Robust Inference with Clustered Data

A. Colin Cameron and Douglas L. Miller
Department of Economics, University of California - Davis.

This version: Feb 10, 2010

Abstract

In this paper we survey methods to control for regression model error that is correlated within groups or clusters, but is uncorrelated across groups or clusters. Then failure to control for the clustering can lead to understatement of standard errors and overstatement of statistical significance, as emphasized most notably in empirical studies by Moulton (1990) and Bertrand, Duflo and Mullainathan (2004). We emphasize OLS estimation with statistical inference based on minimal assumptions regarding the error correlation process. Complications we consider include cluster-specific fixed effects, few clusters, multi-way clustering, more efficient feasible GLS estimation, and adaptation to nonlinear and instrumental variables estimators.

Keywords: Cluster robust, random effects, fixed effects, differences in differences, cluster bootstrap, few clusters, multi-way clusters.

JEL Classification: C12, C21, C23.

This paper is prepared for A. Ullah and D. E. Giles eds., Handbook of Empirical Economics and Finance, forthcoming 2009.

Contents

1 Introduction

In this survey we consider regression analysis when observations are grouped in clusters, with independence across clusters but correlation within clusters. We consider this in settings where estimators retain their consistency, but statistical inference based on the usual cross-section assumption of independent observations is no longer appropriate.

Statistical inference must control for clustering, 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 arises in a much wider range of settings than had been appreciated by microeconometricians. More recently Bertrand, Duflo and Mullainathan (2004) and Kézdi (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. Wooldridge (2003, 2006) provides surveys.

A common solution is to use "cluster-robust" standard errors that rely on weak assumptions – errors are independent but not identically distributed across clusters and can have quite general patterns of within-cluster correlation and heteroskedasticity – provided the number of clusters is large. This correction generalizes that of White (1980) for independent heteroskedastic errors. Additionally, more efficient estimation may be possible using alternative estimators, such as feasible GLS, that explicitly model the error correlation.

The loss of estimator precision due to clustering is presented in section 2, while cluster-robust inference is presented in section 3. The complications of inference given only a few clusters, and inference when there is clustering in more than one direction, are considered in sections 4 and 5. Section 6 presents more efficient feasible GLS estimation when structure is placed on the within-cluster error correlation. In section 7 we consider adaptation to nonlinear and instrumental variables estimators. An empirical example in section 8 illustrates many of the methods discussed in this survey.

2 Clustering and its consequences

Clustering leads to less efficient estimation than if data are independent, and default OLS standard errors need to be adjusted.

2.1 Clustered errors

The linear model with (one-way) clustering is

$$y_{ig} = \mathbf{x}'_{ig}\boldsymbol{\beta} + u_{ig}, \tag{1}$$

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 $i \neq j$

$$\mathrm{E}[u_{ig}u_{jg'}|\mathbf{x}_{ig}, \mathbf{x}_{jg'}] = 0, \text{ unless } g = g'. \tag{2}$$

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 $\mathbf{y}_g = \mathbf{X}_g\boldsymbol{\beta} + \mathbf{u}_g$, 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 $\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$. Given error independence across clusters, this estimator has asymptotic variance matrix

$$\mathrm{V}[\widehat{\boldsymbol{\beta}}] = \left(\mathrm{E}[\mathbf{X}'\mathbf{X}]\right)^{-1}\left(\sum_{g=1}^{G}\mathrm{E}[\mathbf{X}_g'\mathbf{u}_g\mathbf{u}_g'\mathbf{X}_g]\right)\left(\mathrm{E}[\mathbf{X}'\mathbf{X}]\right)^{-1}, \tag{3}$$

rather than the default OLS variance $\sigma_u^2\left(\mathrm{E}[\mathbf{X}'\mathbf{X}]\right)^{-1}$, where $\sigma_u^2 = \mathrm{V}[u_{ig}]$.

2.2 Equicorrelated errors

One way that within-cluster correlation can arise is in the random effects model where the error $u_{ig} = \alpha_g + \varepsilon_{ig}$, where α_g is a cluster-specific error or common shock that is i.i.d. $(0, \sigma_\alpha^2)$, and ε_{ig} is an idiosyncratic error that is i.i.d. $(0, \sigma_\varepsilon^2)$. Then $\mathrm{Var}[u_{ig}] = \sigma_\alpha^2 + \sigma_\varepsilon^2$ and $\mathrm{Cov}[u_{ig}, u_{jg}] = \sigma_\alpha^2$ for $i \neq j$. It follows that the intraclass correlation of the error $\rho_u = \mathrm{Cor}[u_{ig}, u_{jg}] = \sigma_\alpha^2/(\sigma_\alpha^2 + \sigma_\varepsilon^2)$. The correlation is constant across all pairs of errors in a given cluster. This correlation pattern is suitable when observations can be viewed as exchangeable, with ordering not mattering. Leading examples are individuals or households within a village or other geographic unit (such as state), individuals within a household, and students within a school.

If the primary source of clustering is due to such equicorrelated 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 error of the regression, should be inflated by

$$\tau_j \simeq 1 + \rho_{x_j}\rho_u(\bar{N}_g - 1), \tag{4}$$

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. This result for equicorrelated errors is exact if clusters are of equal size; see Kloek (1981) for the special case $\rho_{x_j} = 1$, and Scott and Holt (1982) and Greenwald (1983) for the general result. The efficiency loss, relative to independent observations, is increasing in the within-cluster correlation of both the error and the regressor and in the number of observations in each cluster.

To understand the loss of estimator precision given clustering, consider the sample mean when observations are correlated. In this case the entire sample is viewed as a single cluster. Then

$$\mathrm{V}[\bar{y}] = N^{-2}\left\{\sum\nolimits_{i=1}^{N}\mathrm{V}[u_i] + \sum\nolimits_i\sum\nolimits_{j\neq i}\mathrm{Cov}[u_i, u_j]\right\}. \tag{5}$$

Given equicorrelated errors with $\mathrm{Cov}[y_{ig}, y_{jg}] = \rho\sigma^2$ for $i \neq j$, $\mathrm{V}[\bar{y}] = N^{-2}\{N\sigma^2 + N(N-1)\rho\sigma^2\} = N^{-1}\sigma^2\{1 + \rho(N-1)\}$ compared to $N^{-1}\sigma^2$ in the i.i.d. case. At the extreme $\mathrm{V}[\bar{y}] = \sigma^2$ as $\rho \to 1$ and there is no benefit at all to increasing the sample size beyond $N = 1$.

Similar results are obtained when we generalize to several clusters of equal size (balanced clusters) with regressors that are invariant within cluster, so $y_{ig} = \mathbf{x}_g'\boldsymbol{\beta} + u_{ig}$ where i denotes the i^{th} of N individuals in the sample and g denotes the g^{th} of G clusters, and there are $N_* = N/G$ observations in each cluster. Then OLS estimation of y_{ig} on $\mathbf{x}_g$ is equivalent to OLS estimation in the model $\bar{y}_g = \mathbf{x}_g'\boldsymbol{\beta} + \bar{u}_g$, where $\bar{y}_g$ and $\bar{u}_g$ are the within-cluster averages of the dependent variable and error. If $\bar{u}_g$ is independent and homoskedastic with variance $\sigma^2_{\bar{u}_g}$ then $\mathrm{V}[\widehat{\boldsymbol{\beta}}] = \sigma^2_{\bar{u}_g}\left(\sum_{g=1}^{G}\mathbf{x}_g\mathbf{x}_g'\right)^{-1}$, where the formula for $\sigma^2_{\bar{u}_g}$ varies with the within-cluster correlation of u_{ig}. For equicorrelated errors $\sigma^2_{\bar{u}_g} = N_*^{-1}[1+\rho_u(N_*-1)]\sigma^2_u$ compared to $N_*^{-1}\sigma^2_u$ with independent errors, so the true variance of the OLS estimator is $(1+\rho_u(N_*-1))$ times the default, as given in (4) with $\rho_{x_j} = 1$.

In an influential paper Moulton (1990) pointed out that in many settings the adjustment factor τ_j can be large even if ρ_u is small. He considered a log earnings regression using March CPS data ($N = 18,946$), regressors aggregated at the state level ($G = 49$), and errors correlated within state ($\widehat{\rho}_u = 0.032$). The average group size was $18,946/49 = 387$, $\rho_{x_j} = 1$ for a state-level regressor, so $\tau_j \simeq 1 + 1 \times 0.032 \times 386 = 13.3$. The weak correlation of errors within state was still enough to lead to cluster-corrected standard errors being $\sqrt{13.3} = 3.7$ times larger than the (incorrect) default standard errors, and in this example many researchers would not appreciate the need to make this correction.

2.3 Panel Data

A second way that clustering can arise is in panel data. We assume that observations are independent across individuals in the panel, but the observations for any given individual are correlated over time. Then each individual is viewed as a cluster. The usual notation is to denote the data as y_{it} where i denotes the individual and t the time period. But in our framework (1) the data are denoted y_{ig} where i is the within-cluster subscript (for panel data the time period) and g is the cluster unit (for panel data the individual).

The assumption of equicorrelated errors is unlikely to be suitable for panel data. Instead we expect that the within-cluster (individual) correlation decreases as the time separation increases.

For example, we might consider an AR(1) model with $u_{it} = \rho u_{i,t-1} + \varepsilon_{it}$, where $0 < \rho < 1$ and ε_{it} is i.i.d. $(0, \sigma^2_\varepsilon)$. In terms of the notation in (1), $u_{ig} = \rho u_{i-1,g} + \varepsilon_{ig}$. Then the within-cluster error correlation $\mathrm{Cor}[u_{ig}, u_{jg}] = \rho^{|i-j|}$, and the consequences of clustering are less extreme than in the case of equicorrelated errors.

