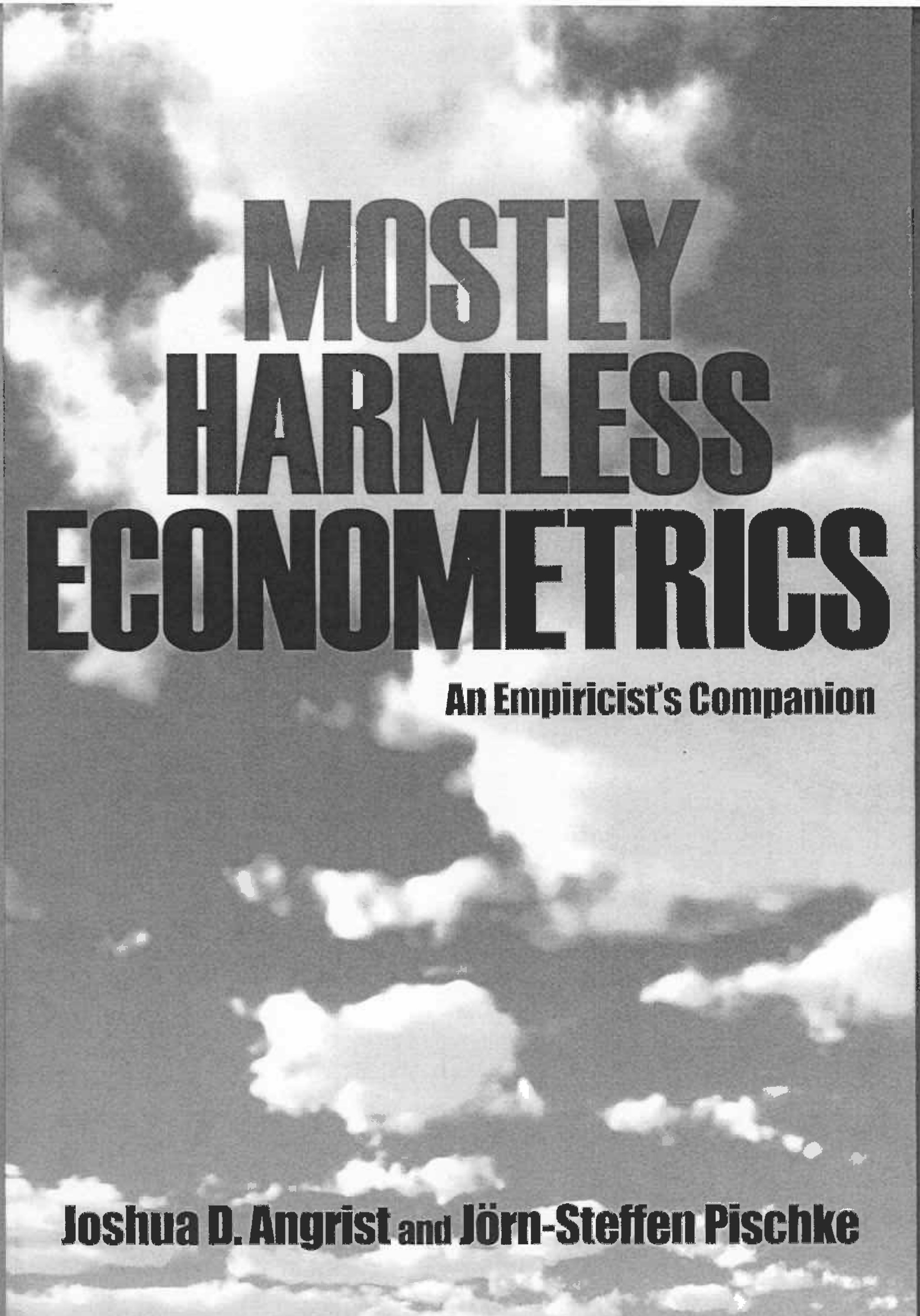
EXHIBIT 11



MOSTLY HARMLESS ECONOMETRICS

An Empiricist's Companion

Joshua D. Angrist and Jörn-Steffen Pischke

Mostly Harmless Econometrics

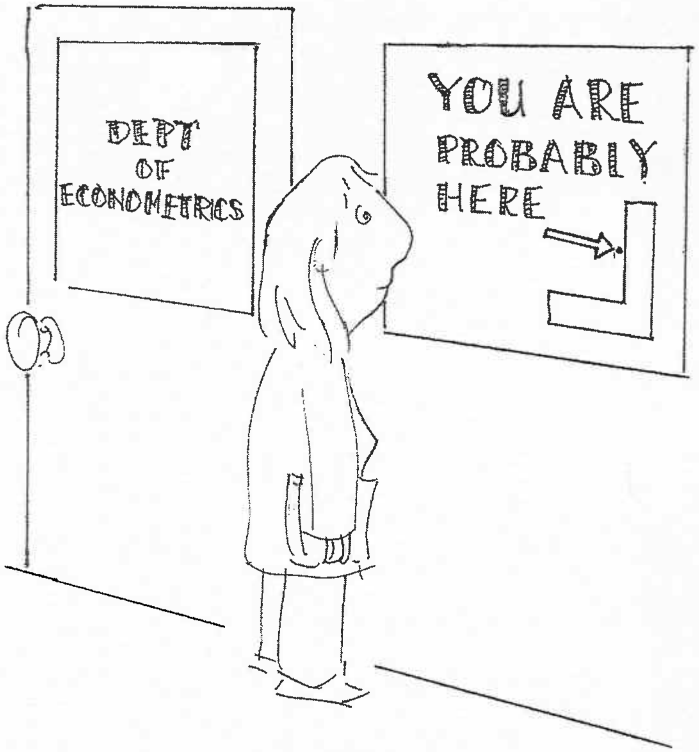
An Empiricist's Companion

Joshua D. Angrist

and

Jörn-Steffen Pischke

PRINCETON UNIVERSITY PRESS ■ PRINCETON AND OXFORD



Chapter 8

Nonstandard Standard Error Issues

> We have normality. I repeat, we have normality. Anything you still can't cope with is therefore your own problem.
>
> Douglas Adams, *The Hitchhiker's Guide to the Galaxy*

Today, software packages routinely compute asymptotic standard errors derived under weak assumptions about the sampling process or underlying model. For example, you get regression standard errors based on formula (3.1.7) using the Stata option `robust`. Robust standard errors improve on old-fashioned standard errors because the resulting inferences are asymptotically valid when the regression residuals are heteroskedastic, as they almost certainly are when regression approximates a nonlinear conditional expectation function (CEF). In contrast, old-fashioned standard errors are derived assuming homoskedasticity. The hangup here is that estimates of robust standard errors can be misleading when the asymptotic approximation that justifies these estimates is not very good. The first part of this chapter looks at the failure of asymptotic inference with robust standard error estimates and some simple palliatives.

A pillar of traditional cross section inference—and the discussion in section 3.1.3—is the assumption that the data are independent. Each observation is treated as a random draw from the same population, uncorrelated with the observation before or after. We understand today that this sampling model is unrealistic and potentially even foolhardy. Much as in the time series studies common in macroeconomics, cross section analysts must worry about correlation between observations. The most important form of dependence arises in data with a group structure—for example, the test scores of children observed within classes or schools. Children in the same school or class tend to have test scores that are correlated, since they are subject to some of the same environmental and family background influences. We call this correlation the clustering problem, or the Moulton problem, after Moulton (1986), who made it famous. A closely related problem is correlation over time in the data sets commonly used to implement differences-in-differences (DD) estimation strategies. For example, studies of state-level minimum wages must confront the fact that state average employment rates are correlated over time. We call this the serial correlation problem, to distinguish it from the Moulton problem.

Researchers plagued by clustering and serial correlation also have to confront the fact that the simplest fixups for these problems, like Stata's `cluster` option, may not be very good. The asymptotic approximation relevant for clustered or serially correlated data relies on a large number of clusters or time series observations. Alas, we are not always blessed with many clusters or long time series. The resulting inference problems are not always insurmountable, though often the best solution is to get more data. Econometric fixups for clustering and serial correlation are discussed in the second part of this chapter. Some of the material in this chapter is hard to work through without matrix algebra, so we take the plunge and switch to a mostly matrix motif.

8.1 The Bias of Robust Standard Error Estimates*

In matrix notation

$$\hat{\beta} = \left[\sum_i \mathrm{X}_i \mathrm{X}_i'\right]^{-1} \sum_i \mathrm{X}_i \mathrm{Y}_i = (X'X)^{-1}X'y,$$

where X is the $N \times \mathrm{K}$ matrix with rows X_i' and y is the $N \times 1$ vector of Y_i's. We saw in section 3.1.3 that $\hat{\beta}$ has an

asymptotically normal distribution. We can write:

$$\sqrt{N}(\hat{\beta} - \beta) \sim N(0, \Omega)$$

where Ω is the asymptotic covariance matrix and $\beta = E[\mathrm{X}_i\mathrm{X}_i']^{-1}E[\mathrm{X}_i\mathrm{Y}_i]$. Repeating (3.1.7), the formula for Ω in this case is

$$\Omega_r = E[\mathrm{X}_i\mathrm{X}_i']^{-1}E[\mathrm{X}_i\mathrm{X}_i'e_i^2]E[\mathrm{X}_i\mathrm{X}_i']^{-1}, \qquad (8.1.1)$$

where $e_i = \mathrm{Y}_i - \mathrm{X}_i'\beta$. When residuals are homoskedastic, the covariance matrix simplifies to $\Omega_c = \sigma^2 E[\mathrm{X}_i\mathrm{X}_i']^{-1}$, where $\sigma^2 = E[e_i^2]$.

We are concerned here with the bias of robust standard error estimates in independent samples (i.e., no clustering or serial correlation). To simplify the derivation of bias, we assume that the regressor vector can be treated as fixed, as it would be if we sampled stratifying on X_i. Nonstochastic regressors gives a benchmark sampling model that is often used to look at finite-sample distributions. It turns out that we miss little of theoretical importance by making this assumption, while simplifying the derivations considerably.

With fixed regressors, we have

$$\Omega_r = \left(\frac{X'X}{N}\right)^{-1}\left(\frac{X'\Psi X}{N}\right)\left(\frac{X'X}{N}\right)^{-1}, \qquad (8.1.2)$$

where

$$\Psi = E[ee'] = diag(\psi_i)$$

is the covariance matrix of residuals. Under homoskedasticity, $\psi_i = \sigma^2$ for all i and we get

$$\Omega_c = \sigma^2\left(\frac{X'X}{N}\right)^{-1}.$$

Asymptotic standard errors are given by the square root of the diagonal elements of Ω_r and Ω_c, after removing the asymptotic normalization by dividing by N.

