

Department of Economics

Working Paper Series

Robust Inference with Multi-way Clustering

Douglas Miller
University of California, Davis

Douglas Miller
University of California, Davis

A. Cameron
University of California, Davis

Jonah Gelbach
University of Arizona

May 01, 2009

Paper # 09-9

In this paper we propose a variance estimator for the OLS estimator as well as for nonlinear estimators such as logit, probit and GMM. This variance estimator enables cluster-robust inference when there is two-way or multi-way clustering that is non-nested. The variance estimator extends the standard cluster-robust variance estimator or sandwich estimator for one-way clustering (e.g. Liang and Zeger (1986), Arellano (1987)) and relies on similar relatively weak distributional assumptions. Our method is easily implemented in statistical packages, such as Stata and SAS, that already offer cluster-robust standard errors when there is one-way clustering. The method is demonstrated by a Monte Carlo analysis for a two-way random effects model; a Monte Carlo analysis of a placebo law that extends the state-year effects example of Bertrand et al. (2004) to two dimensions; and by application to studies in the empirical literature where two-way clustering is present.



Department of Economics
One Shields Avenue
Davis, CA 95616
(530)752-0741

http://www.econ.ucdavis.edu/working_search.cfm

Robust Inference with Multi-way Clustering

A. Colin Cameron,^{*}Jonah B. Gelbach,[†]and Douglas L. Miller[‡]

This version: May 1, 2009
First Version: April 21, 2005

Abstract

In this paper we propose a variance estimator for the OLS estimator as well as for nonlinear estimators such as logit, probit and GMM. This variance estimator enables cluster-robust inference when there is two-way or multi-way clustering that is non-nested. The variance estimator extends the standard cluster-robust variance estimator or sandwich estimator for one-way clustering (e.g. Liang and Zeger (1986), Arellano (1987)) and relies on similar relatively weak distributional assumptions. Our method is easily implemented in statistical packages, such as Stata and SAS, that already offer cluster-robust standard errors when there is one-way clustering. The method is demonstrated by a Monte Carlo analysis for a two-way random effects model; a Monte Carlo analysis of a placebo law that extends the state-year effects example of Bertrand et al. (2004) to two dimensions; and by application to studies in the empirical literature where two-way clustering is present.

Keywords: cluster-robust standard errors; two-way clustering; multi-way clustering.

JEL Classification: C12, C21, C23.

^{*}Dept. of Economics, University of California - Davis.

[†]Dept. of Economics, University of Arizona.

[‡]Dept. of Economics, University of California - Davis. Address for correspondence: Douglas L. Miller, Department of Economics, University of California - Davis, One Shields Ave, Davis, CA 95616. dlmiller@ucdavis.edu

1. Introduction

A key component of empirical research is conducting accurate statistical inference. One challenge to this is the possibility of errors being correlated within cluster. In this paper we propose a variance estimator for commonly used estimators, such as OLS, probit, and logit, that provides cluster-robust inference when there is multi-way non-nested clustering. The variance estimator extends the standard cluster-robust variance estimator for one-way clustering, and relies on similar relatively weak distributional assumptions. Our method is easily implemented in any statistical package that provides cluster-robust standard errors with one-way clustering. An ado file for multi-way clustering in Stata is available at the website www.econ.ucdavis.edu/faculty/dlmiller/statafiles.

Controlling for clustering can be very important, as failure to do so can lead to massively under-estimated standard errors and consequent over-rejection using standard hypothesis tests. Moulton (1986, 1990) demonstrated that this problem arose in a much wider range of settings than had been appreciated by microeconometricians. More recently Bertrand, Duflo and Mullainathan (2004) and Kezdi (2004) emphasized that with state-year panel or repeated cross-section data, clustering can be present even after including state and year effects and valid inference requires controlling for clustering within state. These papers, like most previous analyses, focus on one-way clustering.

For nested two-way or multi-way clustering one simply clusters at the highest level of aggregation. For example, with individual-level data and clustering on both household and state one should cluster on state. Pepper (2002) provides an example.

If instead multi-way clustering is non-nested, the existing approach is to specify a multi-way error components model with iid errors. Moulton (1986) considered clustering due to grouping of three regressors (schooling, age and weeks worked) in a cross-section log earnings regression. Davis (2002) modelled film attendance data clustered by film, theater and time and provided a quite general way to implement feasible GLS even with clustering in many dimensions. But these models impose strong assumptions, including homoskedasticity and errors equicorrelated within cluster. And even the two-way random effects model for linear regression is typically not included in standard econometrics packages.

In this paper we take a less parametric cluster-robust approach that generalizes one-way cluster-robust standard errors to the non-nested multi-way clustering case. One-way “cluster-robust” standard errors rely on weak assumptions – errors are independent but not identically distributed across clusters and can have quite general patterns of within cluster correlation and heteroskedasticity. These standard errors generalize those of White (1980) for independent heteroskedastic errors. Key references include Pfeiffermann and Nathan (1981) for clustered sampling, White (1984) for a multivariate dependent variable, Liang and Zeger (1986) for estimation in a generalized estimating equations setting, and Arellano (1987) and Hansen (2007) for linear panel

models. Wooldridge (2003) provides a survey, and Wooldridge (2002) and Cameron and Trivedi (2005) give textbook treatments.

Our multi-way robust variance estimator is easy to implement. In the two-way clustering case, we obtain three different cluster-robust “variance” matrices for the estimator by one-way clustering in, respectively, the first dimension, the second dimension, and by the intersection of the first and second dimensions (sometimes referred to as first-by-second, as in “state-by-year”, clustering). Then we add the first two variance matrices and subtract the third. In the three-way clustering case there is an analogous formula, with seven one-way cluster robust variance matrices computed and combined.

The method is useful in many applications, including:

1. In a cross-section study clustering may arise at several levels simultaneously. For example a model may have errors that are correlated within region, within industry, and within occupation. This leads to inference problems if there are region-level, industry-level, and occupation-level regressors.
2. Clustering may arise due to discrete regressors. Moulton (1986) considered inference in this case, modelling the error correlation using an error components model. More recently, Card and Lee (2004) argue that in a regression discontinuity framework where the treatment-determining variable is discrete, the observations should be clustered at the level of the right-hand side variable. If additionally interest lies in a “primary” dimension of clustering (e.g., state or village), then there is clustering in more than one dimension.
3. In datasets based on pair-wise observations, researchers may wish to allow for clustering at each node of the pair. For example, Rose and Engel (2002) consider estimation of a gravity model for trade flows using a single cross-section with data on many country-pairs, and are unable to control for the likely two-way error correlation across both the first and second country in the pair.
4. Matched employer-employee studies may wish to allow for clustering at both the employer level as well as the employee level when there are repeated observations at the employee level.
5. Studies that employ the usual one-way cluster robust standard errors may wish to additionally control for clustering due to sample design. For example, clustering may occur at the level of a primary sampling unit in addition to the level of an industry-level regressor.
6. Panel studies that employ the usual one-way cluster robust standard errors may wish to additionally control for panel survey design. For example, the Current

Population Survey (CPS) uses a rotating panel structure, with households resurveyed for a number of months. Researchers using data on households or individuals and concerned about within state-year clustering (correlated errors within state-year along with important state-year variables or instruments) should also account for household-level clustering across the two years of the panel structure. Then they need to account for clustering across both dimensions. A related example is Acemoglu and Pischke (2003), who study a panel of individuals who are affected by region-year policy variables.

7. In a state-year panel setting, we may want to cluster at the state level to permit valid inference if there is within-state autocorrelation in the errors. If there is also geographic-based correlation, a similar issue may be at play with respect to the within-year cross-state errors (Conley 1999). In this case, researchers may wish to cluster at the year level as well as at the state level.
8. More generally this situation arises when there is clustering at both a cross-section level and temporal level. For example, finance applications may call for clustering at the firm level and at the time (e.g., day) level.

There are many other situation-specific applications. Papers that cite earlier drafts of our paper include Baughman and Smith (2007), Beck, Demirguc-Kunt, Laeven, and Levine (2008), Cascio and Schanzenbach (2007), Cuijpers and Peek (2008), Engelhardt and Kumar (2007), Foote (2007), Gow, Ormazabal and Taylor (2008), Gurun, Booth and Zhang (2008), Loughran and Shive (2007), Martin, Mayer and Thoenig (2008), Mitchener and Weidenmier, (2007), Olken and Barron (2007), Peress (2007), Pierce and Snyder (2008), and Rountree, Weston and Allayannis (2008).

Our estimator is qualitatively similar to the ones presented in White and Domowitz (1984), for time series data, and Conley (1999), for spatial data. It is based on a weighted double-sum over all observations of the form $\sum_i \sum_j w(i, j) x_i x_j' \hat{\varepsilon}_i \hat{\varepsilon}_j$. White and Domowitz (1984), considering time series dependence, use a weight $w(i, j) = 1$ for observations “close” in time to one another, and $w(i, j) = 0$ for other observations. Conley (1999) considers the case where observations have spatial locations, and has weights $w(i, j)$ to be decaying to 0 as the distance between observations grows. Our estimator can be expressed algebraically as a special case of the spatial HAC (Heteroscedasticity and Autocorrelation Consistent) estimator presented in Conley (1999). Bester, Conley, and Hansen (2009) explicitly consider a setting with spatial or temporal cross-cluster dependence that dies out. These three papers use mixing conditions to ensure that dependence decays as observations as the spatial or temporal distance between observations grows. Such conditions are not applicable to clustering due to common shocks, which have a factor structure rather than decaying dependence. Thus, we rely on independence of observations that share no clusters in common.

The fifth example introduces consideration of sample design, in which case the most precise statistical inference would control for stratification in addition to clustering. Bhattacharya (2005) provides a comprehensive treatment in a GMM framework. He finds that accounting for stratification tends to reduce reported standard errors, and that this effect can be meaningfully large. In his empirical examples, the stratification effect is largest when estimating (unconditional) means and Lorenz shares, and much smaller when estimating conditional means. Like most econometrics studies, we do not control for the effects of stratification. In so doing there will be some over-estimation of the estimator’s standard deviation.

Since the initial draft of this paper, we have become aware of several independent applications of the multi-way robust estimator. Acemoglu and Pischke (2003) estimate OLS standard errors allowing for clustering at the individual level as well as the region-by-time level. Miglioretti and Heagerty (2006) present results for multi-way clustering in the Generalized Estimating Equations setting, and provide simulation results and an application to a mammogram screening epidemiological study. Petersen (2007) compares a number of approaches for OLS estimation in a finance panel setting, using results by Thompson (2005, 2006) that provides some theory and Monte Carlo evidence for the two-way OLS case with panel data on firms. Fafchamps and Gubert (2006) analyze networks among individuals, where a person-pair is the unit of observation. In this context they describe the two-way robust estimator in the setting of dyadic models.

The methods and supporting theory for two-way and multi-way clustering and for both OLS and quite general nonlinear estimators are presented in Section 2 and in the Appendix. Like the one-way cluster-robust method, our methods assume that the number of clusters goes to infinity. This assumption does become more binding with multi-way clustering. For example, in the two-way case it is assumed that $\min(G, H) \rightarrow \infty$, where there are G clusters in dimension 1 and H clusters in dimension 2. In Section 3 we present two different Monte Carlo experiments. The first is based on a two-way random effects model and some extensions of that model. The second follows the general approach of Bertrand et al. (2004) in investigating a placebo law in an earnings regression, except that in our example the induced error dependence is two-way (over both states and years) rather than one-way. Section 4 presents several empirical examples where we contrast results obtained using conventional one-way clustering to those allowing for two-way clustering. Section 5 concludes.

2. Cluster-Robust Inference

This section emphasizes the OLS estimator, for simplicity. We begin with a review of one-way clustering, before considering in turn two-way clustering and multi-way clustering. The section concludes with extension from OLS to m-estimators, such as probit and logit, and GMM estimators.

2.1. One-Way Clustering

The linear model with one-way clustering is

$$y_{ig} = \mathbf{x}'_{ig}\boldsymbol{\beta} + u_{ig}, \quad (2.1)$$

where i denotes the i^{th} of N individuals in the sample, g denotes the g^{th} of G clusters, $E[u_{ig}|\mathbf{x}_{ig}] = 0$, and error independence across clusters is assumed so that for $i \neq j$

$$E[u_{ig}u_{jg'}|\mathbf{x}_{ig}, \mathbf{x}_{jg'}] = 0, \text{ unless } g = g'. \quad (2.2)$$

Errors for individuals belonging to the same group may be correlated, with quite general heteroskedasticity and correlation.

Grouping observations by cluster, the model can be written as

$$\mathbf{y}_g = \mathbf{X}_g\boldsymbol{\beta} + \mathbf{u}_g, \quad (2.3)$$

where \mathbf{y}_g and \mathbf{u}_g are $N_g \times 1$ vectors, \mathbf{X}_g is an $N_g \times K$ matrix, and there are N_g observations in cluster g . Further stacking over clusters yields

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u},$$

where \mathbf{y} and \mathbf{u} are $N \times 1$ vectors, \mathbf{X} is an $N \times K$ matrix, and $N = \sum_g N_g$.

The OLS estimator is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y} = \left(\sum_{g=1}^G \mathbf{X}'_g \mathbf{X}_g \right)^{-1} \sum_{g=1}^G \mathbf{X}'_g \mathbf{y}_g. \quad (2.4)$$

Under commonly assumed restrictions on moments and heterogeneity of the data, $\sqrt{G}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ has a limit normal distribution with variance matrix

$$\left(\lim_{G \rightarrow \infty} \frac{1}{G} \sum_{g=1}^G E[\mathbf{X}'_g \mathbf{X}_g] \right)^{-1} \left(\lim_{G \rightarrow \infty} \frac{1}{G} \sum_{g=1}^G E[\mathbf{X}'_g \mathbf{u}_g \mathbf{u}'_g \mathbf{X}_g] \right) \left(\lim_{G \rightarrow \infty} \frac{1}{G} \sum_{g=1}^G E[\mathbf{X}'_g \mathbf{X}_g] \right)^{-1}. \quad (2.5)$$

If the primary source of clustering is due to group-level common shocks, a useful approximation is that for the j^{th} regressor the default OLS variance estimate based on $s^2(\mathbf{X}'\mathbf{X})^{-1}$, where s is the standard deviation of the error, should be inflated by $\tau_j \simeq 1 + \rho_{x_j}\rho_u(\bar{N}_g - 1)$, where ρ_{x_j} is a measure of the within cluster correlation of x_j , ρ_u is the within cluster error correlation, and \bar{N}_g is the average cluster size; see Kloeck (1981), Scott and Holt (1982) and Greenwald (1983). Moulton (1986, 1990) pointed out that in many settings the adjustment factor τ_j can be large even if ρ_u is small.

