic in finance literature. So, researchers have tried to address possible biases in the SE.

2. Problem of SE in finance

Independence

- In 2001-2004, 42% of the papers published in JF, JFE and RFS did not adjust the standard errors for possible dependence in the residuals (Petersen, 2009).
→ Petersen shows a difference in performance of the different methods and gives the researchers guidance on how to use these methods.

3. Estimating Standard Errors in Finance Panel Data Sets

Main finding

- 1 In the presence of a firm effect [e.g., $\text{Cov}(X_{it}, X_{it-k}) \neq 0$], SEs are unbiased when estimated clustered by firm and produce correctly sized confidence intervals. And it is still unbiased whether firm effect permanent or temporary.
- 2 In the presence of a time effect [e.g., $\text{Cov}(X_{jt}, X_{it}) \neq 0$], Fama-MacBeth produced unbiased SEs. When there are a sufficient number of clusters, clustered SEs by time are also unbiased.
- 3 When both a firm and a time effect are present in the data, researchers can address one parametrically (e.g., time dummies) and then estimate SEs clustered on the other dimension. Alternatively, researchers can cluster on multiple dimensions when there are sufficient number of clusters in each dimension.

3. Estimating Standard Errors in Finance Panel Data Sets

Presence of fixed firm effect

- The standard regression for a panel data set is $Y_{it} = X_{it}\beta + \varepsilon_{it}$.
- When errors are i.i.d., the asymptotic variance of the estimated coefficient is $AVar[\hat{\beta}_{OLS} - \beta] = \frac{1}{N}(T\sigma_x^2\sigma_\varepsilon^2)(T\sigma_x^2)^{-2} = \frac{\sigma_\varepsilon^2}{\sigma_x^2 NT}$.

3. Estimating Standard Errors in Finance Panel Data Sets

Presence of fixed firm effect

- In relaxing the assumption of independent errors, assume that the data have an unobserved firm effect that is fixed. Thus the residuals consist of a firm-specific component and an idiosyncratic component that is unique to each observation.
 - ▶ $\varepsilon_{it} = \gamma_i + \eta_{it}$ and $X_{it} = \mu_i + \nu_{it}$
- It is called a firm effect when the residuals may be correlated across years for a given firm.
 - ▶ $\text{Corr}(X_{it}, X_{is}) = \rho_x$ and $\text{Corr}(\varepsilon_{it}, \varepsilon_{is}) = \rho_\varepsilon$
 - ▶ $A\text{Var}[\hat{\beta}_{OLS} - \beta] = \frac{\sigma_\varepsilon^2}{\sigma_x^2 NT} (1 + (T - 1)\rho_x \rho_\varepsilon)$
- Therefore, when the data have a fixed firm effect, the OLS standard errors will underestimate the true standard error if and only if both correlation are nonzero.

3. Estimating Standard Errors in Finance Panel Data Sets

Presence of fixed firm effect

- Simulated a panel data set and then estimated the slope coefficient and its standard error. By doing this 5,000 times.
- Each simulated data set contains five thousand observations.
- Across different simulations, he altered the fraction of the variance in the independent variable that is due to the firm effect. This fraction ranges from 0% to 75% in 25% increments. The same thing was done for the residual.

3. Estimating Standard Errors in Finance Panel Data Sets

Presence of fixed firm effect

Table 1
Estimating standard errors with a firm effect OLS and clustered standard errors

		Source of independent variable volatility			
		0%	25%	50%	75%
Source of residual volatility	Avg(β_{OLS})				
	Std(β_{OLS})				
	Avg(SE_{OLS})				
	% Sig(t_{OLS})				
	Avg(SE_C)				
	Avg(SEC)				
Source of residual volatility	0%	1.0004	1.0006	1.0002	1.0001
		0.0286	0.0288	0.0279	0.0283
		0.0283	0.0283	0.0283	0.0283
		[0.0098]	[0.0088]	[0.0094]	[0.0094]
		0.0283	0.0282	0.0282	0.0282
		[0.0108]	[0.0092]	[0.0096]	[0.0098]
	25%	1.0004	0.9997	0.9999	0.9997
		0.0287	0.0353	0.0403	0.0468
		0.0283	0.0283	0.0283	0.0283
		[0.0116]	[0.0348]	[0.0678]	[0.1174]
		0.0283	0.0353	0.0411	0.0463
		[0.0120]	[0.0064]	[0.0112]	[0.0118]
	50%	1.0001	1.0002	1.0007	0.9993
		0.0289	0.0414	0.0508	0.0577
		0.0283	0.0283	0.0283	0.0283
		[0.0124]	[0.0770]	[0.1534]	[0.2076]
		0.0282	0.0411	0.0508	0.0590
		[0.0128]	[0.0114]	[0.0088]	[0.0102]
	75%	1.0000	1.0004	0.9995	1.0016
		0.0285	0.0459	0.0594	0.0698
		0.0283	0.0283	0.0283	0.0283
		[0.0128]	[0.1090]	[0.2230]	[0.2906]
		0.0282	0.0462	0.0589	0.0693
		[0.0128]	[0.0114]	[0.0094]	[0.0112]