In practice, the pieces of the asymptotic covariance matrix are estimated using sample moments. An old-fashioned or conventional covariance matrix estimator is

$$\hat{\Omega}_c = (X'X)^{-1}\hat{\sigma}^2 = (X'X)^{-1}\left(\sum \frac{\hat{e}_i^2}{N}\right),$$

where $\hat{e}_i = \text{Y}_i - \text{X}_i'\hat{\beta}$ is the estimated regression residual and

$$\hat{\sigma}^2 = \sum \frac{\hat{e}_i^2}{N}$$

estimates the residual variance. The corresponding robust covariance matrix estimator is

$$\hat{\Omega}_r = N(X'X)^{-1}\left(\sum \frac{\text{X}_i\text{X}_i'\hat{e}_i^2}{N}\right)(X'X)^{-1}. \qquad (8.1.3)$$

We can think of the middle term as an estimator of the form $\sum \frac{\text{X}_i\text{X}_i'\hat{\psi}_i}{N}$, where $\hat{\psi}_i = \hat{e}_i^2$ estimates ψ_i.

By the law of large numbers and Slutsky's theorem, $N\hat{\Omega}_c$ converges in probability to Ω_c, while $N\hat{\Omega}_r$ converges to Ω_r. But in finite samples, both variance estimators are biased. The bias in $\hat{\Omega}_c$ is well-known from classical least squares theory and easy to correct. Less appreciated is the fact that if the residuals are homoskedastic, the robust estimator is more biased than the conventional estimator, perhaps a lot more. From this we conclude that robust standard errors can be more misleading than conventional standard errors in situations where heteroskedasticity is modest. We also propose a rule of thumb that uses the maximum of old-fashioned and robust standard errors to avoid gross misjudgments of precision.

Our analysis begins with the bias of $\hat{\Omega}_c$. With nonstochastic regressors, we have

$$E[\hat{\Omega}_c] = (X'X)^{-1}\hat{\sigma}^2 = (X'X)^{-1}\left(\sum \frac{E(\hat{e}_i^2)}{N}\right).$$

To analyze $E[\hat{e}_i^2]$, start by expanding $\hat{e} = y - X\hat{\beta}$:

$$\hat{e} = y - X(X'X)^{-1}X'y = [I_N - X(X'X)^{-1}X'](X\beta + e) = Me,$$

where e is the vector of population residuals, $M = I_N - X(X'X)^{-1}X'$ is a nonstochastic residual-maker matrix with

ith row m_i', and I_N is the $N \times N$ identity matrix. Then $\hat{e}_i = m_i'e$, and

$$\begin{aligned} E(\hat{e}_i^2) &= E(m_i'ee'm_i) \\ &= m_i'\Psi m_i. \end{aligned}$$

To simplify further, write $m_i = \ell_i - h_i$, where ℓ_i is the ith column of I_N and $h_i = X(X'X)^{-1}\text{X}_i$, the ith column of the projection matrix $H = X(X'X)^{-1}X'$. Then

$$\begin{aligned} E(\hat{e}_i^2) &= (\ell_i - h_i)'\Psi(\ell_i - h_i) \\ &= \psi_i - 2\psi_i h_{ii} + h_i'\Psi h_i, \qquad (8.1.4) \end{aligned}$$

where h_{ii}, the ith diagonal element of H, satisfies

$$h_{ii} = h_i'h_i = \text{X}_i'(X'X)^{-1}\text{X}_i. \qquad (8.1.5)$$

Parenthetically, h_{ii} is called the *leverage* of the ith observation. Leverage tells us how much pull a particular value of X_i exerts on the regression line. Note that the ith fitted value (ith element of Hy) is

$$\hat{\text{Y}}_i = h_i'y = h_{ii}\text{Y}_i + \sum_{j \neq i} h_{ij}\text{Y}_j. \qquad (8.1.6)$$

A large h_{ii} means that the ith observation has a large impact on the ith predicted value. In a bivariate regression with a single regressor, x_i,

$$h_{ii} = \frac{1}{N} + \frac{(x_i - \overline{x})^2}{\sum (x_j - \overline{x})^2}. \qquad (8.1.7)$$

This shows that leverage increases when x_i is far the mean. In addition to (8.1.6), we know that h_{ii} is a number that lies in the interval [0, 1] and that $\sum_{i=1}^{N} h_{ii} = \text{K}$, the number of regressors (see, e.g., Hoaglin and Welsch, 1978).[1]

[1] The property $\sum_{i=1}^{N} h_{ii} = \text{K}$ comes from the fact that H is idempotent, and so has trace equal to rank. We can also use (8.1.7) to verify that in a bivariate regression, $\sum_{i=1}^{N} h_{ii} = 2$.

Suppose residuals are homoskedastic, so that $\psi_i = \sigma^2$. Then (8.1.4) simplifies to

$$E(\hat{e}_i^2) = \sigma^2[1 - 2h_{ii} + h_i'h_i] = \sigma^2(1 - h_{ii}) < \sigma^2.$$

So $\hat{\Omega}_c$ tends to be too small. Using the properties of h_{ii}, we can go one step further:

$$\sum \frac{E(\hat{e}_i^2)}{N} = \sigma^2 \sum \frac{1 - h_{ii}}{N} = \sigma^2\left(\frac{N - \text{K}}{N}\right).$$

Thus, the bias in $\hat{\Omega}_c$ can be fixed by a simple degrees-of-freedom correction: divide by $N - \text{K}$ instead of N in the formula for $\hat{\sigma}^2$. This correction is used by default in most regression software.

We now want to show that under homoskedasticity, the bias in $\hat{\Omega}_r$ is likely to be worse than the bias in $\hat{\Omega}_c$. The expected value of the robust covariance matrix estimator is

$$E[\hat{\Omega}_r] = N(X'X)^{-1}\left(\sum \frac{X_i X_i' E(\hat{e}_i^2)}{N}\right)(X'X)^{-1}, \qquad (8.1.8)$$

where $E(\hat{e}_i^2)$ is given by (8.1.4). Under homoskedasticity, $\psi_i = \sigma^2$ and we have $E(\hat{e}_i^2) = \sigma^2(1 - h_{ii})$ as in $\hat{\Omega}_c$. It's clear, therefore, that the bias in $\hat{e}_i^2$ tends to pull robust standard errors down. The general expression, (8.1.8), is hard to evaluate, however. Chesher and Jewitt (1987) show that as long as there is not "too much" heteroskedasticity, robust standard errors based on $\hat{\Omega}_r$ are indeed biased downward.[2]

How do we know that $\hat{\Omega}_r$ is likely to be *more* biased than $\hat{\Omega}_c$? Partly this comes from Monte Carlo evidence (e.g., MacKinnon and White, 1985, and our own small study, discussed below). We also prove this here for a bivariate example, where both the dependent variable and the single regressor, x_i, are assumed to be mean zero. (We're assuming the variables here are naturally mean zero so as to avoid the need for a degrees-of-freedom correction due to removal of sample means.) In this case, the estimator of interest is $\hat{\beta}_1 = \frac{\sum x_i Y_i}{\sum x_i^2}$

[2] In particular, as long as the ratio of the largest ψ_i to the smallest ψ_i is less than 2, robust standard errors are biased downward.

and the leverage is $h_{ii} = \frac{x_i^2}{\sum x_i^2}$ (we lose the $\frac{1}{N}$ term in (8.1.7) by dropping the constant). Let $s_x^2 = \frac{\sum x_i^2}{N}$. For the conventional covariance estimator, we have

$$E[\hat{\Omega}_c] = \frac{\sigma^2}{Ns_x^2}\left[\frac{\sum(1 - h_{ii})}{N}\right] = \frac{\sigma^2}{Ns_x^2}\left[1 - \frac{1}{N}\right],$$

so the bias here is small. A simple calculation using (8.1.8) shows that under homoskedasticity, the robust estimator has expectation:

$$\begin{aligned} E[\hat{\Omega}_r] &= \frac{\sigma^2}{Ns_x^2}\sum \frac{(1 - h_{ii})}{N}\left(\frac{x_i^2}{s_x^2}\right) \\ &= \frac{\sigma^2}{Ns_x^2}\sum(1 - h_{ii})h_{ii} = \frac{\sigma^2}{Ns_x^2}\left[1 - \sum h_{ii}^2\right]. \end{aligned}$$

The bias of $\hat{\Omega}_r$ is therefore worse than the bias of $\hat{\Omega}_c$ if $\sum h_{ii}^2 > \frac{1}{N}$, as it is by Jensen's inequality unless the regressor has constant leverage, in which case $h_{ii} = \frac{1}{N}$ for all i.[3]

We can reduce the bias in $\hat{\Omega}_r$ by trying to get a better estimator of ψ_i, say $\hat{\psi}_i$. The estimator $\hat{\Omega}_r$ sets $\hat{\psi}_i = \hat{e}_i^2$, as proposed by White (1980a) and our starting point in this section. The residual variance estimators discussed in MacKinnon and White (1985) include this and three others:

$$\begin{aligned} HC_0 &: \hat{\psi}_i = \hat{e}_i^2 \\ HC_1 &: \hat{\psi}_i = \frac{N}{N - \text{K}}\hat{e}_i^2 \end{aligned}$$

[3] Think of h_{ii} as a random variable with a uniform distribution in the sample. Then

$$E[h_{ii}] = \frac{\sum h_{ii}}{N} = \frac{1}{N},$$

and

$$E[h_{ii}^2] = \frac{\sum h_{ii}^2}{N} > (E[h_{ii}])^2 = \left(\frac{1}{N}\right)^2$$

by Jensen's inequality unless h_{ii} is constant. Therefore $\sum h_{ii}^2 > \frac{1}{N}$. The constant leverage case occurs when x_i^2 is constant.

$$HC_2 : \hat{\psi}_i = \frac{1}{1-h_{ii}}\hat{e}_i^2$$

$$HC_3 : \hat{\psi}_i = \frac{1}{(1-h_{ii})^2}\hat{e}_i^2.$$

HC_1 is a simple degrees of freedom correction as is used for $\hat{\Omega}_c$. HC_2 uses the leverage to give an unbiased estimate of the variance of the ith residual when the residuals are homoskedastic, while HC_3 approximates a jackknife estimator.[4] In the applications we've seen, the estimated standard errors tend to get larger as we go down the list from HC_0 to HC_3, but this is not a theorem.

Time-Out for the Bootstrap

Bootstrapping is a resampling scheme that offers an alternative to inference based on asymptotic formulas. A bootstrap sample is a sample drawn from our own data. In other words, if we have a sample of size N, we treat this sample as if it were the population and draw repeatedly from it (with replacement). The bootstrap sampling distribution is the distribution of an estimator across many draws of this sort. Intuitively, we expect the sampling distribution constructed by sampling from our own data to provide a good approximation to the sampling distribution we are after.

There are many ways to bootstrap regression estimates. The simplest is to draw pairs of $\{Y_i, X_i\}$ values, sometimes called the "pairs bootstrap" or a "nonparametric bootstrap." Alternatively, we can keep the X_i values fixed, draw from the distribution of residuals $(\hat{e}_i)$, and create a new estimate of the dependent variable based on the predicted value and the residual draw for each observation. This procedure, which is a type of parametric bootstrap, mimics a sample drawn with nonstochastic regressors and ensures that X_i and the regression

residuals are independent. On the other hand, we don't want independence if we're interested in standard errors under heteroskedasticity. An alternative residual bootstrap, called the wild bootstrap, draws $X_i'\hat{\beta} + \hat{e}_i$ (which, of course, is just the original Y_i) with probability 0.5, and $X_i'\hat{\beta} - \hat{e}_i$ otherwise (see, e.g., Mammen, 1993, and Horowitz, 1997). This preserves the relationship between residual variances and X_i observed in the original sample, while imposing mean-independence of residuals and regressors, a restriction that improves bootstrap inference when true.

