Jackknife Standard Errors for Clustered Regression

Bruce E. Hansen*
University of Wisconsin†

This version: September 2025

Abstract

This paper presents a theoretical case for replacement of conventional heteroskedasticity-consistent and cluster-robust variance estimators with jackknife variance estimators, in the context of linear regression with heteroskedastic and/or cluster-dependent observations. We examine the bias of variance estimation and the coverage probabilities of confidence intervals. Concerning bias, we show that conventional variance estimators have full downward worst-case bias, while our jackknife variance estimator is never downward biased. Concerning confidence intervals, we show that intervals based on conventional standard errors have worst-case coverage equalling zero, while the jackknife-based confidence interval has coverage probability bounded by the Cauchy distribution, under the auxiliary assumption of normal errors. We also extend the Bell-McCaffrey (2002) student t approximation to our jackknife t-ratio, resulting in confidence intervals with improved coverage probabilities. Our theory holds under broad assumptions, allowing arbitrary cluster sizes, regressor leverage, within-cluster correlation, heteroskedasticity, regression with a single treated cluster, fixed effects, and delete-cluster invertibility failures. Our theoretical findings are consistent with the extensive simulation literature investigating heteroskedasticity-consistent and cluster-robust variance estimation.

*Research support from the NSF and the Phipps Chair are gratefully acknowledged. My thanks to the editor Xavier D'Haultfœuille and four referees for careful and detailed comments. Over the course of this research, I have received helpful comments and suggestions from many individuals, including Andrew Chesher, Harold Chiang, Grant Hillier, Rustam Ibragimov, James MacKinnon, Ulrich Müller, Morten Nielsen, Marc Paolella, Peter Phillips, Matthew Webb, and Thilo Welz. Thanks to J.C. Lazzaro for helpful research assistance.

†Department of Economics, 1180 Observatory Drive, University of Wisconsin, Madison WI 53706.

1 Introduction

Heteroskedasticity-consistent (HC) and cluster-robust (CRVE) variance estimators for regression coefficient estimators are foundational for applied economic analysis. They are used for measuring estimation precision, confidence interval construction, and tests of hypotheses. Unfortunately, under standard conditions, conventional HC and CRVE variance estimators can be fully downward biased, conventional HC and CRVE t-tests can exhibit unbounded size distortions, and conventional HC and CRVE confidence intervals can have coverage rates equal to zero, even when the regression errors are normally distributed. This situation – full downward bias, unbounded size, and zero coverage probability – can be corrected by replacing conventional variance estimators by an appropriate jackknife estimator. The latter is simple to calculate. This simple change – the use of jackknife instead of conventional standard errors – has the result that variance estimation is never downward biased and size distortion is bounded.

Our case for jackknife standard errors is based on two new theoretical insights concerning the linear regression model. First, our jackknife variance estimator is conservative (its expectation is larger than the exact variance), while the standard CRVE variance estimator has arbitrarily large downward worst-case estimation bias. The conservative property of the jackknife estimator holds under broad assumptions, allowing arbitrary cluster sizes, regressor leverage, within-cluster correlation, heteroskedasticity, and delete-cluster invertibility failures. Concerning the latter – delete-cluster invertibility failures – this allows for the context where a regressor is non-zero only for a single cluster; for example, regressions with included cluster-level fixed effects, regressions with cluster-level treatment indicators when only one cluster is treated, and regressions with potentially sparse[1] dummy variables. In contrast, conventional variance estimators (including conventional jackknife variance estimators) can fail miserably in such contexts. A new discovery we highlight is that for jackknife variance estimation to be robust to invertibility failures, it is critical to carefully modify the jackknife formula so not to discard clusters. Existing methods (which make *ad hoc* modifications) fail to be robust.

Our second theoretical finding is that under the assumptions of cluster-dependent normality and a stronger rank condition, the worst-case coverage probability of the CRVE-based confidence interval equals zero, while the worst-case coverage probability of our jackknife-based confidence interval is uniformly bounded and controlled by the Cauchy distribution, where uniformity is across all regressor and covariance matrix configurations. Both of these theoretical results assume that the linear regression model is correctly specified, so that the error has a zero conditional mean.

[1] Binary variables which are non-zero for only a few observations. This often occurs when categorical variables are interacted.

Furthermore, to provide improved finite sample confidence intervals and p-values, we propose a "Satterthwaite" adjusted student t approximation to the finite sample distribution of the jackknife t-ratio. This adjustment is similar to those of Bell and McCaffrey (2002), Imbens and Kolesár (2016), Young (2016), and Pustejovsky and Tipton (2018), but is the first specifically designed for jackknife t-ratios. We present a detailed set of simulation results which demonstrate that our proposed confidence interval has excellent coverage rates in finite samples across a broad range of regression designs. A limitation of this adjustment (similar to the others in this literature) is that it is confined to inference on real-valued parameters and thus excludes joint tests. We recommend that conventional statistical software calculate default confidence intervals and p-values using this Satterthwaite adjustment, instead of the current practice of the student t distribution. The Satterthwaite adjustment is computationally simple, more conservative than student t inference, and more reliable in finite samples.

The family of heteroskedasticity-consistent variance estimators are often written with the labels HC_0, HC_1, HC_2, and HC_3. The HC_0 version was introduced by Eicker (1963), Huber (1967), and White (1980). The degrees of freedom correction known as HC_1 was suggested by Hinkley (1977), and became ubiquitous in applied econometric practice by its designation as the default "`r`" robust option in Stata. Together, HC_0 and HC_1 are known as Eicker-Huber-White (EHW) variance estimators. HC_2 and HC_3 were introduced by MacKinnon and White (1985) as unbiased estimators under homoskedasticity and the jackknife principle, respectively.

The cluster-robust variance estimator (CRVE) was introduced by Liang and Zeger (1986) and Arellano (1987), is available in Stata through its ubiquitous `cluster` standard error option, and currently dominates applied econometric practice. Bell and McCaffrey (2002) introduced two generalizations for clustered regression similar to HC_2 and HC_3. The review by MacKinnon, Nielsen, and Webb (2023a) use the labels CV_1, CV_2, and CV_3 to denote these three estimators. The recognition that finite-sample inference based on EHW confidence intervals can be severely distorted is a recurrent theme in econometrics. Some investigations include Chesher and Jewitt (1987), Chesher (1989), Chesher and Austin (1991), Long and Ervin (2000), and Young (2019). There has been a substantial recent literature proposing improved standard errors over the EHW class under independent sampling. This includes Bera, Suprayitno, and Premaratne (2002), Cattaneo, Jansson, and Newey (2018), and Kline, Saggio, and Sølvsten (2020). These new variance estimators have the advantage that they are (approximately) unbiased, but have as disadvantages that the variance estimators are computationally burdensome in large samples and are not necessarily positive semi-definite. These methods, while promising, have not been generalized to the clustered sampling setting and are not investigated in this paper.

Jackknife standard errors can be paired with conventional critical values (student t or normal) or with alternative methods. The latter include the distributional adjustments of Bell and

McCaffrey (2002), Imbens and Kolesár (2016), and Pustejovsky and Tipton (2018), bootstrap percentile-t methods (Cameron, Gelbach, and Miller (2008)), and conditional critical values (Pötscher and Preinerstorfer (2025)). Our recommendation is that default inference should be based on the Bell-McCaffrey-style distributional adjustment, as it is simple to calculate and has excellent performance. Another excellent option for inference is the cluster wild bootstrap, if paired with jackknife standard errors. The recommendation to use jackknife/HC_3 variance estimators is not new. Authors making this recommendation include Efron and Stein (1981), MacKinnon and White (1985), Andrews (1991), Chesher and Austin (1991), Long and Ervin (2000), and MacKinnon, Nielsen and Webb (2023abc).

Our analysis applies to all regression contexts. One where the inadequacy of conventional approximations has received particular attention is regression with a small number of clusters and/or a small number of treated clusters. This literature includes Conley and Taber (2011), Ibragimov and Müller (2016), Rokicki, Cohen, Fink, Salomon, and Landrum (2018), Ferman and Pinto (2019), Hagemann (2019, 2025), MacKinnon and Webb (2020), Canay, Santos, and Shaikh (2021), and Niccodemi and Wansbeek (2022). These applications may also benefit from jackknife standard errors and our adjusted degrees of freedom approximation, as their application will reduce size distortions.

The organization of the paper is as follows. Section 2 introduces the cluster-robust variance estimator. Section 3 introduces jackknife variance estimation. Section 4 presents results on variance estimation bias. Section 5 presents the Cauchy distribution bound. Section 6 proposes the Satterthwaite adjusted t distributional approximation. Section 7 discusses computational issues. Section 8 presents simulation evidence. Section 9 presents an empirical application, extending Meng, Qian and Yared (2015) and Canay, Santos, and Shaikh (2021). A conclusion is presented in Section 10. Some technical proofs are presented in the Appendix. An Online Appendix presents additional technical proofs, additional asymptotic results, and computation details.

The R code which generates the numerical calculations presented in the paper, as well as Stata and R code to implement the proposed variance estimator and Satterthwaite distributional correction, is posted on the author's webpage `users.ssc.wisc.edu/~bhansen/`. A companion paper, Hansen (2025), presents several additional applications, to further illustrate the empirical relevance of the proposed methods.

A word on notation. We say that the symmetric square matrix $\boldsymbol{A}$ is positive-definite if $x'\boldsymbol{A}x > 0$ for all conformable vectors $x \neq 0$. Similarly, $\boldsymbol{A}$ is positive semi-definite if $x'\boldsymbol{A}x \geq 0$ for all $x \neq 0$. We write $\boldsymbol{A} > \boldsymbol{B}$ if $\boldsymbol{A} - \boldsymbol{B}$ is positive-definite, and write $\boldsymbol{A} \geq \boldsymbol{B}$ if $\boldsymbol{A} - \boldsymbol{B}$ is positive semi-definite.

2 Clustered Regression and the CRVE

The model is a standard clustered sampling regression. The observations are separated into G unbalanced mutually independent clusters. We write the observations on the ith individual in the gth cluster as (Y_{ig}, X_{ig}), for $i = 1, ..., n_g$ and $g = 1 = 1, ..., G$, where Y_{ig} is scalar and X_{ig} is $k \times 1$. We stack the observations by cluster, so that $\boldsymbol{Y}_g = (Y_{1g}, ..., Y_{n_g g})'$ is $n_g \times 1$ and $\boldsymbol{X}_g = (X_{1g}, ..., X_{n_g g})'$ is $n_g \times k$. The number of observations in the gth cluster is n_g and the total number of observations is $n = \sum_{g=1}^{G} n_g$. Stacking the observations conventionally, we obtain the full sample $(\boldsymbol{Y}, \boldsymbol{X})$.

The observations are assumed to satisfy the linear regression model $Y_{ig} = X_{ig}'\beta + e_{ig}$ where β is a $k \times 1$ coefficient vector and e_{ig} is an error. Written at the level of the cluster, the model is

$$\boldsymbol{Y}_g = \boldsymbol{X}_g\beta + \boldsymbol{e}_g \tag{1}$$

$$\mathbb{E}\left[\boldsymbol{e}_g\right] = 0 \tag{2}$$

where $\boldsymbol{e}_g = (e_{1g}, ..., e_{n_g g})'$. It is also sometimes convenient to use the full-sample notation $\boldsymbol{Y} = \boldsymbol{X}\beta + \boldsymbol{e}$. We will treat the regressors as fixed, but all results go through in the random regressor setting by conditioning. Notice that model (1)-(2) specifies the conditional expectation of $\boldsymbol{Y}_g$ as linear in $\boldsymbol{X}_g$, and hence excludes the linear projection model. Define the cluster-level covariance matrices

$$\mathbb{E}\left[\boldsymbol{e}_g\boldsymbol{e}_g'\right] = \boldsymbol{\Sigma}_g. \tag{3}$$

The specification (3) allows the covariance matrices $\boldsymbol{\Sigma}_g$ to be a function of the regressors (and hence conditionally heteroskedastic), and/or to be a function of the cluster g (and hence unconditionally heteroskedastic). We follow the clustering literature and impose no structure on $\boldsymbol{\Sigma}_g$. We also define the full-sample covariance matrix $\boldsymbol{\Sigma} = \text{diag}\{\boldsymbol{\Sigma}_1, ..., \boldsymbol{\Sigma}_G\}$.

The model (1)-(3) includes heteroskedastic regression as the special case where $n_g = 1$ for all g. We call this the "no clustering" or "absence of clustering" case.

Assumption 1 *Model (1)-(3) holds, $\boldsymbol{X}$ is full rank, and $\mathbb{E}[e_{ig}^2] < \infty$ for all i and g.*

We focus on the least squares estimator $\widehat{\beta} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{Y}\right)$. It is well known that under Assumption 1, $\widehat{\beta}$ is unbiased for β with exact covariance matrix

$$\boldsymbol{V} = \text{var}\left[\widehat{\beta}\right] = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\sum_{g=1}^{G}\boldsymbol{X}_g'\boldsymbol{\Sigma}_g\boldsymbol{X}_g\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}.$$

The nearly ubiquitous cluster-robust variance estimator (CRVE_1) for $\boldsymbol{V}$ is

$$\widehat{\boldsymbol{V}}_1 = \frac{G(n-1)}{(G-1)(n-k)} \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \left(\sum_{g=1}^{G} \boldsymbol{X}_g' \widehat{\boldsymbol{e}}_g \widehat{\boldsymbol{e}}_g' \boldsymbol{X}_g \right) \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}, \tag{4}$$

where $\widehat{\boldsymbol{e}}_g = \boldsymbol{Y}_g - \boldsymbol{X}_g \widehat{\beta}$ denotes the least squares residual vector for the gth cluster. The estimator (4) with no degrees of freedom adjustment was introduced by Liang and Zeger (1986) and Arellano (1987). The degrees of freedom correction appearing in (4) is added by the Stata "`cluster`" option. In the absence of clustering, (4) specializes to the HC_1 "heteroskedasticity-robust" or Eicker-Huber-White (EHW) estimator of Eicker (1963), Huber (1967), and White (1980), multiplied by the $n/(n-k)$ degrees of freedom correction suggested by Hinkley (1977).

As an alternative to CRVE_1, Bell and McCaffrey (2002), Imbens and Kolesár (2016), and Kolesár (2023) recommended the CRVE_2 variance estimator

$$\widehat{\boldsymbol{V}}_2 = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \left(\sum_{g=1}^{G} \boldsymbol{X}_g' \boldsymbol{M}_g^{+1/2} \widehat{\boldsymbol{e}}_g \widehat{\boldsymbol{e}}_g' \boldsymbol{M}_g^{+1/2} \boldsymbol{X}_g \right) \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \tag{5}$$

where $\boldsymbol{M}_g^{+1/2}$ is the Moore-Penrose inverse of the square root of the partial projection matrix

$$\boldsymbol{M}_g = \boldsymbol{I}_{n_g} - \boldsymbol{X}_g \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{X}_g'. \tag{6}$$

In the absence of clustering and when all $\boldsymbol{M}_g$ are invertible, CRVE_2 specializes to the HC_2 estimator of MacKinnon and White (1985). The original definitions of HC_2 and CRVE_2 required that all $\boldsymbol{M}_g$ are invertible. As discussed by Kolesár (2023), the generalized inverse in (5) allows CRVE_2 to be defined even when $\boldsymbol{M}_g$ is non-invertible, and this is the implementation in Stata 18 through its `vce(hc2` *clustvar*`)` option.

3 Jackknife Variance Estimation

The jackknife estimator of variance of Tukey (1958) extended to clustered dependence is

$$\widehat{\boldsymbol{V}}_3 = \frac{G-1}{G} \sum_{g=1}^{G} \left(\widehat{\beta}_{-g} - \bar{\beta}\right) \left(\widehat{\beta}_{-g} - \bar{\beta}\right)' \tag{7}$$

where

$$\widehat{\beta}_{-g} = \left(\sum_{j \neq g} \boldsymbol{X}_j' \boldsymbol{X}_j \right)^{-1} \left(\sum_{j \neq g} \boldsymbol{X}_j' \boldsymbol{Y}_j \right) = \left(\boldsymbol{X}'\boldsymbol{X} - \boldsymbol{X}_g'\boldsymbol{X}_g \right)^{-1} \left(\boldsymbol{X}'\boldsymbol{Y} - \boldsymbol{X}_g'\boldsymbol{Y}_g \right) \tag{8}$$

and $\bar{\beta} = \frac{1}{G}\sum_{g=1}^{G}\widehat{\beta}_{-g}$. The delete-one-cluster estimator $\widehat{\beta}_{-g}$ in (8) is obtained by applying least squares to the sample after deleting the observations in cluster g. A variant of $\widehat{\boldsymbol{V}}_3$ is

$$\widehat{\boldsymbol{V}}_4 = \frac{G-1}{G}\sum_{g=1}^{G}\left(\widehat{\beta}_{-g}-\widehat{\beta}\right)\left(\widehat{\beta}_{-g}-\widehat{\beta}\right)', \tag{9}$$

which centers at the full-sample estimator $\widehat{\beta}$ rather than at $\bar{\beta}$. In Stata, $\widehat{\boldsymbol{V}}_3$ and $\widehat{\boldsymbol{V}}_4$ can be calculated by the `vce(jackknife)` and `vce(jackknife,mse)` options.

The estimators (7) and (9) as written are not uniquely defined if there is a cluster g for which $\boldsymbol{X}'\boldsymbol{X}-\boldsymbol{X}_g'\boldsymbol{X}_g$ is noninvertible[2]. The implementation in Stata 18 circumvents this difficulty by excluding from the sums in (7)-(9) any cluster where (8) is undefined[3]. We follow this interpretation; henceforth, we assume that (7)-(9) are implemented with this modification. We describe (7)-(9) as the "conventional" jackknife variance estimators.

As we show in the next section, the conventional estimators $\widehat{\boldsymbol{V}}_3$ and $\widehat{\boldsymbol{V}}_4$ can exhibit downward bias. To eliminate this possibility, we propose the following jackknife estimator:

$$\widehat{\boldsymbol{V}}_5 = \sum_{g=1}^{G}\left(\widehat{\beta}_{-g}-\widehat{\beta}\right)\left(\widehat{\beta}_{-g}-\widehat{\beta}\right)' \tag{10}$$

where

$$\widehat{\beta}_{-g} = \left(\boldsymbol{X}'\boldsymbol{X}-\boldsymbol{X}_g'\boldsymbol{X}_g\right)^{-}\left(\boldsymbol{X}'\boldsymbol{Y}-\boldsymbol{X}_g'\boldsymbol{Y}_g\right) \tag{11}$$

and $\boldsymbol{A}^{-}$ denotes a generalized inverse of $\boldsymbol{A}$. $\widehat{\beta}_{-g}$ is a generalized delete-one-cluster estimator. It is defined for all g and therefore (10) includes all clusters. Most of our results will hold for any choice of generalized inverse; in practice, we recommend the Moore-Penrose inverse. With this choice, $\widehat{\beta}_{-g}$ is the unique minimum-length minimizer of the delete-one-cluster least squares criterion. The Moore-Penrose inverse is computationally stable and is available in all statistical packages. Our estimator $\widehat{\boldsymbol{V}}_5$ differs from $\widehat{\boldsymbol{V}}_3$ and $\widehat{\boldsymbol{V}}_4$ in three respects, all of which contribute to the inequality $\widehat{\boldsymbol{V}}_5 \succ \widehat{\boldsymbol{V}}_4 \succ \widehat{\boldsymbol{V}}_3$. First, $\widehat{\boldsymbol{V}}_5$ does not drop noninvertible clusters. Second, $\widehat{\boldsymbol{V}}_5$ is centered at the full-sample estimator $\widehat{\beta}$ rather than at $\bar{\beta}$. Third, $\widehat{\boldsymbol{V}}_5$ does not have the degrees of freedom correction $(G-1)/G$. In some applications these differences will be negligible, but in others, as we show later, they can be substantial.

Jackknife estimation is ideally suited for the context where the delete-one-cluster estimators $\widehat{\beta}_{-g}$ are well-defined for all clusters, which requires that the matrices $\boldsymbol{X}'\boldsymbol{X}-\boldsymbol{X}_g'\boldsymbol{X}_g$ are invertible for all g. We call this context *clusterwise invertibility* and its failure *clusterwise noninvertibility*. Noninvertibility occurs when deletion of a single cluster renders the regressor design

[2] This is identical to the context where $\boldsymbol{M}_g$ in (6) is noninvertible.

[3] This follows the recommendation in Shao and Tu (1995) for noninvertible bootstrap replications.

matrix singular. Most typically, this occurs when a regressor (or some linear combination of regressors) only takes non-zero values for a single cluster. Examples include regressions with included cluster-level fixed effects, regressions with cluster-level treatment indicators when only one cluster is treated, and saturated regressions with sparse cell proportions. Such regressions are commonplace in applications, so it is desirable for variance estimation to be sufficiently flexible to handle their occurrence, which is presumably the motivation for the "drop noninvertible clusters" modification described above. However, the properties of modifications need to be investigated to preclude undesirable outcomes. When the sample satisfies clusterwise invertibility then the estimators $\widehat{\boldsymbol{V}}_4$ and $\widehat{\boldsymbol{V}}_5$ only differ by the degrees of freedom correction $(G-1)/G$, which is inconsequential when G is large. Under clusterwise noninvertibility, however, they can differ substantially.

The clustered jackknife estimators $\widehat{\boldsymbol{V}}_3$ and $\widehat{\boldsymbol{V}}_4$ were developed by Cochran (1977), Rust and Rao (1996), and Bell and McCaffrey (2002). See MacKinnon, Nielsen, and Webb (2023abc). Stata has codified the modification to delete noninvertible clusters. MacKinnon, Nielsen, and Webb (2023c) propose a generalized delete-one-cluster estimator similar[4] to (11), though they do not investigate its statistical properties.

For alternative algebraic representations see MacKinnon, Nielsen, and Webb (2023b). One we use is based on the delete-one-cluster prediction errors

$$\widehat{\boldsymbol{e}}_{-g} = \boldsymbol{Y}_g - \boldsymbol{X}_g \widehat{\beta}_{-g}. \tag{12}$$

In the proof of Theorem 1 we show that

$$\widehat{\beta}_{-g} - \widehat{\beta} = -\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{X}_g' \widehat{\boldsymbol{e}}_{-g}. \tag{13}$$

This equality holds even under clusterwise noninvertibility. Given (13), we can write (10) as

$$\widehat{\boldsymbol{V}}_5 = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \left(\sum_{g=1}^{G} \boldsymbol{X}_g' \widehat{\boldsymbol{e}}_{-g} \widehat{\boldsymbol{e}}_{-g}' \boldsymbol{X}_g \right) \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}. \tag{14}$$

The jackknife estimators (7) and (10) simplify in the absence of clustering, though existing proposals and analysis assume that $\boldsymbol{M}_g$ are all invertible[5]. Under these conditions the estimator (7) corresponds to that proposed by MacKinnon and White (1985) and the estimator (10) corresponds to that proposed by Andrews (1991) and Davidson and MacKinnon (1993). The latter is known as HC_3 and can be calculated in Stata by the `vce(hc3)` option. When clusterwise invert-

[4] The differences are that they multiply (10) by $(G-1)/G$, and implement (11) via the QR algorithm, mimicking the Stata implementation.

