NBER WORKING PAPER SERIES

ROBUST STANDARD ERRORS IN SMALL SAMPLES: SOME PRACTICAL ADVICE

Guido W. Imbens
Michal Kolesar

Working Paper 18478
http://www.nber.org/papers/w18478

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
October 2012

Financial support for this research was generously provided through NSF grant 0820361. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Robust Standard Errors in Small Samples: Some Practical Advice
Guido W. Imbens and Michal Kolesar
NBER Working Paper No. 18478
October 2012
JEL No. C01

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 the 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 sample sizes of fifty and more. We recommend researchers calculate the Bell-McCaffrey degrees-of-freedom adjustment to assess potential problems with conventional robust standard errors and use the modification as a matter of routine.

Guido W. Imbens
Department of Economics
Littauer Center
Harvard University
1805 Cambridge Street
Cambridge, MA 02138
and NBER
Imbens@stanford.edu

Michal Kolesar
Department of Economics
Harvard University
kolesarm@nber.org

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 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), 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 on higher order expansions (Hausman and Palmer, 2011), others attempt to improve their properties by using resampling methods (MacKinnon and White, 1995; Cameron, Gelbach, and Miller, 2008; Hausman and Palmer, 2011), and some use t-distribution approximations (Bell and McCaffrey, 2002; Donald and Lang, 2007). Few of these alternatives are regularly employed in empirical work. In fact, some researchers argue that for commonly encountered sample sizes (e.g., fifty or more units) the improvements are not necessary because the EHW and LZ standard errors perform well. Perhaps it is also the multitude and the *ad hoc* nature of many of the proposed alternatives to the EHW and LZ procedures, combined with the lack of clear guidance among them, that makes empirical researchers wary of using any of them.

Here we review some of this work and provide specific recommendations for empirical practice. The main recommendation of this paper is that empirical researchers should, as a matter of routine, adopt a particular improvement to the EHW and LZ confidence intervals, due to Bell and McCaffrey (2002), BM from hereon. 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. Here is a brief description of 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 MacKinnon and White, 1985), 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 are typically fixed at the number of observations minus the number of estimated regression parameters. The BM adjustment uses a more sophisticated choice for the degrees of freedom (dof). The dof of the approximating t-distribution, denoted by K_{BM}, is choosen 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}. The BM degrees of freedom is a simple analytic function of the matrix of regressors.

To ease comparisons with other methods we convert this procedure into one that only adjusts the standard errors. What we then refer to as the BM standard error is then $\sqrt{\hat{\mathbb{V}}_{\text{BM}}} = \sqrt{\hat{\mathbb{V}}_{\text{HC2}}} \cdot (t_{0.975}^{K_{\text{BM}}}/t_{0.975}^{\infty})$, where t_q^K is the q-th quantile of the t-distribution with dof equal to K (so that t_q^{∞} is the q-th quantile of the normal distribution and thus $t_{0.975}^{\infty} = 1.96$). A key insight is that the BM dof can differ substantially from the sample size (minus the number of estimated parameters) if the distribution of the covariates is skewed.

We make three specific points in the current paper. One modest novel contribution is to show that the BM proposal is the principled extension from 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. 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 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.

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. In each of these three sections 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: the 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), Fischer (1939), and for a general discussion Scheffé (1970), Wang (1971), Lehman 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 the following linear model with a single binary regressor, allowing for heteroskedasticity:

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

with $X_i \in \{0, 1\}$, and

$$\mathbb{E}[\varepsilon_i | X_i = x] = 0, \qquad \text{and} \quad \text{Var}(\varepsilon_i | X_i = x) = \sigma^2(x).$$

We are interested in $\beta_1 = \text{Cov}(Y_i, X_i)/\text{Var}(X_i) = \mathbb{E}[Y_i|X_i = 1] - \mathbb{E}[Y_i|X_i = 0]$. Because the regressor X_i is binary, the least squares estimator for the slope coefficient β_1 can be written as

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

where, for $x = 0, 1$,

$$\overline{Y}_x = \frac{1}{N_x} \sum_{i: X_i = x} Y_i, \qquad \text{and} \quad N_1 = \sum_{i=1}^{N} X_i, \qquad N_0 = \sum_{i=1}^{N} (1 - X_i).$$

Over repeated samples, conditional on $\mathbf{X} = (X_1, \ldots, X_N)'$, the exact finite sample variance for the least squares estimator $\hat{\beta}_1$ is

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

If in addition we assume normality for ε_i given X_i, $\varepsilon_i | X_i = x \sim \mathcal{N}(0, \sigma^2(x))$, the exact distribution for $\hat{\beta}_1$ conditional on $\mathbf{X}$ is

$$\hat{\beta}_1 | \mathbf{X} \sim \mathcal{N}\left(\beta_1, \frac{\sigma^2(0)}{N_0} + \frac{\sigma^2(1)}{N_1}\right).$$

The problem of how to do inference for β_1 in the absence of knowledge of $\sigma^2(x)$ 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 The 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

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

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 for ε_i given X_i, the t-statistic has an exact t-distribution:

$$t_{\text{homo}} = \left.\frac{\hat\beta_1 - \beta_1}{\sqrt{\hat\sigma^2/N_0 + \hat\sigma^2/N_1}}\right| \mathbf{X} \sim t(N-2). \tag{2.2}$$

This t-distribution with dof equal to $N-2$ can be used for the construction of exact confidence intervals. The exact 95% confidence interval for $\hat\beta_1$, under homoskedasticity, is

$$95\%\ \text{CI}_{\text{homo}} = \left(\hat\beta_1 + t^{N-2}_{0.025} \times \sqrt{\hat{\mathbb{V}}}_{\text{homo}}, \hat\beta_1 + t^{N-2}_{0.975} \times \sqrt{\hat{\mathbb{V}}}_{\text{homo}}\right),$$

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 The Robust EHW Variance Estimator