Bootstrapping is useful as a computer-intensive but otherwise straightforward calculator for asymptotic standard errors. The bootstrap calculator is especially useful when the asymptotic distribution of an estimator is hard to compute or involves a number of steps (e.g., the asymptotic distributions of the quantile regression and quantile treatment effects estimates discussed in chapter 7 require the estimation of densities). Typically, however, we have no problem deriving or evaluating asymptotic formulas for the standard errors of OLS estimates.

More relevant in this context is the use of the bootstrap to improve inference. Improvements in inference potentially come in two forms: (1) a reduction in finite-sample bias in estimators that are consistent (for example, the bias in estimates of robust standard errors) and (2) inference procedures which make use of the fact that the bootstrap sampling distribution of test statistics may be closer to the finite-sample distribution of interest than the relevant asymptotic approximation. These two properties are called asymptotic refinements (see, e.g., Horowitz, 2001).

Here we are mostly interested in use of the bootstrap for asymptotic refinement. The asymptotic distribution of regression estimates is easy enough to compute, but we worry that the traditional robust covariance estimator (HC_0) is biased. The bootstrap can be used to estimate this bias, and then, by a simple transformation, to construct standard error estimates that are less biased. However, for now at least, bootstrap bias correction of regression standard errors is not often used in empirical practice, perhaps because the bias calculation is not

[4] A jackknife variance estimator estimates sampling variance from the empirical distribution generated by omitting one observation at a time. Stata computes HC_1, HC_2, and HC_3. You can also use a trick suggested by Messer and White (1984): divide Y_i and X_i by $\sqrt{\hat{\psi}_i}$ and instrument the transformed model by $X_i/\sqrt{\hat{\psi}_i}$ for your preferred choice of $\hat{\psi}_i$.

automated and perhaps because bootstrap bias corrections introduce extra variability. Also, for simple estimators like regression coefficients, analytic bias corrections such as HC_2 and HC_3 are readily available (e.g., in Stata).

An asymptotic refinement can also be obtained for hypothesis tests (and confidence intervals) based on statistics that are *asymptotically pivotal*. These are statistics that have asymptotic distributions that do not depend on any unknown parameters. An example is a t-statistic: this is asymptotically standard normal. Regression coefficients are not asymptotically pivotal; they have an asymptotic distribution that depends on the unknown residual variance. To refine inference for regression coefficients, you calculate the t-statistic in each bootstrap sample and compare the analogous t-statistic from your original sample to this bootstrap "t-distribution." A hypothesis is rejected if the absolute value of the original t-statistic is above, say, the 95th percentile of the absolute values from the bootstrap distribution.

Theoretical appeal notwithstanding, as applied researchers, we don't like the idea of bootstrapping pivotal statics very much. This is partly because we're not only (or even primarily) interested in formal hypothesis testing: we like to see the standard errors in parentheses under our regression coefficients. These provide a summary measure of precision that can be used to construct confidence intervals, compare estimators, and test any hypothesis that strikes us, now or later. In our view, therefore, practitioners worried about the finite-sample behavior of robust standard errors should focus on bias corrections like HC_2 and HC_3. As we show below, for moderate heteroskedasticity at least, an inference strategy that uses the larger of conventional and bias-corrected standard errors often seems to give us the best of both worlds: reduced bias with a minimal loss of precision.

An Example

For further insight into the differences between robust covariance estimators, we analyze a simple but important example that has featured in earlier chapters in this book. Suppose you

are interested in an estimate of β_1 in the model

$$\mathrm{Y}_i = \beta_0 + \beta_1 \mathrm{D}_i + \varepsilon_i, \qquad (8.1.9)$$

where D_i is a dummy variable. The OLS estimate of β_1 is the difference in means between those with D_i switched on and off. Denoting these subsamples by the subscripts 1 and 0, we have

$$\hat{\beta}_1 = \overline{\mathrm{Y}}_1 - \overline{\mathrm{Y}}_0.$$

For the purposes of this derivation we think of D_i as nonrandom, so that $\sum \mathrm{D}_i = N_1$ and $\sum(1 - \mathrm{D}_i) = N_0$ are fixed. Let $r = N_1/N$.

We know something about the finite-sample behavior of $\hat{\beta}_1$ from statistical theory. If Y_i is normal with equal but unknown variance in both the $\mathrm{D}_i = 1$ and $\mathrm{D}_i = 0$ populations, then the conventional t-statistic for $\hat{\beta}_1$ has a t-distribution. This is the classic two-sample t-test. Heteroskedasticity in this context means that the variances in the $\mathrm{D}_i = 1$ and $\mathrm{D}_i = 0$ populations are different. In this case, the testing problem in small samples becomes surprisingly difficult: the exact small-sample distribution for even this simple problem is unknown.[5] The robust variance estimators HC_0–HC_3 give asymptotic approximations to the unknown finite-sample distribution for the case of unequal variances.

The differences between HC_0, HC_1, HC_2, and HC_3 are differences in how the sample variances in the two groups defined by D_i are processed. Define $S_j^2 = \sum_{\mathrm{D}_i = j}(\mathrm{Y}_i - \overline{\mathrm{Y}}_j)^2$ for $j = 0, 1$. The leverage in this example is

$$h_{ii} = \begin{cases} \frac{1}{N_0} & \text{if } \mathrm{D}_i = 0 \\ \frac{1}{N_1} & \text{if } \mathrm{D}_i = 1 \end{cases}.$$

Using this, it's straightforward to show that the five variance estimators we've been discussing are

$$\textit{Conventional}: \frac{N}{N_0 N_1}\left(\frac{S_0^2 + S_1^2}{N-2}\right) = \frac{1}{Nr(1-r)}\left(\frac{S_0^2 + S_1^2}{N-2}\right)$$

[5] This is called the Behrens-Fisher problem (see, e.g., DeGroot and Schervish, 2001, chap. 8).

$$HC_0\text{:}\ \frac{S_0^2}{N_0^2}+\frac{S_1^2}{N_1^2}$$

$$HC_1\text{:}\ \frac{N}{N-2}\left(\frac{S_0^2}{N_0^2}+\frac{S_1^2}{N_1^2}\right)$$

$$HC_2\text{:}\ \frac{S_0^2}{N_0(N_0-1)}+\frac{S_1^2}{N_1(N_1-1)}$$

$$HC_3\text{:}\ \frac{S_0^2}{(N_0-1)^2}+\frac{S_1^2}{(N_1-1)^2}.$$

The conventional estimator pools subsamples: this is efficient when the two variances are the same. The White (1980a) estimator, HC_0, adds separate estimates of the sampling variances of the means, using the consistent (but biased) variance estimators, $\frac{S_j^2}{N_j}$. The HC_2 estimator uses unbiased estimators of the sample variance for each group, since it makes the correct degrees-of-freedom correction. HC_1 makes a degrees-of-freedom correction outside the sum, which will help but is generally not quite correct. Since we know HC_2 to be the unbiased estimate of the sampling variance under homoskedasticity, HC_3 must be too big.[6] Note that with $r = 0.5$, a case where the regression design is said to be balanced, the conventional estimator equals HC_1 and all five estimators differ little.

A small Monte Carlo study based on (8.1.9) illustrates the pluses and minuses of alternative estimators and the extent to which a simple rule of thumb goes a long way toward ameliorating the bias of the HC class. We choose $N = 30$ to highlight small sample issues, and $r = 0.10$ (10 percent treated), which implies $h_{ii} = \frac{1}{3}$ if $\text{D}_i = 1$ and $h_{ii} = \frac{1}{27}$ if $\text{D}_i = 0$. This is a highly unbalanced design. We draw residuals from the distributions:

$$\varepsilon_i \sim \begin{cases} N(0,\sigma^2) & \text{if } \text{D}_i = 0 \\ N(0,1) & \text{if } \text{D}_i = 1 \end{cases}$$

and report results for three cases. The first has lots of heteroskedasticity, with $\sigma = 0.5$, while the second has relatively little heteroskedasticity, with $\sigma = 0.85$. No heteroskedasticity is the benchmark case.

Table 8.1.1 displays the results. Columns 1 and 2 report means and standard deviations of the various standard error estimates across 25,000 replications of the sampling experiment. The standard deviation of $\hat{\beta}_1$ is the sampling variance we are trying to measure. With lots of heteroskedasticity, as in the upper panel of the table, conventional standard errors are badly biased and, on average, only about half the size of the Monte Carlo sampling variance that constitutes our target. On the other hand, while the robust standard errors perform better, except for HC_3, they are still too small.[7]

The standard errors are themselves estimates and have considerable sampling variability. Especially noteworthy is the fact that the robust standard errors have much higher sampling variability than the conventional standard errors, as can be seen in column 2.[8] The sampling variability of estimated standard errors further increases when we attempt to reduce bias by dividing the residuals by $1-h_{ii}$ (HC_2) or $(1-h_{ii})^2$ (HC_3). The worst case is HC_3, with a standard deviation about 50 percent above the standard deviation of the White (1980a) standard error, HC_0.

The last two columns in the table show empirical rejection rates in a nominal 5 percent test for the hypothesis $\beta_1 = 0$, the population parameter in this case. The test statistics are compared with a normal distribution and to a t-distribution with $N-2$ degrees of freedom. Rejection rates are far too high for all tests, even with HC_3. Using a t-distribution rather than a normal distribution helps only marginally.

[6] In this simple example, HC_2 is unbiased whether or not residuals are homoskedastic.

[7] Although HC_2 is an unbiased estimator of the sampling variance, the mean of the HC_2 standard errors across sampling experiments (0.52) is still below the standard deviation of $\hat{\beta}_1$ (0.59). This comes from the fact that the standard error is the square root of the sampling variance, the sampling variance is itself estimated and hence has sampling variability, and the square root is a concave function.

[8] The large sampling variance of robust standard error estimators is noted by Chesher and Austin (1991). Kauermann and Carroll (2001) propose an adjustment to confidence intervals to correct for this.

