TECHNICAL WORKING PAPER SERIES

ROBUST INFERENCE WITH MULTI-WAY CLUSTERING

A. Colin Cameron
Jonah B. Gelbach
Douglas L. Miller

Technical Working Paper 327
http://www.nber.org/papers/T0327

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
September 2006

This paper has benefitted from presentations at the University of California - Berkely, the University of California - Riverside, and Dartmouth College. Miller gratefully acknowledges funding from the National Institute on Aging, through Grant Number T32-AG00186 to the NBER.

Robust Inference with Multi-way Clustering
A. Colin Cameron, Jonah B. Gelbach, and Douglas L. Miller
NBER Technical Working Paper No. 327
September 2006
JEL No. C14, C21, C52

ABSTRACT

In this paper we propose a new variance estimator for OLS as well as for nonlinear estimators such as logit, probit and GMM, that provcides 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 two studies in the empirical public/labor literature where two-way clustering is present.

A. Colin Cameron
Department of Economics
UC Davis
Davis, CA 95616
accameron@ucdavis.edu

Jonah Gelbach
Department of Economics
University of Maryland
College Park, MD 20742
and
College of Law
Florida State University
425 West Jefferson Street
Tallahassee, FL 32303
and NBER
gelbach@glue.umd.edu

Douglas Miller
Department of Economics
UC Davis
Davis, CA 95616
and NBER
dlmiller@ucdavis.edu

1. Introduction

A key component of empirical research is conducting accurate statistical inference. One challenge to this is the possibility of clustered (or non-independent) errors. In this paper we propose a new 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.[1]

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 analysis, focus on one-way clustering.

In this paper we consider inference when there is nonnested multi-way clustering. The method is useful in many applications, including:

1. Clustering due to sample design may be combined with grouping on a key regressor for reasons other than sample design. For example, clustering may occur at the level of a Primary Sampling Unit as well as at the level of an industry-level regressor.

2. The survey design of 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 (perhaps because of 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.

3. 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 spatial 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.

[1] An ado file for multi-way clustering in Stata is available at the following website: www.econ.ucdavis.edu/faculty/dlmiller/statafiles/index.htm

4. 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. Petersen (2006) compares a number of approaches for OLS estimation in this panel setting.[2]

5. Even in a cross-section study clustering may arise at several levels simultaneously. For example a model may have geographic-level regressors, industry-level regressors and occupation-level regressors.

6. Clustering may arise due to discrete regressors. Moulton (1986) considered inference in this case, 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.

Our method builds on that for one-way cluster-robust inference. Initial controls for one-way clustering relied on strong assumptions on the dgp for the error term, such as a one-way random effects error model. This has been superseded by computation of "cluster-robust" standard errors that rely on much weaker 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 White (1984) for a multivariate dependent variable, Liang and Zeger (1986) for estimation in a generalized estimating equations setting, and Arellano (1987) and Hansen (2005) for the fixed effects estimator in linear panel models. Wooldridge (2003) provides a survey, and Wooldridge (2002) and Cameron and Trivedi (2005) give textbook treatments.

For two-way or multi-way clustering that is nested, 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 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

[2] We thank Mitchell Petersen for sending us a copy of his paper. One of the methods he uses is that proposed in this paper for OLS with two-way clustering. Petersen cites as his source for this method a paper by Thompson (2005) that we were unaware of until after working out our theoretical results and doing substantial Monte Carlo work. Sometime after we described our work to Petersen, he informed us that Thompson (2006) had been posted on the internet. Thompson (2006) correctly derives the formula for OLS in the two-way case, but the theoretical discussion does not address the general multi-way case and nonlinear estimators that we also consider. Thompson's Monte Carlo results are basically consistent with ours, though they are somewhat narrower in scope.

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.

Our new 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 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. 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 two empirical examples, Hersch (1998) using OLS and Gruber and Madrian (1995) using both probit and OLS, where we contrast results obtained using conventional one-way clustering to those allowing for two-way clustering.[3] 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.

[3] We thank Marianne Bertrand, Esther Duflo, Sendhil Mullainathan, and Joni Hersch for assisting us in replicating their data sets.