[5] In the absence of clustering, this means that the leverage values $h_{ii} = X_i' \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} X_i$ all satisfy $h_{ii} < 1$.

ibility fails, the estimators HC_2 and HC_3 are undefined[6]. For more details, see the monographs of Efron (1982) and Shao and Tu (1995).

While this paper is primarily focused on finite sample theory, it is instructive to observe that under the conditions for asymptotic normality of tests construted with the $CRVE_1$ variance estimator (as in Theorem 9 of Hansen and Lee (2019)), tests constructed with the jackknife variance estimator (14) are also asymptotically normal. We present this theory in the Online Appendix.

4 Biased vs Conservative Variance Estimation

In the classical regression model, the classical variance estimator is unbiased for the exact finite sample variance. Not surprisingly, outside the classical model, robust variances estimators are not unbiased. We now examine the worst-case downward bias of the five variance estimators described in the previous sections. We focus on downward bias, as this is the issue which causes undercoverage of confidence intervals and oversized tests. The main contribution of this section is the following result.

Theorem 1 *Under Assumption 1,* $\mathbb{E}\left[\widehat{\boldsymbol{V}}_5\right] \succeq \boldsymbol{V}$.

Theorem 1 shows that our recommended jackknife estimator $\widehat{\boldsymbol{V}}_5$ is *never downward biased*, or *conservative*. This means that in any regression context, and any sample size, we can be confident that the jackknife estimator is not downward biased. We will find that this bias property is important as it is directly connected to the coverage probabilities of confidence intervals.

Theorem 1 holds quite broadly, holding for all sample sizes, regressor matrices, variance matrices, and violations of clusterwise noninvertibility. In particular, the robustness to clusterwise noninvertibility is new and surprising. Theorem 1 augments Theorem 2 of Bell and McCaffrey (2002), which established that $\widehat{\boldsymbol{V}}_4$ is never downward biased when the regressors satisfy clusterwise invertibility and the errors e_{ig} are i.i.d. (that is, when $\boldsymbol{\Sigma} = \boldsymbol{I}_n\sigma^2$).

Theorem 1 is also related to the seminal work of Efron and Stein (1981). Their results are typically described as stating that the same result holds for $\widehat{\boldsymbol{V}}_3$ but this is incorrect. Instead, Efron and Stein's Theorem 2 states that $\widehat{\boldsymbol{V}}_3$ is never-downward-biased as an estimator of $\text{var}\left[\bar{\beta}\right]$, not as an estimator of $\text{var}\left[\widehat{\beta}\right]$. Furthermore, Theorem 2 below shows $\widehat{\boldsymbol{V}}_3$ does not satisfy the never-downward-biased property.

We now explore the worst-case bias properties of the other CRVE variance estimators. For these results we focus on individual coefficient estimates and their variance estimators. For some non-zero $k \times 1$ vector R, define the scalar parameter $\theta = R'\beta$. This includes individual

[6]A word of caution: Stata 18 reports HC_2 and HC_3 standard errors even when the regressor matrix is clusterwise noninvertible by giving these observations a weight of zero in the variance calculation. See Kranz (2024).

coefficients and linear combinations. Its estimator is $\widehat{\theta} = R'\widehat{\beta}$. Under Assumption 1, $\widehat{\theta}$ is unbiased for θ and has exact variance $v^2 = \text{var}\left[\widehat{\theta}\right] = R'\boldsymbol{V}R$. The estimators of v^2 are $\widehat{v}_j^2 = R'\widehat{\boldsymbol{V}}_j R$ for $j = 1,...,5$. Standard errors for $\widehat{\theta}$ are their square roots $\widehat{v}_j = \sqrt{R'\widehat{\boldsymbol{V}}_j R}$. To exclude degeneracy, some of our results require that the variance is strictly positive. For technical reasons, we also restrict the extent of clusterwise noninvertibility.

Assumption 2 $v^2 > 0$.

Assumption 3 *Partition $\boldsymbol{X} = [\boldsymbol{X}_1, \boldsymbol{X}_2]$ and $\boldsymbol{X}_g = \left[\boldsymbol{X}_{1g}, \boldsymbol{X}_{2g}\right]$. Suppose that R loads only on $\boldsymbol{X}_1$. Let $\dot{\boldsymbol{X}}_1$ (with gth component $\dot{\boldsymbol{X}}_{1g}$) be the residual from the linear regression of $\boldsymbol{X}_1$ onto $\boldsymbol{X}_2$. For each $g = 1,...,G$, either (1) $\boldsymbol{X}'\boldsymbol{X} - \boldsymbol{X}_g'\boldsymbol{X}_g$ is invertible, or (2) $\dot{\boldsymbol{X}}_{1g}'\boldsymbol{X}_{2g} = 0$ and $\dot{\boldsymbol{X}}_1'\dot{\boldsymbol{X}}_1 - \dot{\boldsymbol{X}}_{1g}'\dot{\boldsymbol{X}}_{1g}$ is invertible.*

Assumption 3 is essentially identical to the condition of Lemma 1 of Kolesár (2023). It requires that inference only concerns the "well-identified" variables $\boldsymbol{X}_1$. Essentially, it states that after the controls are partialled out, the leave-cluster-out coefficients for $\boldsymbol{X}_1$ are uniquely defined. Assumption 3 holds when $\boldsymbol{X}_2$ are cluster-specific fixed effects, but excludes inference on the fixed effects.

It will be convenient to define sets of models. For fixed k and G let $\mathscr{F}$ be the class of all $(\boldsymbol{X}, \boldsymbol{\Sigma})$ satisfying Assumptions 1-3. Let $\mathscr{F}^* \subset \mathscr{F}$ be the subset where $\boldsymbol{X}$ satisfies clusterwise invertibility. The following shows that the variance estimators other than $\widehat{\boldsymbol{V}}_5$ can be severely biased.

Theorem 2 *Under Assumptions 1-3,*

$$\inf_{(\boldsymbol{X},\boldsymbol{\Sigma})\in\mathscr{F}^*} \frac{\mathbb{E}\left[\widehat{v}_1^2\right]}{v^2} = 0, \tag{15}$$

$$\inf_{(\boldsymbol{X},\boldsymbol{\Sigma})\in\mathscr{F}^*} \frac{\mathbb{E}\left[\widehat{v}_2^2\right]}{v^2} = 0, \tag{16}$$

$$\inf_{(\boldsymbol{X},\boldsymbol{\Sigma})\in\mathscr{F}^*} \frac{\mathbb{E}\left[\widehat{v}_3^2\right]}{v^2} \leq \left(\frac{G-1}{G}\right)^2 < 1, \tag{17}$$

$$\inf_{(\boldsymbol{X},\boldsymbol{\Sigma})\in\mathscr{F}^*} \frac{\mathbb{E}\left[\widehat{v}_4^2\right]}{v^2} = \frac{G-1}{G} < 1, \tag{18}$$

$$\inf_{(\boldsymbol{X},\boldsymbol{\Sigma})\in\mathscr{F}} \frac{\mathbb{E}\left[\widehat{v}_3^2\right]}{v^2} = 0, \tag{19}$$

$$\inf_{(\boldsymbol{X},\boldsymbol{\Sigma})\in\mathscr{F}} \frac{\mathbb{E}\left[\widehat{v}_4^2\right]}{v^2} = 0, \tag{20}$$

$$\inf_{(\boldsymbol{X},\boldsymbol{\Sigma})\in\mathscr{F}} \frac{\mathbb{E}\left[\widehat{v}_5^2\right]}{v^2} = 1. \tag{21}$$

Equation (15) shows that the expected value of the scaled CRVE_1 estimator can be arbitrarily close to zero. This means that there is some distribution under which variance estimation has arbitrarily large downward bias. In words, we say that CRVE_1 is *fully downwardly biased*. The proof of (15) focuses on models which satisfy $\boldsymbol{\Sigma} = \boldsymbol{I}_n$ and clusterwise invertibility, so the result (15) has nothing to do with heteroskedasticity, correlated errors, or invertibility failure. Rather, it is due to extreme regressor leverage (unbalanced regressors and/or cluster sizes). Equation (16) shows a similar result for CRVE_2. The difference between (15) and (16) is that the proof of the latter requires manipulation of the covariance matrices. Thus, CRVE_1 can be fully downward biased due to regressor leverage alone, while the result for CRVE_2 requires non-i.i.d. errors.

Equations (17) and (18) show that the conventional jackknife estimators $\widehat{v}_3^2$ and $\widehat{v}_4^2$ violate the never-downward-biased property, even under clusterwise invertibility. The magnitude of the violation is small when G is large. Regardless, (17) and (18) show that the conventional interpretation[7] of the results of Efron and Stein (1981) is incorrect, and that the never-downward-biased result of Bell and McCaffrey (2002) is not robust to non-i.i.d. errors. Equations (19) and (20) extend the analysis of the conventional jackknife estimators to the class of models allowing clusterwise noninvertibility but satisfying Assumption 3. The results show that under these conditions the conventional estimators $\widehat{v}_3^2$ and $\widehat{v}_4^2$ can be fully downward biased. The proof of these results focus on models which satisfy $\boldsymbol{\Sigma} = \boldsymbol{I}_n$. This means that the full downward bias of the conventional jackknife estimators is entirely a consequence of the deletion of noninvertible clusters, and does not require heteroskedasticity or correlated errors.

Equation (21) examines our recommended jackknife estimator $\widehat{v}_5^2$, allowing for general heteroskedasticity, correlated errors, and clusterwise noninvertibility. As implied by Theorem 1, $\widehat{v}_5^2$ is never downward biased. Equation (21) extends this further and demonstrates that the infimum equals one, meaning that the inequality of Theorem 1 is sharp[8].

Theorem 2 draws a stark contrast between the five variance estimators. The full downward bias of CRVE_1, CRVE_2, and the conventional jackknife estimators means that on average they can be "much too small" relative to the true variance. In contrast, the never-downward-biased property of $\widehat{v}_5^2$ means that there is no situation where it is expected to be "too small", even slightly. The model classes studied in Theorem 2 hold fixed the number of regressors k and number of clusters G, but allow the cluster sizes n_g, regressors $\boldsymbol{X}$, and covariance matrices $\boldsymbol{\Sigma}$ to vary freely. It is important to understand that the statements of Theorem 2 hold for all G. Thus

[7] This distinction was recognized by Efron and Stein (1981). For a further discussion see Section 4.5 of Efron (1982).

[8] While (21) focuses on real-valued parameters, the proof of (21) extends to the full covariance matrix estimator, meaning that $\mathbb{E}\left[\widehat{\boldsymbol{V}}_5\right]$ can be arbitrarily close to $\boldsymbol{V}$.

the bias of CRVE_1, CRVE_2, and the conventional jackknife can be arbitrarily large in both very small and very large samples.

The worst-case downward bias in (15)-(20) is calculated by studying models with extreme leverage, arising when the regressor of interest has variation which is dominated by a single cluster. Intuitively, when a small number of clusters dominate the sample, standard variance estimators are highly biased towards zero. The least squares estimator overfits the dominating clusters, shrinking the residuals for these clusters relative to the true errors. This leads to downward estimation bias. Conventional fixes, such as the degrees of freedom adjustment of CRVE_1, are insufficient to counter the bias.

One suggestion has been for users to examine their model for features which lead to downward bias, such as high leverage points and heterogeneous cluster sizes. Conditional on this information, appropriate standard errors could be selected. We do not recommend this approach. This idea is analogous to selecting EHW standard errors instead of classical standard errors after a test for heteroskedasticity. In general, pre-test estimators have poor finite sample properties due to pre-test bias. It is generally preferred to use robust methods without pre-test selection. Therefore, our recommendation is for the jackknife to be the default method for variance estimation and standard error calculation. Another potential objection to Theorem 2 is that in any given application, $\boldsymbol{X}$ is fixed, so what may be considered relevant is the worst-case bias over $\boldsymbol{\Sigma}$, holding $\boldsymbol{X}$ fixed. However, it is unclear how to calculate the latter bound, and we leave this investigation to future research.

Theorem 2 shows that the jackknife estimator is never downward biased. It is generically upward biased, which we now characterize. From the proof of Theorem 2 we can calculate

$$\mathbb{E}\left[\widehat{\boldsymbol{V}}_5\right] = \boldsymbol{V} + \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\sum_{g=1}^{G} \boldsymbol{X}_g'\boldsymbol{X}_g\boldsymbol{V}_{-g}\boldsymbol{X}_g'\boldsymbol{X}_g\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \tag{22}$$

where $\boldsymbol{V}_{-g} = \text{var}\left[\widehat{\beta}_{-g}\right]$ is the variance of the delete-one-cluster estimator. In well-behaved contexts, $\boldsymbol{V}_{-g} \approx \boldsymbol{V}$ and the bias component on the right side of (22) will be $O\left(\frac{1}{n}\left\|\boldsymbol{V}\right\|\right)$. Thus, in general, the proportionate upward bias will decrease with sample size n. In general, the above expression shows that the upward bias is larger when $\boldsymbol{V}_{-g}$ is much larger than $\boldsymbol{V}$ and $\mathbb{E}\left\|\boldsymbol{X}_g\right\|^4$ is large. These are both related to high leverage. Thus we expect $\widehat{\boldsymbol{V}}_5$ to have largest upward bias under high regressor leverage, which is exactly the context where conventional CRVE estimators are severely downward biased. Equation (22) leaves open the possibility of bias-correction for $\widehat{\boldsymbol{V}}_5$. We do not pursue this idea in this paper.

5 Cauchy Bound

Given a standard error $\widehat{v}_j$ and a critical value c, a confidence interval for θ is

$$\widehat{C}_j(c) = \widehat{\theta} \pm c\widehat{v}_j. \tag{23}$$

The conventional[9] critical value for a nominal $100(1-\alpha)\%$ interval is $c = t_{G-1}^{1-\alpha/2}$, the $1-\alpha/2$ quantile of the student t distribution with $G-1$ degrees of freedom[10]. In finite samples, the actual coverage rate of the confidence interval (23) will deviate from its nominal level. We now present bounds on the exact coverage rate for intervals (23) constructed with different standard errors. As these are finite sample results, we assume that the errors are normally distributed.

Assumption 4 $\boldsymbol{e}_g$ *is distributed* $\mathrm{N}\left(0, \boldsymbol{\Sigma}_g\right)$.

Assumption 4 states that the error vectors $\boldsymbol{e}_g$ are normally distributed with an unrestricted within-cluster covariance structure. As described after (3), this allows the covariance matrices to vary both with the regressors as well as the cluster, and thus allows both unconditional and conditional heteroskedasticity. Normality is a strong assumption, and is not meant to be taken literally. Rather, by studying coverage rates under this assumption we can gain insight into their behavior in finite samples without relying on asymptotic approximations. Alternatively, Assumption 4 could be replaced with the assumption that the cluster sums $\boldsymbol{X}_g'\boldsymbol{e}_g$ are asymptotically normal as the cluster sizes n_g diverge, as in Canay, Santos, and Shaikh (2021).

Under i.i.d. normal errors, it is well known that confidence intervals with classical standard errors have exact coverage $1-\alpha$. What may be less well known (or at least less emphasized in standard curricula) is that this result does not extend to confidence intervals constructed with HC and CRVE standard errors. We now present the second main contribution of the paper, which provides a lower bound on the coverage rate of the jackknife confidence interval (when constructed with our recommended standard error $\widehat{v}_5$). Let ζ denote a random variable with the Cauchy distribution.

Theorem 3 *Under Assumptions 1-4, for any* $c \geq 1$,

$$\mathbb{P}\left[\theta \in \widehat{C}_5(c)\right] \geq \mathbb{P}\left[|\zeta| \leq c\right]. \tag{24}$$

Equation (24) shows that the interval $\widehat{C}_5(c)$ has coverage probability which is uniformly

[9] For example, that used in Stata.

[10] Or $n-k$ degrees of freedom in the absence of clustering.

bounded[11] away from zero. The lower bound is the Cauchy distribution. An important implication is that the finite sample coverage probability of the $\widehat{C}_5(c)$ jackknife confidence interval has bounded distortion from its nominal level. Other than normality, Assumptions 1-4 are broad, including regression models with extreme leverage, within-cluster correlation, heteroskedasticity, and clusterwise noninvertibility. However, Theorem 3 relies on the normality assumption. It is unclear how the result extends to other contexts, or if a similar bound could be attained in a $G \to \infty$ asymptotic framework.

We contrast (24) with the coverage of confidence intervals constructed with other standard errors. To do so, we first state an intermediate result which may be of independent interest.

Theorem 4 *Under Assumptions 1-4, if for some variance estimator $\widehat{v}^2$ and model class $\mathscr{F}_a \subset \mathscr{F}$,*

$$\inf_{(\boldsymbol{X},\boldsymbol{\Sigma})\in\mathscr{F}_a} \frac{\mathbb{E}\left[\widehat{v}^2\right]}{v^2} = 0, \tag{25}$$

then for $\widehat{C}(c) = \widehat{\theta} \pm c\widehat{v}$ and any $0 \leq c < \infty$,

$$\inf_{(\boldsymbol{X},\boldsymbol{\Sigma})\in\mathscr{F}_a} \mathbb{P}\left[\theta \in \widehat{C}(c)\right] = 0.$$

Theorem 4 shows that full downward bias of a variance estimator leads to an interval with zero coverage. This provides a simple method to demonstrate that an interval has zero coverage, or equivalently that the size of a hypothesis test equals one. Theorem 4 leads to the following characterization of confidence intervals constructed with $\widehat{v}_1$, $\widehat{v}_2$, $\widehat{v}_3$, and $\widehat{v}_4$.

Theorem 5 *Under Assumptions 1-4, for $\widehat{C}_j(c) = \widehat{\theta} \pm c\widehat{v}_j$ and any $0 \leq c < \infty$,*

$$\inf_{(\boldsymbol{X},\boldsymbol{\Sigma})\in\mathscr{F}^*} \mathbb{P}\left[\theta \in \widehat{C}_1(c)\right] = 0, \tag{26}$$

$$\inf_{(\boldsymbol{X},\boldsymbol{\Sigma})\in\mathscr{F}^*} \mathbb{P}\left[\theta \in \widehat{C}_2(c)\right] = 0, \tag{27}$$

$$\inf_{(\boldsymbol{X},\boldsymbol{\Sigma})\in\mathscr{F}} \mathbb{P}\left[\theta \in \widehat{C}_3(c)\right] = 0, \tag{28}$$

and

$$\inf_{(\boldsymbol{X},\boldsymbol{\Sigma})\in\mathscr{F}} \mathbb{P}\left[\theta \in \widehat{C}_4(c)\right] = 0. \tag{29}$$

Equation (26) shows that the worst-case coverage (the exact size) of the CRVE_1 confidence interval equals 0, so that coverage can be arbitrarily distorted from the nominal level, and this

[11] The bound (24) requires $c \geq 1$. This is not an important restriction for inference as all conventional critical values exceed 1.

holds for *any* critical value c. This means that the confidence interval $\widehat{C}_1(c)$ does not uniformly[12] achieve any desired coverage probability. The proof focuses on models which satisfy $\boldsymbol{\Sigma} = \boldsymbol{I}_n$ and clusterwise invertibility, so (26) is not due to heteroskedasticity, correlated errors, or invertibility failure. Rather, it is a consequence of extreme regressor leverage (unbalanced regressors and/or cluster sizes). Theorem 5 is similar to Theorem 4.2 of Preinerstorfer and Pötscher (2016), which establishes results similar to (26)-(29) in the context of heteroskedastic linear regression with HC_1-HC_3 standard errors. While Theorem 5 is stated under the normality assumption, normality is not essential (the theorem holds under Assumptions 1-3 alone), as the theorem extends to any enlarged model class which includes normality as a special case.

Equation (27) shows a similar result for the CRVE_2 confidence interval: the worst-case coverage of the CRVE_2 confidence interval equals 0. Again, this demonstrates that coverage can be arbitrarily distorted from the nominal level. The difference with (26) is that the proof of (27) requires the model class to include non-i.i.d. errors.

Equations (28) and (29) show that the conventional jackknife intervals also have worst-case coverage of 0 if the model class is broadened to include clusterwise noninvertibility. This is due to invertibility failure. These results show that the conventional (e.g., Stata) modification, which is explicitly intended to allow regressions with clusterwise noninvertibility, is not actually robust to clusterwise noninvertibility. These results hold under exactly the same conditions as Theorem 3, highlighting a key difference between the conventional jackknife standard errors $\widehat{v}_3$ and $\widehat{v}_4$ and our proposed standard error $\widehat{v}_5$.

The results (26)-(29) should not be surprising given Theorem 2, which showed that the variance estimators $\widehat{v}_1^2$-$\widehat{v}_4^2$ can be arbitrarily downward biased. What is important about these results is that they show that these confidence intervals have no *a priori* guarantee that they are in any sense a confidence interval. Furthermore, the zero coverage rates of (26)-(29) cannot be fixed by simply using a larger critical value c, as these results hold for any finite c.

Returning to the confidence interval $\widehat{C}_5(c)$, equation (24) bounds its smallest coverage probability in *any* regression satisfying Assumptions 1-4. For example, with the conventional $c = 1.96$ critical value, the bound (24) is 0.70. Thus, the finite sample coverage of $\widehat{C}_5$ with $c = 1.96$ can never be less than 70%. Similarly, the finite sample size of a t-test using the standard error $\widehat{v}_5$ and critical value $c = 1.96$ can never be greater than 30%.

Another implication of (24) is that the Cauchy distribution can be used for finite sample inference (substituting the Cauchy for student t critical values). Doing so will produce inferential statements (hypothesis tests and confidence intervals) with uniform size control. This uniformity holds over all regression designs $\boldsymbol{X}$ and error variances $\boldsymbol{\Sigma}$. In practice, however, it is unlikely that researchers will use the Cauchy distribution for inference, as it is exceedingly conservative.

[12] Here, "uniformly" means over all regression designs $\boldsymbol{X}$.

For example, while the 5% normal critical value is 1.96, that for the Cauchy distribution is 12.7. It is difficult to imagine a user declaring a t-ratio equalling 10 to be "insignificant" simply because it is less than the Cauchy critical value. Instead, one practical message of Theorems 3-5 is that conventional EHW and CRVE confidence intervals can have arbitrary coverage distortion, while the jackknife interval has bounded distortion. This is a strong motivation for replacement of the EHW and CRVE standard errors by simple-to-calculate jackknife standard errors.

Another important feature of Theorem 3 is that it is provides the first generally-applicable uniformly-valid confidence interval for clustered and heteroskedastic regression models. As discussed above, combining jackknife standard errors with Cauchy critical values is uniformly valid, regardless of the regressor or error variance structure. The only other proposals providing partial solutions are Ibragimov and Müller (2010, 2016) and Pötscher and Preinerstorfer (2025). The papers of Ibragimov and Müller achieve uniform size control by estimating the model separately on q sub-samples, so as to produce q independent t-ratios. Such sub-sample estimation is not generally feasible, however. In the heteroskedastic regression model, Pötscher and Preinerstorfer (2025) propose numerically maximizing the largest exact critical value $c(\boldsymbol{\Sigma})$ over the class of feasible variance matrices, conditional on the regressors $\boldsymbol{X}$. Such maximization may be feasible when n is very small (their examples set $n = 25$) but is computationally infeasible for typical sample sizes, and is infeasible under clustered dependence.