Table 8.1.1
Monte Carlo results for robust standard error estimates

	Mean	Standard Deviation	Empirical 5% Rejection Rates	
			Normal	t
Parameter Estimate	(1)	(2)	(3)	(4)
A. Lots of heteroskedasticity				
$\hat{\beta}_1$	−.001	.586		
Standard Errors				
Conventional	.331	.052	.278	.257
HC_0	.417	.203	.247	.231
HC_1	.447	.218	.223	.208
HC_2	.523	.260	.177	.164
HC_3	.636	.321	.130	.120
max(HC_0, Conventional)	.448	.172	.188	.171
max(HC_1, Conventional)	.473	.190	.173	.157
max(HC_2, Conventional)	.542	.238	.141	.128
max(HC_3, Conventional)	.649	.305	.107	.097
B. Little heteroskedasticity				
$\hat{\beta}_1$	.004	.600		
Standard Errors				
Conventional	.520	.070	.098	.084
HC_0	.441	.193	.217	.202
HC_1	.473	.207	.194	.179
HC_2	.546	.250	.156	.143
HC_3	.657	.312	.114	.104
max(HC_0, Conventional)	.562	.121	.083	.070
max(HC_1, Conventional)	.578	.138	.078	.067
max(HC_2, Conventional)	.627	.186	.067	.057
max(HC_3, Conventional)	.713	.259	.053	.045
C. No heteroskedasticity				
$\hat{\beta}_1$	−.003	.611		
Standard Errors				
Conventional	.604	.081	.061	.050
HC_0	.453	.190	.209	.193
HC_1	.486	.203	.185	.171
HC_2	.557	.247	.150	.136
HC_3	.667	.309	.110	.100
max(HC_0, Conventional)	.629	.109	.055	.045
max(HC_1, Conventional)	.640	.122	.053	.044
max(HC_2, Conventional)	.679	.166	.047	.039
max(HC_3, Conventional)	.754	.237	.039	.031

Notes: The table reports results from a sampling experiment with 25,000 replications. Columns 1 and 2 shows the mean and standard deviation of estimated *standard errors*, except for the first row in each panel which shows the mean and standard deviation of $\hat{\beta}_1$. The model is as described by (8.1.9), with $\beta_1 = 0$, $r = .1$, $N = 30$, and heteroskedasticity as indicated in the panel headings.

The results with little heteroskedasticity, reported in the second panel, show that conventional standard errors are still too low; this bias is now on the order of 15 precent. HC_0 and HC_1 are also too small, about as before in absolute terms, though they now look worse relative to the conventional standard errors. The HC_2 and HC_3 standard errors are still larger than the conventional standard errors, on average, but empirical rejection rates are higher for these two than for conventional standard errors. This means the robust standard errors are sometimes too small "by accident," an event that happens often enough to inflate rejection rates so that they exceed the conventional rejection rates.

One lesson we can take away from this is that robust standard errors are no panacea. They can be smaller than conventional standard errors for two reasons: the small sample bias we have discussed and their higher sampling variance. We therefore take empirical results where the robust standard errors fall below the conventional standard errors as a red flag. This is very likely due to bias or a chance occurrence that is better discounted. In this spirit, the maximum of the conventional standard error and a robust standard error may be the best measure of precision. This rule of thumb helps on two counts: it truncates low values of the robust estimators, reducing bias, and it reduces variability. Table 8.1.1 shows the empirical rejection rates obtained using max(HC_j, *Conventional*). Rejection rates using this rule of thumb look pretty good in panel B and are considerably better than the rates using robust estimators alone, even with lots of heteroskedasticity, as shown in panel A.[9]

Since there is no gain without pain, there must be some cost to using max(HC_j, *Conventional*). The cost is that the best standard error when there is no heteroskedasticity is the conventional estimate. This is documented in the bottom panel of the table. Use of the maximum inflates standard errors unnecessarily under homoskedasticity, depressing rejection rates. Nevertheless, the table shows that even in this case, rejection

[9] Yang, Hsu, and Zhao (2005) formalize the notion of test procedures based on the maximum of a set of test statistics with differing efficiency and robustness properties.

rates don't go down all that much. We also view an underestimate of precision as being less costly than an overestimate. Underestimating precision, we come away thinking the data are not very informative and that we should try to collect more or improve the research design, while in the latter case we may mistakenly draw important substantive conclusions.

A final comment on this Monte Carlo investigation concerns the small sample size. Labor economists like us are used to working with tens of thousands of observations or more. But sometimes we don't. In a study of the effects of busing on public school students, Angrist and Lang (2004) worked with samples of about 3,000 students grouped in 56 schools. The regressor of interest in this study varied within grade only at the school level, so some of the analysis uses 56 school means. Not surprisingly, therefore, Angrist and Lang (2004) obtained HC_1 standard errors below conventional OLS standard errors when working with school-level data. As a rule, even if you start with the microdata on individuals, when the regressor of interest varies at a higher level of aggregation—a school, state, or some other group or cluster—effective sample sizes are much closer to the number of clusters than to the number of individuals. Inference procedures for clustered data are discussed in detail in the next section.

8.2 Clustering and Serial Correlation in Panels

8.2.1 Clustering and the Moulton Factor

Heteroskedasticity rarely leads to dramatic changes in inference. In large samples where bias is not likely to be a problem, we might see standard errors increase by about 25 percent when moving from the conventional to the HC_1 estimator. In contrast, clustering can make all the difference.

The clustering problem can be illustrated using a simple bivariate model estimated in data with a group structure. Suppose we're interested in the bivariate regression,

$$Y_{ig} = \beta_0 + \beta_1 x_g + e_{ig}, \qquad (8.2.1)$$

where Y_{ig} is the dependent variable for individual i in cluster or group g, with G groups. Importantly, the regressor of interest, x_g, varies only at the group level. For example, data from the STAR experiment analyzed by Krueger (1999) come in the form of Y_{ig}, the test score of student i in class g, and class size, x_g.

Although students were randomly assigned to classes in the STAR experiment, the STAR data are unlikely to be independent across observations. The test scores of students in the same class tend to be correlated because students in the same class share background characteristics and are exposed to the same teacher and classroom environment. It's therefore prudent to assume that, for students i and j in the same class, g,

$$E[e_{ig}e_{jg}] = \rho_e \sigma_e^2 > 0, \qquad (8.2.2)$$

where ρ_e is the residual intraclass correlation coefficient and σ_e^2 is the residual variance.

Correlation within groups is often modeled using an additive random effects model. Specifically, we assume that the residual, e_{ig}, has a group structure,

$$e_{ig} = v_g + \eta_{ig}, \qquad (8.2.3)$$

where v_g is a random component specific to class g and η_{ig} is a mean-zero student-level error component that's left over. We focus here on the correlation problem, so both of these error components are assumed to be homoskedastic. The group-level error component is assumed to capture all within-group correlation, so the η_{ig} are uncorrelated.[10]

When the regressor of interest varies only at the group level, an error structure like (8.2.3) can increase standard errors sharply. This unfortunate fact is not news—Kloek (1981) and

[10] This sort of residual correlation structure is also a consequence of stratified sampling (see, e.g., Wooldridge, 2003). Most of the samples that we work with are close enough to random that we typically worry more about the dependence due to a group structure than clustering due to stratification. Note that there is no GLS estimator for equation 8.2.1 with error structure 8.2.3 because the regressor is fixed within groups. In any case, here as elsewhere we prefer a "fix-the-standard-errors" approach to GLS.

Moulton (1986) both made the point—but it seems fair to say that clustering didn't really become part of the applied econometrics zeitgeist until about 15 years ago.

Given the error structure, (8.2.3), the intraclass correlation coefficient becomes

$$\rho_e = \frac{\sigma_v^2}{\sigma_v^2 + \sigma_\eta^2},$$

where σ_v^2 is the variance of v_g and σ_η^2 is the variance of η_{ig}. A word on terminology: ρ_e is called the *intraclass correlation coefficient* even when the groups of interest are not classrooms.

Let $V_c(\hat{\beta}_1)$ be the conventional OLS variance formula for the regression slope (a diagonal element of Ω_c in the previous section), while $V(\hat{\beta}_1)$ denotes the correct sampling variance given the error structure, (8.2.3). With nonstochastic regressors fixed at the group level and groups of equal size, n, we have

$$\frac{V(\hat{\beta}_1)}{V_c(\hat{\beta}_1)} = 1 + (n-1)\rho_e, \qquad (8.2.4)$$

a formula derived in the appendix to this chapter. We call the square root of this ratio the Moulton factor, after Moulton's (1986) influential study. Equation (8.2.4) tells us how much we overestimate precision by ignoring intraclass correlation. Conventional standard errors become increasingly misleading as n and ρ_e increase. Suppose, for example, that $\rho_e = 1$. In this case, all the errors within a group are the same, so the Y_{ig} values are the same as well. Making a data set larger by copying a smaller one n times generates no new information. The variance $V_c(\hat{\beta}_1)$ should therefore be scaled up from $V_c(\hat{\beta}_1)$ by a factor of n. The Moulton factor increases with group size because with a fixed overall sample size, larger groups mean fewer clusters, in which case there is less independent information in the sample (because the data are independent across clusters but not within).[11]

Even small intraclass correlation coefficients can generate a big Moulton factor. In Angrist and Lavy (2008), for example, 4,000 students are grouped in 40 schools, so the average n is 100. The regressor of interest is school-level treatment status: all students in treated schools were eligible to receive cash awards for passing their matriculation exams. The intraclass correlation in this study fluctuates around .1. Applying formula (8.2.4), the Moulton factor is over 3, so the standard errors reported by default are only one-third what they should be.

Equation (8.2.4) covers an important special case where the regressors are fixed within groups and group size is constant. The general formula allows the regressor, x_{ig}, to vary at the individual level and for different group sizes, n_g. In this case, the Moulton factor is the square root of

$$\frac{V(\hat{\beta}_1)}{V_c(\hat{\beta}_1)} = 1 + \left[\frac{V(n_g)}{\bar{n}} + \bar{n} - 1\right]\rho_x \rho_e, \qquad (8.2.5)$$

where $\bar{n}$ is the average group size, and ρ_x is the intraclass correlation of x_{ig}:

$$\rho_x = \frac{\sum_g \sum_j \sum_{i \neq j} (x_{ig} - \bar{x})(x_{jg} - \bar{x})}{V(x_{ig}) \sum_g n_g(n_g - 1)}.$$

Note that ρ_x does not impose a variance components structure like (8.2.3); here, ρ_x is a generic measure of the correlation of regressors within groups. The general Moulton formula tells us that clustering has a bigger impact on standard errors with variable group sizes and when ρ_x is large. The impact vanishes when $\rho_x = 0$. In other words, if the x_{ig} values are uncorrelated within groups, the grouped error structure does not matter for standard errors. That's why we worry most about clustering when the regressor of interest is fixed within groups.

[11] With nonstochastic regressors and homoscedastic residuals, the Moulton factor is a finite-sample result. Survey statisticians call the Moulton factor the *design effect* because it tells us how much to adjust standard errors in stratified samples for deviations from simple random sampling (Kish, 1965).

We illustrate formula (8.2.5) using the Tennessee STAR example. A regression of kindergartners' percentile score on class size yields an estimate of $-.62$ with a robust (HC_1) standard error of .09. In this case, $\rho_x = 1$ because class size is fixed within classes, while $V(n_g)$ is positive because classes vary in size (in this case, $V(n_g) = 17.1$). The intraclass correlation coefficient for residuals is .31 and the average class size is 19.4. Plugging these numbers into (8.2.5) gives a value of about 7 for $\frac{V(\hat{\beta}_1)}{V_c(\hat{\beta}_1)}$, so that conventional standard errors should be multiplied by a factor of $2.65 = \sqrt{7}$. The corrected standard error is therefore about 0.24.