To see this, consider the variance of the sample mean $\bar{y}$ when $\mathrm{Cov}[y_i, y_j] = \rho^{|i-j|}\sigma^2$. Then (5) yields $\mathrm{V}[\bar{y}] = N^{-1}[1 + 2N^{-1}\sum_{s=1}^{N-1} s\rho^s]\sigma^2_u$. For example, if $\rho = 0.5$ and $N = 10$, then $\mathrm{V}[\bar{y}] = 0.260\sigma^2$ compared to $0.55\sigma^2$ for equicorrelation, using $\mathrm{V}[\bar{y}] = N^{-1}\sigma^2\{1 + \rho(N-1)\}$, and $0.1\sigma^2$ when there is no correlation ($\rho = 0.0$). More generally with several clusters of equal size and regressors invariant within cluster, OLS estimation of y_{ig} on $\mathbf{x}_g$ is equivalent to OLS estimation of $\bar{y}_g$ on $\mathbf{x}_g$, see section 2.2, and with an AR(1) error $\mathrm{V}[\widehat{\boldsymbol{\beta}}] =$

$N_*^{-1}[1+2N_*\sum_{s=1}^{N_*-1}s\rho^s]\sigma_u^2\left(\sum_g \mathbf{x}_g\mathbf{x}_g'\right)^{-1}$, less than $N_*^{-1}[1+\rho_u(N_*-1)]\sigma_u^2\left(\sum_g \mathbf{x}_g\mathbf{x}_g'\right)^{-1}$ with an equicorrelated error.

For panel data in practice, while within-cluster correlations for errors are not constant, they do not dampen as quickly as those for an AR(1) model. The variance inflation formula (4) can still provide a reasonable guide in panels that are short and have high within-cluster serial correlations of the regressor and of the error.

3 Cluster-robust inference for OLS

The most common approach in applied econometrics is to continue with OLS, and then obtain correct standard errors that correct for within-cluster correlation.

3.1 Cluster-robust inference

Cluster-robust estimates for the variance matrix of an estimate are sandwich estimates that are cluster adaptations of methods proposed originally for independent observations by White (1980) for OLS with heteroskedastic errors, and by Huber (1967) and White (1982) for the maximum likelihood estimator.

The cluster-robust estimate of the variance matrix of the OLS estimator, defined in (3), is the sandwich estimate

$$\widehat{\mathrm{V}}[\widehat{\boldsymbol{\beta}}]=(\mathbf{X}'\mathbf{X})^{-1}\widehat{\mathbf{B}}(\mathbf{X}'\mathbf{X})^{-1}, \tag{6}$$

where

$$\widehat{\mathbf{B}}=\left(\sum_{g=1}^{G}\mathbf{X}_g'\widehat{\mathbf{u}}_g\widehat{\mathbf{u}}_g'\mathbf{X}_g\right), \tag{7}$$

and $\widehat{\mathbf{u}}_g=\mathbf{y}_g-\mathbf{X}_g\widehat{\boldsymbol{\beta}}$. 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 $G\rightarrow\infty$.

The estimate of White (1980) for independent heteroskedastic errors is the special case of (7) where each cluster has only one observation (so $G=N$ and $N_g=1$ for all g). It relies on the same intuition that $G^{-1}\sum_{g=1}^{G}\mathrm{E}[\mathbf{X}_g'\mathbf{u}_g\mathbf{u}_g'\mathbf{X}_g]$ is a finite-dimensional ($K\times K$) matrix of averages that can be be consistently estimated as $G\rightarrow\infty$.

White (1984, p.134-142) presented formal theorems that justify use of (7) for OLS with a multivariate dependent variable, a result directly applicable to balanced clusters. Liang and Zeger (1986) proposed this method for estimation for a range of models much wider than OLS; see sections 6 and 7 of their paper for a range of extensions to (7). Arellano (1987) considered the fixed effects estimator in linear panel models, and Rogers (1993) popularized this method in applied econometrics by incorporating it in Stata. Note that (7) does not require specification of a model for $\mathrm{E}[\mathbf{u}_g\mathbf{u}_g']$.

Finite-sample modifications of (7) are typically used, since without modification the cluster-robust standard errors are biased downwards. Stata uses $\sqrt{c}\widehat{\mathbf{u}}_g$ in (7) rather than $\widehat{\mathbf{u}}_g$, with

$$c = \frac{G}{G-1}\frac{N-1}{N-K} \simeq \frac{G}{G-1}. \tag{8}$$

Some other packages such as SAS use $c = G/(G-1)$. This simpler correction is also used by Stata for extensions to nonlinear models. Cameron, Gelbach, and Miller (2008) review various finite-sample corrections that have been proposed in the literature, for both standard errors and for inference using resultant Wald statistics; see also section 6.

The rank of $\widehat{\mathrm{V}}[\widehat{\boldsymbol{\beta}}]$ in (7) can be shown to be at most G, so at most G restrictions on the parameters can be tested if cluster-robust standard errors are used. In particular, in models with cluster-specific effects it may not be possible to perform a test of overall significance of the regression, even though it is possible to perform tests on smaller subsets of the regressors.

3.2 Specifying the clusters

It is not always obvious how to define the clusters.

As already noted in section 2.2, Moulton (1986, 1990) pointed out for statistical inference on an aggregate-level regressor it may be necessary to cluster at that level. For example, with individual cross-sectional data and a regressor defined at the state level one should cluster at the state level if regression model errors are even very mildly correlated at the state level. In other cases the key regressor may be correlated within group, though not perfectly so, such as individuals within household. Other reasons for clustering include discrete regressors and a clustered sample design.

In some applications there can be nested levels of clustering. For example, for a household-based survey there may be error correlation for individuals within the same household, and for individuals in the same state. In that case cluster-robust standard errors are computed at the most aggregated level of clustering, in this example at the state level. Pepper (2002) provides a detailed example.

Bertrand, Duflo and Mullainathan (2004) noted that with panel data or repeated cross-section data, and regressors clustered at the state level, many researchers either failed to account for clustering or mistakenly clustered at the state-year level rather than the state level. Let y_{ist} denote the value of the dependent variable for the i^{th} individual in the s^{th} state in the t^{th} year, and let x_{st} denote a state-level policy variable that in practice will be quite highly correlated over time in a given state. The authors considered the difference-in-differences (DiD) model $y_{ist} = \gamma_s + \delta_t + \beta x_{st} + \mathbf{z}'_{ist}\boldsymbol{\gamma} + u_{it}$, though their result is relevant even for OLS regression of y_{ist} on x_{st} alone. The same point applies if data were more simply observed at only the state-year level (i.e. y_{st} rather than y_{ist}).

In general DiD models using state-year data will have high within-cluster correlation of the key policy regressor. Furthermore there may be relatively few clusters; a complication considered in section 4.

3.3 Cluster-specific fixed effects

A standard estimation method for clustered data is to additionally incorporate cluster-specific fixed effects as regressors, estimating the model

$$y_{ig} = \alpha_g + \mathbf{x}_{ig}'\boldsymbol{\beta} + u_{ig}. \tag{9}$$

This is similar to the equicorrelated error model, except that α_g is treated as a (nuisance) parameter to be estimated. Given N_g finite and $G \to \infty$ the parameters α_g, $g = 1, ..., G$, cannot be consistently estimated. The parameters $\boldsymbol{\beta}$ can still be consistently estimated, with the important caveat that the coefficients of cluster-invariant regressors (x_g rather than x_{ig}) are not identified. (In microeconometrics applications, fixed effects are typically included to enable consistent estimation of a cluster-varying regressor while controlling for a limited form of endogeneity – the regressor x_{ig} may be correlated with the cluster-invariant component α_g of the error term $\alpha_g + u_{ig}$).

Initial applications obtained default standard errors that assume u_{ig} in (9) is i.i.d. $(0, \sigma_u^2)$, assuming that cluster-specific fixed effects are sufficient to mop up any within-cluster error correlation. More recently it has become more common to control for possible within-cluster correlation of u_{ig} by using (7), as suggested by Arellano (1987). Kézdi (2004) demonstrated that cluster-robust estimates can perform well in typical-sized panels, despite the need to first estimate the fixed effects, even when N_g is large relative to G.

It is well-known that there are several alternative ways to obtain the OLS estimator of $\boldsymbol{\beta}$ in (9). Less well-known is that these different ways can lead to different cluster-robust estimates of $\mathrm{V}[\widehat{\boldsymbol{\beta}}]$. We thank Arindrajit Dube and Jason Lindo for bringing this issue to our attention.

The two main estimation methods we consider are the least squares dummy variables (LSDV) estimator, which obtains the OLS estimator from regression of y_{ig} on $\mathbf{x}_{ig}$ and a set of dummy variables for each cluster, and the mean-differenced estimator, which is the OLS estimator from regression of $(y_{ig} - \bar{y}_g)$ on $(\mathbf{x}_{ig} - \bar{\mathbf{x}}_g)$.

These two methods lead to the same cluster-robust standard errors if we apply formula (7) to the respective regressions, or if we multiply this estimate by $G/(G-1)$. Differences arise, however, if we multiply by the small-sample correction c given in (8). Let K denote the number of regressors including the intercept. Then the LSDV model views the total set of regressors to be G cluster dummies and $(K-1)$ other regressors, while the mean-differenced model considers there to be only $(K-1)$ regressors (this model is estimated without an intercept). Then

Model	Finite sample adjustment	Balanced case
LSDV	$c = \frac{G}{G-1}\frac{N-1}{N-G-(k-1)}$	$c \simeq \frac{G}{G-1} \times \frac{N_*}{N_*-1}$
Mean-differenced model	$c = \frac{G}{G-1}\frac{N-1}{N-(k-1)}$	$c \simeq \frac{G}{G-1}.$

In the balanced case $N = N_* G$, leading to the approximation given above if additionally K is small relative to N.

The difference can be very large for small N_*. Thus if $N_* = 2$ (or $N_* = 3$) then the cluster-robust variance matrix obtained using LSDV is essentially 2 times (or 3/2 times) that obtained from estimating the mean-differenced model, and it is the mean-differenced model that gives the correct finite-sample correction.

Note that if instead the error u_{ig} is assumed to be i.i.d. $(0, \sigma_u^2)$, so that default standard errors are used, then it is well-known that the appropriate small-sample correction is $(N-1)/N-G-(K-1)$, i.e. we use $s^2(\mathbf{X}'\mathbf{X})^{-1}$ where $s^2 = (N-G-(K-1))^{-1}\sum_{ig}\widehat{u}_{ig}^2$. In that case LSDV does give the correct adjustment, and estimation of the mean-differenced model will give the wrong finite-sample correction.