The familiar form of the robust EHW variance estimator, given the linear model in (2.1), is

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

In the Behrens-Fisher case with a single binary regressor the component of this matrix corresponding to β_1 simplifies to

$$\hat{\mathbb{V}}_{\text{ehw}} = \frac{\tilde\sigma^2(0)}{N_0} + \frac{\tilde\sigma^2(1)}{N_1}, \qquad \text{where } \tilde\sigma^2(x) = \frac{1}{N_x}\sum_{i:X_i=x}^{N}\left(Y_i - \overline{Y}_x\right)^2, \qquad \text{for } x = 0, 1.$$

The standard, normal-distribution-based, 95% confidence interval based on the robust variance estimator:

$$95\%\ \text{CI}_{\text{ehw}} = \left(\hat\beta_1 - 1.96 \times \sqrt{\hat{\mathbb{V}}}_{\text{ehw}}, \hat\beta_1 + 1.96 \times \sqrt{\hat{\mathbb{V}}}_{\text{ehw}}\right). \tag{2.3}$$

Even if the error term ε_i has a normal distribution, the justification for this confidence interval is asymptotic. Sometimes researchers use a t-distribution with $N-2$ degrees of freedome to calculate the confidence limits, replacing 1.96 in (2.3) by the corresponding quantile of the t-distribution with dof equal to $N-2$, $t_{0.975}^{N-2}$. However, there are no assumptions under which this modification has exact 95% coverage.

2.4 An Unbiased Estimator for the Variance

An alternative to $\hat{\mathbb{V}}_{\mathrm{ehw}}$ is what MacKinnon and White (1985) call the HC2 variance estimator, here denoted by $\hat{\mathbb{V}}_{\mathrm{HC2}}$. 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

$$\hat{\mathbb{V}}_{\mathrm{HC2}} = \frac{\hat{\sigma}^2(0)}{N_0} + \frac{\hat{\sigma}^2(1)}{N_1}, \qquad \text{where } \hat{\sigma}^2(x) = \frac{1}{N_x - 1} \sum_{i:X_i=x}^{N} \left(Y_i - \overline{Y}_x\right)^2, \tag{2.4}$$

for $x = 0, 1$. These conditional variance estimators $\hat{\sigma}^2(x)$ differ from $\tilde{\sigma}^2(x)$ by a factor $N_x/(N_x - 1)$. In combination with the normal approximation to the distribution of the t-statistic, this variance estimator leads to the 95% confidence interval

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

The estimator $\hat{\mathbb{V}}_{\mathrm{HC2}}$ 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 $(\hat{\beta}_1 - \beta_1)/\sqrt{\hat{\mathbb{V}}_{\mathrm{HC2}}}$ 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) . However, whereas in the homoskedastic case the exact distribution is a t-distribution with degrees of freedom equal to $N-2$, the exact sampling distribution of $(\hat{\beta}_1 - \beta_1)/\sqrt{\hat{\mathbb{V}}_{\mathrm{HC2}}}$ does not lend itself to the construction of exact confidence intervals.

In this single-binary-covariate 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 $X_i = 0$ and few units with $X_i = 1$ ($N_0 >> N_1$). In that case $\mathbb{E}[Y_i|X_i = 0]$ is estimated relatively precisely, with variance $\sigma^2(0)/N_0 \approx 0$. As a result the distribution of the t-statistic $(\hat{\beta}_1 - \beta_1)/\sqrt{\hat{\mathbb{V}}_{\mathrm{HC2}}}$ is approximately equal to that of $(\overline{Y}_1 - \mathbb{E}[Y_i|X_i = 1])/\sqrt{\hat{\sigma}^2(1)/N_1}$. 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 $N_0 >> N_1$.

2.5 Degrees of Freedom Adjustment: The Welch and Bell-McCaffrey Solutions

One of the most attractive proposals for 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. Rather than using the sample size minus two as the degrees of freedom for this t-distribution, he suggests using moments of the variance estimator $\hat{\mathbb{V}}_{\text{HC2}}$ to determine the most appropriate value for the degrees of freedom.

Let us describe the Welch suggestion in more detail. Consider the t-statistic in the heteroskedastic case:

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

Note that $\mathbb{E}[\hat{\mathbb{V}}_{\text{HC2}}] = \mathbb{V}$, and that under normality $\hat{\mathbb{V}}_{\text{HC2}}$ is independent of $\hat{\beta}_1 - \beta_1$. Now 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 t_{HC2} would have a t-distribution with dof equal to K, which could be exploited to construct an exact confidence interval. The problem is that there exists no such K such that the scaled distribution of the variance estimator has an exact chi-squared distribution. Welch suggests approximating the scaled distribution of $\hat{\mathbb{V}}_{\text{HC2}}$ by a chi-squared distribution, with the dof choosen to make the approximation as accurate as possible. Under normality, $\hat{\mathbb{V}}_{\text{HC2}}$ is a linear combination of two chi-squared random variables. To be precise, $(N_0 - 1)\hat{\sigma}^2(0)/\sigma^2(0) \sim \mathcal{X}^2(N_0 - 1)$, and $(N_1 - 1)\hat{\sigma}^2(1)/\sigma^2(1) \sim \mathcal{X}^2(N_1 - 1)$, 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

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

Welch' specific suggestion is 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

$$\text{Var}\left(K \cdot \hat{\mathbb{V}}_{\text{HC2}}/\mathbb{V}\right) = 2K, \quad \text{leading to } K^*_{\text{welch}} = \frac{2 \cdot \mathbb{V}^2}{\text{Var}\left(\hat{\mathbb{V}}_{\text{HC2}}\right)} = \frac{\left(\frac{\sigma^2(0)}{N_0} + \frac{\sigma^2(1)}{N_1}\right)^2}{\frac{\sigma^4(0)}{(N_0-1)N_0^2} + \frac{\sigma^4(1)}{(N_1-1)N_1^2}}.$$