The Moulton factor works similarly with 2SLS estimates. In particular, we can use (8.2.5), replacing ρ_x with $\rho_{\hat{x}}$, where $\rho_{\hat{x}}$ is the intraclass correlation coefficient of the first-stage fitted values and ρ_e is the intraclass correlation of the second-stage residuals (Shore-Sheppard, 1996). To understand why this works, recall that conventional standard errors for 2SLS are derived from the residual variance of the second-stage equation divided by the variance of the first-stage fitted values. This is the same asymptotic variance formula as for OLS, with first-stage fitted values playing the role of the regressor.

To conclude, we list and compare solutions to the Moulton problem, starting with the parametric approach described above.

1. Parametric: Fix conventional standard errors using (8.2.5). The intraclass correlations ρ_e and ρ_x are easy to compute and supplied as descriptive statistics in some software packages.[12]
2. Cluster standard errors: Liang and Zeger (1986) generalize the White (1980a) robust covariance matrix to allow for clustering as well as heteroskedasticity. The clustered covariance matrix is

$$\hat{\Omega}_{cl} = (X'X)^{-1}\left(\sum_g X_g'\hat{\Psi}_g X_g\right)(X'X)^{-1}, \text{ where} \qquad (8.2.6)$$

[12] Use Stata's `loneway` command, for example.

$$\hat{\Psi}_g = a\hat{e}_g\hat{e}_g'$$

$$= a\begin{bmatrix} \hat{e}_{1g}^2 & \hat{e}_{1g}\hat{e}_{2g} & \cdots & \hat{e}_{1g}\hat{e}_{n_g g} \\ \hat{e}_{1g}\hat{e}_{2g} & \hat{e}_{2g}^2 & \cdots & \vdots \\ \vdots & \vdots & & \hat{e}_{n_g-1,g}\hat{e}_{n_g g} \\ \hat{e}_{1g}\hat{e}_{n_g g} & \cdots & \hat{e}_{n_g-1,g}\hat{e}_{n_g g} & \hat{e}_{n_g g}^2 \end{bmatrix}.$$

Here, X_g is the matrix of regressors for group g and a is a degrees of freedom adjustment factor similar to that which appears in HC_1. The clustered estimator is consistent as the number of groups gets large given any within-group correlation structure and not just the parametric model in (8.2.3). $\hat{\Omega}_{cl}$ is not consistent with a fixed number of groups, however, even when the group size tends to infinity. Consistency is determined by the law of large numbers, which says that we can rely on sample moments to converge to population moments (section 3.1.3). But here the sums are at the group level and not over individuals. Clustered standard errors are therefore unlikely to be reliable with few clusters, a point we return to below.

3. Use group averages instead of microdata: let $\bar{Y}_g$ be the mean of Y_{ig} in group g. Estimate

$$\bar{Y}_g = \beta_0 + \beta_1 x_g + \bar{e}_g$$

by WLS using the group size as weights. This is equivalent to OLS using micro data but the grouped-equation standard errors reflect the group structure, (8.2.3).[13] Again, the asymptotics here are based on the number of groups and not the group size. Importantly, however, because the group means are close to normally distributed with modest group sizes, we can expect the good finite-sample properties of regression with normal errors to kick in. The standard errors that come out of grouped estimation are therefore likely to be more reliable than clustered standard errors in samples with few clusters.

[13] The grouped residuals are heteroskedastic unless group sizes are equal but this is less important than the fact that the error has a group structure in the microdata.

Grouped-data estimation can be generalized to models with microcovariates using a two-step procedure. Suppose the equation of interest is

$$Y_{ig} = \beta_0 + \beta_1 x_g + \beta_2 W_{ig} + e_{ig}, \qquad (8.2.7)$$

where W_{ig} is a covariate that varies within groups. In step 1, construct the covariate-adjusted group effects, μ_g, by estimating

$$Y_{ig} = \mu_g + \beta_2 W_{ig} + \eta_{ig}.$$

The μ_g, called group effects, are coefficients on a full set of group dummies. The estimated $\hat{\mu}_g$ are group means adjusted for differences in the individual level variable, W_{ig}. Note that, by virtue of (8.2.7) and (8.2.3), $\mu_g = \beta_0 + \beta_1 x_g + \nu_g$. In step 2, therefore, we regress the estimated group effects on group-level variables:

$$\hat{\mu}_g = \beta_0 + \beta_1 x_g + \{\nu_g + (\hat{\mu}_g - \mu_g)\}. \qquad (8.2.8)$$

The efficient GLS estimator for (8.2.8) is WLS, using the reciprocal of the estimated variance of the group-level residual, $\{\nu_g + (\hat{\mu}_g - \mu_g)\}$, as weights. This can be a problem, since the variance of ν_g is not estimated very well with few groups. We might therefore weight by the reciprocal of the variance of the estimated group effects, the group size, or use no weights at all.[14] In an effort to better approximate the relevant finite-sample distribution, Donald and Lang (2007) suggest that inference for grouped equations like (8.2.8) be based on a t-distribution with $G - K$ degrees of freedom.

Note that the grouping approach does not work when x_{ig} varies within groups. Averaging x_{ig} to $\bar{x}_g$ is a version of IV, as we saw in chapter 4. So with micro-variation in the regressor of interest, grouped estimation identifies parameters that differ from the target parameters in a model like (8.2.7).

[14] See Angrist and Lavy (2008) for an example of the latter two weighting schemes.

4. Block bootstrap: In general, bootstrap inference uses the empirical distribution of the data by resampling. But simple random resampling won't do in this case. The trick with clustered data is to preserve the dependence structure in the target population. We can do this by block bootstrapping, that is, drawing blocks of data defined by the groups g. In the Tennessee STAR data, for example, we'd block bootstrap by resampling entire classes instead of individual students.
5. In some cases, you may be able to estimate a GLS or maximum likelihood model based on a version of (8.2.1) combined with a model for the error structure like (8.2.3). This fixes the clustering problem but also changes the estimand unless the CEF is linear, as detailed in section 3.4.1 for LDV models. We therefore prefer other approaches.

Table 8.2.1 compares standard-error fixups in the STAR example. The table reports six estimates: conventional robust standard errors (using HC_1); two versions of corrected standard errors using the Moulton formula (8.2.5), the first using the formula for the intraclass correlation given by Moulton and the second using Stata's estimator from the `loneway` command; clustered standard errors; block-bootstrapped standard errors; and standard errors from weighted estimation at the group level. The coefficient estimate is −.62. In this case, all cluster adjustments deliver similar results, a standard error of about .23. This happy outcome is due in large part to the fact that with 318 classrooms, we have enough clusters for group-level asymptotics to work well. With few clusters, however, things are much dicier, a point we return to at the end of the chapter.

8.2.2 *Serial Correlation in Panels and Difference-in-Difference Models*

Serial correlation—the tendency for one observation to be correlated with those that have gone before—used to be Somebody Else's Problem, specifically, the unfortunate souls who make their living out of time series data (macroeconomists, for

TABLE 8.2.1
Standard errors for class size effects in the STAR data (318 clusters)

Variance Estimator	Std. Err.
Robust (HC_1)	.090
Parametric Moulton correction (using Moulton intraclass correlation)	.222
Parametric Moulton correction (using Stata intraclass correlation)	.230
Clustered	.232
Block bootstrap	.231
Estimation using group means (weighted by class size)	.226

Notes: The table reports standard errors for the estimates from a regression of kindergartners' average percentile scores on class size using the public use data set from Project STAR. The coefficient on class size is −.62. The group level for clustering is the classroom. The number of observations is 5,743. The bootstrap estimate uses 1,000 replications.

example). Applied microeconometricians have therefore long ignored it.[15] But our data often have a time dimension, too, especially in DD models. This fact combined with clustering can have a major impact on statistical inference.

Suppose, as in section 5.2, that we are interested in the effects of a state minimum wage. In this context, the regression version of DD includes additive state and time effects. We therefore get an equation like (5.2.2), repeated below:

$$Y_{ist} = \gamma_s + \lambda_t + \delta D_{st} + \varepsilon_{ist}, \qquad (8.2.9)$$

As before, Y_{ist} is the outcome for individual i in state s in year t and D_{st} is a dummy variable that indicates treatment states in posttreatment periods.

The error term in (8.2.9) reflects the idiosyncratic variation in potential outcomes across people, states, and time. Some of this variation is likely to be common to individuals in the same state and year, for example, a regional business cycle. We can model this common component by thinking of ε_{ist} as the sum of a state-year shock, v_{st}, and an idiosyncratic individual component, η_{ist}. So we have:

$$Y_{ist} = \gamma_s + \lambda_t + \delta D_{st} + v_{st} + \eta_{ist}. \qquad (8.2.10)$$

We assume that in repeated draws across states and over time, $E[v_{st}] = 0$, while $E[\eta_{ist}|s, t] = 0$ by definition.

State-year shocks are bad news for DD models. As with the Moulton problem, state- and time-specific random effects generate a clustering problem that affects statistical inference. But that might be the least of our problems in this case. To see why, suppose we have only two periods and two states, as in the Card and Krueger (1994) New Jersey-Pennsylvania study. The empirical DD estimator is

$$\hat{\delta}_{CK} = (\bar{Y}_{s=NJ,t=Nov} - \bar{Y}_{s=NJ,t=Feb}) - (\bar{Y}_{s=PA,t=Nov} - \bar{Y}_{s=PA,t=Feb}).$$

This estimator is unbiased, since $E[v_{st}] = E[\eta_{ist}] = 0$. On the other hand, assuming we think of probability limits as increasing group size while keeping the choice of states and periods fixed, state-year shocks render $\hat{\delta}_{CK}$ inconsistent:

$$\begin{aligned} &\mathit{plim}\ \hat{\delta}_{CK} \\ &= \delta + \{(v_{s=NJ,t=Nov} - v_{s=NJ,t=Feb}) - (v_{s=PA,t=Nov} - v_{s=PA,t=Feb})\}. \end{aligned}$$

Averaging larger and larger samples within New Jersey and Pennsylvania in a pair of periods does nothing to eliminate the regional shocks specific to a given location and period. With only two states and years, we have no way to distinguish the differences-in-differences generated by a policy

[15] The Somebody Else's Problem (SEP) field, first identified as a natural phenomenon in Adams's *Life, the Universe, and Everything*, is, according to Wikipedia, "a generated energy field that affects perception. . . . Entities within the field will be perceived by an outside observer as 'Somebody Else's Problem,' and will therefore be effectively invisible unless the observer is specifically looking for the entity."

change from the difference-in-differences due to the fact that, say, the New Jersey economy was holding steady in 1992 while Pennsylvania was experiencing a cyclical downturn. The presence of v_{st} amounts to a failure of the common trends assumption discussed in section 5.2.

