Robust Standard Errors in Small Samples: Some Practical Advice*

Guido W. Imbens† Michal Kolesár‡

First Draft: October 2012
This Draft: December 2014

Abstract

In this paper we discuss the properties of confidence intervals for regression parameters based on robust standard errors. We discuss the motivation for a modification suggested by Bell and McCaffrey (2002) to improve the finite sample properties of the confidence intervals based on the conventional robust standard errors. We show that the Bell-McCaffrey modification is a natural extension of a principled approach to the Behrens-Fisher problem, and suggest a further improvement for the case with clustering. We show that these standard errors can lead to substantial improvements in coverage rates even for samples with fifty or more clusters. We recommend researchers routinely calculate the Bell-McCaffrey degrees-of-freedom adjustment to assess potential problems with conventional robust standard errors.

JEL Classification: C14, C21, C52

Keywords: Behrens-Fisher Problem, Robust Standard Errors, Small Samples, Clustering

*Financial support for this research was generously provided through NSF grant 0820361.

†Graduate School of Business, Stanford University, and NBER. Electronic correspondence: imbens@stanford.edu.

‡ Department of Economics and Woodrow Wilson School, Princeton University. Electronic correspondence: mkolesar@princeton.edu.

1 Introduction

It is currently common practice in empirical work to use standard errors and associated confidence intervals that are robust to the presence of heteroskedasticity. The most widely used form of the robust, heteroskedasticity-consistent standard errors is that associated with the work of White (1980) (see also Eicker (1967); Huber (1967)), extended to the case with clustering by Liang and Zeger (1986). The justification for these standard errors and the associated confidence intervals is asymptotic: they rely on large samples for their validity. In small samples the properties of these procedures are not always attractive: the robust (Eicker-Huber-White, or EHW, and Liang-Zeger or LZ, from hereon) variance estimators are biased downward, and the Normal-distribution-based confidence intervals using these variance estimators can have coverage substantially below nominal coverage rates.

There is a large theoretical literature documenting and addressing these small sample problems in the context of linear regression models, some of it reviewed in MacKinnon and White (1985), Angrist and Pischke (2009, Chapter 8), and MacKinnon (2012). A number of alternative versions of the robust variance estimators and confidence intervals have been proposed to deal with these problems. Some of these alternatives focus on reducing the bias of the variance estimators (MacKinnon and White, 1985), some exploit higher order expansions (Hausman and Palmer, 2011), others attempt to improve their properties by using resampling methods (Davidson and Flachaire, 2008; Cameron, Gelbach, and Miller, 2008; Hausman and Palmer, 2011), or data-partitioning (Ibragimov and Müller, 2010), and some use t-distribution approximations (Bell and McCaffrey, 2002; Donald and Lang, 2007). Given the multitude of alternatives, combined with the *ad hoc* nature of some of them, it is not clear, however, how to choose among them. Moreover, some researchers (e.g. Angrist and Pischke, 2009, Chapter 8.2.3) argue that for commonly encountered sample sizes—fifty or more units / fifty or more clusters—using these alternatives is not necessary because the EHW and LZ standard errors perform well.

We make three specific points in this paper. First, we show that a particular improvement to the EHW and LZ confidence intervals, due to Bell and McCaffrey (2002, BM from hereon), is a principled extension of an approach developed by Welch (1951) to a simple, much-studied and well-understood problem, known as the Behrens-Fisher problem (see for a general discussion, Scheffé (1970)). Understanding how the BM proposals and other procedures perform in the simple Behrens-Fisher case provides insights into their general performance. The BM improvement is simple to implement and in small and moderate-sized samples can provide a considerable improvement over the EHW and LZ confidence intervals. We recommend that empirical researchers should, as a matter of routine, use the BM confidence intervals rather than the EHW and LZ confidence intervals.

Second, and this has been pointed out in the theoretical literature before (e.g. Chesher and Jewitt, 1987), without having been appreciated in the empirical literature, problems with the standard robust EHW and LZ variances and confidence intervals can be substantial even with moderately large samples (such as 50 units / clusters) if the distribution of the regressors is skewed. It is the combination of the sample size and the distribution of the regressors that determines the accuracy of the standard robust confidence intervals and the potential benefits from small-sample adjustments.

Third, we suggest a modification of the BM procedure in the case with clustering that further improves the performance of confidence intervals in that case.

Let us briefly describe the BM improvement. Let $\hat{\mathbb{V}}_{\text{EHW}}$ be the standard EHW variance estimator, and let the EHW 95% confidence interval for a parameter β be $\hat{\beta} \pm 1.96\sqrt{\hat{\mathbb{V}}_{\text{EHW}}}$. The BM modification consists of two components, the first removing some of the bias and the second changing the approximating distribution from a Normal distribution to the best fitting t-distribution. First, the commonly used variance estimator $\hat{\mathbb{V}}_{\text{EHW}}$ is replaced by $\hat{\mathbb{V}}_{\text{HC2}}$ (a modification for the general case first proposed by Horn, Horn, and Duncan (1975)), which removes some, and in special cases all, of the bias in $\hat{\mathbb{V}}_{\text{EHW}}$ relative to the true variance $\mathbb{V}$. Second, the distribution of $(\hat{\beta} - \beta)/\sqrt{\hat{\mathbb{V}}_{\text{HC2}}}$ is approximated by a t-distribution. When t-distribution approximations are used in constructing robust confidence intervals, the degrees of freedom (dof) are typically fixed at the number of observations minus the number of estimated regression parameters. The BM dof choice for the approximating t-distribution, denoted K_{BM}, is more sophisticated. It is chosen so that under homoskedasticity the distribution of $K_{\text{BM}} \cdot \hat{\mathbb{V}}_{\text{HC2}}/\mathbb{V}$ has the first two moments in common with a chi-squared distribution with dof equal to K_{BM}, and it is a simple analytic function of the matrix of regressors. To convert the dof adjustment into a procedure that only adjusts the standard errors, we can define the BM standard error as $\sqrt{\hat{\mathbb{V}}_{\text{BM}}} = \sqrt{\hat{\mathbb{V}}_{\text{HC2}}} \cdot (t_{0.975}^{K_{\text{BM}}}/1.96)$, where t_q^K is the q-th quantile of the t-distribution with K dof. A key insight is that K_{BM} can differ substantially from the sample size (minus the number of estimated parameters) if the distribution of the regressors is skewed.

This paper is organized as follows. In the next section we study the Behrens-Fisher problem and the solutions offered by the robust standard error literature specialized to this case. In Section 3 we generalize the results to the general linear regression case, and in Section 4 we study the case with clustering. Along the way, we provide some simulation evidence regarding the performance of the various confidence intervals, using designs previously proposed in the literature. We find that in all these settings the BM proposals perform well relative to the other procedures. Section 5 concludes.

2 The Behrens-Fisher problem: performance of various proposed solutions

In this section we review the Behrens-Fisher problem, which can be viewed as a special case of linear regression with a single binary regressor. For this special case there is a large literature and several attractive methods for constructing confidence intervals with good properties even in very small samples have been proposed. See Behrens (1929), Fisher (1939), and for a general discussion Scheffé (1970), Wang (1971), Lehmann and Romano (2005), and references therein. We discuss the form of the standard variance estimators for this case, and discuss when they perform poorly relative to the methods that are designed especially for this setting.

2.1 The Behrens-Fisher problem

Consider a heteroscedastic linear model with a single binary regressor,

$$Y_i = \beta_0 + \beta_1 \cdot D_i + \varepsilon_i, \tag{2.1}$$

where $D_i \in \{0, 1\}$, $i = 1, \ldots, N$ indexes units, and

$$\mathbb{E}[\varepsilon_i \mid D_i = d] = 0, \quad \text{and} \quad \operatorname{var}(\varepsilon_i \mid D_i = d) = \sigma^2(d).$$

We are interested in $\beta_1 = \operatorname{cov}(Y_i, D_i)/\operatorname{var}(D_i) = \mathbb{E}[Y_i \mid D_i = 1] - \mathbb{E}[Y_i \mid D_i = 0]$. Because the regressor D_i is binary, the least squares estimator for the slope coefficient β_1 is given by a difference between two means,

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

where, for $d = 0, 1$,

$$\overline{Y}_d = \frac{1}{N_d} \sum_{i:D_i=d} Y_i, \qquad \text{and} \qquad N_1 = \sum_{i=1}^{N} D_i, \quad N_0 = \sum_{i=1}^{N} (1 - D_i).$$

The estimator $\hat{\beta}_1$ is unbiased, and, conditional on $\mathbf{D} = (D_1, \ldots, D_N)'$, its exact finite sample variance is

$$\mathbb{V} = \text{var}(\hat{\beta}_1 \mid \mathbf{D}) = \frac{\sigma^2(0)}{N_0} + \frac{\sigma^2(1)}{N_1}.$$

If, in addition, we assume Normality for ε_i given D_i, $\varepsilon_i \mid D_i = d \sim \mathcal{N}(0, \sigma^2(d))$, the exact distribution for $\hat{\beta}_1$ conditional on $\mathbf{D}$ is Normal, $\hat{\beta}_1 \mid \mathbf{D} \sim \mathcal{N}(\beta_1, \mathbb{V})$.

The problem of how to do inference for β_1 in the absence of knowledge of $\sigma^2(d)$ is old, and known as the Behrens-Fisher problem. Let us first review a number of the standard least squares variance estimators, specialized to the case with a single binary regressor.

2.2 Homoskedastic variance estimator

Suppose the errors are homoskedastic, $\sigma^2 = \sigma^2(0) = \sigma^2(1)$, so that the exact variance for $\hat{\beta}_1$ is $\mathbb{V} = \sigma^2(1/N_0 + 1/N_1)$. We can estimate the common error variance σ^2 as

$$\hat{\sigma}^2 = \frac{1}{N-2} \sum_{i=1}^{N} \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 \cdot D_i \right)^2.$$

This variance estimator is unbiased for σ^2, and as a result the estimator for the variance for $\hat{\beta}_1$,

$$\hat{\mathbb{V}}_{\text{homo}} = \frac{\hat{\sigma}^2}{N_0} + \frac{\hat{\sigma}^2}{N_1},$$

is unbiased for the true variance $\mathbb{V}$. Moreover, under Normality of ε_i given D_i, the t-statistic $(\hat{\beta}_1 - \beta_1)/\sqrt{\hat{\mathbb{V}}_{\text{homo}}}$ has an exact t-distribution with $N - 2$ degrees of freedom (dof). Inverting the t-statistic yields an exact 95% confidence interval for $\hat{\beta}_1$ under homoskedasticity,

$$\text{CI}^{95\%}_{\text{homo}} = \left(\hat{\beta}_1 - t^{N-2}_{0.975} \times \sqrt{\hat{\mathbb{V}}_{\text{homo}}},\ \hat{\beta}_1 + t^{N-2}_{0.975} \times \sqrt{\hat{\mathbb{V}}_{\text{homo}}} \right),$$

where t^N_q is the q-th quantile of a t-distribution with dof equal to N. This confidence interval is exact under these two assumptions, Normality and homoskedasticity.

