Journal of Financial Economics 99 (2011) 1–10

Contents lists available at ScienceDirect

Journal of Financial Economics

journal homepage: www.elsevier.com/locate/jfec

Simple formulas for standard errors that cluster by both firm and time ☆

Samuel B. Thompson

Arrowstreet Capital L.P., The John Hancock Tower, 200 Clarendon Street 30th Floor, Boston, MA 02116, USA

ARTICLE INFO

Article history:
Received 13 July 2006
Received in revised form
15 May 2009
Accepted 7 July 2009
Available online 14 October 2010

JEL classification:
C23
G12
G32

Keywords:
Cluster standard errors
Panel data
Finance panel data

ABSTRACT

When estimating finance panel regressions, it is common practice to adjust standard errors for correlation either across firms or across time. These procedures are valid only if the residuals are correlated either across time or across firms, but not across both. This paper shows that it is very easy to calculate standard errors that are robust to simultaneous correlation along two dimensions, such as firms and time. The covariance estimator is equal to the estimator that clusters by firm, plus the estimator that clusters by time, minus the usual heteroskedasticity-robust ordinary least squares (OLS) covariance matrix. Any statistical package with a clustering command can be used to easily calculate these standard errors.

1. Introduction

A typical finance panel data set contains observations on multiple firms across multiple time periods. Although OLS standard errors will be consistent as long as the regression residuals are uncorrelated across both firms and months, such uncorrelatedness is unlikely to hold in a finance panel. For example, market-wide shocks will induce correlation between firms at a moment in time, and persistent firm-specific shocks will induce correlation across time. Furthermore, persistent common shocks, like business cycles, can induce correlation between different firms in different years.

A number of techniques are available for adjusting standard errors for correlation along a single dimension. Fama and MacBeth (1973) propose a sequential time-series of cross-sections procedure that produces standard errors robust to correlation between firms at a moment in time. Huber (1967) and Rogers (1983) show how to compute "clustered" standard errors which are robust either to correlation across firms at a moment in time or to correlation within a firm across time. None of these techniques correctly adjusts standard errors for simultaneous correlation across both firms and time. If one clusters by firm, observations may be correlated within each firm, but must be independent across firms. If one clusters by time, observations may be correlated within each time period, but correlation across time periods is ruled out.

This paper describes a method for computing standard errors that are robust to correlation along two dimensions. To make the discussion concrete, we call one

☆ I thank Eugene Fama, Megan MacGarvie, Antti Petajisto, Mitchell Petersen and Christopher Polk for helpful comments. The comments of two anonymous referees led to revisions which significantly improved the paper. I owe special thanks to John Campbell and Tuomo Vuolteenaho. The idea of computing a forward-looking Herfindahl-Hirschman Index (used in the empirical application) was communicated to me by Tuomo Vuolteenaho.

E-mail address: sambthompson@gmail.com

0304-405X/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.jfineco.2010.08.016

dimension time, and the other firm, but the results trivially generalize to any two-dimensional panel data setting. In addition, these standard errors are easy to compute. In the simplest case, we have firm and time effects, but no persistent common shocks. In this case, the variance estimate for an OLS estimator $\widehat{\beta}$ is

$$\widehat{\mathrm{Var}}(\widehat{\beta}) = \widehat{\mathbf{V}}_{firm} + \widehat{\mathbf{V}}_{time,0} - \widehat{\mathbf{V}}_{white,0},$$

where $\widehat{\mathbf{V}}_{firm}$ and $\widehat{\mathbf{V}}_{time,0}$ are the estimated variances that cluster by firm and time, respectively, and $\widehat{\mathbf{V}}_{white,0}$ is the usual heteroskedasticity-robust OLS variance matrix (White, 1980).[1] Thus, any statistical package with a clustering command (e.g., STATA) can be used to easily calculate these standard errors. The paper also provides valid standard errors for the more complicated case which allows for persistent common shocks.

This paper also discusses the pros and cons of double-clustered standard errors. I analyze the standard error formulas using the familiar trade-off between bias and variance. The various standard error formulas are estimates of true, unknown standard errors. The more robust formulas have less bias, but more estimation variance. The lower bias improves the performance of test statistics, but the increased variance can lead to size distortions. I use Jensen's inequality to show that, when sample sizes are small, the more robust standard errors lead us to find statistical significance even when it does not exist.

When is the bias reduction likely to be important? I argue that double clustering is likely to be most helpful in data sets with the following characteristics: the regression errors include significant time and firm components, the regressors themselves include significant firm and time components, and the number of firms and time periods is not too different. So, if the regressors vary by time but not by firm, then clustering by time may be good enough, and double clustering may not make a large difference. If there are far more firms than time periods, clustering by time eliminates most of the bias unless within-firm correlations are much larger than within-time period correlations.

I also point out special considerations related to persistent common shocks. Correcting for correlations between different firms in different time periods involves estimating autocovariances between residuals. As Hurwicz (1950) and many subsequent authors have shown, autocovariance estimates are biased downward. Thus, standard errors that correct for persistent common shocks will tend to be biased downward. Eliminating the bias requires a large number of time periods.

I use a Monte Carlo to evaluate how large sample sizes must be in practice. When I apply pure double clustering, and do not adjust for persistent common shocks, the standard errors are reliable in data sets with at least 25 firms observed over 25 time periods. When I correct for persistent common shocks, the number of time periods should be greater than 50.

This leads to reasonably simple advice for applied researchers. Double clustering is worth doing because it is an easy robustness check, and the standard error estimates are accurate in small samples. However, we should not expect it to make a big difference in all data sets, especially when there are far more firms than time periods. I do not make as strong a case for adjusting for persistent common shocks. The standard error formulas are a bit more complicated, and a larger number of time periods is needed for the estimates to be accurate.

2. Firm effects, time effects, and persistent common shocks

Consider the panel regression

$$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it}. \tag{1}$$

y_{it} is the dependent variable, ε_{it} is the error term, $\mathbf{x}_{it}$ is the covariate vector, and $\boldsymbol{\beta}$ is the coefficient vector. We have $i=1,\ldots,N$ firms observed over $t=1,\ldots,T$ time periods. More generally, index i could refer to any unit of observation, such as an industry- or country-level observation, and t could refer to any other unit. I write in terms of firms and time periods because it makes the discussion more concrete. The errors may be heteroskedastic, but must have zero conditional mean, so $\mathrm{E}(\varepsilon_{it}|\mathbf{x}_{it}) = 0$. We make the following assumptions about the correlations between errors.

Assumption 1. *Firm effects*: The errors may exhibit *firm effects*, meaning that errors may have arbitrary correlation across time for a particular firm: $\mathrm{E}(\varepsilon_{it}\varepsilon_{ik}|\mathbf{x}_{it},\mathbf{x}_{ik}) \neq 0$ for all $t \neq k$.

Assumption 2. *Time effects*: The errors may exhibit *time effects*, meaning that errors may have arbitrary correlation across firms at a moment in time: $\mathrm{E}(\varepsilon_{it}\varepsilon_{jt}|\mathbf{x}_{it},\mathbf{x}_{jt}) \neq 0$ for $i \neq j$.

Assumption 3. *Persistent common shocks*: The errors may exhibit *persistent common shocks*, meaning that we allow some correlation between different firms in different time periods, but these shocks die out over time, and may be ignored after L periods. So $\mathrm{E}(\varepsilon_{it}\varepsilon_{jk}|\mathbf{x}_{it},\mathbf{x}_{jk}) = 0$ if $i \neq j$ and $|t-k| > L$.