2.1. One-Way Clustering

The linear model with one-way clustering is

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

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 $i \neq j$

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

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

$$\mathbf{y}_g = \mathbf{X}_g\boldsymbol{\beta} + \mathbf{u}_g, \tag{2.3}$$

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

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

Given error independence across clusters, this estimator has variance matrix (conditional on regressors)

$$\mathrm{V}[\widehat{\boldsymbol{\beta}}] = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\left(\sum_{g=1}^{G}\mathbf{X}'_g\boldsymbol{\Omega}_g\mathbf{X}_g\right)\left(\mathbf{X}'\mathbf{X}\right)^{-1}, \tag{2.5}$$

which depends on the unknown cluster error variance matrices

$$\boldsymbol{\Omega}_g = \mathrm{V}[\mathbf{u}_g|\mathbf{X}_g] = \mathrm{E}[\mathbf{u}_g\mathbf{u}'_g|\mathbf{X}_g]. \tag{2.6}$$

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 standard deviation of the residual distribution, should be inflated by $\tau_j \simeq 1 + \rho_{x_j}\rho_u(\bar{N}_g - 1)$, where ρ_{x_j} is 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.

A cluster-robust variance matrix estimate is

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

where $\widehat{\mathbf{u}}_g = \mathbf{y}_g - \mathbf{X}_g \widehat{\boldsymbol{\beta}}$. This provides a consistent estimate of the variance matrix if $G^{-1} \sum_{g=1}^{G} \mathbf{X}_g' \widehat{\mathbf{u}}_g \widehat{\mathbf{u}}_g' \mathbf{X}_g - G^{-1} \sum_{g=1}^{G} \mathrm{E}[\mathbf{X}_g' \boldsymbol{\Omega}_g \mathbf{X}_g] \xrightarrow{p} \mathbf{0}$ as $G \to \infty$. White (1984, p.134-142) presents 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), which considered the case $N_g = 1$. Note that (2.7) does not require specification of a model for $\boldsymbol{\Omega}_g$, and thus it permits quite general forms of $\boldsymbol{\Omega}_g$.

A helpful informal presentation of (2.7) is that

$$\widehat{\mathrm{V}}[\widehat{\boldsymbol{\beta}}] = (\mathbf{X}'\mathbf{X})^{-1} \widehat{\mathbf{B}} (\mathbf{X}'\mathbf{X})^{-1}, \tag{2.8}$$

where the central matrix

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

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 $\widehat{\mathbf{B}}$ in (2.9) as being given by

$$\widehat{\mathbf{B}} = \mathbf{X}'(\widehat{\mathbf{u}} \widehat{\mathbf{u}}' .* \mathbf{S}^G)\mathbf{X} \tag{2.10}$$

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 $\Delta^G \Delta^{G\prime}$ 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 $\widehat{\mathbf{B}}$ 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

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

where we assume that for $i \neq j$

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

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, $\boldsymbol{\Omega} = \mathrm{V}[\mathbf{u}|\mathbf{X}]$ 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 $\widehat{\mathbf{u}}\widehat{\mathbf{u}}'$ where the i^{th} and j^{th} observations share a cluster in one specified dimension, we keep those elements of $\widehat{\mathbf{u}}\widehat{\mathbf{u}}'$ where the i^{th} and j^{th} observations share a cluster in any dimension. Then

$$\widehat{\mathbf{B}} = \mathbf{X}'(\widehat{\mathbf{u}}\widehat{\mathbf{u}}' .* \mathbf{S}^{GH})\mathbf{X}, \tag{2.13}$$

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 $\widehat{\mathbf{B}}$ 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 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 $g \in \{1, 2, ..., G\}$, $\mathbf{S}^H$ 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

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

Substituting (2.14) into (2.7) yields

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

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

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

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

For one-way clustering small-sample modifications of (2.7) are typically used, since without modification the cluster-robust standard errors are biased downwards.[4] For example, Stata uses $\sqrt{c}\widehat{\mathbf{u}}_g$ in (2.7) 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

[4] Cameron, Gelbach, and Miller (2006) review various small-sample corrections that have been proposed in the literature, for both standard errors and for inference using resultant Wald statistics.

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.

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. Because this problem is more likely to arise when the third covariance matrix is relatively large, using the Stata-style formula for residual adjustment reduces the likelihood of estimating a negative variance. This is because it uses a smaller (inflationary) adjustment to the standard errors in the third matrix, since $I < G$ or H. Our experience suggests that this problem is infrequent, and primarily occurs when there are very few clusters and when there is actually no need to cluster in more than one dimension. Should negative variances arise in practice, a researcher may need to fall back on a more ad hoc rule of thumb such as using the maximum of the standard errors obtained from one-way clustering along each possible dimension.

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. But adapting the method to robustify feasible GLS estimators that model clustering may be difficult. For example, if a one-way random effects model that clusters over g is specified, then the resulting residuals $\widehat{u}^*_{igh}$ from the transformed model will asymptotically retain the same broad correlation pattern over g, but not over h.