This choice for K is not feasible because K^*_{welch} depends on unknown quantities, namely, the conditional variances $\sigma^2(x)$. 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 $\sigma^2(x)$ are replaced by the estimates $\hat{\sigma}^2(x)$. 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 (2002) propose a slightly different degrees of freedom adjustment. For the Behrens-Fisher problem (regression with a single binary covariate) the BM modification is minor, but it has considerable attraction in settings with more general distributions of covariates. The BM adjustment is based on assuming homoskedasticity. In that case the Welch dof simplifies 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 varies less across repeated samples and as a result tends to be more accurate then the Welch adjustment which can be conservative in settings with noisy estimates of the conditional error variances. The associated 95% confidence interval is now

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

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 $N_0 >> N_1$, 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 A Small Simulation Study based on a Angrist-Pischke design

Now let us 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 $\beta_0 = \beta_1 = 0$ (note that the results are invariant to the values for β_0 and β_1). The distribution of the disturbances is normal,

$$\varepsilon_i | X_i = x \sim \mathcal{N}(0, \sigma^2(x)), \qquad \text{for} \quad x = 0, 1,$$

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 the Welch or BM dof adjustments. For the five designs the Welch dof correction is quite substantial. Consider the first design with $\sigma(0)/\sigma(1) = 0.5$. Then the value for the infeasible Welch dof is $K^*_{\text{welch}} = 2.1$. Given that $t^2_{0.975} = 4.30$, compared to the normal 0.975 quantile $t^\infty_{0.975} = 1.96$, this leads to an adjustment in the standard errors by a factor of 2.2. For the other four designs the infeasible Welch dof values are equal to 2.3, 2.5, 2.7, and 4.1 respectively, in each case leading to substantial changes in the confidence intervals even though the overall sample size is substantial.

When we implement the Welch and BM degrees-of-freedom adjustments the adjusted-degrees-of-freedom are not necessarily integer. In that case we use the distribution for the t-distribution defined 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 $\alpha = K_{\text{BM}}/2$ and $\beta = 2$. We include 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, two confidence intervals based on $\hat{\mathbb{V}}_{\text{ehw}}$, again either using the normal or the t-distribution with $N - 2$ dof. Next, six confidence intervals based on $\hat{\mathbb{V}}_{\text{HC2}}$. First among these is the one with the normal distribution, next the t-distribution with degrees of freedom equal to $N - 2$, K_{welch}, K^*_{welch}, and K_{BM}. Finally, a resampling method, specifically the wild bootstrap, discussed in more detail in Appendix A. Next we include two confidence intervals based on $\hat{\mathbb{V}}_{\text{HC3}}$ (see Appendix A for more details), either using the normal distribution or the wild bootstrap. Finally we include normal distribution based confidence intervals based on the maximum of $\hat{\mathbb{V}}_{\text{homo}}$ and $\hat{\mathbb{V}}_{\text{ehw}}$, and one based on the maximum of $\hat{\mathbb{V}}_{\text{homo}}$ and $\hat{\mathbb{V}}_{\text{HC2}}$. For each of the analytic variance estimators we use 1,000,000 replications. For those based on the wild bootstrap we use 100,000 replications and 10,000 draws from the wild bootstrap distribution.

Table 1 presents the simulation results for the Angrist-Pischke design. For each of the variance estimators we report coverage properties for nominal 95% confidence intervals. We also report the median of the standard errors over the simulations. In cases where the confidence intervals are based on t-distributions with K degrees of freedom, we multiply the standard error by $t^{K}_{0.975}/t^{\infty}_{0.975}$, to make the standard errors comparable. For the variance estimators included in the Angrist-Pischke design our simulation 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 to all others. Consider the case with $\sigma(0) = 0.5$. The coverage rate for the normal-distribution confidence interval based on $\hat{\mathbb{V}}_{\text{ehw}}$ is 0.77. Using the unbiased variance estimator $\hat{\mathbb{V}}_{\text{HC2}}$ raise that to 0.82, but only using the t-distribution approximation with Welch or BM degrees of freedom gets that close to the nominal level. An interesting aspect of the Welch dof calculation is that it leads to confidence intervals that are typically substantially wider, and have substantially more variation in their width. For the two confidence intervals based on K_{welch} and K_{BM}, the median widths of the confidence intervals are 6.45 and 3.71, but the 0.95 quantile of the widths are 14.89 and 7.59. 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. One of the attractions of the BM intervals is that it avoids this variation.

For comparison purposes we report in Table 2 the results for a simulation exercise with a balanced design where $N_0 = N_1 = N/2 = 15$. Here the actual coverage rates are close to nominal coverage rates for essentially all procedures: for a sample size of 30 with a balanced design asymptotic, normal-distribution-based, approximations are fairly accurate and refinements are unnecessary. Note that $K_{\text{BM}} = 28$, and $t^{28}_{0.975} = 2.05$, close to the 1.96 for the normal distribution, so the BM dof correction is unlikely to be important here.

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 The Set Up

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

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

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:

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

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

$$\mathbb{V} = \operatorname{Var}\left(\hat{\beta}\middle|\mathbf{X}\right) = \left(\sum_{i=1}^{N} X_i X_i'\right)^{-1} \left(\sum_{i=1}^{N} \sigma^2(X_i) \cdot X_i X_i'\right) \left(\sum_{i=1}^{N} X_i X_i'\right)^{-1},$$

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

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

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

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

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

3.2 The Bias-adjusted Variance Estimator

In Section 2 we discussed the bias of the robust variance estimator in the case with a single binary covariate. In that case there was a simple modification of the EHW variance estimator that removes all bias. In the general regression case the bias-adjustment is more complicated. Here we focus on a particular adjustment for the bias due to MacKinnon and White (1985). In the special case with only a single binary covariate this adjustment is identical to that used in Section 2. It does not, however, remove all bias in general.