The earliest work posited a model for the cluster error variance matrices

$$\mathbf{\Omega}_g = V[\mathbf{u}_g|\mathbf{X}_g] = E[\mathbf{u}_g\mathbf{u}_g'|\mathbf{X}_g], \quad (2.6)$$

in which case $E[\mathbf{X}_g'\mathbf{u}_g\mathbf{u}_g'\mathbf{X}_g] = E[\mathbf{X}_g'\mathbf{\Omega}_g\mathbf{X}_g]$ can be estimated given a consistent estimate $\hat{\mathbf{\Omega}}_g$ of $\mathbf{\Omega}_g$, and feasible GLS estimation is then additionally possible.

Current applied studies instead use the cluster-robust variance matrix estimate

$$\hat{V}[\hat{\beta}] = (\mathbf{X}'\mathbf{X})^{-1} \left(\sum_{g=1}^G \mathbf{X}_g' \hat{\mathbf{u}}_g \hat{\mathbf{u}}_g' \mathbf{X}_g \right) (\mathbf{X}'\mathbf{X})^{-1}, \quad (2.7)$$

where $\hat{\mathbf{u}}_g = \mathbf{y}_g - \mathbf{X}_g\hat{\beta}$. This provides a consistent estimate of the variance matrix if $G^{-1} \sum_{g=1}^G \mathbf{X}_g' \hat{\mathbf{u}}_g \hat{\mathbf{u}}_g' \mathbf{X}_g - G^{-1} \sum_{g=1}^G E[\mathbf{X}_g' \mathbf{u}_g \mathbf{u}_g' \mathbf{X}_g] \xrightarrow{p} \mathbf{0}$ as $G \rightarrow \infty$. White (1984, p.134-142) presented formal theorems for a multivariate dependent variable, directly applicable to balanced clusters. Liang and Zeger (1986) proposed this method for estimation in a generalized estimating equations setting, Arellano (1987) for the fixed effects estimator in linear panel models, and Rogers (1993) popularized this method in applied econometrics by incorporating it in Stata. The method generalizes White (1980), who considered the case $N_g = 1$. Most recently, Hansen (2007) provides asymptotic theory for panel data where $T \rightarrow \infty$ ($N_g \rightarrow \infty$ in the notation above) in addition to $N \rightarrow \infty$ ($G \rightarrow \infty$ in the notation above). Note that (2.7) does not require specification of a model for $\mathbf{\Omega}_g$, and thus it permits quite general forms of $\mathbf{\Omega}_g$.

A helpful informal presentation of (2.7) is that

$$\hat{V}[\hat{\beta}] = (\mathbf{X}'\mathbf{X})^{-1} \hat{\mathbf{B}} (\mathbf{X}'\mathbf{X})^{-1}, \quad (2.8)$$

where the central matrix

$$\begin{aligned} \hat{\mathbf{B}} &= \sum_{g=1}^G \mathbf{X}_g' \hat{\mathbf{u}}_g \hat{\mathbf{u}}_g' \mathbf{X}_g \\ &= \mathbf{X}' \begin{bmatrix} \hat{\mathbf{u}}_1 \hat{\mathbf{u}}_1' & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \hat{\mathbf{u}}_2 \hat{\mathbf{u}}_2' & & \vdots \\ \vdots & & \ddots & \mathbf{0} \\ \mathbf{0} & \cdots & \cdots & \hat{\mathbf{u}}_G \hat{\mathbf{u}}_G' \end{bmatrix} \mathbf{X} \\ &= \mathbf{X}' \left(\hat{\mathbf{u}} \hat{\mathbf{u}}' . * \begin{bmatrix} \mathbf{E}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{E}_2 & & \vdots \\ \vdots & & \ddots & \\ \mathbf{0} & \cdots & \cdots & \mathbf{E}_G \end{bmatrix} \right) \mathbf{X}, \end{aligned} \quad (2.9)$$

where $*$ denotes element-by-element multiplication and \mathbf{E}_g is an $(N_g \times N_g)$ matrix of ones.

More generally we can view $\hat{\mathbf{B}}$ in (2.9) as being given by

$$\hat{\mathbf{B}} = \mathbf{X}'(\hat{\mathbf{u}}\hat{\mathbf{u}}' * \mathbf{S}^G)\mathbf{X} \quad (2.10)$$

where \mathbf{S}^G is an $N \times N$ indicator, or selection, matrix with ij^{th} entry equal to one if the i^{th} and j^{th} observation belong to the same cluster and equal to zero otherwise. \mathbf{S}^G in turn equals $\Delta^G \Delta^{G'}$ where Δ^G is an $N \times G$ matrix with ig^{th} entry equal to one if the i^{th} observation belongs to cluster g and equal to zero otherwise. The (a, b) -th element of $\hat{\mathbf{B}}$ is $\sum_{i=1}^N \sum_{j=1}^N x_{ia} x_{jb} \hat{u}_i \hat{u}_j \mathbf{1}[i, j \text{ in same cluster}]$, where $\hat{u}_i = y_i - \mathbf{x}_i' \boldsymbol{\beta}$.

An intuitive explanation of the asymptotic theory is that the indicator matrix \mathbf{S}^G must zero out a large amount of $\hat{\mathbf{u}}\hat{\mathbf{u}}'$, or, asymptotically equivalently, $\mathbf{u}\mathbf{u}'$. Here there are $N^2 = (\sum_{g=1}^G N_g)^2$ terms in $\hat{\mathbf{u}}\hat{\mathbf{u}}'$ and all but $\sum_{g=1}^G N_g^2$ of these are zeroed out. For fixed N_g , $\sum_{g=1}^G N_g^2 / N^2 \rightarrow 0$ as $G \rightarrow \infty$. In particular, for balanced clusters $N_g = N/G$, so $\sum_{g=1}^G N_g^2 / N^2 = 1/G \rightarrow 0$ as $G \rightarrow \infty$.

2.2. Two-Way Clustering

Now consider situations where each observation may belong to more than one “dimension” of groups. For instance, if there are two dimensions of grouping, each individual will belong to a group $g \in \{1, 2, \dots, G\}$, as well as to a group $h \in \{1, 2, \dots, H\}$, and we have

$$y_{igh} = \mathbf{x}_{igh}' \boldsymbol{\beta} + u_{igh}, \quad (2.11)$$

where we assume that for $i \neq j$

$$E[u_{igh} u_{jg'h'} | \mathbf{x}_{igh}, \mathbf{x}_{jg'h'}] = 0, \text{ unless } g = g' \text{ or } h = h'. \quad (2.12)$$

If errors belong to the same group (along either dimension), they may have an arbitrary correlation. For non-nested two-way clustering, which we consider, $\boldsymbol{\Omega} = V[\mathbf{u}|\mathbf{X}]$ can no longer be written as a block diagonal matrix.

The intuition for the variance estimator in this case is a simple extension of (2.10) for one-way clustering. Instead of keeping only those elements of $\hat{\mathbf{u}}\hat{\mathbf{u}}'$ where the i^{th} and j^{th} observations share a cluster in one specified dimension, we keep those elements of $\hat{\mathbf{u}}\hat{\mathbf{u}}'$ where the i^{th} and j^{th} observations share a cluster in any dimension. Then

$$\hat{\mathbf{B}} = \mathbf{X}'(\hat{\mathbf{u}}\hat{\mathbf{u}}' * \mathbf{S}^{GH})\mathbf{X}, \quad (2.13)$$

where \mathbf{S}^{GH} is an $N \times N$ indicator matrix with ij^{th} entry equal to one if the i^{th} and j^{th} observation share any cluster, and equal to zero otherwise. Now, the (a, b) -th element of $\hat{\mathbf{B}}$ is $\sum_{i=1}^N \sum_{j=1}^N x_{ia} x_{jb} \hat{u}_i \hat{u}_j \mathbf{1}[i, j \text{ share any cluster}]$.

$\widehat{\mathbf{B}}$ and hence $\widehat{\mathbf{V}}[\widehat{\boldsymbol{\beta}}]$ can be calculated using matrix algebra. The $N \times N$ selection matrix \mathbf{S}^{GH} may be large in some problems, however, and even if N is manageable many users will prefer to use readily available software that calculates cluster-robust standard errors for one-way clustering.

This is done by defining three $N \times N$ indicator matrices: \mathbf{S}^G with ij^{th} entry equal to one if the i^{th} and j^{th} observation belong to the same cluster $g \in \{1, 2, \dots, G\}$, \mathbf{S}^H with ij^{th} entry equal to one if the i^{th} and j^{th} observation belong to the same cluster $h \in \{1, 2, \dots, H\}$, and $\mathbf{S}^{G \cap H}$ with ij^{th} entry equal to one if the i^{th} and j^{th} observation belong to both the same cluster $g \in \{1, 2, \dots, G\}$ and the same cluster $h \in \{1, 2, \dots, H\}$. Then

$$\mathbf{S}^{GH} = \mathbf{S}^G + \mathbf{S}^H - \mathbf{S}^{G \cap H},$$

so

$$\widehat{\mathbf{B}} = \mathbf{X}'(\widehat{\mathbf{u}}\widehat{\mathbf{u}}' * \mathbf{S}^G)\mathbf{X} + \mathbf{X}'(\widehat{\mathbf{u}}\widehat{\mathbf{u}}' * \mathbf{S}^H)\mathbf{X} - \mathbf{X}'(\widehat{\mathbf{u}}\widehat{\mathbf{u}}' * \mathbf{S}^{G \cap H})\mathbf{X}. \quad (2.14)$$

Substituting (2.14) into (2.8) yields

$$\begin{aligned} \widehat{\mathbf{V}}[\widehat{\boldsymbol{\beta}}] &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\widehat{\mathbf{u}}\widehat{\mathbf{u}}' * \mathbf{S}^G)\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &\quad + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\widehat{\mathbf{u}}\widehat{\mathbf{u}}' * \mathbf{S}^H)\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &\quad - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\widehat{\mathbf{u}}\widehat{\mathbf{u}}' * \mathbf{S}^{G \cap H})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}. \end{aligned} \quad (2.15)$$

The three components can be separately computed by

1. OLS regression of \mathbf{y} on \mathbf{X} with variance matrix estimate computed using clustering on $g \in \{1, 2, \dots, G\}$;
2. OLS regression of \mathbf{y} on \mathbf{X} with variance matrix estimate computed using clustering on $h \in \{1, 2, \dots, H\}$; and
3. OLS regression of \mathbf{y} on \mathbf{X} with variance matrix estimate computed using clustering on $(g, h) \in \{(1, 1), \dots, (G, H)\}$.

Given these three components, $\widehat{\mathbf{V}}[\widehat{\boldsymbol{\beta}}]$ is computed as the sum of the first and second components, minus the third component.

2.3. Practical considerations

In much the same way that robust inference in the presence of one-way clustering requires empirical researchers to know which “way” is the one where clustering may be important, our discussion presumes that the researcher knows what “ways” will be potentially important for clustering in her application.

It would be useful to have an objective way to determine which, and how many, dimensions require allowance for clustering. Unfortunately, we are presently unaware

of a systematic, data-driven approach to this issue. From the discussion after (2.5) a necessary condition for a dimension (e.g., state, industry, occupation, year) to exhibit clustering is that there be correlation in the errors within that dimension of the data. This effect is exacerbated by regressors that also exhibit correlation in that dimension. The impact of controlling for two-way clustering is likely to be greatest when both the regressor of interest and the error, conditional on the other regressors, are correlated over two dimensions.

In principle, we believe that one could formulate tests based on conditional moments, similar to the White (1980) test for heteroskedasticity. Such an approach would likely involve using sample covariances of $\mathbf{X}'\hat{\mathbf{u}}$ terms within dimensions to test the null hypothesis that the average of such covariances is zero. Rejecting this null would be sufficient, though not necessary, to reject the null hypothesis of no clustering in a dimension. The difficulty in making such a test operational is finding a way to partial out, or otherwise hold constant, the part of the covariance average in one dimension that overlaps with other dimensions. While we think this could be a fruitful direction for future research, we also believe that this topic is too far afield for detailed treatment in the present paper.

Small-sample modifications of (2.7) for one-way clustering are typically used, since without modification the cluster-robust standard errors are biased downwards. Cameron, Gelbach, and Miller (2008) review various small-sample corrections that have been proposed in the literature, for both standard errors and for inference using resultant Wald statistics. For example, Stata uses $\sqrt{c}\hat{\mathbf{u}}_g$ in (2.7) rather than $\hat{\mathbf{u}}_g$, with $c = \frac{G}{G-1} \frac{N-1}{N-K} \simeq \frac{G}{G-1}$. Similar corrections may be used for two-way clustering. One method is to use the Stata formula throughout, in which case the errors in the three components are multiplied by, respectively, $c_1 = \frac{G}{G-1} \frac{N-1}{N-K}$, $c_2 = \frac{H}{H-1} \frac{N-1}{N-K}$ and $c_3 = \frac{I}{I-1} \frac{N-1}{N-K}$ where I equals the number of unique clusters formed by the intersection of the H groups and the G groups. A second is to use a constant $c = \frac{J}{J-1} \frac{N-1}{N-K}$ where $J = \min(G, H)$. We use the first of these methods in our OLS simulations and applications.

Some statistical packages, for example Stata, permit separate estimation of the variance matrices using stored estimation results. In this case one need only estimate β and invert $(\mathbf{X}'\mathbf{X})$ once. As a result estimating $\hat{\mathbf{V}}[\hat{\beta}]$ often adds little computational time over that of one-way cluster-robust inference.