The solution to the inconsistency induced by random shocks in differences in differences models is to analyze samples including multiple time periods or many states (or both). For example, Card (1992) uses 51 states to study minimum wage changes, while Card and Krueger (2000) take another look at the New Jersey-Pennsylvania experiment with a longer monthly time series of payroll data. With multiple states or periods, we can hope that the v_{st} average out to zero. As in the first part of this chapter on the Moulton problem, the inference framework in this context relies on asymptotic distribution theory with many groups and not on group size (or, at least, not on group size alone). The most important inference issue then becomes the behavior of v_{st}. In particular, if we are prepared to assume that shocks are independent across states and over time—that is, that they are serially uncorrelated—we are back to the plain vanilla Moulton problem in section 8.2.1, in which case clustering standard errors by state × year should generate valid inferences. But in most cases, the assumption that v_{st} is serially uncorrelated is hard to defend. Almost certainly, for example, regional shocks are highly serially correlated: if things are bad in Pennsylvania in one month, they are likely to be about as bad in the next.

The consequences of serial correlation for clustered panels are highlighted by Bertrand, Duflo, and Mullainathan (2004) and Kézdi (2004). Any research design with a group structure where the group means are correlated can be said to have the serial correlation problem. The upshot of recent research on serial correlation in data with a group structure is that, just as we must adjust our standard errors for the correlation within groups induced by the presence of v_{st}, we must further adjust for serial correlation in the v_{st} themselves. There are a number of ways to do this, not all equally effective in all situations. It seems fair to say that the question of how best to approach the serial correlation problem is currently under study, and a consensus has not yet emerged.

The simplest and most widely applied approach is to pass the clustering buck one level higher. In the state-year example, we can report Liang and Zeger (1986) standard errors clustered by state instead of by state and year (e.g., using Stata `cluster`). This might seem odd at first blush, since the model controls for state effects. The state effect, γ_s, in (8.2.10) removes the state mean of v_{st}, which we denote by $\bar{v}_s$. Nevertheless, $v_{st} - \bar{v}_s$ is probably still serially correlated. Clustering standard errors at the state level takes account of this, since the one-level-up clustered covariance estimator allows for unrestricted residual correlation within clusters, including the time series correlation in $v_{st} - \bar{v}_s$. This is a quick and easy fix.[16] The problem here is that passing the buck up one level reduces the number of clusters. And asymptotic inference supposes we have a large number of clusters because we need many states or periods to estimate the correlation between $v_{st} - \bar{v}_s$ and $v_{st-1} - \bar{v}_s$ reasonably well. A paucity of clusters can lead to biased standard errors and misleading inferences.

8.2.3 *Fewer than 42 Clusters*

Bias from few clusters is a risk in both the Moulton and the serial correlation contexts because in both cases, inference is cluster-based. With few clusters, we tend to underestimate either the serial correlation in a random shock like v_{st} or the intraclass correlation, ρ_e, in the Moulton problem. The relevant dimension for counting clusters in the Moulton problem is the number of groups, G. In a DD scenario where you'd like to cluster on state or some other cross-sectional dimension, the relevant dimension for counting clusters is the number of states or cross-sectional groups. Therefore, following Douglas Adams's dictum that the ultimate answer to life, the universe, and everything is 42, we believe the question is: How many clusters are enough for reliable inference using the standard cluster adjustment derived from (8.2.6)?

If 42 is enough for the standard cluster adjustment to be reliable, and if less is too few, then what should you do when

[16] Arellano (1987) appears to have been the first to suggest higher-level clustering for models with a panel structure.

the cluster count is low? First-best is to get more clusters by collecting more data. But sometimes we're too lazy for that, or the number of groups is naturally fixed, so other ideas are detailed below. It's worth noting at the outset that not all of these ideas are equally well-suited for the Moulton and serial correlation problems.

1. Bias correction of clustered standard errors: Clustered standard errors are biased in small samples because $E(\hat{e}_g\hat{e}_g') \neq E(e_g e_g') = \Psi_g$, just as with the residual covariance matrix in section 8.1. Usually, $E(\hat{e}_g\hat{e}_g')$ is too small. One solution is to inflate residuals in the hopes of reducing bias. Bell and McCaffrey (2002) suggest a procedure (called bias-reduced linearization, or BRL) that adjusts residuals by

$$\hat{\Psi}_g = a\tilde{e}_g\tilde{e}_g'$$
$$\tilde{e}_g = A_g\hat{e}_g$$

where A_g solves

$$A_g'A_g = (I - H_g)^{-1},$$
$$H_g = X_g(X'X)^{-1}X_g',$$

and a is a degrees-of-freedom correction.

This is a version of HC_2 for the clustered case. BRL works for the straight-up Moulton problem with few clusters but for technical reasons cannot be used for the typical DD serial correlation problem.[17]

2. Recognizing that the fundamental unit of observation is a cluster and not an individual unit within clusters, Bell and McCaffrey (2002) and Donald and Lang (2007) suggest that inference be based on a t-distribution with $G - K$ degrees of freedom rather than on the standard normal distribution. For small G, this makes a difference: confidence intervals will be wider, thereby avoiding some mistakes. Cameron, Gelbach, and Miller (2008) report Monte Carlo examples where the combination of a BRL adjustment and the use of t-tables works well.
3. Donald and Lang (2007) argue that estimation using group means works well with small G in the Moulton problem, and even better when inference is based on a t-distribution with $G - K$ degrees of freedom. But, as we discussed in section 8.2.1, for grouped estimation the regressor should be fixed within groups. The level of aggregation is the level at which you'd like to cluster, such as schools in Angrist and Lavy (2008). For serial correlation, this is the state, but state averages cannot be used to estimate a model with a full set of state effects. Also, since treatment status varies within states, averaging up to the state level averages the regressor of interest as well, changing the rules of the game in a way we may not like (the estimator becomes IV using group dummies as instruments). The group means approach is therefore out of bounds for the serial correlation problem. Note also that if the grouped residuals are heteroskedastic, and you therefore use robust standard errors, you may have to worry about bias of the form discussed in section 8.1. In some cases, heteroskedasticity in the grouped residuals can be fixed by weighting by the group size. But weighting changes the estimand when the CEF is nonlinear, so the case for weighting is not open and shut (Angrist and Lavy, 1999, chose not to weight school-level averages because the variation in their study comes mostly from small schools). Weighted or not, a conservative approach when working with group-level averages is to use our rule of thumb from section 8.1: take the larger of robust and conventional standard errors as your measure of precision.

[17] The matrix A_g is not unique; there are many such decompositions. Bell and McCaffrey (2002) use the symmetric square root of $(I - H_g)^{-1}$, or

$$A_g = R\Lambda^{1/2},$$

where R is the matrix of eigenvectors of $(I - H_g)^{-1}$ and $\Lambda^{1/2}$ is the diagonal matrix of the square roots of the eigenvalues. One problem with the Bell and McCaffrey adjustment is that $(I - H_g)$ may not be of full rank, and hence the inverse may not exist for all designs. This happens, for example, when one of the regressors is a dummy variable that is one for exactly one of the clusters, and zero otherwise. This scenario occurs in the panel DD model discussed by Bertrand et al. (2004), which includes a full set of state dummies and clusters by state.

4. Cameron, Gelbach, and Miller (2008) report that some forms of a block bootstrap work well with small numbers of groups, and that the block bootstrap typically outperforms Stata-clustered standard errors. This appears to be true both for the Moulton and serial correlation problems. But Cameron, Gelbach, and Miller (2008) focus on rejection rates using (pivotal) test statistics, while we like to see standard errors.
5. Parametric corrections: For the Moulton problem, this amounts to use of the Moulton factor. With serial correlation, this means correcting your standard errors for first-order serial correlation at the group level. Based on our sampling experiments with the Moulton problem and a reading of the literature, parametric approaches may work well, and better than the nonparametric cluster estimator (8.2.6), especially if the parametric model is not too far off (see, e.g., Hansen, 2007b, which also proposes a bias correction for estimates of serial correlation parameters). Unfortunately, however, beyond the greenhouse world of controlled Monte Carlo studies, we're unlikely to know whether parametric assumptions are a good fit.

Alas, the bottom line here is not entirely clear, nor is the more basic question of when few clusters are fatal for inference. The severity of the resulting bias seems to depend on the nature of your problem, in particular whether you confront straight-up Moulton or serial correlation issues. Aggregation to the group level as in Donald and Lang (2007) seems to work well in the Moulton case as long as the regressor of interest is fixed within groups and there is not too much underlying heteroskedasticity. At a minimum, you'd like to show that your conclusions are consistent with the inferences that arise from an analysis of group averages, since this is a conservative and transparent approach. Angrist and Lavy (2008) use BRL standard errors to adjust for clustering at the school level but validate this approach by showing that key results come out the same using covariate-adjusted group averages.

As far as serial correlation goes, most of the evidence suggests that when you are lucky enough to do research on U.S. states, giving 51 clusters, you are on reasonably safe ground with a naive application of Stata's `cluster` command at the state level. But you might have to study Canada, which offers only 10 clusters in the form of provinces, well below 42. Hansen (2007a) finds that Liang and Zeger (1986) (Stata-clustered) standard errors are reasonably good at correcting for serial correlation in panels, even in the Canadian scenario. Hansen also recommends use of a t-distribution with $G - K$ degrees of freedom for critical values.

Clustering problems have forced applied microeconometricians to eat a little humble pie. Proud of working with large microdata sets, we like to sneer at macroeconomists toying with small time series samples. But he who laughs last laughs best: if the regressor of interest varies only at a coarse group level, such as over time or across states or countries, then it's the macroeconomists who have had the most realistic mode of inference all along.

8.3 Appendix: Derivation of the Simple Moulton Factor

Write

$$y_g = \begin{bmatrix} Y_{1g} \\ Y_{2g} \\ \vdots \\ Y_{n_g g} \end{bmatrix} \qquad e_g = \begin{bmatrix} e_{1g} \\ e_{2g} \\ \vdots \\ e_{n_g g} \end{bmatrix}$$

and

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_G \end{bmatrix} \qquad x = \begin{bmatrix} \iota_1 x_1 \\ \iota_2 x_2 \\ \vdots \\ \iota_G x_G \end{bmatrix} \qquad e = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_G \end{bmatrix},$$

where ι_g is a column vector of n_g ones and G is the number of groups. Note that

$$E(ee') = \Psi = \begin{bmatrix} \Psi_1 & 0 & \cdots & 0 \\ 0 & \Psi_2 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & \Psi_G \end{bmatrix}$$

$$\Psi_g = \sigma_e^2 \begin{bmatrix} 1 & \rho_e & \cdots & \rho_e \\ \rho_e & 1 & & \vdots \\ \vdots & & \ddots & \rho_e \\ \rho & \cdots & \rho_e & 1 \end{bmatrix} = \sigma_e^2 \left[(1-\rho_e)I + \rho_e \iota_g \iota_g' \right],$$

where $\rho_e = \frac{\sigma_v^2}{\sigma_v^2 + \sigma_\eta^2}$.