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

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

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 $\sigma^2 \cdot (1 - \mathbf{P}_{ii})$. 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 MacKinnon and White (1985) introduce as HC2:

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

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 $\hat{\mathbb{V}}_{\mathrm{HC2},k} = e_{L,k}'\hat{\mathbb{V}}_{\mathrm{HC2}}e_{L,k}$, where $e_{L,k}$ is an L-vector with kth element equal to one and all other elements equal to zero. The 95% confidence interval, based on the normal approximation, is then given by

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

3.3 The Degrees of Freedom Adjustment

BM, building on Satterthwaite (1946), suggest approximating the distribution of

$$\frac{\hat{\beta}_k - \beta_k}{\sqrt{\hat{\mathbb{V}}_{\mathrm{HC2},k}}},$$

by a t-distribution instead of a normal distribution. 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. Note that under homoskedasticity, $\mathbb{E}[\hat{\mathbb{V}}_{\mathrm{HC2}}] = \mathbb{V}$ and 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, and we choose K to match the second moment. Moreover, 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). Let λ_i be the weight for the i-th chi-squared random variable, so we can write

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

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

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

To characterize the weights, define, the $N \times N$ matrix $\mathbf{G}$, with i-th column equal to

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

Then the λ_i are the eigenvalues of the $N \times N$ matrix

$$\sigma^2 \cdot \mathbf{G}'\mathbf{G}.$$

Note that because of the form of (3.2), the value of K_{BM} does not depend on σ^2 even though the weights λ_i do depend on σ^2. Note also that the dof adjustment may be different for different elements of parameter β. Formally, the BM 95% confidence interval is:

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

The BM contribution over the earlier Sattherthwaite (1946) work is to base the dof calculation on the homoskedastic case with $\Omega = \sigma^2 \cdot I_N$. 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 $\hat{\Omega} = \text{diag}(\hat{\varepsilon}_i^2/(1-\mathbf{P}_{ii}))$. This often leads to substantially conservative confidence intervals.

There is a somewhat subtle difference between between the binary and the general regressor case. Applying the BM solution for the general case, given in (3.2), to data with a single binary regressor, leads to the same value as applying the BM solution for the binary case, given in (2.6). Similarly, applying the infeasible Sattherthwaite solution, based on the eigenvalues of $\mathbf{G}\Omega\mathbf{G}$, to binary regressor data, leads to the same dof as applying the infeasible Welch solution K^*_{welch}. However, applying the feasible Sattherthwaite solution to the case with a binary regressor does *not* lead to the feasible Welch solution. In the case with a single binary regressor, the Welch proposal for the dof calculation given in (2.5) is numerically identical to

$$K_{\text{welch}} = \left(\sum_{i=1}^{N} \lambda_{\text{welch},i}\right)^2 \Bigg/ \sum_{i=1}^{N} \lambda^2_{\text{welch},i},$$

where the weights $\lambda_{\text{welch},i}$ are the eigenvalues of

$$\mathbf{G}'\hat{\Omega}_{\text{welch}}\mathbf{G},$$

with $\hat{\Omega}_{\text{welch}}$ the diagonal matrix

$$\hat{\Omega}_{\text{welch},ij} = \begin{cases} \hat{\sigma}^2(0) & \text{if } X_i = 0, i = j \\ \hat{\sigma}^2(1) & \text{if } X_i = 1, i = j \\ 0 & \text{if } i \neq j. \end{cases}$$

The Welch dof solution is *not* equal to the dof based on the eigenvalues of $\mathbf{G}'\hat{\Omega}\mathbf{G}$, with $\hat{\Omega} = \text{diag}(\hat{\varepsilon}_i/(1-\mathbf{P}_{ii}))$, even though the estimated variances are the same:

$$(\mathbf{X}'\mathbf{X})^{-1}\left(\mathbf{X}'\hat{\Omega}_{\text{welch}}\mathbf{X}\right)(\mathbf{X}'\mathbf{X})^{-1} = (\mathbf{X}'\mathbf{X})^{-1}\left(\mathbf{X}'\hat{\Omega}\mathbf{X}\right)(\mathbf{X}'\mathbf{X})^{-1}.$$

3.4 A Small Simulation Study (Cragg, 1983)

We carry out a small simulation study based on designs by Cragg (1983). The model is

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

with $\beta_0 = 1$, $\beta_1 = 1$, and

$$\ln(X_i) \sim \mathcal{N}(0,1), \qquad \text{and} \;\; \varepsilon_i | X_i = x \sim \mathcal{N}\left(0, \gamma_0 + \gamma_1 \cdot x + \gamma_2 \cdot x^2\right).$$

Two designs were used: $(\gamma_0, \gamma_1, \gamma_2) = (0.6, 0.3, 0.0)$, and $(\gamma_0, \gamma_1, \gamma_2) = (0.3, 0.2, 0.1)$. The latter case exhibits considerable heteroskedasticity. The median value of $\sigma(x)$ is 0.77, the 0.025 quantile is 0.57, and the 0.975 quantile is 2.60. This particularly impacts the quality of the confidence intervals based on the feasible Sattherthwaite dof adjustment with $\hat{\Omega}$. For comparison we also include the standard errors and coverage rates based on the infeasible Sattherthwaite dof adjustment with the eigenvalue calculations based on $\mathbf{G}'\Omega\mathbf{G}$. We report results for two sample sizes, $N = 25$ and $N = 100$. For each design and each of the analytic variance estimators we use 1,000,000 replications, and 100,000 for the wild bootstrap with 10,000 bootstrap replications. The results are in presented in Table 3. For comparison, see Table III, panel 2, page 760 in Cragg (1983).