A practical matter that can arise when implementing the two-way robust estimator is that the resulting variance estimate $\hat{\mathbf{V}}[\hat{\beta}]$ may have negative elements on the diagonal. Using the Stata-style formula for residual adjustment reduces the likelihood of estimating a negative variance, especially when the number of clusters is small, because this problem is more likely to arise when the third covariance matrix is relatively large and the Stata-style formula uses a smaller (inflationary) adjustment to the standard errors in the third matrix, since $I > \max(G, H)$.

In some applications with fixed effects, $\widehat{V}[\widehat{\beta}]$ may be non positive-definite, but the subcomponent of $\widehat{V}[\widehat{\beta}]$ associated with the regressors of interest may be positive-definite. In some statistical package programs this may lead to a reported error, even though inference is appropriate for the parameters of interest. Our informal observation is that this issue is most likely to arise when clustering is done over the same groups as the fixed effects. In that case eliminating the fixed effects by differencing, rather than directly estimating them, leads to a positive definite matrix for the remaining coefficients.

In some applications and simulations it can still be the case that the variance-covariance matrix is not positive-semidefinite. A positive-semidefinite matrix can be created by employing a technique used in the time series HAC literature, such as in Politis (2007). This involves three steps. First decompose the variance matrix defined in (2.15) into the product of its eigenvectors and eigenvalues: $\widehat{V}[\widehat{\beta}] = U\Lambda U'$, with U containing the eigenvectors of \widehat{V} , and $\Lambda = \text{Diag}[\lambda_1, \dots, \lambda_d]$ containing the eigenvalues of \widehat{V} . Then create $\Lambda^+ = \text{Diag}[\lambda_1^+, \dots, \lambda_d^+]$, with $\lambda_j^+ = \max(0, \lambda_j)$, and use $\widehat{V}^+[\widehat{\beta}] = U\Lambda^+U'$ as the variance estimate. In some of our simulations with a small number of clusters ($G, H = 10$) we very occasionally obtained a non positive-semidefinite variance matrix and dropped that draw from our Monte Carlo analysis. When we instead use the above method we find that in the problematic draws the negative eigenvalue is small, $\widehat{V}^+[\widehat{\beta}]$ always yields a positive definite variance matrix estimate, and keeping all draws (using $\widehat{V}^+[\widehat{\beta}]$ where necessary) leads to results very similar to those reported in the Monte Carlos below.

Most empirical studies with clustered data estimate by OLS, ignoring potential efficiency gains due to modeling heteroskedasticity and/or clustering and estimating by feasible GLS. The method outlined in this paper can be adapted to weighted least squares that accounts for heteroskedasticity, as the resulting residuals \widehat{u}_{igh}^* from the transformed model will asymptotically retain the same broad correlation pattern over g and h . It can also be adapted to robustify a one-way random effects feasible GLS estimator that clusters over g , say, when there is also correlation over h . Then the random effects transformation will induce some correlation across h and h' between transformed errors u_{igh}^* and $u_{ig'h'}^*$, but this correlation is negligible as $G \rightarrow \infty$ and $H \rightarrow \infty$.

In some applications researchers will wish to include fixed effects in one or both dimensions. We do not formally address this complication. However, we note that given our assumption that $G \rightarrow \infty$ and $H \rightarrow \infty$, each fixed effect is estimated using many observations. We think that this is likely to mitigate the incidental parameters problem in nonlinear models such as the probit model. We find in practice that the main consequence of including fixed effects is a reduction in within cluster correlation.

As discussed in the appendix, the relative importance of the third variance term in (2.15) varies with the type of application. For a classic two-way random effects model

with common shocks and no fixed effects the third term is dominated by the first two variance terms asymptotically. But in other applications (for example with iid errors) the third term can be of similar order to the first two terms asymptotically. In both cases we recommend use of all three terms in finite-sample practice.

2.4. Multi-Way Clustering

Our approach generalizes to clustering in more than two dimensions. We now give a quite general treatment that requires some new notation and definitions.

Suppose there are D dimensions within which clustering must be accounted for. For example, if we want to cluster on industry, occupation, and state, then $D = 3$. Let G_d denote the number of clusters in dimension d . Let the D -vector $\boldsymbol{\delta}_i = \boldsymbol{\delta}(i)$, where the function $\boldsymbol{\delta} : \{1, 2, \dots, N\} \rightarrow \times_{d=1}^D \{1, 2, \dots, G_d\}$ lists the cluster membership in each dimension for each observation. For example, if $\boldsymbol{\delta}_i = (5, 8, 2)$ then there are three dimensions and the i^{th} observation is in the fifth cluster in the first dimension, the eighth cluster in the second dimension, and the second cluster in the third dimension. Thus $\mathbf{1}[i, j \text{ share a cluster}] = 1 \Leftrightarrow \delta_{id} = \delta_{jd}$ for some $d \in \{1, 2, \dots, D\}$, where δ_{id} denotes the d^{th} element of $\boldsymbol{\delta}_i$.

Now let \mathbf{r} be a D -vector, with d^{th} coordinate equal to r_d , and define the set $R \equiv \{\mathbf{r}: r_d \in \{0, 1\}, d = 1, 2, \dots, D, \mathbf{r} \neq \mathbf{0}\}$, where the exclusion of the vector $\mathbf{0}$ means that R has $2^D - 1$ elements. For example, for $D = 3$ we have $R = \{(1, 0, 0), (0, 1, 0), (0, 0, 1), (1, 1, 0), (1, 0, 1), (0, 1, 1), (1, 1, 1)\}$. Elements of the set R can be used to index all cases in which two observations share a cluster in at least one dimension. To see how, define the indicator function $I_{\mathbf{r}}(i, j) \equiv \mathbf{1}[r_d \delta_{id} = r_d \delta_{jd}, \forall d]$. This function tells us whether observations i and j have identical cluster membership for *all* dimensions d such that $r_d = 1$. For example, with $D = 3$ and $\mathbf{r} = (1, 1, 0)$, $I_{\mathbf{r}}(i, j) = 1$ if and only if $(\delta_{i1}, \delta_{i2}) = (\delta_{j1}, \delta_{j2})$, so that i and j are in the same group in dimensions 1 and 2 (regardless of whether $\delta_{i3} = \delta_{j3}$). Suppose that the three clustering dimensions are industry, occupation, and U.S. state. For the vector $\mathbf{r} = (1, 1, 0)$, $I_{\mathbf{r}}(i, j) = 1$ if and only if the two observations share an industry and an occupation, regardless of whether or not they share a U.S. state. Similarly, if $\mathbf{r} = (1, 1, 1)$, $I_{\mathbf{r}}(i, j) = 1$ if and only if the two observations share industry, occupation, and U.S. state. Define $I(i, j) = 1$ if and only if $I_{\mathbf{r}}(i, j) = 1$ for some $\mathbf{r} \in R$. Thus, $I(i, j) = 1$ if and only if the two observations share *at least* one dimension.

Now define the $2^D - 1$ matrices

$$\tilde{\mathbf{B}}_{\mathbf{r}} \equiv \sum_{i=1}^N \sum_{j=1}^N \mathbf{x}_i \mathbf{x}_j' \hat{u}_i \hat{u}_j I_{\mathbf{r}}(i, j), \quad \mathbf{r} \in R. \quad (2.16)$$

For example, if $D = 2$, then $\hat{\mathbf{B}}$ in (2.14) can be expressed in the new notation as $\hat{\mathbf{B}} = \tilde{\mathbf{B}}_{(1,0)} + \tilde{\mathbf{B}}_{(0,1)} - \tilde{\mathbf{B}}_{(1,1)}$. And if $D = 3$ and $\mathbf{r} = (1, 1, 0)$, then $\tilde{\mathbf{B}}_{\mathbf{r}}$ is the middle matrix

we get when we cluster on the variable $I_{(1,1,0)}$; when the first two dimensions are industry and occupation, this is the matrix we get when we cluster on industry-occupation cells.

Our proposed estimator may be written as

$$\widehat{V}[\widehat{\beta}] = (\mathbf{X}'\mathbf{X})^{-1}\widetilde{\mathbf{B}}(\mathbf{X}'\mathbf{X})^{-1}, \quad (2.17)$$

where

$$\widetilde{\mathbf{B}} \equiv \sum_{\|\mathbf{r}\|=k, \mathbf{r} \in R} (-1)^{k+1} \widetilde{\mathbf{B}}_{\mathbf{r}}. \quad (2.18)$$

Thus we sum over all possible values of $\|\mathbf{r}\| = \sum_d r_d$. Cases in which the matrix $\widetilde{\mathbf{B}}_{\mathbf{r}}$ involves clustering on an odd number of dimensions are added, while those involving clustering on an even number are subtracted (note that $\|\mathbf{r}\| \leq D$ for all $\mathbf{r} \in R$).

As an example, when $D = 3$, $\widetilde{\mathbf{B}}$ may be written as

$$\left(\widetilde{\mathbf{B}}_{(1,0,0)} + \widetilde{\mathbf{B}}_{(0,1,0)} + \widetilde{\mathbf{B}}_{(0,0,1)} \right) - \left(\widetilde{\mathbf{B}}_{(1,1,0)} + \widetilde{\mathbf{B}}_{(1,0,1)} + \widetilde{\mathbf{B}}_{(0,1,1)} \right) + \widetilde{\mathbf{B}}_{(1,1,1)}.$$

Each of the first three matrices clusters on exactly one dimension. In some cases, observation pairs are in the same cluster in dimensions one and two; thus if we included only the first three matrices, we would double-count these pairs. Thus we cluster on each of the three combinations of two dimensions and subtract the resulting matrices, eliminating double-counting of such pairs. However, some observation pairs share the same cluster in all three dimensions; if we stopped after the first six matrices, these pairs would be included three times and excluded three times, so that they would not be accounted for. Hence we add back the seventh matrix, which is the clustering matrix for observation pairs sharing the same cluster on all dimensions (e.g., industry-occupation cells within state).

To prove that this approach is identical to the earlier one, so that $\widetilde{\mathbf{B}} = \widehat{\mathbf{B}}$ identically, it is sufficient to show that (i) no observation pair with $I(i, j) = 0$ is included, and (ii) the covariance term corresponding to each observation pair with $I(i, j) = 1$ is included exactly once in $\widetilde{\mathbf{B}}$. The first result is immediate, since $I(i, j) = 0$ if and only if $I_{\mathbf{r}}(i, j) = 0$ for all \mathbf{r} (see above). The second result follows because it is straightforward to show by induction that when $I(i, j) = 1$,

$$\sum_{\|\mathbf{r}\|=k, \mathbf{r} \in R} (-1)^{k+1} I_{\mathbf{r}}(i, j) = 1.$$

(Actually, the first result also follows using this expression, since the left hand side is 0 when all $I_{\mathbf{r}}(i, j) = 0$.) This fact, which can also be shown to be an application of the inclusion-exclusion principle for set cardinality, ensures that $\widetilde{\mathbf{B}}$ and $\widehat{\mathbf{B}}$ are numerically identical in every sample.

As a practical matter, the inclusion-exclusion approach may be computationally dominated by direct computation of (2.16) whenever D is relatively large. This is because the computational cost of this approach grows at rate $2^D - 1$. However, our experience suggests that when D is small (*e.g.*, 2 or 3), it may be quicker to use the inclusion-exclusion approach.

A related concern is the possibility of a curse of dimensionality with multi-way clustering. This could arise in a setting with many dimensions of clustering, and in which one or more dimensions have few clusters. The square design (where each dimension has the same number of clusters) with orthogonal dimensions (for example, 30 states by 30 years by 30 industries) has the least independence of observations. In this setting on average a fraction $\frac{D}{G}$ observations will be potentially related to one another. While this has a multiplier of D , it always decays at a rate G (since D is fixed). We suggest an ad-hoc rule of thumb for approximating sufficient numbers of clusters - if G_1 would be a sufficient number with one-way clustering, then DG_1 should be a sufficient number with D -way clustering. In the rectangular case (*e.g.* with 20 years and 50 states and 200 industries) the curse of dimensionality is lessened.

2.5. Multi-way Clustering for m-estimators and GMM Estimators

The preceding analysis considered the OLS estimator. More generally we consider multi-way clustering for other (nonlinear) regression estimators commonly used in econometrics.

We begin with an m-estimator that solves

$$\sum_{i=1}^N \mathbf{h}_i(\hat{\boldsymbol{\theta}}) = \mathbf{0}. \quad (2.19)$$

Examples include nonlinear least squares estimation, maximum likelihood estimation, and instrumental variables estimation in the just-identified case. For the probit MLE $\mathbf{h}_i(\boldsymbol{\beta}) = (y_i - \Phi(\mathbf{x}_i'\boldsymbol{\beta}))\phi(\mathbf{x}_i'\boldsymbol{\beta})\mathbf{x}_i/[\Phi(\mathbf{x}_i'\boldsymbol{\beta})(1 - \Phi(\mathbf{x}_i'\boldsymbol{\beta}))]$, where $\Phi(\cdot)$ and $\phi(\cdot)$ are the standard normal cdf and density.

Under standard assumptions, $\hat{\boldsymbol{\theta}}$ is asymptotically normal with estimated variance matrix

$$\hat{\mathbf{V}}[\hat{\boldsymbol{\theta}}] = \hat{\mathbf{A}}^{-1}\hat{\mathbf{B}}\hat{\mathbf{A}}'^{-1},$$

where $\hat{\mathbf{A}} = \sum_i \frac{\partial \mathbf{h}_i}{\partial \boldsymbol{\theta}'} \Big|_{\hat{\boldsymbol{\theta}}}$ or $\hat{\mathbf{A}} = \sum_i \mathbf{E} \left[\frac{\partial \mathbf{h}_i}{\partial \boldsymbol{\theta}'} \right] \Big|_{\hat{\boldsymbol{\theta}}}$, and $\hat{\mathbf{B}}$ is an estimate of $\mathbf{V}[\sum_i \mathbf{h}_i]$.