The results of Theorems 4 and 5 are related to previous impossibility results. Most famously, Bahadur and Savage (1956) show that any confidence interval for the mean of an unknown distribution (without restrictions on the class of distributions) has an exact size of 0. Similar impossibility results are obtained (in different contexts) by Romano (2004), Hirano and Porter (2012), and Bertanha and Moreira (2020). The common feature of these results is the *impossibility* of confidence interval construction. In contrast, Theorem 5 shows that specific (commonly used) confidence intervals have an exact size of 0, while Theorem 3 shows that a properly constructed jackknife confidence interval has correct size. Therefore, our results are not impossibility results, and we do not describe them as such. Our results also bear resemblance to the following: Dufour (1997), who shows that when a parameter is locally almost unidentified, valid confidence intervals must be unbounded with positive probability; Dufour (2003), who shows that inference on the median of symmetric distributions can only be achieved via a sign test; and Coudin and Dufour (2009), who extend this analysis to median regression.

6 Satterthwaite Approximation

In this section we present a computationally attractive method for critical value construction with improved coverage relative to conventional methods. The method is based on the

Satterthwaite (1946) approximation for a weighted sum of chi-squares, and is similar to those proposed by Bell and McCaffrey (2002), Imbens and Kolesár (2016), Young (2016), and Pustejovsky and Tipton (2018) for HC_1 and HC_2 confidence intervals. Our proposed Satterthwaite $100(1-\alpha)\%$ confidence interval for θ is

$$\widetilde{C}_5 = \widehat{\theta} \pm \frac{t_K^{1-\alpha/2}\widehat{v}_5}{a} \tag{30}$$

where $\widehat{v}_5$ is the jackknife standard error, $t_K^{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the student t distribution with K degrees of freedom, and a is a scale adjustment. The degrees of freedom K and scale adjustment a are sample-dependent and given below by the formula (34)-(39).

Our proposed Satterthwaite p-value for a test of $\theta = \theta_0$ can be calculated using either the student t or the F distribution. Let $T = \left|\widehat{\theta} - \theta\right| / \widehat{v}_5$. Our proposed Satterthwaite p-value is

$$p = 2\left(1 - G(aT;K)\right) = 1 - F\left(a^2T^2;1,K\right) \tag{31}$$

where $G(x;K)$ is the student t distribution with K degrees of freedom and $F(x;1,K)$ is the F distribution with degrees of freedom $(1,K)$. As for the confidence interval (30), the p-value (31) makes both a degrees of freedom adjustment (through K) and a scale adjustment (through a).

The foundation is the following representation for the distribution of the jackknife t-ratio.

Theorem 6 *Under Assumptions 1-4,*

$$\frac{\widehat{\theta} - \theta}{\widehat{v}_5} = \frac{\xi_0}{\sqrt{\sum_{g=1}^{G} \lambda_g \xi_g^2}} \tag{32}$$

where $\xi_0, ..., \xi_G$ *are* $\mathrm{N}(0,1)$, $\xi_1, ..., \xi_G$ *are mutually independent,* λ_g *are the eigenvalues of*

$$\boldsymbol{D} = \left(\boldsymbol{S} + \boldsymbol{U}\boldsymbol{X}'\boldsymbol{\Sigma}\boldsymbol{X}\boldsymbol{U}' - \boldsymbol{V}\boldsymbol{U}' - \boldsymbol{U}\boldsymbol{V}'\right)/\nu^2, \tag{33}$$

$\nu^2 = R'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{\Sigma}\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}R$, $\boldsymbol{U}$ *and* $\boldsymbol{V}$ *are the* $G \times k$ *matrices*

$$\boldsymbol{U} = \begin{bmatrix} \boldsymbol{U}_1' \\ \vdots \\ \boldsymbol{U}_G' \end{bmatrix}, \qquad \boldsymbol{V} = \begin{bmatrix} \boldsymbol{V}_1' \\ \vdots \\ \boldsymbol{V}_G' \end{bmatrix}$$

with rows $\boldsymbol{V}_g = \boldsymbol{X}_g'\boldsymbol{\Sigma}_g\boldsymbol{T}_g$ *and* $\boldsymbol{U}_g = \left(\boldsymbol{X}'\boldsymbol{X} - \boldsymbol{X}_g'\boldsymbol{X}_g\right)^{-}\boldsymbol{X}_g'\boldsymbol{X}_g\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}R$, *where* $\boldsymbol{T}_g = \boldsymbol{Z}_g + \boldsymbol{X}_g\boldsymbol{U}_g$, *and* $\boldsymbol{S}$ *is the* $G \times G$ *diagonal matrix with diagonal elements* $S_g = \boldsymbol{T}_g'\boldsymbol{\Sigma}_g\boldsymbol{T}_g$.

Theorem 6 shows that the t-ratio can be written as the ratio of a normal random variable to the square root of a weighted sum of chi-square random variables. The weights are a computable function of the regressors $\boldsymbol{X}$ and covariance matrices $\boldsymbol{\Sigma}_g$. As the covariance matrices are unknown, we approximate the above distribution by replacing the unknown $\boldsymbol{\Sigma}_g$ with a reference choice $\boldsymbol{\Sigma}_g^0\sigma^2$ for arbitrary scale σ^2. Following the previous literature, we recommend $\boldsymbol{\Sigma}_g^0 = \boldsymbol{I}_{n_g}$, but any other choice can be implemented. The choice $\boldsymbol{\Sigma}_g^0 = \boldsymbol{I}_{n_g}$ has the special implication that in (32), the random variables ξ_0 and $\xi_1^2, ..., \xi_G^2$ are independent. For other choices of $\boldsymbol{\Sigma}_g^0$ this independence does not generally hold.

The Satterthwaite approximation simplifies the distribution (32). It replaces the weighted sum of chi-squares in the denominator of (32) by $a^2\chi_K^2/K$, where χ_K^2 is a single chi-square random variable with K degrees of freedom, and a and K are free parameters selected so that the first two moments of $\sum_{g=1}^G \lambda_g\xi_g^2$ and $a^2\chi_K^2/K$ match. This matching is achieved by setting

$$a = \sqrt{\sum_{g=1}^G \lambda_g} \tag{34}$$

and

$$K = \frac{\left(\sum_{g=1}^G \lambda_g\right)^2}{\sum_{g=1}^G \lambda_g^2}. \tag{35}$$

Furthermore, when $\boldsymbol{\Sigma}_g^0 \neq \boldsymbol{I}_{n_g}$, the variables ξ_0 and χ_K^2 are approximated by independent versions. With these simplifications, (32) equals T_K/a, where T_K is a student t random variable with K degrees of freedom. This leads to the confidence interval and p-value (30) and (31). An important limitation is that the adjustment is explicitly limited to inference on real-valued parameters. It does not apply to tests of joint hypotheses and (to my knowledge) has not been extended to joint tests. If feasible, this would be a useful extension.

In what sense is the Satterthwaite approximation (30)-(31) an approximation? Take the reference model $\boldsymbol{\Sigma}_g^0 = \boldsymbol{I}_{n_g}$. We calculate[13] that the difference between the coverage probabilities is bounded by the uniform absolute difference between the distributions of $\sum_{g=1}^G \lambda_g\xi_g^2$ and $a^2\chi_K^2/K$. Khuri (1995) showed that this difference converges to 0 as $G \to \infty$. This is consistent with the extensive simulations of Bodenham and Adams (2016), who showed that this difference is numerically less than 0.01 when $G \geq 100$.

We can also show that the adjustments vanish in well-balanced large samples, so the confidence interval and p-value (30) and (31) converge to those based on the standard normal. Let

[13]Let the distribution functions of ξ_0^2, $\sum_{g=1}^G \lambda_g\xi_g^2$, and $a^2\chi_K^2/K$ be denoted by F, G, and G_K, with associated density functions f, g, and g_K. Let $T^2 = \xi_0^2/\sum_{g=1}^G \lambda_g\xi_g^2$ and $T_K^2 = \xi_0^2/a^2\chi_K^2/K$. Let $\delta = \sup_x |G(x) - G_K(x)|$. Then $\left|\mathbb{P}\left[T^2 \leq x\right] - \mathbb{P}\left[T_K^2 \leq xa^2\right]\right| = \left|\int F(xq)\left(g(q) - g_K(q)\right)dq\right| = x\left|\int f(xq)\left(G(q) - G_K(q)\right)dq\right| \leq \delta x \int f(xq)dq = \delta$.

$\|\boldsymbol{A}\|$ denote the spectral matrix norm.

Theorem 7 *If* $\left\|\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\right\| = O(1)$ *and* $n^{-1}\max_{g\leq G}\left\|\boldsymbol{X}_g\right\|^2 = o(1)$ *as* $n\to\infty$*, and the coefficients (34) and (35) are calculated with* $\boldsymbol{\Sigma}_g^0 = \boldsymbol{I}_{n_g}$*, then* $a\to 1$ *and* $K\to\infty$ *as* $n\to\infty$*.*

Theorem 7 shows that under mild conditions (the design matrix is uniformly non-singular and no single cluster dominates the sample) the scale adjustment coefficient a converges to 1, and the degrees of freedom parameter K diverges to infinity. Consequently, the interval (30) collapses to that based on the standard normal distribution in well-balanced large samples.

7 Computational Issues

The expressions (34)-(35) are written in terms of the eigenvalues of $\boldsymbol{D}$. In practice, computation of the eigenvalues is not necessary. Alternative computational expressions are as follows. Let D_{gh} denote the ghth element of $\boldsymbol{D}$. Then (34)-(35) equal

$$a = \sqrt{\operatorname{tr}[\boldsymbol{D}]} = \sqrt{\sum_{g=1}^{G} D_{gg}} \tag{36}$$

$$K = \frac{(\operatorname{tr}[\boldsymbol{D}])^2}{\operatorname{tr}[\boldsymbol{D}\boldsymbol{D}]} = \frac{\left(\sum_{g=1}^{G} D_{gg}\right)^2}{\sum_{g=1}^{G}\sum_{h=1}^{G} D_{gh}^2}. \tag{37}$$

These expressions are computationally preferred to (34)-(35), and are convenient for small G.

For large G, explicit calculation of the $G\times G$ matrix $\boldsymbol{D}$ should be avoided. In this case we can replace $\operatorname{tr}[\boldsymbol{D}]$ and $\operatorname{tr}[\boldsymbol{D}\boldsymbol{D}]$ in (36)-(37) with matrix operations. We focus on the reference choice $\boldsymbol{\Sigma}_g^0 = \boldsymbol{I}_{n_g}$, in which case $v^2 = R'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}R$, $\boldsymbol{V}_g^0 = \boldsymbol{X}_g'\boldsymbol{T}_g$, and $S_g = \boldsymbol{T}_g'\boldsymbol{T}_g$. Then, expressions[14] for the components needed for (36)-(37) are

$$\operatorname{tr}[\boldsymbol{D}] = \left(\operatorname{tr}[\boldsymbol{S}\boldsymbol{S}] - \operatorname{tr}\left[\boldsymbol{V}'\boldsymbol{U}\right]\right)/v^2, \tag{38}$$

$$\begin{aligned}\operatorname{tr}[\boldsymbol{D}\boldsymbol{D}] = \big(&\operatorname{tr}[\boldsymbol{S}\boldsymbol{S}] - 2\operatorname{tr}\left(\boldsymbol{V}'\boldsymbol{S}\boldsymbol{U}\right) + \operatorname{tr}\left(\boldsymbol{U}'\boldsymbol{U}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{U}'\boldsymbol{U}\boldsymbol{X}'\boldsymbol{X}\right)\\ &-4\operatorname{tr}\left(\boldsymbol{U}'\boldsymbol{U}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{U}'\boldsymbol{V}\right) + 2\operatorname{tr}\left(\boldsymbol{U}'\boldsymbol{V}\boldsymbol{U}'\boldsymbol{V}\right) + 2\operatorname{tr}\left(\boldsymbol{V}'\boldsymbol{V}\boldsymbol{U}'\boldsymbol{U}\right)\big)/v^4.\end{aligned} \tag{39}$$

The expressions (38)-(39) can be evaluated without explicit computation of $\boldsymbol{D}$. Numerically, we find that the expressions (38)-(39) are (roughly) computationally faster than (36)-(37)

[14] The expressions (38)-(39) use the assumption that the generalized inverse $\boldsymbol{A}^-$ is a reflective generalized inverse (and thus satisfies $\boldsymbol{A}^-\boldsymbol{A}\boldsymbol{A}^- = \boldsymbol{A}^-$) which is true of the Moore-Penrose inverse but not all generalized inverses.

if $G > k$, while (36)-(37) are computationally faster when $G < k$. The derivation of (38)-(39) is presented in the Online Appendix. For clarity, we summarize our recommended procedure for variance estimation, standard errors, confidence intervals, and p-value construction.

1. Calculate the jackknife variance $\widehat{\boldsymbol{V}}_5$ from (10) using the delete-cluster estimator (11) and the Moore-Penrose generalized inverse.

2. For any linear coefficient $\theta = R'\beta$:

   (a) Set $\widehat{\theta} = R'\widehat{\beta}$.

   (b) Calculate its standard error as $\widehat{v}_5 = \sqrt{R'\widehat{\boldsymbol{V}}_5 R}$.

   (c) Calculate the reference adjustment parameters a and K using $\boldsymbol{\Sigma}_g^0 = \boldsymbol{I}_{n_g}$ with the expressions (36)-(37) if $G \leq k$ and (38)-(39) if $G > k$.

   (d) Calculate the critical value $t_K^{1-\alpha/2}$ from the student t distribution with degrees of freedom K.

   (e) Calculate the confidence interval $\widetilde{C}_5$ using (30).

   (f) Calculate the p-value for a test of $\theta = \theta_0$ using (31).

The adjustment parameters a and K are specific to the linear combination R, and hence specific to each individual coefficient. Therefore, the above steps need to be repeated for each element of β in order to calculate intervals and p-values for all coefficients. In the above procedure, we use the scale adjustment a to adjust the intervals and p-values via (30) and (31). A computationally equivalent method is to adjust the standard errors, by calculating and reporting $\widehat{v}_5^* = \widehat{v}_5 / a$, say. I do not recommend this second method, however, as the adjusted standard errors $\widehat{v}_5^*$ do not satisfy the "never downward biased" property of $\widehat{v}_5$. As the standard error is of first-order importance, and many users only examine coefficient estimates and standard errors, it is prudent to report the standard errors without adjustment, along with the adjusted intervals and p-values. This will result in more reliable interpretation of empirical results.

It is common in panel/clustering contexts to include cluster-specific fixed effects. This possibility is allowed in our framework either by explicit inclusion of fixed effect dummy variables in the regressor matrix $\boldsymbol{X}$, or by specifying $\boldsymbol{Y}_g$ and $\boldsymbol{X}_g$ as within-transformed. In the latter case, equation (1) represents the model *after* the within transformation has been applied. Note that in this case, the covariance matrix (3) is that of the within-transformed errors, not the original equation errors. As (3) is unstructured and allows $\boldsymbol{\Sigma}_g$ to be singular, this is without loss of generality. When the fixed effects are applied at the same level as the error clustering, our recommendation is to first apply the within transformation and then apply least squares estimation to

the within transformed variables. This is both computationally and theoretically preferred. It is theoretically preferred because the model after the within transformation is clusterwise invertible, so inference on any linear coefficient is valid under our theory. In contrast, in the model with the included fixed effect dummy variables, only inference on linear coefficients satisfying Assumption 3 is justified under our theory.

8 Simulation Evidence

We present a simulation experiment to investigate performance. The experiment concerns the clustered linear regression (1)-(3). The goal is 95% confidence intervals for the slope coefficients. Our baseline model is the simple regression $\boldsymbol{Y}_g = \mathbf{1}_g\alpha + \boldsymbol{X}_g\beta + \boldsymbol{e}_g$ with $\boldsymbol{X}_g$ a single $n_g \times 1$ regressor. In our baseline specification, the cluster sizes are homogeneous with $n_g = 10$ observations per cluster. Within this model we consider six designs which vary the distributions of the regressors $\boldsymbol{X}_g$ and the heteroskedasticity of the equation errors. Let $\boldsymbol{I}_{n_g}$ denote the $n_g \times n_g$ identity matrix, $\mathbf{1}_g = (1,1,...,1)'$ the $n_g \times 1$ vector of ones, and $\boldsymbol{h}_g = (1,-1,1,-1,...)'$ the $n_g \times 1$ vector containing alternating ± 1.

The designs for the regressors are:

1. Normal with Clustered Dependence: $\boldsymbol{X}_g \sim \mathrm{N}\left(\mathbf{1}_g, \boldsymbol{I}_{n_g} + \mathbf{1}_g\mathbf{1}_g'\right)$.
2. LogNormal with Clustered Dependence: $\boldsymbol{X}_g \sim \exp\left(\mathrm{N}\left(0, \boldsymbol{I}_{n_g} + \mathbf{1}_g\mathbf{1}_g'\right)\right)$.
3. Dummy: $X_{ig} = \begin{cases} 10, \text{ if } g = 1 \text{ and } i \le 2, \text{ or } g = 2 \text{ and } i = 1 \\ 1, \text{ otherwise} \end{cases}$.

In the first design the regressors have a cluster-level random effects covariance matrix. In the second design the regressors are log-normally distributed with the same cluster-level random effects covariance matrix. In the third design the regressor is binary, taking the value 10 for three observations and the value 1 for all other observations. As documented by Chesher and Jewitt (1987) for the LogNormal design, and Imbens and Kolesár (2016) for the Dummy designs, these two designs produce highly leveraged regressor matrices and standard robust covariance matrix estimators can be highly biased.

The designs for the errors are:

1. Normal with Clustered Dependence: $\boldsymbol{e}_g = \boldsymbol{u}_g \sim \mathrm{N}\left(0, \boldsymbol{I}_{n_g} + \mathbf{1}_g\mathbf{1}_g' + \boldsymbol{h}_g\boldsymbol{h}_g'\right)$.
2. Heteroskedastic: $\boldsymbol{e}_g = \boldsymbol{X}_g \odot \boldsymbol{u}_g$ where $\odot$ denotes Hadamard product.

In the first design, the errors are normally distributed with a cluster-level factor covariance matrix. The second design adds conditional heteroskedasticity. The three regressor designs and two error designs combine for a total of six designs. For each design, the number of clusters G is varied among $\{6, 12, 40, 100\}$. As there are 10 observations per cluster, the associated total sample sizes are $\{60, 120, 400, 1000\}$.

For each simulation replication we estimate the coefficients by least squares. We calculate the five standard errors described in the text: CRVE_1 ($\widehat{v}_1$), CRVE_2 ($\widehat{v}_2$), the two conventional jackknife estimators ($\widehat{v}_3$ and $\widehat{v}_4$), and our recommended estimator $\widehat{v}_5$. We calculate confidence intervals using eight methods. The first five are conventional, based on the five standard errors and conventional student t critical values. Thus, given each standard error $\widehat{v}_j$, we form the confidence interval $\widehat{\beta} \pm t_{G-1}^{0.975} \widehat{v}_j$ where $t_{G-1}^{0.975}$ is the 0.975 quantile of the t_{G-1} distribution. We use the $t_{G-1}^{0.975}$ critical value as this is the current implementation in Stata for cluster-robust inference. The next is the Bell and McCaffrey (2002) adjusted t interval (BM) based on the CRVE_2 standard error. It equals $\widehat{\beta} \pm t_K^{0.975} \widehat{v}_2$ where K is calculated similar to (35) under the reference model $\boldsymbol{\Sigma}_g^0 = \boldsymbol{I}_{n_g}$.

The next interval is the restricted wild cluster bootstrap symmetric percentile-t interval, using the standard errors $\widehat{v}_5$ and 999 bootstrap replications. The wild clustered bootstrap is proposed and studied by Cameron, Gelbach, and Miller (2008), Djogbenou, MacKinnon, and Nielsen (2019), Canay, Santos, and Shaikh (2021), and MacKinnon, Nielsen and Webb (2023b). The latter paper provides considerable simulation evidence regarding the performance of several implementations, and their evidence strongly supports using this implementation, which they call WCR-V. We follow this implementation, though using $\text{N}(0,1)$ auxiliary bootstrap errors[15], and constructing confidence intervals through test inversion. We provide a thorough description of our implementation in the Online Appendix.

Our final interval is our Satterthwaite interval (30) using the reference model $\boldsymbol{\Sigma}_g^0 = \boldsymbol{I}_{n_g}$.

We compute the actual coverage probability of these nominal 95% intervals by simulation with 20,000 replications. These estimates are precise, as their standard errors are all less than 0.003. We also compute the average length of these intervals, though we only report this length for the final four confidence intervals (conventional with jackknife $\widehat{v}_5$, Bell-McCaffrey, wild bootstrap, and Satterthwaite).

We report the results for the baseline regression model in Table 1. The top block reports the results for $G = 6$. The first five columns are the coverage probabilities of the conventional confidence intervals. We can see that the conventional CRVE_1 confidence interval has substantial

[15] MacKinnon, Nielsen and Webb (2023b) recommend Rademacher auxiliary errors. This works poorly in our simulation design due to the small cluster sizes and heavy regressor leverage. Consequently, we use normal auxiliary errors, which has much better performance.