Qualitatively the results are similar to those for the Angrist-Pischke design. The robust variance estimators $\hat{\mathbb{V}}_{\text{ehw}}$ and the bias-adjusted version $\hat{\mathbb{V}}_{\text{HC2}}$ do not perform well unless the confidence intervals are based on t-distributions with the $K_{\text{Satterthwaite}}$ or K_{BM} 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 in the Presence of 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:

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

where $i = 1, \ldots, N$ indexes units. 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 $S_i \in \{1, \ldots, S\}$ denote the cluster unit i belongs to. We assume that

$$\mathbb{E}[\varepsilon|\mathbf{X}] = 0, \qquad \text{and} \;\; \mathbb{E}[\varepsilon\varepsilon'|\mathbf{X}] = \Omega,$$

where,

$$\Omega_{ij} = \begin{cases} \omega_{ij} & \text{if } S_i = S_j, \\ 0 & \text{otherwise.} \end{cases}$$

Let $\hat{\beta}$ be the least squares estimator, and let $\hat{\varepsilon}_i = Y_i - X_i'\hat{\beta}$ be the residual. Let $\underline{\hat{\varepsilon}}_s$ 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 $N_s \times N$ matrix $(I_N - \mathbf{P})_s$ to consist of the N_s rows of the $N \times N$ matrix $(I_N - \mathbf{P})$ corresponding to cluster s.

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

$$\hat{\mathbb{V}}_{\mathrm{lz}} = \left(\sum_{s=1}^{S} \mathbf{X}_s'\mathbf{X}_s\right)^{-1} \sum_{s=1}^{S} \mathbf{X}_s'\underline{\hat{\varepsilon}}_s\underline{\hat{\varepsilon}}_s'\mathbf{X}_s \left(\sum_{s=1}^{S} \mathbf{X}_s'\mathbf{X}_s\right)^{-1}.$$

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

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

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

$$\hat{\mathbb{V}}_{\mathrm{lz2}} = \left(\sum_{s=1}^{S} \mathbf{X}_s'\mathbf{X}_s\right)^{-1} \sum_{s=1}^{S} \mathbf{X}_s'(I_{N_s} - \mathbf{P}_{ss})^{-1/2}\underline{\hat{\varepsilon}}_s\underline{\hat{\varepsilon}}_s'((I_{N_s} - \mathbf{P}_{ss})^{-1/2})'\mathbf{X}_s \left(\sum_{s=1}^{S} \mathbf{X}_s'\mathbf{X}_s\right)^{-1},$$

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 formalize the degrees-of-freedom adjustment, define the $N \times S$ matrix $\mathbf{G}$ with s-th column equal to the N-vector $\mathbf{G}_s$ defined as

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

Then the dof adjustment is given by

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

where λ_i are the eigenvalues of $\mathbf{G}'\mathbf{G}$. The 95% confidence interval is now

$$95\%\ \mathrm{CI}^{\mathrm{cluster}}_{\mathrm{BM}} = \left(\hat{\beta}_k + t^{K_{\mathrm{BM}}}_{0.025} \times \sqrt{\hat{\mathbb{V}}_{\mathrm{lz2},k}}, \hat{\beta}_k + t^{K_{\mathrm{BM}}}_{0.975} \times \sqrt{\hat{\mathbb{V}}_{\mathrm{lz2},k}}\right). \tag{4.2}$$

We also consider a slightly different version of the dof adjustment. In principle we would like to use the eigenvalues of the matrix $\mathbf{G}'\Omega\mathbf{G}$, where $\Omega = \mathbb{E}[\varepsilon\varepsilon|\mathbf{X}]$. It is difficult to estimate ω accurately, which motivated BM to use $\sigma^2 \cdot I_N$ instead. In the clustering case, however, we have some structure on Ω that may be useful. We estimate a model for Ω where

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

We estimate σ^2_ν as the average of the product of the residuals for units with $S_i = S_j$, and $i \neq j$, and then estimate σ^2_ε as the average of the square of the residuals minus $\hat{\sigma}^2_\nu$. We then calculate the $\tilde{\lambda}_i$ as the eigenvalues of $\mathbf{G}'\hat{\Omega}\mathbf{G}$, and

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

4.1 A Small Simulation Study

We carry out a small simulation study following designs first used in Cameron, Gelbach, and Miller (2008). The model is the same as in (4.1), with a scalar covariate:

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

with $\beta_0 = \beta_1 = 0$. We consider five specific designs. In the first design, $X_i = V_{S_i} + W_i$ and $\varepsilon_i = \nu_{S_i} + \eta_i$. The V_s, W_i, ν_s, η_i are all normally distributed, with mean zero and unit variance. In the first design there are $S = 10$ clusters, with $N_s = 30$ units in each cluster. In the second design we have $S = 5$ clusters, again with $N_s = 30$ in each cluster. In the third design 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 fourth design we introduce heteroskedasticity, with $\eta_i|\mathbf{X} \sim N(0, 0.9X_i^2)$, and in the fifth design, the covariate is fixed within the clusters: $W_i = 0$ and $V_s \sim \mathcal{N}(0, 2)$.

For each design and each of the analytic variance estimators we use 1,000,000 replications, and 100,000 for the wild bootstrap, with 10,000 bootstrap replications.

5 Conclusion

Although there is a substantial literature documenting the poor properties of the conventional robust standard errors in small samples, in practice 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. We discuss the connection to the Behrens-Fisher problem, and suggest a minor modification for the case with clustering.

References

Angrist, J., and S. Pischke., (2009), *Mostly Harmless Econometrics*, Princeton University Press, Princeton, NJ.

Behrens, W., (1929). "Ein Beitrag zur Fehlerberechnung weniigen Beobachtungen," *Landswirth, Jahrbucher*, 68, 807-837.

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

Bester, C., T. Conley, and C. Hansen, (2009), "Inference with Dependent Data Using Clustering Covariance Matrix Estimators," Unpublished Manuscript, University of Chicago Business School.

Cameron, C., J. Gelbach, and D. Miller, (2008), "Bootstrap-based Improvements for Inference with Clustered Errors," *Review of Economics and Statistics* , Vol. 90(3): 414-427.