An alternative variance estimator after estimation of (9) is a heteroskedastic-robust estimator, which permits the error u_{ig} in (9) to be heteroskedastic but uncorrelated across both i and g. Stock and Watson (2008) show that applying the method of White (1980) after mean-differenced estimation of (9) leads, surprisingly, to inconsistent estimates of $\mathrm{V}[\widehat{\boldsymbol{\beta}}]$ if the number of observations N_g in each cluster is small (though it is correct if $N_g = 2$). The bias comes from estimating the cluster-specific means rather than being able to use the true cluster-means. They derive a bias-corrected formula for heteroskedastic-robust standard errors. Alternatively, and more simply, the cluster-robust estimator gives a consistent estimate of $\mathrm{V}[\widehat{\boldsymbol{\beta}}]$ even if the errors are only heteroskedastic, though this estimator is more variable than the bias-corrected estimator proposed by Stock and Watson.

3.4 Many observations per cluster

The preceding analysis assumes the number of observations within each cluster is fixed, while the number of clusters goes to infinity.

This assumption may not be appropriate for clustering in long panels, where the number of time periods goes to infinity. Hansen (2007a) derived asymptotic results for the standard one-way cluster-robust variance matrix estimator for panel data under various assumptions. We consider a balanced panel of N individuals over T periods, so there are NT observations in N clusters with T observations per cluster. When $N \rightarrow \infty$ with T fixed (a short panel), as we have assumed above, the rate of convergence for the OLS estimator $\widehat{\boldsymbol{\beta}}$ is $\sqrt{N}$. When both $N \rightarrow \infty$ and $T \rightarrow \infty$ (a long panel with $N_* \rightarrow \infty$), the rate of convergence of $\widehat{\boldsymbol{\beta}}$ is $\sqrt{N}$ if there is no mixing (his Theorem 2) and $\sqrt{NT}$ if there is mixing (his Theorem 3). By mixing we mean that the correlation becomes damped as observations become further apart in time.

As illustrated in section 2.3, if the within-cluster error correlation of the error diminishes as errors are further apart in time, then the data has greater informational content. This is reflected in the rate of convergence increasing from $\sqrt{N}$ (determined by the number of cross-sections) to $\sqrt{NT}$ (determined by the total size of the panel). The latter rate is the rate we expect if errors were independent within cluster.

While the rates of convergence differ in the two cases, Hansen (2007a) obtains the same asymptotic variance for the OLS estimator, so (7) remains valid.

3.5 Survey design with clustering and stratification

Clustering routinely arises in complex survey data. Rather than randomly draw individuals from the population, the survey may be restricted to a randomly-selected subset of primary sampling units (such as a geographic area) followed by selection of people within that geographic area. A common approach in microeconometrics is to control for the resultant clustering by computing cluster-robust standard errors that control for clustering at the level of the primary sampling unit, or at a more aggregated level such as state.

The survey methods literature uses methods to control for clustering that predate the references in this paper. The loss of estimator precision due to clustering is called the design effect: "The design effect or Deff is the ratio of the actual variance of a sample to the variance of a simple random sample of the same number of elements" (Kish (1965), p.258)). Kish and Frankel (1974) give the variance inflation formula (4) assuming equicorrelated errors in the non-regression case of estimation of the mean. Pfeffermann and Nathan (1981) consider the more general regression case.

The survey methods literature additionally controls for another feature of survey data – stratification. More precise statistical inference is possible after stratification. For the linear regression model, survey methods that do so are well-established and are incorporated in specialized software as well as in some broad-based packages such as Stata.

Bhattacharya (2005) provides a comprehensive treatment in a GMM framework. He finds that accounting for stratification tends to reduce estimated 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 via regression.

The current common approach of microeconometrics studies is to ignore the (beneficial) effects of stratification. In so doing there will be some over-estimation of estimator standard errors.

4 Inference with few clusters

Cluster-robust inference asymptotics are based on $G \to \infty$. Often, however, cluster-robust inference is desired but there are only a few clusters. For example, clustering may be at the regional level but there are few regions (e.g. Canada has only ten provinces). Then several different finite-sample adjustments have been proposed.

4.1 Finite-sample adjusted standard errors

Finite-sample adjustments replace $\widehat{\mathbf{u}}_g$ in (7) with a modified residual $\widetilde{\mathbf{u}}_g$. The simplest is $\widetilde{\mathbf{u}}_g = \sqrt{G/(G-1)}\widehat{\mathbf{u}}_g$, or the modification of this given in (8). Kauermann and Carroll (2001) and Bell and McCaffrey (2002) use $\widetilde{\mathbf{u}}_g^* = [\mathbf{I}_{N_g} - \mathbf{H}_{gg}]^{-1/2}\widehat{\mathbf{u}}_g$, where $\mathbf{H}_{gg} = \mathbf{X}_g(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_g'$. This transformed residual leads to $\mathrm{E}[\widehat{\mathrm{V}}[\widehat{\boldsymbol{\beta}}]] = \mathrm{V}[\widehat{\boldsymbol{\beta}}]$ in the special case that $\boldsymbol{\Omega}_g = \mathrm{E}[\mathbf{u}_g\mathbf{u}_g'] = \sigma^2\mathbf{I}$.

Bell and McCaffrey (2002) also consider use of $\widetilde{\mathbf{u}}_g^+ = \sqrt{G/(G-1)}[\mathbf{I}_{N_g} - \mathbf{H}_{gg}]^{-1}\widehat{\mathbf{u}}_g$, which can shown to equal the (clustered) jackknife estimate of the variance of the OLS estimator. These adjustments are analogs of the HC2 and HC3 measures of MacKinnon and White (1985) proposed for heteroskedastic-robust standard errors in the nonclustered case.

Angrist and Lavy (2002) found that using $\widetilde{\mathbf{u}}_g^+$ rather than $\widetilde{\mathbf{u}}_g$ increased cluster-robust standard errors by $10-50$ percent in an application with $G=30$ to 40.

Kauermann and Carroll (2001), Bell and McCaffrey (2002), Mancl and DeRouen (2001), and McCaffrey, Bell and Botts (2001) also consider the case where $\boldsymbol{\Omega}_g \neq \sigma^2\mathbf{I}$ is of known functional form, and present extension to generalized linear models.

4.2 Finite-sample Wald tests

For a two-sided test of $H_0 : \beta_j = \beta_j^0$ against $H_a : \beta_j \neq \beta_j^0$, where β_j is a scalar component of $\boldsymbol{\beta}$, the standard procedure is to use Wald test statistic $w = \left(\widehat{\beta}_j - \beta_j^0\right)/s_{\widehat{\beta}_j}$, where $s_{\widehat{\beta}_j}$ is the square root of the appropriate diagonal entry in $\widehat{\mathrm{V}}[\widehat{\boldsymbol{\beta}}]$. This "t" test statistic is asymptotically normal under H_0 as $G \rightarrow \infty$, and we reject H_0 at significance level 0.05 if $|w| > 1.960$.

With few clusters, however, the asymptotic normal distribution can provide a poor approximation, even if an unbiased variance matrix estimator is used in calculating $s_{\widehat{\beta}_j}$. The situation is a little unusual. In a pure time series or pure cross-section setting with few observations, say $N = 10$, β_j is likely to be very imprecisely estimated so that statistical inference is not worth pursuing. By contrast, in a clustered setting we may have N sufficiently large that β_j is reasonably precisely estimated, but G is so small that the asymptotic normal approximation is a very poor one.

We present two possible approaches: basing inference on the T distribution with degrees of freedom determined by the cluster, and using a cluster bootstrap with asymptotic refinement. Note that feasible GLS based on a correctly specified model of the clustering, see section 6, will not suffer from this problem.

4.3 T-distribution for inference

The simplest small-sample correction for the Wald statistic is to use a T distribution, rather than the standard normal. As we outline below in some cases the T_{G-L} distribution might be used, where L is the number of regressors that are invariant within cluster. Some packages for some commands do use the T distribution. For example, Stata uses $G-1$ degrees of freedom for t-tests and $F-$tests based on cluster-robust standard errors.

Such adjustments can make quite a difference. For example with $G = 10$ for a two-sided test at level 0.05 the critical value for T_9 is 2.262 rather than 1.960, and if $w = 1.960$ the p-value based on T_9 is 0.082 rather than 0.05. In Monte Carlo simulations by Cameron, Gelbach, and Miller (2008) this technique works reasonably well. At the minimum one should use the T distribution with $G-1$ degrees of freedom, say, rather than the standard normal.

Donald and Lang (2007) provide a rationale for using the T_{G-L} distribution. If clusters are balanced and all regressors are invariant within cluster then the OLS estimator in the model $y_{ig} = \mathbf{x}_g'\boldsymbol{\beta} + u_{ig}$ is equivalent to OLS estimation in the grouped model $\bar{y}_g = \mathbf{x}_g'\boldsymbol{\beta} + \bar{u}_g$. If $\bar{u}_g$ is i.i.d. normally distributed then the Wald statistic is T_{G-L} distributed, where $\widehat{\mathrm{V}}[\widehat{\boldsymbol{\beta}}] = s^2(X'X)^{-1}$ and $s^2 = (G-K)^{-1}\sum_g \widehat{\bar{u}_g}^2$. Note that $\bar{u}_g$ is i.i.d. normal in the random effects model if the error components are i.i.d. normal.

Donald and Lang (2007) extend this approach to additionally include regressors $\mathbf{z}_{ig}$ that vary within clusters, and allow for unbalanced clusters. They assume a random effects model with normal i.i.d. errors. Then feasible GLS estimation of $\boldsymbol{\beta}$ in the model

$$y_{ig} = \mathbf{x}_g'\boldsymbol{\beta} + \mathbf{z}_{ig}'\boldsymbol{\gamma} + \alpha_s + \varepsilon_{is}, \tag{10}$$

is equivalent to the following two-step procedure. First do OLS estimation in the model $y_{ig} = \delta_g + \mathbf{z}_{ig}'\boldsymbol{\gamma} + \varepsilon_{ig}$, where δ_g is treated as a cluster-specific fixed effect. Then do FGLS of $\bar{y}_g - \bar{\mathbf{z}}_g'\widehat{\boldsymbol{\gamma}}$ on $\mathbf{x}_g$. Donald and Lang (2007) give various conditions under which the resulting Wald statistic based on $\widehat{\beta}_j$ is T_{G-L} distributed. These conditions require that if $\mathbf{z}_{ig}$ is a regressor then $\bar{\mathbf{z}}_g$ in the limit is constant over g, unless $N_g \rightarrow \infty$. Usually $L = 2$, as the only regressors that do not vary within clusters are an intercept and a scalar regressor x_g.