Table 1: Baseline Regression Model. Coverage of Nominal 95% Confidence Intervals for β

		Coverage Probability								Interval Length			
	Cr. Value	Conventional t_{G-1}					BM	Wild	Satt	Jack	BM	Wild	Satt
	St. Error	$\widehat{v}_1$	$\widehat{v}_2$	$\widehat{v}_3$	$\widehat{v}_4$	$\widehat{v}_5$	$\widehat{v}_2$	$\widehat{v}_5$	$\widehat{v}_5$	$\widehat{v}_5$	$\widehat{v}_2$	$\widehat{v}_5$	$\widehat{v}_5$
$G = 6$	Normal	0.91	0.93	0.95	0.95	0.96	0.95	0.91	0.96	1.4	1.3	0.5	1.4
Clustered	LogNorm	0.85	0.91	0.95	0.96	0.96	0.99	0.95	0.99	0.7	0.8	0.3	1.1
	Dummy	0.86	0.89	0.93	0.93	0.94	1.00	0.96	1.00	0.8	2.0	0.4	2.2
	Normal	0.89	0.91	0.93	0.93	0.94	0.93	0.90	0.94	3.2	2.9	2.7	3.2
Hetero.	LogNorm	0.57	0.70	0.86	0.87	0.89	0.90	0.92	0.93	5.8	6.3	16	9.1
	Dummy	0.70	0.77	0.84	0.84	0.85	0.95	0.95	0.95	7.2	19	48	21
$G = 12$	Normal	0.92	0.93	0.95	0.95	0.95	0.95	0.92	0.95	0.8	0.8	0.2	0.8
Clustered	LogNorm	0.84	0.89	0.94	0.94	0.94	0.99	0.95	0.99	0.3	0.4	0.0	0.5
	Dummy	0.75	0.81	0.88	0.88	0.89	1.00	0.95	1.00	0.6	2.2	0.5	2.2
	Normal	0.91	0.92	0.93	0.93	0.94	0.94	0.92	0.94	1.9	1.8	1.2	1.9
Hetero.	LogNorm	0.61	0.73	0.86	0.86	0.88	0.92	0.93	0.94	4.2	5.3	11	7.5
	Dummy	0.64	0.74	0.82	0.82	0.83	0.95	0.94	0.95	6.2	20	51	22
$G = 40$	Normal	0.94	0.94	0.95	0.95	0.95	0.95	0.94	0.95	0.4	0.4	0.0	0.4
Clustered	LogNorm	0.86	0.90	0.93	0.93	0.93	0.97	0.94	0.98	0.1	0.1	0.0	0.2
	Dummy	0.65	0.75	0.83	0.83	0.84	1.00	0.94	0.99	0.6	2.2	0.5	2.3
	Normal	0.93	0.94	0.94	0.94	0.94	0.94	0.94	0.94	1.0	1.0	0.3	1.0
Hetero.	LogNorm	0.70	0.79	0.88	0.88	0.89	0.93	0.93	0.95	3.0	3.6	5.4	4.9
	Dummy	0.60	0.71	0.81	0.81	0.81	0.95	0.95	0.95	5.7	22	52	22
$G = 100$	Normal	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.2	0.2	0.0	0.2
Clustered	LogNorm	0.89	0.91	0.93	0.93	0.93	0.97	0.93	0.97	0.1	0.1	0.0	0.1
	Dummy	0.61	0.72	0.81	0.81	0.81	0.98	0.94	0.97	0.6	2.3	0.5	2.3
	Normal	0.94	0.94	0.95	0.95	0.95	0.94	0.94	0.95	0.6	0.6	0.1	0.6
Hetero.	LogNorm	0.76	0.83	0.89	0.89	0.90	0.93	0.93	0.95	2.3	2.6	3.2	3.4
	Dummy	0.59	0.70	0.80	0.80	0.80	0.95	0.95	0.95	5.6	22	52	22

under-coverage in most designs. The worst case is LogNormal regressors with heteroskedastic error variances, where the interval has only 57% coverage. The CRVE_2 confidence interval has slightly better coverage, but still substantially undercovers in most designs. The jackknife $\widehat{v}_5$ confidence interval has better coverage, but undercovers (as low as 85%) in the leveraged/heteroskedastic designs.

The next column is the coverage probability of the BM confidence interval, which has improved coverage relative to the conventional methods. Its coverage rates are somewhat sensitive to the design, being excessively conservative (100% coverage) in some designs and undercovering (90%) in others. The next column is the coverage probability of the wild bootstrap. The coverage rates are generally excellent when compared to conventional methods. The intervals undercover (90%) in some designs. There are no cases of excessive conservative coverage. The

following column is the coverage probability of the Satterthwaite interval. The coverage rates are excellent, uniformly exceeding 93%. The intervals are excessively conservative (100%) in some designs.

Table 2: Regression on Dummy. Coverage of Nominal 95% Confidence Intervals for β

		Coverage Probability								Interval Length			
	Cr. Value	Conventional t_{G-1}					BM	Wild	Satt	Jack	BM	Wild	Satt
	St. Error	$\widehat{v}_1$	$\widehat{v}_2$	$\widehat{v}_3$	$\widehat{v}_4$	$\widehat{v}_5$	$\widehat{v}_2$	$\widehat{v}_5$	$\widehat{v}_5$	$\widehat{v}_5$	$\widehat{v}_2$	$\widehat{v}_5$	$\widehat{v}_5$
$G=6$	Normal	0.90	0.93	0.93	0.94	0.96	0.95	0.91	0.96	1.4	1.3	0.5	1.4
Clustered	LogNorm	0.85	0.91	0.92	0.93	0.96	0.99	0.95	0.99	0.7	0.8	0.3	1.1
	Dummy	0.81	0.86	0.77	0.80	0.91	1.00	0.97	0.99	0.6	2.1	0.4	2.1
	Normal	0.89	0.91	0.90	0.91	0.94	0.93	0.90	0.94	3.2	2.9	2.8	3.2
Hetero.	LogNorm	0.58	0.70	0.79	0.81	0.89	0.90	0.92	0.93	5.9	6.5	16	9.1
	Dummy	0.71	0.78	0.67	0.69	0.85	0.96	0.95	0.96	6.9	22	48	22
$G=12$	Normal	0.92	0.93	0.94	0.94	0.95	0.95	0.92	0.95	0.8	0.8	0.2	0.8
Clustered	LogNorm	0.83	0.89	0.92	0.92	0.95	0.99	0.94	0.99	0.3	0.4	0.0	0.5
	Dummy	0.73	0.81	0.73	0.74	0.87	1.00	0.96	0.99	0.5	2.1	0.4	2.1
	Normal	0.91	0.92	0.93	0.93	0.94	0.94	0.92	0.94	1.9	1.9	1.2	1.9
Hetero.	LogNorm	0.61	0.73	0.83	0.84	0.88	0.92	0.93	0.94	4.3	5.3	11	7.5
	Dummy	0.66	0.74	0.66	0.67	0.82	0.95	0.95	0.95	5.9	22	48	22
$G=40$	Normal	0.94	0.94	0.95	0.95	0.95	0.95	0.94	0.95	0.4	0.4	0.0	0.4
Clustered	LogNorm	0.86	0.90	0.93	0.93	0.93	0.98	0.94	0.98	0.1	0.1	0.0	0.2
	Dummy	0.67	0.77	0.70	0.71	0.85	0.99	0.95	0.98	0.5	2.1	0.4	2.1
	Normal	0.93	0.94	0.94	0.94	0.94	0.94	0.94	0.94	1.0	1.0	0.3	1.0
Hetero.	LogNorm	0.70	0.79	0.87	0.88	0.89	0.93	0.93	0.95	3.0	3.6	5.5	4.9
	Dummy	0.62	0.72	0.65	0.65	0.81	0.95	0.95	0.95	5.4	23	48	23
$G=100$	Normal	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.2	0.2	0.0	0.2
Clustered	LogNorm	0.89	0.91	0.93	0.93	0.93	0.97	0.93	0.97	0.1	0.1	0.0	0.1
	Dummy	0.66	0.76	0.70	0.70	0.84	0.98	0.95	0.97	0.5	2.1	0.4	2.1
	Normal	0.94	0.94	0.94	0.94	0.95	0.94	0.94	0.95	0.6	0.6	0.1	0.6
Hetero.	LogNorm	0.76	0.83	0.89	0.89	0.90	0.93	0.93	0.95	2.3	2.6	3.2	3.4
	Dummy	0.61	0.71	0.65	0.65	0.80	0.95	0.95	0.95	5.3	23	49	23

It is instructive to use the results to separate the effect of the jackknife standard error from the effect of the Satterthwaite correction. Take the LogNormal heteroskedastic design. Moving from CRVE_1 to jackknife standard errors improves the coverage rates from 57% to 89%, and the Satterthwaite correction improves the coverage from 89% to 93%. Take the Dummy heteroskedastic design. Moving from CRVE_1 to jackknife improves the coverage rates from 70% to 85%, and the Satterthwaite further improves the coverage to 95%. In general, we see that the largest effect is due to the use of jackknife standard errors, with the Satterthwaite correction a secondary, but important, effect.

Table 3: Regression on Dummy. Coverage of Nominal 95% Confidence Intervals for γ

		Coverage Probability								Interval Length			
	Cr. Value	Conventional t_{G-1}					BM	Wild	Satt	Jack	BM	Wild	Satt
	St. Error	$\widehat{v}_1$	$\widehat{v}_2$	$\widehat{v}_3$	$\widehat{v}_4$	$\widehat{v}_5$	$\widehat{v}_2$	$\widehat{v}_5$	$\widehat{v}_5$	$\widehat{v}_5$	$\widehat{v}_2$	$\widehat{v}_5$	$\widehat{v}_5$
$G = 6$	Normal	0.63	0.66	0.68	0.68	1.00	0.73	1.00	1.00	6.1	3.0	6.5	11
Clustered	LogNorm	0.64	0.66	0.66	0.66	1.00	0.71	1.00	1.00	5.9	2.9	6.1	11
	Dummy	0.64	0.66	0.65	0.65	1.00	0.76	1.00	1.00	5.8	3.0	5.9	11
	Normal	0.70	0.73	0.74	0.75	1.00	0.79	0.98	1.00	8.4	4.5	15	14
Hetero.	LogNorm	0.72	0.77	0.79	0.80	1.00	0.81	0.99	1.00	18	12	111	34
	Dummy	0.86	0.91	0.85	0.87	1.00	0.95	1.00	1.00	14	11	42	27
$G = 12$	Normal	0.48	0.49	0.51	0.51	1.00	0.52	1.00	1.00	4.4	1.6	5.0	14
Clustered	LogNorm	0.47	0.47	0.48	0.48	1.00	0.50	1.00	1.00	4.4	1.6	4.9	15
	Dummy	0.47	0.49	0.48	0.48	1.00	0.77	1.00	1.00	4.4	2.8	4.9	13
	Normal	0.58	0.60	0.61	0.62	1.00	0.62	1.00	1.00	6.1	2.6	12	19
Hetero.	LogNorm	0.62	0.67	0.71	0.72	1.00	0.69	1.00	1.00	15	9	121	48
	Dummy	0.77	0.86	0.79	0.80	1.00	0.97	1.00	1.00	11	14	35	33
$G = 40$	Normal	0.27	0.28	0.28	0.28	1.00	0.28	1.00	1.00	3.6	0.8	4.1	19
Clustered	LogNorm	0.26	0.26	0.26	0.26	1.00	0.27	1.00	1.00	3.6	0.7	4.1	19
	Dummy	0.29	0.33	0.30	0.31	1.00	0.84	1.00	1.00	3.7	3.4	4.2	15
	Normal	0.38	0.38	0.39	0.39	1.00	0.39	1.00	1.00	4.8	1.3	10	25
Hetero.	LogNorm	0.49	0.54	0.59	0.59	1.00	0.55	1.00	1.00	14	6.3	132	67
	Dummy	0.68	0.82	0.74	0.74	1.00	0.99	1.00	1.00	9.5	24	31	38
$G = 100$	Normal	0.17	0.17	0.17	0.17	1.00	0.18	1.00	1.00	3.4	0.5	3.8	20
Clustered	LogNorm	0.16	0.16	0.16	0.16	1.00	0.16	1.00	1.00	3.4	0.5	3.8	20
	Dummy	0.22	0.27	0.24	0.24	1.00	0.84	1.00	1.00	3.5	3.8	3.9	15
	Normal	0.26	0.26	0.26	0.26	1.00	0.26	1.00	1.00	4.5	0.8	9.0	27
Hetero.	LogNorm	0.42	0.47	0.52	0.52	1.00	0.47	1.00	1.00	13	4.6	120	72
	Dummy	0.66	0.81	0.72	0.73	1.00	0.99	1.00	1.00	9.1	33	31	39

The final four columns report the average length of four confidence interval methods: conventional with jackknife $\widehat{v}_5$, Bell-McCaffrey, wild bootstrap, and Satterthwaite. We focus on these four as they are the only methods with reasonable coverage rates. The lengths vary as expected. In cases where the confidence intervals have similar coverage rates, then the interval lengths are also similar. In cases where the confidence intervals have differing coverage rates, the interval lengths differ similarly. In most cases the lengths of the most conservative (Satterthwaite) interval is not much longer than the BM and wild bootstrap intervals. If we compare the wild bootstrap with the Satterthwaite intervals, we can see considerable differences, with cases where the wild bootstrap has shorter intervals, and other cases (the leveraged/heteroskedastic designs) where the wild bootstrap produces much longer intervals. We do not have a good explanation for this phenomenon, and believe it is a interesting topic for future research.

Table 4: Heterogeneous Cluster Sizes. Coverage of Nominal 95% Confidence Intervals for β

		Coverage Probability								Interval Length			
	Cr. Value	Conventional t_{G-1}					BM	Wild	Satt	Jack	BM	Wild	Satt
	St. Error	$\widehat{v}_1$	$\widehat{v}_2$	$\widehat{v}_3$	$\widehat{v}_4$	$\widehat{v}_5$	$\widehat{v}_2$	$\widehat{v}_5$	$\widehat{v}_5$	$\widehat{v}_5$	$\widehat{v}_2$	$\widehat{v}_5$	$\widehat{v}_5$
$G=6$	Normal	0.75	0.87	0.95	0.96	0.97	0.97	0.93	0.98	1.7	1.6	0.9	2.0
Clustered	LogNorm	0.79	0.89	0.95	0.96	0.97	0.99	0.95	0.99	1.0	1.2	0.8	1.5
	Dummy	0.77	0.84	0.90	0.91	0.92	0.99	0.95	0.99	0.6	1.7	0.3	1.8
	Normal	0.74	0.82	0.89	0.89	0.91	0.93	0.91	0.93	4.1	4.5	6.1	5.1
Hetero.	LogNorm	0.48	0.63	0.86	0.86	0.88	0.89	0.93	0.91	6.3	7.4	21	9.6
	Dummy	0.69	0.76	0.83	0.83	0.85	0.95	0.95	0.95	7.0	20	49	21
$G=12$	Normal	0.67	0.83	0.95	0.95	0.96	0.96	0.93	0.98	1.1	1.1	0.5	1.5
Clustered	LogNorm	0.75	0.86	0.95	0.95	0.96	0.99	0.95	0.99	0.5	0.7	0.2	0.8
	Dummy	0.70	0.80	0.88	0.89	0.90	0.99	0.95	0.99	0.5	1.8	0.3	1.9
	Normal	0.68	0.79	0.87	0.87	0.89	0.92	0.91	0.93	2.9	3.2	4.0	3.9
Hetero.	LogNorm	0.50	0.63	0.85	0.85	0.86	0.90	0.93	0.93	4.7	6.2	15	8.4
	Dummy	0.64	0.73	0.81	0.82	0.82	0.95	0.95	0.95	6.0	21	49	21
$G=40$	Normal	0.54	0.78	0.96	0.96	0.96	0.96	0.94	0.99	0.8	0.8	0.4	1.1
Clustered	LogNorm	0.73	0.86	0.95	0.96	0.96	0.99	0.95	0.99	0.2	0.3	0.0	0.4
	Dummy	0.64	0.77	0.87	0.87	0.88	0.99	0.95	0.99	0.5	1.8	0.3	1.9
	Normal	0.62	0.74	0.86	0.86	0.86	0.89	0.91	0.93	2.4	2.5	3.1	3.2
Hetero.	LogNorm	0.55	0.68	0.85	0.86	0.86	0.92	0.93	0.94	3.5	4.7	9.0	6.6
	Dummy	0.60	0.71	0.81	0.81	0.81	0.95	0.95	0.95	5.6	22	49	22
$G=100$	Normal	0.46	0.76	0.97	0.97	0.97	0.97	0.95	0.99	0.8	0.7	0.3	1.0
Clustered	LogNorm	0.71	0.85	0.96	0.96	0.96	0.98	0.94	0.99	0.2	0.2	0.0	0.3
	Dummy	0.64	0.76	0.87	0.87	0.87	0.99	0.94	0.99	0.5	1.8	0.3	1.9
	Normal	0.57	0.71	0.84	0.84	0.84	0.87	0.91	0.92	2.2	2.3	2.8	3.0
Hetero.	LogNorm	0.57	0.69	0.86	0.86	0.86	0.92	0.93	0.94	3.0	3.8	6.4	5.4
	Dummy	0.59	0.70	0.80	0.80	0.80	0.95	0.95	0.95	5.4	22	49	22

The following blocks are for $G=12$, $G=40$, and $G=100$. Qualitatively, the results are similar to the $G=6$ case. For some designs and methods the coverage probabilities improve slightly as G increases, but in other cases the coverage probabilities worsen. The best coverage is obtained by the Satterthwaite interval, with coverage uniformly exceeding 94%. The BM and wild bootstrap have coverage uniformly exceeding 92%.

We expand the analysis by examining a model with clusterwise noninvertibility. This is $\boldsymbol{Y}_g = \mathbf{1}_g\alpha + \boldsymbol{X}_g\beta + \boldsymbol{D}_g\gamma + \boldsymbol{e}_g$ with $\boldsymbol{D}_g$ a dummy indicator for the first cluster. This model is cluster-level treatment with a single treated cluster. The regression is clusterwise noninvertible as the least squares estimator is undefined when the first cluster is omitted. In this model we examine confidence intervals for both β and γ. For this model we consider the same six designs as in the baseline model for the distributions of the regressors and error variances.

Table 5: LogNormal Errors. Coverage of Nominal 95% Confidence Intervals for β

		Coverage Probability								Interval Length			
	Cr. Value	Conventional t_{G-1}					BM	Wild	Satt	Jack	BM	Wild	Satt
	St. Error	$\widehat{v}_1$	$\widehat{v}_2$	$\widehat{v}_3$	$\widehat{v}_4$	$\widehat{v}_5$	$\widehat{v}_2$	$\widehat{v}_5$	$\widehat{v}_5$	$\widehat{v}_5$	$\widehat{v}_2$	$\widehat{v}_5$	$\widehat{v}_5$
$G=6$	Normal	0.93	0.94	0.96	0.96	0.97	0.96	0.93	0.97	1.2	1.1	0.5	1.2
Clustered	LogNorm	0.86	0.92	0.96	0.96	0.97	0.99	0.96	0.99	0.6	0.7	0.3	1.0
	Dummy	0.85	0.89	0.92	0.92	0.94	1.00	0.97	1.00	0.6	1.7	0.5	1.8
	Normal	0.84	0.86	0.89	0.89	0.91	0.89	0.86	0.90	2.6	2.4	2.7	2.7
Hetero.	LogNorm	0.55	0.68	0.84	0.85	0.86	0.88	0.91	0.91	4.9	5.3	16	8
	Dummy	0.65	0.72	0.79	0.80	0.81	0.94	0.93	0.94	6.0	15	50	17
$G=12$	Normal	0.94	0.95	0.96	0.96	0.97	0.96	0.94	0.97	0.7	0.7	0.2	0.7
Clustered	LogNorm	0.86	0.91	0.95	0.95	0.96	0.99	0.96	0.99	0.3	0.3	0.0	0.4
	Dummy	0.74	0.80	0.86	0.86	0.87	1.00	0.96	1.00	0.5	1.8	0.5	1.0
	Normal	0.86	0.87	0.88	0.88	0.89	0.89	0.87	0.89	1.6	1.6	1.1	1.6
Hetero.	LogNorm	0.60	0.70	0.83	0.83	0.84	0.89	0.90	0.92	3.5	4.4	10	6.2
	Dummy	0.59	0.68	0.77	0.78	0.78	0.94	0.93	0.94	5.1	17	50	18
$G=40$	Normal	0.95	0.96	0.96	0.96	0.96	0.96	0.95	0.96	0.4	0.3	0.0	0.3
Clustered	LogNorm	0.88	0.91	0.94	0.94	0.94	0.98	0.94	0.98	0.1	0.1	0.0	0.1
	Dummy	0.62	0.71	0.79	0.80	0.80	1.00	0.93	0.99	0.5	1.9	0.5	1.9
	Normal	0.89	0.90	0.90	0.90	0.91	0.90	0.90	0.90	0.9	0.9	0.3	0.9
Hetero.	LogNorm	0.68	0.76	0.85	0.85	0.85	0.90	0.90	0.93	2.5	3.1	5.4	4.1
	Dummy	0.56	0.66	0.76	0.76	0.76	0.94	0.94	0.94	4.7	18	53	19
$G=100$	Normal	0.96	0.96	0.96	0.96	0.96	0.96	0.95	0.96	0.2	0.2	0.0	0.2
Clustered	LogNorm	0.90	0.92	0.94	0.94	0.94	0.97	0.93	0.97	0.0	0.1	0.0	0.1
	Dummy	0.58	0.68	0.77	0.77	0.77	0.99	0.92	0.98	0.5	1.9	0.5	1.9
	Normal	0.92	0.92	0.92	0.92	0.92	0.92	0.92	0.92	0.6	0.6	0.1	0.6
Hetero.	LogNorm	0.75	0.81	0.87	0.87	0.88	0.91	0.91	0.93	2.0	2.3	3.0	2.9
	Dummy	0.55	0.65	0.76	0.76	0.76	0.94	0.94	0.94	4.6	19	53	19

Table 2 presents the results for confidence intervals for β. In general, the results are similar to those of Table 1 but with one important difference. In Table 2 we see a meaningful divergence in performance between the intervals based on conventional jackknife standard errors $\widehat{v}_3$ and $\widehat{v}_4$ and those based on our recommended jackknife standard error $\widehat{v}_5$. In Table 1, these three methods were nearly identical; in Table 2 we can see that the two conventional intervals exhibit substantial undercoverage in many designs. This difference is due to the treatment of noninvertible clusters. This strongly supports our recommendation for $\widehat{v}_5$ over $\widehat{v}_3$ and $\widehat{v}_4$.

Next, we examine Table 3, which presents the results for the dummy variable coefficient γ in the regression model $\boldsymbol{Y}_g = \mathbf{1}_g\alpha + \boldsymbol{X}_g\beta + \boldsymbol{D}_g\gamma + \boldsymbol{e}_g$. This is a treacherous context, as we are essentially making inference for a coefficient based on a single cluster. The results in Table 3 reveal that the conventional inference methods fail, and worsen as the sample size increases. The Bell-

McCaffrey adjusted interval similarly fails. The exceptions (which generally produce coverage rates of 100%) are the three intervals (conventional, wild bootstrap, and Satterthwaite) using the jackknife $\widehat{v}_5$ standard error. The reason why the intervals based on the jackknife standard error $\widehat{v}_5$ have coverage rates of 100% is because in this context $\widehat{v}_5$ tends to approximately equal the coefficient estimate $\widehat{\gamma}$, so the confidence interval always covers the true value of 0. Coverage is conservative, but at least there is no tendency towards false significance. We can also see that the average lengths of these three feasible intervals are different from one another, with heterogeneous rankings. Overall, the message of Table 3 is that in the extreme context of inference on a dummy indicator for a single treated cluster, the Bell-McCaffrey interval is not robust, the wild bootstrap and Satterthwaite intervals are robust, but can be highly conservative.

We next explore the impact of heterogeneous cluster sizes. In Tables 1-3, cluster sizes were all set at $n_g = 10$. In our next experiment we set two clusters to have $2 + 4G$ observations, with the remainder with 2 observations each. This means that the total number of observations is the same as before, but the samples are dominated by two large clusters. Otherwise, the experiment follows the baseline model. The results are presented in Table 4. Most inference methods are adversely affected, with coverage rates meaningfully lower in Table 4 relative to Table 1. Generally good coverage is obtained by the Satterthwaite confidence interval. Interval lengths follow similar patterns as in Table 1.