2.3. 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 $\boldsymbol{\delta} : \{1, 2, ..., N\} \rightarrow \times_{d=1}^{D}\{1, 2, ..., G_d\}$ 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. We will say that $d_i \approx d_j$ if and only if $\delta_{id} = \delta_{jd}$ for some $d \in \{1, 2, ..., D\}$, where δ_{id} denotes the d^{th} element of $\boldsymbol{\delta}_i$. Thus $\mathbf{1}[i, j \text{ in same cluster for some } d] \Leftrightarrow \delta_i \approx \delta_j$.

Now let $\mathbf{r}$ be a D-vector, with d^{th} coordinate equal to r_d, and define the set $R \equiv \{\mathbf{r}\colon r_d \in \{0, 1\}, d = 1, 2, ..., D\} - \mathbf{0}$, where the subtraction of the vector $\mathbf{0}$ means that R has $2^D - 1$ elements. For example, for $D = 3$ we have $R = \{(1, 0, 0),\ (0, 1, 0),$

$(0,0,1)$, $(1,1,0)$, $(1,0,1)$, $(0,1,0)$, $(1,1,1)\}$. 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 referenced by $\mathbf{r}$. 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 $\delta_{i3} = \delta_{j3}$). Define $I(i,j) = 1$ if and only if $I_{\mathbf{r}}(i,j) = 1$ for some $\mathbf{r} \in R$.

Now define the $2^D - 1$ matrices

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

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

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

where

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

Thus we sum over all possible values of $\parallel \mathbf{r} \parallel = \sum_d r_d$. 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 $\parallel \mathbf{r} \parallel \leq D$ 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 (2.13) 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 some dimensions have few clusters. However, this need not necessarily be the case, as it depends on the nature of the clustering. For example, with cross-section data and clustering due to grouped regressors, it is quite possible in a sample of several thousand observations to have several grouped regressors each taking, say fifty distinct values, as should be clear from the application in Section 4.1.

2.4. 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

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

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, $\widehat{\boldsymbol{\theta}}$ is asymptotically normal with estimated variance matrix

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

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

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

For one-way clustering $\widehat{\mathbf{B}} = \sum_{g=1}^G \widehat{\mathbf{h}}_g\widehat{\mathbf{h}}_g'$ where $\widehat{\mathbf{h}}_g$ is the sum of $\widehat{\mathbf{h}}_i$ for those observations in cluster g. 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. This is likely to be the case for a probit model, for example, but less likely for a Tobit model, for example. In the latter case the estimated variance matrix is still as above, but the distribution of the estimator will be instead centered on a pseudo-true value.

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

$$\widehat{\mathrm{V}}[\widehat{\boldsymbol{\theta}}] = \widehat{\mathbf{A}}^{-1}\widetilde{\mathbf{B}}\widehat{\mathbf{A}}^{-1}, \tag{2.20}$$

where as usual

$$\widehat{\mathbf{A}} = \sum_i \frac{\partial \mathbf{h}_i}{\partial \boldsymbol{\theta}}\bigg|_{\widehat{\boldsymbol{\theta}}}, \tag{2.21}$$

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

$$\widetilde{\mathbf{B}} \equiv \sum_{\|\mathbf{r}\|=k,\ \mathbf{r}\in R} (-1)^{k+1}\widetilde{\mathbf{B}}_r, \tag{2.22}$$

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

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

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 $g \in \{1, 2, ..., G\}$;

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

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

Given these three components, $\widehat{\mathrm{V}}[\widehat{\boldsymbol{\beta}}]$ 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 underdispersion 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 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, $(y_1, \mathbf{X}_1), ..., (y_G, \mathbf{X}_G)$, in step 2 do a pairs cluster bootstrap that resamples with replacement from the H clusters, $(y_1, \mathbf{X}_1), ..., (y_H, \mathbf{X}_H)$, and in step 3 do a pairs cluster bootstrap that resamples with replacement using clustering on $(g, h) \in \{(1, 1), ..., (G, H)\}$. 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. (2006) for a discussion of clustering and asymptotic refinement in the one-way case).

