Robust Inference with Multi-way Clustering

A. Colin Cameron,*Jonah B. Gelbach,†and Douglas L. Miller‡

This version: May 15, 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 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 underestimated 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. These models impose strong assumptions, including homoskedasticity and errors equicorrelated within cluster.

In this paper we take a less parametric approach that generalizes one-way cluster-robust stan-

dard errors to the non-nested multi-way clustering case. One-way "cluster-robust" standard errors generalize those of White (1980) for independent heteroskedastic errors. Key references include Pfeffermann 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 (2008) 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. Empirical papers that cite earlier drafts of our paper include Baughman and Smith (2007), Beck, Demirguc-Kunt, Laeven, and Levine (2008), Cascio and Schanzenbach (2007), Cujjpers 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, (2008), 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 $\sum_i \sum_j w(i,j)\, x_i x_j' \widehat{\varepsilon}_i \widehat{\varepsilon}_j$. 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 specifies weights $w(i,j)$ that decay 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 beween 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, leading to conservative statistical inference.

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 screeing epidemiological study. Petersen (2009) compares a number of approaches for OLS estimation in a finance panel setting, using results by Thompson (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 $\min(G, H) \to \infty$, 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. 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 $y_{ig} = \mathbf{x}_{ig}'\boldsymbol{\beta} + u_{ig}$, where i denotes the i^{th} of N individuals in the sample, g denotes the g^{th} of G clusters, $\mathrm{E}[u_{ig}|\mathbf{x}_{ig}] = 0$, and error independence across clusters is assumed so that for $i \neq j$

$$\mathrm{E}[u_{ig}u_{jg'}|\mathbf{x}_{ig}, \mathbf{x}_{jg'}] = 0, \text{ unless } g = g'. \tag{2.1}$$

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

Grouping observations by cluster, so $\mathbf{y}_g = \mathbf{X}_g\boldsymbol{\beta} + \mathbf{u}_g$, and stacking over clusters yields $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$, where $\mathbf{y}$ and $\mathbf{u}$ are $N \times 1$ vectors, and $\mathbf{X}$ is an $N \times K$ matrix.

The OLS estimator is

$$\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y} = \left(\sum_{g=1}^{G}\mathbf{X}_g'\mathbf{X}_g\right)^{-1}\sum_{g=1}^{G}\mathbf{X}_g'\mathbf{y}_g, \tag{2.2}$$

where $\mathbf{X}_g$ has dimension $N_g \times K$ and $\mathbf{y}_g$ has dimension $N_g \times 1$, with N_g observations in cluster g.

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

$$\left(\lim_{G\to\infty}\frac{1}{G}\sum_{g=1}^{G}\mathrm{E}\left[\mathbf{X}_g'\mathbf{X}_g\right]\right)^{-1}\left(\lim_{G\to\infty}\frac{1}{G}\sum_{g=1}^{G}\mathrm{E}\left[\mathbf{X}_g'\mathbf{u}_g\mathbf{u}_g'\mathbf{X}_g\right]\right)\left(\lim_{G\to\infty}\frac{1}{G}\sum_{g=1}^{G}\mathrm{E}\left[\mathbf{X}_g'\mathbf{X}_g\right]\right)^{-1}. \tag{2.3}$$

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 $s^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}$, where s is the estimated 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 Kloek (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 $\mathbf{\Omega}_g = \mathrm{V}[\mathbf{u}_g|\mathbf{X}_g] = \mathrm{E}[\mathbf{u}_g\mathbf{u}_g'|\mathbf{X}_g]$, in which case $\mathrm{E}[\mathbf{X}_g'\mathbf{u}_g\mathbf{u}_g'\mathbf{X}_g] = \mathrm{E}[\mathbf{X}_g'\mathbf{\Omega}_g\mathbf{X}_g]$ can be estimated given a consistent estimate $\widehat{\mathbf{\Omega}}_g$ of $\mathbf{\Omega}_g$, and feasible GLS estimation is then additionally possible.