Now

$$X'X = \sum_g n_g x_g x_g'$$

$$X'\Psi X = \sum_g x_g \iota_g' \Psi_g \iota_g x_g'.$$

But

$$x_g \iota_g' \Psi_g \iota_g x_g' = \sigma_e^2 x_g \iota_g' \begin{bmatrix} 1+(n_g-1)\rho_e \\ 1+(n_g-1)\rho_e \\ \cdots \\ 1+(n_g-1)\rho_e \end{bmatrix} x_g'$$

$$= \sigma_e^2 n_g \left[1+(n_g-1)\rho_e \right] x_g x_g'.$$

Let $\tau_g = 1+(n_g-1)\rho_e$, so we get

$$x_g \iota_g' \Psi_g \iota_g x_g' = \sigma_e^2 n_g \tau_g x_g x_g'$$

$$X'\Psi X = \sigma_e^2 \sum_g n_g \tau_g x_g x_g'.$$

With this in hand, we can write

$$V(\hat\beta) = (X'X)^{-1} X'\Psi X (X'X)^{-1}$$

$$= \sigma_e^2 \Big(\sum_g n_g x_g x_g'\Big)^{-1} \sum_g n_g \tau_g x_g x_g' \Big(\sum_g n_g x_g x_g'\Big)^{-1}.$$

We want to compare this with the standard OLS covariance estimator

$$V_c(\hat\beta) = \sigma_e^2 \Big(\sum_g n_g x_g x_g'\Big)^{-1}.$$

If the group sizes are equal, $n_g = n$ and $\tau_g = \tau = 1+(n-1)\rho_e$, so that

$$V(\hat\beta) = \sigma_e^2 \tau \Big(\sum_g n x_g x_g'\Big)^{-1} \sum_g n x_g x_g' \Big(\sum_g n x_g x_g'\Big)^{-1}$$

$$= \sigma_e^2 \tau \Big(\sum_g n x_g x_g'\Big)^{-1}$$

$$= \tau V_c(\hat\beta),$$

which implies (8.2.4).

EXHIBIT 12

Proving Antitrust Damages

Legal and Economic Issues

Second Edition

SECTION OF ANTITRUST LAW

AMERICAN BAR ASSOCIATION
Defending Liberty
Pursuing Justice

Proving Antitrust Damages

Legal and Economic Issues

Second Edition

SECTION OF ANTITRUST LAW

AMERICAN BAR ASSOCIATION
Defending Liberty
Pursuing Justice

coefficient estimate and $se(\hat{\beta})$ is the standard error of the coefficient estimate. This ratio is called a *t-statistic.*[58]

If the hypothesis is correct and the true underlying coefficient is in fact zero, then the t-statistic should not be very far from zero. If the t-statistic turns out to be far from zero, it would cast doubt on the truth of the hypothesis. How do we determine whether the t-statistic is "far" from zero? We can calculate the probability that the t-statistic achieves a certain value, if the hypothesis were true. For example, if the hypothesis were true, there is about a 90 percent probability that the t-statistic will fall between 1.7 and -1.7 and about a 95 percent probability that the t-statistic will fall between 2 and -2. Thus, if the hypothesis were true, there would be only a 5 percent probability that the t-statistic we observe would be either greater than 2 or less than -2. Accordingly, if we observe a t-statistic greater than 2 or less than -2, the data would appear to be inconsistent with the hypothesis (because such an outcome is quite unlikely if the hypothesis were in fact true).

Indeed, if the absolute value of the t-statistic that the economist calculates exceeds two, then the hypothesis that the true underlying coefficient equals zero typically would be said to be *rejected* at the 5 percent significance level and the result typically would be termed *statistically significant.*[59] This result often is also expressed by saying that the coefficient is "statistically significantly different from zero (at the 5 percent level of significance)." The 5 percent level of significance (and the corresponding 95 percent confidence interval) is often used by economists and statisticians when conducting hypothesis tests, but other levels of significance, such as 1 percent or 10 percent, are also sometimes used.

As an example of these techniques of statistical inference, suppose that the coefficient estimate on the PERIOD variable in the price regression was 0.50, which would imply that prices were $0.50 higher during the alleged conspiracy period as compared to outside that period, holding constant the variables COST and DEMAND (and assuming correct model specification). Suppose further that the standard error of the coefficient estimate on PERIOD is 0.20. In this case, the 95 percent confidence interval would be approximately $0.10 to $0.90—one can be 95 percent confident that the true underlying coefficient on PERIOD lies

[58] *See* GREENE, *supra* note 16, at 249.

[59] *Id.*

within this interval, assuming the validity and reliability of the econometric model and approach.

Similarly, the standard error can be used to test the hypothesis that the coefficient on the PERIOD value is zero, which implies that price was not higher during the alleged conspiracy period (holding constant COST and DEMAND). This test would be conducted by calculating the t-statistic: 0.50/0.20 = 2.50. Since the calculated t-statistic of 2.50 is greater than two, the coefficient on PERIOD would be said to be statistically significantly different from zero at the 5 percent level of significance, and the hypothesis that the price was no higher during the alleged conspiracy period would be rejected.

4. Estimation of the Standard Errors

Correct statistical inference requires not only good estimates of the coefficients of the model, but also good estimates of the standard errors of these coefficient estimates.[60] For example, a t-test of whether a coefficient β is zero is conducted by forming the t-statistic "$\hat{\beta}/ se(\hat{\beta})$," discussed in the last section. This t-statistic can be invalid and lead to incorrect statistical inference if the standard error of the coefficient estimate, $se(\hat{\beta})$, is itself inconsistently estimated.

Consistent estimation of the standard errors requires that the properties of the error term of the regression be properly taken into account. For example, standard errors frequently are estimated assuming the error term of the regression is *independently and identically distributed.* That is, the error for each observation reflecting the impact of unmeasured factors is assumed to be from the same distribution (or pattern) of possible errors, and each error is statistically independent of the others (they are not correlated to each other).[61] If in fact the error terms are *correlated* with each other (i.e., not independent) or not identically distributed, then the resulting standard error estimates generally will be inconsistent.[62]

Correlation among the errors of different observations can arise in various situations. For example, suppose the data sample is a *time series*, i.e., the data were generated by observing the variables (e.g., price, cost,

[60] *See* WOOLDRIDGE, *supra* note 3, at 57.

[61] *Id.* at 54-57.

[62] *Id.*

and demand) at various points over time (e.g., on a monthly basis).[63] In such a case, the error in one month might well be related to the errors in adjacent months, since the unobserved economic factors that appear in the error term might themselves exhibit correlation over time. This correlation of errors over time is called *serial correlation.*[64]

As another example, suppose the data sample is a *cross-section/time series*, or *panel data* set, where the variables are observed at various points of time separately for each of a number of units of observation such as individual customers.[65] Each customer's data are a time series. Therefore, the error terms for a given customer may exhibit serial correlation. In addition, each customer may have idiosyncratic factors that affect the price it paid, but are unobserved in the data. These factors would be present in all of the errors across time for that customer, which would be a further cause of correlation among the errors for a given customer. This effect is called an *unobserved individual-specific effect* (where the "individual" refers to the unit of observation, e.g., a customer).[66]

Importantly, the correlation among errors need not be confined to errors that pertain to the same customer. For example, the error terms for all customers within the same time period also may be correlated. Unobserved economic factors may affect all customers' prices at a given point in time and therefore these common factors will appear in the errors of all of the customers in a given time period. Similarly, if these unobserved factors are themselves serially correlated, then the error for one customer in one month will be correlated with the error for another customer in another month. Therefore, there may be correlation among the errors both within and between units of observation in a panel data set.[67]

There can be substantial consequences from estimating the standard errors for the coefficient estimates as if the errors were uncorrelated when they are in fact correlated. With positive correlation between the error terms, the incorrectly estimated standard errors generally will be

[63] *See* GREENE, *supra* note 16, at 97.

[64] *Id.* at 525.

[65] *Id.* at 98.

[66] *See* WOOLDRIDGE, *supra* note 3, at 248.

[67] This problem is widely recognized in the econometrics literature. *See* Brent R. Moulton, *An Illustration of the Pitfall in Estimating the Effects of Aggregate Variables on Micro Units*, 72 REV. ECON. STAT. 334 (1990); Marianne Bertrand et al., *How Much Should We Trust Differences-in-Differences Estimates?*, 119 Q. J. OF ECON. 249 (2004).

biased downward, making the regression coefficients seem to be more precisely estimated than they really are. As a result, a statistical test on the coefficients may yield what appears to be a statistically significant result but is not.[68]

To see why this is so, suppose that the estimate of the coefficient on a price-fixing conspiracy dummy variable is 0.15 and that the standard error, estimated by incorrectly ignoring correlation among the errors, is 0.05. The (incorrect) t-statistic would then be 0.15/0.05 = 3 and the hypothesis that the alleged conspiracy had no effect on prices would be strongly rejected. But, the standard error is too low because it did not account for correlation in the error terms. Suppose that the standard error is re-estimated correctly to account for correlation in the error terms, and this correct standard error is 0.15. Then, the correctly calculated t-statistic is only 0.15/0.15 = 1. Because the correct t-statistic is less than two, the hypothesis that the alleged conspiracy had no effect would not be rejected at conventional levels of statistical significance. If the standard error had not been corrected, the wrong inference would have been made.

Econometricians have a variety of methods for consistently estimating the standard errors when correlation among the errors exists. In a time series context (discussed in more detail below), various non-parametric procedures (procedures that do not impose any functional form on the correlation) have become widely used.[69] In a panel data context (also discussed in more detail below) these procedures may be used, and other methods have been proposed as well.[70] Some of these panel data procedures are easily implemented.[71]

[68] *Id.*

[69] *See* Donald W. Andrews, *Autocorrelation and Heteroskedasticity Consistent Covariance Matrix Estimation*, 59 ECONOMETRICA 817 (1991); Donald W. Andrews & J. Christopher Monahan, *An Improved Autocorrelation and Heteroskedasticity Consistent Covariance Matrix Estimator*, 60 ECONOMETRICA 953 (1992). The resulting standard errors are also consistent in the presence of *heteroskedasticity* (the variance of the error term differs across observations). Many statistical software packages implement the Newey-West procedure, which is one example of such a procedure to obtain autocorrelation and heteroskedasticity consistent standard errors. *See* Whitney Newey & Kenneth West, *A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix*, 55 ECONOMETRICA 703 (1987).

[70] *See* Bertrand et al., *supra* note 67.