3. Estimating Standard Errors in Finance Panel Data Sets

Presence of fixed firm effect

- An alternative way to estimate the regression coefficients and standard errors when the residuals are not independent is the Fama-MacBeth approach (Fama and MacBeth, 1973).
- The Fama-MacBeth standard errors are biased in exactly the same way as the OLS estimates.
 - ▶ $AVar[\hat{\beta}_{OLS} - \beta] = \frac{\sigma_{\varepsilon}^2}{\sigma_x^2 NT} (1 + (T - 1)\rho_x \rho_{\varepsilon})$
- In both cases, the magnitude of the bias is a function of the serial correlation of both the independent variable and the residual.

3. Estimating Standard Errors in Finance Panel Data Sets

Presence of fixed firm effect

Table 2
Estimating standard errors with a firm effect Fama-MacBeth standard errors

		Source of independent variable volatility			
		0%	25%	50%	75%
Ave(R_{FM}) Std(β_{FM}) Avg(SE_{FM}) % Sig(1_{FM})					
Source of residual volatility	0%	1.0004	1.0006	1.0002	1.0001
		0.0287	0.0288	0.0280	0.0283
		0.0276	0.0276	0.0277	0.0275
		[0.0288]	[0.0304]	[0.0236]	[0.0294]
	25%	1.0004	0.9997	0.9998	0.9997
		0.0288	0.0354	0.0403	0.0469
		0.0275	0.0268	0.0259	0.0250
		[0.0336]	[0.0758]	[0.1202]	[0.1918]
	50%	1.0000	1.0002	1.0007	0.9993
		0.0289	0.0415	0.0509	0.0578
		0.0276	0.0259	0.0238	0.0219
		[0.0330]	[0.1264]	[0.2460]	[0.3388]
	75%	1.0000	1.0004	0.9995	1.0016
		0.0286	0.0460	0.0595	0.0699
		0.0277	0.0248	0.0218	0.0183
		[0.0310]	[0.1778]	[0.3654]	[0.4994]

- Same with OLS, Moving down the diagonal of Table 2 from upper left to bottom right, the true standard error increases but the standard error estimated by Fama-MacBeth actually shrinks.

3. Estimating Standard Errors in Finance Panel Data Sets

Presence of fixed firm effect

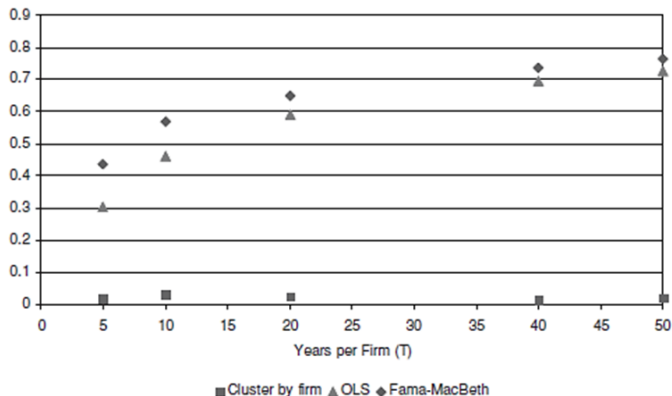


Figure 3
Bias in estimated standard errors as a function of years per cluster

- The bias in OLS and Fama-Macbeth standard errors is highly sensitive to the number of time periods (years) used in the estimation.