Wooldridge (2006) presents an expansive exposition of the Donald and Lang approach. Additionally, Wooldridge proposes an alternative approach based on minimum distance estimation. He assumes that δ_g in $y_{ig} = \delta_g + \mathbf{z}_{ig}'\boldsymbol{\gamma} + \varepsilon_{ig}$ can be adequately explained by $\mathbf{x}_g$ and at the second step uses minimum chi-square methods to estimate $\boldsymbol{\beta}$ in $\widehat{\delta}_g = \alpha + \mathbf{x}_g'\boldsymbol{\beta}$. This provides estimates of $\boldsymbol{\beta}$ that are asymptotically normal as $N_g \rightarrow \infty$ (rather than $G \rightarrow \infty$). Wooldridge argues that this leads to less conservative statistical inference. The χ^2 statistic from the minimum distance method can be used as a test of the assumption that the δ_g do not depend in part on cluster-specific random effects. If this test fails, the researcher can then use the Donald and Lang approach, and use a T distribution for inference.

An alternate approach for correct inference with few clusters is presented by Ibragimov and Muller (2010). Their method is best suited for settings where model identification, and central limit theorems, can be applied separately to observations in each cluster. They propose separate estimation of the key parameter within each group. Each group's estimate is then a draw from a normal distribution with mean around the truth, though perhaps with separate variance for each group. The separate estimates are averaged, divided by the sample standard deviation of these estimates, and the test statistic is compared against critical values from a T distribution. This approach has the strength of offering correct inference even with few clusters. A limitation is that it requires identification using only within-group variation, so that the group estimates are independent of one another. For example, if state-year data y_{st} are used and the state is the cluster unit, then the regressors cannot use any regressor z_t such as a time dummy that varies over time but not states.

4.4 Cluster bootstrap with asymptotic refinement

A cluster bootstrap with asymptotic refinement can lead to improved finite-sample inference.

For inference based on $G \rightarrow \infty$, a two-sided Wald test of nominal size α can be shown to have true size $\alpha + O(G^{-1})$ when the usual asymptotic normal approximation is used. If instead an appropriate bootstrap with asymptotic refinement is used, the true size is $\alpha + O(G^{-3/2})$. This is closer to the desired α for large G, and hopefully also for small G. For a one-sided test or a nonsymmetric two-sided test the rates are instead, respectively, $\alpha + O(G^{-1/2})$ and $\alpha + O(G^{-1})$.

Such asymptotic refinement can be achieved by bootstrapping a statistic that is asymptotically pivotal, meaning the asymptotic distribution does not depend on any unknown parameters. For this reason the Wald t-statistic w is bootstrapped, rather than the estimator $\widehat{\beta}_j$ whose distribution depends on $\mathrm{V}[\widehat{\beta}_j]$ which needs to be estimated. The pairs cluster bootstrap procedure does B iterations where at the b^{th} iteration: (1) form G clusters $\{(\mathbf{y}_1^*, \mathbf{X}_1^*), ..., (\mathbf{y}_G^*, \mathbf{X}_G^*)\}$ by resampling with replacement G times from the original sample of clusters; (2) do OLS estimation with this resample and calculate the Wald test statistic $w_b^* = (\widehat{\beta}_{j,b}^* - \widehat{\beta}_j)/s_{\widehat{\beta}_{j,b}^*}$ where $s_{\widehat{\beta}_{j,b}^*}$ is the cluster-robust standard error of $\widehat{\beta}_{j,b}^*$, and $\widehat{\beta}_j$ is the OLS estimate of β_j from the original sample. Then reject H_0 at level α if and only if the original sample Wald statistic w is such that $w < w_{[\alpha/2]}^*$ or $w > w_{[1-\alpha/2]}^*$ where $w_{[q]}^*$ denotes the q^{th} quantile of $w_1^*, ..., w_B^*$.

Cameron, Gelbach, and Miller (2008) provide an extensive discussion of this and related bootstraps. If there are regressors which contain few values (such as dummy variables), and if there are few clusters, then it is better to use an alternative design-based bootstrap that additionally conditions on the regressors, such as a cluster Wild bootstrap. Even then bootstrap methods, unlike the method of Donald and Lang, will not be appropriate when there are very few groups, such as $G = 2$.

4.5 Few treated groups

Even when G is sufficiently large, problems arise if most of the variation in the regressor is concentrated in just a few clusters. This occurs if the key regressor is a cluster-specific binary treatment dummy and there are few treated groups.

Conley and Taber (2010) examine a differences-in-differences (DiD) model in which there are few treated groups and an increasing number of control groups. If there are group-time random effects, then the DiD model is inconsistent because the treated groups random effects are not averaged away. If the random effects are normally distributed, then the model of Donald and Lang (2007) applies and inference can use a T distribution based on the number of treated groups. If the group-time shocks are not random, then the T distribution may be a poor approximation. Conley and Taber (2010) then propose a novel method that uses the distribution of the untreated groups to perform inference on the treatment parameter.

5 Multi-way clustering

Regression model errors can be clustered in more than way. For example, they might be correlated across time within a state, and across states within a time period. When the groups are nested (for example, households within states), one clusters on the more aggregate group; see section 3.2. But when they are non-nested, traditional cluster inference can only deal with one of the dimensions.

In some applications it is possible to include sufficient regressors to eliminate error correlation in all but one dimension, and then do cluster-robust inference for that remaining dimension. A leading example is that in a state-year panel of individuals (with dependent variable y_{ist}) there may be clustering both within years and within states. If the within-year clustering is due to shocks that are the same across all individuals in a given year, then including year fixed effects as regressors will absorb within-year clustering and inference then need only control for clustering on state.

When this is not possible, the one-way cluster robust variance can be extended to multi-way clustering.

5.1 Multi-way cluster-robust inference

The cluster-robust estimate of $\mathrm{V}[\widehat{\boldsymbol{\beta}}]$ defined in (6)-(7) can be generalized to clustering in multiple dimensions. Regular one-way clustering is based on the assumption that $\mathrm{E}[u_i u_j | \mathbf{x}_i, \mathbf{x}_j] = 0$, unless observations i and j are in the same cluster. Then (7) sets $\widehat{\mathbf{B}} = \sum_{i=1}^{N} \sum_{j=1}^{N} \mathbf{x}_i \mathbf{x}_j' \widehat{u}_i \widehat{u}_j \mathbf{1}[i, j$ in same cluster], where $\widehat{u}_i = y_i - \mathbf{x}_i'\widehat{\boldsymbol{\beta}}$ and the indicator function $\mathbf{1}[A]$ equals 1 if event A occurs and 0 otherwise. In multi-way clustering, the key assumption is that $\mathrm{E}[u_i u_j | \mathbf{x}_i, \mathbf{x}_j] = 0$, unless observations i and j share any cluster dimension. Then the multi-way cluster robust estimate of $\mathrm{V}[\widehat{\boldsymbol{\beta}}]$ replaces (7) with $\widehat{\mathbf{B}} = \sum_{i=1}^{N} \sum_{j=1}^{N} \mathbf{x}_i \mathbf{x}_j' \widehat{u}_i \widehat{u}_j \mathbf{1}[i, j$ share any cluster].

For two-way clustering this robust variance estimator is easy to implement given software that computes the usual one-way cluster-robust estimate. 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. Then add the first two variance matrices and, to account for double-counting, subtract the third. Thus

$$\widehat{\mathrm{V}}_{\text{two-way}}[\widehat{\boldsymbol{\beta}}] = \widehat{\mathrm{V}}_1[\widehat{\boldsymbol{\beta}}] + \widehat{\mathrm{V}}_2[\widehat{\boldsymbol{\beta}}] - \widehat{\mathrm{V}}_{1\cap 2}[\widehat{\boldsymbol{\beta}}], \tag{11}$$

where the three component variance estimates are computed using (6)-(7) for the three different ways of clustering. Similar methods for additional dimensions, such as three-way clustering, are detailed in Cameron, Gelbach, and Miller (2010).

This method relies on asymptotics that are in the number of clusters of the dimension with the fewest number. This method is thus most appropriate when each dimension has many clusters. Theory for two-way cluster robust estimates of the variance matrix is presented in Cameron, Gelbach, and Miller (2006, 2010), Miglioretti and Heagerty (2006), and

Thompson (2006). Early empirical applications that independently proposed this method include Acemoglu and Pischke (2003), and Fafchamps and Gubert (2007).

5.2 Spatial correlation

The multi-way robust clustering estimator is closely related to the field of time-series and spatial heteroskedasticity and autocorrelation variance estimation.

In general $\widehat{\mathbf{B}}$ in (7) has the form $\sum_i \sum_j w(i,j)\, \mathbf{x}_i \mathbf{x}_j' \widehat{u}_i \widehat{u}_j$. For multi-way clustering the weight $w(i,j) = 1$ for observations who share a cluster, and $w(i,j) = 0$ otherwise. In White and Domowitz (1984), the weight $w(i,j) = 1$ for observations "close" in time to one another, and $w(i,j) = 0$ for other observations. Conley (1999) considers the case where observations have spatial locations, and has weights $w(i,j)$ decaying to 0 as the distance between observations grows.

A distinguishing feature between these papers and multi-way clustering is that White and Domowitz (1984) and Conley (1999) use mixing conditions (to ensure decay of dependence) as observations grow apart in time or distance. These conditions are not applicable to clustering due to common shocks. Instead the multi-way robust estimator relies on independence of observations that do not share any clusters in common.

There are several variations to the cluster-robust and spatial or time-series HAC estimators, some of which can be thought of as hybrids of these concepts.

The spatial estimator of Driscoll and Kraay (1998) treats each time period as a cluster, additionally allows observations in different time periods to be correlated for a finite time difference, and assumes $T \rightarrow \infty$. The Driscoll-Kraay estimator can be thought of as using weight $w(i,j) = 1 - D(i,j)/(D_{\max} + 1)$, where $D(i,j)$ is the time distance between observations i and j, and $D_{\max}$ is the maximum time separation allowed to have correlation.

An estimator proposed by Thompson (2006) allows for across-cluster (in his example firm) correlation for observations close in time in addition to within-cluster correlation at any time separation. The Thompson estimator can be thought of as using $w(i,j) = \mathbf{1}[i,j$ share a firm, or $D(i,j) \leq D_{\max}]$. It seems that other variations are likely possible.