[71] For example, Stata, a popular econometrics software package, includes a "cluster" option for calculating standard errors assuming unspecified

These procedures produce consistent estimates of the standard errors even when there is no correlation among the error terms. In other words, they work well in both situations. Thus, these procedures have become used more generally in practice.[72] When there is good reason to suspect the existence of correlation among the errors, such procedures should be used to avoid making incorrect statistical inferences.[73]

Finally, heteroskedasticity can also create problems in accurately measuring standard errors. Heteroskedasticity occurs when the variance of the error term varies across observations.[74] This condition is another violation of the independent and identically distributed error term assumption that can cause the traditional standard error calculation to be inconsistent. Again a well-known and widely used technique exists for calculating standard errors that are robust to heteroskedasticity (White standard errors).[75] This technique is also easily implemented in many econometric packages.[76]

5. Choice of Explanatory Variables

As discussed above, the explanatory variables in an econometric model represent economic factors that influence the dependent variable.[77] An important question is which explanatory variables to include in the model. Answering this question should begin with economic theory combined with qualitative knowledge about the industry. For example, if the dependent variable is price, economic theory suggests that demand drivers, cost factors, and industry capacity, among other things, are potential explanatory variables.[78] Industry knowledge would suggest specific variables that would appropriately represent these factors. If a

within-group (cluster) correlation between the error terms. *See* 3 STATA PRESS, BASE REFERENCE MANUAL 81 (2007).

[72] *See* GREENE, *supra* note 16, at 465 ("The . . . Newey-West estimator [is] becoming ubiquitous in the econometrics literature").

[73] If one is not sure, there are various tests one can run to test the hypothesis of no correlation in the error terms. *See* WOOLDRIDGE, *supra* note 3, at 130, 279, 420-449; GREENE, *supra* note 16, at 538-42.

[74] *See* GREENE, *supra* note 16, at 499.

[75] Hal White, *A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity*, 48 ECONOMETRICA 817, 838 (1980).

[76] For example, Stata includes an option for calculating White standard errors. *See* STATA, *supra* note 71, at 81.

[77] *See* part A of this chapter.

[78] *See* Baker & Rubinfeld, *supra* note 2, at 391.

product is used as an input by downstream industries, the level of production in those industries might drive the demand for the product.[79] The prices of the inputs used to produce the product could be important to costs.

a. Too Many or Too Few Variables?

The number of explanatory variables suggested by economic theory and industry knowledge often will be large. Is it best to include all of the explanatory variables, or should one try to pare back the number of variables in order to have a simpler model? The downside to including extraneous explanatory variables in a regression is that the coefficients may be less precisely estimated. However, these estimates will still be unbiased.[80] Moreover, the effect on precision of having additional variables will often be small when the sample size is large.

Mistakenly excluding important explanatory variables in an attempt at simplicity, on the other hand, can result in an *omitted variable bias.* Omitted variable bias arises when important explanatory variables that have been omitted from the regression model are correlated with included explanatory variables. Because the omitted variables are in the error term, the result will be a correlation between the included explanatory variables and the error term.[81] This misspecification will bias the resulting coefficient estimates, and make these estimates

[79] Care needs to be taken that the explanatory variables used in a least squares regression are *exogenous*, or uncorrelated with the error term, to the extent possible. *See* WOOLDRIDGE, *supra* note 3, at 50-51. For example, if the intermediate product in question represents a large share of the downstream industry's costs, the amount of downstream production may be affected by the price of the intermediate good. If the impact of the price of the intermediate good on the sales of the downstream product is substantial, then downstream production could be considered *endogenous* (correlated with the error term) rather than exogenous. *See* Baker & Rubinfeld, *supra* note 2, at n.17. Methods of detecting and dealing with endogeneity are discussed later in this chapter.

[80] More specifically, ordinary least squares is still the best linear unbiased estimator as demonstrated by the Gauss-Markov Theorem, one of the more famous theorems in statistics. *See* GREENE, *supra* note 16, at 246. This means that not only is least squares unbiased, but also it is the most efficient (i.e., most precise) among linear unbiased estimators.

[81] *See* WOOLDRIDGE, *supra* note 3, at 61-62.

EXHIBIT 13

Cameron, A. Colin; Miller, Douglas L.

Working Paper

Robust inference with clustered data

Working Papers, University of California, Department of Economics, No. 10,7

Provided in Cooperation with:
University of California, Davis, Department of Economics

UCDAVIS
DEPARTMENT OF ECONOMICS
Working Paper Series

Robust Inference with Clustered Data

A. Colin Cameron
Douglas L. Miller

April 06, 2010

Paper # 10-7

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.

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

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

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 $\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y}$. 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 $\rho \rightarrow 1$ 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

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

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 $G \to \infty$.

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 $G \to \infty$.

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 $N \to \infty$ 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 $N \to \infty$ and $T \to \infty$ (a long panel with $N_* \to \infty$), 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 $G \rightarrow \infty$. 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 $\mathbf{\Omega}_g \neq \sigma^2\mathbf{I}$ 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 $\sum_i \sum_j w\left(i,j\right) \mathbf{x}_i \mathbf{x}_j' \widehat{u}_i \widehat{u}_j$. For multi-way clustering the weight $w\left(i,j\right) = 1$ for observations who share a cluster, and $w\left(i,j\right) = 0$ otherwise. In White and Domowitz (1984), the weight $w\left(i,j\right) = 1$ for observations "close" in time to one another, and $w\left(i,j\right) = 0$ for other observations. Conley (1999) considers the case where observations have spatial locations, and has weights $w\left(i,j\right)$ 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 $T \to \infty$. The Driscoll-Kraay estimator can be thought of as using weight $w\left(i,j\right) = 1 - D\left(i,j\right)/(D_{\max} + 1)$, where $D\left(i,j\right)$ 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 $w\left(i,j\right) = \mathbf{1}[i,j$ share a firm, or $D\left(i,j\right) \leq D_{\max}]$. 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 $(\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\,\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{y}$, 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 $\mathbf{W}_g = \text{Diag}[\mathbf{w}_{kg}]$ 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 $\widehat{\mathbf{D}}_g = \partial \mathbf{m}_g'(\boldsymbol{\beta})/\partial\boldsymbol{\beta}\big|_{\widehat{\boldsymbol{\beta}}}$, $\widehat{\mathbf{D}} = [\widehat{\mathbf{D}}_1,...,\widehat{\mathbf{D}}_G]'$, $\widehat{\mathbf{u}}_g = \mathbf{y}_g - \mathbf{m}_g(\widehat{\boldsymbol{\beta}})$, 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 $G \to \infty$.

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 $(\widehat{\mathbf{D}}'\widehat{\boldsymbol{\Omega}}^{-1}\widehat{\mathbf{D}})^{-1}$.

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

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

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

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

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 $G \to \infty$ then there is an incidental parameters problem, as there are N_g parameters $\alpha_1, ..., \alpha_G$ to estimate and $G \to \infty$. 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

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

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 $\widehat{\boldsymbol{\theta}}_{\text{GMM}}$ 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

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

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 $G \to \infty$. 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.

EXHIBIT 14

UNITED STATES DISTRICT COURT
SOUTHERN DISTRICT OF NEW YORK
------------------------------x

UNITED STATES OF AMERICA,

Plaintiff,

v. 12 Civ. 2826 (DLC)

APPLE, INC., *et al.*,

Defendants.

------------------------------x

May 23, 2013
2:30 p.m.

Before:

HON. DENISE L. COTE,

District Judge

APPEARANCES

UNITED STATES DEPARTMENT OF JUSTICE
Attorneys for Plaintiff
BY: MARK W. RYAN
DANIEL McCUAIG
LAWRENCE BUTERMAN
CARRIE SYME

OFFICE OF THE ATTORNEY GENERAL OF TEXAS
Attorneys for State of Texas and Liaison counsel for plaintiff States
BY: ERIC LIPMAN
GABRIEL R. GERVEY
DAVID M. ASHTON

OFFICE OF THE ATTORNEY GENERAL OF CONNECTICUT
Attorneys for State of Connecticut and Liaison counsel for plaintiff States
BY: W. JOSEPH NIELSEN
GARY M. BECKER

OFFICE OF THE ATTORNEY GENERAL OF OHIO
Attorneys for State of Ohio
BY: EDWARD J. OLSZEWSKI

GIBSON, DUNN & CRUTCHER
Attorneys for Defendant Apple
BY: ORIN SNYDER
LISA RUBIN
DANIEL FLOYD
DANIEL SWANSON
CYNTHIA RICHMAN
-and-
O'MELVENEY & MYERS
BY: HOWARD HEISS

(In open court)

THE DEPUTY CLERK: United States of America v. Apple Inc. and others.

Counsel for the government, please state your name for the record.

MR. RYAN: Mark Ryan for the United States, your Honor. Good afternoon.

THE DEPUTY CLERK: For the plaintiff.

THE COURT: Excuse me one second. Anyone else for the United States?

MR. BUTERMAN: Good afternoon, your Honor. Lawrence Buterman.

MR. MCCUAIG: Daniel McCuaig, your Honor.

MS. SYME: Carrie Syme, your Honor.

THE COURT: For the plaintiff States. For Texas.

MR. LIPMAN: Good afternoon, your Honor. Eric Lipman.

MR. GERVEY: Good afternoon, your Honor. Gabriel Gervey.

MR. ASHTON: David Ashton, your Honor.

THE COURT: For Connecticut.

MR. NIELSEN: Joe Nielsen, your Honor.

MR. BECKER: Gary Becker, your Honor.

THE COURT: Anyone else for the States?

MR. OLSZEWSKI: Yes. Edward Olszewski for Ohio, Attorney General's Office.

THE COURT: For Apple.

MR. SNYDER: Good afternoon. Orin Snyder for Apple.

MS. RUBIN: Good afternoon, your Honor. Lisa Rubin for Apple.

MR. SWANSON: Good afternoon, your Honor. Dan Swanson for Apple.

MS. RICHMAN: Good afternoon, your Honor. Cynthia Richman for Apple.

MR. HEISS: Good afternoon, your Honor. Howard Heiss for Apple.

MR. FLOYD: Good afternoon, your Honor. Daniel Floyd for Apple.

THE COURT: Is there anyone else who needs to place their appearance on the record?

MR. SNYDER: No, your Honor. Thank you.

THE COURT: Thank you, Mr. Snyder.

To assist our court reporter, and me, perhaps, I am going to ask you if you speak please to identify yourself briefly for the record before you speak.

Welcome, everyone. This is our final pretrial conference. We have a long agenda today to get ourselves ready for our trial which begins on June 3rd, as you know. I want to address the following topics, and you may have some additional issues as well. I want to talk about the schedule we will follow during the trial and the procedures generally that will apply in this non-jury trial. I want to go through your witness list, make sure we understand who is actually going to be called to testify and clarify who is going to be in the courtroom and subject to cross-examination. I want to talk about time limits and whether those are appropriate here. We have motions in limine that I am prepared to address. I want to talk about the state law claims and the extent to which they will be litigated and under what standard. I want to talk about the depositions and the way we are going to approach deposition evidence that parties have offered as part of their pretrial order. I want to deal with objections to exhibits, including potentially authenticity objections. I want to talk about third-party redactions. We have gotten some submissions there and I want to make sure we know what procedure we are going to follow with respect to those.

So then, of course, I won't end this conference without -- and maybe I will just start this conference that way. I am working hard. My staff is working hard on this case. I am sure counsel is working hard on this case to be prepared for our June 3rd trial. So if for any reason this case settles and I can put down my pen and turn to something else, I would like a call, night or day, at the chamber's telephone number because it will affect how I spend my time. So thank you so much for that.

So let's talk about our schedule. We will begin at