3. Estimating Standard Errors in Finance Panel Data Sets

Presence of Time effect

- How the techniques work in the presence of a time effect, data sets containing only a time effect were generated (observations on different firms within the same year are correlated).
 - ▶ $\varepsilon_{it} = \delta_t + \eta_{it}$ and $X_{it} = \zeta_t + \nu_{it}$
- Simulated a panel data set (5,000 obs) and then estimated the slope coefficient and its standard error. By doing this 5,000 times.
- Across different simulations, he altered the fraction of the variance in the independent variable that is due to the time effect. This fraction ranges from 0% to 75% in 25% increments. The same thing was done for the residual.

3. Estimating Standard Errors in Finance Panel Data Sets

Presence of Time effect

Table 3
Estimating standard errors with a time effect: OLS and clustered standard errors

		Source of independent variable volatility			
		0%	25%	50%	75%
Avg(β_{OLS})					
Std(β_{OLS})					
Avg(SE_{OLS})					
% Sig(Γ_{OLS})					
Avg(SE_c)					
% Sig(Γ_c)					
Source of residual volatility	0%	1.0004	1.0002	1.0006	0.9994
		0.0286	0.0291	0.0293	0.0314
		0.0283	0.0288	0.0295	0.0306
		[0.0098]	[0.0094]	[0.0088]	[0.0114]
		0.0277	0.0276	0.0275	0.0270
		[0.0330]	[0.0304]	[0.0348]	[0.0520]
	25%	1.0006	1.0043	0.9962	0.9996
		0.0284	0.1490	0.2148	0.2874
		0.0279	0.0284	0.0289	0.0300
		[0.0114]	[0.6064]	[0.7270]	[0.7874]
		0.0268	0.1297	0.1812	0.2305
		[0.0320]	[0.0360]	[0.0506]	[0.0736]
	50%	0.9996	0.9997	0.9919	1.0079
		0.0276	0.2138	0.3015	0.3986
		0.0274	0.0278	0.0282	0.0292
		[0.0100]	[0.7298]	[0.8096]	[0.8536]
		0.0258	0.1812	0.2546	0.3248
		[0.0294]	[0.0458]	[0.0596]	[0.0756]
	75%	1.0002	0.9963	0.9970	0.9908
		0.0273	0.2620	0.3816	0.4927
		0.0267	0.0271	0.0276	0.0284
		[0.0110]	[0.7994]	[0.8586]	[0.8790]
		0.0244	0.2215	0.3141	0.3986
		[0.0322]	[0.0402]	[0.0588]	[0.0768]

- The clustered standard errors are much more accurate, but unlike the results with the firm effect, they underestimate the true standard error. The problem arises due to the limited number of clusters.

3. Estimating Standard Errors in Finance Panel Data Sets

Presence of Time effect

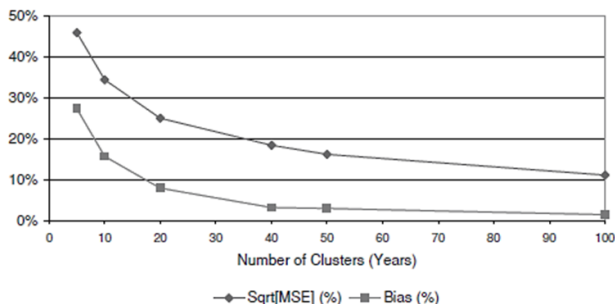


Figure 5
True standard errors and clustered standard errors as a function of cluster size (T)

- Data sets of five thousand observations were simulated with the number of years (or clusters) ranging from 5 to 100. In all of the simulations, $\rho_x = \rho_\varepsilon = 0.25$.
- $Bias = SE_c - \hat{SE}$
- $MSE = E[(\hat{SE} - SE_{true})^2]$

3. Estimating Standard Errors in Finance Panel Data Sets

Presence of Time effect

Table 4
Estimating standard errors with a time effect: Fama-MacBeth standard errors

		Source of independent variable volatility			
		0%	25%	50%	75%
Avg(β_{FM}) Std(β_{FM}) Avg(SE_{FM}) % Sig(1_{FM})					
Source of residual volatility	0%	1.0004	1.0004	1.0007	0.9991
		0.0287	0.0331	0.0396	0.0573
		0.0278	0.0318	0.0390	0.0553
		[0.0310]	[0.0312]	[0.0252]	[0.0338]
	25%	1.0005	1.0003	1.0006	0.9999
		0.0252	0.0284	0.0343	0.0496
		0.0239	0.0276	0.0338	0.0480
		[0.0376]	[0.0296]	[0.0284]	[0.0294]
	50%	1.0000	1.0002	1.0006	1.0007
		0.0200	0.0231	0.0280	0.0394
		0.0195	0.0227	0.0276	0.0387
		[0.0254]	[0.0304]	[0.0272]	[0.0278]
	75%	1.0001	0.9996	1.0000	0.9999
		0.0142	0.0161	0.0200	0.0285
		0.0138	0.0159	0.0196	0.0276
		[0.0308]	[0.0302]	[0.0284]	[0.0300]

- Fama-MacBeth produces unbiased standard errors and correctly sized confidence intervals. This is not surprising since it was designed for just such a setting.

