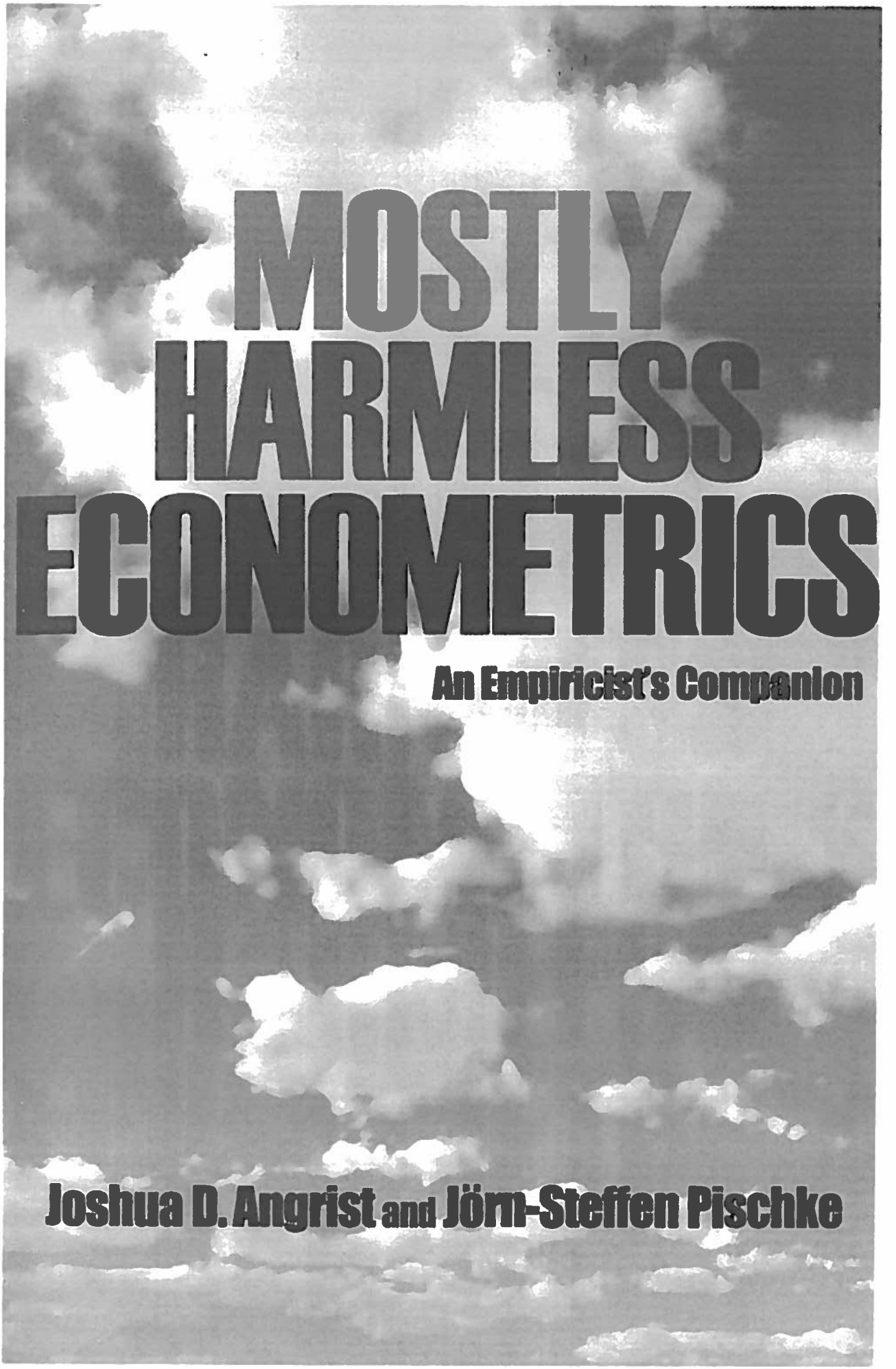
Exhibit 4


MOSTLY
HARMLESS
ECONOMETRICS
An Empiricist's Companion
Joshua D. Angrist and Jörn-Steffen Pischke

Published by Princeton University Press, 41 William Street,
Princeton, New Jersey 08540
In the United Kingdom: Princeton University Press,
6 Oxford Street, Woodstock, Oxfordshire OX20 1TW

Library of Congress Cataloging-in-Publication Data

Angrist, Joshua David.
Mostly harmless econometrics : an empiricist's companion /
Joshua D. Angrist, Jörn-Steffen Pischke.
p. cm.
Includes bibliographical references and index.
ISBN 978-0-691-12034-8 (hardcover : alk. paper) —
ISBN 978-0-691-12035-5 (pbk. : alk. paper) 1. Econometrics.
2. Regression analysis. I. Pischke, Jörn-Steffen. II. Title.
HB139.A54 2008
330.01'5195—dc22 2008036265