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

$$\widehat{\mathrm{V}}[\widehat{\boldsymbol{\beta}}] = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\left(\sum_{g=1}^{G}\mathbf{X}_g'\widehat{\mathbf{u}}_g\widehat{\mathbf{u}}_g'\mathbf{X}_g\right)\left(\mathbf{X}'\mathbf{X}\right)^{-1}, \tag{2.4}$$

where $\widehat{\mathbf{u}}_g = \mathbf{y}_g - \mathbf{X}_g\widehat{\boldsymbol{\beta}}$. This provides a consistent estimate of the variance matrix if $G^{-1}\sum_{g=1}^{G}\mathbf{X}_g'\widehat{\mathbf{u}}_g\widehat{\mathbf{u}}_g'\mathbf{X}_g - G^{-1}\sum_{g=1}^{G}\mathrm{E}[\mathbf{X}_g'\mathbf{u}_g\mathbf{u}_g'\mathbf{X}_g] \xrightarrow{p} \mathbf{0}$ as $G \to \infty$. 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. Most recently, Hansen (2007) provides asymptotic theory for panel data where $T \to \infty$ ($N_g \to \infty$ in the notation above) in addition to $N \to \infty$ ($G \to \infty$ in the notation above). Note that (2.4) 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 (2.4) is that

$$\widehat{\mathrm{V}}[\widehat{\boldsymbol{\beta}}] = (\mathbf{X}'\mathbf{X})^{-1}\widehat{\mathbf{B}}(\mathbf{X}'\mathbf{X})^{-1}, \tag{2.5}$$

where the central matrix

$$\begin{aligned}
\widehat{\mathbf{B}} &= \sum_{g=1}^{G}\mathbf{X}_g'\widehat{\mathbf{u}}_g\widehat{\mathbf{u}}_g'\mathbf{X}_g \qquad (2.6)\\
&= \mathbf{X}'\begin{bmatrix} \widehat{\mathbf{u}}_1\widehat{\mathbf{u}}_1' & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \widehat{\mathbf{u}}_2\widehat{\mathbf{u}}_2' & & \vdots \\ \vdots & & \ddots & \mathbf{0} \\ \mathbf{0} & \cdots & \cdots & \widehat{\mathbf{u}}_G\widehat{\mathbf{u}}_G' \end{bmatrix}\mathbf{X} \\
&= \mathbf{X}'\left(\widehat{\mathbf{u}}\widehat{\mathbf{u}}'.*\mathbf{S}^G\right)\mathbf{X},
\end{aligned}$$

where $.*$ denotes element-by-element multiplication and $\mathbf{S}^G$ 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. The (a,b)-th element of $\widehat{\mathbf{B}}$ is $\sum_{i=1}^{N}\sum_{j=1}^{N} x_{ia}x_{jb}\widehat{u}_i\widehat{u}_j\mathbf{1}[i,j \text{ in same cluster}]$, where $\widehat{u}_i = y_i - \mathbf{x}_i'\widehat{\boldsymbol{\beta}}$.

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

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 $g \in \{1,2,...,G\}$, as well as to a group $h \in \{1,2,...,H\}$, and we have $y_{igh} = \mathbf{x}_{igh}'\boldsymbol{\beta} + u$, where we assume that for $i \neq j$

$$\mathrm{E}[u_{igh}u_{jg'h'}|\mathbf{x}_{igh},\mathbf{x}_{jg'h'}] = 0, \text{ unless } g = g' \text{ or } h = h'. \tag{2.7}$$

If errors belong to the same group (along either dimension), they may have an arbitrary correlation.

The intuition for the variance estimator in this case is a simple extension of (2.6) for one-way clustering. We keep those elements of $\widehat{\mathbf{u}}\widehat{\mathbf{u}}'$ where the i^{th} and j^{th} observations share a cluster in <u>any</u> dimension. Then

$$\widehat{\mathbf{B}} = \mathbf{X}'(\widehat{\mathbf{u}}\widehat{\mathbf{u}}'.*\mathbf{S}^{GH})\mathbf{X}, \tag{2.8}$$

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 $\widehat{\mathbf{B}}$ is $\sum_{i=1}^{N}\sum_{j=1}^{N} x_{ia}x_{jb}\widehat{u}_i\widehat{u}_j\mathbf{1}[i,j \text{ share any cluster}]$.