3. Estimating Standard Errors in Finance Panel Data Sets

Presence of Fixed Firm and Time effect

- If the data have only a firm effect, SEs clustered by firm are unbiased. On the other hand, only a time effect, the Fama-MacBeth is better than SEs clustered by time when there are few clusters(years) and equally good when the number of clusters is sufficiently large.
- If panel data sets have both fixed firm and time effect, there are two choices.
 - ① Since the time effect is fixed, the time dummies completely remove the correlation between observations in the same time period. Then, there is only a firm effect left and researchers can use SEs clustered by firm. But, The parametric approach only works when the dependence is correctly specified.
 - ② Cluster on two dimensions simultaneously (e.g., High beta firms will respond more to systematic shocks across year. since it is not solved by time dummy).

3. Estimating Standard Errors in Finance Panel Data Sets

Presence of Fixed Firm and Time effect

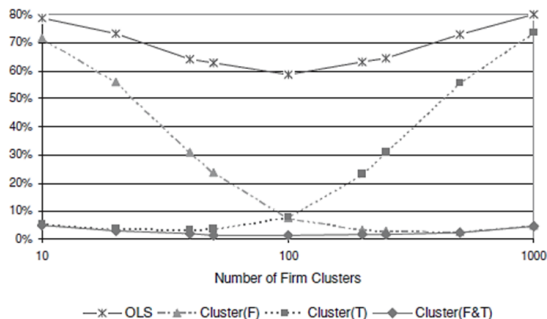


Figure 7

Rejection rates in the presence of a firm and a time effect for t -statistics clustered by firm, by time, or both

- $Var(\gamma) = Var(\delta) = Var(\eta)$ and $Var(\mu) = Var(\zeta) = Var(\nu)$
- The figure graphs the fraction of t -statistics that are statistically significant at the 1% level (greater than 2.58 in absolute value).
- Nine data sets are then simulated. The number of firms and time periods range from 10 to 1,000 so that the total number of observations is always 10,000.

3. Estimating Standard Errors in Finance Panel Data Sets

Presence of Temporary Firm effect

- The analysis thus far has assumed that the firm effect is fixed, but it may not always be true in the data. The dependence between residuals may decay as the time between them increases.
- A data structure is simulated that includes both a permanent component (a fixed firm effect) and a temporary component (nonfixed firm effect) that is assumed to be a first-order autoregressive process.

► $\eta_{it} = \zeta_{it}$ if $t = 1$, $\eta_{it} = \phi\eta_{it-1} + \sqrt{(1 - \phi^2)}$ if $t > 1$

3. Estimating Standard Errors in Finance Panel Data Sets

Presence of Temporary Firm effect

Table 5
Estimated standard errors with a nonfixed firm effect

Panel A: OLS and clustered standard errors

	I	II	III	IV
Avg($\hat{\rho}_{OLS}$)				
Std($\hat{\rho}_{OLS}$)				
Avg($\hat{SE}_{OLS}$)				
% Sig(T_{OLS})				
Avg($\hat{SE}_c$)				
% Sig(T_c)				
ρ_X/ρ_e	0.50/0.50	0.00/0.00	0.25/0.25	0.60/0.35
ϕ_X/ϕ_e	0.00/0.00	0.90/0.90	0.75/0.75	0.99/0.81
OLS	0.9994 0.0513 0.0283 [0.1578] 0.0508 [0.0114]	1.0001 0.0659 0.0283 [0.2746] 0.0668 [0.0086]	1.0009 0.0566 0.0283 [0.1996] 0.0569 [0.0104]	0.9991 0.0677 0.0253 [0.3302] 0.0670 [0.0098]
OLS with firm dummies	1.0007 0.0299 0.0298 [0.0096] 0.0298 [0.0100]	1.0003 0.0517 0.0298 [0.1382] 0.0516 [0.0098]	1.0013 0.0442 0.0298 [0.0802] 0.0441 [0.0092]	1.0046 0.1881 0.1101 [0.1288] 0.1886 [0.0108]

- Once the firm effects are included, the OLS SEs are unbiased (column I). If the firm effect decays over time, the firm dummies no longer fully capture the within-cluster dependence, and SEs are still biased. But, the clustered SEs are unbiased with and without the firm dummies.