2.3 Robust EHW variance estimator

The familiar form of the robust Eicker-Huber-White (EHW) variance estimator, given the linear model (2.1), is

$$\left(\sum_{i=1}^{N} X_i X_i'\right)^{-1} \left(\sum_{i=1}^{N} \left(Y_i - X_i\hat{\beta}\right)^2 X_i X_i'\right) \left(\sum_{i=1}^{N} X_i X_i'\right)^{-1},$$

where $X_i = (1, D_i)'$. In the Behrens-Fisher case with a single binary regressor the component of this matrix corresponding to β_1 simplifies to

$$\hat{\mathbb{V}}_{\text{EHW}} = \frac{\tilde{\sigma}^2(0)}{N_0} + \frac{\tilde{\sigma}^2(1)}{N_1}, \quad \text{where} \quad \tilde{\sigma}^2(d) = \frac{1}{N_d} \sum_{i:D_i=d}^{N} \left(Y_i - \overline{Y}_d\right)^2, \quad d = 0, 1. \tag{2.2}$$

The estimators $\tilde{\sigma}^2(d)$ are downward-biased in finite samples, and so $\hat{\mathbb{V}}_{\text{EHW}}$ is also a downward-biased estimator of the variance. Using a Normal approximation to the t-statistic based on this variance estimator, we obtain the standard EHW 95% confidence interval,

$$\text{CI}^{95\%}_{\text{EHW}} = \left(\hat{\beta}_1 - 1.96 \times \sqrt{\hat{\mathbb{V}}_{\text{EHW}}}, \hat{\beta}_1 + 1.96 \times \sqrt{\hat{\mathbb{V}}_{\text{EHW}}}\right). \tag{2.3}$$

The justification for the Normal approximation is asymptotic even if the error term ε_i has a Normal distribution, and requires both $N_0, N_1 \to \infty$. Sometimes researchers use a t-distribution with $N-2$ dof to calculate the confidence limits, replacing 1.96 in (2.3) by $t^{N-2}_{0.975}$. However, there are no assumptions under which this modification has exact 95% coverage

2.4 Unbiased variance estimator

An alternative to $\hat{\mathbb{V}}_{\text{EHW}}$ is what MacKinnon and White (1985) call the HC2 variance estimator, which we denote by $\hat{\mathbb{V}}_{\text{HC2}}$. In general, this correction removes only part of the bias, but in the single binary regressor (Behrens-Fisher) case the MacKinnon-White HC2 correction removes the entire bias. Its form in this case is

$$\hat{\mathbb{V}}_{\text{HC2}} = \frac{\hat{\sigma}^2(0)}{N_0} + \frac{\hat{\sigma}^2(1)}{N_1}, \quad \text{where} \quad \hat{\sigma}^2(d) = \frac{1}{N_d - 1} \sum_{i:D_i=d}^{N} \left(Y_i - \overline{Y}_d\right)^2, \quad d = 0, 1. \tag{2.4}$$

These conditional variance estimators $\hat{\sigma}^2(d)$ differ from the EHW estimator $\tilde{\sigma}^2(d)$ by a factor $N_d/(N_d-1)$. In combination with the Normal approximation to the distribution of the t-statistic, this variance estimator leads to the 95% confidence interval

$$\text{CI}^{95\%}_{\text{HC2}} = \left(\hat{\beta}_1 - 1.96 \times \sqrt{\hat{\mathbb{V}}}_{\text{HC2}}, \hat{\beta}_1 + 1.96 \times \sqrt{\hat{\mathbb{V}}}_{\text{HC2}}\right).$$

The estimator $\hat{\mathbb{V}}_{\text{HC2}}$ is unbiased for $\mathbb{V}$, but the resulting confidence interval is still not exact. Just as in the homoskedastic case, the sampling distribution of the t-statistic $(\hat{\beta}_1 - \beta_1)/\sqrt{\hat{\mathbb{V}}_{\text{HC2}}}$ is in this case not Normally distributed in small samples, even if the underlying errors are Normally distributed (and thus $(\hat{\beta}_1 - \beta_1)/\sqrt{\mathbb{V}}$ has an exact standard Normal distribution). Whereas in the homoskedastic case, the t-statistic has an exact t-distribution with $N-2$ dof, here the exact distribution of the t-statistic does not lend itself to the construction of exact confidence intervals: the distribution of $\hat{\mathbb{V}}_{\text{HC2}}$ not chi-squared, but a weighted sum of two chi-squared distributions with weights that depend on $\sigma^2(d)$.

In this single-binary-regressor case it is easy to see that in some cases $N-2$ will be a poor choice for the degrees of freedom for the approximating t-distribution. Suppose that there are many units with $D_i = 0$ and few units with $D_i = 1$ ($N_0 \gg N_1$). In that case $\mathbb{E}[Y_i \mid D_i = 0]$ is estimated relatively precisely, with variance $\sigma^2(0)/N_0 \approx 0$. As a result the distribution of the t-statistic $(\hat{\beta}_1 - \beta_1)/\sqrt{\hat{\mathbb{V}}_{\text{HC2}}}$ is approximately equal to that of $(\overline{Y}_1 - \mathbb{E}[Y_i \mid D_i = 1])/\sqrt{\hat{\sigma}^2(1)/N_1}$. The latter has, under Normality, an exact t-distribution with dof equal to $N_1 - 1$, substantially different from the t-distribution with $N - 2 = N_0 + N_1 - 2$ dof if $N_0 \gg N_1$.

2.5 Degrees of freedom adjustment: Welch and Bell-McCaffrey solutions

One popular and attractive approach to deal with the Behrens-Fisher problem is due to Welch (1951). Welch suggests approximating the distribution of the t-statistic $(\hat{\beta}_1 - \beta_1)/\sqrt{\hat{\mathbb{V}}_{\text{HC2}}}$ by a t-distribution with dof adjusted to reflect the variability of the variance estimator $\hat{\mathbb{V}}_{\text{HC2}}$. To describe this adjustment in more detail, consider the t-statistic in the heteroskedastic case:

$$t_{\text{HC2}} = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\hat{\mathbb{V}}_{\text{HC2}}}} = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\hat{\sigma}^2(0)/N_0 + \hat{\sigma}^2(1)/N_1}}.$$

Suppose there was a constant K such that the distribution of $K \cdot \hat{\mathbb{V}}_{\text{HC2}}/\mathbb{V}$ had a chi-squared distribution with dof equal to K. Then, under Normality, because $\hat{\mathbb{V}}_{\text{HC2}}$ is independent of $\hat{\beta}_1 - \beta_1$, t_{HC2} would have a t-distribution with dof equal to K, which could be exploited to construct an exact confidence interval. Unfortunately, there is no value of K that makes $K \cdot \hat{\mathbb{V}}_{\text{HC2}}/\mathbb{V}$ exactly chi-squared distributed. Welch therefore suggests approximating the scaled distribution of $\hat{\mathbb{V}}_{\text{HC2}}$ by a chi-squared distribution, with the dof parameter K chosen to make the approximation as accurate as possible. In particular, Welch proposes to choose the dof parameter K such that $K \cdot \hat{\mathbb{V}}_{\text{HC2}}/\mathbb{V}$ has the first two moments in common with a chi-squared distribution with dof equal to K. Because irrespective of the value for K, $\mathbb{E}[K \cdot \hat{\mathbb{V}}_{\text{HC2}}/\mathbb{V}] = K$, this amounts to choosing K such that $\text{var}(K \cdot \hat{\mathbb{V}}_{\text{HC2}}/\mathbb{V}) = 2K$. To find this value of K, note that under Normality, $\hat{\mathbb{V}}_{\text{HC2}}$ is a linear combination of two chi-squared random variables. To be precise, $(N_0 - 1)\hat{\sigma}^2(0)/\sigma^2(0) \sim \chi^2(N_0 - 1)$, and $(N_1 - 1)\hat{\sigma}^2(1)/\sigma^2(1) \sim \chi^2(N_1 - 1)$, and $\hat{\sigma}^2(0)$ and $\hat{\sigma}^2(1)$ are independent of each other and of $\hat{\beta}_1 - \beta_1$. Hence it follows that

$$\text{var}\left(\hat{\mathbb{V}}_{\text{HC2}}\right) = \frac{2\sigma^4(0)}{(N_0 - 1)N_0^2} + \frac{2\sigma^4(1)}{(N_1 - 1)N_1^2},$$

which leads to

$$K^*_{\text{Welch}} = \frac{2 \cdot \mathbb{V}^2}{\text{var}\left(\hat{\mathbb{V}}_{\text{HC2}}\right)} = \frac{\left(\frac{\sigma^2(0)}{N_0} + \frac{\sigma^2(1)}{N_1}\right)^2}{\frac{\sigma^4(0)}{(N_0-1)N_0^2} + \frac{\sigma^4(1)}{(N_1-1)N_1^2}} = \frac{\left(\frac{1}{N_0}\frac{\sigma^2(0)}{\sigma^2(1)} + \frac{1}{N_1}\right)^2}{\frac{1}{(N_0-1)N_0^2}\frac{\sigma^4(0)}{\sigma^4(1)} + \frac{1}{(N_1-1)N_1^2}}$$

This choice for K is not feasible because K^*_{Welch} depends on the unknown ratio of the conditional variances $\sigma^2(0)/\sigma^2(1)$. In the feasible version we approximate the distribution of t_{HC2} by a t-distribution with dof equal to

$$K_{\text{Welch}} = \left(\frac{\hat{\sigma}^2(0)}{N_0} + \frac{\hat{\sigma}^2(1)}{N_1}\right)^2 \bigg/ \left(\frac{\hat{\sigma}^4(0)}{(N_0 - 1)N_0^2} + \frac{\hat{\sigma}^4(1)}{(N_1 - 1)N_1^2}\right), \tag{2.5}$$

where the unknown $\sigma^2(d)$ are replaced by the estimates $\hat{\sigma}^2(d)$. Wang (1971) presents some exact results for the difference between the coverage of confidence intervals based on the Welch procedures and the nominal levels, showing that the Welch intervals perform extremely well in very small samples.