Chesher, A., and G. Austin, (1991). "The finite-sample distributions of heteroskedasticity robust Wald statistics," *Journal of Econometrics*, 47, 153-173.

Chesher, A., and I. Jewitt, (1987), "The Bias of a Heteroskedasticity Consistent Covariance Matrix Estimator," *Econometrica*, 55(5): 1217-1222.

Cragg, J., (1983), "More Efficient Estimation in the Presence of Heteroscedasticity of Unknown Form," *Econometrica* , Vol. 51(3): 751-763.

Diggle, P., P. Heagerty, K.-Y. Liang, and S. Zeger, (2002), *Analysis of Longitudinal Data*, Oxford University Press, Oxford.

Donald, S. and K. Lang, (2007), "Inference with Difference in Differences and Other Panel Data," *Review of Economics and Statistics*, Vol. 89(2): 221-233.

Eicker, F., (1967), "Limit Theorems for Regressions with Unequal and Dependent Errors," *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1, 59:82. Berkeley, University of California Press.

Fisher, R., (1939), "The Comparisons of Samples with Possibly Unequal Variances," *Annals of Eugenics*, 9, 174-180.

Hansen, C., (2009), "Generalized Least Squares Inference in Panel and Multilevel Models with Serial Correlation and Fixed Effects," *Journal of Econometrics*, 140(2), 670-694.

Hausman, J., and C. Palmer, (2011), "Heteroskedasticity-Robust Inference in Finite Samples," NBER Working Paper 17698, http://www.nber.org/papers/w17698.

Horn, S., and R. Horn, (1975), "Comparisons of Estimators of Heteroscedastic Variances in Linear Models," *Journal of the American Statistical Association*, 70(352), 872-879.

Horn, S., R. Horn, and D. Duncan, (1975), "Estimating Heteroscedastic Variances in Linear Models," *Journal of the American Statistical Association*, 70(350), 380-385.

Huber, P. (1967), "The Behavior of Maximum Likelihood Estimates Under Nonstandard Conditions," in *Proceedings of the Fifth Berkeley Symposium in Mathematical Statistics*, Vol. 1, Berkeley, University of California Press, 221-233.

Ibragimov, R., and U. Müller, (2009), "t-statistic based correlation and heterogeneity robust inference" Unpublished Manuscript, Department of Economics, Harvard University.

Lehmann, E., and J. Romano, (2005), *Testing Statistical Hypotheses*, Springer.

Liang, K., and S. Zeger, (1986), "Longitudinal Data Analysis Using Generalized Linear Models," *Biometrika*, 73(1): 13-22.

Liu, R., (1988), "Bootstrap Procedures under Some Non-iid Models," *Annals of Statistics*, 16, 1696-1708.

MacKinnon, J., (2002), "Bootstrap Inference in Econometrics," *Canadian Journal of Economics*, 35 615-645.

MacKinnon, J., (2012), "Thirty Years of Heteroskedasticity-Robust Inference," in *Recent Advances and Future Directions in Causality, Prediction, and Specification Analysis*, ed. Xaiohong Chen and Norman R. Swanson, New York, Springer, p 437–461.

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

Mammen, E., (1993). "Bootstrap and wild bootstrap for high dimensional linear models," *Annals of Statistics*, 21, 255-285.

Mantel, N., (1967), "The Detection of Disease Clustering and a Generalized Regression Approach," *Cancer Research*, 27(2):209-220.

Moulton, B., (1986), "Random Group Effects and the Precision of Regression Estimates," *Journal of Econometrics*, 32, 385-397.

Moulton, B., (1990), "An Illustration of a Pitfall in Estimating the Effects of Aggregate Variables on Micro Units," *Review of Economics and Statistics*, 334-338.

Moulton, B., and W. Randolph, (1989) "Alternative Tests of the Error Component Model," *Econometrica, Vol. 57, No. 3, 685-693.*

Pan, W., and M. Wall, (2002) "Small-sample Adjustments in Using the Sandwich Variance Estimator in Generalized Estimating Equation," *Statistics in Medicine, Vol. 51, 1429-1411.*

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

Scheffé, H., (1970), "Practical Solutions to the Behrens-Fisher Problem," *Journal of the American Statistical Association*, Vol. 65, No. 332, 1501-1508.

Stock, J., and M. Watson., (2008), "Heteroskedasticity-Robust Standard Errors for Fixed Effect Panel Data Regression," *Econometrica*, 76(1): 155-174.

Wang, Y. , (1971), "The Probabilities of the Type I Errors of the Welch Tests for the Behrens-Fisher Problem," *Journal of the American Statistical Association*, Vol. 66 , No. 335, 605-608.

Welch, B. (1951), "The Generalization of 'Students' Problem When Several Different Population Variances are Involved," *Biometrika*, Vol 34: 28-35.

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

William S. (1998), *Comparison of standard errors for robust, cluster, and standard estimators, StataCorp* http://www.stata.com/support/faqs/stat/cluster.html

Wooldridge, J., (2002), *Econometric Analysis of Cross Section and Panel Data*, MIT Press, Cambridge, MA.

Wu, C. (1986), "Jackknife, Bootstrap and Other Resampling Methods in Regression Analysis," (with discussion) *Annals of Statistics*, 14, 1261-1295.

Appendix A Other methods

A.1 HC3

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

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

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

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

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

A.2 The Wild Bootstrap

The simple bootstrap where we resample N units picked with replacement from the original sample is unlikely to perform well. In particular in cases where either N_0 or N_1 is small, the additional noise introduced by variation in the number of $X_i = 0$ units sampled is likely to adversely affect the properties of the corresponding confidence intervals. In this literature 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; MacKinnon, 2011).