To understand the difference between time effects, firm effects, and persistent common shocks, consider the following data-generating process:

$$\varepsilon_{it} = \boldsymbol{\theta}_i'\mathbf{f}_t + \eta_{it} + u_{it},$$

$$\eta_{it} = \varphi\eta_{i,t-1} + \varsigma_{it}, \quad \eta_{i0} = 0. \tag{2}$$

$\mathbf{f}_t$ is a vector of random factors common to all firms, and $\boldsymbol{\theta}_i$ is a vector of factor loadings specific to firm i. u_{it} and ς_{it} are random shocks, uncorrelated across both firm and time. The η_{it} term generates firm effects—shocks specific to firm i. $\boldsymbol{\theta}_i'\mathbf{f}_t$ generates both time effects and persistent common shocks. When $\mathbf{f}_t$ is uncorrelated across time, we have time effects but no persistent common shocks—firms are correlated with one another at a moment in time, but

[1] The double-clustering problem was also solved in Cameron, Gelbach, and Miller (2006). I was unaware of their paper while working on these results. Their paper was made available on the Web at roughly the same time as this one.

different firms in different time periods are uncorrelated. When $\mathbf{f}_t$ is persistent, we have both time effects and persistent common shocks. We assume that the autocorrelations for $\mathbf{f}_t$ disappear after L months.[2]

2.1. Examples

The assumptions cover many interesting corporate finance applications. Consider a capital structure regression as in Petersen (2009), where the dependent variable is the ratio of firm debt to assets. The residual probably includes a firm-specific effect (our η_{it} term) as well as common persistent business-cycle shocks that affect all firms (our f_t term). Later in this paper, I consider profitability regressions as in Fama and French (2000), where the dependent variable is profitability measured as the ratio of firm earnings to book value of equity. The residual probably includes firm-specific components, as well as common components that vary over time.

Other examples come from the literature that links a country's growth rate of output to its financial development. A country's growth rate is probably influenced by country-specific and business-cycle shocks. Rajan and Zingales (1998) run regressions at the country and industry level, and use country and industry dummies to control for common effects. Papers such as Larrain (2006) and Li, Morck, Yang, and Yeung (2004) estimate country-level panel regressions.

For an asset-pricing example, consider predictive panel regressions with overlapping returns. The dependent variable y_{it} is a J-period overlapping return, so $y_{it}=\sum_{k=1}^{J} R_{i,t+k}$, where $R_{i,t}$ is the return on company i's stock in month t. The regression errors will likely contain shocks common to many stocks, and the overlapping structure of the dependent variable will induce correlations across different firms in different time periods. Thus, is it likely that we will have time effects and persistent common shocks, but we may be able to rule out firm effects.

Predictive regressions with overlapping returns have been well studied in the univariate case. For example, Hansen and Hodrick (1980) show how to calculate correct standard errors when predicting a univariate time-series of exchange rates. The panel regression case is less well understood. Cohen, Polk, and Vuolteenaho (2003) is an example of a paper that handles the problem carefully—the formulas in this paper generalize and simplify their calculations.

2.2. Alternative approaches

This paper provides standard error formulas that correctly handle these examples. In order to better understand the usefulness of this result, let us consider other approaches that an applied researcher might take. One approach would be to use the usual standard errors that do not adjust for correlation between observations. Petersen (2009) and many other have shown that can lead to standard errors that are too small. Small standard errors lead to large t-statistics, and the researcher will see statistical significance even when it does not exist.

Another approach is to cluster along a single dimension. Similarly, we could use the standard errors of Fama and MacBeth (1973), since they also solve the single-clustering problem [see Petersen (2009) for further explanation]. Again, this can lead to understated standard errors. Consider an application to model firm profitability. We might cluster by time, meaning that we allow firms to be correlated with one another at a moment in time. This will ignore persistent firm-specific effects. The residual may contain unobserved components that cause one company to be persistently more profitable than others.

One way to simultaneously handle firm and time effects is to use firm and time dummies. For example, we could cluster the standard errors by time and include firm fixed effects (e.g., we could include firm-specific dummy variables in the regression). This is a sensible procedure that will work well in many cases. However, fixed effects will not handle many relevant forms of correlated errors. In our example data-generating process (Eq. 2), the time effect has the factor structure $\boldsymbol{\theta}_i'\mathbf{f}_t$ and the firm effect follows the autoregressive process η_{it}. Time dummies will not correctly model the factor structure if the loadings $\boldsymbol{\theta}_i$ vary across firms, and firm dummies will not correctly model the autoregressive process.

Another limitation with firm or time fixed effects is that they limit the kinds of covariates that can be included. If we use time dummies, we cannot include macroeconomic variables in the regression, since they are collinear with the dummies. Similarly, dummies can significantly increase the standard errors when the covariate does not vary much along a dimension. For example, consider a regression where the covariate is the yield on a firm's long-term debt. If the firms in our sample have similar credit quality, this covariate may vary significantly across time, but may not vary much across firms at a moment in time. While it is possible to include time dummies in this regression, they will be highly correlated with the covariate and therefore, will cause the standard errors to increase.

3. Standard error formulas

What is the variance of the OLS estimator? The estimator satisfies

$$\widehat{\beta}-\beta=\mathbf{H}^{-1}\left[\sum_{i,t}\mathbf{u}_{it}\right],$$

with $\mathbf{u}_{it}=\mathbf{x}_{it}\varepsilon_{it}$ and $\mathbf{H}=\sum_{i,t}\mathbf{x}_{it}\mathbf{x}_{it}'$. In large samples, the estimator variance can be approximated by $\mathbf{H}^{-1}\mathbf{G}\mathbf{H}^{-1}$, where $\mathbf{G}=\mathrm{Var}[\sum_{i,t}\mathbf{u}_{it}]$. The term $\mathbf{G}$ may be written as

$$\mathbf{G}=\sum_{i,j,t,k}\mathrm{E}(\mathbf{u}_{it}\mathbf{u}_{jk}').$$

[2] More generally, shocks to $\mathbf{f}_t$ could decay slowly but not completely disappear after L periods. For example, $\mathbf{f}_t$ could follow a first-order autoregressive process. While this would violate the assumption, I assume that after some time the correlation between shocks is small enough that it can be ignored. Autoregressive processes could be handled by allowing the lag length L to grow with the sample size (see, for example, Newey and West, 1987).