BM propose a slightly different degrees of freedom adjustment. For the Behrens-Fisher problem (regression with a single binary regressor) the BM modification is minor, but it has considerable attraction in settings with more general distributions of regressors. The BM adjustment simplifies the Welch dof K^*_{Welch} by assuming homoskedasticity, leading to

$$K_{\text{BM}} = \frac{\left(\frac{\sigma^2}{N_0} + \frac{\sigma^2}{N_1}\right)^2}{\frac{\sigma^4}{(N_0-1)N_0^2} + \frac{\sigma^4}{(N_1-1)N_1^2}} = \frac{(N_0+N_1)^2(N_0-1)(N_1-1)}{N_1^2(N_1-1)+N_0^2(N_0-1)}. \tag{2.6}$$

Because the BM dof does not depend on the conditional variances, it is non-random conditional on the regressors, and as a result tends to be more accurate then the Welch adjustment in settings with noisy estimates of the conditional error variances. The associated 95% confidence interval is now

$$\text{CI}^{95\%}_{\text{BM}} = \left(\hat{\beta}_1 - t^{K_{\text{BM}}}_{0.975} \times \sqrt{\hat{\mathbb{V}}}_{\text{HC2}}, \hat{\beta}_1 + t^{K_{\text{BM}}}_{0.975} \times \sqrt{\hat{\mathbb{V}}}_{\text{HC2}}\right). \tag{2.7}$$

This is the interval we recommend researchers use in practice.

To gain some intuition for the BM dof adjustment, consider some special cases. First, if $N_0 \gg N_1$, then $K_{\text{BM}} \approx N_1 - 1$. As we have seen before, as $N_0 \to \infty$, using $N_1 - 1$ as the degrees of freedom leads to exact confidence intervals under Normally distributed errors. If the two subsamples are equal size, $N_0 = N_1 = N/2$, then $K_{\text{BM}} = N - 2$. Thus, if the two subsamples are approximately equal size, the often-used dof adjustment of $N-2$ is appropriate, but if the distribution is very skewed, this adjustment is likely to be inadequate.

2.6 Small simulation study based on Angrist-Pischke design

To see how relevant the small sample adjustments are in practice, we conduct a small simulation study based on a design previously used by Angrist and Pischke (2009). The sample size is $N = 30$, with $N_1 = 3$ and $N_0 = 27$. The parameter values are $\beta_0 = \beta_1 = 0$ (the results are invariant to the values for β_0 and β_1). The distribution of the disturbances is Normal,

$$\varepsilon_i \mid D_i = d \sim \mathcal{N}(0, \sigma^2(d)), \quad d = 0, 1. \tag{2.8}$$

with $\sigma^2(1) = 1$. Angrist and Pischke report results for three choices for $\sigma(0)$: $\sigma(0) \in \{0.5, 0.85, 1\}$. We add the complementary values $\sigma(0) \in \{1.18, 2\}$, where $1.18 \approx 1/0.85$. Angrist and Pischke report results for a number of variance estimators, including some where they take the maximum of $\hat{\mathbb{V}}_{\text{homo}}$ and $\hat{\mathbb{V}}_{\text{EHW}}$ or $\hat{\mathbb{V}}_{\text{HC2}}$, but they do not consider the Welch or BM dof adjustments.

We consider the following confidence intervals. First, two intervals based on the homoskedastic variance estimator $\hat{\mathbb{V}}_{\text{homo}}$, using either the Normal distribution or a t-distribution with $N-2$ dof. Next, four confidence intervals based on $\hat{\mathbb{V}}_{\text{EHW}}$. The first two again use either the Normal or the t-distribution with $N-2$ dof. The last two are based on the wild bootstrap, a resampling method discussed in more detail in Appendix A. The first one of these methods (denoted "wild") is based on the percentile-t method of obtaining the confidence interval. The second confidence interval (denoted "wild$_0$") consists of all null hypotheses $H_0\colon \beta_1 = \beta_1^0$ that were not rejected by wild bootstrap tests that impose the null hypothesis when calculating the wild bootstrap distribution (see Appendix A for details). This method involves a numerical search, and is therefore computationally intensive. Next, seven confidence intervals based on $\hat{\mathbb{V}}_{\text{HC2}}$, using: Normal distribution, t-distribution with $N-2$ dof, the two versions of the wild bootstrap, K_{Welch}, K^*_{Welch}, and K_{BM}. We also include a confidence interval based on $\hat{\mathbb{V}}_{\text{HC3}}$ (see Appendix A for more details). Finally, we include confidence intervals based on the maximum of $\hat{\mathbb{V}}_{\text{homo}}$ and $\hat{\mathbb{V}}_{\text{EHW}}$, and on the maximum of $\hat{\mathbb{V}}_{\text{homo}}$ and $\hat{\mathbb{V}}_{\text{HC2}}$, both using the Normal distribution.

Table 1 presents the simulation results. For each of the variance estimators we report coverage probabilities for nominal 95% confidence intervals, and the median of the standard errors over the simulations. To make the standard errors comparable, we multiply the square root of the variance estimators by $t^K_{0.975}/t^\infty_{0.975}$ in cases where the confidence intervals are based on t-distributions with K degrees of freedom. We also report the mean K^*_{Welch}, K_{Welch} and K_{BM} dof adjustments, which are substantial in these designs. For instance, in the first design, with $\sigma(0)/\sigma(1) = 0.5$, the infeasible Welch dof is $K^*_{\text{Welch}} = 2.1$, indicating that the EHW standard errors may not be reliable: the dof correction leads to an adjustment in the standard errors by a factor[1] of $t^{2.1}_{0.975}/t^\infty_{0.957} = 4.11/1.96 = 2.1$. Indeed, the coverage rate for Normal-distribution confidence interval based on $\hat{\mathbb{V}}_{\text{EHW}}$ is 0.77, and it's 0.82 based on the unbiased variance estimator $\hat{\mathbb{V}}_{\text{HC2}}$.

For the variance estimators included in the Angrist-Pischke design our simulation

[1] To implement the degrees-of-freedom adjustment with non-integer dof K, we define the t-distribution as the ratio of two random variables, one a random variable with a standard (mean zero, unit variance) Normal distribution and and the second a random variable with a gamma distribution with parameters $\alpha = K/2$ and $\beta = 2$.

results are consistent with theirs. However, the three confidence intervals based on the (feasible and infeasible) Welch and BM degrees of freedom adjustments are superior in terms of coverage. The confidence intervals based on the wild bootstrap with the null imposed also perform well although they undercover somewhat at $\sigma(0) = 0.5$, and are very conservative and wide at $\sigma(0) = 2$: their median length is about 45% greater than that of BM.

An attractive feature of the BM correction is that the confidence intervals have substantially more variation in their width relative to the Welch confidence intervals. For instance, with $\sigma(0) = 1$, the median widths of the confidence intervals based on K_{Welch} and K_{BM} are 3.5 and 3.7 (and the Welch confidence interval slightly undercovers), but the 0.95 quantile of the widths are 7.1 and 6.5. The attempt to base the approximating chi-square distribution on the heteroskedasticity consistent variance estimates leads to a considerable increase in the variability of the width of the confidence intervals (this is evidenced in the variability of K_{Welch}, which has variance between 2.6 and 7.5 depending on the design). Moreover, because conditional on the regressors, the BM critical value is fixed, size-adjusted power of tests based on the BM correction coincides with that of tests based on HC2 and the Normal distribution, while, as evidenced by the simulation results, its size properties are superior.

To investigate the importance of the assumption of the Normality for these results, we also consider a design with log-Normal errors, $\epsilon_i \mid D_i = d \sim \sigma(d)L_i$, where L_i is a log-Normal random variable, recentered and rescaled so that it has mean zero and variance one. The results are reported in Table 2. Here the BM intervals perform substantially better than Welch intervals. The undercoverage of the remaining confidence intervals except the wild bootstrap with the null imposed is even more severe than with Normal errors. The wild bootstrap intervals, however, again tend to be very conservative and wide for larger values of $\sigma(0)$.

For comparison, we also report in Table 3 the results for a simulation exercise with a balanced design where $N_0 = N_1 = N/2 = 15$, and Normal errors. Here $K_{\text{BM}} = 28$ across the designs, and since $t^{28}_{0.975} = 2.05$ is close to the 1.96, it suggests that refinements are not important here. Indeed, the actual coverage rates are close to nominal coverage rates for essentially all procedures: for a sample size of 30 and balanced design, the asymptotic Normal-distribution-based approximations are fairly accurate.

3 Linear regression with general regressors

Now let us look at the general regression case, allowing for multiple regressors, and regressors with other than binomial distributions.

3.1 Setup

We have an L-dimensional vector of regressors X_i, and a linear model

$$Y_i = X_i'\beta + \varepsilon_i, \quad \text{with} \qquad \mathbb{E}\left[\varepsilon_i|X_i\right] = 0, \qquad \text{var}\left(\varepsilon_i|X_i\right) = \sigma^2(X_i).$$

Let $\mathbf{X}$ be the $N \times L$ dimensional matrix with ith row equal to X_i', and let $\mathbf{Y}$ and ε be the N-vectors with ith elements equal to Y_i and ε_i respectively. The ordinary least squares estimator is given by

$$\hat{\beta} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\left(\mathbf{X}'\mathbf{Y}\right) = \left(\sum_{i=1}^{N} X_i X_i'\right)^{-1}\left(\sum_{i=1}^{N} X_i Y_i\right).$$

Without assuming homoskedasticity, the exact variance for $\hat{\beta}$ conditional on $\mathbf{X}$ is

$$\mathbb{V} = \text{var}(\hat{\beta} \mid \mathbf{X}) = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\sum_{i=1}^{N}\sigma^2(X_i)X_iX_i'\left(\mathbf{X}'\mathbf{X}\right)^{-1},$$

with k-th diagonal element $\mathbb{V}_k$. For the general regression case the EHW robust variance estimator is

$$\hat{\mathbb{V}}_{\text{EHW}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\sum_{i=1}^{N}\left(Y_i - X_i\hat{\beta}\right)^2 X_iX_i'\left(\mathbf{X}'\mathbf{X}\right)^{-1},$$

with k-th diagonal element $\hat{\mathbb{V}}_{\text{EHW},k}$. Using a Normal distribution, the associated 95% confidence interval for β_k is

$$\text{CI}_{\text{EHW}}^{95\%} = \left(\hat{\beta}_k - 1.96 \times \sqrt{\hat{\mathbb{V}}_{\text{EHW},k}}, \hat{\beta}_k + 1.96 \times \sqrt{\hat{\mathbb{V}}_{\text{EHW},k}}\right).$$

This robust variance estimator and the associated confidence intervals are widely used in empirical work.