Here we briefly describe the wild bootstrap in the regression setting, and then in the cluster setting. First consider the regression setting. Let $\hat{\beta}_0$ and $\hat{\beta}_1$ be the least squares estimates in the original sample, and $\hat{\varepsilon} = Y_i - \hat{\beta}_0 - X_i \cdot \hat{\beta}_1$ be the estimated residuals, and let $\hat{\mathbb{V}}$ be a variance estimator, either $\hat{\mathbb{V}}_{\mathrm{ehw}}$, or $\hat{\mathbb{V}}_{\mathrm{HC2}}$, or $\hat{\mathbb{V}}_{\mathrm{HC3}}$. In the wild bootstrap the regressor values are fixed in the resampling. In the b-th bootstrap sample, the value for the i-th outcome is

$$Y_{i,b} = \hat{\beta}_0 + X_i \cdot \hat{\beta}_1 + U_{i,b} \cdot \hat{\varepsilon}_i,$$

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, here we only consider this particular choice.)

Here we use a symmetric version of the bootstrap. In the b-th bootstrap sample we calculate the t-statistic

$$t_b = \frac{\hat{\beta}_{b,1} - \hat{\beta}_1}{\sqrt{\hat{\mathbb{V}}_b}},$$

for some variance estimator $\hat{\mathbb{V}}_b$. Over all the bootstrap samples we calculate the 0.95 quantile of the distribution of $|t_b|$ (which, because of the symmetry of the design, are approximately equal to minus the 0.025 and the 0.975 quantile of the distribution of t_b. Let this quantile be $q_{0.95}^{\text{wild}}$. We use this quantile instead of 1.96 to construct the confidence interval as

$$95\%\ \text{CI}_{\text{HC2}} = \left(\hat{\beta}_1 - q_{0.95}^{\text{wild}} \times \sqrt{\hat{\mathbb{V}}}, \hat{\beta}_1 + q_{0.95}^{\text{wild}} \times \sqrt{\hat{\mathbb{V}}}\right). \tag{A.2}$$

The wild bootstrap standard errors reported in the table are $\sqrt{\hat{\mathbb{V}}_{\text{HC2}}}(q_{0.95}^{\text{wild}}/1.96)$.

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 $\text{pr}(U_{s,b} = 1) = \text{pr}(U_{s,b} = -1) = 0.5$. The bootstrap value for the outcome for unit i in cluster s is then

$$Y_{is,b} = \hat{\beta}_0 + \hat{\beta}_1 \cdot X_{is} + U_{s,b} \cdot \varepsilon_{is},$$

with the values for X_{is} and ε_{is} remaining fixed across the bootstrap replications.

Table 1: Coverage Rates and Standard Errors, Angrist-Pischke Unbalanced Design, $N_0 = 27,\ N_1 = 3$

		Design I		Design II		Design III		Design IV		Design V	
		$\sigma(0) = 0.50$		$\sigma(0) = 0.85$		$\sigma(0) = 1.00$		$\sigma(0) = 1.18$		$.00\ \ \sigma(0) = 2$	
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.
$\hat{\mathbb{V}}_{\text{homo}}$	∞	0.73	0.33	0.90	0.52	0.94	0.60	0.97	0.70	1.00	1.17
	$N-2$	0.75	0.34	0.92	0.54	0.95	0.63	0.97	0.73	1.00	1.22
$\hat{\mathbb{V}}_{\text{ehw}}$	∞	0.77	0.40	0.79	0.42	0.81	0.44	0.82	0.45	0.87	0.55
	$N-2$	0.78	0.42	0.81	0.44	0.82	0.46	0.83	0.47	0.88	0.57
$\hat{\mathbb{V}}_{\text{HC2}}$	∞	0.82	0.49	0.84	0.51	0.85	0.52	0.86	0.53	0.90	0.62
	$N-2$	0.84	0.51	0.86	0.53	0.86	0.54	0.87	0.56	0.91	0.65
	K_{Welch}	0.93	1.00	0.92	0.93	0.92	0.90	0.93	0.87	0.93	0.80
	K^*_{Welch}	0.96	1.04	0.97	1.02	0.97	1.00	0.97	0.97	0.97	0.87
	K_{BM}	0.95	0.90	0.96	0.94	0.97	0.95	0.98	0.98	0.99	1.14
	wild	0.90	0.76	0.90	0.74	0.91	0.73	0.91	0.72	0.92	0.73
$\hat{\mathbb{V}}_{\text{HC3}}$	∞	0.87	0.60	0.89	0.61	0.89	0.62	0.90	0.63	0.92	0.71
	wild	0.91	0.78	0.91	0.77	0.92	0.77	0.92	0.76	0.93	0.77
$\max\{\hat{\mathbb{V}}_{\text{homo}}, \hat{\mathbb{V}}_{\text{ehw}}\}$	∞	0.82	0.41	0.92	0.54	0.95	0.62	0.97	0.71	1.00	1.17
$\max\{\hat{\mathbb{V}}_{\text{homo}}, \hat{\mathbb{V}}_{\text{HC2}}\}$	∞	0.86	0.49	0.93	0.57	0.95	0.64	0.97	0.73	1.00	1.17

Table 2: Coverage Rates and Standard Errors, Angrist-Pischke Balanced Design, $N_0 = 15$, $N_1 = 15$