Foote (2007) contrasts the two-way cluster-robust and these other variance matrix estimators in the context of a macroeconomics example. Petersen (2009) contrasts various methods for panel data on financial firms, where there is concern about both within firm correlation (over time) and across firm correlation due to common shocks.

6 Feasible GLS

When clustering is present and a correct model for the error correlation is specified, the feasible GLS estimator is more efficient than OLS. Furthermore, in many situations one can obtain a cluster-robust version of the standard errors for the FGLS estimator, to guard against misspecification of model for the error correlation. Many applied studies nonetheless use the OLS estimator, despite the potential expense of efficiency loss in estimation.

6.1 FGLS and cluster-robust inference

Suppose we specify a model for $\boldsymbol{\Omega}_g = \mathrm{E}[\mathbf{u}_g\mathbf{u}_g'|\mathbf{X}_g]$, such as within-cluster equicorrelation. Then the GLS estimator is $\left(\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{y}$, where $\boldsymbol{\Omega} = \mathrm{Diag}[\boldsymbol{\Omega}_g]$. Given a consistent estimate $\widehat{\boldsymbol{\Omega}}$ of $\boldsymbol{\Omega}$, the feasible GLS estimator of $\boldsymbol{\beta}$ is

$$\widehat{\boldsymbol{\beta}}_{\mathrm{FGLS}} = \left(\sum\nolimits_{g=1}^{G}\mathbf{X}_g'\widehat{\boldsymbol{\Omega}}_g^{-1}\mathbf{X}_g\right)^{-1}\sum\nolimits_{g=1}^{G}\mathbf{X}_g'\widehat{\boldsymbol{\Omega}}_g^{-1}\mathbf{y}_g. \tag{12}$$

The default estimate of the variance matrix of the FGLS estimator, $\left(\mathbf{X}'\widehat{\boldsymbol{\Omega}}^{-1}\mathbf{X}\right)^{-1}$, is correct under the restrictive assumption that $\mathrm{E}[\mathbf{u}_g\mathbf{u}_g'|\mathbf{X}_g] = \boldsymbol{\Omega}_g$.

The cluster-robust estimate of the asymptotic variance matrix of the FGLS estimator is

$$\widehat{\mathrm{V}}[\widehat{\boldsymbol{\beta}}_{\mathrm{FGLS}}] = \left(\mathbf{X}'\widehat{\boldsymbol{\Omega}}^{-1}\mathbf{X}\right)^{-1}\left(\sum\nolimits_{g=1}^{G}\mathbf{X}_g'\widehat{\boldsymbol{\Omega}}_g^{-1}\widehat{\mathbf{u}}_g\widehat{\mathbf{u}}_g'\widehat{\boldsymbol{\Omega}}_g^{-1}\mathbf{X}_g\right)\left(\mathbf{X}'\widehat{\boldsymbol{\Omega}}^{-1}\mathbf{X}\right)^{-1}, \tag{13}$$

where $\widehat{\mathbf{u}}_g = \mathbf{y}_g - \mathbf{X}_g\widehat{\boldsymbol{\beta}}_{\mathrm{FGLS}}$. This estimator requires that $\mathbf{u}_g$ and $\mathbf{u}_h$ are uncorrelated, for $g \neq h$, but permits $\mathrm{E}[\mathbf{u}_g\mathbf{u}_g'|\mathbf{X}_g] \neq \boldsymbol{\Omega}_g$. In that case the FGLS estimator is no longer guaranteed to be more efficient than the OLS estimator, but it would be a poor choice of model for $\boldsymbol{\Omega}_g$ that led to FGLS being less efficient.

Not all econometrics packages compute this cluster-robust estimate. In that case one can use a pairs cluster bootstrap (without asymptotic refinement). Specifically B times form G clusters $\{(\mathbf{y}_1^*, \mathbf{X}_1^*), ..., (\mathbf{y}_G^*, \mathbf{X}_G^*)\}$ by resampling with replacement G times from the original sample of clusters, each time compute the FGLS estimator, and then compute the variance of the B FGLS estimates $\widehat{\boldsymbol{\beta}}_1, ..., \widehat{\boldsymbol{\beta}}_B$ as $\widehat{\mathrm{V}}_{\mathrm{boot}}[\widehat{\boldsymbol{\beta}}] = (B-1)^{-1}\sum_{b=1}^{B}(\widehat{\boldsymbol{\beta}}_b - \overline{\widehat{\boldsymbol{\beta}}})(\widehat{\boldsymbol{\beta}}_b - \overline{\widehat{\boldsymbol{\beta}}})'$. Care is needed, however, if the model includes cluster-specific fixed effects; see, for example, Cameron and Trivedi (2009, p.421).

6.2 Efficiency gains of feasible GLS

Given a correct model for the within-cluster correlation of the error, such as equicorrelation, the feasible GLS estimator is more efficient than OLS. The efficiency gains of FGLS need not necessarily be great. For example, if the within-cluster correlation of all regressors is unity (so $\mathbf{x}_{ig} = \mathbf{x}_g$) and $\bar{u}_g$ defined in section 2.3 is homoskedastic, then FGLS is equivalent to OLS so there is no gain to FGLS.

For equicorrelated errors and general $\mathbf{X}$, Scott and Holt (1982) provide an upper bound to the maximum proportionate efficiency loss of OLS compared to the variance of the FGLS estimator of $1/\left[1+\frac{4(1-\rho_u)[1+(N_{\max}-1)\rho_u}{(N_{\max}\times\rho_u)^2}\right]$, $N_{\max} = \max\{N_1, ..., N_G\}$. This upper bound is increasing in the error correlation ρ_u and the maximum cluster size $N_{\max}$. For low ρ_u the maximal efficiency gain for can be low. For example, Scott and Holt (1982) note that for $\rho_u = .05$ and $N_{\max} = 20$ there is at most a 12% efficiency loss of OLS compared to FGLS. But for $\rho_u = 0.2$ and $N_{\max} = 50$ the efficiency loss could be as much as 74%, though this depends on the nature of $\mathbf{X}$.

6.3 Random effects model

The one-way random effects (RE) model is given by (1) with $u_{ig} = \alpha_g + \varepsilon_{ig}$, where α_g and ε_{ig} are i.i.d. error components; see section 2.2. Some algebra shows that the FGLS estimator in (12) can be computed by OLS estimation of $(y_{ig} - \widehat{\lambda}\bar{y}_i)$ on $(\mathbf{x}_{ig} - \widehat{\lambda}\bar{\mathbf{x}}_i)$ where $\widehat{\lambda} = 1 - \widehat{\sigma}_\varepsilon / \sqrt{\widehat{\sigma}_\varepsilon^2 + N_g \widehat{\sigma}_\alpha^2}$. Applying the cluster-robust variance matrix formula (7) for OLS in this transformed model yields (13) for the FGLS estimator.

The RE model can be extended to multi-way clustering, though FGLS estimation is then more complicated. In the two-way case, $y_{igh} = \mathbf{x}'_{igh}\boldsymbol{\beta} + \alpha_g + \delta_h + \varepsilon_{igh}$. For example, Moulton (1986) considered clustering due to grouping of regressors (schooling, age and weeks worked) in a log earnings regression. In his model he allowed for a common random shock for each year of schooling, for each year of age, and for each number of weeks worked. Davis (2002) modelled film attendance data clustered by film, theater and time. Cameron and Golotvina (2005) modelled trade between country-pairs. These multi-way papers compute the variance matrix assuming $\boldsymbol{\Omega}$ is correctly specified.

6.4 Hierarchical linear models

The one-way random effects model can be viewed as permitting the intercept to vary randomly across clusters. The hierarchical linear model (HLM) additionally permits the slope coefficients to vary. Specifically

$$y_{ig} = \mathbf{x}'_{ig}\boldsymbol{\beta}_g + u_{ig}, \tag{14}$$

where the first component of $\mathbf{x}_{ig}$ is an intercept. A concrete example is to consider data on students within schools. Then y_{ig} is an outcome measure such as test score for the i^{th} student in the g^{th} school. In a two-level model the k^{th} component of $\boldsymbol{\beta}_g$ is modelled as $\beta_{kg} = \mathbf{w}'_{kg}\gamma_k + v_{kg}$, where $\mathbf{w}_{kg}$ is a vector of school characteristics. Then stacking over all K components of $\boldsymbol{\beta}$ we have

$$\boldsymbol{\beta}_g = \mathbf{W}_g\boldsymbol{\gamma} + \mathbf{v}_j, \tag{15}$$

where $\mathbf{W}_g = \mathrm{Diag}[\mathbf{w}_{kg}]$ and usually the first component of $\mathbf{w}_{kg}$ is an intercept.

The random effects model is the special case $\boldsymbol{\beta}_g = (\beta_{1g}, \boldsymbol{\beta}_{2g})$ where $\beta_{1g} = 1 \times \gamma_1 + v_{1g}$ and $\beta_{kg} = \gamma_k + 0$ for $k > 1$, so v_{1g} is the random effects model's α_g. The HLM model additionally allows for random slopes $\boldsymbol{\beta}_{2g}$ that may or may not vary with level-two observables $\mathbf{w}_{kg}$. Further levels are possible, such as schools nested in school districts.

The HLM model can be re-expressed as a mixed linear model, since substituting (15) into (14) yields

$$y_{ig} = (\mathbf{x}'_{ig}\mathbf{W}_g)\boldsymbol{\gamma} + \mathbf{x}'_{ig}\mathbf{v}_g + u_{ig}. \tag{16}$$

The goal is to estimate the regression parameter $\boldsymbol{\gamma}$ and the variances and covariances of the errors u_{ig} and $\mathbf{v}_g$. Estimation is by maximum likelihood assuming the errors $\mathbf{v}_g$ and u_{ig} are normally distributed. Note that the pooled OLS estimator of $\boldsymbol{\gamma}$ is consistent but is less efficient.

HLM programs assume that (15) correctly specifies the within-cluster correlation. One can instead robustify the standard errors by using formulae analogous to (13), or by the cluster bootstrap.