3.2 Bias-adjusted variance estimator

In Section 2 we discussed the bias of the robust variance estimator in the case with a single binary regressor. In that case there was a simple modification of the EHW variance estimator that removes all bias. In the general regression case it is not possible to remove all bias in general. We focus on a particular adjustment for the bias first proposed by MacKinnon and White (1985) (see also Horn, Horn, and Duncan, 1975). In the special case with only a single binary regressor this adjustment is identical to that used in Section 2. Let $\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ be the $N \times N$ projection matrix, with i-th column denoted by $\mathbf{P}_i = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}X_i$ and (i,i)-th element denoted by $P_{ii} = X_i'(\mathbf{X}'\mathbf{X})^{-1}X_i$. Let Ω be the $N \times N$ diagonal matrix with i-th diagonal element equal to $\sigma^2(X_i)$, and let $e_{N,i}$ be the N-vector with i-th element equal to one and all other elements equal to zero. Let I_N be the $N \times N$ identity matrix. The residuals $\hat{\varepsilon}_i = Y_i - X_i'\hat{\beta}$ can be written as

$$\hat{\varepsilon}_i = \varepsilon_i - e_{N,i}'\mathbf{P}\varepsilon = e_{N,i}'(I_N - \mathbf{P})\varepsilon, \quad \text{or, in vector form,} \quad \hat{\varepsilon} = (I_N - \mathbf{P})\varepsilon.$$

The expected value of the square of the i-th residual is

$$\mathbb{E}\left[\hat{\varepsilon}_i^2\right] = \mathbb{E}\left[(e_{N,i}'(I_N - \mathbf{P})\varepsilon)^2\right] = (e_{N,i} - \mathbf{P}_i)'\Omega(e_{N,i} - \mathbf{P}_i),$$

which, under homoskedasticity reduces to $\sigma^2(1-\mathbf{P}_{ii})$. This in turn implies that $\hat{\varepsilon}_i^2/(1-\mathbf{P}_{ii})$ is unbiased for $\mathbb{E}\left[\varepsilon_i^2\right]$ under homoskedasticity. This is the motivation for the variance estimator that MacKinnon and White (1985) introduce as HC2:

$$\hat{\mathbb{V}}_{\text{HC2}} = (\mathbf{X}'\mathbf{X})^{-1} \sum_{i=1}^{N} \frac{\left(Y_i - X_i\hat{\beta}\right)^2}{1 - \mathbf{P}_{ii}} X_i X_i' \left(\mathbf{X}'\mathbf{X}\right)^{-1}. \tag{3.1}$$

Suppose we want to construct a confidence interval for β_k, the k-th element of β. The variance of $\hat{\beta}_k$ is estimated as $\hat{\mathbb{V}}_{\text{HC2},k}$, the kth diagonal element of $\hat{\mathbb{V}}_{\text{HC2}}$. The 95% confidence interval, based on the Normal approximation, is then given by

$$\text{CI}_{\text{HC2}}^{95\%} = \left(\hat{\beta}_k - 1.96 \times \sqrt{\hat{\mathbb{V}}_{\text{HC2},k}}, \hat{\beta}_k + 1.96 \times \sqrt{\hat{\mathbb{V}}_{\text{HC2},k}}\right).$$

3.3 Degrees of freedom adjustment

BM, building on Satterthwaite (1946), suggest approximating the distribution of the t-statistic $t_{\text{HC2}} = (\hat{\beta}_k - \beta_k)/\sqrt{\hat{\mathbb{V}}_{\text{HC2},k}}$ by a t-distribution instead of a Normal distribution.

Like in the binary Behrens-Fisher case, the degrees of freedom K are chosen so that under homoskedasticity ($\Omega = \sigma^2 I_N$) the first two moments of $K \cdot (\hat{\mathbb{V}}_{\mathrm{HC2},k}/\mathbb{V}_k)$ are equal to those of a chi-squared distribution with degrees of freedom equal to K. Under homoskedasticity, $\hat{\mathbb{V}}_{\mathrm{HC2}}$ is unbiased, and thus thus $\mathbb{E}[\hat{\mathbb{V}}_{\mathrm{HC2},k}] = \mathbb{V}_k$, so that the first moment of $K \cdot (\hat{\mathbb{V}}_{\mathrm{HC2},k}/\mathbb{V}_k)$ is always equal to to that of a chi-squared distribution with dof equal to K. Therefore, we choose K to match the second moment. Under Normality, $\hat{\mathbb{V}}_{\mathrm{HC2},k}$ is a linear combination of N independent chi-squared one random variables (with some of the coefficients equal to zero),

$$\hat{\mathbb{V}}_{\mathrm{HC2},k} = \sum_{i=1}^{N} \lambda_i \cdot Z_i, \quad \text{where} \quad Z_i \sim \chi^2(1), \quad \text{all } Z_i \text{ independent},$$

where the weights λ_i are eigenvalues of the $N \times N$ matrix $\sigma^2 \cdot \mathbf{G}'\mathbf{G}$, with the i-th column of the $N \times N$ matrix $\mathbf{G}$, equal to

$$\mathbf{G}_i = \frac{1}{\sqrt{1 - \mathbf{P}_{ii}}} (e_{N,i} - \mathbf{P}_i) X_i' (\mathbf{X}'\mathbf{X})^{-1} e_{L,k}.$$

Given these weights, the BM dof that match the first two moments of $K \cdot (\hat{\mathbb{V}}_{\mathrm{HC2},k}/\mathbb{V}_k)$ to that of a chi-squared K distribution is given by

$$K_{\mathrm{BM}} = \frac{2 \cdot \mathbb{V}_k^2}{\mathrm{var}\left(\hat{\mathbb{V}}_{\mathrm{HC2},k}\right)} = \left(\sum_{i=1}^{N} \lambda_i\right)^2 \Big/ \sum_{i=1}^{N} \lambda_i^2. \tag{3.2}$$

The value of K_{BM} only depends on the regressors (through the matrix $\mathbf{G}$) and not on σ^2 even though the weights λ_i do depend on σ^2. In particular, the effective dof will be smaller if the distribution of the regressors is skewed. Note also that the dof adjustment may be different for different elements of parameter β. The resulting 95% confidence interval is

$$\mathrm{CI}_{\mathrm{BM}}^{95\%} = \left(\hat{\beta}_k + t_{0.025}^{K_{\mathrm{BM}}} \times \sqrt{\hat{\mathbb{V}}_{\mathrm{HC2},k}}, \hat{\beta}_k + t_{0.975}^{K_{\mathrm{BM}}} \times \sqrt{\hat{\mathbb{V}}_{\mathrm{HC2},k}}\right).$$

In general, the weights λ_i that set the moments of the chi-squared approximation equal to those of the normalized variance are the eigenvalues of $\mathbf{G}'\Omega\mathbf{G}$. These weights are not feasible, because Ω is not known in general. The feasible version of the Sattherthwaite dof suggestion replaces Ω by $\hat{\Omega} = \mathrm{diag}(\hat{\varepsilon}_i^2/(1 - \mathbf{P}_{ii}))$. However, because $\hat{\Omega}$ is a noisy estimator of the conditional variance, the resulting confidence intervals are often substantially conservative. By basing the dof calculation on the homoskedastic case with $\Omega = \sigma^2 \cdot I_N$, the BM adjustment avoids this problem.

If there is a single binary regressor, the BM solution for the general case (3.2) reduces to that in the binary case, (2.6). Similarly, the infeasible Sattherthwaite solution, based on the eigenvalues of $\mathbf{G}\Omega\mathbf{G}$, reduces to the infeasible Welch solution K^*_{Welch}. In contrast, applying the feasible Sattherthwaite solution to the case with a binary regressor does *not* lead to the feasible Welch solution because the feasible Welch solution implicitly uses an estimator for Ω different from $\hat{\Omega}$.

The performance of the Sattherthwaite and BM confidence intervals is similar to that of the Welch and BM confidence intervals in the binary case.[2] In particular, if the design of regressors is skewed (for example, if the regressor of interest has a log-Normal distribution), then the robust variance estimators $\hat{\mathbb{V}}_{\text{EHW}}$ and the bias-adjusted version $\hat{\mathbb{V}}_{\text{HC2}}$ based on a normal distribution or a t-distribution with $N-2$ dof may undercover substantially even when $N \approx 100$. In contrast, the Sattherthwaite and BM confidence intervals control size even in small samples, because any skewness is captured in the matrix $\mathbf{G}$, leading to appropriate dof adjustments. The K_{BM} dof adjustment leads to much narrower confidence intervals with much less variation, so again that is the superior choice in this setting.

4 Robust variance estimators with clustering

In this section we discuss the extensions of the variance estimators discussed in the previous sections to the case with clustering. The model is:

$$Y_i = X_i'\beta + \varepsilon_i, \tag{4.1}$$

There are S clusters. In cluster s the number of units is N_s, with the overall sample size $N = \sum_{s=1}^{S} N_s$. Let $S_i \in \{1, \dots, S\}$ denote the cluster unit i belongs to. We assume that the errors ϵ_i are uncorrelated between clusters, but there may be arbitrary correlation within a cluster,

$$\mathbb{E}[\varepsilon \mid \mathbf{X}] = 0, \qquad \mathbb{E}[\varepsilon\varepsilon' \mid \mathbf{X}] = \Omega, \qquad \Omega_{ij} = \begin{cases} \omega_{ij} & \text{if } S_i = S_j, \\ 0 & \text{otherwise.} \end{cases}$$

If $\omega_{ij} = 0$ for $i \neq j$ (that is, each unit is in its own cluster), the setup reduces to that in Section 3.

[2]See an earlier version of this paper (Imbens and Kolesár, 2012) for simulation evidence.