		Design I $\sigma(0) = 0.50$		Design II $\sigma(0) = 0.85$		Design III $\sigma(0) = 1.00$		Design IV $\sigma(0) = 1.18$		Design V $\sigma(0) = 2.00$	
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.
$\hat{\mathbb{V}}_{\text{homo}}$	∞	0.94	0.28	0.94	0.33	0.94	0.36	0.94	0.39	0.94	0.57
	$N-2$	0.95	0.30	0.95	0.35	0.95	0.38	0.95	0.41	0.95	0.59
$\hat{\mathbb{V}}_{\text{ehw}}$	∞	0.93	0.27	0.93	0.32	0.93	0.35	0.93	0.38	0.93	0.55
	$N-2$	0.94	0.29	0.94	0.34	0.94	0.36	0.94	0.40	0.94	0.57
$\hat{\mathbb{V}}_{\text{HC2}}$	∞	0.94	0.28	0.94	0.33	0.94	0.36	0.94	0.39	0.94	0.57
	$N-2$	0.95	0.30	0.95	0.35	0.95	0.38	0.95	0.41	0.95	0.59
	K_{Welch}	0.95	0.30	0.95	0.35	0.95	0.38	0.95	0.41	0.95	0.60
	K^*_{Welch}	0.95	0.30	0.95	0.35	0.95	0.38	0.95	0.41	0.95	0.60
	K_{BM}	0.95	0.30	0.95	0.35	0.95	0.38	0.95	0.41	0.95	0.59
	wild	0.95	0.30	0.95	0.35	0.95	0.38	0.95	0.41	0.95	0.60
$\hat{\mathbb{V}}_{\text{HC3}}$	∞	0.94	0.29	0.95	0.35	0.95	0.37	0.95	0.41	0.94	0.59
	wild	0.95	0.30	0.95	0.35	0.95	0.38	0.95	0.41	0.95	0.60
$\max\{\hat{\mathbb{V}}_{\text{homo}}, \hat{\mathbb{V}}_{\text{ehw}}\}$	∞	0.94	0.28	0.94	0.33	0.94	0.36	0.94	0.39	0.94	0.57
$\max\{\hat{\mathbb{V}}_{\text{homo}}, \hat{\mathbb{V}}_{\text{HC2}}\}$	∞	0.94	0.28	0.94	0.33	0.94	0.36	0.94	0.39	0.94	0.57

Table 3: Coverage Rates and Standard Errors, Cragg Design

		$(\gamma_0, \gamma_1, \gamma_2) = (0.6, 0.3, 0.0)$				$(\gamma_0, \gamma_1, \gamma_2) = (0.3, 0.2, 0.1)$			
		Design I		Design II		Design III		Design IV	
		$N = 25$		$N = 100$		$N = 25$		$N = 100$	
variance estimator	dist/ dof	cov rate	med s.e.	cov rate	med. s.e.	cov rate	med. s.e.	cov rate	med. s.e.
$\hat{\mathbb{V}}_{\text{homo}}$	∞	0.81	0.12	0.76	0.05	0.63	0.12	0.51	0.06
	$N-2$	0.83	0.13	0.76	0.05	0.65	0.13	0.51	0.06
$\hat{\mathbb{V}}_{\text{ehw}}$	∞	0.74	0.11	0.83	0.07	0.67	0.14	0.78	0.11
	$N-2$	0.76	0.12	0.83	0.07	0.69	0.15	0.78	0.11
$\hat{\mathbb{V}}_{\text{HC2}}$	∞	0.82	0.14	0.87	0.07	0.77	0.17	0.84	0.12
	$N-2$	0.84	0.15	0.87	0.07	0.79	0.18	0.84	0.12
	$K_{\text{Satterthwaite}}$	0.96	0.30	0.96	0.12	0.96	0.47	0.97	0.23
	$K^*_{\text{Satterthwaite}}$	0.98	0.31	0.97	0.12	0.99	0.51	0.98	0.24
	K_{BM}	0.96	0.23	0.94	0.09	0.94	0.28	0.93	0.15
	wild	0.79	0.15	0.87	0.09	0.78	0.21	0.88	0.15
$\hat{\mathbb{V}}_{\text{HC3}}$	∞	0.89	0.18	0.90	0.08	0.87	0.23	0.90	0.13
	wild	0.81	0.17	0.88	0.09	0.81	0.24	0.89	0.16
$\max\{\hat{\mathbb{V}}_{\text{homo}}, \hat{\mathbb{V}}_{\text{ehw}}\}$	∞	0.83	0.13	0.85	0.07	0.71	0.15	0.78	0.11
$\max\{\hat{\mathbb{V}}_{\text{homo}}, \hat{\mathbb{V}}_{\text{HC2}}\}$	∞	0.87	0.15	0.88	0.08	0.79	0.18	0.84	0.12

Table 4: Coverage Rates and Standard Errors, Cameron-Gelbach-Miller Clustering Design

variance estimator	dist/ dof	Design I cov rate	Design I med s.e.	Design II cov rate	Design II med s.e.	Design III cov rate	Design III med s.e.	Design IV cov rate	Design IV med s.e.	Design V cov rate	Design V med s.e.
$\hat{\mathbb{V}}_{\text{homo}}$	∞	0.47	0.06	0.52	0.08	0.50	0.06	0.53	0.07	0.36	0.06
	$S-1$	0.53	0.07	0.69	0.11	0.56	0.07	0.59	0.08	0.41	0.07
$\hat{\mathbb{V}}_{\text{lz}}$	∞	0.79	0.12	0.73	0.13	0.84	0.13	0.84	0.14	0.81	0.18
	$S-1$	0.85	0.14	0.86	0.18	0.89	0.14	0.89	0.16	0.86	0.21
$\hat{\mathbb{V}}_{\text{STATA}}$	∞	0.81	0.13	0.78	0.15	0.87	0.13	0.86	0.15	0.83	0.19
	$S-1$	0.87	0.15	0.90	0.21	0.91	0.15	0.91	0.17	0.88	0.22
$\hat{\mathbb{V}}_{\text{lz2}}$	∞	0.87	0.15	0.84	0.17	0.89	0.14	0.89	0.16	0.88	0.22
	$S-1$	0.91	0.17	0.93	0.24	0.93	0.17	0.93	0.18	0.91	0.26
	K_{BM}	0.94	0.19	0.95	0.27	0.94	0.18	0.94	0.20	0.96	0.34
	$K^*_{\text{Satterthwaite}}$	0.98	0.25	0.98	0.34	0.97	0.21	0.96	0.23	0.96	0.34
	K_{IK}	0.97	0.24	0.97	0.32	0.97	0.20	0.96	0.22	0.96	0.34
	wild	0.91	0.19	0.91	0.29	0.93	0.18	0.93	0.20	0.89	0.27