6.5 Serially correlated errors models for panel data

If N_g is small, the clusters are balanced, and it is assumed that $\mathbf{\Omega}_g$ is the same for all g, say $\mathbf{\Omega}_g = \mathbf{\Omega}$, then the FGLS estimator in (12) can be used without need to specify a model for $\mathbf{\Omega}$. Instead we can let $\widehat{\mathbf{\Omega}}$ have ij^{th} entry $G^{-1}\sum_{g=1}^{G}\widehat{u}_{ig}\widehat{u}_{jg}$, where $\widehat{u}_{ig}$ are the residuals from initial OLS estimation.

This procedure was proposed for short panels by Kiefer (1980). It is appropriate in this context under the assumption that variances and autocovariances of the errors are constant across individuals. While this assumption is restrictive, it is less restrictive than, for example, the AR(1) error assumption given in section 2.3.

In practice two complications can arise with panel data. First, there are $T\left(T-1\right)/2$ off-diagonal elements to estimate and this number can be large relative to the number of observations NT. Second, if an individual-specific fixed effects panel model is estimated, then the fixed effects lead to an incidental parameters bias in estimating the off-diagonal covariances. This is the case for differences-in-differences models, yet FGLS estimation is desirable as it is more efficient than OLS. Hausman and Kuersteiner (2008) present fixes for both complications, including adjustment to Wald test critical values by using a higher-order Edgeworth expansion that takes account of the uncertainty in estimating the within-state covariance of the errors.

A more commonly-used model specifies an AR(p) model for the errors. This has the advantage over the preceding method of having many fewer parameters to estimate in $\mathbf{\Omega}$, though is a more restrictive model. Of course, one can robustify using (13). If fixed effects are present, however, then there is again a bias (of order N_g^{-1}) in estimation of the AR(p) coefficients due to the presence of fixed effects. Hansen (2007b) obtains bias-corrected estimates of the AR(p) coefficients and uses these in FGLS estimation.

Other models for the errors have also been proposed. For example if clusters are large, we can allow correlation parameters to vary across clusters.

7 Nonlinear and instrumental variables estimators

Relatively few econometrics papers consider extension of the complications discussed in this paper to nonlinear models; a notable exception is Wooldridge (2006).

7.1 Population-averaged models

The simplest approach to clustering in nonlinear models is to estimate the same model as would be estimated in the absence of clustering, but then base inference on cluster-robust standard errors that control for any clustering. This approach requires the assumption that the estimator remains consistent in the presence of clustering.

For commonly-used estimators that rely on correct specification of the conditional mean, such as logit, probit and Poisson, one continues to assume that $\mathrm{E}[y_{ig}|\mathbf{x}_{ig}]$ is correctly-specified. The model is estimated ignoring any clustering, but then sandwich standard errors that control for clustering are computed. This pooled approach is called a population-averaged approach because rather than introduce a cluster effect α_g and model $\mathrm{E}[y_{ig}|\mathbf{x}_{ig},\alpha_g]$, see section 7.2, we directly model $\mathrm{E}[y_{ig}|\mathbf{x}_{ig}] = \mathrm{E}_{\alpha_g}[\mathrm{E}[y_{ig}|\mathbf{x}_{ig},\alpha_g]]$ so that α_g has been averaged out.

This essentially extends pooled OLS to, for example, pooled probit. Efficiency gains analogous to feasible GLS are possible for nonlinear models if one additionally specifies a reasonable model for the within-cluster correlation.

The generalized estimating equations (GEE) approach, due to Liang and Zeger (1986), introduces within-cluster correlation into the class of generalized linear models (GLM). A conditional mean function is specified, with $\mathrm{E}[y_{ig}|\mathbf{x}_{ig}] = m(\mathbf{x}_{ig}'\boldsymbol{\beta})$, so that for the g^{th} cluster

$$\mathrm{E}[\mathbf{y}_g|\mathbf{X}_g] = \mathbf{m}_g(\boldsymbol{\beta}), \tag{17}$$

where $\mathbf{m}_g(\boldsymbol{\beta}) = [m(\mathbf{x}_{1g}'\boldsymbol{\beta}), ..., m(\mathbf{x}_{N_g g}'\boldsymbol{\beta})]'$ and $\mathbf{X}_g = [\mathbf{x}_{1g}, ..., \mathbf{x}_{N_g g}]'$. A model for the variances and covariances is also specified. First given the variance model $\mathrm{V}[y_{ig}|\mathbf{x}_{ig}] = \phi h(m(\mathbf{x}_{ig}'\boldsymbol{\beta})$ where ϕ is an additional scale parameter to estimate, we form $\mathbf{H}_g(\boldsymbol{\beta}) = \mathrm{Diag}[\phi h(m(\mathbf{x}_{ig}'\boldsymbol{\beta})]$, a diagonal matrix with the variances as entries. Second a correlation matrix $\mathbf{R}(\boldsymbol{\alpha})$ is specified with ij^{th} entry $\mathrm{Cor}[y_{ig}, y_{jg}|\mathbf{X}_g]$, where $\boldsymbol{\alpha}$ are additional parameters to estimate. Then the within-cluster covariance matrix is

$$\boldsymbol{\Omega}_g = \mathrm{V}[\mathbf{y}_g|\mathbf{X}_g] = \mathbf{H}_g(\boldsymbol{\beta})^{1/2}\mathbf{R}(\boldsymbol{\alpha})\mathbf{H}_g(\boldsymbol{\beta})^{1/2} \tag{18}$$

$\mathbf{R}(\boldsymbol{\alpha}) = \mathbf{I}$ if there is no within-cluster correlation, and $\mathbf{R}(\boldsymbol{\alpha}) = \mathbf{R}(\rho)$ has diagonal entries 1 and off diagonal entries ρ in the case of equicorrelation. The resulting GEE estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{GEE}}$ solves

$$\sum\nolimits_{g=1}^{G} \frac{\partial \mathbf{m}_g'(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \widehat{\boldsymbol{\Omega}}_g^{-1}(\mathbf{y}_g - \mathbf{m}_g(\boldsymbol{\beta})) = \mathbf{0}, \tag{19}$$

where $\widehat{\boldsymbol{\Omega}}_g$ equals $\boldsymbol{\Omega}_g$ in (18) with $\mathbf{R}(\boldsymbol{\alpha})$ replaced by $\mathbf{R}(\widehat{\boldsymbol{\alpha}})$ where $\widehat{\boldsymbol{\alpha}}$ is consistent for $\boldsymbol{\alpha}$. The cluster-robust estimate of the asymptotic variance matrix of the GEE estimator is

$$\widehat{\mathrm{V}}[\widehat{\boldsymbol{\beta}}_{\mathrm{GEE}}] = \left(\widehat{\mathbf{D}}'\widehat{\boldsymbol{\Omega}}^{-1}\widehat{\mathbf{D}}\right)^{-1} \left(\sum\nolimits_{g=1}^{G} \mathbf{D}_g'\widehat{\boldsymbol{\Omega}}_g^{-1}\widehat{\mathbf{u}}_g\widehat{\mathbf{u}}_g'\widehat{\boldsymbol{\Omega}}_g^{-1}\mathbf{D}_g\right) \left(\mathbf{D}'\widehat{\boldsymbol{\Omega}}^{-1}\mathbf{D}\right)^{-1}, \tag{20}$$

where $\widehat{\mathbf{D}}_g = \partial \mathbf{m}_g'(\boldsymbol{\beta})/\partial \boldsymbol{\beta}|_{\widehat{\boldsymbol{\beta}}}$, $\widehat{\mathbf{D}} = [\widehat{\mathbf{D}}_1, ..., \widehat{\mathbf{D}}_G]'$, $\widehat{\mathbf{u}}_g = \mathbf{y}_g - \mathbf{m}_g(\widehat{\boldsymbol{\beta}})$, and now $\widehat{\boldsymbol{\Omega}}_g = \mathbf{H}_g(\widehat{\boldsymbol{\beta}})^{1/2}\mathbf{R}(\widehat{\boldsymbol{\alpha}})\mathbf{H}_g(\widehat{\boldsymbol{\beta}})^{1/2}$. The asymptotic theory requires that $G \rightarrow \infty$.

The result (20) is a direct analog of the cluster-robust estimate of the variance matrix for FGLS. Consistency of the GEE estimator requires that (17) holds, i.e. correct specification of the conditional mean (even in the presence of clustering). The variance matrix defined in

(18) permits heteroskedasticity and correlation. It is called a "working" variance matrix as subsequent inference based on (20) is robust to misspecification of (18). If (18) is assumed to be correctly specified then the asymptotic variance matrix is more simply $(\widehat{\mathbf{D}}'\widehat{\mathbf{\Omega}}^{-1}\widehat{\mathbf{D}})^{-1}$.

For likelihood-based models outside the GLM class, a common procedure is to perform ML estimation under the assumption of independence over i and g, and then obtain cluster-robust standard errors that control for within-cluster correlation. Let $f(y_{ig}|\mathbf{x}_{ig},\boldsymbol{\theta})$ denote the density, $\mathbf{s}_{ig}(\boldsymbol{\theta}) = \partial \ln f(y_{ig}|\mathbf{x}_{ig},\boldsymbol{\theta})/\partial\boldsymbol{\theta}$, and $\mathbf{s}_g(\boldsymbol{\theta}) = \sum_i \mathbf{s}_{ig}(\boldsymbol{\theta})$. Then the MLE of $\boldsymbol{\theta}$ solves $\sum_g \sum_i \mathbf{s}_{ig}(\boldsymbol{\theta}) = \sum_g \mathbf{s}_g(\boldsymbol{\theta}) = \mathbf{0}$. A cluster-robust estimate of the variance matrix is

$$\widehat{\mathrm{V}}[\widehat{\boldsymbol{\beta}}_{\mathrm{ML}}] = \left(\sum\nolimits_g \partial\mathbf{s}_g(\boldsymbol{\theta})/\partial\boldsymbol{\theta}'|_{\widehat{\boldsymbol{\theta}}}\right)^{-1} \left(\sum\nolimits_g \mathbf{s}_g(\widehat{\boldsymbol{\theta}})\mathbf{s}_g(\widehat{\boldsymbol{\theta}})'\right) \left(\sum\nolimits_g \partial\mathbf{s}_g(\boldsymbol{\theta})/\partial\boldsymbol{\theta}'|_{\widehat{\boldsymbol{\theta}}}\right)^{-1}. \qquad (21)$$

This method generally requires that $f(y_{ig}|\mathbf{x}_{ig},\boldsymbol{\theta})$ is correctly specified even in the presence of clustering.