Finally we consider GMM estimation for over-identified models. Then $\widehat{\boldsymbol{\theta}}$ minimizes

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

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

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

where $\widehat{\mathbf{A}}$ is defined in (2.21), and $\widetilde{\mathbf{B}}$ is an estimate of $\mathrm{V}[\sum_i \mathbf{h}_i]$ 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

The preceding section provides methods to obtain cluster-robust standard errors given multi-way clustering. In this section we analyze the size performance of Wald tests based on these standard errors, rather than on the standard errors per se, in two different settings for two-way clustering. We compare our Wald test to those 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

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

where $\beta_0 = \beta_1 = \beta_2 = 1$ throughout, while the regressors x_{1gh} and x_{2gh} and the errors u_{gh} vary with the experiment performed, as described below. The subscript i does not appear in (3.1) as i is redundant since we use rectangular designs with exactly one observation drawn from each (g, h) pair, leading to $G \times H$ observations. 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 $\widehat{\beta}_1$ 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

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

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 se$[\widehat{\beta}_1]$ and 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 (2.7) 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 $\mathrm{V}[u]$ 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 generally use 2,000 simulations, which yields a 95% confidence interval of (4.0%, 6.0%) for the rejection rate, given that the true rejection rate is 5%. For methods 1-3 with larger designs, specifically $G \times H > 1600$, we use only 1,000 simulations due to computational cost; the 95% confidence interval is (3.6%, 6.4%).

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 fatter 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. (2006)). 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 final 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 $G = H = 10$.

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 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 $\widehat{\beta}_1$ 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 (2.6)) 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 $(G, H) = (10, 10)$. The high over-rejection rates for inference on β_2 even exceed those when iid errors are assumed.

The 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 final two columns show considerable improvement for the two-way cluster-robust method when T critical values are instead used. 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 $(G, H) = (10, 10)$, 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.

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 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 to simulation variability as the rejection rates for $(G, H) = (90, 90)$ 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 $\beta_1$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 $(G, H) = (10, 10)$ 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 $(G, H) = (10, 10)$.

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. 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 (2005) 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 data on 1,358,623 employed women are from the 1979-1999 Current Population Surveys, with log earnings as the outcome of interest. The model estimated is

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

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 binary state-year-specific regressor, $\mathbf{x}_{ist}$ are individual characteristics, and some models may include state-specific effects δ_s and time-specific effects γ_t. Here $G = 50$ and $H = 21$ and, unlike in Section 3.1, there are many observations per (g, h) cell.[5]

Interest lies in inference on α, the coefficient of a randomly-assigned "placebo policy" dummy 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. Specifically, the placebo law for a state-year cell is generated by $d_{st} = \mathbf{1}[\varepsilon_s + 0.333\varepsilon_t > 0]$, with ε_s and ε_t iid $\mathcal{N}[0, 1]$. 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. In each of 1,000 simulations 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 $|\widehat{\alpha}|/\text{se}[\widehat{\alpha}] > 1.96$. Given the design used here, $\widehat{\alpha}$ is consistent and Table 4 will display asymptotic rejection rates of 5%, provided that the correct standard error is calculated. Because we use the actual data for the y_{ist} and $\mathbf{x}_{ist}$, the error terms u_{ist} will be correlated within states and over time to an extent typical for many repeated cross-section studies of earnings in the U.S. If this correlation is non-trivial, then testing $\widehat{\alpha} = 0$ will asymptotically reject more than 5% of the time.

The first column of Table 4 considers regression on only d_{st} (and an intercept). 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 (93%) due to failure to control for clustering. The standard one-way cluster-robust cluster methods partly mitigate this, though the rejection rates still exceed 17%. 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. The two-way cluster-robust method does best, with rejection rate of 8.9%. 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%). Thus we see the empirical relevance of the fact that the $T(H - 1)$ distribution is not exactly correct here.

We see in the second column of Table 4 that the results change little when we partial out a quartic in age and four education dummies. Evidently the correlation over states and time in the error u_{ist} is little changed by inclusion of these regressors, leading to

[5] In most of their simulations Bertand 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 (here over one million observations).

little change to inference on α.

The correlation in the error u_{ist} does change substantially with the inclusion of year fixed effects. In particular, the rejection rate with year fixed effects included as regressors and one-way clustering on state is 7.9% (which is close to the rejection rate of 9.1% that we estimate when we do not partial out year fixed effects but do use two-way clustering on state and year). Evidently for these data the correlation over states is well modelled by a time-specific fixed effect that is state-invariant. But for other types of data, such as financial data on firms over time, this need not be the case.

The results show that when we include both year and state dummies as regressors, we need to account for clustering: the rejection rate using heteroskedastic-robust standard errors is 27%. But now all cluster-robust methods (one-way and two-way) give roughly similar rejection rates. Surprisingly, clustering on state-year cell, rather than just state, works best.