$\widehat{\mathbf{B}}$ and hence $\widehat{\mathrm{V}}[\widehat{\boldsymbol{\beta}}]$ can be calculated directly. However, $\widehat{\mathrm{V}}[\widehat{\boldsymbol{\beta}}]$ can also be represented as the sum of one-way cluster-robust matricies. 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 $g \in \{1,2,...,G\}$, $\mathbf{S}^H$ with ij^{th} entry equal to one if the i^{th} and j^{th} observation belong to the same cluster $h \in \{1,2,...,H\}$, 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

$g \in \{1, 2, ..., G\}$ and the same cluster $h \in \{1, 2, ..., H\}$. Then $\mathbf{S}^{GH} = \mathbf{S}^G + \mathbf{S}^H - \mathbf{S}^{G \cap H}$ so

$$\widehat{\mathbf{B}} = \mathbf{X}'(\widehat{\mathbf{u}}\widehat{\mathbf{u}}'. * \mathbf{S}^G)\mathbf{X} + \mathbf{X}'(\widehat{\mathbf{u}}\widehat{\mathbf{u}}'. * \mathbf{S}^H)\mathbf{X} - \mathbf{X}'(\widehat{\mathbf{u}}\widehat{\mathbf{u}}'. * \mathbf{S}^{G \cap H})\mathbf{X}. \tag{2.9}$$

Substituting (2.9) into (2.5) yields

$$\begin{aligned} \widehat{\mathrm{V}}[\widehat{\boldsymbol{\beta}}] &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\widehat{\mathbf{u}}\widehat{\mathbf{u}}'. * \mathbf{S}^G)\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &\quad + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\widehat{\mathbf{u}}\widehat{\mathbf{u}}'. * \mathbf{S}^H)\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &\quad - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\widehat{\mathbf{u}}\widehat{\mathbf{u}}'. * \mathbf{S}^{G \cap H})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}, \end{aligned} \tag{2.10}$$

or

$$\widehat{\mathrm{V}}[\widehat{\boldsymbol{\beta}}] = \widehat{\mathrm{V}}^G[\widehat{\boldsymbol{\beta}}] + \widehat{\mathrm{V}}^H[\widehat{\boldsymbol{\beta}}] - \widehat{\mathrm{V}}^{G \cap H}[\widehat{\boldsymbol{\beta}}]. \tag{2.11}$$

The three components can be separately computed by OLS regression of $\mathbf{y}$ on $\mathbf{X}$ with variance matrix estimates based on: (1) clustering on $g \in \{1, 2, ..., G\}$; (2) clustering on $h \in \{1, 2, ..., H\}$; and (3) clustering on $(g, h) \in \{(1, 1), ..., (G, H)\}$. $\widehat{\mathrm{V}}[\widehat{\boldsymbol{\beta}}]$ is 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. We are presently unaware of a systematic, data-driven approach to this issue. From the discussion after (2.3) a necessary condition for a dimension 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.

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 $\mathbf{X}'\widehat{\mathbf{u}}$ 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.

Small-sample modifications of (2.4) 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 $\sqrt{c}\widehat{\mathbf{u}}_g$ in (2.4) rather than $\widehat{\mathbf{u}}_g$, 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 simulations and applications.

A practical matter that can arise when implementing the two-way robust estimator is that the resulting variance estimate $\widehat{\mathrm{V}}[\widehat{\boldsymbol{\beta}}]$ may have negative elements on the diagonal. In some applications with fixed effects, $\widehat{\mathrm{V}}[\widehat{\boldsymbol{\beta}}]$ may be non positive-definite, but the subcomponent of $\widehat{\mathrm{V}}[\widehat{\boldsymbol{\beta}}]$ 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 uses the eigendecomposition of the estimated variance matrix and converts any negative eigenvalue(s) to zero. Specifically, decompose the variance matrix into the product of its eigenvectors and eigenvalues: $\widehat{\mathrm{V}}[\widehat{\boldsymbol{\beta}}] = U\Lambda U'$, with U containing the eigenvectors of $\widehat{\mathrm{V}}$, and $\Lambda = \mathrm{Diag}[\lambda_1, ..., \lambda_d]$ containing the eigenvalues of

$\widehat{\mathrm{V}}$. Then create $\Lambda^+ = \mathrm{Diag}[\lambda_1^+, ..., \lambda_d^+]$, with $\lambda_j^+ = \max(0, \lambda_j)$, and use $\widehat{\mathrm{V}}^+[\widehat{\boldsymbol{\beta}}] = U\Lambda^+U'$ 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, $\widehat{\mathrm{V}}^+[\widehat{\boldsymbol{\beta}}]$ always yields a positive definite variance matrix estimate, and keeping all draws (using $\widehat{\mathrm{V}}^+[\widehat{\boldsymbol{\beta}}]$ 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 $G \to \infty$ and $H \to \infty$.