In the case of a (mis)specified density that is in the linear exponential family, as in GLM estimation, the MLE retains its consistency under the weaker assumption that the conditional mean $\mathrm{E}[y_{ig}|\mathbf{x}_{ig},\boldsymbol{\theta}]$ is correctly specified. In that case the GEE estimator defined in (19) additionally permits incorporation of a model for the correlation induced by the clustering.

7.2 Cluster-specific effects models

An alternative approach to controlling for clustering is to introduce a group-specific effect.

For conditional mean models the population-averaged assumption that $\mathrm{E}[y_{ig}|\mathbf{x}_{ig}] = m(\mathbf{x}_{ig}'\boldsymbol{\beta})$ is replaced by

$$\mathrm{E}[y_{ig}|\mathbf{x}_{ig},\alpha_g] = g(\mathbf{x}_{ig}'\boldsymbol{\beta} + \alpha_g), \qquad (22)$$

where α_g is not observed. The presence of α_g will induce correlation between y_{ig} and y_{jg}, $i \neq j$. Similarly, for parametric models the density specified for a single observation is $f(y_{ig}|\mathbf{x}_{ig},\boldsymbol{\beta},\alpha_g)$ rather than the population-averaged $f(y_{ig}|\mathbf{x}_{ig},\boldsymbol{\beta})$.

In a fixed effects model the α_g are parameters to be estimated. If asymptotics are that N_g is fixed while $G \rightarrow \infty$ then there is an incidental parameters problem, as there are N_g parameters $\alpha_1, ..., \alpha_G$ to estimate and $G \rightarrow \infty$. In general this contaminates estimation of $\boldsymbol{\beta}$ so that $\widehat{\boldsymbol{\beta}}$ is a inconsistent. Notable exceptions where it is still possible to consistently estimate $\boldsymbol{\beta}$ are the linear regression model, the logit model, the Poisson model, and a nonlinear regression model with additive error (so (22) is replaced by $\mathrm{E}[y_{ig}|\mathbf{x}_{ig},\alpha_g] = g(\mathbf{x}_{ig}'\boldsymbol{\beta}) + \alpha_g$). For these models, aside from the logit, one can additionally compute cluster-robust standard errors after fixed effects estimation.

We focus on the more commonly-used random effects model that specifies α_g to have density $h(\alpha_g|\boldsymbol{\eta})$ and consider estimation of likelihood-based models. Conditional on α_g, the joint density for the g^{th} cluster is $f(y_{1g}, ..., |\mathbf{x}_{N_g g},\boldsymbol{\beta},\alpha_g) = \prod_{i=1}^{N_g} f(y_{ig}|\mathbf{x}_{ig},\boldsymbol{\beta},\alpha_g)$. We then integrate out α_g to obtain the likelihood function

$$L(\boldsymbol{\beta},\boldsymbol{\eta}|\mathbf{y},\mathbf{X}) = \prod\nolimits_{g=1}^{G} \left\{ \int \left(\prod\nolimits_{i=1}^{N_g} f(y_{ig}|\mathbf{x}_{ig},\boldsymbol{\beta},\alpha_g) \right) dh(\alpha_g|\boldsymbol{\eta}) \right\}. \qquad (23)$$

In some special nonlinear models, such as a Poisson model with α_g being gamma distributed, it is possible to obtain a closed-form solution for the integral. More generally this is not the case, but numerical methods work well as (23) is just a one-dimensional integral. The usual assumption is that α_g is distributed as $\mathcal{N}[0, \sigma_\alpha^2]$. The MLE is very fragile and failure of any assumption in a nonlinear model leads to inconsistent estimation of $\boldsymbol{\beta}$.

The population-averaged and random effects models differ for nonlinear models, so that $\boldsymbol{\beta}$ is not comparable across the models. But the resulting average marginal effects, that integrate out α_g in the case of a random effects model, may be similar. A leading example is the probit model. Then $\mathrm{E}[y_{ig}|\mathbf{x}_{ig}, \alpha_g] = \Phi(\mathbf{x}_{ig}'\boldsymbol{\beta} + \alpha_g)$, where $\Phi(\cdot)$ is the standard normal c.d.f. Letting $f(\alpha_g)$ denote the $\mathcal{N}[0, \sigma_\alpha^2]$ density for α_g, we obtain $\mathrm{E}[y_{ig}|\mathbf{x}_{ig}] = \int \Phi(\mathbf{x}_{ig}'\boldsymbol{\beta} + \alpha_g) f(\alpha_g) d\alpha_g = \Phi(\mathbf{x}_{ig}'\boldsymbol{\beta}/\sqrt{1+\sigma_\alpha^2})$; see Wooldridge (2002, p.470). This differs from $\mathrm{E}[y_{ig}|\mathbf{x}_{ig}] = \Phi(\mathbf{x}_{ig}'\boldsymbol{\beta})$ for the pooled or population-averaged probit model. The difference is the scale factor $\sqrt{1+\sigma_\alpha^2}$. However, the marginal effects are similarly rescaled, since $\partial \Pr[y_{ig} = 1|\mathbf{x}_{ig}]/\partial \mathbf{x}_{ig} = \phi(\mathbf{x}_{ig}'\boldsymbol{\beta}/\sqrt{1+\sigma_\alpha^2}) \times \boldsymbol{\beta}/\sqrt{1+\sigma_\alpha^2}$, so in this case PA probit and random effects probit will yield similar estimates of the average marginal effects; see Wooldridge (2002, 2006).

7.3 Instrumental variables

The cluster-robust formula is easily adapted to instrumental variables estimation. It is assumed that there exist instruments $\mathbf{z}_{ig}$ such that $u_{ig} = y_{ig} - \mathbf{x}_{ig}'\boldsymbol{\beta}$ satisfies $\mathrm{E}[u_{ig}|\mathbf{z}_{ig}] = 0$. If there is within-cluster correlation we assume that this condition still holds, but now $\mathrm{Cov}[u_{ig}, u_{jg}|\mathbf{z}_{ig}, \mathbf{z}_{jg}] \neq 0$.

Shore-Sheppard (1996) examines the impact of equicorrelated instruments and group-specific shocks to the errors. Her model is similar to that of Moulton, applied to an IV setting. She shows that IV estimation that does not model the correlation will understate the standard errors, and proposes either cluster-robust standard errors or FGLS.

Hoxby and Paserman (1998) examine the validity of overidentification (OID) tests with equicorrelated instruments. They show that not accounting for within-group correlation can lead to mistaken OID tests, and they give a cluster-robust OID test statistic. This is the GMM criterion function with a weighting matrix based on cluster summation.

A recent series of developments in applied econometrics deals with the complication of weak instruments that lead to poor finite-sample performance of inference based on asymptotic theory, even when sample sizes are quite large; see for example the survey by Andrews and Stock (2007), and Cameron and Trivedi (2005, 2009). The literature considers only the non-clustered case, but the problem is clearly relevant also for cluster-robust inference. Most papers consider only i.i.d. case errors. An exception is Chernozhukov and Hansen (2008) who suggest a method based on testing the significance of the instruments in the reduced form that is heteroskedastic-robust. Their tests are directly amenable to adjustments that allow for clustering; see Finlay and Magnusson (2009).

7.4 GMM

Finally we consider generalized methods of moments (GMM) estimation.

Suppose that we combine moment conditions for the g^{th} cluster, so $\mathrm{E}[\mathbf{h}_g(\mathbf{w}_g, \boldsymbol{\theta})] = \mathbf{0}$ where $\mathbf{w}_g$ denotes all variables in the cluster. Then the GMM estimator $\widehat{\boldsymbol{\theta}}_{\mathrm{GMM}}$ with weighting matrix $\mathbf{W}$ minimizes $\left(\sum_g \mathbf{h}_g\right)' \mathbf{W} \left(\sum_g \mathbf{h}_g\right)$, where $\mathbf{h}_g = \mathbf{h}_g(\mathbf{w}_g, \boldsymbol{\theta})$. Using standard results in, for example, Cameron and Trivedi (2005, p.175) or Wooldridge (2002, p.423), the variance matrix estimate is

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

where $\widehat{\mathbf{A}} = \sum_g \partial \mathbf{h}_g / \partial \boldsymbol{\theta}'|_{\widehat{\boldsymbol{\theta}}}$ and a cluster-robust variance matrix estimate uses $\widehat{\mathbf{B}} = \sum_g \widehat{\mathbf{h}}_g \widehat{\mathbf{h}}_g'$. This assumes independence across clusters and $G \rightarrow \infty$. Bhattacharya (2005) considers stratification in addition to clustering for the GMM estimator.

Again a key assumption is that the estimator remains consistent even in the presence for clustering. For GMM this means that we need to assume that the moment condition holds true even when there is within-cluster correlation. The reasonableness of this assumption will vary with the particular model and application at hand.

8 Empirical Example

To illustrate some empirical issues related to clustering, we present an application based on a simplified version of the model in Hersch (1998), who examined the relationship between wages and job injury rates. We thank Joni Hersch for sharing her data with us. Job injury rates are observed only at occupation levels and industry levels, inducing clustering at these levels. In this application we have individual-level data from the Current Population Survey on 5,960 male workers working in 362 occupations and 211 industries. For most of our analysis we focus on the occupation injury rate coefficient.

In column 1 of Table 1, we present results from linear regression of log wages on occupation and industry injury rates, potential experience and its square, years of schooling, and indicator variables for union, nonwhite, and 3 regions. The first three rows show that standard errors of the OLS estimate increase as we move from default (row 1) to White heteroskedastic-robust (row 2) to cluster-robust with clustering on occupation (row 3). A priori heteroskedastic-robust standard errors may be larger or smaller than the default. The clustered standard errors are expected to be larger. Using formula (4) yields inflation factor $\sqrt{1 + 1 \times 0.207 \times (5960/362 - 1)} = 2.05$, as the within-cluster correlation of model residuals is 0.207, compared to an actual inflation of $0.516/0.188 = 2.74$.