Our inference theory is developed under the assumption of normally distributed errors, which raises the question if the coverage rates are sensitive to departures to normality. To explore this question, we repeat the analysis of the baseline model, but with the errors drawn from a LogNormal distribution, centered and scaled to have a mean of zero and variance of one. We sample the regression errors from this skewed distribution, with the same cluster covariances as in the baseline model. The results are displayed in Table 5. Comparing Tables 1 and 5, we can see that the coverage rates of most methods worsen in Table 5. The worst case coverage of the Bell-McCaffrey interval is 88%, of the wild bootstrap is 86%, and of the Satterthwaite interval is 89%. These coverage rates improve, however, as G increases.

It is instructive to contrast the coverage rates of Tables 1-5 with the lower bound of Theorem 3. Recall, the latter states under normal errors and inference on the slope β, the coverage rates of a nominal 95% interval using the jackknife $\widehat{v}_5$ can never be lower than 70%. Indeed, the coverage rates for the jackknife interval uniformly exceed this bound; in fact, the coverage rates are never lower than 80% under normal errors. In contrast, the intervals using $\widehat{v}_1$ have coverage rates as low as 46% for β, and those using $\widehat{v}_2$ have coverage rates as low as 63%.

In unreported simulations, we explored the impact of inclusion of a large number of extraneous regressors. Cattaneo, Jansson and Newey (2018) showed that the downward bias of conventional variance estimators is increasing in the number of extraneous regressors, while

jackknife variance estimators can be upward biased. While our theory allows for an arbitrary number of regressors under normal errors, it is unclear how the results extend to the case of nonnormal errors. Our simulation results are qualitatively similar to those in the reported tables. We also explored the performance of alternative inference methods, including the pairs clustered bootstrap and alternative wild bootstrap methods. None of these alternatives had good coverage accuracy, and so are not reported. We also explored the performance of the Satterthwaite interval calculated with alternative reference covariance matrix models (for example, $\boldsymbol{\Sigma}_g = \boldsymbol{X}_g \boldsymbol{X}_g'$). These produced qualitatively similar results, but more conservative, than the reported Satterthwaite interval.

In summary, we are able to draw the following conclusions from the simulation evidence. First, the standard error which produces intervals with the best coverage rates is $\widehat{\nu}_5$. Second, the simple interval $\widehat{\beta} \pm t_{G-1}^{1-\alpha/2} \widehat{\nu}_5$ has good coverage rates in many contexts, but can undercover in extreme designs. Third, the adjusted interval $\widehat{\beta} \pm t_K^{1-\alpha/2} \widehat{\nu}_5 / a$ has excellent (but sometimes conservative) coverage in all contexts examined. Fourth, the restricted wild bootstrap with jackknife standard errors also is an excellent option for inference. Fifth, issues such as clusterwise invertibility should not be handled by *ad hoc* computational implementations, but rather by methods justified by theoretical insight. In particular, the jackknife should be implemented without the discarding of iterations with noninvertible design matrices.

9 Empirical Illustration

We illustrate the applicability of the methods with an empirical example. We follow Canay, Santos, and Shaikh (2021) by revisiting an application by Meng, Qian, and Yared (2015) into the causes of the Chinese Great Famine between 1958 and 1960. Their regressions (Table 2 of Meng-Qian-Yared) take the form $Y = Z_1\beta_1 + Z_2\beta_2 + W'\gamma + e$ for $G = 19$ provinces between 1953 and 1982, where Y equals the log of deaths in the province, Z_1 equals the log of predicted grain production, Z_2 equals the product of Z_1 and an indicator for a famine year, and W are other controls. The focus is on the coefficient sum $\beta_1 + \beta_2$. The authors report six specifications which vary the sample period (1953-1982 vs 1953-1965), the provinces (19 vs 23), and replacing predicted with reported grain production. Canay, Santos, and Shaikh (2021) use this application to illustrate hypothesis testing using the cluster wild bootstrap. In contrast, we are interested in standard error calculation, confidence interval construction, and hypothesis testing. Following these authors, we cluster by province.

We estimate the same six regression specifications as Meng, Qian, and Yared (2015) and focus on the coefficient sum $\beta_1 + \beta_2$. In Table 6 we report the least squares estimates $\widehat{\beta}_1 + \widehat{\beta}_2$ plus three standard errors: CRVE_1, CRVE_2, and jackknife. What you can see from the table is that

in some of the specifications there are considerable differences between the three standard errors, and in particular between the jackknife and the other two. The discrepancies between the CRVE_1 and jackknife standard errors range from 10% (in specification #1) to 66% (in specification #3). These are large and substantial differences.

We next construct 95% adjusted confidence intervals for the coefficient sum $\beta_1+\beta_2$. We start by calculating the adjustment coefficients K and a for each of the six specifications, and report these coefficients in Table 6. The values for the adjusted degrees of freedom K range between 4 and 6, which are all small. The values for the scale adjustment a range between 1.17 and 1.26. Together, these are used to construct the confidence intervals, which are reported in the table. Take, for example, specification #1, where $\widehat{\beta}_1+\widehat{\beta}_2 = 0.141$, $\widehat{v}_5 = 0.066$, $K = 4.18$, and $a = 1.21$. The 95% critical value from the t distribution with $K = 4.18$ degrees of freedom is 2.73. The 95% confidence interval is therefore $0.141 \pm 2.73 \times 0.066/1.21 = [-0.01, 0.29]$. The confidence intervals are wide, indicating uncertainty about the value of the coefficient sum. The intervals do not vary greatly across the six specifications, indicating that the result is reasonably robust to the specification.

We also construct and report adjusted p-values for t-tests of the hypothesis $\beta_1+\beta_2 = 0$. The t-statistic using the jackknife standard error is $0.141/0.066 = 2.16$. The adjusted p-value is $1 - F\left(1.21^2 \times 2.16^2; 1, 4.18\right) = 0.058$. None of the six p-values are significant at the 5% level. This contrasts with the p-values reported by Meng, Qian, and Yared (2015), which were all statistically significant, some greatly so. Several of our p-values are similar to those calculated by the wild cluster bootstrap as reported by Canay, Santos, and Shaikh (2021), which we report in the final row of the table. For example, for the baseline specification #1, our p-value of 0.058 is nearly identical to their wild studentized p-value of 0.061.

10 Conclusion

Heteroskedasticity-consistent and cluster-robust standard errors are routinely reported in applied econometric practice. It is prudent to coalesce on simple yet well-behaved methods which produce reliable inference across reasonable estimation settings. It is our contention that jackknife variance estimators are superior to conventional (EHW and CRVE_1) estimators, based on our analysis of worst-case downward bias and confidence interval coverage. They are also computationally simple to implement. Furthermore, we recommend that conventional statistical software calculate default confidence intervals and p-values using the Satterthwaite adjustment, as it is computationally simple, more conservative than student t inference, and more reliable in finite samples.

Table 6: China's Great Famine, 1959-1961

	Dependent variable: log deaths in year $t+1$					
	Constructed grain production				Reported grain production	
	19 provinces		23 provinces		19 provinces	
	1953-1982	1953-1965	1953-1982	1953-1965	1953-1982	1953-1965
	(1)	(2)	(3)	(4)	(5)	(6)
$\widehat{\beta}_1+\widehat{\beta}_2$	0.141	0.098	0.115	0.094	0.113	0.089
CRVE_1	(0.060)	(0.053)	(0.037)	(0.037)	(0.063)	(0.059)
CRVE_2	[0.061]	[0.056]	[0.039]	[0.037]	[0.068]	[0.063]
Jackknife	⟨0.066⟩	⟨0.066⟩	⟨0.061⟩	⟨0.052⟩	⟨0.079⟩	⟨0.073⟩
K	4.18	3.95	5.34	5.03	5.18	5.97
a	1.21	1.26	1.21	1.23	1.18	1.17
Interval	$[-0.01, 0.29]$	$[-0.05, 0.24]$	$[-0.01, 0.24]$	$[-0.01, 0.20]$	$[-0.06, 0.28]$	$[-0.06, 0.24]$
p-value	0.058	0.135	0.069	0.076	0.151	0.200
CSS p-value	0.061	0.072	0.029	0.030	0.141	0.171

Notes: All regressions include log total population, log urban population, and year fixed effects. Standard errors are clustered by province. The CSS p-value is that reported in Canay, Santos, and Shaikh (2021).

A Appendix: Technical Proofs

Let $\lambda_{\min}(\boldsymbol{A})$ and $\lambda_{\max}(\boldsymbol{A})$ denote the smallest and largest eigenvalue of a Hermitian matrix $\boldsymbol{A}$, $\|\boldsymbol{A}\| = \left(\lambda_{\max}\left(\boldsymbol{A}'\boldsymbol{A}\right)\right)^{1/2}$ denote the spectral matrix norm, and $\|\boldsymbol{A}\|_F = \left(\text{tr}\left[\boldsymbol{A}'\boldsymbol{A}\right]\right)^{1/2}$ denote the Frobenius matrix norm.

Proof of Theorem 1: The estimator (11) with any generalized inverse is a least-squares minimizer and solves the first order condition

$$\left(\boldsymbol{X}'\boldsymbol{X} - \boldsymbol{X}_g'\boldsymbol{X}_g\right)\widehat{\beta}_{-g} = \left(\boldsymbol{X}'\boldsymbol{Y} - \boldsymbol{X}_g'\boldsymbol{Y}_g\right).$$

Pre-multiplying by $\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$, rearranging, and using (12), we obtain

$$\begin{aligned}\widehat{\beta}_{-g} &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{Y} - \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}_g'\boldsymbol{Y}_g + \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}_g'\boldsymbol{X}_g\widehat{\beta}_{-g} \\ &= \widehat{\beta} - \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}_g'\left(\boldsymbol{Y}_g - \boldsymbol{X}_g\widehat{\beta}_{-g}\right) \\ &= \widehat{\beta} - \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}_g'\widehat{\boldsymbol{e}}_{-g},\end{aligned}$$

which is (13) as claimed. By definition (12) and model (1), the prediction errors equal $\widehat{\boldsymbol{e}}_{-g} =$

$\boldsymbol{e}_g - \boldsymbol{X}_g\left(\widehat{\beta}_{-g} - \beta\right)$. Squaring and expanding,

$$\widehat{\boldsymbol{e}}_{-g}\widehat{\boldsymbol{e}}_{-g}' = \boldsymbol{e}_g\boldsymbol{e}_g' - \boldsymbol{e}_g\left(\widehat{\beta}_{-g} - \beta\right)'\boldsymbol{X}_g' - \boldsymbol{X}_g\left(\widehat{\beta}_{-g} - \beta\right)\boldsymbol{e}_g' + \boldsymbol{X}_g\left(\widehat{\beta}_{-g} - \beta\right)\left(\widehat{\beta}_{-g} - \beta\right)'\boldsymbol{X}_g' \tag{40}$$

$$\geq \boldsymbol{e}_g\boldsymbol{e}_g' - \boldsymbol{e}_g\left(\widehat{\beta}_{-g} - \beta\right)'\boldsymbol{X}_g' - \boldsymbol{X}_g\left(\widehat{\beta}_{-g} - \beta\right)\boldsymbol{e}_g'. \tag{41}$$

The inequality holds because the final term in (40) is positive semi-definite.

The first term in (41) has expectation $\boldsymbol{\Sigma}_g$. Observe that $\boldsymbol{e}_g$ is independent of $\widehat{\beta}_{-g} - \beta$ and mean zero, so the expectation of the second and third terms in (41) equals zero. We deduce that

$$\mathbb{E}\left[\widehat{\boldsymbol{e}}_{-g}\widehat{\boldsymbol{e}}_{-g}'\right] \geq \mathbb{E}\left[\boldsymbol{e}_g\boldsymbol{e}_g'\right] = \boldsymbol{\Sigma}_g. \tag{42}$$

Using expression (14) and inequality (42),

$$\begin{aligned}\mathbb{E}\left[\widehat{\boldsymbol{V}}_5\right] &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\sum_{g=1}^{G}\boldsymbol{X}_g'\mathbb{E}\left[\widehat{\boldsymbol{e}}_{-g}\widehat{\boldsymbol{e}}_{-g}'\right]\boldsymbol{X}_g\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \\ &\geq \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\sum_{g=1}^{G}\boldsymbol{X}_g'\boldsymbol{\Sigma}_g\boldsymbol{X}_g\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} = \boldsymbol{V}.\end{aligned}$$

This is the stated result. ■

Proof of Theorem 2, equation (15): Set $\boldsymbol{Z}_g = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}_g R$. Without loss of generality, normalize $\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{I}_k$ and $R'R = 1$. Set $c_g = \boldsymbol{Z}_g'\boldsymbol{Z}_g$ and observe that $\sum_{g=1}^{G} c_g = 1$. Assume that $\boldsymbol{\Sigma}_g = \boldsymbol{I}_{n_g}$. Under these conditions, $v^2 = 1$. In the model class $\mathscr{F}^*$ the value of c_1 is freely varying in (0,1). The CRVE_1 estimator is

$$\widehat{v}_1^2 = d\sum_{g=1}^{G}\boldsymbol{Z}_g'\widehat{\boldsymbol{e}}_g\widehat{\boldsymbol{e}}_g'\boldsymbol{Z}_g \tag{43}$$

where $d = G(n-1)/(G-1)(n-k)$. We can calculate that under these conditions

$$\mathbb{E}\left[\widehat{\boldsymbol{e}}_g\widehat{\boldsymbol{e}}_g'\right] = \boldsymbol{M}_g. \tag{44}$$

Observe that since $\boldsymbol{I}_k - RR'$ is a projection matrix, it is positive semi-definite, and then

$$\boldsymbol{X}_g\boldsymbol{X}_g' = \boldsymbol{Z}_g\boldsymbol{Z}_g' + \boldsymbol{X}_g\left(\boldsymbol{I}_k - RR'\right)\boldsymbol{X}_g' \succeq \boldsymbol{Z}_g\boldsymbol{Z}_g'.$$

It follows that

$$\boldsymbol{M}_g = \boldsymbol{I}_{n_g} - \boldsymbol{X}_g\boldsymbol{X}_g' \preceq \boldsymbol{I}_{n_g} - \boldsymbol{Z}_g\boldsymbol{Z}_g'. \tag{45}$$

Taking expectations of (43) using (44), (45), $\boldsymbol{Z}_g'\boldsymbol{Z}_g = c_g$, and $\sum_{g=1}^G c_g = 1$, we find that

$$\begin{aligned}\frac{\mathbb{E}\left[\widehat{v}_1^2\right]}{v^2} &\leq d\sum_{g=1}^G \boldsymbol{Z}_g'\left(\boldsymbol{I}_{n_g} - \boldsymbol{Z}_g\boldsymbol{Z}_g'\right)\boldsymbol{Z}_g \\ &= d\left(1 - \sum_{g=1}^G c_g^2\right) \\ &\leq d\left(1 - c_1^2\right). \end{aligned} \tag{46}$$

The assumptions are a special case of the model class $\mathscr{F}^*$. Inequality (46) implies that the left side of (15) is weakly smaller than the infimum of the right-hand side of (46) over c_1. Thus

$$\inf_{(\boldsymbol{X},\boldsymbol{\Sigma})\in\mathscr{F}^*} \frac{\mathbb{E}\left[\widehat{v}_1^2\right]}{v^2} \leq d \inf_{0<c_1<1}\left(1 - c_1^2\right) = 0. \tag{47}$$

The rightmost equality in (47) is attained as $c_1 \to 1$. Since the left side of (47) is non-negative, the equation hold as an equality. This verifies (15). ■

Proof of Theorem 2, equation (16): We use the same normalizations and notation as in the proof of (15). We replace the assumption on the covariance matrices with the assumption that $\boldsymbol{\Sigma}_1 = \boldsymbol{I}_{n_1}$ and $\boldsymbol{\Sigma}_g = 0$ for $g > 1$, which is extreme heteroskedasticity. Under these conditions, you can calculate that $\boldsymbol{V} = \boldsymbol{X}_1'\boldsymbol{X}_1$ and $v^2 = c_1$. It will be convenient to define

$$\delta = \max_{2\leq g\leq G} \left\|\boldsymbol{X}_g\right\|^2 \geq \frac{1 - c_1}{G - 1}. \tag{48}$$

The inequality in (48) holds because

$$\delta \geq \sum_{g=2}^G \lambda_{\max}\left(\boldsymbol{X}_g'\boldsymbol{X}_g\right) \geq \lambda_{\max}\left(\boldsymbol{I}_k - \boldsymbol{X}_1'\boldsymbol{X}_1\right) = 1 - \lambda_{\min}\left(\boldsymbol{X}_1'\boldsymbol{X}_1\right) \geq 1 - c_1.$$

The CRVE$_2$ estimator for v^2 can be written as

$$\widehat{v}_2^2 = \sum_{g=1}^G \boldsymbol{Z}_g'\boldsymbol{M}_g^{-1/2}\widehat{\boldsymbol{e}}_g\widehat{\boldsymbol{e}}_g'\boldsymbol{M}_g^{-1/2}\boldsymbol{Z}_g. \tag{49}$$

We can calculate that for $g = 1$,

$$\mathbb{E}\left[\widehat{\boldsymbol{e}}_1\widehat{\boldsymbol{e}}_1'\right] = \boldsymbol{I}_{n_1} - 2\boldsymbol{X}_1\boldsymbol{X}_1' + \boldsymbol{X}_1\boldsymbol{X}_1'\boldsymbol{X}_1\boldsymbol{X}_1' = \boldsymbol{M}_1\boldsymbol{M}_1, \tag{50}$$

and for $g > 1$,

$$\mathbb{E}\left[\widehat{\boldsymbol{e}}_g \widehat{\boldsymbol{e}}_g'\right] = \boldsymbol{X}_g \boldsymbol{X}_1' \boldsymbol{X}_1 \boldsymbol{X}_g'. \tag{51}$$

Taking expectations of (49) and using (50)-(51)

$$\begin{aligned}\mathbb{E}\left[\widehat{v}_2^2\right] &= \sum_{g=1}^{G} \boldsymbol{Z}_g' \boldsymbol{M}_g^{-1/2} \mathbb{E}\left[\widehat{\boldsymbol{e}}_g \widehat{\boldsymbol{e}}_g'\right] \boldsymbol{M}_g^{-1/2} \boldsymbol{Z}_g \\ &= \boldsymbol{Z}_1' \boldsymbol{M}_1^{-1/2} \boldsymbol{M}_1 \boldsymbol{M}_1 \boldsymbol{M}_1^{-1/2} \boldsymbol{Z}_1 + \sum_{g=2}^{G} \boldsymbol{Z}_g' \boldsymbol{M}_g^{-1/2} \boldsymbol{X}_g \boldsymbol{X}_1' \boldsymbol{X}_1 \boldsymbol{X}_g' \boldsymbol{M}_g^{-1/2} \boldsymbol{Z}_g. \end{aligned} \tag{52}$$

The first term on the right side of (52) equals

$$\boldsymbol{Z}_1' \boldsymbol{M}_1 \boldsymbol{Z}_1 \leq \boldsymbol{Z}_1' \left(\boldsymbol{I}_{n_1} - \boldsymbol{Z}_1 \boldsymbol{Z}_1'\right) \boldsymbol{Z}_1 = c_1 - c_1^2, \tag{53}$$

where the inequality is (45). We observe that for $g > 1$

$$\left\| \boldsymbol{M}_g^{-1} \right\| = \frac{1}{\lambda_{\min}\left(\boldsymbol{M}_g\right)} = \frac{1}{1 - \lambda_{\max}\left(\boldsymbol{X}_g' \boldsymbol{X}_g\right)} = \frac{1}{1 - \left\| \boldsymbol{X}_g \right\|^2} \leq \frac{1}{1-\delta}. \tag{54}$$

Using the fact $\boldsymbol{X}_1' \boldsymbol{X}_1 \preceq \boldsymbol{X}' \boldsymbol{X} = \boldsymbol{I}_k$, the quadratic inequality, (48), and (54), the second term on the right side of (52) is smaller than

$$\sum_{g=2}^{G} \boldsymbol{Z}_g' \boldsymbol{M}_g^{-1/2} \boldsymbol{X}_g \boldsymbol{X}_g' \boldsymbol{M}_g^{-1/2} \boldsymbol{Z}_g \leq \frac{\delta}{(1-\delta)} \sum_{g=2}^{G} \boldsymbol{Z}_g' \boldsymbol{Z}_g = \frac{\delta\,(1-c_1)}{(1-\delta)}. \tag{55}$$

Combining (52), (53), (55), and $\nu^2 = c_1$ we find that

$$\frac{\mathbb{E}\left[\widehat{v}_2^2\right]}{\nu^2} \leq \frac{c_1 - c_1^2}{c_1} + \frac{\delta\,(1-c_1)}{c_1\,(1-\delta)}. \tag{56}$$

The assumptions we have made are a special case of the model class $\mathscr{F}^*$. Therefore, the left side of (16) is weakly smaller than the infimum of (56) over (c_1, δ). Hence

$$\begin{aligned}\inf_{(\boldsymbol{X},\boldsymbol{\Sigma})\in\mathscr{F}^*} \frac{\mathbb{E}\left[\widehat{v}_2^2\right]}{\nu^2} &\leq \inf_{c_1,\delta}\left(1 - c_1 + \frac{\delta\,(1-c_1)}{c_1\,(1-\delta)}\right) \\ &= \inf_{c_1}\left(1 - c_1 + \frac{(1-c_1)^2}{c_1\,(G-2+c_1)}\right) = 0. \end{aligned} \tag{57}$$

The first equality in the second line is obtained by setting δ at its lower bound $(1-c_1)/(G-1)$ from (48). The final equality is attained as $c_1 \to 1$. Equation (57) implies (16), as claimed. ■

Proof of Theorem 2, equation (17): To show (17) we calculate an upper bound in the example

used in the proof of (16). We adopt the assumptions made therein. The model class $\mathscr{F}^*$ imposes clusterwise invertibility. As shown by equation (19) of MacKinnon, Nielsen, and Webb (2023b), clusterwise invertibility implies

$$R'\left(\widehat{\beta}-\widehat{\beta}_{-g}\right)=R'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}_g'\boldsymbol{M}_g^{-1}\widehat{\boldsymbol{e}}_g=\boldsymbol{Z}_g'\boldsymbol{M}_g^{-1}\widehat{\boldsymbol{e}}_g. \tag{58}$$

Consequently,

$$\begin{aligned}\widehat{v}_3^2 &= \left(\frac{G-1}{G}\right)\sum_{g=1}^{G} R'\left(\widehat{\beta}_{-g}-\widehat{\beta}\right)\left(\widehat{\beta}_{-g}-\widehat{\beta}\right)' R-(G-1)\,R'\left(\widehat{\beta}-\bar{\beta}\right)\left(\widehat{\beta}-\bar{\beta}\right)' R \\ &= \left(\frac{G-1}{G}\right)^2 \boldsymbol{Z}_1'\boldsymbol{M}_1^{-1}\widehat{\boldsymbol{e}}_1\widehat{\boldsymbol{e}}_1'\boldsymbol{M}_1^{-1}\boldsymbol{Z}_1 && (59)\\ &+\left(\frac{G-1}{G}\right)\sum_{g=2}^{G}\boldsymbol{Z}_g'\boldsymbol{M}_g^{-1}\widehat{\boldsymbol{e}}_g\widehat{\boldsymbol{e}}_g'\boldsymbol{M}_g^{-1}\boldsymbol{Z}_g && (60)\\ &-2\left(\frac{G-1}{G^2}\right)\sum_{g=2}^{G}\boldsymbol{Z}_g'\boldsymbol{M}_g^{-1}\widehat{\boldsymbol{e}}_g\widehat{\boldsymbol{e}}_1'\boldsymbol{M}_1^{-1}\boldsymbol{Z}_1 && (61)\\ &-\left(\frac{G-1}{G^2}\right)\left(\sum_{g=2}^{G}\boldsymbol{Z}_g'\boldsymbol{M}_g^{-1}\widehat{\boldsymbol{e}}_g\right)\left(\sum_{g=2}^{G}\widehat{\boldsymbol{e}}_g'\boldsymbol{M}_g^{-1}\boldsymbol{Z}_g\right). && (62)\end{aligned}$$

Using (50), the expectation of (59) equals $((G-1)/G)^2$ times

$$\boldsymbol{Z}_1'\boldsymbol{M}_1^{-1}\mathbb{E}\left[\widehat{\boldsymbol{e}}_1\widehat{\boldsymbol{e}}_1'\right]\boldsymbol{M}_1^{-1}\boldsymbol{Z}_1=\boldsymbol{Z}_1'\boldsymbol{M}_1^{-1}\boldsymbol{M}_1\boldsymbol{M}_1\boldsymbol{M}_1^{-1}\boldsymbol{Z}_1=\boldsymbol{Z}_1'\boldsymbol{Z}_1=c_1. \tag{63}$$

Using (51) and (54) the expectation of (60) equals $((G-1)/G)$ times

$$\sum_{g=2}^{G}\boldsymbol{Z}_g'\boldsymbol{M}_g^{-1}\mathbb{E}\left[\widehat{\boldsymbol{e}}_g\widehat{\boldsymbol{e}}_g'\right]\boldsymbol{M}_g^{-1}\boldsymbol{Z}_g=\sum_{g=2}^{G}\boldsymbol{Z}_g'\boldsymbol{M}_g^{-1}\boldsymbol{X}_g\boldsymbol{X}_1'\boldsymbol{X}_1\boldsymbol{X}_g'\boldsymbol{M}_g^{-1}\boldsymbol{Z}_g\leq\frac{(1-c_1)\,\delta}{(1-\delta)^2}, \tag{64}$$

where the inequality follows by similar same steps as for (55).