Computation of $\hat{\mathbf{B}}$ varies with assumptions about clustering. Given independence over i , $\mathbf{V}[\sum_i \mathbf{h}_i] = \sum_i \mathbf{V}[\mathbf{h}_i]$ and $\hat{\mathbf{B}} = \sum_{i=1}^N \hat{\mathbf{h}}_i \hat{\mathbf{h}}_i'$, where $\hat{\mathbf{h}}_i = \mathbf{h}_i(\hat{\boldsymbol{\theta}})$. Note that for OLS $\hat{\mathbf{h}}_i = \hat{u}_i \hat{\mathbf{x}}_i$, so $\hat{\mathbf{B}} = \sum_{i=1}^N \hat{u}_i^2 \hat{\mathbf{x}}_i \hat{\mathbf{x}}_i'$, leading to White's heteroskedastic consistent estimate.

For one-way clustering $\hat{\mathbf{B}} = \sum_{g=1}^G \hat{\mathbf{h}}_g \hat{\mathbf{h}}_g'$ where $\hat{\mathbf{h}}_g = \sum_{i=1}^{N_g} \hat{\mathbf{h}}_{ig}$. Clustering may or may not lead to parameter inconsistency, depending on whether $\mathbf{E}[\mathbf{h}_i(\boldsymbol{\theta})] = \mathbf{0}$ in the

presence of clustering. As an example consider a probit model with one-way clustering. One approach, called a population-averaged approach in the statistics literature is to assume that $E[y_{ig}|\mathbf{x}_{ig}] = \Phi(\mathbf{x}'_{ig}\boldsymbol{\beta})$, even in the presence of clustering. An alternative approach is a random effects approach. Let $y_{ig} = 1$ if $y_{ig}^* > 0$ where $y_{ig}^* = \mathbf{x}'_{ig}\boldsymbol{\beta} + \varepsilon_g + \varepsilon_{ig}$, where the idiosyncratic error $\varepsilon_{ig} \sim \mathcal{N}[0, 1]$ as usual and the cluster-specific error $\varepsilon_g \sim \mathcal{N}[0, \sigma_g^2]$. Then it can be shown that $E[y_{ig}|\mathbf{x}_{ig}] = \Phi(\mathbf{x}'_{ig}\boldsymbol{\beta}/\sqrt{1 + \sigma_g^2})$, so that the moment condition is no longer $E[y_{ig}|\mathbf{x}_{ig}] = \Phi(\mathbf{x}'_{ig}\boldsymbol{\beta})$. When $E[\mathbf{h}_i(\boldsymbol{\theta})] \neq \mathbf{0}$ the estimated variance matrix is still as above, but the distribution of the estimator will be instead centered on a pseudo-true value (White, 1982). For the probit model the average partial effect is nonetheless consistently estimated (Wooldridge 2002, pg. 471).

Our concern is with multiway clustering. The analysis of the preceding section carries through, with $\hat{u}_i\mathbf{x}_i$ in (2.16) replaced by $\hat{\mathbf{h}}_i$. Then $\hat{\boldsymbol{\theta}}$ is asymptotically normal with estimated variance matrix

$$\hat{\mathbf{V}}[\hat{\boldsymbol{\theta}}] = \hat{\mathbf{A}}^{-1}\tilde{\mathbf{B}}\hat{\mathbf{A}}'^{-1}, \quad (2.20)$$

where as usual

$$\hat{\mathbf{A}} = \sum_i \frac{\partial \mathbf{h}_i}{\partial \boldsymbol{\theta}'} \bigg|_{\hat{\boldsymbol{\theta}}}, \quad (2.21)$$

or $\hat{\mathbf{A}} = \sum_i E \left[\frac{\partial \mathbf{h}_i}{\partial \boldsymbol{\theta}'} \right] \bigg|_{\hat{\boldsymbol{\theta}}}$, and now

$$\tilde{\mathbf{B}} \equiv \sum_{\|\mathbf{r}\|=k, \mathbf{r} \in R} (-1)^{k+1} \tilde{\mathbf{B}}_{\mathbf{r}}, \quad (2.22)$$

as in (2.18), with the $2^D - 1$ matrices $\tilde{\mathbf{B}}_{\mathbf{r}}$ defined analogously to (2.16) as

$$\tilde{\mathbf{B}}_{\mathbf{r}} \equiv \sum_{i=1}^N \sum_{j=1}^N \hat{\mathbf{h}}_i \hat{\mathbf{h}}'_j I_{\mathbf{r}}(i, j), \quad \mathbf{r} \in R. \quad (2.23)$$

Implementation is similar to before. For example, for two-way clustering in the probit model estimate the three components separately by

1. Probit regression of \mathbf{y} on \mathbf{X} with variance matrix estimate computed using clustering on $g \in \{1, 2, \dots, G\}$;
2. Probit regression of \mathbf{y} on \mathbf{X} with variance matrix estimate computed using clustering on $h \in \{1, 2, \dots, H\}$; and
3. Probit regression of \mathbf{y} on \mathbf{X} with variance matrix estimate computed using clustering on $(g, h) \in \{(1, 1), \dots, (G, H)\}$.

Given these three components, $\widehat{V}[\widehat{\beta}]$ is computed as the sum of the first and second components, minus the third component.

Commonly-used examples of nonlinear estimators to which this method can be applied are nonlinear-least squares, just-identified instrumental variables estimation, logit, probit and Poisson. In the case of Poisson, for example, the method controls for under-dispersion or overdispersion in addition to multiway clustering.

The standard small-sample correction for standard errors of these nonlinear estimators in the one-way clustering case leads to use of $\sqrt{c_{\mathbf{r}}}\widehat{\mathbf{h}}_i$ rather than $\widehat{\mathbf{h}}_i$ in (2.23), where $c_{\mathbf{r}} = G_{\mathbf{r}}/(G_{\mathbf{r}} - 1)$ and $G_{\mathbf{r}}$ is the number of clusters defined by \mathbf{r} . We use this adjustment, which is used in Stata, in our probit application in Section 4.2.

If the estimator under consideration is one for which a package does not provide one-way cluster-robust standard errors it is possible to implement our procedure using several one-way clustered bootstraps. In the two-way clustered probit example above, in step 1 do a pairs cluster bootstrap that resamples with replacement from the G clusters, $(y_1, \mathbf{X}_1), \dots, (y_G, \mathbf{X}_G)$, in step 2 do a pairs cluster bootstrap that resamples with replacement from the H clusters, $(y_1, \mathbf{X}_1), \dots, (y_H, \mathbf{X}_H)$, and in step 3 do a pairs cluster bootstrap that resamples with replacement using clustering on $(g, h) \in \{(1, 1), \dots, (G, H)\}$. The resulting three separate variance matrix estimates are then combined as before – add the first two and subtract the third. This bootstrap provides the same level of asymptotic approximation as that without bootstrap, and does not additionally provide an asymptotic refinement (see Cameron et al. (2008) for a discussion of clustering and asymptotic refinement in the one-way case).

Finally we consider GMM estimation for over-identified models. A leading example is linear two stage least squares with more instruments than endogenous regressors. Then $\widehat{\theta}$ minimizes

$$Q(\theta) = \left(\sum_{i=1}^N \mathbf{h}_i(\theta) \right)' \mathbf{W} \left(\sum_{i=1}^N \mathbf{h}_i(\theta) \right),$$

where \mathbf{W} is a symmetric positive definite weighting matrix. Under standard regularity conditions $\widehat{\theta}$ is asymptotically normal with estimated variance matrix

$$\widehat{V}[\widehat{\theta}] = \left(\widehat{\mathbf{A}}' \mathbf{W} \widehat{\mathbf{A}} \right)^{-1} \widehat{\mathbf{A}}' \mathbf{W} \widetilde{\mathbf{B}} \mathbf{W} \widehat{\mathbf{A}} \left(\widehat{\mathbf{A}}' \mathbf{W} \widehat{\mathbf{A}} \right)^{-1}, \quad (2.24)$$

where $\widehat{\mathbf{A}}$ is defined in (2.21), and $\widetilde{\mathbf{B}}$ is an estimate of $V[\sum_i \mathbf{h}_i]$ that can be computed using (2.22) and (2.23).

The procedure is qualitatively the same as for OLS and m-estimation. In the two-way clustering case, we obtain three different cluster-robust variance matrices for the GMM estimator by one-way clustering in, respectively, the first dimension, the second dimension, and after grouping by the intersection of the first and second dimensions. Then we add the first two variance matrices and subtract the third.

3. Monte Carlo Exercises

In this section we analyze the size performance of Wald tests based on standard errors, rather than on the standard errors per se, in two different settings for two-way clustering. We compare the rejection rates of Wald tests based on alternative standard error estimates and, in the first example, investigate the performance of our asymptotically-justified method when there are few clusters.

3.1. Monte Carlo based on Two-way Random Effects Errors

The first Monte Carlo exercise is based on a two-way random effects model for the errors. This has the advantage of providing a more parsimonious competitor, a Moulton-type correction that assumes the error process is that of a two-way random effects model. We eventually introduce group-level heteroskedasticity into the errors that can be accommodated by our two-way cluster-robust method, but not by the other methods.

We consider the following data generating process for two-way clustering

$$y_{igh} = \beta_0 + \beta_1 x_{1igh} + \beta_2 x_{2igh} + u_{igh}, \quad (3.1)$$

where $\beta_0 = \beta_1 = \beta_2 = 1$ throughout. The regressors x_{1igh} and x_{2igh} and the errors u_{igh} vary with the experiment performed, as described below. We use rectangular designs with exactly one observation drawn from each (g, h) pair, leading to $G \times H$ observations. The subscript i in (3.1) is then redundant, and is suppressed in the subsequent discussion. The first ten designs are square with $G = H$ varying from 10 to 100 in increments of 10, and the remaining designs are rectangular with $G < H$.

We consider inference based on the OLS slope coefficients $\hat{\beta}_1$ and $\hat{\beta}_2$, reporting empirical rejection probabilities for asymptotic two-sided tests of whether $\beta_1 = 1$ or $\beta_2 = 1$. That is we report in adjacent columns the percentage of times

$$t_1 = \left| \frac{\hat{\beta}_1 - 1}{\text{se}[\hat{\beta}_1]} \right| \geq 1.96, \text{ and } t_2 = \left| \frac{\hat{\beta}_2 - 1}{\text{se}[\hat{\beta}_2]} \right| \geq 1.96.$$

Since the Wald test statistic is asymptotically normal, asymptotically rejection should occur 5% of the time. As a small-sample adjustment for two-way cluster-robust standard errors, discussed below, we also report rejection rates when the critical value is $t_{.025; \min(G, H) - 1}$.

The standard errors $\text{se}[\hat{\beta}_1]$ and $\text{se}[\hat{\beta}_2]$ used to construct the Wald statistics are computed in several ways:

1. Assume iid errors: This uses the “default” variance matrix estimate $\hat{s}^2(\mathbf{X}'\mathbf{X})^{-1}$.

2. One-way cluster-robust (cluster on first group): This uses one-way cluster-robust standard errors, based on (2.7) with small-sample modification, that correct for clustering on the first grouping $g \in \{1, 2, \dots, G\}$ but not the second grouping.
3. Two-way random effects correction: This assumes a two-way random effects model for the error and gives Moulton-type corrected standard errors calculated from $\widehat{V}[\widehat{\beta}] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\widehat{\Omega}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$, where $\widehat{\Omega}$ is a consistent estimate of $V[u]$ based on assuming two-way random effects errors ($u_{gh} = \varepsilon_g + \varepsilon_h + \varepsilon_{gh}$ where the three error components are iid).
4. Two-way cluster-robust: This is the method of this paper, given in (2.15), that allows for two-way clustering but does not restrict it to follow a two-way random effects model.

Tables 1-3 use 2,000 simulations, which yields a 95% confidence interval of (4.0%, 6.0%) for the Monte Carlo rejection rate, given that the true rejection rate is 5%.

3.1.1. Dgp with no clustering

Table 1 reports results for a dgp with iid errors and regressors. Specifically $u_{gh} = \varepsilon_{gh} \sim \mathcal{N}[0, 1]$, $x_{1gh} \sim \mathcal{N}[0, 1]$, $x_{2gh} \sim \mathcal{N}[0, 1]$.

Here all four methods are asymptotically valid, since the errors are not clustered. This fact is reflected by simulations with the largest sample, the $G = H = 100$ row, presented in bold in Table 1. The rejection rates for the four methods range from 4.7% to 6.1%, with one case marginally outside the already-mentioned simulation confidence intervals.

We now consider in detail inference with smaller numbers of clusters. Then rejection rates may exceed 5%, as even with a Gaussian dgp, the Wald test statistic has a distribution with fatter tails than the standard normal, due to the need to estimate the unknown error variance (even if the standard error estimate is unbiased).

The Wald test based on assuming iid errors is exactly T distributed with $(GH - 3)$ degrees of freedom under the current dgp, so that even in the smallest design with $G = H = 10$ the theoretical rejection rate is 5.3% (since $\Pr[|t| > 1.96 | t \sim T(97)] = 0.053$), still quite close to 5%. Results in Table 1 reflect this fact, with rejection rates in the first two columns ranging from 4.1% to 6.7%.

Exact finite-sample results are not available for the other methods. For one-way clustering a common small-sample correction is to use the $T(G - 1)$ distribution, though this may still not be fat enough in the tails (see, for example, Cameron et al. (2008)). For a regressor that is cluster-invariant, Donald and Lang (2007) support the $T(G - L)$ distribution, with L the number of cluster-invariant regressors and often $L = 2$ (the cluster-invariant regressor and the intercept). Assuming a $T(G - 1)$ distribution, with

$G = 10$ the rejection rate should be 8.2% (since $\Pr[|t| > 1.96 | t \sim T(9)] = 0.082$), which can be compared with the actual one-way rejection rates that range from 7.0% to 9.7% for various rows of Table 1 with $G = 10$.

Wald tests based on standard errors computed using a two-way random effects model have rejection rates in Table 1 that are qualitatively similar to those assuming iid errors. This is expected as the random effects method has little loss of degrees of freedom as just two additional variance parameters need to be computed. A T distribution with degrees of freedom close to the number of observations, essentially a standard normal, may provide a good approximation.

The next two columns of Table 1 present Wald tests based on two-way cluster-robust standard errors. From the first two rows of the table, with a small number of clusters the test over-rejects considerably when standard normal critical values are used.