Under the error assumptions, we can simplify the formula as

$$\mathbf{G} = \mathbf{G}_{firm} + \mathbf{G}_{time,0} - \mathbf{G}_{white,0} + \sum_{l=1}^{L}(\mathbf{G}_{time,l} + \mathbf{G}'_{time,l}) - \sum_{l=1}^{L}(\mathbf{G}_{white,l} + \mathbf{G}'_{white,l}),$$

with

$$\mathbf{G}_{firm} \equiv \sum_{i} \mathrm{E}(\mathbf{c}_i \mathbf{c}_i'), \quad \mathbf{G}_{time,l} \equiv \sum_{t} \mathrm{E}(\mathbf{s}_t \mathbf{s}_{t+l}'),$$
$$\mathbf{G}_{white,l} \equiv \sum_{i,t} \mathrm{E}(\mathbf{u}_{it} \mathbf{u}_{i,t+l}').$$

$\mathbf{c}_i = \sum_t \mathbf{u}_{it}$ is the sum over all observations for firm i, and $\mathbf{s}_t = \sum_i \mathbf{u}_{it}$ is the sum over all observations for time t. The variance can be consistently estimated by $\widehat{\mathrm{Var}}(\hat{\beta}) = \hat{\mathbf{V}}_{firm} + \hat{\mathbf{V}}_{time,0} - \hat{\mathbf{V}}_{white,0} + \sum_{l=1}^{L}(\hat{\mathbf{V}}_{time,l} + \hat{\mathbf{V}}'_{time,l}) - \sum_{l=1}^{L}(\hat{\mathbf{V}}_{white,l} + \hat{\mathbf{V}}'_{white,l})$, where

$$\hat{\mathbf{V}}_{firm} \equiv \mathbf{H}^{-1} \sum_{i} \hat{\mathbf{c}}_i \hat{\mathbf{c}}_i' \mathbf{H}^{-1},$$
$$\hat{\mathbf{V}}_{time,l} \equiv \mathbf{H}^{-1} \sum_{t} \hat{\mathbf{s}}_t \hat{\mathbf{s}}_{t+l}' \mathbf{H}^{-1},$$
$$\hat{\mathbf{V}}_{white,l} \equiv \mathbf{H}^{-1} \sum_{t} \sum_{i} \hat{\mathbf{u}}_{it} \hat{\mathbf{u}}_{i,t+l}' \mathbf{H}^{-1}. \quad (3)$$

$\hat{\mathbf{u}}_{it} = \mathbf{x}_{it}\hat{\varepsilon}_{it}$, $\hat{\mathbf{c}}_i = \sum_t \hat{\mathbf{u}}_{it}$, $\hat{\mathbf{s}}_t = \sum_i \hat{\mathbf{u}}_{it}$, and $\hat{\varepsilon}_{it}$ is the residual $y_{it} - \mathbf{x}_{it}'\hat{\boldsymbol{\beta}}$. $\hat{\mathbf{V}}_{firm}$ is the usual formula for standard errors clustered by firm, $\hat{\mathbf{V}}_{time,0}$ is the usual formula for standard errors clustered by time, and $\hat{\mathbf{V}}_{white,0}$ are the usual OLS standard errors robust to heteroskedasticity. The $\hat{\mathbf{V}}_{time,l}$ and $\hat{\mathbf{V}}_{white,l}$ terms for $l \geq 1$ correct for persistent shocks common to many firms.

We subtract $\hat{\mathbf{V}}_{white,0}$ to correct for double-counting the within-firm variance. Both $\hat{\mathbf{V}}_{firm}$ and $\hat{\mathbf{V}}_{time,0}$ sum over the cross product $\hat{\mathbf{u}}_{it}\hat{\mathbf{u}}_{it}'$. Since that cross product appears in $\hat{\mathbf{V}}_{white,0}$, we eliminate the double-counting by subtraction. Similarly, $\hat{\mathbf{V}}_{firm}$, the variance that clusters by firm, includes all residual cross products $\hat{\mathbf{u}}_{it}\hat{\mathbf{u}}_{i,t+l}'$ within firm i. These cross products also appear in $\hat{\mathbf{V}}_{time,l}$, so we subtract $\hat{\mathbf{V}}_{white,l}$ to avoid double-counting.

The $\hat{\mathbf{V}}_{time,l}$ terms are less familiar. $\hat{\mathbf{V}}_{time,l}$ estimates $\mathbf{H}^{-1}\sum_t \mathrm{Cov}(\mathbf{s}_t, \mathbf{s}_{t+l})\mathbf{H}^{-1}$, a weighted autocovariance between time clusters. The autocovariances are induced by the persistent common shocks. Why do the $\hat{\mathbf{V}}_{time,l}$ terms appear twice? Recall that, when we take the variance of a univariate sum, the result is the sum of variances plus two times the sum of covariances. In the vector case, we have a similar result:

$$\mathrm{Var}\left(\sum_{t=1}^{T} \mathbf{s}_t\right) = \sum_{t} \mathrm{Var}(\mathbf{s}_t) + \sum_{t=1}^{T-1}\sum_{l=1}^{T-t} \mathrm{Cov}(\mathbf{s}_t, \mathbf{s}_{t+l}) + \sum_{t=1}^{T-1}\sum_{l=1}^{T-t} \mathrm{Cov}(\mathbf{s}_t, \mathbf{s}_{t+l})'.$$

The covariance terms $\mathrm{Cov}(\mathbf{s}_t, \mathbf{s}_{t+l})$ appear twice, just as they do in the univariate case.

Special case (*Double-clustering, but no persistent common shocks*): If the residuals do not contain persistent common shocks, then L=0 and the variance estimator becomes $\widehat{\mathrm{Var}}(\hat{\beta}) = \hat{\mathbf{V}}_{firm} + \hat{\mathbf{V}}_{time,0} - \hat{\mathbf{V}}_{white,0}$. This estimator is trivially calculated using a statistical package that has a built-in clustering command.[3]

Special case (*Persistent common shocks, but no double-clustering*): Consider the predictive regression where y_{it} is an L-period overlapping return, so $y_{it} = \sum_{k=1}^{L} R_{i,t+k}$, where $R_{i,t}$ is the return on company i's stock in month t. This regression is typically run under the assumption that there is no persistent firm-specific shock. In this case, the variance estimator becomes $\widehat{\mathrm{Var}}(\hat{\beta}) = \hat{\mathbf{V}}_{time,0} + \sum_{l=1}^{L}(\hat{\mathbf{V}}_{time,l} + \hat{\mathbf{V}}'_{time,l})$.

Asymptotic consistency: Asymptotic consistency of the standard errors is demonstrated in Appendix A. Consistency requires that both N and T become large. I assume that $T = \alpha N$, where α is a positive constant, and then take the probabilistic limit as $N \to \infty$. The relative magnitudes do not matter; for example consistency holds if N is twice as big as T, as long as both approach infinity. However, consistency will not necessarily hold if T goes to infinity while N is fixed, or if N goes to infinity while T is fixed. The intuition for this result is that we need T to become large for $\hat{\mathbf{V}}_{time,l}$ to be consistent, and we need N to become large for $\hat{\mathbf{V}}_{firm}$ to be consistent. When either T or N are small, there may be too much sampling variability in $\hat{\mathbf{V}}_{time,l}$ or $\hat{\mathbf{V}}_{firm}$.

4. When should we use robust standard errors?

Is there a downside to double-clustering the standard errors? Should we always adjust standard errors to handle persistent common shocks? In fact, it is not always best to use the "most robust" standard error formula. The various standard error formulas are estimates of true, unknown standard errors. In this section, I point out that the more robust standard error formulas tend to have less bias, but more variance. The lower bias improves the performance of test statistics. But the increased variance often leads us to find statistical significance even when it does not exist (e.g., we erroneously reject a true null hypothesis).

4.1. Bias

More robust standard errors have less bias. When is this effect likely to be important? Consider a researcher who uses single-clustered standard errors, and is considering double-clustering and adjusting for persistent common shocks. When will the more robust formulas make a difference? In this section, I argue that the researcher should think about three features of the data set: the distribution of the errors, the distribution of the regressors, and relative number of observations along the two clustering dimensions.

To make the discussion more concrete, consider a few scenarios.

[3] For example, in STATA we would issue the command "reg y x, cluster(firm)" to compute $\hat{\mathbf{V}}_{firm}$, "reg y x, cluster(time)" to compute $\hat{\mathbf{V}}_{time,0}$, and "reg y x, robust" to compute $\hat{\mathbf{V}}_{white,0}$. Here, "y" is the dependent variable, "x" is the single regressor (we could have more than one), "firm" is an index number unique to each firm, and "time" is an index number unique to each time period.