We calculate that for $g\neq 1$

$$\mathbb{E}\left[\widehat{\boldsymbol{e}}_g\widehat{\boldsymbol{e}}_1'\right]=-\boldsymbol{X}_g\left(\boldsymbol{I}_k-\boldsymbol{X}_1'\boldsymbol{X}_1\right)\boldsymbol{X}_1'. \tag{65}$$

Using the Woodbury identity,

$$\begin{aligned}\boldsymbol{X}_1'\boldsymbol{M}_1^{-1}\boldsymbol{X}_1 &= \boldsymbol{X}_1'\boldsymbol{X}_1+\boldsymbol{X}_1'\boldsymbol{X}_1\left(\boldsymbol{I}_k-\boldsymbol{X}_1'\boldsymbol{X}_1\right)^{-1}\boldsymbol{X}_1'\boldsymbol{X}_1 \\ &= -\boldsymbol{I}_k+\left(\boldsymbol{I}_k-\boldsymbol{X}_1'\boldsymbol{X}_1\right)^{-1}\preceq\left(\boldsymbol{I}_k-\boldsymbol{X}_1'\boldsymbol{X}_1\right)^{-1}.\end{aligned} \tag{66}$$

Combining (65), (66), and (54), the expectation of (61) equals $2(G-1)/G^2$ times

$$R'\sum_{g=2}^{G}\boldsymbol{X}_g'\boldsymbol{M}_g^{-1}\boldsymbol{X}_g\left(\boldsymbol{I}_k-\boldsymbol{X}_1'\boldsymbol{X}_1\right)\boldsymbol{X}_1'\boldsymbol{M}_1^{-1}\boldsymbol{X}_1R \leq \sum_{g=2}^{G}R'\boldsymbol{X}_g'\boldsymbol{M}_g^{-1}\boldsymbol{X}_gR$$
$$\leq \frac{1}{(1-\delta)}\sum_{g=2}^{G}\boldsymbol{Z}_g'\boldsymbol{Z}_g = \frac{(1-c_1)}{(1-\delta)}. \tag{67}$$

Together, (59)-(64), (67), $\nu^2=c_1$, and the fact that (62) is non-positive, imply

$$\frac{\mathbb{E}\left[\widehat{\nu}_3^2\right]}{\nu^2} \leq \left(\frac{G-1}{G}\right)^2+\left(\frac{G-1}{G}\right)\frac{(1-c_1)\,\delta}{c_1\,(1-\delta)^2}+2\left(\frac{G-1}{G^2}\right)\frac{(1-c_1)}{c_1\,(1-\delta)}. \tag{68}$$

As in the proof of (16), the left side of (17) is weakly smaller than the infimum of (68) over (c_1,δ). By similar steps as in (57),

$$\inf_{(\boldsymbol{X},\boldsymbol{\Sigma})\in\mathscr{F}^*}\frac{\mathbb{E}\left[\widehat{\nu}_3^2\right]}{\nu^2} \leq \inf_{c_1}\left(\left(\frac{G-1}{G}\right)^2+\frac{(G-1)^2}{G}\frac{(1-c_1)^2}{c_1\,(G-2+c_1)^2}+2\left(\frac{G-1}{G}\right)^2\frac{(1-c_1)}{c_1\,(G-2+c_1)}\right)=\left(\frac{G-1}{G}\right)^2.$$

The final equality is attained as $c_1\to 1$. This is (17), as claimed. ■

Proof of Theorem 2, equation (18): Since the model is clusterwise invertible, $\widehat{\nu}_4^2=\left(\frac{G-1}{G}\right)\widehat{\nu}_5^2$. The result follows from (21), which we establish below. ■

Proof of Theorem 2, equations (19) and (20): We calculate an upper bound for the left side of (19)-(20) assuming a single cluster ($g=1$) is noninvertible and the remaining clusters ($g>1$) are invertible. The noninvertibility is due to the inclusion of a fixed effect dummy variable (only for this single cluster), and it is assumed that R does not load on this variable. Thus Assumption 3 holds. Otherwise, we adopt the assumptions used in the proof of (15), the notation (48), and inequalities (48)-(54).

Because cluster $g=1$ is noninvertible, it is discarded from the calculation of $\widehat{\nu}_3^2$ and $\widehat{\nu}_4^2$. Therefore the latter equals

$$\begin{aligned}\widehat{\nu}_4^2 &= \left(\frac{G-2}{G-1}\right)\sum_{g=2}^{G}R'\left(\widehat{\beta}_{-g}-\widehat{\beta}\right)\left(\widehat{\beta}_{-g}-\widehat{\beta}\right)'R\\ &= \left(\frac{G-2}{G-1}\right)\sum_{g=2}^{G}\boldsymbol{Z}_g'\boldsymbol{M}_g^{-1}\widehat{\boldsymbol{e}}_g\widehat{\boldsymbol{e}}_g'\boldsymbol{M}_g^{-1}\boldsymbol{Z}_g\\ &\leq \sum_{g=2}^{G}\boldsymbol{Z}_g'\boldsymbol{M}_g^{-1}\widehat{\boldsymbol{e}}_g\widehat{\boldsymbol{e}}_g'\boldsymbol{M}_g^{-1}\boldsymbol{Z}_g,\end{aligned}$$

the equality using (58) (which holds for $g\geq 2$ because these are clusterwise invertible), and the

inequality is $(G-2)/(G-1) \leq 1$. Taking expectations, using $\nu^2 = 1$, (44), and then (54)

$$\frac{\mathbb{E}\left[\widehat{\nu}_4^2\right]}{\nu^2} \leq \sum_{g=2}^{G} \boldsymbol{Z}_g' \boldsymbol{M}_g^{-1} \boldsymbol{Z}_g \leq \frac{1}{1-\delta} \sum_{g=2}^{G} \boldsymbol{Z}_g' \boldsymbol{Z}_g = \frac{1-c_1}{1-\delta}. \tag{69}$$

The assumptions we have made are a special case of the model class $\mathscr{F}$. Therefore, the left side of (18) is weakly smaller than the infimum of (69) over (c_1, δ). Hence

$$\inf_{(\boldsymbol{X},\boldsymbol{\Sigma})\in\mathscr{F}} \frac{\mathbb{E}\left[\widehat{\nu}_4^2\right]}{\nu^2} \leq \inf_{c_1,\delta}\left(\frac{1-c_1}{1-\delta}\right) = \inf_{c_1} \frac{(1-c_1)(G-1)}{G-2+c_1} = 0.$$

This implies (18) as claimed. Since $\widehat{\nu}_3^2 \leq \widehat{\nu}_4^2$ we find that (19) holds as well. ■

Proof of Theorem 2, equation (21): Theorem 1 implies

$$\inf_{(\boldsymbol{X},\boldsymbol{\Sigma})\in\mathscr{F}} \frac{\mathbb{E}\left[\widehat{\nu}_5^2\right]}{\nu^2} = \inf_{(\boldsymbol{X},\boldsymbol{\Sigma})\in\mathscr{F}} \frac{R'\mathbb{E}\left[\widehat{\boldsymbol{V}}_5\right]R}{R'\boldsymbol{V}R} \geq \inf_{(\boldsymbol{X},\boldsymbol{\Sigma})\in\mathscr{F}} \frac{R'\boldsymbol{V}R}{R'\boldsymbol{V}R} = 1. \tag{70}$$

To show this is a strict equality we calculate an upper bound for the left side of (70) in the context of the example from the proof of (17). Using the calculations therein, including (63) and (64)

$$\widehat{\nu}_5^2 = \sum_{g=1}^{G} \boldsymbol{Z}_g' \boldsymbol{M}_g^{-1} \widehat{\boldsymbol{e}}_g \widehat{\boldsymbol{e}}_g' \boldsymbol{M}_g^{-1} \boldsymbol{Z}_g.$$

Using (63) and (64), its expectation satisfies $\mathbb{E}\left[\widehat{\nu}_5^2\right]/\nu^2 \leq 1 + (1-c_1)\delta/[(1-\delta)^2 c_1]$. This implies

$$\inf_{(\boldsymbol{X},\boldsymbol{\Sigma})\in\mathscr{F}} \frac{\mathbb{E}\left[\widehat{\nu}_5^2\right]}{\nu^2} \leq \inf_{c_1,\delta}\left[1 + \frac{(1-c_1)\delta}{(1-\delta)^2 c_1}\right] = \inf_{c_1}\left[1 + \frac{(1-c_1)^2(G-1)}{(G-2+c_1)^2 c_1}\right] = 1.$$

Combined with (70) this yields (21) as stated. ■

Proof of Theorem 3: Define the delete-one-cluster operator $\widetilde{\boldsymbol{M}}$, which is the $n \times n$ matrix with $g\,j$th block

$$\widetilde{\boldsymbol{M}}_{gj} = \begin{cases} \boldsymbol{I}_{n_g} & g = j \\ -\boldsymbol{X}_g\left(\boldsymbol{X}'\boldsymbol{X} - \boldsymbol{X}_g'\boldsymbol{X}_g\right)^{-}\boldsymbol{X}_j' & g \neq j. \end{cases}$$

This operator has the algebraic property that it creates delete-one-cluster prediction errors and

is the jackknife analog of the least squares annihilation matrix. Define $\boldsymbol{Z}_g = \boldsymbol{X}_g \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} R$,

$$\bar{\boldsymbol{Z}} = \begin{bmatrix} \boldsymbol{Z}_1 & 0_{n_1\times 1} & \cdots & 0_{n_1\times 1} \\ 0_{n_2\times 1} & \boldsymbol{Z}_2 & \cdots & 0_{n_2\times 1} \\ \vdots & \vdots & \ddots & \vdots \\ 0_{n_G\times 1} & 0_{n_G\times 1} & \cdots & \boldsymbol{Z}_G \end{bmatrix}, \tag{71}$$

and

$$\boldsymbol{B} = \widetilde{\boldsymbol{M}}' \bar{\boldsymbol{Z}}. \tag{72}$$

The matrices $\bar{\boldsymbol{Z}}$ and $\boldsymbol{B}$ are $n \times G$. We now show that under Assumption 3,

$$\boldsymbol{B}'\boldsymbol{X} = 0. \tag{73}$$

The left side of (73) is a $G \times 1$ vector with gth element

$$\begin{aligned} &R' \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{X}_g' \boldsymbol{X}_g \left[\boldsymbol{I}_{n_g} - \left(\boldsymbol{X}'\boldsymbol{X} - \boldsymbol{X}_g'\boldsymbol{X}_g\right)^- \left(\boldsymbol{X}'\boldsymbol{X} - \boldsymbol{X}_g'\boldsymbol{X}_g\right)\right] \\ &= R' \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{X}_g' \boldsymbol{X}_g + R' \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \left(\boldsymbol{X}'\boldsymbol{X} - \boldsymbol{X}_g'\boldsymbol{X}_g\right) \left(\boldsymbol{X}'\boldsymbol{X} - \boldsymbol{X}_g'\boldsymbol{X}_g\right)^- \left(\boldsymbol{X}'\boldsymbol{X} - \boldsymbol{X}_g'\boldsymbol{X}_g\right) \\ &- R' \left(\boldsymbol{X}'\boldsymbol{X} - \boldsymbol{X}_g'\boldsymbol{X}_g\right)^- \left(\boldsymbol{X}'\boldsymbol{X} - \boldsymbol{X}_g'\boldsymbol{X}_g\right) \\ &= R' \left[\boldsymbol{I}_{n_g} - \left(\boldsymbol{X}'\boldsymbol{X} - \boldsymbol{X}_g'\boldsymbol{X}_g\right)^- \left(\boldsymbol{X}'\boldsymbol{X} - \boldsymbol{X}_g'\boldsymbol{X}_g\right)\right]. \end{aligned} \tag{74}$$

The final equality uses the generalized inverse property $\boldsymbol{A}\boldsymbol{A}^-\boldsymbol{A} = \boldsymbol{A}$.

If part (1) of Assumption 3 holds, (74) equals zero. Suppose part (2) holds. As inference concerns the coefficients on $\boldsymbol{X}_1$ only, least squares estimation and inference is unaffected if we replace $\boldsymbol{X}_1$ in $\boldsymbol{X}$ with the residual $\dot{\boldsymbol{X}}_1$. Then

$$\boldsymbol{X}'\boldsymbol{X} - \boldsymbol{X}_g'\boldsymbol{X}_g = \begin{bmatrix} \dot{\boldsymbol{X}}_1'\dot{\boldsymbol{X}}_1 - \dot{\boldsymbol{X}}_{1g}'\dot{\boldsymbol{X}}_{1g} & 0 \\ 0 & \boldsymbol{X}_2'\boldsymbol{X}_2 - \boldsymbol{X}_{2g}'\boldsymbol{X}_{2g} \end{bmatrix}.$$

Since $\dot{\boldsymbol{X}}_1'\dot{\boldsymbol{X}}_1 - \dot{\boldsymbol{X}}_{1g}'\dot{\boldsymbol{X}}_{1g}$ is invertible, (74) equals

$$R' \begin{bmatrix} 0 & 0 \\ 0 & \boldsymbol{I} - \left(\boldsymbol{X}_2'\boldsymbol{X}_2 - \boldsymbol{X}_{2g}'\boldsymbol{X}_{2g}\right)^- \left(\boldsymbol{X}_2'\boldsymbol{X}_2 - \boldsymbol{X}_{2g}'\boldsymbol{X}_{2g}\right) \end{bmatrix} = 0,$$

the final equality because R only loads on $\boldsymbol{X}_1$. Hence under Assumption 3, (73) holds, as claimed.

It will be useful to observe that $\operatorname{rank}(\boldsymbol{B}) \le \max[G, n-k]$. This holds because, firstly, $\operatorname{rank}(\boldsymbol{B}) \le G$ as $\boldsymbol{B}$ is $n \times G$, and secondly, because $\operatorname{rank}(\boldsymbol{B}) \le n-k$ as (73) shows that $G \times n$ $\boldsymbol{B}'$ has null space

$\boldsymbol{X}$, which is of dimension k.

We next write the coefficient estimator and squared standard error $\widehat{v}_5^2$ as explicit linear/quadratic functions of the regression error vector $\boldsymbol{e}$. First,

$$\widehat{\theta} - \theta = R' \left(\boldsymbol{X}'\boldsymbol{X} \right)^{-1} \boldsymbol{X}'\boldsymbol{e} = \boldsymbol{Z}'\boldsymbol{e} \tag{75}$$

where $\boldsymbol{Z} = \boldsymbol{X} \left(\boldsymbol{X}'\boldsymbol{X} \right)^{-1} R$ is the vector of stacked $\boldsymbol{Z}_g$. Second, let $\widehat{\boldsymbol{e}}_-$ denote the stacked prediction errors $\widehat{\boldsymbol{e}}_{-g}$, which satisfies $\widehat{\boldsymbol{e}}_- = \widetilde{\boldsymbol{M}}\boldsymbol{Y}$. Using (14), the definitions of $\boldsymbol{Z}_g$ and $\bar{\boldsymbol{Z}}$, the equation $\widehat{\boldsymbol{e}}_- = \widetilde{\boldsymbol{M}}\boldsymbol{Y}$, (72), $\boldsymbol{Y} = \boldsymbol{X}\beta + \boldsymbol{e}$, and finally (73), we can write

$$\begin{aligned}
\widehat{v}_5^2 &= R' \left(\boldsymbol{X}'\boldsymbol{X} \right)^{-1} \left(\sum_{g=1}^{G} \boldsymbol{X}_g' \widehat{\boldsymbol{e}}_{-g} \widehat{\boldsymbol{e}}_{-g}' \boldsymbol{X}_g \right) \left(\boldsymbol{X}'\boldsymbol{X} \right)^{-1} R \\
&= \sum_{g=1}^{G} \boldsymbol{Z}_g' \widehat{\boldsymbol{e}}_{-g} \widehat{\boldsymbol{e}}_{-g}' \boldsymbol{Z}_g \qquad (76) \\
&= \widehat{\boldsymbol{e}}_-' \bar{\boldsymbol{Z}} \bar{\boldsymbol{Z}}' \widehat{\boldsymbol{e}}_- \\
&= \boldsymbol{Y}' \widetilde{\boldsymbol{M}}' \bar{\boldsymbol{Z}} \bar{\boldsymbol{Z}}' \widetilde{\boldsymbol{M}} \boldsymbol{Y} \\
&= \left(\boldsymbol{X}\beta + \boldsymbol{e} \right)' \boldsymbol{B}\boldsymbol{B}' \left(\boldsymbol{X}\beta + \boldsymbol{e} \right) \\
&= \boldsymbol{e}'\boldsymbol{B}\boldsymbol{B}'\boldsymbol{e}. \qquad (77)
\end{aligned}$$

It will be useful to observe that Theorem 1 and (77) imply that

$$v^2 \leq \mathbb{E}\left[\widehat{v}_5^2 \right] = \mathbb{E}\left[\boldsymbol{e}'\boldsymbol{B}\boldsymbol{B}'\boldsymbol{e} \right] = \mathbb{E}\left[\operatorname{tr}\left(\boldsymbol{B}\boldsymbol{B}'\boldsymbol{e}\boldsymbol{e}' \right) \right] = \operatorname{tr}\left(\boldsymbol{B}\boldsymbol{B}'\boldsymbol{\Sigma} \right). \tag{78}$$

Since $\boldsymbol{e} \sim \mathrm{N}(0, \boldsymbol{\Sigma})$ we can write $\boldsymbol{e} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{\psi}$ with $\boldsymbol{\psi} \sim \mathrm{N}(0, \boldsymbol{I}_n)$. Using $\widehat{C}_5(c) = \widehat{\theta} \pm \widehat{v}_5 c$, (75), (77), and $\boldsymbol{e} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{\psi}$, we find

$$\begin{aligned}
\mathbb{P}\left[\theta \in \widehat{C}_5(c) \right] &= \mathbb{P}\left[\left(\widehat{\theta} - \theta \right)^2 \leq c^2 \widehat{v}_5^2 \right] \\
&= \mathbb{P}\left[\boldsymbol{e}'\boldsymbol{Z}\boldsymbol{Z}'\boldsymbol{e} \leq c^2 \boldsymbol{e}'\boldsymbol{B}\boldsymbol{B}'\boldsymbol{e} \right] \\
&= \mathbb{P}\left[0 \leq \boldsymbol{e}' \left(c^2 \boldsymbol{B}\boldsymbol{B}' - \boldsymbol{Z}\boldsymbol{Z}' \right) \boldsymbol{e} \right] \\
&= \mathbb{P}\left[0 \leq \boldsymbol{\psi}'\boldsymbol{C}\boldsymbol{\psi} \right] \qquad (79)
\end{aligned}$$

where

$$\boldsymbol{C} = \boldsymbol{\Sigma}^{1/2} \left(c^2 \boldsymbol{B}\boldsymbol{B}' - \boldsymbol{Z}\boldsymbol{Z}' \right) \boldsymbol{\Sigma}^{1/2}. \tag{80}$$

By the spectral decomposition, $\boldsymbol{C} = \boldsymbol{H}\boldsymbol{\Lambda}\boldsymbol{H}'$ where $\boldsymbol{\Lambda} = \operatorname{diag}\{\lambda_1, ..., \lambda_n\}$, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ are the eigenvalues of $\boldsymbol{C}$, and $\boldsymbol{H}'\boldsymbol{H} = \boldsymbol{I}_n$. Define $\boldsymbol{\xi} = \boldsymbol{H}'\boldsymbol{\psi} \sim \mathrm{N}(0, \boldsymbol{I}_n)$ and make the partition $\boldsymbol{\xi} =$

$(\xi_1, ..., \xi_n)'$. Set $\lambda_0 = -\lambda_n$ and $\xi_0^2 = \xi_n^2$. We find that (79) equals

$$\mathbb{P}\left[0 \leq \boldsymbol{\psi}' \boldsymbol{H} \boldsymbol{\Lambda} \boldsymbol{H}' \boldsymbol{\psi}\right] = \mathbb{P}\left[0 \leq \boldsymbol{\xi}' \boldsymbol{\Lambda} \boldsymbol{\xi}\right] = \mathbb{P}\left[\lambda_0 \xi_0^2 \leq \sum_{j=1}^{n-1} \lambda_j \xi_j^2\right]. \tag{81}$$

Next, we establish the following properties of the eigenvalues:

$$\lambda_j \geq 0 \text{ for } j < n \tag{82}$$

$$\lambda_0 \geq 0 \tag{83}$$

$$\lambda_0 \leq \nu^2 \tag{84}$$

$$\bar{\lambda} \stackrel{\text{def}}{=} \sum_{j=1}^{n-1} \lambda_j \geq c^2 \lambda_0 \tag{85}$$

with (85) holding for $c \geq 1$.