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 $G \to \infty$ and $H \to \infty$, 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.

2.4. Multi-Way Clustering

Our approach generalizes to clustering in more than two dimensions. Suppose there are D dimensions of clustering. Let G_d denote the number of clusters in dimension d. Let the D-vector $\boldsymbol{\delta}_i = \boldsymbol{\delta}(i)$, where the function $\boldsymbol{\delta} : \{1, 2, ..., N\} \to \times_{d=1}^{D}\{1, 2, ..., G_d\}$ lists the cluster membership in each dimension for each observation. Thus $\mathbf{1}[i, j \text{ share a cluster}] = 1 \Leftrightarrow \delta_{id} = \delta_{jd}$ for some

$d \in \{1, 2, ..., D\}$, where δ_{id} denotes the d^{th} element of $\boldsymbol{\delta}_i$.

Let $\mathbf{r}$ be a D-vector, with d^{th} coordinate equal to r_d, and define the set $R \equiv \{\mathbf{r}\text{: } r_d \in \{0, 1\}, d = 1, 2, ..., D, r \neq \mathbf{0}\}$. 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 $r_d = 1$. 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.

Finally, define the $2^D - 1$ matrices

$$\widetilde{\mathbf{B}}_{\mathbf{r}} \equiv \sum_{i=1}^{N} \sum_{j=1}^{N} \mathbf{x}_i \mathbf{x}_j' \widehat{u}_i \widehat{u}_j I_{\mathbf{r}}(i, j), \quad \mathbf{r} \in R. \tag{2.12}$$

Our proposed estimator may then be written as $\widehat{\mathrm{V}}[\widehat{\boldsymbol{\beta}}] = (\mathbf{X}'\mathbf{X})^{-1} \widetilde{\mathbf{B}} (\mathbf{X}'\mathbf{X})^{-1}$, where

$$\widetilde{\mathbf{B}} \equiv \sum_{\|\mathbf{r}\|=k,\ \mathbf{r} \in R} (-1)^{k+1} \widetilde{\mathbf{B}}_r. \tag{2.13}$$

Cases in which the matrix $\widetilde{\mathbf{B}}_r$ 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

$$\left(\widetilde{\mathbf{B}}_{(1,0,0)} + \widetilde{\mathbf{B}}_{(0,1,0)} + \widetilde{\mathbf{B}}_{(0,0,1)}\right) - \left(\widetilde{\mathbf{B}}_{(1,1,0)} + \widetilde{\mathbf{B}}_{(1,0,1)} + \widetilde{\mathbf{B}}_{(0,1,1)}\right) + \widetilde{\mathbf{B}}_{(1,1,1)}. \tag{2.14}$$

To prove that $\widetilde{\mathbf{B}} = \widehat{\mathbf{B}}$ identically, where $\widehat{\mathbf{B}}$ is the $D-$dimensional analog of (2.8), 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 $I(i, j) = 1$, $\sum_{\|\mathbf{r}\|=k,\ \mathbf{r} \in R} (-1)^{k+1} I_{\mathbf{r}}(i, j) = 1$. This fact, which is an application of the inclusion-exclusion principle for set cardinality, ensures that $\widetilde{\mathbf{B}}$ and $\widehat{\mathbf{B}}$ are identical in every sample.

One potential 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. These procedures are qualitatively the same as for OLS. In the two-way clustering case, analogous to (2.11) we obtain three different cluster-robust variance matrices and add the first two variance matrices and subtract the third.