Scenario #1: The researcher should double-cluster, but instead single-clusters by firm. The double-clustered formula is $\widehat{\mathbf{V}}_{firm}+\widehat{\mathbf{V}}_{time,0}-\widehat{\mathbf{V}}_{white,0}$, while the single-clustered formula is $\widehat{\mathbf{V}}_{firm}$. Thus, the researcher omits $\widehat{\mathbf{V}}_{time,0}-\widehat{\mathbf{V}}_{white,0}$.

Scenario #2: The researcher should double-cluster, but instead single-clusters by time. The researcher omits $\widehat{\mathbf{V}}_{firm}-\widehat{\mathbf{V}}_{white,0}$.

Scenario #3: The researcher should adjust for persistent common shocks but fails to do so. The researcher omits $\widehat{\mathbf{V}}_{time,l}-\widehat{\mathbf{V}}_{white,l}$ for $l\geq 1$.

These scenarios show us that robust standard errors are most helpful when terms like $\widehat{\mathbf{V}}_{firm}-\widehat{\mathbf{V}}_{white,0}$, $\widehat{\mathbf{V}}_{time,0}-\widehat{\mathbf{V}}_{white,0}$, and $\widehat{\mathbf{V}}_{time,l}-\widehat{\mathbf{V}}_{white,l}$ are large. When are these terms large?

Start with scenario #1. The bias comes from omitting the time clustering. In large samples we can approximate the bias with

$$\mathrm{E}\widehat{\mathbf{V}}_{time,0}-\mathrm{E}\widehat{\mathbf{V}}_{white,0}\approx\mathbf{H}^{-1}\sum_{t}\sum_{i\neq j}\mathrm{Cov}(\mathbf{x}_{it}\varepsilon_{it},\mathbf{x}_{jt}\varepsilon_{jt})\mathbf{H}^{-1}. \tag{4}$$

This formula tells us three things: the distribution of the errors matters, the distribution of the regressors matters, and the balance between the number of observations on firms and time periods matters. Let us first consider the distribution of the errors, then come back to the other points. If, conditional on the regressors, errors are not correlated across firms, then there is no bias in scenario #1. To put it in mathematical terms, we know that $\mathrm{E}[\varepsilon_{it}\varepsilon_{jt}|\mathbf{x}_{it},\mathbf{x}_{jt}]=0$ implies $\mathrm{Cov}(\mathbf{x}_{it}\varepsilon_{it},\mathbf{x}_{jt}\varepsilon_{jt})=0$. We can make similar statements for the other scenarios. If the errors are conditionally uncorrelated across time then there is no bias in scenario #2, and if the errors are not conditionally autocorrelated then there is no bias in scenario #3.

Next consider the distribution of the regressors. Suppose that the errors exhibit strong time effects, so $\mathrm{Cov}(\varepsilon_{it},\varepsilon_{jt})>0$, but the regressors $\mathbf{x}_{it}$ and $\mathbf{x}_{jt}$ are independent of each other and of the errors. Then $\mathrm{Cov}(\mathbf{x}_{it}\varepsilon_{it},\mathbf{x}_{jt}\varepsilon_{jt})=0$, and we do not need to cluster by time, even though the residuals have strong time effects. The same argument holds for scenario #2. When we fail to cluster by firm, we omit the term $\widehat{\mathbf{V}}_{firm}-\widehat{\mathbf{V}}_{white,0}$. This term is a sum over $\mathrm{Cov}(\mathbf{x}_{it}\varepsilon_{it},\mathbf{x}_{ik}\varepsilon_{ik})$—the covariance between observations on the same firm in different time periods. If the regressors are not correlated across time, clustering by firm will not affect the standard errors, even if the errors have significant firm components. The same argument can also be applied to scenario #3. We need to adjust for persistent common shocks if the regressors are correlated across time. Otherwise the adjustment is not important.

To better understand this point, consider a few examples. Suppose the regressor is the growth rate of gross domestic product. This does not vary at all by firm, but varies across time and is persistent. Omitting the corrections for time effects and persistent common shocks may lead to bias. Thus, the researcher should worry about scenarios #1 and #3, but #2 is less important. Now suppose that the regressor is the return on the aggregate stock market. This does not vary by firm, but is not persistent. So scenario #1 is important, but #2 and #3 may not be a problem. Next consider the dividend yield of a firm. This varies by both firm and time. However, if most of the variation is across firms, and not across time, then omitting the firm clustering will create more bias than omitting the time clustering.

Double-clustering is most useful when both scenarios #1 and #2 lead to bias. In this case, clustering in either dimension will not eliminate the bias. This is likely to happen when the regressors exhibit both time and firm effects. For example, a regressor like the dividend yield has both time and firm variation, although most of the variation may be at the firm level. Another important case is a multivariate regression where some regressors vary by time, and some vary by firm. For example, if one regressor is the dividend yield, and another regressor is growth in gross domestic product, then time clustering alone would not get the standard errors right for the dividend yield, and firm clustering alone would not get the standard errors right for the other regressor. The only way to get both standard errors right is to double-cluster.

Finally, consider the relative number of firms and time periods in a data set. Suppose that we have 1,000 firms observed over ten years, for 10,000 total observations. The OLS formula $\widehat{\mathbf{V}}_{white,0}$ relies on the assumption that we have 10,000 uncorrelated observations. Probably we have far fewer. If we cluster by time, we allow arbitrary correlation between observations in the ten time periods, thus, we assume that we have only ten uncorrelated observations. If we cluster by firm, we assume 1,000 uncorrelated observations. Going from 10,000 observations to ten probably has a bigger effect than going from 10,000 to 1,000. In this example, omitting the time clustering is likely to be more important than omitting the firm clustering.

The general point is that, all else equal, it is more important to cluster along the dimension with fewer observations. If we have a sample with ten firms and 1,000 time periods, the bigger bias reduction will probably come from clustering by firm.

In fact, we can make a stronger statement—if the dimensions are extremely unbalanced, we do not need to double-cluster at all. If we fix the number of observations in one dimension, and let the number of observations in the other become very large, the bias disappears (so long as we single-cluster on the less-numerous dimension). Here is a mathematical statement of this claim when N is fixed and T is very large:

$$\lim_{T\to\infty,N\text{ fixed}}\frac{\widehat{\mathbf{V}}_{firm}+\widehat{\mathbf{V}}_{time,0}-\widehat{\mathbf{V}}_{white,0}}{\widehat{\mathbf{V}}_{firm}}=1.$$

This is a counter-intuitive result. To understand it, recall that our estimate $\widehat{\beta}$ becomes more precise as we add more independent observations. The estimator variance reflects this and converges to zero in large samples. The term $\widehat{\mathbf{V}}_{time,0}$ is constructed based on the assumption that observations in different time periods are independent. Thus, it will converge to zero as T becomes large,

whether or not the assumption is in fact true.[4] Likewise, $\widehat{\mathbf{V}}_{white,0}$ converges to zero as the total number of observations becomes large. In contrast, $\widehat{\mathbf{V}}_{firm}$ relies on the assumption that observations are independent across firms. If we fix N, $\widehat{\mathbf{V}}_{firm}$ will not converge to zero. Stated loosely, as T becomes large, we average away noise due to variation across time, but we do not average away noise due to variation across firms.