Let $\hat{\beta}$ be the least squares estimator, and let $\hat{\varepsilon}_i = Y_i - X_i'\hat{\beta}$ be the residual. Let $\hat{\boldsymbol{\varepsilon}}_s$ be the N_s dimensional vector with the residuals in cluster s, let $\mathbf{X}_s$ the $N_s \times L$ matrix with ith row equal to the value of X_i' for the ith unit in cluster s, and let $\mathbf{X}$ be the $N \times L$ matrix constructed by stacking $\mathbf{X}_1$ through $\mathbf{X}_S$. Define the $N \times N_s$ matrix $\mathbf{P}_s = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_s'$, the $N_s \times N_s$ matrix $\mathbf{P}_{ss} = \mathbf{X}_s(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_s'$, and define the $N \times N_s$ matrix $(I_N - \mathbf{P})_s$ to consist of the N_s columns of the $N \times N$ matrix $(I_N - \mathbf{P})$ corresponding to cluster s.

The exact variance of $\hat{\beta}$ conditional on $\mathbf{X}$ is given by

$$\mathbb{V} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\Omega\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}.$$

The standard robust variance estimator, due to Liang and Zeger (1986) (see also Diggle, Heagerty, Liang, and Zeger, 2002), is

$$\hat{\mathbb{V}}_{\mathrm{LZ}} = (\mathbf{X}'\mathbf{X})^{-1} \sum_{s=1}^{S} \mathbf{X}_s'\hat{\boldsymbol{\varepsilon}}_s\hat{\boldsymbol{\varepsilon}}_s'\mathbf{X}_s \, (\mathbf{X}'\mathbf{X})^{-1} .$$

Often a simple multiplicative adjustment is used, for example in STATA, to reduce the bias of the LZ variance estimator:

$$\hat{\mathbb{V}}_{\mathrm{STATA}} = \frac{N-1}{N-L} \cdot \frac{S}{S-1} \cdot (\mathbf{X}'\mathbf{X})^{-1} \sum_{s=1}^{S} \mathbf{X}_s'\hat{\boldsymbol{\varepsilon}}_s\hat{\boldsymbol{\varepsilon}}_s'\mathbf{X}_s \, (\mathbf{X}'\mathbf{X})^{-1} .$$

The main component of this adjustment is typically the $S/(S-1)$ factor, because in many applications, $(N-1)/(N-L)$ is close to one.

The bias-reduction modification developed by Bell and McCaffrey (2002), analogous to the HC2 bias reduction of the original Eicker-Huber-White variance estimator, is

$$\hat{\mathbb{V}}_{\mathrm{LZ2}} = (\mathbf{X}'\mathbf{X})^{-1} \sum_{s=1}^{S} \mathbf{X}_s'(I_{N_s} - \mathbf{P}_{ss})^{-1/2}\hat{\boldsymbol{\varepsilon}}_s\hat{\boldsymbol{\varepsilon}}_s'(I_{N_s} - \mathbf{P}_{ss})^{-1/2}\mathbf{X}_s \, (\mathbf{X}'\mathbf{X})^{-1} ,$$

where $(I_{N_s} - \mathbf{P}_{ss})^{-1/2}$ is the inverse of the symmetric square root of $(I_{N_s} - \mathbf{P}_{ss})$. For each of the variance estimators, let $\hat{\mathbb{V}}_{\mathrm{LZ},k}$, $\hat{\mathbb{V}}_{\mathrm{STATA},k}$ and $\hat{\mathbb{V}}_{\mathrm{LZ2},k}$ are the k-th diagonal elements of $\hat{\mathbb{V}}_{\mathrm{LZ}}$, $\hat{\mathbb{V}}_{\mathrm{STATA}}$ and $\hat{\mathbb{V}}_{\mathrm{LZ2}}$ respectively.

To define the degrees-of-freedom adjustment, let $\mathbf{G}$ denote the $N \times S$ matrix with s-th column equal to the N-vector

$$\mathbf{G}_s = (I_N - \mathbf{P})_s(I_{N_s} - \mathbf{P}_{ss})^{-1/2}\mathbf{X}_s \, (\mathbf{X}'\mathbf{X})^{-1} e_{L,k}.$$

Then the dof adjustment is given by

$$K_{\rm BM}=\frac{\left(\sum_{i=1}^{N}\lambda_i\right)^2}{\sum_{i=1}^{N}\lambda_i^2}.$$

where λ_i are the eigenvalues of $\mathbf{G}'\mathbf{G}$. If each unit is in its own cluster (so there is no clustering), this adjustment reduces to the adjustment given in (3.2). The 95% confidence interval is given by

$$\mathrm{CI}^{95\%}_{\rm cluster,BM}=\left(\hat{\beta}_k+t^{K_{\rm BM}}_{0.025}\times\sqrt{\hat{\mathbb{V}}_{{\rm LZ2},k}},\hat{\beta}_k+t^{K_{\rm BM}}_{0.975}\times\sqrt{\hat{\mathbb{V}}_{{\rm LZ2},k}}\right). \tag{4.2}$$

We also consider a slightly different version of the dof adjustment. In principle, we would like to use the eigenvalues of the matrix $\mathbf{G}'\Omega\mathbf{G}$, so that the first two moments of $K\cdot\hat{\mathbb{V}}_{{\rm LZ2},k}/\mathbb{V}_k$ match that of $\chi^2(K)$. It is difficult to estimate Ω accurately without any restrictions, which motivated BM to use $\sigma^2\cdot I_N$ instead. In the clustering case, however, it is attractive to put a random-effects structure on the errors as in Moulton (1986, 1990) and estimate a model for Ω where

$$\Omega_{ij}=\begin{cases}\sigma^2_\varepsilon & \text{if } i=j,\\ \rho & \text{if } i\neq j, S_i=S_j.\\ 0 & \text{otherwise}\end{cases}$$

We estimate σ_ν as the average of the product of the residuals for units with $S_i=S_j$, and $i\neq j$

$$\rho=\frac{1}{n-m}\left(\sum_{s=1}^{S}\sum_{i:\,S_i=s}\sum_{j:\,S_j=s}\hat{\varepsilon}_i\hat{\varepsilon}_j-\sum_i\hat{\varepsilon}_i^2\right),$$

where $m=\sum_s N_s^2$, and N_s is the number of observations in cluster S, and we estimate σ^2_ε as the average of the square of the residuals, $\hat{\sigma}^2_\epsilon=N^{-1}\sum_{i=1}^N\hat{\epsilon}_i^2$. We then calculate the $\tilde{\lambda}_i$ as the eigenvalues of $\mathbf{G}'\hat{\Omega}\mathbf{G}$, and set

$$K_{\rm IK}=\frac{\left(\sum_{i=1}^{N}\tilde{\lambda}_i\right)^2}{\sum_{i=1}^{N}\tilde{\lambda}_i^2}.$$

4.1 Small simulation study

We carry out a small simulation study. The first sets of designs is corresponds to the designs first used in Cameron, Gelbach, and Miller (2008). The baseline model (design I) is the same as in (4.1), with a scalar regressor:

$$Y_i = \beta_0 + \beta_1 \cdot X_i + \varepsilon_i,$$

with $\beta_0 = \beta_1 = 0$, $X_i = V_{S_i} + W_i$ and $\varepsilon_i = \nu_{S_i} + \eta_i$, with V_s, W_i, ν_s, η_i are all Normally distributed, with mean zero and unit variance. There there are $S = 10$ clusters, with $N_s = 30$ units in each cluster. In design II, we have $S = 5$ clusters, again with $N_s = 30$ in each cluster. In design III, there there are again $S = 10$ clusters, half with $N_s = 10$ and half with $N_s = 50$. In the fourth and fifth design we return to the design with $S = 10$ clusters and $N_s = 30$ units per cluster. In the design IV we introduce heteroskedasticity, with $\eta_i|\mathbf{X} \sim N(0, 0.9X_i^2)$, and in the design V, the regressor is fixed within the clusters: $W_i = 0$ and $V_s \sim \mathcal{N}(0, 2)$. All five designs correspond to those in Cameron, Gelbach, and Miller (2008).

We consider the following confidence intervals. First, two intervals based on the homoskedastic variance estimator $\hat{\mathbb{V}}_{\text{homo}}$ that ignores clustering, using either the Normal distribution or a t-distribution with $S - 1$ dof. Next, four confidence intervals based on $\hat{\mathbb{V}}_{\text{LZ}}$. The first two again use either the Normal or the t-distribution with $S - 1$ dof. The last two are based on the wild bootstrap, a resampling method discussed in more detail in Appendix A. The first one of these methods (denoted "wild") is based on the percentile-t method of obtaining the confidence interval. The second confidence interval (denoted "wild$_0$") consists of all null hypotheses $H_0\colon \beta_1 = \beta_1^0$ that were not rejected by wild bootstrap tests that impose the null hypothesis when calculating the wild bootstrap distribution (see Appendix A for details). This method involves a numerical search, and is therefore computationally intensive. Next, we report two confidence intervals based on $\hat{\mathbb{V}}_{\text{STATA}}$, using the Normal distribution and the t-distribution with $N - 1$ dof. Finally, we report seven confidence intervals based on $\hat{\mathbb{V}}_{\text{LZ2}}$, using: Normal distribution, t-distribution with $S - 1$ dof, the two versions of the wild bootstrap, K_{BM}, K_{IK}, and the infeasible Sattherthwaite dof $K^*_{\text{Satt.}}$ that uses eigenvalues of the matrix $\mathbf{G}'\Omega\mathbf{G}$ to compute the dof correction.

Table 4 presents the simulation results. As in the simulations in Section 2, we report coverage probabilities and normalized standard errors for each estimator, and we also report report the mean $K^*_{\text{Satt.}}$, K_{BM} and K_{IK} dof adjustments, which are substantial in these designs. The K_{IK} dof adjustment yields confidence intervals that are closer to

$K^*_{\text{Satt.}}$, which yields slight improvements in coverage. Overall, however, for the BM and IK methods are superior in terms of coverage to all other methods. Although using $S-1$ dof rather than a Normal approximation improves coverage for $\hat{\mathbb{V}}_{\text{LZ}}, \hat{\mathbb{V}}_{\text{STATA}}$ and $\hat{\mathbb{V}}_{\text{LZ2}}$, the confidence intervals still undercover. The wild bootstrap with the null imposes does better than these methods, although it results in very wide confidence intervals in design II with only 5 clusters. In design III, the unbalanced cluster size means that the distribution of the regressor is more skewed than in design I, and leads to one less effective dof (3.1 rather than 4.1 for K_{IK}, for instance), and consequently to more severe undercoverage of the standard confidence interval.