We begin with an m-estimator that solves $\sum_{i=1}^{N} \mathbf{h}_i(\widehat{\boldsymbol{\theta}}) = \mathbf{0}$. 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, $\widehat{\boldsymbol{\theta}}$ is asymptotically normal with estimated variance matrix

$$\widehat{\mathrm{V}}[\widehat{\boldsymbol{\theta}}] = \widehat{\mathbf{A}}^{-1}\widehat{\mathbf{B}}\widehat{\mathbf{A}}'^{-1}, \tag{2.15}$$

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

For one-way clustering $\widehat{\mathbf{B}} = \sum_{g=1}^{G} \widehat{\mathbf{h}}_g\widehat{\mathbf{h}}_g'$ where $\widehat{\mathbf{h}}_g = \sum_{i=1}^{N_g} \widehat{\mathbf{h}}_{ig}$. Clustering may or may not lead to parameter inconsistency, depending on whether $\mathrm{E}[\mathbf{h}_i(\boldsymbol{\theta})] = \mathbf{0}$ 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 $\mathrm{E}[y_{ig}|\mathbf{x}_{ig}] = \Phi(\mathbf{x}_{ig}'\boldsymbol{\beta})$, even in the presence of clustering. An alternative approach is a random effects approach. Let $y_{ig} = 1$ if $y_{ig}^* > 0$

where $y_{ig}^* = \mathbf{x}_{ig}'\boldsymbol{\beta} + \varepsilon_g + \varepsilon_{ig}$, 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 $\mathrm{E}[y_{ig}|\mathbf{x}_{ig}] = \Phi(\mathbf{x}_{ig}'\boldsymbol{\beta}/\sqrt{1+\sigma_g^2})$, so that the moment condition is no longer $\mathrm{E}[y_{ig}|\mathbf{x}_{ig}] = \Phi(\mathbf{x}_{ig}'\boldsymbol{\beta})$. When $\mathrm{E}[\mathbf{h}_i(\boldsymbol{\theta})] \neq \mathbf{0}$ 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 $\widehat{u}_i\mathbf{x}_i$ in (2.12) replaced by $\widehat{\mathbf{h}}_i$. Then $\widehat{\boldsymbol{\theta}}$ is asymptotically normal with estimated variance matrix $\widehat{\mathrm{V}}[\widehat{\boldsymbol{\theta}}] = \widehat{\mathbf{A}}^{-1}\widetilde{\mathbf{B}}\widehat{\mathbf{A}}'^{-1}$, with $\widehat{\mathbf{A}}$ defined as in (2.15) and $\widetilde{\mathbf{B}}$ defined as in (2.13), with matrices $\widetilde{\mathbf{B}}_r$ defined analogously as

$$\widetilde{\mathbf{B}}_{\mathbf{r}} \equiv \sum_{i=1}^{N}\sum_{j=1}^{N} \widehat{\mathbf{h}}_i\widehat{\mathbf{h}}_j' I_{\mathbf{r}}(i,j), \quad \mathbf{r} \in R. \tag{2.16}$$

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. Each of the one-way cluster robust matrices is estimated by a pairs cluster bootstrap that resamples with replacement from the appropriate cluster dimension. They are then combined as if they had been estimated analytically.

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 $\widehat{\boldsymbol{\theta}}$ minimizes $Q(\boldsymbol{\theta}) = \left(\sum_{i=1}^N \mathbf{h}_i(\boldsymbol{\theta})\right)' \mathbf{W} \left(\sum_{i=1}^N \mathbf{h}_i(\boldsymbol{\theta})\right)$, where $\mathbf{W}$ is a symmetric positive definite weighting matrix. Under standard regularity conditions $\widehat{\boldsymbol{\theta}}$ is asymptotically normal with estimated variance matrix

$$\widehat{\mathrm{V}}[\widehat{\boldsymbol{\theta}}] = \left(\widehat{\mathbf{A}}'\mathbf{W}\widehat{\mathbf{A}}\right)^{-1} \widehat{\mathbf{A}}'\mathbf{W}\widetilde{\mathbf{B}}\mathbf{W}\widehat{\mathbf{A}} \left(\widehat{\mathbf{A}}'\mathbf{W}\widehat{\mathbf{A}}\right)^{-1}, \tag{2.17}$$

where $\widehat{\mathbf{A}} = \sum_i \left.\frac{\partial \mathbf{h}_i}{\partial \boldsymbol{\theta}'}\right|_{\widehat{\boldsymbol{\theta}}}$, and $\widetilde{\mathbf{B}}$ is an estimate of $\mathrm{V}[\sum_i \mathbf{h}_i]$ that can be computed using (2.13) and (2.16).

3. Monte Carlo Exercises

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

The first Monte Carlo exercise is based on a two-way random effects model for the errors with an additional heteroscedastic component.