I should not over-sell this point. To be clear, the effect of clustering is determined by an interaction between the number of observations in each dimension and the magnitude of the correlation between observations. If there is no firm effect, so observations on a given firm are uncorrelated across time periods, then we do not need to cluster by firm, even if there are 1,000 time periods and only ten firms. However, if the data have significant firm and time effects, then it is probably more important to cluster along the dimension with fewer observations.

This analysis suggests that double-clustering is most important when the number of firms and time periods are not too different. For example, Fama and French (2000) predict firm-level profitability in a panel with thousands of firms and roughly 35 years of annual accounting data. In this case, clustering by time is probably good enough as the increase in bias from failing to also cluster by firm will likely be small. In the empirical application later in this paper, I run profitability regressions at the industry level. There are far fewer industries than firms, and in that application double-clustering significantly changes the standard errors relative to single-clustering by time.

4.2. Variance

In many cases of interest, the more robust standard error estimates have higher variances. In some simple cases we can verify this statement with analytic results. Consider a regression with a single regressor where the errors are independent, so that we do not need to cluster. When N and T are both large, the single-clustered standard error estimate always has a higher variance than the OLS standard errors. In more complicated cases we can carry out simulations. For example, consider a regression model with a single independent and identically distributed (iid) standard normal regressor and iid standard normal errors. Suppose we have ten firms observed over ten time periods, for 100 observations total. This is a model where the OLS standard errors are appropriate, and we do not need to single-cluster, double-cluster, or adjust for persistent common shocks. I generate 10,000 samples from this model and calculate the various standard error estimates. White standard errors (with no clustering) had a simulation standard deviation of 1.4%, and single-clustered standard errors had simulation standard deviations of 2.6%, whether clustering was done by firm or time. The double-clustered standard errors that exclude persistent common shocks had a simulation standard deviation of 3.2%, and when allowing for persistent common shocks with L=2 lags, the standard deviation was 3.6%. From this simple experiment we see that more robust standard error estimates tend to have more sampling variability.

Increasing the variance of a standard error estimate may lead us to see statistical significance where it does not exist. This result comes from Jensen's inequality. Suppose we have a single coefficient, and we test the null hypothesis that $\beta=\beta_{null}$. We reject the null for large values of the t-statistic $\widehat{t}=|\widehat{\beta}-\beta_{null}|/se(\widehat{\beta})$. If $\widehat{\beta}$ and $se(\widehat{\beta})$ are independent, then as the variance of $se(\widehat{\beta})$ increases, the expected value of the test statistics $\mathrm{E}(\widehat{t})$ will rise. Larger test statistics mean that we too often reject a true null hypothesis. In many cases of interest, $\widehat{\beta}$ and $se(\widehat{\beta})$ are independent or are close to independent. For example, in a regression with independent normal errors, the estimate $\widehat{\beta}$ is asymptotically independent of all the standard error estimators proposed in this paper. If we repeat the simulation experiment in the previous paragraph, we get approximately zero correlations between the coefficient estimate and all the standard error estimates.

For the clustered standard errors, the variance of the standard error estimate becomes large as the number of clusters decreases. Clustered standard errors are estimated by averaging across clusters. Few clusters means a small number of terms in the average, and thus more estimation error. Consider, for example, the standard errors that cluster by time:

$$\widehat{\mathbf{V}}_{time,0}\equiv\mathbf{H}^{-1}\sum_{t}\widehat{\mathbf{s}}_t\widehat{\mathbf{s}}_t'\mathbf{H}^{-1}.$$

This is an average over the products $\widehat{\mathbf{s}}_t\widehat{\mathbf{s}}_t'$. As we increase T, we increase the number of terms in the average, and the variance of the standard error estimate declines. Thus, we need large T to shrink the variance of $\widehat{\mathbf{V}}_{time,l}$, and we need large N to shrink the variance of $\widehat{\mathbf{V}}_{firm}$. If either T or N are small, then double-clustered standard errors can do more harm than good.

Consider the situation where a researcher sees very different results when going from single- to double-clustered standard errors. The researcher may take this as evidence that it is important to cluster along both dimensions. But if there are too few clusters in either the time or firm dimension, the double-clustered standard error estimate will be noisy. The different results could be spurious and due to noise. Thus, double clustering makes sense only when we have sufficient clusters along both dimensions. How many clusters do we need in practice? In the next section, I investigate this question with a Monte Carlo experiment.

4.3. Added bias when adjusting for persistent common shocks

There is a special consideration when adjusting for persistent common shocks. The estimator is biased, and it

[4] To see this mathematically, return to the formula for the estimator in Eq. (3):

$$\widehat{\mathbf{V}}_{time,0}\equiv\mathbf{H}^{-1}\sum_{t}\widehat{\mathbf{s}}_t\widehat{\mathbf{s}}_t'\mathbf{H}^{-1}.$$

$\mathbf{H}$ is $\sum\mathbf{x}_{it}\mathbf{x}_{it}'$, a summation over NT terms. $\widehat{\mathbf{s}}_t$ is a summation over N terms, so $\sum_t\widehat{\mathbf{s}}_t\widehat{\mathbf{s}}_t'$ is a summation over TN^2 terms. Therefore, $\mathrm{E}\widehat{\mathbf{V}}_{time,0}\leq K(N^{-1}T^{-1})(TN^2)(N^{-1}T^{-1})=K(T^{-1})$, where K is a finite constant. Since $\widehat{\mathbf{V}}_{time,0}$ is non-negative, this inequality implies that $\widehat{\mathbf{V}}_{time,0}$ converges to zero as T becomes large. This holds whether or not observations are truly independent across time.

is not a simple thing to fix this bias. We handle persistent common shocks with the terms $\widehat{\mathbf{V}}_{time,l}$ and $\widehat{\mathbf{V}}_{white,l}$. Both of these terms involve estimates of autocorrelations: $\widehat{\mathbf{V}}_{time,l}$ uses $\sum_t \widehat{\mathbf{s}}_t \widehat{\mathbf{s}}_{t+l}'$, and $\widehat{\mathbf{V}}_{white,l}$ uses $\sum_t \widehat{\mathbf{u}}_{it} \widehat{\mathbf{u}}_{i,t+l}'$. In general, estimates of non-negative autocorrelations are biased downward. Thus, we have identified two factors that cause these standard errors to falsely reject true null hypotheses—they have larger estimation variances than simpler formulas, and they exhibit downward bias.

We can see the bias in a simple example. Suppose we have data $\{Z_t\}_{t=1}^T$. The first-order autocovariance is $\text{Cov}(Z_t,Z_{t-1}) = \text{E}Z_tZ_{t-1} - (\text{E}Z_t)^2$. We estimate this with

$$\widehat{\text{Cov}}(Z_t,Z_{t-1}) = (T-1)^{-1}\sum_{t\geq 2} Z_tZ_{t-1} - \overline{Z}^2,$$

where $\overline{Z} = T^{-1}\sum_{t\geq 1} Z_t$. The expectation of the estimate is

$$\text{E}\widehat{\text{Cov}}(Z_t,Z_{t-1}) = \text{E}Z_tZ_{t-1} - \text{E}\overline{Z}^2.$$

If $\text{E}\overline{Z}^2 = (\text{E}\overline{Z})^2$, then this estimate is unbiased. But Jensen's inequality tells us that $\text{E}\overline{Z}^2 \geq (\text{E}\overline{Z})^2$, which shrinks the estimate. The downward bias is present even when the true autocorrelation is zero—in this case the expected value of the estimate is negative.