In this example, the two-way cluster-robust method works comparatively well regardless of whether or not state and year fixed effects are included as regressors. But once both these effects are included the one-way clustering methods also work well.

4. Empirical examples

In the two empirical examples 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 example considers two-way clustering in a cross-section setting. The second considers a rotating panel, and considers probit estimation in addition to OLS. Since, unlike Section 3, there is no benchmark for rejection rates of Wald tests, we compare computed standard errors across various methods.

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

In a cross-section study clustering may arise at several levels simultaneously. 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.

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

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

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 (there are 211 industries) and $rocc_{ih}$ is the injury rate for occupation (there are 387 occupations). 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.

The 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. 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.

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 do 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

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

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

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, $\mathbf{x}_{ist}$ denotes individual-level controls, and state fixed effects and year fixed effects are also included. The data are on 39,063 men aged 55-64 years from the 1980-90 March CPS.

One natural dimension for clustering is the state-year group 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. 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.[6] We use White heteroskedastic standard errors, which differ little from those assuming iid errors, as the benchmark.

[6] 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

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.

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.[7]

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. At the same time, we also note in some settings the impact of the method is modest. The impact is likely to be greatest when both the regressor and the error are correlated over two dimensions. But the Hersch (1995) example 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. Moreover, 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.

exceptions. The basic probit estimates provide point estimates and (nonclustered) standard errors that are broadly similar to those reported in their paper.

[7] 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. (2006), can be adapted to two-way clustering.

A. Appendix

For two-way clustering and $G, H \to \infty$ we establish consistency and obtain the limit distribution of the OLS estimator.

The model is

$$y_{ghi} = \mathbf{x}'_{ghi}\boldsymbol{\beta} + \mathbf{u}_{ghi},$$

for the i^{th} of N_{gh} observations in cluster (g, h). The OLS estimator $\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ yields

$$\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} = \left(T^{-1} \sum_{g=1}^{G} \sum_{h=1}^{H} \sum_{i \in C_{gh}} \mathbf{x}_{ghi}\mathbf{x}'_{ghi} \right)^{-1} T^{-1} \sum_{g=1}^{G} \sum_{h=1}^{H} \sum_{i \in C_{gh}} \mathbf{x}_{ghi}u_{ghi}, \tag{A.1}$$

where $T = GH$ is the total number of clusters (not observations), C_{gh} denotes the observations in cluster (g, h), and we assume a finite number N_{gh} of observations in each cluster.

The first term in (A.1) is the average $T^{-1}\sum_{g=1}^{G}\sum_{h=1}^{H}\mathbf{Z}_{gh}$ where $\mathbf{Z}_{gh} = \sum_{i \in C_{gh}} \mathbf{x}_{ghi}\mathbf{x}'_{ghi}$ is a $K \times K$ matrix for cluster (g, h). This has finite nonzero probability limit if $\mathrm{E}[\mathbf{x}_{ghi}\mathbf{x}'_{ghi}]$ is nonzero and bounded from above and N_{gh} is finite. The second term in (A.1) can be more compactly written as

$$T^{-1} \sum_{g=1}^{G} \sum_{h=1}^{H} \mathbf{z}_{gh},$$

where $\mathbf{z}_{gh} = \sum_{i \in C_{gh}} \mathbf{x}_{ghi}u_{ghi}$ is a $K \times 1$ vector for cluster (g, h). This term has mean $\mathrm{E}\Big[T^{-1}\sum_g \sum_h \mathbf{z}_{gh}\Big] = \mathbf{0}$ if $\mathrm{E}[u_{ghi}|\mathbf{x}_{ghi}] = 0$ as then $\mathrm{E}[\mathbf{z}_{ghi}] = \mathbf{0}$, and variance

$$\begin{aligned}
&\mathrm{V}\left[T^{-1}\textstyle\sum_{g=1}^{G}\sum_{h=1}^{H}\mathbf{z}_{gh}\right] \\
&= \mathrm{E}\left[T^{-2}\textstyle\sum_{g=1}^{G}\sum_{h=1}^{H}\sum_{g'=1}^{G}\sum_{h'=1}^{H}\mathbf{z}_{gh}\mathbf{z}'_{g'h'}\right] \\
&= T^{-2}\textstyle\sum_g\sum_h\sum_{h'}\mathrm{E}[\mathbf{z}_{gh}\mathbf{z}'_{gh'}] + T^{-2}\sum_h\sum_g\sum_{g'}\mathrm{E}[\mathbf{z}_{gh}\mathbf{z}'_{g'h}] - T^{-2}\sum_g\sum_h\mathrm{E}[\mathbf{z}_{gh}\mathbf{z}'_{gh}],
\end{aligned}$$