To further investigate the effect of the skewness of the regressors, we consider additional simulation designs, which are reported in Table 5. The baseline design (design VI), is the same as design I, except there are 50 clusters, with 6 observations in each cluster. Here, like in the balanced design in section 2, the dof correction is not important, and all methods perform well. Next, in design VII we consider a log-normal distribution of the regressor, $V_s \sim \exp(\mathcal{N}(0,1))$, $W_i = 0$. Here, the dof correction matters, and standard methods undercover substantially in spite of there being as many as 50 clusters. Finally, we consider three designs similar to the unbalanced designs in Section 2. There are three treated states with $X_i = 1$, and $X_i = 0$ for observations in the remaining states. In design VII, the errors are drawn as in the baseline design, with both ν_{S_i} and η_i standard Normal. In design IX, $\nu_{S_i} \mid X_i = x \sim \mathcal{N}(0, \sigma_\nu(x))$, with $\sigma_\nu(1) = 2$ and $\sigma_\nu(0) = 1$. The final design (design X) is the same except $\sigma_\nu(1) = 1$ and $\sigma_\nu(0) = 2$. Again, in these designs the standard methods undercover due to the skewness of the regressors despite the relatively large number of clusters. In contrast, both the IK and the BM adjustment work well.

5 Conclusion

Although there is a substantial literature documenting the poor properties of the conventional robust standard errors in small samples, in practice many researchers continue to use the EHW and LZ robust standard errors. Here we discuss one of the proposed modifications, due to Bell and McCaffrey (2002), and argue that it should be used more widely, even in moderately sized samples, especially when the distribution of the covariates is skewed. The modification is straightforward to implement. It consists of two components. First, it removes some of the bias in the EHW variance estimator. Second, it uses a degrees-of-freedom adjustment that matches the moments of the variance estimator to one of a chi-squared distribution. The dof adjustment depends on the sample size and the joint distribution of the covariates, and differs by covariate. We discuss the connection to the Behrens-Fisher problem, and suggest a minor modification for the case with clustering.

References

Angrist, J. D., and J.-S. Pischke (2009): *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press, Princeton, NJ.

Behrens, W. U. (1929): "Ein Beitrag zur Fehlerberechnung bei wenigen Beobachtungen," *Landwirtschaftliche Jahrbücher*, 68, 807–837.

Bell, R. M., and D. F. McCaffrey (2002): "Bias Reduction in Standard Errors for Linear Regression with Multi-Stage Samples," *Survey Methodology*, 28(2), 169–181.

Cameron, A. C., J. B. Gelbach, and D. L. Miller (2008): "Bootstrap-based improvements for inference with clustered errors," *The Review of Economics and Statistics*, 90(3), 414–427.

Chesher, A., and I. Jewitt (1987): "The bias of a heteroskedasticity consistent covariance matrix estimator," *Econometrica*, 55(5), 1217–1222.

Davidson, R., and E. Flachaire (2008): "The wild bootstrap, tamed at last," *Journal of Econometrics*, 146(1), 162–169.

Diggle, P., P. Heagerty, K.-Y. Liang, and S. L. Zeger (2002): *Analysis of longitudinal data*. Oxford University Press, Oxford.

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.

Eicker, F. (1967): "Limit theorems for regressions with unequal and dependent errors," in *Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability*, ed. by L. M. Le Cam, and J. Neyman, vol. 1, pp. 59–82, Berkeley, CA. University of California Press.

Fisher, R. A. (1939): "The comparison of samples with possibly unequal variances," *Annals of Eugenics*, 9(2), 174–180.

Hausman, J. A., and C. J. Palmer (2011): "Heteroskedasticity-Robust Inference in Finite Samples," NBER Working Paper 17698.

Horn, S. D., R. A. Horn, and D. B. Duncan (1975): "Estimating heteroscedastic variances in linear models," *Journal of the American Statistical Association*, 70(350), 380–385.

Huber, P. J. (1967): "The behavior of maximum likelihood estimates under nonstandard conditions," in *Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability*, ed. by L. M. Le Cam, and J. Neyman, vol. 1, pp. 221–233, Berkeley, CA. University of California Press.

IBRAGIMOV, R., AND U. K. MÜLLER (2010): "t-statistic based correlation and heterogeneity robust inference," *Journal of Business & Economic Statistics*, 28(4), 453–468.

IMBENS, G. W., AND M. KOLESÁR (2012): "Robust Standard Errors in Small Samples: Some Practical Advice," NBER Working Paper 18478.

LEHMANN, E. L., AND J. P. ROMANO (2005): *Testing statistical hypotheses*. Springer, third edn.

LIANG, K.-Y., AND S. L. ZEGER (1986): "Longitudinal data analysis for generalized linear models," *Biometrika*, 73(1), 13–22.

LIU, R. Y. (1988): "Bootstrap procedures under some non-iid models," *The Annals of Statistics*, 16(4), 1696–1708.

MACKINNON, J. G. (2002): "Bootstrap inference in econometrics," *The Canadian Journal of Economics / Revue Canadienne d'Economique*, 35(4), 615–645.

——— (2012): "Thirty Years of Heteroskedasticity-Robust Inference," in *Recent Advances and Future Directions in Causality, Prediction, and Specification Analysis*, ed. by X. Chen, and N. R. Swanson, pp. 437–461. Springer, New York.

MACKINNON, J. G., AND H. WHITE (1985): "Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties," *Journal of Econometrics*, 29(3), 305–325.

MAMMEN, E. (1993): "Bootstrap and wild bootstrap for high dimensional linear models," *The Annals of Statistics*, 21(1), 255–285.

MOULTON, B. R. (1986): "Random group effects and the precision of regression estimates," *Journal of Econometrics*, 32(3), 385–397.

——— (1990): "An Illustration of a Pitfall in Estimating the Effects of Aggregate Variables on Micro Units," *The Review of Economics and Statistics*, 72(2), 334–338.

SATTERTHWAITE, F. E. (1946): "An Approximate Distribution of Estimates of Variance Components," *Biometrics Bulletin*, 2(6), 110–114.

SCHEFFÉ, H. (1970): "Practical solutions of the Behrens-Fisher problem," *Journal of the American Statistical Association*, 65(332), 1501–1508.

WANG, Y. Y. (1971): "Probabilities of the Type I Errors of the Welch Tests for the Behrens-Fisher Problem," *Journal of the American Statistical Association*, 66(335), 605–608.

WELCH, B. L. (1951): "On the comparison of several mean values: An alternative approach," *Biometrika*, 38(3), 330–336.

WHITE, H. (1980): "A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity," *Econometrica*, 48(4), 817–838.

Appendix A Other methods

A.1 HC3

A second alternative to the EHW variance estimator is $\hat{\mathbb{V}}_{\text{HC3}}$. We use the version discussed in MacKinnon (2012):

$$\hat{\mathbb{V}}_{\text{HC3}} = \left(\sum_{i=1}^{N} X_i X_i'\right)^{-1} \left(\sum_{i=1}^{N} \frac{\left(Y_i - X_i\hat{\beta}\right)^2}{(1-P_{ii})^2} X_i X_i'\right) \left(\sum_{i=1}^{N} X_i X_i'\right)^{-1}. \tag{A.1}$$

Compared to $\hat{\mathbb{V}}_{\text{HC2}}$ this variance estimator has the square of $1 - P_{ii}$ in the denominator. In the binary regressor case this leads to:

$$\hat{\mathbb{V}}_{\text{HC3}} = \sigma^2(0)\frac{N_0}{(N_0-1)^2} + \sigma^2(1)\frac{N_1}{(N_1-1)^2}.$$

In simple cases this leads to an upwardly biased estimator for the variance.

A.2 Wild bootstrap

Although the confidence intervals based on the standard nonparametric bootstrap (where we resample N units picked with replacement from the original sample) have better coverage than the EHW confidence intervals, they can still suffer from substantial undercoverage if the distribution of the regressors is skewed or if the sample size is small (see, for instance, MacKinnon (2002) or Cameron, Gelbach, and Miller (2008) for simulation evidence). The problem is that the additional noise introduced by variation in the regressors adversely affects the the properties of the corresponding confidence intervals. Researchers have therefore focused on alternative resampling methods. One that has been proposed as an attractive choice is the wild bootstrap (Liu, 1988; Mammen, 1993; Cameron, Gelbach, and Miller, 2008; Davidson and Flachaire, 2008; MacKinnon, 2002, 2012).

There are several ways to implement the wild bootstrap. Here we focus on two methods based on resampling the t statistic. We first describe the two methods in the regression setting, and then in the cluster setting.

Suppose that we wish to test the hypothesis that $H_0\colon \beta_\ell = \beta_\ell^0$. Let $\hat{\beta}$ be the least squares estimate in the original sample, let $\hat{\varepsilon} = Y_i - X_i'\hat{\beta}$ be the estimated residuals, and let $\hat{\mathbb{V}}$ be a variance estimator, either $\hat{\mathbb{V}}_{\text{EHW}}$, or $\hat{\mathbb{V}}_{\text{HC2}}$, or $\hat{\mathbb{V}}_{\text{HC3}}$. Let $\hat{t} = (\hat{\beta}_\ell - \beta_\ell^0)/\sqrt{\hat{\mathbb{V}}}$ denote the t-statistic.

In the wild bootstrap the regressor values are fixed in the resampling. For the first method, the value of the i-th outcome in the bth bootstrap replication is redrawn as

$$Y_{i,b} = X_i'\hat{\beta}_1 + U_{i,b} \cdot \hat{\varepsilon}_i,$$

where $U_{i,b}$ is a binary random variable with $\mathrm{pr}(U_{i,b} = 1) = \mathrm{pr}(U_{i,b} = -1) = 1/2$, with $U_{i,b}$ independent across i and b. (Other distributions for $U_{i,b}$ are also possible; we focus on this particular choice following Cameron, Gelbach, and Miller (2008).) The second method we consider "imposes the null" when redrawing the outcomes. In particular, letting $\tilde{\beta}(\beta_\ell^0)$ denote the value of the restricted least squares estimate that minimizes the sum of squared residuals subject to $\beta_\ell = \beta_\ell^0$. Then the i-th outcome in the bth bootstrap replication is redrawn as

$$Y_{i,b} = X_i'\tilde{\beta}(\beta_\ell^0) + U_{i,b} \cdot (Y_i - X_i'\tilde{\beta}(\beta_\ell^0))$$

Once the new outcomes are redrawn, for each bootstrap sample $(Y_{i,b}, X_i)_{i=1}^n$, calculate the t-statistic as

$$t_b^1 = \frac{\hat{\beta}_{b,\ell} - \hat{\beta}_\ell}{\sqrt{\hat{\mathbb{V}}_b}},$$

if using the first method, or as

$$t_b^2(\beta_\ell^0) = \frac{\hat{\beta}_{b,\ell} - \beta_\ell^0}{\sqrt{\hat{\mathbb{V}}_b}},$$

if using the second method, where $\hat{\mathbb{V}}_b$ is some variance estimator. We focus on a symmetric version of the critical values. In particular, over all the bootstrap samples, set the critical value to $q_{0.95}(|t^1|)$, the 0.95 quantile of the distribution of $|t_b^1|$ (or $q_{0.95}(|t^2(\beta_0)|)$ if using the second method). Reject the null if $|\hat{t}|$ is greater than the critical value.