The bias of autocorrelation estimates is an old and unsolved statistical problem that dates at least back to Hurwicz (1950). Nickell (1981) points out some of the problems this effect causes in panel regressions. Stambaugh (1999) shows that it makes conventional inference in predictive time-series regressions unreliable. Petersen (2009) shows that the effect leads to bias in adjusted Fama-MacBeth standard errors. Petersen's critique applies here as well.

Two useful facts about the bias are that it increases with the magnitude of the correlation, and it disappears as the sample becomes large. Therefore, we expect these standard errors to perform well when the correlations are close to zero, and the sample size is large. Of course, if the correlations are low then we can just ignore them and use simpler formulas. The only case where these standard errors are unambiguously preferred is when the correlations are significant and the sample size is large enough to correct the bias. I perform Monte Carlos to see how big the samples need to be.

5. Monte Carlo experiments

In this section, I use Monte Carlo simulations to investigate the small-sample performance of the robust standard errors. I simulate 5,000 draws from the panel regression,

$$y_{it} = \beta_0 + \beta_1 x_{1,it} + \beta_2 x_{2,it} + \varepsilon_{it},$$

with $\beta_0 = 0$ and $\beta_1 = \beta_2 = 1$. The simulation is repeated for various sample sizes and error dependencies. For each sample, I estimate the regression and carry out two-sided t-tests of the nulls that $\beta_1 = 1$ and $\beta_2 = 1$. Table 1 reports rejection frequencies for t-tests constructed from many different variance estimators: the usual OLS variance estimator, $\widehat{\mathbf{V}}_{white,0}$, the estimator that clusters by firm, $\widehat{\mathbf{V}}_{firm,0}$, the estimator that clusters by time, $\widehat{\mathbf{V}}_{time,0}$, the estimator that clusters by both firm and time but does not allow persistent common shocks, $\widehat{\mathbf{V}}_{firm} + \widehat{\mathbf{V}}_{time,0} - \widehat{\mathbf{V}}_{white,0}$, and the full variance estimator that clusters by firm and time and allows persistent common shocks (with $L=2$). The Monte Carlo also considers fixed-effects regressions—I run the regression with firm fixed effects and cluster the standard errors by time, and run the regression with time fixed effects and cluster the standard errors by firm. Since the null hypothesis is true, we prefer standard errors that deliver rejection frequencies close to 5%.

I consider three different data-generating processes.

Panel A: The errors and regressors are distributed $N(0,1)$ and independent across both i and t.

Panel B: x_1 has time effects, x_2 has firm effects, and the errors have both. There are no persistent common shocks. $x_{1,it} = \xi_t$, where $\xi_t \sim N(0,1)$ and independent across t. $x_{2,it} = \eta_{it}$, where $\eta_{it} = 0.9\eta_{i,t-1} + \varsigma_{it}$, with $\varsigma_{it} \sim N(0,1)$ and independent across i and t. $\varepsilon_{it} = \tilde{\xi}_t + \tilde{\eta}_{it}$, where $\tilde{\xi}_t$ and $\tilde{\eta}_{it}$ have the same distributions as ξ_t and η_{it}.

Panel C: x_1 has a persistent common shock, x_2 has firm effects, and the errors have a persistent common shock but no firm effects. $\varepsilon_{it} = \theta_i f_t + u_{it}$, where $\theta_i \sim N(0,0.25)$ and independent across firms, $f_t = 0.5f_{t-1} + v_t$, with $v_t \sim N(0,1)$ and independent across time. $x_{1,it} = \theta_i \tilde{f}_t$, where $\tilde{f}_t$ has the same distribution as f_t, and θ_i is the same loading used to generate ε_{it}. $x_{2,it} = \eta_i$, where $\eta_i \sim N(0,1)$ and independent across firms.

Notice that, in Panels B and C, the regressors exhibit correlations similar to those in the errors. As argued in Section 4.1, double-clustering matters most when both the regressors and the errors exhibit time and firm effects.

In Panel A all the variance estimators are valid and should deliver rejection frequencies of 5%. Instead we see that the simpler formulas get the size right, but the more robust formulas over-reject in small samples. For example, in the simulation with $T=25$ and $N=50$, the standard errors that are robust to persistent common shocks reject a true null at least 12% of the time. The size distortion diminishes, but does not disappear, when we go to a sample with 100 time periods. This is consistent with the arguments made in Section 4.2: in small samples the more robust formulas have higher estimation noise, and via Jensen's inequality this causes us to over-reject a true null hypothesis.

In Panel B we need to double-cluster. The OLS rejection frequencies are all at least 40%. Single-clustering by firm gets the size right for β_2 but not for β_1. Likewise, single-clustering by time gets the size right for β_1 but not for β_2. This happens because x_1 has only time effects, x_2 has only firm effects, and the two regressors are uncorrelated. Thus, even though the errors have both firm and time effects, single-clustering works for either β_1 or β_2. In order to get the size right for both regressors, we need to double-cluster. The fixed effects do not help much. The time fixed effects are collinear with x_1, so we cannot estimate β_1. The firm fixed effects do not capture the actual firm dynamics, which follow an autoregressive process.

In Panel C we need to use standard errors robust to persistent common shocks. However, from our results in Panel A we know that these standard errors have poor small-sample properties. We see a similar effect here—there are size distortions for all sample sizes, and they are smallest in the biggest sample. It is worth

Table 1
Monte Carlo comparison of standard error formulas.

The table shows results of a Monte Carlo evaluation of various standard error formulas. I simulate 5,000 samples from the regression model $y_{it} = \beta_0 + \beta_1 x_{1it} + \beta_2 x_{2it} + \varepsilon_{it}$ with $\beta_0 = 0$ and $\beta_1 = \beta_2 = 1$. For each sample, I estimate the regression and carry out two-sided t-tests of the nulls that $\beta_1 = 1$ and $\beta_2 = 1$. The table reports rejection frequencies from various estimator variance formulas: (1) "OLS std errors" denotes V_{white}, (2) "cluster by firm" denotes V_{firm}, (3) "cluster by time" denotes V_{time}, (4) "cluster by firm, time FE" denotes time fixed effects with V_{firm}, (5) "cluster by time, firm FE" denotes firm fixed effects with V_{time}, (6) "cluster by both firm and time (not robust to persistent common shocks)" denotes $V_{firm} + V_{time} - V_{white}$, and (7) "cluster by both firm and time (robust to persistent common shocks, L=2)" denotes $V_{firm} + V_{time} - V_{white}$ + {corrections for persistent common shocks with L=2}.