The next two columns present rejection rates when the critical value is instead that from a T distribution with $\min(G, H) - 1$ degrees of freedom. The motivation is that for one-way cluster-robust standard errors a common small-sample adjustment is to use critical values from the T distribution with $G - 1$ degrees of freedom. This leads to rejection rates of no more than 7.2% for all designs except the smallest with $G = H = 10$.

The final two columns present results when G group 1 and H group 2 fixed effect dummies are additionally included as regressors in the fitted model. Two-way cluster robust standard errors continue to show good performance.

3.1.2. Dgp with two-way clustered homoskedastic errors

Table 2 reports results for a dgp with two-way random effect errors and with clustered regressors. Specifically, $u_{gh} = \varepsilon_g + \varepsilon_h + \varepsilon_{gh}$ where the three errors are iid $\mathcal{N}[0, 1]$, the regressor x_{1gh} is the sum of an iid $\mathcal{N}[0, 1]$ draw and a g^{th} cluster-specific $\mathcal{N}[0, 1]$ draw, and similarly x_{2gh} is the sum of an iid $\mathcal{N}[0, 1]$ draw and an h^{th} cluster-specific $\mathcal{N}[0, 1]$ draw. The intraclass correlation coefficient for errors that share one but not two clusters is 0.33.

Here both the third and fourth methods are asymptotically valid. With two-way clustering the second method will generally fail, but for our particular dgp, one-way cluster-robust standard errors (with clustering on group 1) will be valid for inference on β_1 but not β_2 . Specifically, here the regressor x_{1gh} is correlated over only g (and not h), so that for inference on β_1 it is necessary to control for clustering only over g , even though the error is also correlated over h . If the regressor x_{1gh} was additionally correlated over h , even mildly so, then the one-way standard errors for $\hat{\beta}_1$ would also be incorrect.

Simulations for the largest sample, the $G = H = 100$ row presented in bold in Table 2, confirm these assertions. The rejection rates for the third and fourth methods, and the second method for β_1 , range from 3.6% to 6.4%.

For Wald tests based on the erroneous assumption of iid errors there is considerable over-rejection, and we observe the well-known result (presented after (2.5)) that the over-rejection is increasing in the number of observations within each cluster, while it is invariant in the number of clusters. For example, with 20 group 1 clusters the rejection rates for tests on β_1 are 34.0%, 50.8% and 62.9%, respectively, as the number of observations in each cluster (which equals the number of group 2 clusters in our design) increases from 20 to 50 and to 100, while the corresponding rejection rates for tests on β_2 are 32.3%, 33.1% and 33.9%.

Controlling for clustering by using standard one-way cluster-robust standard errors that cluster on group 1 leads to rejection rates for β_1 that go to 5% as the number of clusters increases, though there is a high rejection rate of 13.7% when $G = H = 10$. The high over-rejection rates for inference on β_2 even exceed those when iid errors are assumed.

The two-way random effects correction does very well. This is to be expected as this corresponds to the dgp, and because in finite samples the Wald test is close to T distributed with many degrees of freedom (roughly the number of observations).

The next two columns of Table 2 show that the two-way cluster-robust correction with standard normal critical values does fine for large number of clusters, but there is considerable over-rejection when there are few clusters.

The next two columns show considerable improvement for the two-way cluster-robust method when T critical values are used in place of standard normal critical values. The rejection rate is less than 9% for all designs except those with 10 clusters. And even with 10 clusters the rejection rate falls as the number of clusters in the other dimension rises. Thus the rejection rate for tests on β_1 is 12.6% when $G = H = 10$, 10.2% when $(G, H) = (10, 50)$ and 9.2% when $(G, H) = (10, 100)$. This fact suggests that our design with only one observation per (g, h) cluster may be especially challenging.

The final two columns of Table 2 show that the two-way cluster robust standard errors continue to perform well after adding group specific fixed effect dummies as additional regressors.

3.1.3. Dgp with two-way clustered heteroskedastic errors

Table 3 considers a dgp with heteroskedastic two-way random effect errors and clustered regressors. Specifically, $u_{gh} = \varepsilon_g + \varepsilon_h + \varepsilon_{gh}$ where ε_g and ε_h are again $\mathcal{N}[0, 1]$ but now ε_{gh} is $\mathcal{N}[0, |x_{1gh} \times x_{2gh}|]$, while the regressors are distributed as in the dgp for Table 2. This dgp induces heteroskedasticity, so the Moulton-type standard error estimator that assumes homoskedastic error components is inconsistent and will lead to Wald tests with rejection rate different from 5%. Note that compared to the Table 2 dgp, the variances of the cluster components ε_g and ε_h are unchanged while the variance of ε_{gh} has increased. This reduces the correlation of u_{gh} over g and h , so that rejection rates

of methods that do not control for clustering will not be as high as in Table 2.

Here the first three methods will in general fail. As already mentioned in section 3.1.2, however, inference on β_1 (but not β_2) is valid using the second method, due to the particular dgp used here. The fourth method remains asymptotically valid.

The simulations with the largest sample, the $G = H = 100$ row presented in bold in Table 3, confirm these expectations. The two-way cluster-robust method has rejection rates between 6% and 7% that may be high due in part to simulation variability as the rejection rates for $G = H = 90$ are between 5% and 6%. All the other methods, (except the one-way cluster-robust for β_1 with clustering on group 1), have rejection rates for one or both of β_1 and β_2 that exceed 9%.

Assuming iid errors, the first two columns of Table 3 display over-rejection rates that are lower than those in Table 2, due to lower correlation in the errors as already explained.

The next two columns are qualitatively similar to those in Table 2 – controlling for one-way clustering on group 1 improves inference on β_1 , but tests on β_2 over-reject even more than when iid errors are assumed.

The Moulton-type two-way effects method clearly fails when heteroskedasticity is present. The lowest rejection rate in Table 3 is 8.5%, and the rejection rates generally exceed those assuming iid errors.

The two-way cluster robust standard errors are clearly able to control for both two-way clustering and heteroskedasticity. When standard normal critical values are used there is some over-rejection for small numbers of clusters, as in earlier Tables, but except for $G = H = 10$ the rejection rates are lower than if the Moulton-type correction is used. Once T critical values are used, the two-way cluster-robust method’s rejection rates are always lower than using the Moulton-type standard errors, and they are always less than 10% except for the smallest design with $G = H = 10$. It is not clear whether the small-sample correction of Bell and McCaffrey (2002) for the variance of the OLS estimator with one-way clustering, used in Angrist and Lavy (2002) and Cameron et al. (2008), can be adapted to two-way clustering.

As in Tables 1 and 2, the final two columns show continued good performance when group specific dummies are additionally included as regressors.

Our results are based on the assumption that the group size N_{gh} is finite (see the Appendix). However, it does not necessarily need to be small compared to G or H . We have estimated models similar to this dgp with $G = H = 30$, where we have varied the cell sizes (observations per $g \times h$ cell) from 1, as in Tables 1-3, to 1000. In these simulations we have also added separate iid $N(0,1)$ errors to each of x_{1igh} , x_{2igh} and u_{igh} . Results (not reported) indicate that the two-way robust estimator continues to perform well across the various cell sizes.

3.2. Monte Carlo Based on Errors Correlated over Time and States

We now consider an example applicable to panel and repeated cross-section data, with errors that are correlated over both states and time. Correlation over states at a given point in time may occur, for example, if there are common shocks, while correlation over time for a given state typically reduces with lag length. An example of this sort of situation is found in Foote (2007). This is unlike the preceding section random effects model that assumes constant autocorrelation.

One possibility is to adapt the random effects model to allow dampening serial correlation in the error, similar to the *dgp* used by Kezdi (2004) and Hansen (2007) in studying one-way clustering, with addition of a common shock.

Instead we follow Bertrand et al. (2004) in using actual data, augmented by a variation of their randomly-generated “placebo law” policy that produces a regressor correlated over both states and time.

The original data are for 1,358,623 employed women from the 1979-1999 Current Population Surveys, with log earnings as the outcome of interest. For each simulation, we randomly draw 50 U.S. states from the original data (and re-label the states from 1 to 50). The model estimated is

$$y_{ist} = \alpha d_{st} + \mathbf{x}'_{ist}\boldsymbol{\beta} + \delta_s + \gamma_t + u_{ist}, \quad (3.2)$$

where y_{ist} is individual log-earnings, the grouping is by state and time (with indices s and t corresponding to g and h in Section 2), d_{st} is a state-year-specific regressor, and \mathbf{x}_{ist} are individual characteristics. Here $G = 50$ and $H = 21$ and, unlike in Section 3.1, there are many (on average 1294) observations per (g, h) cell. For some estimations we include state-specific fixed effects δ_s and time-specific fixed effects γ_t (70 dummies), as our *d.g.p.* enables these fixed effects to be identified. In most of their simulations Bertrand et al. (2004) run regressions on data aggregated into state-year cells, to reduce computation time for their many simulations. Here we work with the individual-level data in part to demonstrate the feasibility of our methods for large data sets (over one million observations).

Interest lies in inference on α , the coefficient of a randomly-assigned “placebo policy” variable. Bertrand et al. (2004) consider one-way clustering, with d_{st} generated to be correlated within state (i.e., over time for a given state). Here we extend their approach to induce two-way clustering, with within-time clustering as well as within-state clustering. The placebo law for a state-year cell is generated by $d_{st} = \varepsilon_{st}^s + 2\varepsilon_{st}^t$. The variable ε_{st}^s is a within-state AR(1) variable, $\varepsilon_{st}^s = 0.6\varepsilon_{st-1}^s + u_{st}^s$, with $u_{st}^s \text{ iid } \mathcal{N}[0, 1]$, and is generated independently from all other variables. ε_{st}^s is independent across states. Similarly, the variable ε_{st}^t is a within-year AR(1) variable, $\varepsilon_{st}^t = 0.6\varepsilon_{s-1,t}^t + u_{st}^t$, correlated over states, with $u_{st}^t \text{ iid } \mathcal{N}[0, 1]$, and also independent from other variables. Here the index s ranges from 1-50 based on the order that the states were drawn from the original

data. This law is the same for all individuals within a state-year cell. This dgp ensures that d_{st} and $d_{s't'}$ are dependent if and only if at least one of $s = s'$ or $t = t'$ holds. Because we draw the full time-series for each state, the outcome variables (and hence the errors) are autocorrelated over time within a state. We also add in a wage shock $y_{ist}^{new} = y_{ist}^{original} + 0.01 \cdot w_{st}^t$, with w_{st}^t generated similarly to (but independent of) ε_{st}^t , that is correlated over states. In each of 2,000 simulations we draw the 50 states' worth of individual data, wages are adjusted with w_{st}^t , the variable d_{st} is randomly generated, model (3.2) is estimated, and the null hypothesis that $\alpha = 0$ is rejected at significance level 0.05 if $|\hat{\alpha}|/\text{se}[\hat{\alpha}] > 1.96$. Given the design used here, $\hat{\alpha}$ is consistent, and the correct asymptotic rejection rates for the simulation results in Table 4 will be 5%, provided that a consistent estimate of the standard error is used.

The first column of Table 4 considers regression on d_{st} and individual controls (a quartic in age and four education dummies, without the fixed effects δ_s and γ_t). Since log earnings y_{ist} are correlated over both time and state and d_{st} is a generated regressor uncorrelated with y_{ist} , the error u_{ist} is correlated over both time and state. Using heteroskedastic-robust standard errors leads to a very large rejection rate (92%) due to failure to control for clustering. The standard one-way cluster-robust cluster methods partly mitigate this, though the rejection rates still exceed 19%. Note that, as argued by Bertrand et al. (2004), clustering on the 50 states does better than clustering on the 1,050 state-year cells. In this example, clustering on year also shows improvements over clustering on state-year cells. We present results from the two-way cluster-robust method in the last row. As before, we use standard Stata degrees-of-freedom corrections for each component of the variance estimator. The two-way variance estimator does best, with rejection rate of 7.2%. This rate is still higher than 5%, in part due to use of critical values from asymptotic theory. Assuming a $T(H - 1)$ distribution, with $H = 21$ the rejection rate should be 6.4% (since $\Pr[|t| > 1.96 | t \sim T(20)] = 0.064$), and with 1,000 simulations a 95% confidence interval is (4.9%, 7.9%). The dgp studied here is thus might be well approximated by a $T(H - 1)$ distribution.

For the second column of Table 4, we add state fixed effects. The inclusion of state fixed effects does not improve rejection rates for heteroskedasticity robust, clustering on state-year cells, or clustering on state. Clustering on year does somewhat better. As in the first column, two-way robust clustering does best, with rejection rates of 6.9%.

For the third column of Table 4, we add year (but not state) fixed effects. In this setting the results for clustering on state-by-year and for clustering on state improve markedly. However, when clustering on state we still reject 12% of the time, which is not close to the two-way cluster robust rejection rate of 7.6%.

In column four we include both year and state dummies as regressors. For the models using heteroskedastic-robust standard errors the rejection rate is 79%. Clustering on just state-year cells results in rejection rates of 13.9%, which is similar to those from

clustering on state (15%). As before, two-way clustering does best, with rejection rates of 7.1%. In this example the two-way cluster-robust method works well regardless of whether or not state and year fixed effects are included as regressors, and gives the best results of the methods considered.

4. Empirical examples

In this section we contrast results obtained using conventional one-way cluster-robust standard errors to those using our method that controls for two-way (or multi-way) clustering. The first and third examples consider two-way clustering in a cross-section setting. The second considers a rotating panel, and considers probit estimation in addition to OLS.

We compare computed standard errors and p-values across various methods. In contrast to the section 3 simulations, there is no benchmark for the rejection rates.

4.1. Hersch - Cross-Section with Two-way Clustering

We consider a cross-section study of wages with clustering at both the industry and occupation level. Ideally one would obtain cluster-robust standard errors that control for both sources of clustering, but previous researchers have been restricted to the choice of one or the other. In this example there are 5,960 individuals in 211 industries and 387 occupations.