Column 2 of Table 1 illustrates analysis with few clusters, when analysis is restricted to the 1,594 individuals who work in the ten most common occupations in the dataset. From rows 1-3 the standard errors increase, due to fewer observations, and the variance inflation factor is larger due to a larger average group size, as suggested by formula (4). Our concern is that with $G = 10$ the usual asymptotic theory requires some adjustment. The Wald two-sided test statistic for a zero coefficient on occupation injury rate is $-2.751/0.994 = 2.77$. Rows 4-6 of column 2 report the associated p-value computed in three ways. First, $p = 0.006$ using standard normal critical values (or the T with $N - K = 1584$ degrees of freedom). Second, $p = 0.022$ using a T-distribution based on $G - 1 = 9$ degrees of freedom. Third, when we perform a pairs cluster percentile-T bootstrap, the p-value increases to 0.110. These changes illustrate the importance of adjusting for few clusters in conducting inference. The large increase in p-value with the bootstrap may in part be because the first two p-values are based on cluster-robust standard errors with finite-sample bias; see section 4.1.This may also explain why the RE model standard errors in rows 8-10 of column 2 exceed the OLS cluster-robust standard error in row 3 of column 2.

We next consider multi-way clustering. Since both occupation-level and industry-level regressors are included we should compute two-way cluster-robust standard errors. Comparing row 7 of column 1 to row 3, the standard error of the occupation injury rate coefficient changes little from 0.516 to 0.515. But there is a big impact for the coefficient of the industry injury rate. In results not reported in the table, the standard error of the industry injury rate coefficient increases from 0.563 when we cluster on only occupation to 1.015 when we cluster on both occupation and industry.

If the clustering within occupations is due to common occupation-specific shocks, then a random effects (RE) model may provide more efficient parameter estimates. From row 8 of column 1 the default RE standard error is 0.308, but if we cluster on occupation this increases to 0.536 (row 10). For these data there is apparently no gain compared to OLS (see row 3).

Finally we consider a nonlinear example, probit regression with the same data and regressors, except the dependent variable is now a binary outcome equal to one if the hourly wage exceeds twelve dollars. The results given in column 3 are qualitatively similar to those in column 1. Cluster-robust standard errors are 2-3 times larger, and two-way cluster robust are slightly larger still. The parameters $\boldsymbol{\beta}$ of the random effects probit model are rescalings of those of the standard probit model, as explained in section 7.2. The rescaled coefficient is -5.119, as $\widehat{\alpha}_g$ has estimated variance 0.279. This is smaller than the probit coefficient, though this difference may just reflect noise in estimation.

9 Conclusion

Cluster-robust inference is possible in a wide range of settings. The basic methods were proposed in the 1980's, but are still not yet fully incorporated into applied econometrics, especially for estimators other than OLS. Useful references on cluster-robust inference for the practitioner include the surveys by Wooldridge (2003, 2006), the texts by Wooldridge (2002) and Cameron and Trivedi (2005) and, for implementation in Stata, Nichols and Schaffer (2007) and Cameron and Trivedi (2009).

10 References

Acemoglu, D., and J.-S. Pischke (2003), "Minimum Wages and On-the-job Training," *Research in Labor Economics*, 22, 159-202.

Andrews, D.W.K., and J.H. Stock (2007), "Inference with Weak Instruments," in R. Blundell, W.K. Newey, and T. Persson, eds., *Advances in Economics and Econometrics, Theory and Applications: Ninth World Congress of the Econometric Society*, Vol. III, Ch.3, Cambridge, Cambridge University Press.

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.

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.

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 (2006), "Robust Inference with Multi-Way Clustering," NBER Technical Working Paper 0327.

Cameron, A.C., Gelbach, J.G., and D.L. Miller (2010), "Robust Inference with Multi-Way Clustering," *Journal of Business and Economic Statistics*, forthcoming.

Cameron, A.C., Gelbach, J.G., and D.L. Miller (2008), "Bootstrap-Based Improvements for Inference with Clustered Errors," *Review of Economics and Statistics*, 90, 414-427.

Cameron, A.C., and N. Golotvina (2005), "Estimation of Country-Pair Data Models Controlling for Clustered Errors: with International Trade Applications," U.C.-Davis Economics Department Working Paper No. 06-13.

Cameron, A.C., and P.K. Trivedi (2005), *Microeconometrics: Methods and Applications,* Cambridge, Cambridge University Press.

Cameron, A.C., and P.K. Trivedi (2009), *Microeconometrics using Stata,* College Station, TX, Stata Press.

Chernozhukov, V., and C. Hansen (2008), "The Reduced Form: A Simple Approach to Inference with Weak Instruments," *Economics Letters*, 100, Pages 68-71.

Conley, T.G. (1999), "GMM Estimation with Cross Sectional Dependence," *Journal of Econometrics*, 92, 1-45.

Conley, T.G., and C. Taber (2010), "Inference with 'Difference in Differences' with a Small Number of Policy Changes," *Review of Economics and Statistics*, forthcoming.

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.

Driscoll, J.C. and A.C. Kraay (1998), "Consistent Covariance Matrix Estimation with Spatially Dependent Panel Data," *The Review of Economics and Statistics*, 80(4), 549-560.

Fafchamps, M., and F. Gubert (2007), "The Formation of Risk Sharing Networks," *Journal of Development Economics*, 83, 326-350.

Finlay, K. and L.M. Magnusson (2009), "Implementing Weak Instrument Robust Tests for a General Class of Instrumental-Variables Models," *Stata Journal*, 9, 398-421.

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.

Greenwald, B.C. (1983), "A General Analysis of Bias in the Estimated Standard Errors of Least Squares Coefficients," *Journal of Econometrics*, 22, 323-338.

Hansen, C. (2007a), "Asymptotic Properties of a Robust Variance Matrix Estimator for Panel Data when T is Large," *Journal of Econometrics*, 141, 597-620.

Hansen, C. (2007b), "Generalized Least Squares Inference in Panel and Multi-level Models with Serial Correlation and Fixed Effects," *Journal of Econometrics*, 141, 597-620.

Hausman, J. and G. Kuersteiner (2008), "Difference in Difference Meets Generalized Least Squares: Higher Order Properties of Hypotheses Tests," *Journal of Econometrics*, 144, 371-391.

Hersch, J. (1998), "Compensating Wage Differentials for Gender-Specific Job Injury Rates," *American Economic Review*, 88, 598-607.

Hoxby, C. and M.D. Paserman (1998), "Overidentification Tests with Group Data," NBER Technical Working Paper 0223.

Huber, P.J. (1967), "The Behavior of Maximum Likelihood Estimates under Nonstandard Conditions," in *Proceedings of the Fifth Berkeley Symposium*, J. Neyman (Ed.), 1, 221-233, Berkeley, CA, University of California Press.

Huber, P.J. (1981), *Robust Statistics*, New York, John Wiley.

Ibragimov, R. and U.K. Muller (2010), "T-Statistic Based Correlation and Heterogeneity Robust Inference," *Journal of Business and Economic Statistics*, forthcoming.

Kauermann, G. and R.J. Carroll (2001), "A Note on the Efficiency of Sandwich Covariance Matrix Estimation," *Journal of the American Statistical Association*, 96, 1387-1396.

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.

Kiefer, N.M. (1980), "Estimation of fixed effect models for time series of cross-sections with arbitrary intertemporal covariance," *Journal of Econometrics*, 214, 195-202.

Kish, L. (1965), *Survey Sampling*, New York, John Wiley.

Kish, L., and Frankel (1974), "Inference from Complex Surveys with Discussion", *Journal of the Royal Statistical Society*, Series B, 36, 1-37.

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.

MacKinnon, J.G., and H. White (1985), "Some Heteroskedasticity-Consistent Covariance Matrix Estimators with Improved Finite Sample Properties," *Journal of Econometrics*, 29, 305-325.

Mancl, L.A. and T.A. DeRouen, "A Covariance Estimator for GEE with Improved Finite-Sample Properties," *Biometrics*, 57, 126-134.

McCaffrey, D.F., Bell, R.M., and C.H. Botts (2001), "Generalizations of bias Reduced Linearization," *Proceedings of the Survey Research Methods Sectio*n, American Statistical Association.

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.

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.

Nichols, A., and M.E. Schaffer (2007), "Clustered Standard Errors in Stata," United Kingdom Stata Users' Group Meetings, July 2007.

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.

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

Pfeffermann, D., and G. Nathan (1981), "Regression analysis of data from a cluster sample," *Journal of the American Statistical Association*, 76, 681-689.

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.

Shore-Sheppard, L. (1996), "The Precision of Instrumental Variables Estimates with Grouped Data," Princeton University Industrial Relations Section Working Paper 374.

Stock, J.H. and M.W. Watson (2008), "Heteroskedasticity-robust Standard Errors for Fixed Effects Panel Data Regression," *Econometrica*, 76, 155-174.

Thompson, S. (2006), "Simple Formulas for Standard Errors that Cluster by Both Firm and Time," 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.

Wooldridge, J.M. (2006), "Cluster-Sample Methods in Applied Econometrics: An Extended Analysis," Department of Economics, Michigan State University.

Table 1 - Occupation injury rate and Log Wages

Impacts of varying ways of dealing with clustering

		1	2	3
			10 Largest	
		Main Sample	Occupations	Main Sample
		Linear	Linear	Probit
	OLS (or Probit) coefficient on Occupation Injury Rate	-2.158	-2.751	-6.978
1	Default (iid) std. error	0.188	0.308	0.626
2	White-robust std. error	0.243	0.320	1.008
3	Cluster-robust std. error (Clustering on Occupation)	0.516	0.994	1.454
4	P-value based on (3) and Standard Normal		0.006	
5	P-value based on (3) and T(10-1)		0.022	
6	P-value based on Percentile-T Pairs Bootstrap (999 replications)		0.110	
7	Two-way (Occupation and Industry) robust std. error	0.515		1.516
	Random effects Coefficient on Occupation Injury Rate	-1.652	-2.669	-5.789
8	Default std. error	0.357	1.429	1.106
9	White-robust std. error	0.579	2.058	
10	Cluster-robust std. error (Clustering on Occupation)	0.536	2.148	
	Number of observations (N)	5960	1594	5960
	Number of Clusters (G)	362	10	362
	Within-Cluster correlation of errors (rho)	0.207	0.211	

Notes: Coefficients and standard errors multiplied by 100. Regression covariates include Occupation Injurty rate, Industry Injury rate, Potential experience, Potential experience squared, Years of schooling, and indicator variables for union, nonwhite, and three regions. Data from Current Population Survey, as described in Hersch (1998). Std. errs. in rows 9 and 10 are from bootstraps with 400 replications. Probit outcome is wages >= $12/hour.