Panel A: Both regressors and ε_{it} are iid N(0,1) across both i and t, so OLS error assumptions are satisfied

	T=25, N=50		T=50, N=50		T=100, N=100	
	β_1	β_2	β_1	β_2	β_1	β_2
OLS std errors	0.049	0.048	0.050	0.047	0.049	0.056
Cluster by firm	0.053	0.056	0.056	0.053	0.052	0.055
Cluster by time	0.058	0.065	0.057	0.056	0.052	0.059
Cluster by firm, time FE	0.057	0.056	0.060	0.058	0.051	0.058
Cluster by time, firm FE	0.065	0.067	0.060	0.059	0.051	0.062
Cluster by both firm and time (not robust to persistent common shocks)	0.069	0.070	0.062	0.064	0.054	0.059
Cluster by both firm and time (robust to persistent common shocks, L=2)	0.127	0.123	0.100	0.100	0.069	0.078

Panel B: Errors have both time and firm effects, and persistent common shocks. x_{1it} has time fixed effects (and no persistent common shock), and x_{2it} has firm effects that follow an autoregressive process

	T=25, N=50		T=50, N=50		T=100, N=100	
	β_1	β_2	β_1	β_2	β_1	β_2
OLS std errors	0.560	0.425	0.519	0.482	0.640	0.478
Cluster by firm	0.695	0.058	0.635	0.059	0.707	0.053
Cluster by time	0.093	0.531	0.074	0.533	0.056	0.501
Cluster by firm, time FE	–	0.058	–	0.058	–	0.054
Cluster by time, firm FE	0.093	0.369	0.074	0.466	0.056	0.477
Cluster by both firm and time (not robust to persistent common shocks)	0.105	0.066	0.081	0.061	0.060	0.055
Cluster by both firm and time (robust to persistent common shocks, L=2)	0.174	0.103	0.113	0.080	0.076	0.062

Panel C: Errors have both time and firm effects, and persistent common shocks. x_{1it} has time effects with persistent common shocks, and x_{2it} has firm fixed effects

	T=25, N=50		T=50, N=50		T=100, N=100	
	β_1	β_2	β_1	β_2	β_1	β_2
OLS std errors	0.747	0.563	0.765	0.673	0.809	0.750
Cluster by firm	0.906	0.074	0.916	0.074	0.931	0.056
Cluster by time	0.169	0.865	0.155	0.906	0.139	0.924
Cluster by firm, time FE	0.654	0.074	0.672	0.074	0.726	0.056
Cluster by time, firm FE	0.167	–	0.157	–	0.140	–
Cluster by both firm and time (not robust to persistent common shocks)	0.176	0.087	0.162	0.081	0.143	0.058
Cluster by both firm and time (robust to persistent common shocks, L=2)	0.201	0.127	0.145	0.100	0.098	0.067

pointing out that, since x_2 has only firm effects, we can get the right test size for β_2 by clustering on firm, and we do not need to adjust for persistent common shocks.

The Monte Carlos are generally supportive of using robust standard errors. Single-clustered standard errors cannot handle regressions where one regressor has significant time effects and another has significant firm effects. If we are willing to accept false rejections of up to 10% in a test with 5% size, then double-clustering works well so long as we have more than 25 observations on both firms and time periods. Correcting for persistent common shocks requires between 50 and 100 time periods.

6. Application to modeling industry profitability

I demonstrate the standard errors with an application to modeling industry profitability. I consider the hypothesis that profits are higher in more concentrated industries, and measure concentration with a forward-looking variant of the Herfindahl-Hirschman Index (HHI) (Hirschman, 1964). The HHI is a widely used measure of industry concentration. For example, the U.S. Department of Justice uses the index to help determine whether a merger is anticompetitive (see USDOJ and FTC, 1997).

The HHI is usually calculated from historical sales data. I calculate a variant based on market capitalization. We have M industries, indexed by $m=1,\ldots,M$. For each industry m, the value of the index is

$$HHI_{m,t} = 100^2 \sum_{i \in I_m} \left(\frac{\text{market cap of firm } i \text{ in industry } m \text{ at time } t}{\text{total market cap of all firms in industry } m \text{ at time } t} \right)^2,$$

where I_m is the set of all firms i in industry m. While sales data are backward looking, market capitalization is a forward-looking measure of earnings (and payouts).

I test the predictive power of our index in a panel data set of US industries. The sample consists of annual Compustat data observed from 1964 to 2007. I predict $ROA_{m,t}$, a measure of industry profitability given by the ratio of earnings to assets. Industries are based on four-digit Standard Industrial Classification (SIC) codes. The panel is unbalanced—there are 434 SIC categories and 44 years, for a total of 15,066 observations. The regression is

$$ROA_{m,t} = \beta_0 + \beta_1 \ln(HHI_{m,t-1}) + \beta_2 PB_{m,t-1} + \beta_3 DB_{m,t-1} + \beta_4 \overline{ROA}_{t-1} + \varepsilon_{m,t}.$$

PB and *DB* are industry-level price-to-book and dividend-to-book ratios, respectively. Fama and French (2000) identify these variables as strongly contemporaneously associated with firm-level profitability. My regression is predictive, but takes their contemporaneous regression model as a starting point. Fama and French (2000) argue that high price-to-book ratios indicate that the market predicts higher future earnings, and that high dividend-to-book ratios indicate that company management anticipates higher earnings. $\overline{ROA}_t$ is the return on assets for the market as a whole in year t. It controls for market-wide trends in profitability. The data and regressors are described in more detail in Appendix B.

Notice that this example uses industry-level rather than firm-level data. Even though the text of this paper has mostly referred to the clustering dimensions as firms and time periods, the results trivially generalize to clustering along any two dimensions. I picked this example in part because it is one where double-clustering is likely to make a significant difference. If I used an example with firm-level data, there would be many more firms than time periods. As discussed in Section 4.1, double-clustering is most useful when the number of observations in each dimension is not too far apart. This data set has 434 industries and 43 years, but US firm-level data would have many more than 434 firms.

Results appear in Table 2. *HHI* positively predicts industry profitability, indicating that profits tend to be persistently higher in more concentrated industries. Like Fama and French (2000), I find that *PB* and *DB* are significant positive predictors of profitability. $\overline{ROA}$ is also positive, suggesting that higher-than-usual profits one year predict higher-than-usual profits the next.

It is interesting to compare the effects of single- and double-clustering. For the regressor ln(*HHI*), double-clustered standard errors are similar to standard errors from single-clustering on industry. For $\overline{ROA}$, double-clustering gives results similar to single-clustering by year. Industry concentration varies a lot between industries, and does not vary as much over time within an industry. In contrast, $\overline{ROA}$ varies across time, but does not vary at all across industries. As discussed in Section 4.1, clustering makes a big difference when both the error and the regressor are correlated within the clustering dimension. We need to cluster by industry to get the right standard error for ln(*HHI*), and we need to cluster by time to get the right standard error for $\overline{ROA}$. Double-clustering gets the right standard error for both.

Correcting for persistent common shocks generally increases the standard errors. The largest effect is for $\overline{ROA}$, and the smallest effect is for ln(*HHI*). This makes sense, since $\overline{ROA}$ is positively correlated across time, while ln(*HHI*) has very weak time effects.

7. Conclusion

This paper derives easy-to-compute formulas for standard errors that cluster by both firm and time. Both the statistical theory and the Monte Carlo results suggest that

Table 2
Application to modeling industry profitability.

The table shows a regression to model industry profitability. The dependent variable is $ROA_{m,t}$, the ratio of earnings-to-assets in industry m in year t. The regressors are as follows. $\ln(HHI_{m,t-1})$ is the log of the Hefindahl-Hirschman concentration index for the industry, computed from market caps. Price/Book equity$_{m,t-1}$ is the price-to-book ratio. Dividends/Book equity$_{m,t-1}$ is the dividends-to-book ratio. Market ROA_{t-1} is the market-wide (not industry-specific) ratio of earnings-to-assets. "Estimate" denotes ordinary-least-squares estimates. t-Statistics are presented for different standard error formulas: (1) "White" denotes V_{white}, (2) "single- clustered, time" denotes V_{time}, (3) "single-clustered, firm" denotes V_{firm}, (3) "double-clustered, $L=0$" denotes $V_{time} + V_{firm} - V_{white}$, and (4) "double-clustered, $L=2$" denotes $V_{firm} + V_{time} - V_{white}$ + {corrections for persistent common shocks} with $L=2$. See Appendix B for details about the data.

			t-Statistics			
			Single-clustered		Double-clustered	
Regressor	Estimate	White	Time	Industry	$L=0$	$L=2$
$\ln(HHI_{m,t-1})$	0.0049	10.530	9.988	4.314	4.274	4.610
Price/Book equity$_{m,t-1}$	0.0072	17.670	5.604	11.062	5.212	3.394
Dividends/Book equity$_{m,t-1}$	0.3171	20.700	9.877	10.083	7.505	5.482
Market ROA_{t-1}	1.0631	32.720	8.667	19.958	8.195	4.924
Intercept	−0.0521	−14.461	−9.148	−5.847	−5.240	−4.714
R-squared: 19.29%						

simultaneously clustering by firms and time leads to significantly more accurate inference in finance panels. Monte Carlo experiments suggest that, as long as we do not allow for persistent common shocks, clustering on both firm and time works adequately when we have at least 25 firms and time periods. However, allowing for persistent common shocks requires a larger number of time periods.