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

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

where $\beta_0 = \beta_1 = \beta_2 = 1$. 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 five designs are square with $G = H$ varying from 10 to 100, and the remaining designs are rectangular with $G < H$.

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 errors

$$u_{gh} = \varepsilon_g + \varepsilon_h + \varepsilon_{gh}, \tag{3.2}$$

where ε_g and ε_h are $\mathcal{N}[0, 1]$ and independent of both regressors, and $\varepsilon_{gh} \sim \mathcal{N}[0, |x_{1gh} \times x_{2gh}|]$ is conditionally heteroskedastic.

We consider inference based on the OLS slope coefficients $\widehat{\beta}_1$ and $\widehat{\beta}_2$, 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 $t_1 = |\widehat{\beta}_1 - 1|/\text{se}[\widehat{\beta}_1] \geq 1.96$, and $t_2 = |\widehat{\beta}_2 - 1|/\text{se}[\widehat{\beta}_2] \geq 1.96$. 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, we also report rejection rates when the critical value is $t_{.025;\min(G,H)-1}$. Donald and Lang (2007) suggest using the $T(G - L)$ distribution, with L the number of cluster-invariant regressors.

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

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

2. One-way cluster-robust (cluster on first group): This uses one-way cluster-robust standard errors, based on (2.4) with small-sample modification, that correct for clustering on the first grouping $g \in \{1, 2, ..., G\}$ 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 $\widehat{\mathrm{V}}[\widehat{\boldsymbol{\beta}}] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\widehat{\Omega}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$, where $\widehat{\Omega}$ is a consistent estimate of $\mathrm{V}[\mathbf{u}|\mathbf{X}]$ 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 (2.11), that allows for two-way clustering but does not restrict it to follow a two-way random effects model.

Here the first three methods will in general fail. However, inference on β_1 (but not β_2) is valid using the second method, due to the particular dgp used here. Specifically, the regressor x_{1gh} is correlated over only g (and not h), so that for inference on β_1 it is sufficient 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, then one-way standard errors for $\widehat{\beta}_1$ would also be incorrect. The fourth method is asymptotically valid.

Table 1 reports results based on 2000 simulations. This yields a 95% confidence interval of (4.0%, 6.0%) for the Monte Carlo rejection rate, given that the true rejection rate is 5%.

We first consider all but the last two columns of Table 1. The simulations with the largest sample, the $G = H = 100$ row presented in bold in Table 1, confirm expectations. The two-way cluster-robust method performs well. 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%. Controlling for one-way clustering on group 1 improves inference on β_1, but the tests on β_2 then over-reject even more than when iid errors are assumed. The Moulton-type two-way effects method fails when heteroskedasticity is present, with lowest rejection rate in Table 1 of 8.8% and

rejection rates that generally exceed even those assuming iid errors.

The two-way cluster robust standard errors are 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, 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.

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 Table 1, 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.

We have also examined alternative d.g.p.s in which the errors and regressors are distributed i.i.d., and d.g.p.s with the classical homoscedastic two-way random-effects design. Results from these simulations can be found in our working paper (Cameron, Gelbach, and Miller 2009).

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 decreases with lag length.

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

$$y_{ist} = \alpha d_{st} + \mathbf{x}_{ist}'\boldsymbol{\beta} + \delta_s + \gamma_t + u_{ist}, \tag{3.3}$$

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 (69 dummies), as our d.g.p. enables these fixed effects to be identified. In most of their simulations Bertand et al. (2004) run regressions on data aggregated into state-year cells. Here we work with the individual-level data in part to demonstrate the feasibility of our methods for large data sets.

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

$$d_{st} = d_{st}^s + 2d_{st}^t. \tag{3.4}$$

The variable d_{st}^s is a within-state AR(1) variable $d_{st}^s = 0.6d_{st-1}^s + v_{st}^s$, with v_{st}^s iid $\mathcal{N}[0, 1]$, and is generated independently from all other variables. d_{st}^s is independent across states. Similarly, the variable d_{st}^t is a within-year AR(1) variable, $d_{st}^t = 0.6d_{s-1,t}^t + v_{st}^t$, correlated over states, with v_{st}^t 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 $y_{ist}^{new} = y_{ist}^{original} + 0.01 w_{st}^t$, with w_{st}^t generated similarly to (but independent of) d_{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 (3.3) is estimated, and the null hypothesis that $\alpha = 0$ is rejected at significance level 0.05 if $|\widehat{\alpha}|/\text{se}[\widehat{\alpha}] > 1.96$. Given the design used here, $\widehat{\alpha}$ is consistent, and the correct asymptotic rejection rates for the simulation results in Table 2 will be 5%, provided that a consistent estimate of the standard error is used.