Let $\lambda_j(\boldsymbol{D})$ denote the jth largest eigenvalue of a Hermitian matrix $\boldsymbol{D}$. By a corollary of the Weyl eigenvalue inequality for Hermitian matrices (Corollary 4.3.15 of Horn and Johnson (2013)), for any Hermitian matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, and any j,

$$\lambda_j(\boldsymbol{A}) + \lambda_{\min}(\boldsymbol{B}) \leq \lambda_j(\boldsymbol{A} + \boldsymbol{B}) \leq \lambda_{\max}(\boldsymbol{A}) + \lambda_j(\boldsymbol{B}). \tag{86}$$

Using the definition (80) and lower bound in (86), for each $1 \leq j < n$,

$$\begin{aligned}
\lambda_j &\geq \lambda_{\min}\left(c^2 \boldsymbol{\Sigma}^{1/2} \boldsymbol{B}\boldsymbol{B}' \boldsymbol{\Sigma}^{1/2}\right) + \lambda_j\left(-\boldsymbol{\Sigma}^{1/2} \boldsymbol{Z}\boldsymbol{Z}' \boldsymbol{\Sigma}^{1/2}\right) \\
&= c^2 \lambda_{\min}\left(\boldsymbol{\Sigma}^{1/2} \boldsymbol{B}\boldsymbol{B}' \boldsymbol{\Sigma}^{1/2}\right) - \lambda_{n-j+1}\left(\boldsymbol{\Sigma}^{1/2} \boldsymbol{Z}\boldsymbol{Z}' \boldsymbol{\Sigma}^{1/2}\right) = 0,
\end{aligned}$$

the final equality since $\boldsymbol{\Sigma}^{1/2} \boldsymbol{B}\boldsymbol{B}' \boldsymbol{\Sigma}^{1/2}$ and $\boldsymbol{\Sigma}^{1/2} \boldsymbol{Z}\boldsymbol{Z}' \boldsymbol{\Sigma}^{1/2}$ are positive semi-definite with deficient rank. (Indeed, $\text{rank}\left(\boldsymbol{\Sigma}^{1/2} \boldsymbol{B}\boldsymbol{B}' \boldsymbol{\Sigma}^{1/2}\right) \leq \text{rank}(\boldsymbol{B}) \leq \max[G, n-k] < n$ and $\text{rank}\left(\boldsymbol{\Sigma}^{1/2} \boldsymbol{Z}\boldsymbol{Z}' \boldsymbol{\Sigma}^{1/2}\right) = 1$.) This is (82).

Using the lower bound in (86) and similar reasoning,

$$\begin{aligned}
\lambda_0 &= \lambda_{\max}(-\boldsymbol{C}) \\
&= \lambda_{\max}\left(-c^2 \boldsymbol{\Sigma}^{1/2} \boldsymbol{B}\boldsymbol{B}' \boldsymbol{\Sigma}^{1/2} + \boldsymbol{\Sigma}^{1/2} \boldsymbol{Z}\boldsymbol{Z}' \boldsymbol{\Sigma}^{1/2}\right) \\
&\geq c^2 \lambda_{\max}\left(-\boldsymbol{\Sigma}^{1/2} \boldsymbol{B}\boldsymbol{B}' \boldsymbol{\Sigma}^{1/2}\right) + \lambda_{\min}\left(\boldsymbol{\Sigma}^{1/2} \boldsymbol{Z}\boldsymbol{Z}' \boldsymbol{\Sigma}^{1/2}\right) \\
&= -c^2 \lambda_{\min}\left(\boldsymbol{\Sigma}^{1/2} \boldsymbol{B}\boldsymbol{B}' \boldsymbol{\Sigma}^{1/2}\right) + \lambda_{\min}\left(\boldsymbol{\Sigma}^{1/2} \boldsymbol{Z}\boldsymbol{Z}' \boldsymbol{\Sigma}^{1/2}\right) = 0.
\end{aligned}$$

This is (83).

Using the upper bound in (86),

$$\begin{aligned}\lambda_0 &= \lambda_{\max}\left(-c^2\boldsymbol{\Sigma}^{1/2}\boldsymbol{B}\boldsymbol{B}'\boldsymbol{\Sigma}^{1/2} + \boldsymbol{\Sigma}^{1/2}\boldsymbol{Z}\boldsymbol{Z}'\boldsymbol{\Sigma}^{1/2}\right)\\ &\leq \lambda_{\max}\left(-c^2\boldsymbol{\Sigma}^{1/2}\boldsymbol{B}\boldsymbol{B}'\boldsymbol{\Sigma}^{1/2}\right) + \lambda_{\max}\left(\boldsymbol{\Sigma}^{1/2}\boldsymbol{Z}\boldsymbol{Z}'\boldsymbol{\Sigma}^{1/2}\right)\\ &= -c^2\lambda_{\min}\left(\boldsymbol{\Sigma}^{1/2}\boldsymbol{B}\boldsymbol{B}'\boldsymbol{\Sigma}^{1/2}\right) + \boldsymbol{Z}'\boldsymbol{\Sigma}\boldsymbol{Z}\\ &= \nu^2,\end{aligned}$$

since $\boldsymbol{Z}'\boldsymbol{\Sigma}\boldsymbol{Z} = \nu^2$. This is (84).

Using the fact that the trace of a matrix equals the sum of its eigenvalues, $\lambda_n = -\lambda_0$, (80), (78), and $\boldsymbol{Z}'\boldsymbol{\Sigma}\boldsymbol{Z} = \nu^2$,

$$\begin{aligned}\bar{\lambda} &= \operatorname{tr}(\boldsymbol{C}) + \lambda_0\\ &= c^2\operatorname{tr}\left(\boldsymbol{\Sigma}^{1/2}\boldsymbol{B}\boldsymbol{B}'\boldsymbol{\Sigma}^{1/2}\right) - \operatorname{tr}\left(\boldsymbol{\Sigma}^{1/2}\boldsymbol{Z}\boldsymbol{Z}'\boldsymbol{\Sigma}^{1/2}\right) + \lambda_0\\ &= c^2\operatorname{tr}\left(\boldsymbol{B}\boldsymbol{B}'\boldsymbol{\Sigma}\right) - \boldsymbol{Z}'\boldsymbol{\Sigma}\boldsymbol{Z} + \lambda_0\\ &\geq c^2\nu^2 - \nu^2 + \lambda_0\\ &= c^2\lambda_0 + (\nu^2 - \lambda_0)(c^2 - 1) \geq c^2\lambda_0,\end{aligned}$$

The final inequality holds by (84) and $c^2 \geq 1$. This is (85). We have established (82)-(85) as stated.

From (79) and (81) we have

$$\mathbb{P}\left[\theta \in \widehat{C}_5(c)\right] = \mathbb{P}\left[\lambda_0\xi_0^2 \leq \sum_{j=1}^{n-1}\lambda_j\xi_j^2\right]. \tag{87}$$

From (82)-(83), $\lambda_j \geq 0$ for $j = 0, ..., n-1$. Suppose $\lambda_0 = 0$. Then (87) equals 1, which satisfies (24). Now suppose $\lambda_0 > 0$. For $j \geq 1$ define $w_j = \lambda_j/\bar{\lambda}$, which satisfy $\sum_{j=1}^{n-1} w_j = 1$ and $w_j \geq 0$. Then (87) equals

$$\mathbb{P}\left[\frac{\xi_0^2}{\sum_{j=1}^{n-1} w_j\xi_j^2} \leq \frac{\bar{\lambda}}{\lambda_0}\right] \geq \mathbb{P}\left[\zeta^2 \leq \frac{\bar{\lambda}}{\lambda_0}\right] \geq \mathbb{P}\left[\zeta^2 \leq c^2\right] = \mathbb{P}\left[|\zeta| \leq c\right].$$

The first inequality is the right-side inequality of Corollary 2 of Makshanov and Shalaevski (1986). The second inequality is (85). This satisfies (24), and we have shown the latter holds whether $\lambda_0 = 0$ or $\lambda_0 > 0$, completing the proof. ■

The proofs of Theorems 4-7 are in the Online Appendix

References

[1] Andrews, Donald W. K. (1991): "Asymptotic normality of series estimators for nonparametric and semiparametric regression models," *Econometrica*, 59, 307-345.

[2] Arellano, Manuel (1987): "Computing robust standard errors for within groups estimators," *Oxford Bulletin of Economics and Statistics* 49, 431-434.

[3] Bahadur, R. R. and Leonard J. Savage (1956): "The nonexistence of certain statistical procedures in nonparametric problems," *Annals of Mathematical Statistics*, 27, 1115-1122.

[4] Bell, Robert M., and Daniel F. McCaffrey (2002): "Bias reduction in standard errors for linear regression with multi-stage samples," *Survey Methodology*, 28, 169-181.

[5] Bera, Anil K., Totok Suprayitno, and Gamini Premaratne (2002): "On some heteroskedasticity-robust estimators of variance-covariance matrix of the least-squares estimators," *Journal of Statistical Planning and Inference*, 108, 121-136.

[6] Bertanha, Marinho and Marcelo J. Moreira (2020): "Impossible inference in econometrics: Theory and applications," *Journal of Econometrics*, 218, 247-270.

[7] Bodenham, Dean A. and Niall M. Adams (2016): "A comparison of efficient approximations for a weighted sum of chi-squared random variables," *Statistics and Computing*, 26, 917-928.

[8] Cameron, A. Colin, Jonah B. Gelbach, and Douglas L. Miller (2008): "Bootstrap-based improvements for inference with clustered errors," *Review of Economics and Statistics*, 90, 414-427.

[9] Canay, Ivan A., Andres Santos, and Azeem M. Shaikh (2021): "The wild bootstrap with a small number of large clusters," *Review of Economics and Statistics*, 103, 346-363.

[10] Cattaneo, Matias D., Michael Jansson, and Whitney K. Newey (2018): "Inference in linear regression models with many covariates and heteroskedasticity," *Journal of the American Statistical Association*, 113, 1350-1361.

[11] Chesher, Andrew D. (1989): "Hájek inequalities, measures of leverage, and the size of heteroskedasticity robust Wald tests," *Econometrica*, 57, 971-977.

[12] Chesher, Andrew D. and Gerard Austin (1991): "The finite-sample distributions of heteroskedasticity robust Wald statistics," *Journal of Econometrics*, 47, 153-173.

[13] Chesher, Andrew D. and Ian D. Jewitt (1987): "The bias of the heteroskedasticity consistent covariance matrix estimator," *Econometrica*, 55, 1217-1272.

[14] Cochran, William G. (1977): *Sampling Techniques*, 3rd Edition, Wiley.

[15] Conley, Timothy G. and Christopher R. Taber (2011): "Inference with 'difference in differences' with a small number of policy changes," *Review of Economics and Statistics*, 93, 113-125.

[16] Coudin, Elise and Jean-Marie Dufour (2009): "Finite-sample distribution-free inference in linear median regressions under heteroskedasticity and nonlinear dependence of unknown form," *Econometrics Journal*, 12, S19-S49.

[17] Davidson, Russell, and James G. MacKinnon (1993): *Estimation and Inference in Econometrics*, Oxford University Press.

[18] Djogbenou, Antoine. A., James G. MacKinnon, and Morten Ørregaard Nielsen (2019): "Asymptotic theory and wild bootstrap inference with clustered errors," *Journal of Econometrics*, 212, 393-412.

[19] Dufour, Jean-Marie (1997): "Some impossibility theorems in econometrics, with applications to structural and dynamic models," *Econometrica*, 65, 1365-1389.

[20] Dufour, Jean-Marie (2003): "Identification, weak instruments and statistical inference in econometrics," *Canadian Journal of Economics*, 36, 767-808.

[21] Efron, Bradley (1982): *The Jackknife, the Bootstrap, and Other Resampling Plans*, Society for Industrial and Applied Mathematics.

[22] Efron, Bradley, and Charles Stein (1981): "The jackknife estimate of variance," *The Annals of Statistics*, 9, 586-596.

[23] Eicker, Friedhelm (1963): "Asymptotic normality and consistency of the least squares estimators for families of linear regressions," *Annals of Mathematical Statistics*, 34, 447-456.

[24] Ferman, Bruno and Cristine Pinto (2019): "Inference in differences-in-differences with few treated groups and heteroskedasticity," *Review of Economics and Statistics*, 101, 452-467.

[25] Hagemann, Andreas (2019): "Placebo inference on treatment effects when the number of clusters is small," *Journal of Econometrics*, 213, 190-209.

[26] Hagemann, Andreas (2025): "Inference with a single treated cluster," *Review of Economic Studies*, forthcoming.

[27] Hansen, Bruce E. (2025): "Standard errors for difference-in-difference regression," *Journal of Applied Econometrics*, 40, 291-309.

[28] Hansen, Bruce E. and Seojeong Lee (2019): "Asymptotic theory for clustered samples," *Journal of Econometrics*, 210, 268-290.

[29] Hinkley, David V. (1977): "Jackknifing in unbalanced situations," *Technometrics*, 19, 285-292.

[30] Hirano, Kei and Jack R. Porter (2012): "Impossibility results for nondifferentiable functionals," *Econometrica*, 80, 1769-1790.

[31] Horn, Roger A. and Charles R. Johnson (2013): *Matrix Analysis*, Second Edition, Cambridge University Press.

[32] Huber, Peter J. (1967): "The behavior of maximum likelihood estimates under nonstandard conditions," *Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability*, Lucien M. Le Cam and Jerzy Neyman, editors, 1, 221-223.

[33] Ibragimov, Rustam and Ulrich K. Müller (2010): "t-statistic based correlation and heterogeneity robust inference," *Journal of Business and Economic Statistics*, 28, 453-468.

[34] Ibragimov, Rustam and Ulrich K. Müller (2016): "Inference with a few heterogeneous clusters," *Review of Economics and Statistics*, 98, 83-96.

[35] Imbens, Guido W. and Michal Kolesár (2016): "Robust standard errors in small samples: Some practical advice," *Review of Economics and Statistics*, 98, 701-712.

[36] Khuri, Andre I. (1995): "A measure to evaluate the closeness of Satterthwaite's approximation," *Biometrical Journal*, 37, 547-563.

[37] Kline, Patrick, Raffaele Saggio, and Mikkel Sølvsten (2020): "Leave-out estimation of variance components," *Econometrica*, 88, 1859-1898.

[38] Kolesár, Michal (2023): "Robust standard errors in small samples," unpublished R vignette.

[39] Kranz, Sebastian (2024): "From replications to revelations: Heteroskedasticity-robust inference," unpublished manuscript.

[40] Liang, Kung-Yee, and Scott L. Zeger (1986): "Longitudinal Data Analysis Using Generalized Linear Models," *Biometrika*, 73, 13-22.

[41] Long, J. Scott, and Laurie H. Ervin (2000): "Using heteroscedasticity consistent standard errors in the linear regression model," *The American Statistician*, 54, 217-224.

[42] MacKinnon, James G., and Matthew D. Webb (2020): "Randomization inference for difference-in-differences with few treated clusters," *Journal of Econometrics*, 218, 435-450.

[43] MacKinnon, James G. and Halbert White (1985): "Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties," *Journal of Econometrics*, 29, 305-325.

[44] MacKinnon, James G., Morten Ørregaard Nielsen, and Matthew D. Webb (2023a): "Cluster-robust inference: A guide to empirical practice," *Journal of Econometrics*, 232, 272-299.

[45] MacKinnon, James G., Morten Ørregaard Nielsen, and Matthew D. Webb (2023b): "Fast and reliable jackknife and bootstrap methods for cluster-robust inference," *Journal of Applied Econometrics*, 38, 671-694.

[46] MacKinnon, James G., Morten Ørregaard Nielsen, and Matthew D. Webb (2023c): "Leverage, influence, and the jackknife in clustered regression models: Reliable inference using summclust," *Stata Journal*, 4, 942-982.

[47] Makshanov, A. V. and O. V. Shalaevski (1986): "Some problems of asymptotic approximation of distributions," *Journal of Mathematical Sciences*, 34, 1433-1445.

[48] Meng, Xin, Nancy Qian, and Pierre Yared (2015): "The institutional causes of China's Great Famine, 1959-1961," *Review of Economic Studies*, 82, 1568-1611.

[49] Niccodemi, Gianmaria and Tom Wansbeek (2022): "A new estimator for standard errors with a few unbalanced clusters," *Econometrics*, 10, 6.

[50] Pötscher, Benedikt M. and David Preinerstorfer (2025): "Valid heteroskedasticity robust testing," *Econometric Theory*, 41, 249-301.

[51] Preinerstorfer, David and Benedikt M. Pötscher (2016): "On size and power of heteroskedasticity and autocorrelation robust tests," *Econometric Theory*, 32, 261-358.

[52] Pustejovsky, James E. and Elizabeth Tipton (2018): "Small sample methods for cluster-robust variance estimation and hypothesis testing in fixed effects models," *Journal of Business and Economic Statistics*, 36, 672-683.

[53] Rokicki, Slawa, Jessica Cohen, Günther Fink, Joshua A. Salomon, and Mary Beth Landrum (2018): "Inference with difference-in-differences with a small number of groups: A review, simulation study, and empirical application using SHARE data," *Medical Care*, 56, 97-105.

[54] Romano, Joseph P. (2004): "On non-parametric testing, the uniform behavior of the t-test, and related problems," *Scandinavian Journal of Statistics*, 31, 567-584.

[55] Rust, Keith F. and J. N. K. Rao (1996): "Variance estimation for complex surveys using replication techniques," *Statistical Methods in Medical Research*, 5, 283-310.

[56] Satterthwaite, F. E. (1946): "An approximate distribution of estimates of variance components," *Biometrics Bulletin*, 2, 110-114.

[57] Shao, Jun and Dongsheng Tu (1995): *The Jackknife and Bootstrap*, Springer.

[58] Tukey, John (1958): "Bias and confidence in not quite large samples," *Annals of Mathematical Statistics*, 29, 614.

[59] White, Halbert (1980): "A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity," *Econometrica*, 48, 817-838.

[60] Young, Alwyn (2016): "Improved, nearly exact, statistical inference with robust and clustered covariance matrices using effective degrees of freedom corrections," unpublished manuscript, London School of Economics.

[61] Young, Alwyn (2019): "Channeling Fisher: Randomization tests and the statistical insignificance of seemingly significant experimental results," *Quarterly Journal of Economics*, 134, 557-598.

Jackknife Standard Errors for Clustered Regression Online Appendix

Bruce E. Hansen

September 2025

This appendix provides extra material on three topics. First, we provide proofs of Theorems 4-7 and equations (38)-(39) from the main body. Second, we provide an asymptotic theory which shows that statistics constructed with the cluster jackknife variance estimator are asymptotic normal. Third, we provide details concerning our implementation of the wild bootstrap in the numerical simulation.

Proofs of Results from the Main Text

Proof of Theorem 4: Set $\varepsilon > 0$. Let $Q = \left(\widehat{\theta} - \theta\right)^2 / v^2 \sim \chi^2_1$. Let q be the εth quantile of the χ^2_1 distribution and set $\eta = q/c^2$. Define the events $A = \left\{\widehat{v}^2/v^2 \leq \eta\right\}$ and $B = \left\{\left(\widehat{\theta} - \theta\right)^2 / \widehat{v}^2 \leq c^2\right\}$. They jointly imply the event $\left\{Q \leq c^2\eta\right\}$. Thus

$$\mathbb{P}\left[B \cap A\right] \leq \mathbb{P}\left[Q \leq c^2\eta\right] = \varepsilon. \tag{88}$$

Pick $(\boldsymbol{X}, \boldsymbol{\Sigma}) \in \mathscr{F}_a$ so that

$$\frac{\mathbb{E}\left[\widehat{v}^2\right]}{v^2} \leq \eta\varepsilon, \tag{89}$$

which is feasible by (25). By Markov's inequality and (89),

$$\mathbb{P}\left[B \cap A^c\right] \leq \mathbb{P}\left[A^c\right] = \mathbb{P}\left[\frac{\widehat{v}^2}{v^2} > \eta\right] \leq \frac{\mathbb{E}\left[\widehat{v}^2\right]}{\eta v^2} \leq \varepsilon. \tag{90}$$

Equations (88) and (90) imply that

$$\mathbb{P}\left[\theta \in \widehat{C}(c)\right] = \mathbb{P}\left[B\right] = \mathbb{P}\left[B \cap A\right] + \mathbb{P}\left[B \cap A^c\right] \leq 2\varepsilon.$$

As ε is arbitrary this establishes the stated result. ■

Proof of Theorem 5: Results (26), (27), (28), and (29) follow from Theorem 2, equations (15), (16), (19), and (20), combined with Theorem 4. ■

Proof of Theorem 6: Set $\xi_0 = (\widehat{\theta} - \theta)/v \sim \mathrm{N}(0,1)$. As in the proof of Theorem 3, write $\boldsymbol{e} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{\psi}$ with $\boldsymbol{\psi} \sim \mathrm{N}(0, \boldsymbol{I}_n)$. Using this and (77), we find $\widehat{v}_5^2/v^2 = \boldsymbol{\psi}'\boldsymbol{A}\boldsymbol{\psi}$ where $\boldsymbol{A} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{B}\boldsymbol{B}'\boldsymbol{\Sigma}^{1/2}/v^2$. The matrix $\boldsymbol{A}$ has rank at most G, and its non-zero eigenvalues are the same as those of $\bar{\boldsymbol{A}} = \boldsymbol{B}'\boldsymbol{\Sigma}\boldsymbol{B}/v^2$. Make the spectral decomposition $\boldsymbol{A} = \boldsymbol{H}_1\boldsymbol{\Lambda}\boldsymbol{H}_1'$ where $\boldsymbol{H}_1$ is $n \times G$, satisfies $\boldsymbol{H}_1'\boldsymbol{H}_1 = \boldsymbol{I}_G$, and $\boldsymbol{\Lambda}$ has the eigenvalues of $\bar{\boldsymbol{A}}$. Set $\boldsymbol{\xi} = \boldsymbol{H}_1'\boldsymbol{\psi} \sim \mathrm{N}(0, \boldsymbol{I}_G)$. We find

$$\frac{\widehat{\theta} - \theta}{\widehat{v}_5} = \frac{(\widehat{\theta} - \theta)/v}{\sqrt{\widehat{v}_5^2/v^2}} = \frac{\xi_0}{\sqrt{\sum_{g=1}^G \lambda_g \xi_g^2}}.$$

This is (32). The proof is completed by demonstrating that $\bar{\boldsymbol{A}} = \boldsymbol{D}$ as defined in (33).