The first method does not impose the null hypothesis when redrawing the outcomes, or calculating the critical value, so that $q_{0.95}(|t^1|)$ does not depend on which β_ℓ^0 is being tested. Therefore, to construct a 95% confidence interval, we simply replace the standard 1.96 critical value by $q_{0.95}^{\text{wild}}$,

$$\mathrm{CI}_{\text{wild}}^{95\%} = \left(\hat{\beta}_\ell - q_{0.95}(|t^1|) \times \sqrt{\hat{\mathbb{V}}}, \hat{\beta}_1 + q_{0.95}(|t^1|) \times \sqrt{\hat{\mathbb{V}}}\right). \tag{A.2}$$

We denote this confidence interval as "wild" in the simulations.For the second method, the confidence interval consists of all points b such that the null $H_0\colon\, \beta_\ell = b$ is not rejected:

$$\mathrm{CI}^{95\%}_{\mathrm{wild0}} = \left\{ b\colon\, |\hat\beta_\ell - b|/\sqrt{\hat{\hat{\mathbb{V}}}} \le q^{\mathrm{wild}}_{0.95}(t^2(b)) \right\}.$$

We denote this confidence interval as "wild_0" in the simulations. Because constructing this confidence interval involves testing many null hypotheses, the method it is computationally intensive. The wild bootstrap standard errors reported in the tables defined as the length of the bootstrap confidence interval divided by 2×1.96.

For the cluster version of the wild bootstrap, the bootstrap variable $U_{s,b}$ is indexed by the cluster only. Again the distribution of $U_{s,b}$ is binary with values -1 and 1, and probability $\mathrm{pr}(U_{s,b} = 1) = \mathrm{pr}(U_{s,b} = -1) = 0.5$. The bootstrap value for the outcome for unit i in cluster s is then

$$Y_{is,b} = X_{is}'\hat\beta + U_{s,b} \cdot \hat\varepsilon_{is}$$

for the first method, and

$$Y_{is,b} = X_{is}'\tilde\beta(\beta_{0,\ell}) + U_{s,b} \cdot (Y_{is} - X_{is}'\tilde\beta(\beta_{0,\ell}))$$

for the second method that imposes the null, with the covariates X_{is} remaining fixed across the bootstrap replications.

Table 1: *Coverage rates and normalized standard errors for different confidence intervals in the Behrens-Fisher problem. Angrist-Pischke unbalanced design, $N_0 = 27$, $N_1 = 3$, Normal errors.*

		I		II		III		IV		V	
	$\sigma(0)$	0.5		0.85		1		1.18		2	
Panel 1: Coverage rates and median standard errors											
variance estimator	dist/dof	cov rate	med s.e.	cov rate	med s.e.	cov rate	med s.e.	cov rate	med s.e.	cov rate	med s.e.
$\hat{\mathbb{V}}_{\text{homo}}$	∞	72.5	0.33	90.2	0.52	94.0	0.60	96.7	0.70	99.8	1.17
	$N-2$	74.5	0.34	91.5	0.54	95.0	0.63	97.4	0.73	99.8	1.22
$\hat{\mathbb{V}}_{\text{EHW}}$	∞	76.8	0.40	79.3	0.42	80.5	0.44	81.8	0.45	86.6	0.55
	$N-2$	78.3	0.42	80.9	0.44	82.0	0.46	83.3	0.47	88.1	0.57
	wild	89.6	0.73	89.4	0.70	89.6	0.69	89.9	0.68	91.8	0.69
	wild$_0$	89.7	0.51	97.5	0.74	98.7	0.85	99.5	0.98	99.9	1.64
$\hat{\mathbb{V}}_{\text{HC2}}$	∞	82.5	0.49	84.4	0.51	85.2	0.52	86.2	0.53	89.8	0.62
	$N-2$	83.8	0.51	85.6	0.53	86.5	0.54	87.4	0.56	91.0	0.65
	wild	90.5	0.76	90.4	0.74	90.5	0.73	90.8	0.72	92.4	0.72
	wild$_0$	89.7	0.51	97.5	0.74	98.7	0.85	99.4	0.98	99.9	1.64
	K^*_{Welch}	96.1	1.02	96.8	0.98	97.0	0.95	97.1	0.93	96.7	0.87
	K_{Welch}	93.1	1.00	92.5	0.93	92.4	0.90	92.5	0.87	93.5	0.80
	K_{BM}	94.7	0.90	96.4	0.94	97.0	0.95	97.6	0.98	99.1	1.14
$\hat{\mathbb{V}}_{\text{HC3}}$	∞	87.2	0.60	88.6	0.61	89.2	0.62	89.9	0.63	92.4	0.71
	$N-2$	88.2	0.62	89.5	0.64	90.1	0.65	90.8	0.66	93.4	0.74
$\max(\hat{\mathbb{V}}_{\text{homo}}, \hat{\mathbb{V}}_{\text{EHW}})$	∞	82.2	0.41	91.8	0.54	94.7	0.62	97.0	0.71	99.8	1.17
$\max(\hat{\mathbb{V}}_{\text{homo}}, \hat{\mathbb{V}}_{\text{HC2}})$	∞	86.1	0.49	93.2	0.57	95.4	0.64	97.3	0.73	99.8	1.17
Panel 2: Mean effective dof											
	K^*_{Welch}	2.1		2.3		2.5		2.7		4.1	
	K_{Welch}	2.8		3.8		4.4		5.1		8.6	
	K_{BM}	2.5		2.5		2.5		2.5		2.5	

Notes: "cov. rate" refers to coverage of nominal 95% confidence intervals (in percentages), and "med. s.e." refers to standard errors normalized by $t^{K}_{0.975}/t^{\infty}_{0.975}$. Variance estimators and degrees-of-freedom (dof) adjustments are described in the text, wild bootstrap confidence intervals ("wild" and "wild$_0$") are described in Appendix A.2.
Results are based on 1,000,000 replications, except for wild wild bootstrap-based confidence intervals, which use 100,000 replications, and 1,000 bootstrap draws in each replication.

Table 2: *Coverage rates and normalized standard errors for different confidence intervals in the Behrens-Fisher problem. Angrist-Pischke unbalanced design, $N_0 = 27$, $N_1 = 3$, log-Normal errors.*

		I		II		III		IV		V	
	$\sigma(0)$	0.5		0.85		1		1.18		2	
variance estimator	dist/dof	cov rate	med s.e.	cov rate	med s.e.	cov rate	med s.e.	cov rate	med s.e.	cov rate	med s.e.
$\hat{\mathbb{V}}_{\text{homo}}$	∞	75.9	0.26	91.8	0.41	93.3	0.47	94.4	0.55	97.0	0.91
	$N-2$	78.2	0.27	92.6	0.43	93.9	0.49	94.9	0.58	97.3	0.96
$\hat{\mathbb{V}}_{\text{EHW}}$	∞	66.8	0.22	73.4	0.26	76.7	0.27	80.6	0.30	91.1	0.41
	$N-2$	68.2	0.23	75.0	0.27	78.4	0.29	82.4	0.31	92.6	0.42
	wild	76.1	0.36	78.8	0.36	81.1	0.36	84.2	0.38	94.3	0.46
	wild_0	95.2	0.42	99.0	0.63	99.1	0.72	99.2	0.83	99.5	1.36
$\hat{\mathbb{V}}_{\text{HC2}}$	∞	71.3	0.26	77.2	0.29	80.2	0.31	83.7	0.33	93.3	0.44
	$N-2$	72.5	0.27	78.6	0.31	81.7	0.33	85.3	0.35	94.5	0.46
	wild	77.2	0.38	79.7	0.38	81.8	0.38	84.7	0.39	94.4	0.48
	wild_0	95.2	0.42	99.0	0.63	99.2	0.72	99.3	0.83	99.5	1.36
	K_{Welch}	79.9	0.47	82.2	0.44	84.3	0.44	87.1	0.44	95.7	0.52
	K^*_{Welch}	90.1	0.54	95.8	0.57	97.2	0.57	98.3	0.58	98.9	0.62
	K_{BM}	87.2	0.48	94.9	0.54	97.2	0.57	98.8	0.61	99.7	0.81
$\hat{\mathbb{V}}_{\text{HC3}}$	∞	75.4	0.31	80.6	0.34	83.2	0.36	86.4	0.38	94.9	0.49
	$N-2$	76.5	0.33	81.9	0.36	84.6	0.37	87.8	0.40	95.9	0.51
$\max(\hat{\mathbb{V}}_{\text{homo}}, \hat{\mathbb{V}}_{\text{EHW}})$	∞	85.7	0.30	97.8	0.44	98.4	0.50	98.6	0.58	98.8	0.93
$\max(\hat{\mathbb{V}}_{\text{homo}}, \hat{\mathbb{V}}_{\text{HC2}})$	∞	86.9	0.33	98.5	0.46	99.0	0.52	99.2	0.60	99.3	0.94

Panel 2: Mean effective dof

	I	II	III	IV	V
K^*_{Welch}	2.1	2.3	2.5	2.7	4.1
K_{Welch}	4.9	7.5	8.5	9.7	14.0
K_{BM}	2.5	2.5	2.5	2.5	2.5

Notes: "cov. rate" refers to coverage of nominal 95% confidence intervals (in percentages), and "med. s.e." refers to standard errors normalized by $t^K_{0.975}/t^\infty_{0.975}$. Variance estimators and degrees-of-freedom (dof) adjustments are described in the text, wild bootstrap confidence intervals ("wild" and "wild_0") are described in Appendix A.2.