The first column of Table 2 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%. Clustering on the 50 states does better than clustering on the 1,050 state-year cells. 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. 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 thus might be well approximated by a $T(H-1)$ distribution.

For the second column of Table 2, 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 2, 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. 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. In this example there are 5,960 individuals in 211 industries and 387 occupations. The model is

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

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. She is able to control only for one source of clustering - industry or occupation - and not both simultaneously.

We replicate results for column 4 of Panel B of Table 1 of Hersch (1998), with both *rind* and *rocc* included as regressors. We report several 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 3 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*. In this application it is it is most important to cluster on industry for *rind*, and to cluster on occupation for *rocc*.

Our two-way cluster-robust method permits clustering on both industry and occupation. 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.

4.2. Rose and Engel - bilateral trade model

A common setting for two-way clustering 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.

4.3. Other examples

We have also examined the importance of clustering in the context of CPS rotating panel design. In an applicaion based on Gruber and Madrian's (1995) study of health insurance availability and retirement, we examine the importance of clustering on state-year cell (359 clusters) and by household (26,383 clusters). In this particular application the impact of two-way clustering is modest compared to clustering at either level. For more details see Cameron, Gelbach and Miller (2009).

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).

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 greatest when the errors are correlated over two or more dimensions and, in addition, the regressors of interest are 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. The Hersch 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 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 $\mathrm{E}[\mathbf{z}_i(\boldsymbol{\theta}_0)] = \mathbf{0}$ 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 $\widetilde{\boldsymbol{\theta}}$ that solves $\sum_{i=1}^N \mathbf{z}_i(\widetilde{\boldsymbol{\theta}}) = \mathbf{0}$. Given two-way clustering with typical cluster (g,h), $\mathbf{z}_i(\boldsymbol{\theta}) = \mathbf{z}_{igh}(\boldsymbol{\theta})$ and

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

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

$$\mathbf{z}_{gh}(\boldsymbol{\theta}) = \sum_{i \in C_{gh}} \mathbf{z}_{igh}(\boldsymbol{\theta}) \tag{A.2}$$

combines observations in cluster (g,h).

For models with $m > q$, the more general GMM estimator $\widehat{\boldsymbol{\theta}}$ maximizes $Q(\boldsymbol{\theta}) = \left(N^{-1}\sum_{i=1}^N \mathbf{z}_i(\boldsymbol{\theta})\right)' \mathbf{W} \left(N^{-1}\sum_{i=1}^N \mathbf{z}_i(\boldsymbol{\theta})\right)$, 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 $\widehat{\boldsymbol{\theta}}$ is consistent for $\boldsymbol{\theta}_0$, that $G \to \infty$ and $H \to \infty$ at the same rate, so that $G/H \to$ 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

$$\sqrt{G}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}[\mathbf{0},\ \left(\mathbf{A}_0'\mathbf{W}_0\mathbf{A}_0\right)^{-1}\mathbf{A}_0'\mathbf{W}_0\mathbf{B}_0\mathbf{W}_0\mathbf{A}_0\left(\mathbf{A}_0'\mathbf{W}_0\mathbf{A}_0\right)^{-1}], \tag{A.3}$$

where $\mathbf{A}_0 = \lim \mathrm{E}\left[(GH)^{-1}\sum_{i=1}^{N}\partial \mathbf{z}_{i0}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}'\right]$ and

$$\mathbf{B}_0 = \lim\ \mathrm{E}\left[G^{-1}H^{-2}\sum\nolimits_{i=1}^{N}\sum\nolimits_{j=1}^{N}\mathbf{z}_i(\boldsymbol{\theta})\mathbf{z}_j(\boldsymbol{\theta})'\right]. \tag{A.4}$$