We base our application on Hersch’s (1998) study of compensating wage differentials. Using industry and occupation injury rates merged into CPS data, Hersch examines the relationship between injury risk and wages for men and women. The model is

$$y_{igh} = \alpha + \mathbf{x}'_{igh}\beta + \gamma \times rind_{ig} + \delta \times rocc_{ih} + u_{igh}, \quad (4.1)$$

where y_{igh} is individual log-wage rate, \mathbf{x}_{igh} includes individual characteristics such as education, race, and union status, $rind_{ig}$ is the injury rate for individual i ’s industry and $rocc_{ih}$ is the injury rate for occupation. In this application, as in many similar applications, it is not possible to include industry and occupation fixed effects, because then the coefficients of the key regressors $rind$ and $rocc$ cannot be identified. Hersch emphasizes the importance of using cluster-robust standard errors, noting that they are considerably larger than heteroskedastic-robust standard errors. But she is able to control only for one source of clustering - industry or occupation - and not both simultaneously. Instead she separately reports regressions with just $rind$ as a regressor with clustering on industry, with just $rocc$ as a regressor with clustering on occupation, and with both $rind$ and $rocc$ as regressors with clustering on just industry.

We replicate results for column 4 of Panel B of Table 3 of Hersch (1998), with both $rind$ and $rocc$ included as regressors, using data on 5,960 male workers. We

report a wider array of estimated standard errors: default standard errors assuming iid errors, White heteroskedastic-robust, one-way cluster-robust by industry, one-way cluster-robust by occupation, and our preferred two-way cluster-robust with clustering on both industry and occupation. We also present (in brackets) p-values from a test of each coefficient being equal to zero.

The first results given in our Table 5 show that heteroskedastic-robust standard errors differ little from standard errors based on the assumption of iid errors. The big change arises when clustering is appropriately accounted for. One-way cluster-robust standard errors with clustering on industry lead to substantially larger standard errors for *rind* (0.643 compared to 0.397 for heteroskedastic-robust), though clustering on industry has little effect on those for *rocc*. One-way cluster-robust standard errors with clustering on occupation yield substantially larger standard errors for *rocc* (0.363 compared to 0.260 for heteroskedastic-robust), with a lesser effect for those for *rind*. Clearly for *rind* it is best to cluster on industry, and for *rocc* it is best to cluster on occupation.

Our two-way cluster-robust method permits clustering on both industry and occupation. It is to be expected that the increase in the standard error for *rind* will be greatest when compared to one-way clustering on occupation (rather than industry), and for *rocc* the increase will be largest when compared to one-way clustering on industry (rather than occupation). This is indeed the case. For *rind*, the two-way cluster-robust standard error is ten percent larger than that based on one-way clustering at the industry level, and is forty-five percent larger than that based on one-way clustering on occupation. The p-value for a test of zero on the coefficient on *rind* goes from 0.0001 (when clustering on Occupation) to 0.0070. For *rocc*, the two-way standard error is little different from that based on clustering on occupation, but it is forty percent larger than that based on clustering on industry. The p-value on a similar test for *rocc* goes from 0.0639 (when clustering on Industry) to 0.1927.

In this application it is obvious that for *rind* it is most important to cluster on industry, while for *rocc* it is most important to cluster on occupation. Our method provides a way to simultaneously do both. For the industry injury rate this makes a substantial difference. The standard error of *rind* increases from 0.40 without control for clustering to 0.64 with one-way clustering on industry, and then increases further to 0.70 with two-way clustering on both industry and occupation. This application nicely illustrates the importance of using our procedure when we are interested in estimating coefficients for multiple variables having different intraclass correlation coefficients in different clustering dimensions.

4.2. Gruber and Madrian - Rotating Panel

In this example we use data on 39,063 men and cluster by state-year cell (359 clusters) and by household (26,383 clusters). The latter clustering is unconventional and accommodates the rotating panel of the CPS.

Specifically, many surveys taken on a regular basis involve a panel-type structure for households, which are resurveyed for several months. The U.S. Current Population Survey (CPS) uses a specific rotation scheme to survey households: a household is surveyed for four consecutive months, then not surveyed for the next eight months, and then surveyed again for four more months. Then any study that uses the CPS data for more than one time period will have households appearing more than once in the data set (unless the time periods are more than 15 months apart).

Household errors can be expected to be correlated from one period to the next. This correlation is typically ignored, due to a perceived need to control first for other sources of error correlation (note that any control for clustering on region, such as on state, will subsume household error correlation).

In this example we use similar data to that in Gruber and Madrian's (1995) study of health insurance availability and retirement. The probit model estimated is

$$\Pr[y_{ist} = 1] = \Phi(\alpha d_{st} + \mathbf{x}'_{ist}\boldsymbol{\beta} + \delta_s + \gamma_t), \quad (4.2)$$

where y_{ist} is a binary variable for whether or not retired in the past year, the key regressor d_{st} is a state-year policy variable that equals the number of months in a state-year of mandated continuation of health insurance coverage after job separation, and \mathbf{x}_{ist} denotes individual-level controls. State fixed effects and year fixed effects are also included. Given the large number of observations available to estimate each fixed effect, the well-known incidental parameters problem for probit models is unlikely to be important. In addition to estimating probit models, we also estimate linear probability (OLS) models. For comparability, we present for the probit model the average marginal effect, and its estimated standard error.

One natural dimension for clustering is the state-year group (539 clusters) since this reflects the variation in d_{st} . Given the rotating design, if a household is in a given year's March CPS, it is likely to also appear in the data set in the previous year or in the subsequent year. If household outcomes are correlated from one year to the next, then the household identifier serves as a natural second dimension for clustering (26,383 clusters). The maximum possible increase in standard errors due to error correlation at the household level is about forty percent (corresponding to a doubling of the variance estimate: $\sqrt{2} = 1.41$). This would occur under the strong assumptions that all households appear in two consecutive years, that the errors for the same household are perfectly correlated across the two years, that d_{st} for the same household is perfectly correlated across the two years (i.e., d_{st} is time invariant), and

that already accounted for state-year correlation is negligible. The difference turns out to be considerably less than that here.

Our results are given in Table 6. We use White heteroskedastic standard errors, which differ little from those assuming iid errors, as the benchmark. We have come close to replicating Gruber and Madrian’s data, but we have not done not so exactly. The means of key variables in our data set are close to those in their 1993 and 1995 papers, with small exceptions. The basic probit estimates provide point estimates and (nonclustered) standard errors that are broadly similar to those reported in their paper.

For the probit estimator, the standard error increases by 9.2% when we control for one-way clustering at the state-year level ($6.265/5.732 = 1.092$) and by 2.3% when we control for one-way clustering at the household level. When we allow for two-way clustering (with state-year as one dimension and household as the other dimension), the standard error increases by 11.5% which in this example coincides with the sum of the two-separate one-way clustering corrections. A more common correction for these data would be one-way clustering on state, which leads to a smaller 5.2% increase in the standard error.

The results for OLS estimation of this model are qualitatively similar. The standard errors increase by 11.1% using one-way clustering on state-year, by 2.6% using one-way clustering on household, and by 13.3% using two-way clustering on state-year and household.

4.3. Rose and Engel - bilateral trade model

A common setting for two-way clustering arises is paired or dyadic data, such as that on trade flows between pairs of countries. Cameron and Golotvina (2005) show the importance of controlling for two-way clustering, and propose FGLS estimation based on the assumption of iid country random effects. Here we instead apply our more robust method to an example in their paper, which replicates the fitted model given in the first column of Table 3 of Rose and Engel (2002).

The data are a single cross-section on trade flows between 98 countries with 3262 unique country pairs. A gravity model is fitted for the natural logarithm of bilateral trade. The coefficient of the log product of real GDP (estimated slope = 0.867) has heteroskedastic-robust standard error of 0.013, reported by Rose and Engel (2002), and average one-way clustered standard error of 0.031, where we average the one-way standard error with clustering on the first country in the country pair and the one-way standard error clustering on the second country in the country pair. Using the methods proposed in this paper, the two-way robust standard error is 0.043. This is 36% larger than the average one-way cluster robust standard error, and 230% larger than the White robust standard error. Note that if country specific effects are included (for each of the two countries in the country pair) as a possible way to control for the clustering, then

the coefficient of the log product of real GDP is no longer identified.

For the coefficient on log distance (estimated slope = -1.367), we obtain standard errors of 0.035 (heteroskedastic robust), 0.078 (average of one-way clustered standard errors), and 0.106 (two-way robust). Roughly similar proportionate increases in the standard errors are obtained for the coefficients of the other regressors in the model.

Allowing for two-way robust clustering impacts the estimated standard errors by a considerable magnitude.

4.4. Other examples

In this section we discuss details of two studies that implement the two-way clustering method proposed in this paper.

Foote (2007) re-investigated Shimer’s (2001) influential finding of a (surprising) negative correlation between a U.S. state’s annual unemployment rate (dependent variable) and the share of the state’s labor force that is young. Even with relatively high migration by the young, a state’s youth share is highly autocorrelated over time; correlation in regional socioeconomic conditions also imply that youth shares will be correlated across states within year. Similar two-way correlation is expected for residual state-level unemployment rates.

In the subset of his results that exactly replicates Shimer’s OLS specification (Panel A, column (1) of his Table I), Foote finds that clustering at the state level, which most researchers likely would do in the wake of Bertrand et al. (2004), raises the estimated standard error from 0.18 to 0.39. Using our method to cluster at both the state and year levels yields as dramatic an increase in the estimated standard error from 0.39 to 0.61, even with state and year fixed effects included as regressors. Clustering on year alone, which would be an uncommon approach, yields a 0.50 estimate. A qualitatively similar pattern of changes in estimated standard errors is obtained for a specification that instruments the state’s youth share (Foote’s Panel B).

Cascio and Schanzenbach (2007) use data from the state of Tennessee’s Project Star random-assignment experiment to study the relationship between educational outcomes and a child’s age relative to the rest of her class, holding constant the child’s own age. In their typical regression model, an educational outcome is measured at the student-year-classroom level. Each classroom contains multiple students, so there is the potential for classroom-level clustering. At the same time, students may be observed in multiple years, each time in a different classroom, so there is the potential for student-level clustering. Since any given student’s relative age depends deterministically on the ages of other students in the class, there will be (negative) autocorrelation of relative-age variables within class; moreover, students who are young for their class in one year will also tend to be young in the subsequent year’s class, so there will be (positive) autocorrelation of relative-age variables within the student dimension. Again, there is

no way to construct a variable that subsumes all potential multi-way clustering.

According to the authors, (unreported) standard error estimates using our estimator were “slightly more precise” (p. 26) than estimates that cluster only at the classroom level. Such an increase in precision is expected if there is clustering at the classroom level, given the negative dependence of students’ relative-age measures within classroom.

5. Conclusion

There are many empirical applications where a researcher needs to make statistical inference controlling for clustering in errors in multiple non-nested dimensions, under less restrictive assumptions than those of a multi-way random effects model. In this paper we offer a simple procedure that allows researchers to do this.

Our two-way or multi-way cluster-robust procedure is straightforward to implement. As a small-sample correction we propose adjustments to both standard errors and Wald test critical values that are analogous to those often used in the case of one-way cluster-robust inference. Then inference appears to be reasonably accurate except in the smallest design with ten clusters in each dimension.

In a variety of Monte Carlo experiments and replications, we find that accounting for multi-way clustering can have important quantitative impacts on the estimated standard errors and associated p-values. For perspective we note that if our method leads to an increase of 20% in the reported standard errors, then a t-statistic of 1.96 with a p-value of 0.050 becomes a t-statistic of 1.63 with a p-value of 0.103. Even modest changes in standard errors can have large effects on statistical inference.

The impact of controlling for multi-way clustering is likely to be greatest when the errors are correlated over two or more dimensions. When this is the case, then the impact of the errors’ correlation may be magnified if in addition the regressors of interest are also correlated over the same dimensions. This is especially likely to be the case when the research design precludes fixed effects along each of the dimensions, as in the Hersch (1995) example. This example also illustrates that even if the regressor is most clearly correlated over only one dimension, controlling for error correlation in the second dimension can also make a difference. However, we also note that in some settings, such as the Gruber-Madrian replication, the impact of the method is modest.

In general a researcher will not know *ex ante* how important it is to allow for multi-way clustering, just as in the one-way case. Our method provides a way to control for multi-way clustering that is a simple extension of established methods for one-way clustering, and it should be of considerable use to applied researchers.

6. Acknowledgements

This paper has benefitted from comments from the Editor, an Associate Editor, and two referees, and from presentations at The Australian National University, U.C. - Berkeley, U.C. - Riverside, Dartmouth College, MIT, PPIC, and Stanford University. Miller gratefully acknowledges funding from the National Institute on Aging, through Grant Number T32-AG00186 to the NBER. We thank Marianne Bertrand, Esther Duflo, Sendhil Mullainathan, and Joni Hersch for assisting us in replicating their data sets. We thank Peter Hansen, David Neumark, and Mitchell Petersen for helpful comments, particularly for referring us to relevant literature.

A. Appendix

We present results for the general case of GMM estimation. Estimation is based on the moment condition

$$\mathbb{E}[\mathbf{z}_i(\boldsymbol{\theta}_0)] = \mathbf{0},$$

for observation i , where $\boldsymbol{\theta}$ is a $q \times 1$ parameter vector $\boldsymbol{\theta}$ and \mathbf{z} is an $m \times 1$ vector with $m \geq q$. Examples include OLS with $\mathbf{z}_i = (y_i - \mathbf{x}_i' \boldsymbol{\beta}) \mathbf{x}_i$, linear IV with $\mathbf{z}_i = (y_i - \mathbf{x}_i' \boldsymbol{\beta}) \mathbf{w}_i$ where \mathbf{w}_i are instruments for \mathbf{x}_i , and the logit MLE with $\mathbf{z}_i = (y_i - \Lambda(\mathbf{x}_i' \boldsymbol{\beta})) \mathbf{x}_i \boldsymbol{\beta}$.