Results are based on 1,000,000 replications, except for wild wild bootstrap-based confidence intervals, which use 100,000 replications, and 1,000 bootstrap draws in each replication.

Table 3: *Coverage rates and normalized standard errors for different confidence intervals in the Behrens-Fisher problem. Angrist-Pischke balanced design, $N_0 = 15$, $N_1 = 15$, Normal errors.*

		I		II		III		IV		V	
	$\sigma(0)$	0.5		0.85		1		1.18		2	
variance estimator	dist/dof	cov rate	med s.e.	cov rate	med s.e.	cov rate	med s.e.	cov rate	med s.e.	cov rate	med s.e.
$\hat{\mathbb{V}}_{\text{homo}}$	∞	93.7	0.28	94.0	0.33	94.0	0.36	94.0	0.39	93.7	0.57
	$N-2$	94.7	0.30	95.0	0.35	95.0	0.38	95.0	0.41	94.7	0.59
$\hat{\mathbb{V}}_{\text{EHW}}$	∞	92.8	0.27	93.1	0.32	93.1	0.35	93.1	0.38	92.8	0.55
	$N-2$	93.9	0.29	94.2	0.34	94.2	0.36	94.2	0.40	93.9	0.57
	wild	94.9	0.30	94.9	0.35	95.0	0.38	95.0	0.41	94.9	0.60
	wild_0	94.8	0.30	95.0	0.35	95.0	0.38	95.0	0.41	94.9	0.60
$\hat{\mathbb{V}}_{\text{HC2}}$	∞	93.7	0.28	94.0	0.33	94.0	0.36	94.0	0.39	93.7	0.57
	$N-2$	94.7	0.30	95.0	0.35	95.0	0.38	95.0	0.41	94.7	0.59
	wild	94.9	0.30	94.8	0.35	94.8	0.38	94.8	0.41	94.8	0.60
	wild_0	94.9	0.30	95.0	0.35	94.9	0.38	95.0	0.41	94.8	0.60
	K_{Welch}	95.0	0.30	95.1	0.35	95.1	0.38	95.1	0.41	95.0	0.60
	K^*_{Welch}	95.0	0.30	95.0	0.35	95.0	0.38	95.0	0.41	95.0	0.60
	K_{BM}	94.7	0.30	95.0	0.35	95.0	0.38	95.0	0.41	94.7	0.59
$\hat{\mathbb{V}}_{\text{HC3}}$	∞	94.5	0.29	94.8	0.35	94.8	0.37	94.8	0.41	94.5	0.59
	$N-2$	95.4	0.31	95.7	0.36	95.7	0.39	95.7	0.43	95.4	0.61
$\max(\hat{\mathbb{V}}_{\text{homo}}, \hat{\mathbb{V}}_{\text{EHW}})$	∞	93.7	0.28	94.0	0.33	94.0	0.36	94.0	0.39	93.7	0.57
$\max(\hat{\mathbb{V}}_{\text{homo}}, \hat{\mathbb{V}}_{\text{HC2}})$	∞	93.7	0.28	94.0	0.33	94.0	0.36	94.0	0.39	93.7	0.57

Panel 2: Mean effective dof

		I	II	III	IV	V
	K^*_{Welch}	20.6	27.3	26.4	27.3	20.6
	K_{Welch}	21.0	26.0	28.0	26.0	21.0
	K_{BM}	28.0	28.0	28.0	28.0	28.0

Notes: "cov. rate" refers to coverage of nominal 95% confidence intervals (in percentages), and "med. s.e." refers to standard errors normalized by $t^K_{0.975}/t^\infty_{0.975}$. Variance estimators and degrees-of-freedom (dof) adjustments are described in the text, wild bootstrap confidence intervals ("wild" and "wild_0") are described in Appendix A.2.
Results are based on 1,000,000 replications, except for wild wild bootstrap-based confidence intervals, which use 100,000 replications, and 1,000 bootstrap draws in each replication.

Table 4: *Coverage rates and normalized standard errors for different confidence intervals with clustering. Cameron-Gelbach-Miller designs with 10 clusters.*

		I		II		III		IV		V	
		Baseline		5 clusters		Unbalanced cluster size		Hetero-skedasticity		X_i fixed in cluster	
Panel 1: Coverage rates and median standard errors											
variance estimator	dist/dof	cov rate	med s.e.	cov rate	med s.e.	cov rate	med s.e.	cov rate	med s.e.	cov rate	med s.e.
$\hat{\mathbb{V}}_{\text{homo}}$	∞	51.3	0.06	53.0	0.08	46.6	0.06	71.1	0.18	36.1	0.06
$\hat{\mathbb{V}}_{\text{LZ}}$	∞	84.7	0.12	73.9	0.13	79.6	0.12	85.7	0.26	81.7	0.18
	$S-1$	89.5	0.14	86.9	0.19	85.2	0.14	90.2	0.31	86.4	0.21
	wild	92.5	0.17	89.8	0.28	90.2	0.18	92.6	0.36	88.7	0.26
	wild$_0$	94.2	0.17	94.0	1.33	93.4	0.17	94.3	0.36	94.3	0.37
$\hat{\mathbb{V}}_{\text{STATA}}$	∞	86.7	0.13	78.8	0.15	81.9	0.13	87.6	0.28	83.6	0.19
	$S-1$	91.1	0.15	90.3	0.21	87.2	0.15	91.8	0.32	88.1	0.22
$\hat{\mathbb{V}}_{\text{LZ2}}$	∞	89.2	0.14	84.7	0.17	87.2	0.15	89.1	0.29	87.7	0.22
	$S-1$	93.0	0.16	93.6	0.24	91.3	0.17	92.8	0.34	91.4	0.26
	wild	92.6	0.18	90.9	0.29	91.2	0.19	92.8	0.37	88.6	0.27
	wild$_0$	94.0	0.17	93.9	1.33	93.7	0.18	94.4	0.36	94.3	0.37
	$K^*_{\text{Satt.}}$	96.9	0.20	97.7	0.34	97.9	0.25	96.2	0.40	96.6	0.35
	K_{BM}	94.4	0.17	95.3	0.27	94.4	0.19	94.2	0.36	96.6	0.35
	K_{IK}	96.7	0.20	97.1	0.33	97.4	0.24	94.7	0.37	96.6	0.35
Panel 2: Mean effective dof											
	$K^*_{\text{Satt.}}$	4.0		2.3		2.9		4.6		3.4	
	K_{BM}	6.6		3.3		5.1		6.6		3.4	
	K_{IK}	4.1		2.4		3.1		5.7		3.4	

Notes: "cov. rate" refers to coverage of nominal 95% confidence intervals (in percentages), and "med. s.e." refers to standard errors normalized by $t^K_{0.975}/t^\infty_{0.975}$. Variance estimators and degrees-of-freedom (dof) adjustments are described in the text, wild bootstrap confidence intervals ("wild" and "wild$_0$") are described in Appendix A.2.
Results are based on 100,000 replications, except for wild wild bootstrap-based confidence intervals, which use 10,000 replications, and 500 bootstrap draws in each replication.

Table 5: *Coverage rates and normalized standard errors for different confidence intervals with clustering. 50 clusters.*

		VI		VII		VIII		IX		X	
		Baseline, balanced covariates		Log-normal regressors		3 treated clusters, $\frac{\sigma_\nu(1)}{\sigma_\nu(0)}=1$		3 treated clusters, $\frac{\sigma_\nu(1)}{\sigma_\nu(0)}=2$		3 treated clusters, $\frac{\sigma_\nu(1)}{\sigma_\nu(0)}=1/2$	
Panel 1: Coverage rates and median standard errors											
variance estimator	dist/dof	cov rate	med s.e.	cov rate	med s.e.	cov rate	med s.e.	cov rate	med s.e.	cov rate	med s.e.
$\hat{\mathbb{V}}_{\text{homo}}$	∞	80.8	0.06	69.8	0.04	43.6	0.35	69.9	0.34	86.1	0.52
$\hat{\mathbb{V}}_{\text{LZ}}$	∞	93.0	0.08	86.0	0.07	76.2	0.82	78.6	0.45	83.6	0.52
	$S-1$	93.7	0.08	86.9	0.07	77.0	0.84	79.5	0.46	84.4	0.53
	wild	94.3	0.09	86.9	0.08	89.9	1.52	89.0	0.76	90.4	0.71
	wild_0	94.4	0.09	95.4	0.12	94.3	1.23	99.8	1.12	99.9	2.08
$\hat{\mathbb{V}}_{\text{STATA}}$	∞	93.4	0.08	86.5	0.07	76.7	0.83	79.1	0.46	84.1	0.52
	$S-1$	94.0	0.09	87.3	0.07	77.5	0.85	80.0	0.47	84.9	0.54
$\hat{\mathbb{V}}_{\text{LZ2}}$	∞	93.7	0.08	89.7	0.08	82.0	0.99	83.8	0.54	87.4	0.60
	$S-1$	94.3	0.09	90.3	0.08	82.7	1.02	84.5	0.56	88.1	0.61
	wild	94.1	0.09	86.4	0.08	90.6	1.55	90.2	0.80	91.3	0.76
	wild_0	94.3	0.09	95.6	0.12	94.5	1.23	99.7	1.11	99.9	2.08
	$K^*_{\text{Satt.}}$	95.3	0.09	97.1	0.11	95.7	2.11	96.6	1.07	98.8	1.19
	K_{BM}	94.7	0.09	97.1	0.11	94.9	1.95	96.6	1.07	98.7	1.18
	K_{IK}	95.2	0.09	97.1	0.11	94.9	1.95	96.6	1.07	98.7	1.18
Panel 2: Mean effective dof											
	$K^*_{\text{Satt.}}$	20		5.4		2.1		2.3		2.2	
	K_{BM}	28		5.4		2.3		2.3		2.3	
	K_{IK}	20		5.4		2.3		2.3		2.3	

Notes: "cov. rate" refers to coverage of nominal 95% confidence intervals (in percentages), and "med. s.e." refers to standard errors normalized by $t^K_{0.975}/t^\infty_{0.975}$. Variance estimators and degrees-of-freedom (dof) adjustments are described in the text, wild bootstrap confidence intervals ("wild" and "wild_0") are described in Appendix A.2.
Results are based on 100,000 replications, except for wild wild bootstrap-based confidence intervals, which use 10,000 replications, and 500 bootstrap draws in each replication.