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

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

$$\begin{aligned}
& \mathrm{E}\left[\ G^{-1}H^{-2}\sum\nolimits_{i=1}^{N}\sum\nolimits_{j=1}^{N}\mathbf{z}_i(\boldsymbol{\theta})\mathbf{z}_j(\boldsymbol{\theta})'\right] \qquad (A.5)\\
= \ & \mathrm{E}\left[G^{-1}H^{-2}\sum\nolimits_{g=1}^{G}\sum\nolimits_{h=1}^{H}\sum\nolimits_{g'=1}^{G}\sum\nolimits_{h'=1}^{H}\mathbf{z}_{gh}(\boldsymbol{\theta})\mathbf{z}_{g'h'}(\boldsymbol{\theta})'\right]\\
= \ & G^{-1}H^{-2}\sum\nolimits_{g}\sum\nolimits_{h}\sum\nolimits_{h'}\mathrm{E}[\mathbf{z}_{gh}\mathbf{z}_{gh'}']\\
& +G^{-1}H^{-2}\sum\nolimits_{h}\sum\nolimits_{g}\sum\nolimits_{g'}\mathrm{E}[\mathbf{z}_{gh}\mathbf{z}_{g'h}']\\
& -G^{-1}H^{-2}\sum\nolimits_{g}\sum\nolimits_{h}\mathrm{E}[\mathbf{z}_{gh}\mathbf{z}_{gh}'],
\end{aligned}$$

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. Each of the cross-product terms $\mathbf{z}_{gh}\mathbf{z}_{gh'}' = \sum_{i\in C_{gh}}\sum_{j\in C_{gh'}}\mathbf{z}_{ghi}(\boldsymbol{\theta})\mathbf{z}_{gh'j}(\boldsymbol{\theta})$ is an $N_{gh}\times N_{gh}$ matrix. We assume that $\mathrm{E}[\mathbf{z}_{igh}(\boldsymbol{\theta})\mathbf{z}_{jgh'}(\boldsymbol{\theta})]$ is bounded away from zero and bounded from above. Then $\mathrm{E}[\mathbf{z}_{gh}\mathbf{z}_{gh'}']$ is bounded, given N_{gh} fixed, and $G^{-1}H^{-2}\sum_g\sum_h\sum_{h'}\mathrm{E}[\mathbf{z}_{gh}\mathbf{z}_{gh'}']$ 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 $\mathrm{E}[\mathbf{z}_{igh}(\boldsymbol{\theta})\mathbf{z}_{jgh'}(\boldsymbol{\theta})]$ 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 $\mathrm{E}[\mathbf{z}_{gh}\mathbf{z}'_{gh'}] = 0$ 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, however, we obtain the same asymptotic variance matrix for $\widehat{\boldsymbol{\beta}}$.

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 $N \to \infty$ with T fixed (a short panel), the rate of convergence is $\sqrt{N}$. When both $N \to \infty$ and $T \to \infty$ (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. Demirguc-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., Gelbach, J.G., and D.L. Miller (2009), "Robust Inference with Multi-way Clustering," U.C.-Davis Economics Department Working Paper No. 09-??.

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 (2008), "Regression Discontinuity Inference with Specification Error,"

Journal of Econometrics, 142(2), 655-674.

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

Martin, P., T. Mayer, and M. Thoenig (2008), "Make Trade Not War?", *Review of Economic Studies*, 75, 865–900.

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.

Mitchener, K.J., and M.D. Weidenmier (2007), "The Baring Crisis and the Great Latin American Meltdown of the 1890s," *Journal of Economic History*, 68, 462-500.

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. (2009), "Estimating Standard Errors in Finance Panel Data Sets: Comparing Approaches," *Review of Financial Studies*, 22, 435-480.

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," *Management Science*, 54, 1891 - 1903.

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*, 90(3), 237-251.

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. (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: a random effect common to each group, and a heteroscedastic component.

		Assumption about errors in construction of Variance											
Number of Group 1 Clusters	Number of Group 2 Clusters	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%
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%
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: 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
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 estimator:				
Heteroscedasticity 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_t), with e_s a state-specific autoregressive component and e_t a year-specific "spatial" autoregressive component. The outcome is also modified by an independent year-specific autoregressive component. See text for details. 2000 Monte Carlo replications

Table 3
Replication of Hersch (1998)

		Variable: Industry Injury Rate		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.