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. Assume $\mathrm{E}[\mathbf{z}_{gh}\mathbf{z}'_{g'h'}] < \mathbf{P}$ which will be the case if $\mathrm{V}[\mathbf{x}_{ghi}u_{ghi}]$ is bounded and N_{gh} is finite. Then counting the number of terms in the first two triple sums, we have

$$\begin{aligned}
\mathrm{V}\left[T^{-1}\sum_{g=1}^{G}\sum_{h=1}^{H}\mathbf{z}_{gh}\right] &< T^{-2}(GH^2\mathbf{P} + HG^2\mathbf{P}) \\
&= \left(\frac{1}{G} + \frac{1}{H}\right)\mathbf{P} \quad \text{as } T = GH \\
&\to \mathbf{0} \quad \text{if } G \to \infty \text{ and } H \to \infty.
\end{aligned}$$

This establishes convergence in mean square to zero of $T^{-1}\sum_g \sum_h \mathbf{z}_{gh}$ and hence convergence in probability to zero of the second term in (A.1). We conclude that $\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta} \xrightarrow{p} \mathbf{0}$, so OLS is consistent.

For the limit distribution it is convenient to assume that $G \to \infty$ and $H \to \infty$ at the same rate, so that $G/H \to$ constant. We rescale $(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})$ in (A.1) by $\sqrt{G}$, so that

$$\sqrt{G}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}) = \left((GH)^{-1}\sum_{g=1}^{G}\sum_{h=1}^{H}\sum_{i\in C_{gh}} \mathbf{x}_{ghi}\mathbf{x}'_{ghi} \right)^{-1} G^{-1/2}H^{-1}\sum_{g=1}^{G}\sum_{h=1}^{H}\sum_{i\in C_{gh}} \mathbf{x}_{ghi}u_{ghi}, \tag{A.2}$$

using $T=GH$. For the second term in (A.2) we have already shown that $\mathrm{V}\left[(GH)^{-1}\sum_g\sum_h \mathbf{z}_{gh}\right]$ converges to zero at rate $1/G$, where $\mathbf{z}_{gh}=\sum_{i\in C_{gh}} \mathbf{x}_{ghi}u_{ghi}$, so $\mathrm{V}\left[G^{-1/2}H^{-1}\sum_g\sum_h \mathbf{z}_{gh}\right]$ converges to the matrix

$$\begin{aligned}
\mathbf{V}_G &= \mathrm{V}[G^{-1/2}H^{-1}\sum_g\sum_h\sum_{i\in C_{gh}} \mathbf{x}_{ghi}u_{ghi}] \qquad\qquad (A.3)\\
&= G^{-1}H^{-2}\sum_g\sum_h\sum_{h'} \mathrm{E}[\mathbf{z}_{gh}\mathbf{z}'_{gh'}]\\
&\quad + G^{-1}H^{-2}\sum_h\sum_g\sum_{g'} \mathrm{E}[\mathbf{z}_{gh}\mathbf{z}'_{g'h}]\\
&\quad - G^{-1}H^{-2}\sum_g\sum_h \mathrm{E}[\mathbf{z}_{gh}\mathbf{z}'_{gh}].
\end{aligned}$$

Then applying a central limit theorem to the second term in (A.2) we have

$$\mathbf{V}_G^{-1/2}G^{-1/2}H^{-1}\sum_g\sum_h\sum_{i\in C_{gh}} \mathbf{x}_{ghi}u_{ghi} \xrightarrow{d} \mathcal{N}[\mathbf{0},\ \mathbf{I}].$$

The first term in (A.2) has already been assumed to have finite probability limit, say $\mathbf{M}_G = \operatorname{plim} T^{-1}\mathbf{X}'\mathbf{X}$. Combining it follows that

$$\sqrt{G}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}) \xrightarrow{p} \mathcal{N}[\mathbf{0},\ \mathbf{M}_G^{-1}\mathbf{V}_G\mathbf{M}_G^{-1}]. \tag{A.4}$$

Equation (2.15) in the main text and, equivalently, equations (2.16)–(2.18) with $D=2$ provide implementation of (A.4). The quantity $(\mathbf{X}'\mathbf{X})^{-1}$ in these formulae arises due to estimation of $\mathbf{M}_G^{-1} = (\operatorname{plim} T^{-1}\mathbf{X}'\mathbf{X})^{-1}$. The three different components of $\widehat{\mathbf{B}}$ in (2.14) or of $\widetilde{\mathbf{B}}$ in (2.18) arise due to estimation of the three components of $\mathbf{V}_G$ defined in (A.3).