For models with $m = q$, such as OLS, logit, and just-identified IV we need only use the m-estimator $\tilde{\boldsymbol{\theta}}$ that solves

$$\sum_{i=1}^N \mathbf{z}_i(\tilde{\boldsymbol{\theta}}) = \mathbf{0}. \quad (\text{A.1})$$

Given two-way clustering with typical cluster (g, h) , $\mathbf{z}_i(\boldsymbol{\theta}) = \mathbf{z}_{igh}(\boldsymbol{\theta})$ and

$$\begin{aligned} \sum_{i=1}^N \mathbf{z}_i(\boldsymbol{\theta}) &= \sum_{g=1}^G \sum_{h=1}^H \sum_{i \in C_{gh}} \mathbf{z}_{igh}(\boldsymbol{\theta}) \\ &= \sum_{g=1}^G \sum_{h=1}^H \mathbf{z}_{gh}(\boldsymbol{\theta}), \end{aligned} \quad (\text{A.2})$$

where C_{gh} denotes the observations in cluster (g, h) , and

$$\mathbf{z}_{gh}(\boldsymbol{\theta}) = \sum_{i \in C_{gh}} \mathbf{z}_{igh}(\boldsymbol{\theta}) \quad (\text{A.3})$$

combines observations in cluster (g, h) .

For models with $m > q$, the more general GMM estimator $\hat{\boldsymbol{\theta}}$ maximizes

$$Q(\boldsymbol{\theta}) = \left(N^{-1} \sum_{i=1}^N \mathbf{z}_i(\boldsymbol{\theta}) \right)' \mathbf{W} \left(N^{-1} \sum_{i=1}^N \mathbf{z}_i(\boldsymbol{\theta}) \right), \quad (\text{A.4})$$

where \mathbf{W} is an $m \times m$ full rank symmetric weighting matrix with $\mathbf{W} \xrightarrow{p} \mathbf{W}_0$. The GMM estimator reduces to the m-estimator when $m = q$, for any choice of \mathbf{W} .

We assume that $\hat{\boldsymbol{\theta}}$ is consistent for $\boldsymbol{\theta}_0$, that $G \rightarrow \infty$ and $H \rightarrow \infty$ at the same rate, so that $G/H \rightarrow \text{constant}$, and that the number N_{gh} of observations in cluster (g, h) is not growing with G or H . Note that $N_{gh} = 1$ is possible. As discussed below, we consider a rate of convergence \sqrt{G} , so that

$$\sqrt{G}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}[\mathbf{0}, (\mathbf{A}_0' \mathbf{W}_0 \mathbf{A}_0)^{-1} \mathbf{A}_0' \mathbf{W}_0 \mathbf{B}_0 \mathbf{W}_0 \mathbf{A}_0 (\mathbf{A}_0' \mathbf{W}_0 \mathbf{A}_0)^{-1}], \quad (\text{A.5})$$

where

$$\mathbf{A}_0 = \lim \mathbb{E} \left[(GH)^{-1} \sum_{i=1}^N \partial \mathbf{z}_{i0}(\boldsymbol{\theta}) / \partial \boldsymbol{\theta}' \right] \quad (\text{A.6})$$

and

$$\mathbf{B}_0 = \lim E \left[G^{-1} H^{-2} \sum_{i=1}^N \sum_{j=1}^N \mathbf{z}_i(\boldsymbol{\theta}) \mathbf{z}_j(\boldsymbol{\theta})' \right]. \quad (\text{A.7})$$

For $m = q$ the result simplifies to $\sqrt{G}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}[\mathbf{0}, (\mathbf{A}_0' \mathbf{B}_0 \mathbf{A}_0)^{-1}]$.

We now simplify \mathbf{B}_0 under the assumption of two-way clustering. Since $\sum_i \mathbf{z}_i(\boldsymbol{\theta}) = \sum_g \sum_h \mathbf{z}_{gh}(\boldsymbol{\theta})$ we have

$$\begin{aligned} & E \left[G^{-1} H^{-2} \sum_{i=1}^N \sum_{j=1}^N \mathbf{z}_i(\boldsymbol{\theta}) \mathbf{z}_j(\boldsymbol{\theta})' \right] \\ &= E \left[G^{-1} H^{-2} \sum_{g=1}^G \sum_{h=1}^H \sum_{g'=1}^G \sum_{h'=1}^H \mathbf{z}_{gh}(\boldsymbol{\theta}) \mathbf{z}_{g'h'}(\boldsymbol{\theta})' \right] \\ &= G^{-1} H^{-2} \sum_g \sum_h \sum_{h'} E[\mathbf{z}_{gh} \mathbf{z}_{gh'}'] \\ &\quad + G^{-1} H^{-2} \sum_h \sum_g \sum_{g'} E[\mathbf{z}_{gh} \mathbf{z}_{g'h}'] \\ &\quad - G^{-1} H^{-2} \sum_g \sum_h E[\mathbf{z}_{gh} \mathbf{z}_{gh}'], \end{aligned} \quad (\text{A.8})$$

where the first triple sum uses dependence if $g = g'$, the second triple sum uses dependence if $h = h'$, and the third double sum subtracts terms when $g = g'$ and $h = h'$ which are double counted as they appear in both of the first two sums.

Consider the first triple sum which has GH^2 terms. The cross-product term $\mathbf{z}_{gh} \mathbf{z}_{gh'}' = \sum_{i \in C_{gh}} \sum_{j \in C_{gh'}} \mathbf{z}_{ghi}(\boldsymbol{\theta}) \mathbf{z}_{gh'j}(\boldsymbol{\theta})$ is an $N_{gh} \times N_{gh}$ matrix. We assume that $E[\mathbf{z}_{igh}(\boldsymbol{\theta}) \mathbf{z}_{jgh'}(\boldsymbol{\theta})]$ is bounded away from zero and bounded from above. Then $E[\mathbf{z}_{gh} \mathbf{z}_{gh'}']$ is bounded, given N_{gh} fixed, and $G^{-1} H^{-2} \sum_g \sum_h \sum_{h'} E[\mathbf{z}_{gh} \mathbf{z}_{gh'}']$ is bounded. Similarly for the second term. The third term has only GH terms so this third term goes to zero.

The above analysis assumes that $E[\mathbf{z}_{igh}(\boldsymbol{\theta}) \mathbf{z}_{jgh'}(\boldsymbol{\theta})]$ is bounded away from zero. This will be the case for common shocks such as the standard two-way random effects model. But it need not always be the case. As an extreme example, suppose $N_{gh} = 1$ and that there is no clustering; i.e., each observation is independent. Then $E[\mathbf{z}_{gh} \mathbf{z}_{gh'}'] = 0$ unless $h = h'$ and so the first sum has only GH nonzero terms, and similarly for the other two terms. The triple sum is of order GH , rather than GH^2 , and the rate of convergence of the estimator becomes a faster \sqrt{GH} rather than \sqrt{G} . This is the rate expected for estimation based on GH independent observations.

More generally the triple sum is of order GH , rather than GH^2 , if the dependence of observations in common cluster g goes to zero as clusters h and h' become further apart, as is the case with declining time series dependence or spatial dependence. Then in \mathbf{B}_0 we normalize by (GH) and the rate of convergence of the estimator becomes a faster \sqrt{GH} rather than \sqrt{G} . Regardless of the rate of convergence we obtain the same asymptotic variance matrix for $\hat{\boldsymbol{\beta}}$.

Qualitatively similar differences in rates of convergence are obtained by Hansen (2007) for the standard one-way cluster-robust variance matrix estimator for panel data. When $N \rightarrow \infty$ with T fixed (a short panel), the rate of convergence is \sqrt{N} . When both $N \rightarrow \infty$ and $T \rightarrow \infty$ (a long panel), the rate of convergence is \sqrt{N} if there is no mixing (his Theorem 2) and \sqrt{NT} if there is mixing (his Theorem 3). While the rates of convergence differ in the two cases, he obtains the same asymptotic variance for the OLS estimator.

References

- Acemoglu, D., and J.-S. Pischke (2003), “Minimum Wages and On-the-job Training,” *Research in Labor Economics*, 22, 159-202.
- Angrist, J.D., and V. Lavy (2002), “The Effect of High School Matriculation Awards: Evidence from Randomized Trials,” NBER Working Paper No. 9389.
- Arellano, M. (1987), “Computing Robust Standard Errors for Within-Group Estimators,” *Oxford Bulletin of Economics and Statistics*, 49, 431-434.
- Baughman R., and K. Smith (2007), “The Labor Market for Direct Care Workers,” New England Public Policy Center Working Paper No. 07-4, Federal Reserve Bank of Boston.
- Beck, T., A. Demircuc-Kunt, L. Laeven, and R. Levine (2008), “Finance, Firm Size, and Growth,” *Journal of Money, Credit, and Banking*, 40, 1379-1405.
- Bell, R.M., and D.F. McCaffrey (2002), “Bias Reduction in Standard Errors for Linear Regression with Multi-Stage Samples,” *Survey Methodology*, 169-179.
- Bertrand, M., E. Duflo, and S. Mullainathan (2004), “How Much Should We Trust Differences-in-Differences Estimates?,” *Quarterly Journal of Economics*, 119, 249-275.
- Bester, C. A., Conley, T.G., and C.B. Hansen (2009), “Inference with Dependent Data Using Cluster Covariance Estimators,” mimeo, University of Chicago Graduate School of Business, January 2009.
- Bhattacharya, D. (2005), “Asymptotic Inference from Multi-stage Samples,” *Journal of Econometrics*, 126, 145-171.
- Cameron, A.C., Gelbach, J.G., and D.L. Miller (2008), “Bootstrap-Based Improvements for Inference with Clustered Errors,” *Review of Economics and Statistics* 90, 414-427.
- Cameron, A.C., and N. Golotvina (2005), “Estimation of Country-Pair Data Models Controlling for Clustered Errors: with International Trade Applications,” U.C.-Davis Economics Department Working Paper No. 06-13.
- Cameron, A.C., and P.K. Trivedi (2005), *Microeconometrics: Methods and Applications*, Cambridge, Cambridge University Press.
- Card, D., and D.S. Lee (2004), “Regression Discontinuity Inference with Specification Error,” Center for Labor Economics Working Paper No. 74, U.C.-Berkeley.
- Cascio, E., and D.W. Schanzenbach (2007), “First in the Class? Age and the Education Production Function,” NBER Working Paper No. 13663.

- Conley, T.G. (1999), "GMM Estimation with Cross Sectional Dependence," *Journal of Econometrics*, 92, 1-45.
- Cuijpers, R., and E. Peek (2008), "The Economic Consequences of the Choice between Quarterly and Semiannual Reporting", available at SSRN: <http://ssrn.com/abstract=1125321>.
- Davis, P. (2002), "Estimating Multi-Way Error Components Models with Unbalanced Data Structures," *Journal of Econometrics*, 106, 67-95.
- Donald, S.G. and K. Lang. (2007), "Inference with Difference-in-Differences and Other Panel Data," *The Review of Economics and Statistics*, 89(2), 221-233.
- Engelhardt, G. V., and A. Kumar (2007), "The Repeal of the Retirement Earnings Test and the Labor Supply of Older Men," Boston College Center for Retirement Research wp2007-01,
- Fafchamps, M., and F. Gubert (2006), "The Formation of Risk Sharing Networks," mimeo, April 2006.
- Foote, C.L. (2007), "Space and Time in Macroeconomic Panel Data: Young Workers and State-Level Unemployment Revisited," Working Paper No. 07-10, Federal Reserve Bank of Boston..
- Gow, I.D., G Ormazabal, and D.J. Taylor (2008), "Correcting for Cross-Sectional and Time-Series Dependence in Accounting Research," mimeo, Stanford Graduate School of Business.
- Greenwald, B.C. (1983), "A General Analysis of Bias in the Estimated Standard Errors of Least Squares Coefficients," *Journal of Econometrics*, 22, 323-338.
- Gruber, J., and B. C. Madrian (1993), "Health-Insurance Availability and Early Retirement: Evidence from the Availability of Continuation Coverage," NBER Working Paper No. 4594.
- Gruber, J., and B. C. Madrian (1995), "Health-Insurance Availability and the Retirement Decision," *American Economic Review*, 85, 938-948.
- Gurun, U.G., G.G. Booth, and H.H. Zhang (2008), "Global Financial Networks and Trading in Emerging Bond Markets," available at SSRN: <http://ssrn.com/abstract=1105962>.
- Hansen, C. (2007), "Asymptotic Properties of a Robust Variance Matrix Estimator for Panel Data when T is Large," *Journal of Econometrics*, 141, 597-620.
- Hersch, J. (1998), "Compensating Wage Differentials for Gender-Specific Job Injury Rates," *American Economic Review*, 88, 598-607.