British Library Cataloging in-Publication Data is available

This book has been composed in Sabon
with Hel. Neue Cond. family display

Illustrations by Karen Norberg

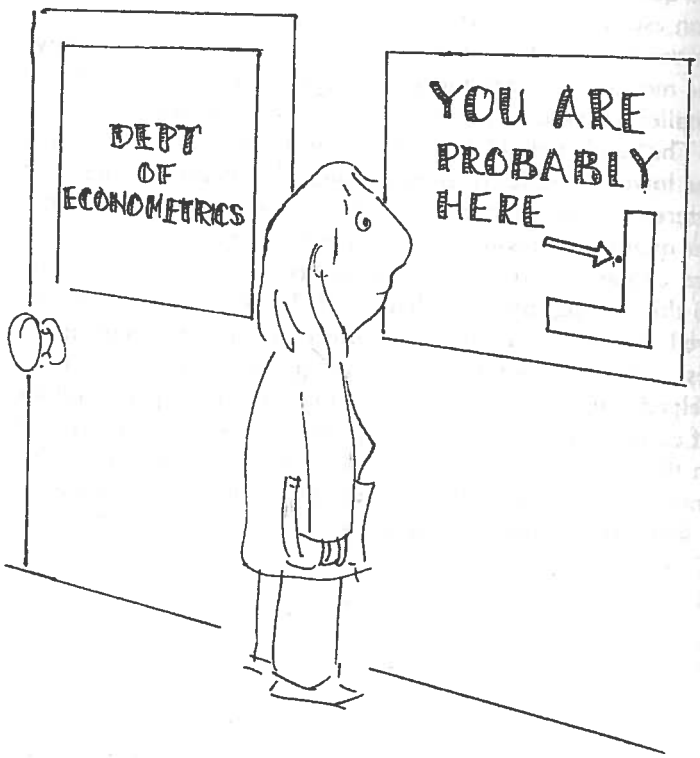
Printed on acid-free paper. ∞

press.princeton.edu

Printed in the United States of America

5 7 9 10 8 6

CONTENTS



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 $\hat{e}_i = \mathrm{Y}_i - \mathrm{X}_i'\hat{\beta}$ 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

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

We can think of the middle term as an estimator of the form $\sum \frac{\mathrm{X}_i\mathrm{X}_i'\hat{\psi}_i}{N}$, 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 $h_i = X(X'X)^{-1}\mathrm{X}_i$, the ith column of the projection matrix $H = X(X'X)^{-1}X'$. Then

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

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

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

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

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

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,

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

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 $\sum_{i=1}^{N} h_{ii} = \mathrm{K}$, the number of regressors (see, e.g., Hoaglin and Welsch, 1978).[1]

[1]The property $\sum_{i=1}^{N} h_{ii} = \mathrm{K}$ 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:

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

Thus, the bias in $\hat{\Omega}_c$ can be fixed by a simple degrees-of-freedom correction: divide by $N - \kappa$ 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

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

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 the single regressor, $\tilde{x}_i$, is assumed to be in deviations-from-means form, so there is a single coefficient. In this case, the estimator of interest is $\hat{\beta}_1 = \frac{\sum \tilde{x}_i y_i}{\sum \tilde{x}_i^2}$ and the leverage is

[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.

$h_{ii} = \frac{\tilde{x}_i^2}{\sum \tilde{x}_i^2}$ (we lose the $\frac{1}{N}$ term in (8.1.7) by partialing out the constant). Let $s_x^2 = \frac{\sum \tilde{x}_i^2}{N}$. For the conventional covariance estimator, we have

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

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

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

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:

$$HC_0 : \hat{\psi}_i = \hat{e}_i^2$$

$$HC_1 : \hat{\psi}_i = \frac{N}{N - \kappa} \hat{e}_i^2$$

[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 $(\tilde{x}_i)^2$ 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 $\{\mathrm{Y}_i, \mathrm{X}_i\}$ 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

[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 $\mathrm{X}_i/\sqrt{\hat{\psi}_i}$ for your preferred choice of $\hat{\psi}_i$.

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 $\mathrm{X}_i'\hat{\beta} + \hat{e}_i$ (which, of course, is just the original Y_i) with probability 0.5, and $\mathrm{X}_i'\hat{\beta} - \hat{e}_i$ 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

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

$$Y_i = \beta_0 + \beta_1 D_i + \varepsilon_i, \tag{8.1.9}$$

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

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

For the purposes of this derivation we think of D_i as nonrandom, so that $\sum D_i = N_1$ and $\sum (1 - D_i) = N_0$ 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 $D_i = 1$ and $D_i = 0$ 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 $D_i = 1$ and $D_i = 0$ 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 $S_j^2 = \sum_{D_i=j} (Y_i - \overline{Y}_j)^2$ for $j = 0, 1$. The leverage in this example is

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

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

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

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

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

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

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

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

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 $D_i = 1$ and $h_{ii} = \frac{1}{27}$ if $D_i = 0$. This is a highly unbalanced design. We draw residuals from the distributions:

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

and report results for three cases. The first has lots of heteroskedasticity, with $\sigma = 0.5$, while the second has relatively

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

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.

[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

Parameter Estimate	Mean (1)	Standard Deviation (2)	Empirical 5% Rejection Rates: Normal (3)	Empirical 5% Rejection Rates: t (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,

$$Y_{ig} = \beta_0 + \beta_1 x_g + e_{ig}, \tag{8.2.1}$$

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,

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

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,

$$e_{ig} = v_g + \eta_{ig}, \tag{8.2.3}$$

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

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

where σ_ν^2 is the variance of ν_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]

[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).

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

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

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

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

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.

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

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

[12]Use Stata's `oneway` 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

$$\mathrm{Y}_{ig} = \beta_0 + \beta_1 x_g + \beta_2 \mathrm{W}_{ig} + e_{ig}, \tag{8.2.7}$$

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

$$\mathrm{Y}_{ig} = \mu_g + \beta_2 \mathrm{W}_{ig} + \eta_{ig}.$$

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:

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

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 $\mathrm{G} - \mathrm{K}$ 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:

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

[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."

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, υ_{st}, and an idiosyncratic individual component, η_{ist}. So we have:

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

We assume that in repeated draws across states and over time, $E[\upsilon_{st}] = 0$, 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

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

This estimator is unbiased, since $E[\upsilon_{st}] = E[\eta_{ist}] = 0$. 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:

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

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

change from the difference-in-dfferences 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]

[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.

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.

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, 2007a, 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 (2007b) 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

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

and

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

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).