References

Angrist and Lavy (2002), “The Effect of High School Matriculation Awards: Evidence from Randomized Trials”, NBER Working Paper Number 9389.

Arellano, M. (1987), “Computing Robust Standard Errors for Within-Group Estimators,” *Oxford Bulletin of Economics and Statistics*, 49, 431-434.

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.

Cameron, A.C., Gelbach, J.G., and D.L. Miller (2006), “Bootstrap-Based Improvements for Inference with Clustered Errors,” Working Paper 06-21, Department of Economics, University of California - Davis.

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 #74, U.C.-Berkeley.

Conley, T.G. (1999) “GMM Estimation with Cross Sectional Dependence”

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 4594.

Gruber, J., and B. C. Madrian (1995), “Health-Insurance Availability and the Retirement Decision,” *American Economic Review*, 85, 938-948.

Hansen, Christian (2005), “Asymptotic Properties of a Robust Variance Matrix Estimator for Panel Data when T is Large,” forthcoming, *Journal of Econometrics.*

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.

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.

Petersen, M. (2006), "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.

Rogers, W.H. (1993), "Regression Standard Errors in Clustered Samples," *Stata Technical Bulletin*, 13, 19-23.

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," post.economics.harvard.edu/faculty/thompson/papers/stderrs3jul2006.pdf.

White, H. (1980), "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," *Econometrica*, 48, 817-838.

White, H. (1984), *Asymptotic Theory for Econometricians*, San Diego, Academic Press.

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

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

Note: The null hypothesis should be rejected 5% of the time. Number of monte carlo simulations is 2000, except that for methods 1-3 it is 1000 when G*H > 1600.

Table 2

Rejection probabilities for a true null hypothesis

True model: random effects on both Group1 and Group 2

		Assumption about errors in construction of Variance									
		Assume iid errors		One-way cluster robust (cluster on group1)		Two-way random effects		Two-way cluster-robust		Two-way cluster-robust, T critical values	
Number of Group 1 Clusters	Number of Group 2 Clusters										
10	10	23.4%	23.7%	13.7%	33.2%	6.2%	6.6%	17.4%	17.4%	12.6%	13.4%
20	20	34.0%	32.3%	8.6%	42.7%	5.7%	5.3%	10.3%	9.6%	8.7%	7.6%
30	30	39.7%	41.7%	7.4%	50.6%	5.2%	5.6%	8.6%	9.1%	7.8%	7.7%
40	40	47.7%	47.6%	8.7%	55.2%	6.5%	5.4%	9.0%	9.3%	7.9%	8.1%
50	50	50.0%	50.4%	6.0%	58.8%	4.9%	4.8%	7.0%	6.7%	6.3%	6.2%
60	60	54.5%	56.3%	6.4%	64.1%	5.6%	6.5%	6.7%	6.1%	5.9%	5.6%
70	70	54.2%	57.6%	5.5%	64.8%	4.8%	6.0%	6.4%	6.5%	5.9%	6.0%
80	80	61.1%	60.9%	6.5%	67.0%	4.9%	4.7%	6.3%	7.0%	5.7%	6.5%
90	90	63.4%	60.7%	5.6%	67.9%	5.0%	5.4%	6.0%	6.7%	5.8%	6.3%
100	**100**	**62.2%**	**60.4%**	**5.8%**	**67.9%**	**5.3%**	**3.6%**	**6.4%**	**5.3%**	**6.1%**	**5.1%**
10	50	49.9%	21.3%	13.3%	33.4%	8.9%	4.0%	15.0%	9.3%	10.2%	5.8%
20	50	50.8%	33.1%	9.8%	44.5%	6.7%	4.5%	9.3%	8.1%	8.2%	6.2%
10	100	63.0%	21.0%	14.1%	31.7%	10.4%	3.3%	14.2%	8.1%	9.2%	4.6%
20	100	62.9%	33.9%	10.0%	43.7%	6.2%	3.7%	9.2%	6.3%	7.7%	4.6%
50	100	63.9%	54.0%	6.6%	60.5%	5.3%	3.6%	6.9%	6.9%	6.4%	6.1%

Note: See Table 1.

Table 3

Rejection probabilities for a true null hypothesis

True model: a random effects common to each Group, and a heterscedastic component.