- Kézdi, G. (2004), “Robust Standard Error Estimation in Fixed-Effects Models,” Robust Standard Error Estimation in Fixed-Effects Panel Models,” *Hungarian Statistical Review*, Special Number 9, 95-116.
- Kloek, T. (1981), “OLS Estimation in a Model where a Microvariable is Explained by Aggregates and Contemporaneous Disturbances are Equicorrelated,” *Econometrica*, 49, 205-07.
- Liang, K.-Y., and S.L. Zeger (1986), “Longitudinal Data Analysis Using Generalized Linear Models,” *Biometrika*, 73, 13-22.
- Loughran, T., and S. Shive (2007), “The Impact of Venture Capital Investments On Industry Performance,” mimeo, University of Notre Dame, <http://www.dsh.fsu.edu/fin/shivepaper.pdf>
- Miglioretti, D.L., and P.J. Heagerty (2006), “Marginal Modeling of Nonnested Multilevel Data using Standard Software,” *American Journal of Epidemiology*, 165(4), 453-463.
- Martin, P., T. Mayer, and M. Thoenig (2008), “Make Trade Not War?,” *Review of Economic Studies*, 75, 865–900.
- Mitchener, K.J., and M.D. Weidenmier (2007), “The Baring Crisis and the Great Latin American Meltdown of the 1890s,” NBER Working Paper No. 13403.
- Moulton, B.R. (1986), “Random Group Effects and the Precision of Regression Estimates,” *Journal of Econometrics*, 32, 385-397.
- Moulton, B.R. (1990), “An Illustration of a Pitfall in Estimating the Effects of Aggregate Variables on Micro Units,” *Review of Economics and Statistics*, 72, 334-38.
- Olken, B.A., and P. Barron (2007), “The Simple Economics of Extortion: Evidence from Trucking in Aceh,” BREAD Working Paper No. 151.
- Petersen, M. (2007), “Estimating Standard Errors in Finance Panel Data Sets: Comparing Approaches,” unpublished manuscript, Northwestern University.
- Pepper, J.V. (2002), “Robust Inferences from Random Clustered Samples: An Application using Data from the Panel Study of Income Dynamics,” *Economics Letters*, 75, 341-5.
- Peress, J. (2007), “Product Market Competition, Insider Trading and Stock Market Efficiency”, INSEAD, working paper, <http://faculty.insead.edu/peress/personal/competition.pdf>
- Pfeffermann, D., and G. Nathan (1981), “Regression analysis of data from a cluster sample,” *Journal of the American Statistical Association*, 76, 681-689.
- Pierce, L., and J. Snyder (2008), “Ethical Spillovers in Firms: Evidence from Vehicle

- Emissions Testing,” available at SSRN: <http://ssrn.com/abstract=1157996>.
- Politis, D.N. (2007), “Higher-order accurate, positive semi-definite estimation of large-sample covariance and spectral density matrices,” mimeo, U.C.-San Diego.
- Rogers, W.H. (1993), “Regression Standard Errors in Clustered Samples,” *Stata Technical Bulletin*, 13, 19-23.
- Rose, A. C. and C. Engel (2002), “Currency Unions and International Integration,” *Journal of Money, Credit and Banking*, 34(4), 1067-1089.
- Rountree, B.R., J.P. Weston, and G.S. Allayannis (2008), “Do Investors Value Smooth Performance?,” *Journal of Financial Economics*, forthcoming, available at SSRN: <http://ssrn.com/abstract=1105077>.
- Scott, A.J., and D. Holt (1982), “The Effect of Two-Stage Sampling on Ordinary Least Squares Methods,” *Journal of the American Statistical Association*, 77, 848-854.
- Thompson, S. (2005), “A Simple Formula for Standard Errors that Cluster by Both Firm and Time,” unpublished manuscript.
- Thompson, S. (2006), “Simple Formulas for Standard Errors that Cluster by Both Firm and Time,” available at SSRN: <http://ssrn.com/abstract=914002>.
- White, H. (1980), “A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity,” *Econometrica*, 48, 817-838.
- White, H. (1982), “Maximum Likelihood Estimation of Misspecified Models,” *Econometrica*, 50, 1-25.
- White, H. (1984), *Asymptotic Theory for Econometricians*, San Diego, Academic Press.
- White, H, and I. Domowitz (1984), “Nonlinear Regression with Dependent Observations,” *Econometrica*, 52, 143-162.
- Wooldridge, J.M. (2002), *Econometric Analysis of Cross Section and Panel Data*, Cambridge, MA, MIT Press.
- Wooldridge, J.M. (2003), “Cluster-Sample Methods in Applied Econometrics,” *American Economic Review*, 93, 133-138.

Table 1
Rejection probabilities for a true null hypothesis

True model: iid errors

Number of Group 1 Clusters	Number of Group 2 Clusters	Assumption about errors in construction of Variance											
		Assume independent errors		One-way cluster robust (cluster on group1)		Two-way random effects		Two-way cluster- robust		Two-way cluster- robust, T critical values		Group fixed effects, Two-way cluster- robust	
10	10	5.6%	6.7%	9.0%	9.7%	4.4%	5.8%	13.3%	14.9%	9.9%	11.5%	10.1%	9.9%
20	20	6.5%	4.8%	7.9%	5.7%	5.9%	4.9%	9.0%	7.6%	7.2%	5.9%	6.8%	7.4%
30	30	5.1%	4.8%	5.9%	5.9%	5.1%	5.0%	6.8%	7.1%	6.1%	6.1%	6.7%	5.6%
40	40	5.3%	4.8%	5.8%	5.5%	5.3%	4.7%	6.6%	6.1%	5.5%	5.4%	5.9%	6.1%
50	50	5.0%	5.3%	5.7%	6.1%	5.2%	5.8%	5.7%	6.5%	4.9%	5.6%	5.8%	6.3%
60	60	4.9%	5.6%	5.7%	5.8%	4.9%	5.4%	6.5%	6.0%	6.0%	5.2%	5.4%	5.9%
70	70	4.7%	5.3%	4.3%	5.6%	4.9%	5.5%	5.9%	6.0%	5.7%	5.7%	5.5%	5.0%
80	80	4.6%	5.3%	5.2%	6.0%	4.7%	5.2%	6.0%	6.4%	5.6%	5.9%	4.7%	6.0%
90	90	5.1%	5.1%	5.7%	5.7%	5.2%	4.9%	5.7%	5.6%	5.4%	5.4%	5.6%	4.4%
100	100	4.7%	5.1%	4.9%	5.3%	4.7%	5.2%	5.1%	6.1%	4.9%	5.7%	5.1%	6.2%
10	50	5.6%	4.0%	8.6%	7.3%	5.5%	4.1%	8.9%	9.7%	5.9%	6.8%	7.3%	6.6%
20	50	5.2%	5.1%	6.2%	7.3%	5.3%	5.1%	6.7%	7.4%	5.2%	5.8%	6.5%	5.8%
10	100	4.8%	5.1%	7.0%	7.7%	4.7%	5.0%	8.9%	8.9%	5.9%	5.8%	5.8%	6.8%
20	100	4.1%	3.8%	5.8%	5.0%	3.9%	3.7%	6.7%	7.7%	5.4%	5.9%	5.8%	6.6%
50	100	5.7%	4.8%	6.2%	5.0%	5.7%	4.8%	6.1%	4.6%	5.6%	4.3%	5.4%	5.3%

Note: The null hypothesis should be rejected 5% of the time. Number of monte carlo simulations is 2000.

Table 2
Rejection probabilities for a true null hypothesis

True model: random effects on both Group1 and Group 2

Number of Group 1 Clusters	Number of Group 2 Clusters	Assumption about errors in construction of Variance										Group fixed effects, Two-way cluster- robust	
		Assume independent errors		One-way cluster robust (cluster on group1)		Two-way random effects		Two-way cluster- robust		Two-way cluster- robust, T critical values			
10	10	23.4%	23.7%	13.7%	33.2%	6.2%	6.6%	17.4%	17.4%	12.6%	13.4%	9.0%	9.1%
20	20	34.0%	32.3%	8.6%	42.7%	5.7%	5.3%	10.3%	9.6%	8.7%	7.6%	7.0%	8.2%
30	30	39.7%	41.7%	7.4%	50.6%	5.2%	5.6%	8.6%	9.1%	7.8%	7.7%	6.5%	7.4%
40	40	47.7%	47.6%	8.7%	55.2%	6.5%	5.4%	9.0%	9.3%	7.9%	8.1%	6.2%	5.5%
50	50	50.0%	50.4%	6.0%	58.8%	4.9%	4.8%	7.0%	6.7%	6.3%	6.2%	6.1%	5.1%
60	60	54.5%	56.3%	6.4%	64.1%	5.6%	6.5%	6.7%	6.1%	5.9%	5.6%	5.0%	5.3%
70	70	54.2%	57.6%	5.5%	64.8%	4.8%	6.0%	6.4%	6.5%	5.9%	6.0%	4.9%	5.6%
80	80	61.1%	60.9%	6.5%	67.0%	4.9%	4.7%	6.3%	7.0%	5.7%	6.5%	5.9%	5.0%
90	90	63.7%	60.4%	5.4%	67.0%	4.9%	5.0%	5.9%	6.1%	5.5%	5.8%	5.4%	5.3%
100	100	62.2%	60.4%	5.8%	67.9%	5.3%	3.6%	6.4%	5.3%	6.1%	5.1%	5.4%	4.9%
10	50	49.9%	21.3%	13.3%	33.4%	8.9%	4.0%	15.0%	9.3%	10.2%	5.8%	7.1%	6.5%
20	50	50.8%	33.1%	9.8%	44.5%	6.7%	4.5%	9.3%	8.1%	8.2%	6.2%	6.4%	5.3%
10	100	63.0%	21.0%	14.1%	31.7%	10.4%	3.3%	14.2%	8.1%	9.2%	4.6%	5.8%	6.6%
20	100	62.9%	33.9%	10.0%	43.7%	6.2%	3.7%	9.2%	6.3%	7.7%	4.6%	6.1%	6.2%
50	100	63.4%	60.7%	5.6%	67.9%	5.0%	5.4%	6.0%	6.7%	5.8%	6.3%	4.8%	5.9%

Note: See Table 1.

Table 3**Rejection probabilities for a true null hypothesis****True model: a random effects common to each Group, and a hetercedastic component.**

Number of Group 1 Clusters	Number of Group 2 Clusters	Assumption about errors in construction of Variance											
		Assume independent errors		One-way cluster robust (cluster on group1)		Two-way random effects		Two-way cluster- robust		Two-way cluster- robust, T critical values		Group fixed effects, Two-way cluster-robust	
10	10	8.0%	7.9%	15.7%	9.8%	15.9%	14.3%	18.4%	16.5%	14.5%	12.9%	8.6%	8.9%
20	20	7.0%	5.4%	9.5%	7.1%	13.0%	10.8%	11.9%	10.9%	10.3%	8.8%	7.1%	5.9%
30	30	5.9%	6.9%	7.0%	8.1%	9.7%	10.8%	8.2%	9.2%	7.1%	8.0%	6.2%	6.0%
40	40	5.9%	6.5%	5.4%	7.3%	8.7%	9.7%	6.8%	7.8%	5.9%	7.1%	5.6%	5.4%
50	50	7.9%	7.1%	6.6%	7.5%	9.7%	8.9%	6.7%	6.1%	6.1%	5.4%	5.4%	5.3%
60	60	7.6%	8.2%	6.0%	8.5%	8.8%	9.4%	7.1%	7.0%	6.3%	6.5%	5.9%	5.7%
70	70	9.3%	8.6%	6.6%	9.1%	9.6%	9.9%	7.3%	6.2%	6.8%	5.9%	6.3%	5.5%
80	80	10.3%	9.0%	6.0%	10.2%	9.5%	9.0%	6.7%	6.8%	5.9%	6.2%	5.0%	5.1%
90	90	9.9%	9.1%	5.4%	10.2%	9.4%	8.1%	5.3%	6.6%	5.2%	6.0%	5.5%	6.1%
100	100	11.6%	10.5%	6.1%	11.2%	9.6%	9.0%	6.4%	6.9%	6.0%	6.4%	4.4%	5.4%
10	50	8.1%	5.6%	12.9%	8.8%	12.9%	12.3%	13.7%	9.8%	9.6%	5.9%	6.1%	6.0%
20	50	7.6%	7.5%	7.9%	8.1%	10.5%	11.5%	9.2%	8.6%	7.6%	6.6%	5.2%	6.7%
10	100	10.0%	6.4%	10.4%	9.4%	10.1%	13.0%	11.3%	10.0%	7.3%	6.8%	7.5%	6.2%
20	100	11.7%	5.3%	9.2%	6.7%	10.8%	10.1%	9.4%	6.4%	7.7%	4.5%	5.1%	6.2%
50	100	11.2%	8.1%	6.7%	8.7%	9.9%	10.0%	6.9%	6.8%	6.1%	6.2%	6.1%	5.2%

Note: See Table 1.

Table 4

**Rejection probabilities for a true null hypothesis
Monte Carlos with micro (CPS) data**

	RHS control variables			
	quartic in age, 4 education dummies	quartic in age, 4 education dummies, state fixed effects	quartic in age, 4 education dummies, year fixed effects	quartic in age, 4 education dummies, state and year fixed effects
Standard error assumption:				
Heterscedasticity robust	91.6%	92.1%	82.2%	79.0%
One-way cluster robust (cluster on state-by-year cell)	19.8%	22.4%	13.1%	13.9%
One-way cluster robust (cluster on state)	16.2%	17.0%	12.0%	15.0%
One-way cluster robust (cluster on year)	10.2%	8.9%	8.7%	7.6%
Two-way cluster-robust (cluster on state and year)	7.2%	6.9%	7.6%	7.1%

Note: Data come from 1.3 million employed women from the 1979-1999 March CPS. Table reports rejection rates for testing a (true) null hypothesis of zero on the coefficient of fake treatments. The "treatments" are generated as $t = e_s + 2 e_y$, with e_s a state-specific autoregressive component and e_y a year-specific "spatial" autoregressive component. The outcome is also modified by an independent year-specific "spatial" autoregressive component. See text for details. 2000 Monte Carlo replications

Table 5
Replication of Hersch (1998)

		Industry Injury Rate	Variable Occupation Injury Rate		
Estimated slope coefficient:		-1.894	-0.465		
Estimated standard errors and p-values:	Default (iid)	(0.415)	{0.0000}	(0.235)	{0.0478}
	Heteroscedastic robust	(0.397)	{0.0000}	(0.260)	{0.0737}
	One-way cluster on Industry	(0.643)	{0.0032}	(0.251)	{0.0639}
	One-way cluster on Occupation	(0.486)	{0.0001}	(0.363)	{0.2002}
	Two-way clustering	(0.702)	{0.0070}	(0.357)	{0.1927}

Note: Replication of Hersch (1998), pg 604, Table 3, Panel B, Column 4. Standard errors in parentheses. P-values from a test of each coefficient equal to zero in brackets. Data are 5960 observations on working men from the Current Population Survey. Both columns come from the same regression. There are 211 industries and 387 occupations in the data set.

Table 6
Replication of Gruber and Madrian (1995)

		Model	
		Probit coefficient	Probit average marginal effect
			OLS
Estimated slope coefficient (* 1000):		13.264	1.644
Estimated marginal effect (* 1000)			1.544
Estimated standard errors (* 1000):	Default (iid)	(5.709)	(0.665)
	Heteroscedastic robust	(5.732)	(0.668)
	One-way cluster on state-year	(6.265)	(0.729)
	One-way cluster on household id	(5.866)	(0.683)
	One-way cluster on hhid-by-state-year	(5.732)	(0.668)
	Two-way clustering	(6.389)	(0.718)
	One-way cluster on State	(6.030)	(0.703)

Note: Replication of Gruber and Madrian (1995), pg 943, Table 3, Model 1, Column 1. Standard errors in parentheses. Data are 39,063 observations on 55-64 year-old men from the 1980-1990 Current Population Surveys.