This paper leaves a number of issues unresolved. The standard errors that correct for persistent common shocks do not behave well in small samples. Further work could be done to improve their small-sample performance. There is also more work to be done with the pure double-clustering problem, some of which has already been carried out by Cameron, Gelbach, and Miller (2006). They show how to extend two-way clustering to clustering along more dimensions. They also describe how to apply these methods to nonlinear estimators.

Appendix A. Demonstration of asymptotic consistency

To establish consistency, normalize the estimator as

$$\widehat{\mathrm{Var}}[N^{-1/2}T^{-1}\mathbf{H}\widehat{\beta}] = N^{-1}\sum_i W_{firm,i} + T^{-1}\sum_t (W_{time,0,t} - W_{ols,0,t}) + \sum_l \left[T^{-1}\sum_t (W_{time,l,t} - W_{ols,l,t} + W'_{time,l,t} - W'_{ols,l,t}) \right],$$

where $W_{firm,i} = T^{-2}\sum_{t,k} \mathbf{x}_{it}\widehat{\varepsilon}_{it}\widehat{\varepsilon}_{ik}\mathbf{x}'_{ik}$, $W_{time,l,t} = \alpha^{-1}N^{-2}\sum_{i,j}\mathbf{x}_{it}\widehat{\varepsilon}_{it}\widehat{\varepsilon}_{j,t+l}\mathbf{x}'_{j,t+l}$, and $W_{white,l,t} = \alpha^{-1}N^{-2}\sum_i \mathbf{x}_{it}\widehat{\varepsilon}_{it}\widehat{\varepsilon}_{i,t+l}\mathbf{x}'_{i,t+l}$. Tedious manipulations lead to the results that

$$\mathrm{cov}(W_{firm,i}, W_{firm,j}) = \mathrm{O}(T^{-1}) \quad \text{for } i \neq j,$$

and

$$\mathrm{cov}(W_{time,l,t}, W_{time,l,k}) = \mathrm{O}(N^{-1}) \quad \text{for } |t-k| > L.$$

Therefore, we can show by direct calculation that

$$\lim_{T\to\infty} \mathrm{Var}\left[T^{-1}\sum_t W_{time,l,t} \right] = 0,$$

which implies that $T^{-1}\sum_t W_{time,l,t}$ converges to its expectation in mean square. A similar argument demonstrates that $N^{-1}\sum_i W_{firm,i}$ and $N^{-1}\sum_i W_{white,l,t}$ converge to their expectations. Consistency of the standard errors follows.

Appendix B. Data construction for empirical application

The data used for the application to forecasting firm profitability come from Compustat. Data construction details and Compustat codes follow.

Firm-level earnings are Compustat item IB, earnings before extraordinary items. Firm-level assets are item AT. Firm-level liabilities are item LT. Book value is calculated as Assets–Liabilities–Preferred Stock. To calculate the value of preferred stock, I use the redemption value of preferred stock (item PSTKRV). If that is not available I use the liquidating value (PSTKL), and if that is not available the carrying value (UPSTK) is used. Market capitalization is common shares outstanding (CSHO) multiplied by the closing price at the end of the fiscal year (PRCC_F). Dividends are item DVC.

I also carried out the empirical analysis adjusting earnings and book value for deferred income taxes and investment tax credits, as in Fama and French (2000). The results did not meaningfully change.

Industry-level ratios were calculated by aggregating firm-level ratios. Aggregation is carried out at the four-digit Standard Industrial Classification (SIC) level. Industry-year pairs that contain only one firm are excluded. The results are not sensitive to inclusion of the single-firm industries. They are also not sensitive to screening out industry-year pairs with five firms. Before calculating industry-level data, I first drop all observations with book values less than $5 million and assets less than $10 million.

To calculate industry-level return on assets, I calculate the firm-level earnings-to-assets ratio, then winsorize within each yearly cross-section at the 1% and 99% percentiles. Industry-level ROA is the asset-weighted average of firm-level ratios. Similarly, market-wide ROA is calculated as the asset-weighted average over the entire market.

To calculate industry-level dividends-to-book, I calculate the firm-level dividends-to-book, winsorize at yearly 1% and 99% percentiles, and form the book-weighted average of firm-level ratios. Industry-level market-to-book is calculated in the same way.

References

Cameron, C., Gelbach, J., Miller, D., 2006. Robust inference with multi-way clustering. NBER Technical Working Paper no. 327.

Cohen, R., Polk, C., Vuolteenaho, T., 2003. The value spread. Journal of Finance 58, 609–641.

Fama, E., MacBeth, J., 1973. Risk, return, and equilibrium. Journal of Political Economy 81, 607–636.

Fama, E., French, K., 2000. Forecasting profitability and earnings. Journal of Business 73, 161–175.

Hansen, L., Hodrick, R., 1980. Forward exchange rates as optimal predictors of future spot rates: an econometric analysis. Journal of Political Economy 88, 829–853.

Hirschman, A., 1964. The paternity of an index. The American Economic Review 54, 761–762.

Huber, P., 1967. The behavior of the maximum likelihood estimates under nonstandard conditions. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1; 1967, pp. 221–233.

Hurwicz, L., 1950. Least-squares bias in time series. In: Koopmans, T. (Ed.), Statistical Inference in Dynamic Economic Models. John Wiley and Sons, New York, pp. 365–383.

Larrain, B., 2006. Do banks affect the level and composition of industrial volatility? Journal of Finance 61, 1897–1925

Li, K., Morck, R., Yang, F., Yeung, B., 2004. Firm-specific variation and openness in emerging markets. Review of Economics and Statistics 86, 658–669.

Newey, W., West, K., 1987. A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. Econometrica 55, 703–708.

Nickell, S., 1981. Biases in dynamic models with fixed effects. Econometrica 49, 1417–1426.

Petersen, M., 2009. Estimating standard errors in finance panel data sets: comparing approaches. Review of Financial Studies 22, 435–480.

Rajan, R., Zingales, L., 1998. Financial dependence and growth. American Economic Review 88, 559–586.

Rogers, W., 1983. Analyzing Complex Survey Data. Rand Corporation Memorandum, Santa Monica, CA.

Stambaugh, R., 1999. Predictive regressions. Journal of Financial Economics 54, 375–421.

United States Department of Justice and the Federal Trade Commission, 1997. Horizontal Merger Guidelines. Available at: 〈http://www.usdoj.gov/atr/public/guidelines/hmg.pdf〉.

White, H., 1980. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. Econometrica 48, 817–838.