		Assumption about errors in construction of Variance									
		Assume iid errors		One-way cluster robust (cluster on group1)		Two-way random effects		Two-way cluster-robust		Two-way cluster-robust, T critical values	
Number of Group 1 Clusters	Number of Group 2 Clusters										
10	10	7.3%	6.6%	11.9%	8.9%	13.8%	13.2%	16.3%	15.2%	12.3%	11.8%
20	20	7.5%	7.0%	9.0%	8.1%	12.9%	12.4%	10.3%	11.3%	8.6%	8.5%
30	30	7.0%	6.7%	7.5%	7.2%	11.0%	10.7%	9.2%	9.2%	8.6%	8.1%
40	40	7.3%	7.7%	7.3%	8.7%	10.4%	11.0%	8.9%	7.8%	7.7%	7.1%
50	50	6.5%	8.0%	6.0%	8.8%	9.4%	9.9%	8.1%	7.5%	7.3%	6.7%
60	60	9.0%	7.2%	5.4%	7.3%	10.3%	8.8%	6.0%	6.4%	5.4%	5.8%
70	70	8.7%	9.5%	6.3%	10.7%	10.2%	9.8%	6.6%	6.8%	6.2%	6.5%
80	80	9.2%	10.0%	7.0%	9.5%	9.3%	9.5%	6.2%	7.4%	5.9%	7.1%
90	90	8.8%	10.9%	4.2%	10.7%	8.6%	9.8%	6.0%	5.7%	5.7%	5.4%
100	**100**	**11.3%**	**9.8%**	**6.4%**	**11.5%**	**9.0%**	**8.5%**	**6.8%**	**6.9%**	**6.3%**	**6.7%**
10	50	8.5%	7.4%	13.1%	10.4%	13.2%	14.7%	12.6%	10.6%	8.4%	7.3%
20	50	8.6%	7.0%	8.9%	7.9%	10.9%	11.8%	9.5%	7.7%	7.2%	6.1%
10	100	10.0%	5.9%	11.6%	9.1%	10.7%	12.9%	12.8%	9.0%	9.2%	5.3%
20	100	10.0%	6.3%	8.5%	7.2%	9.5%	11.3%	9.6%	7.9%	8.0%	6.1%
50	100	12.8%	8.7%	7.8%	8.4%	10.9%	11.0%	6.9%	6.9%	6.0%	6.4%

Note: See Table 1.

Table 4
Rejection probabilities for a true null hypothesis
Monte Carlos with micro (CPS) data

	RHS control variables			
	none	quartic in age, 4 education dummies	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	92.8%	90.8%	91.2%	27.0%
One-way cluster robust (cluster on state-by-year cell)	31.0%	30.5%	62.5%	6.1%
One-way cluster robust (cluster on state)	17.1%	17.1%	7.9%	9.2%
One-way cluster robust (cluster on year)	17.6%	18.6%	76.5%	8.1%
Two-way cluster-robust (cluster on state and year)	8.9%	9.1%	9.0%	10.2%

Note: Data come from 1979-1999 March CPS. Table reports rejection rates for testing a (true) null hypothesis of zero on the coefficient of fake dummy treatments. The "treatments" are generated as: 1(e_s + .333 * e_t > 0), with e_s a state-level N(0,1) and e_t a year-level N(0,1). 1000 Monte Carlo replications

Table 5

Replication of Hersch (1998)

		Variable	
		Industry Injury Rate	Occupation Injury Rate
Estimated slope coefficient:		-1.894	-0.465
Estimated standard errors:	Default (iid)	(0.415)	(0.235)
	Heteroscedastic robust	(0.397)	(0.260)
	One-way cluster on Industry	(0.643)	(0.251)
	One-way cluster on Occuptation	(0.486)	(0.363)
	Two-way clustering	(0.702)	(0.357)

Note: Replication of Hersch (1998), pg 604, Table 3, Panel B, Column 4. Standard errors in parentheses. 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	OLS
Estimated slope coefficient (* 1000):		13.264	1.644
Estimated standard errors (* 1000):	Default (iid)	(5.709)	(0.675)
	Heteroscedastic robust	(5.732)	(0.684)
	One-way cluster on state-year	(6.265)	(0.759)
	One-way cluster on household id	(5.866)	(0.702)
	One-way cluster on hhid-by-state-year	(5.732)	(0.685)
	Two-way clustering	(6.389)	(0.775)
	One-way cluster on State	(6.030)	(0.718)

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.