3. Estimating Standard Errors in Finance Panel Data Sets

Summary

- If a firm exhibits correlation across years (whether permanent or temporary), we must use clustered standard errors by firm to avoid underestimating the standard errors.
- If observations on different firms within the same year are correlated, use Fama-Macbeth SEs or clustered SEs by year (only sufficient sample case).
- If observations have both fixed effects, cluster on two dimensions simultaneously(=double clustering).
- In stata, `reg y x, vce(clustered firm)`

4. Number of Clusters

- The paper dealt with previously leads to question about the number of enough clusters to get unbiased estimator.
 - ▶ Too small clusters can cause downward bias.
- There is no absolute method.
 - ▶ Simulations of Hansen (JoE, 2007) suggest that 10 clusters is enough when using clusters to account for serial correlation.
 - ▶ Bertrand, et al (QJE 2004) suggest 50 clusters are sufficient in DID setting.

5. Double Clustering

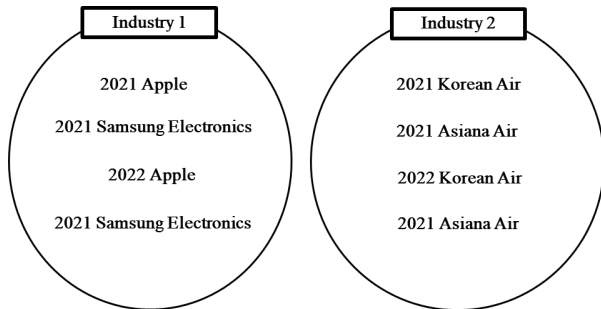
Thomson(JFE,2011)

- Thomson(JFE,2011) derives easy-to-compute formulas for SEs that cluster by both firm and time, suggesting simultaneously clustering by firms and time leads to significantly more accurate inference in finance panels.
- The double-clustering SEs are reliable in data sets with at least 25 firms observed over 25 time periods. However, they said that when there are far more firms than time periods or regressor vary by time but not by firm, it can't make a big difference.
- We can use a double-clustering easily in stata.; `reg y x, vce(cluster gvkey year)`

5. Double Clustering

Practice

- Industry clustering

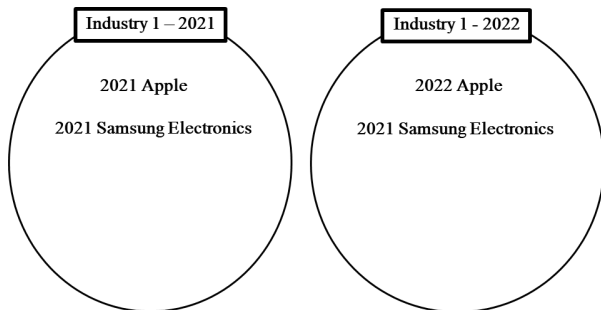


- Allows that regressors or errors of firms within the same industry are correlated.
- Allows that regressors or errors of the same firms across different years are correlated.
- Not allows correlation between firms in different industries.

5. Double Clustering

Practice

- Industry-Year clustering

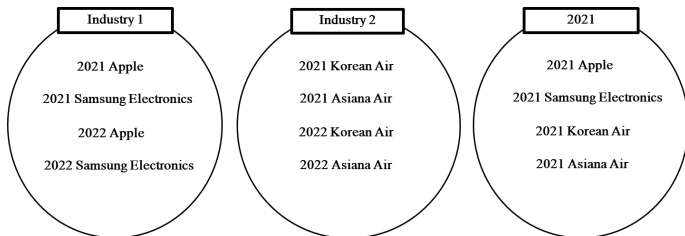


- Allows that regressors or errors of firms within the same industry and the same year are correlated.
- However, even if firms belong to the same industry, they are uncorrelated across different years.

5. Double Clustering

Practice

- Double clustering: Industry and Year



- Allows that regressors or errors of firms within the same year are correlated.