Recalling definitions (71)-(72), with a little algebra we can write

$$\boldsymbol{B} = \begin{bmatrix} \boldsymbol{Z}_1 & -\boldsymbol{X}_1\boldsymbol{U}_2 & \cdots & -\boldsymbol{X}_1\boldsymbol{U}_G \\ -\boldsymbol{X}_2\boldsymbol{U}_1 & \boldsymbol{Z}_2 & \cdots & -\boldsymbol{X}_2\boldsymbol{U}_G \\ \vdots & \vdots & \ddots & \vdots \\ -\boldsymbol{X}_G\boldsymbol{U}_1 & -\boldsymbol{X}_G\boldsymbol{U}_2 & \cdots & \boldsymbol{Z}_G \end{bmatrix} = \bar{\boldsymbol{T}} - \boldsymbol{X}\boldsymbol{U}',$$

where

$$\bar{\boldsymbol{T}} = \begin{bmatrix} \boldsymbol{T}_1 & 0_{n_1 \times 1} & \cdots & 0_{n_1 \times 1} \\ 0_{n_2 \times 1} & \boldsymbol{T}_2 & \cdots & 0_{n_2 \times 1} \\ \vdots & \vdots & \ddots & \vdots \\ 0_{n_G \times 1} & 0_{n_G \times 1} & \cdots & \boldsymbol{T}_G \end{bmatrix}.$$

Then

$$\begin{aligned} \bar{\boldsymbol{A}} &= \left(\bar{\boldsymbol{T}}' - \boldsymbol{U}\boldsymbol{X}'\right)\boldsymbol{\Sigma}\left(\bar{\boldsymbol{T}} - \boldsymbol{X}\boldsymbol{U}'\right)/v^2 \\ &= \left(\bar{\boldsymbol{T}}'\boldsymbol{\Sigma}\bar{\boldsymbol{T}} + \boldsymbol{U}\boldsymbol{X}'\boldsymbol{\Sigma}\boldsymbol{X}\boldsymbol{U}' - \boldsymbol{U}\boldsymbol{X}'\boldsymbol{\Sigma}\bar{\boldsymbol{T}} - \bar{\boldsymbol{T}}'\boldsymbol{\Sigma}\boldsymbol{X}\boldsymbol{U}'\right)/v^2 \\ &= \left(\boldsymbol{S} + \boldsymbol{U}\boldsymbol{X}'\boldsymbol{\Sigma}\boldsymbol{X}\boldsymbol{U}' - \boldsymbol{U}\boldsymbol{V}' - \boldsymbol{V}\boldsymbol{U}'\right)/v^2 = \boldsymbol{D} \end{aligned} \tag{91}$$

as claimed. The third equality holds because of the following relationships

$$\begin{aligned} \bar{\boldsymbol{T}}'\boldsymbol{\Sigma}\bar{\boldsymbol{T}} &= \mathrm{diag}\left\{\boldsymbol{T}_g'\boldsymbol{\Sigma}_g\boldsymbol{T}_g\right\} = \mathrm{diag}\left\{S_g\right\} = \boldsymbol{S} \\ \bar{\boldsymbol{T}}'\boldsymbol{\Sigma}\boldsymbol{X} &= \boldsymbol{V}. \end{aligned}$$

This completes the derivation. ■

Proof of Theorem 7: Using the Schwarz matrix inequality and the assumptions,

$$\max_{g\leq G}\left\|\boldsymbol{I}_{n_g}-\boldsymbol{M}_g\right\| = \max_{g\leq G}\left\|\boldsymbol{X}_g\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}_g'\right\| \leq \left\|\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\right\| \frac{1}{n}\max_{g\leq G}\left\|\boldsymbol{X}_g\right\|^2 \leq o(1). \tag{92}$$

As $\boldsymbol{I}_{n_g}-\boldsymbol{M}_g$ is positive semi-definite,

$$\max_{g\leq G}\left\|\boldsymbol{I}_{n_g}-\boldsymbol{M}_g\right\| = \max_{g\leq G}\lambda_{\max}\left(\boldsymbol{I}_{n_g}-\boldsymbol{M}_g\right) = 1-\min_{g\leq G}\lambda_{\min}(\boldsymbol{M}_g).$$

Equation (92) implies that $\min_{g\leq G}\lambda_{\min}(\boldsymbol{M}_g)\to 1$. As $\boldsymbol{M}_g$ is positive semi-definite, this implies that (at least in sufficiently large samples) $\boldsymbol{M}_g$ are uniformly invertible. Together, this implies

$$\max_{g\leq G}\left\|\boldsymbol{M}_g^{-1}-\boldsymbol{I}_{n_g}\right\| = \max_{g\leq G}\lambda_{\max}(\boldsymbol{M}_g^{-1}-\boldsymbol{I}_{n_g}) = \max_{g\leq G}\lambda_{\max}(\boldsymbol{M}_g^{-1})-1 = \frac{1}{\min_{g\leq G}\lambda_{\min}(\boldsymbol{M}_g)}-1\to 0. \tag{93}$$

Under $\boldsymbol{\Sigma}_g=\boldsymbol{I}_{n_g}$,

$$a^2=\sum_{g=1}^{G}\lambda_g=\operatorname{tr}[\boldsymbol{D}]=\frac{\mathbb{E}\left[\widehat{v}^2\mid\boldsymbol{X}\right]}{v^2}$$

and

$$v^2=R'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}R=\sum_{g=1}^{G}\boldsymbol{Z}_g'\boldsymbol{Z}_g.$$

MacKinnon, Nielsen, and Webb (2023b) established that when $\boldsymbol{M}_g$ are invertible, $\widehat{\boldsymbol{e}}_{-g}=\boldsymbol{M}_g^{-1}\widehat{\boldsymbol{e}}_g$. Using (76) and $\mathbb{E}\left[\widehat{\boldsymbol{e}}_g\widehat{\boldsymbol{e}}_g'\right]=\boldsymbol{M}_g$ under $\boldsymbol{\Sigma}_g=\boldsymbol{I}_{n_g}$

$$\begin{aligned}\mathbb{E}\left[\widehat{v}^2\mid\boldsymbol{X}\right] &= \sum_{g=1}^{G}\boldsymbol{Z}_g'\boldsymbol{M}_g^{-1}\mathbb{E}\left[\widehat{\boldsymbol{e}}_g\widehat{\boldsymbol{e}}_g'\mid\boldsymbol{X}\right]\boldsymbol{M}_g^{-1}\boldsymbol{Z}_g\\ &= \sum_{g=1}^{G}\boldsymbol{Z}_g'\boldsymbol{M}_g^{-1}\boldsymbol{Z}_g\\ &= \sum_{g=1}^{G}\boldsymbol{Z}_g'\boldsymbol{Z}_g+\sum_{g=1}^{G}\boldsymbol{Z}_g'\left(\boldsymbol{M}_g^{-1}-\boldsymbol{I}_{n_g}\right)\boldsymbol{Z}_g.\end{aligned}$$

Hence

$$\left|a^2-1\right|\leq\frac{\left|\sum_{g=1}^{G}\boldsymbol{Z}_g'\left(\boldsymbol{M}_g^{-1}-\boldsymbol{I}_{n_g}\right)\boldsymbol{Z}_g\right|}{\sum_{g=1}^{G}\boldsymbol{Z}_g'\boldsymbol{Z}_g}\leq\max_{1\leq g\leq G}\left\|\boldsymbol{M}_g^{-1}-\boldsymbol{I}_{n_g}\right\|=o(1)$$

by (93). This establishes $a\to 1$ as claimed.

We next consider K. Notice that $K=a^4/\left\|\boldsymbol{D}\right\|_F^2$. From (91) we can deduce that

$$\boldsymbol{D}=\left(\bar{\boldsymbol{Z}}+\overline{\boldsymbol{X}\boldsymbol{U}}-\boldsymbol{X}\boldsymbol{U}'\right)'\left(\bar{\boldsymbol{Z}}+\overline{\boldsymbol{X}\boldsymbol{U}}-\boldsymbol{X}\boldsymbol{U}'\right)/v_0^2$$

where $\tilde{\boldsymbol{Z}}$ is defined in (71) and $\overline{\boldsymbol{XU}} = \operatorname{diag}\left(\boldsymbol{X}_g \boldsymbol{U}_g\right)$. By the strong Schwarz matrix inequality and the triangle inequality

$$\begin{aligned}
n^2 \|\boldsymbol{D}\|_F &= \left\| \tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}} + \tilde{\boldsymbol{Z}}'\left(\overline{\boldsymbol{XU}} - \boldsymbol{XU}'\right) + \left(\overline{\boldsymbol{XU}} - \boldsymbol{XU}'\right)'\tilde{\boldsymbol{Z}} + \left(\overline{\boldsymbol{XU}} - \boldsymbol{XU}'\right)'\left(\overline{\boldsymbol{XU}} - \boldsymbol{XU}'\right) \right\|_F \\
&\leq \left\| \tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}} \right\|_F + 2\left\| \tilde{\boldsymbol{Z}}'\left(\overline{\boldsymbol{XU}} - \boldsymbol{XU}'\right) \right\|_F + \left\| \left(\overline{\boldsymbol{XU}} - \boldsymbol{XU}'\right)'\left(\overline{\boldsymbol{XU}} - \boldsymbol{XU}'\right) \right\|_F \\
&\leq \left\| \tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}} \right\|_F + 4\left\| \tilde{\boldsymbol{Z}} \right\|_F \left\| \boldsymbol{XU}' \right\|_F + 4\left\| \boldsymbol{XU}' \right\|_F^2.
\end{aligned}$$

We calculate that

$$\left\| \tilde{\boldsymbol{Z}} \right\|_F^2 = \sum_{g=1}^{G} \boldsymbol{Z}_g' \boldsymbol{Z}_g = R'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} R = n^2 \tag{94}$$

and

$$\begin{aligned}
\left\| \tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}} \right\|_F^2 &= \sum_{g=1}^{G} \left(\boldsymbol{Z}_g' \boldsymbol{Z}_g\right)^2 \\
&\leq \left(\sum_{g=1}^{G} \boldsymbol{Z}_g' \boldsymbol{Z}_g \right) \max_{g \leq G} \boldsymbol{Z}_g' \boldsymbol{Z}_g \\
&= n^2 \max_{g \leq G} R'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{X}_g' \boldsymbol{X}_g \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} R \\
&\leq n^2 R'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} R \max_{g \leq G} \left\| \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1/2} \boldsymbol{X}_g' \boldsymbol{X}_g \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1/2} \right\| \\
&= n^4 \max_{g \leq G} \left\| \boldsymbol{X}_g \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{X}_g' \right\| \\
&\leq n^4 o(1).
\end{aligned}$$

The second inequality is the quadratic inequality. The final inequality is (92).

Using $\boldsymbol{U}_g = \left(\boldsymbol{X}'\boldsymbol{X} - \boldsymbol{X}_g'\boldsymbol{X}_g\right)^{-1} \boldsymbol{X}_g'\boldsymbol{Z}_g$, we calculate that

$$\begin{aligned}
\left\| \boldsymbol{XU}' \right\|_F^2 &= \operatorname{tr}\left(\boldsymbol{UX}'\boldsymbol{XU}'\right) = \sum_{g=1}^{G} \boldsymbol{U}_g' \boldsymbol{X}'\boldsymbol{X}\boldsymbol{U}_g \\
&= \sum_{g=1}^{G} \boldsymbol{Z}_g' \boldsymbol{X}_g \left(\boldsymbol{X}'\boldsymbol{X} - \boldsymbol{X}_g'\boldsymbol{X}_g\right)^{-1} \boldsymbol{X}'\boldsymbol{X} \left(\boldsymbol{X}'\boldsymbol{X} - \boldsymbol{X}_g'\boldsymbol{X}_g\right)^{-1} \boldsymbol{X}_g'\boldsymbol{Z}_g \\
&= \sum_{g=1}^{G} \boldsymbol{Z}_g' \boldsymbol{X}_g \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{X}_g'\boldsymbol{Z}_g + 2\sum_{g=1}^{G} \boldsymbol{Z}_g' \boldsymbol{X}_g \boldsymbol{H}_g \boldsymbol{X}_g'\boldsymbol{Z}_g + \sum_{g=1}^{G} \boldsymbol{Z}_g' \boldsymbol{X}_g \boldsymbol{H}_g \boldsymbol{X}'\boldsymbol{X}\boldsymbol{H}_{\boldsymbol{g}} \boldsymbol{X}_g'\boldsymbol{Z}_g
\end{aligned} \tag{95}$$

where $\boldsymbol{H}_g = \left(\boldsymbol{X}'\boldsymbol{X} - \boldsymbol{X}_g'\boldsymbol{X}_g\right)^{-1} - \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$. The first component on the right of (95) is bounded by

$$\sum_{g=1}^{G} \boldsymbol{Z}_g'\boldsymbol{Z}_g \max_{g \leq G} \left\| \boldsymbol{X}_g \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{X}_g' \right\| = n^2 o(1)$$

by the quadratic inequality, (94), and (92).

We establish below that

$$n \max_{g \leq G} \left\| \boldsymbol{H}_g \right\| \leq o(1) \tag{96}$$

and

$$\max_{g \leq G} \left\| \boldsymbol{H}_g \boldsymbol{X}' \boldsymbol{X} \right\| \leq 1 + o(1). \tag{97}$$

Then the second component on the right side of (95) is bounded by

$$2 \sum_{g=1}^{G} \boldsymbol{Z}_g' \boldsymbol{Z}_g \max_{g \leq G} \left\| \boldsymbol{X}_g \right\|^2 \max_{g \leq G} \left\| \boldsymbol{H}_g \right\| \leq \nu^2 o(1)$$

by the quadratic inequality, the Schwarz matrix inequality, (94), the assumption on $\boldsymbol{X}_g$, and (96). The third component on the right side of (95) is bounded by

$$\sum_{g=1}^{G} \boldsymbol{Z}_g' \boldsymbol{Z}_g \max_{g \leq G} \left\| \boldsymbol{X}_g \right\|^2 \max_{g \leq G} \left\| \boldsymbol{H}_g \boldsymbol{X}' \boldsymbol{X} \right\| \max_{g \leq G} \left\| \boldsymbol{H}_g \right\| \leq \nu^2 o(1)$$

by the quadratic inequality, the Schwarz matrix inequality, (94), the assumption on $\boldsymbol{X}_g$, (96), and (97). Together, (95) is bounded by $\nu^2 o(1)$.

We have shown that

$$\| \boldsymbol{D} \|_F \leq \frac{\left\| \bar{\boldsymbol{Z}}' \bar{\boldsymbol{Z}} \right\|_F + 4 \left\| \bar{\boldsymbol{Z}}' \bar{\boldsymbol{Z}} \right\|_F \left\| \boldsymbol{X} \boldsymbol{U}' \right\|_F + 4 \left\| \boldsymbol{X} \boldsymbol{U}' \right\|_F^2}{\nu^2} \to 0.$$

Hence $K = a / \| \boldsymbol{D} \|_F^2 \to \infty$ as claimed.

The proof is completed by demonstrating (96)-(97). By the Woodbury identity

$$\left(\boldsymbol{X}' \boldsymbol{X} - \boldsymbol{X}_g' \boldsymbol{X}_g \right)^{-1} = \left(\boldsymbol{X}' \boldsymbol{X} \right)^{-1} + \left(\boldsymbol{X}' \boldsymbol{X} \right)^{-1} \boldsymbol{X}_g' \boldsymbol{M}_g^{-1} \boldsymbol{X}_g \left(\boldsymbol{X}' \boldsymbol{X} \right)^{-1}. \tag{98}$$

Thus

$$\begin{aligned} n \max_{g \leq G} \left\| \boldsymbol{H}_g \right\| &= n \max_{g \leq G} \left\| \left(\boldsymbol{X}' \boldsymbol{X} \right)^{-1} \boldsymbol{X}_g' \boldsymbol{M}_g^{-1} \boldsymbol{X}_g \left(\boldsymbol{X}' \boldsymbol{X} \right)^{-1} \right\| \\ &\leq \frac{1}{n} \left\| \left(\frac{1}{n} \boldsymbol{X}' \boldsymbol{X} \right)^{-1} \right\|^2 \max_{g \leq G} \left\| \boldsymbol{X}_g \right\|^2 \max_{g \leq G} \left\| \boldsymbol{M}_g^{-1} \right\| \leq o(1) \end{aligned}$$

by the Schwarz matrix inequality, the assumptions, and (93). This is (96).

Using (98), $\boldsymbol{H}_g\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{I}_k + \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}_g'\boldsymbol{M}_g^{-1}\boldsymbol{X}_g$. Thus by the triangle inequality,

$$\begin{aligned}\max_{g\leq G}\left\|\boldsymbol{H}_g\boldsymbol{X}'\boldsymbol{X}\right\| &\leq 1 + \max_{g\leq G}\left\|\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}_g'\boldsymbol{M}_g^{-1}\boldsymbol{X}_g\right\| \\ &\leq 1 + \frac{1}{n}\left\|\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\right\|\max_{g\leq G}\left\|\boldsymbol{X}_g\right\|^2\max_{g\leq G}\left\|\boldsymbol{M}_g^{-1}\right\| \leq 1 + o(1).\end{aligned}$$

This is (97), and completes the proof. ■

Proof of Equations (38)-(39): Equation (33) with $\boldsymbol{\Sigma}_0 = \boldsymbol{I}_{n_g}$ is

$$\boldsymbol{D} = \left(\boldsymbol{S} + \boldsymbol{U}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{U}' - \boldsymbol{V}\boldsymbol{U}' - \boldsymbol{U}\boldsymbol{V}'\right)/\nu^2$$

with $\nu^2 = R'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}R$. Expressions (38)-(39) follow by standard matrix operations plus the following two equalities:

$$\operatorname{tr}\left[\boldsymbol{U}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{U}'\right] = \operatorname{tr}\left[\boldsymbol{U}\boldsymbol{V}'\right] \tag{99}$$

$$\operatorname{tr}\left[\boldsymbol{S}\boldsymbol{U}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{U}'\right] = \operatorname{tr}\left[\boldsymbol{S}\boldsymbol{U}\boldsymbol{V}'\right]. \tag{100}$$

Both expressions follow from

$$\boldsymbol{U}_g'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{U}_g = \boldsymbol{U}_g'\boldsymbol{V}_g \tag{101}$$

which we now show. The assumption that $\left(\boldsymbol{X}'\boldsymbol{X} - \boldsymbol{X}_g'\boldsymbol{X}_g\right)^-$ is a reflective generalized inverse implies that it satisfies

$$\begin{aligned}\boldsymbol{U}_g'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{U}_g &= \boldsymbol{U}_g'\left(\boldsymbol{X}'\boldsymbol{X} - \boldsymbol{X}_g'\boldsymbol{X}_g\right)\boldsymbol{U}_g + \boldsymbol{U}_g'\boldsymbol{X}_g'\boldsymbol{X}_g\boldsymbol{U}_g \\ &= R'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}_g'\boldsymbol{X}_g\left(\boldsymbol{X}'\boldsymbol{X} - \boldsymbol{X}_g'\boldsymbol{X}_g\right)^-\left(\boldsymbol{X}'\boldsymbol{X} - \boldsymbol{X}_g'\boldsymbol{X}_g\right)\left(\boldsymbol{X}'\boldsymbol{X} - \boldsymbol{X}_g'\boldsymbol{X}_g\right)^-\boldsymbol{X}_g'\boldsymbol{X}_g\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}R \\ &+ \boldsymbol{U}_g'\boldsymbol{X}_g'\boldsymbol{X}_g\boldsymbol{U}_g \\ &= \boldsymbol{U}_g'\boldsymbol{X}_g'\boldsymbol{X}_g\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}R + \boldsymbol{U}_g'\boldsymbol{X}_g'\boldsymbol{X}_g\boldsymbol{U}_g \\ &= \boldsymbol{U}_g'\boldsymbol{V}_g\end{aligned}$$

where the final line uses the definition of $\boldsymbol{V}_g$ under $\boldsymbol{\Sigma}_g^0 = \boldsymbol{I}_{n_g}$. This is (101) as claimed.

Take (99)-(100). Using (101),

$$\operatorname{tr}\left[\boldsymbol{U}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{U}'\right] = \sum_{g=1}^{G}\boldsymbol{U}_g'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{U}_g = \sum_{g=1}^{G}\boldsymbol{U}_g'\boldsymbol{V}_g = \operatorname{tr}\left[\boldsymbol{U}\boldsymbol{V}'\right].$$

This is (99). Similarly, again using (101)

$$\operatorname{tr}\left[\boldsymbol{S}\boldsymbol{U}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{U}'\right] = \sum_{g=1}^{G} S_g \boldsymbol{U}_g' \boldsymbol{X}'\boldsymbol{X}\boldsymbol{U}_g = \sum_{g=1}^{G} S_g \boldsymbol{U}_g' \boldsymbol{V}_g = \operatorname{tr}\left[\boldsymbol{S}_0 \boldsymbol{U}\boldsymbol{V}'\right].$$

This is (100). ■

Asymptotic Theory

The model is the linear regression with clustered errors

$$\boldsymbol{Y}_g = \boldsymbol{X}_g \beta + \boldsymbol{e}_g. \tag{102}$$

The main text treated the regressors as fixed. For our asymptotic theory we instead assume that $(\boldsymbol{Y}_g, \boldsymbol{X}_g)$ are jointly random. Define the unconditional covariance matrix components

$$\begin{aligned}
\boldsymbol{Q}_n &= \frac{1}{n}\sum_{g=1}^{G} \mathbb{E}\left[\boldsymbol{X}_g'\boldsymbol{X}_g\right] \\
\boldsymbol{\Omega}_n &= \frac{1}{n}\sum_{g=1}^{G} \mathbb{E}\left[\boldsymbol{X}_g'\boldsymbol{e}_g\boldsymbol{e}_g'\boldsymbol{X}_g\right] \\
\boldsymbol{V}_n &= \boldsymbol{Q}_n^{-1}\boldsymbol{\Omega}_n\boldsymbol{Q}_n^{-1}.
\end{aligned}$$

An important feature of cluster asymptotic theory is that we allow the possibility of non-standard rates of convergence, which can arise due to within-cluster dependence and non-homogeneous cluster sizes. Thus, we allow the possibility that the covariance matrices $\boldsymbol{\Omega}_n$ and $\boldsymbol{V}_n$, or some sub-components of these matrices, increase with n, rather than converge to constant matrices as they do under non-clustered i.i.d. sampling.

We use the following conditions, which correspond to Theorem 9 of Hansen and Lee (2019), which estalished results for test statistics constructed with the CRVE_1 variance estimator. Let (Y_{ig}, X_{ig}) denote a single observation. Let $\lambda_{\min}(\boldsymbol{A})$ and $\lambda_{\max}(\boldsymbol{A})$ denote the smallest and largest eigenvalue of a Hermitian matrix $\boldsymbol{A}$, let $\|\boldsymbol{a}\| = \left(\boldsymbol{a}'\boldsymbol{a}\right)^{1/2}$ denote the Euclidean norm for a vector $\boldsymbol{a}$, and let $\|\boldsymbol{A}\| = \left(\lambda_{\max}\left(\boldsymbol{A}'\boldsymbol{A}\right)\right)^{1/2}$ denote the spectral norm of a matrix $\boldsymbol{A}$.

Assumption 5 *The clusters $(\boldsymbol{Y}_g, \boldsymbol{X}_g)$ are mutually independent across g. For some $2 \le r < s < \infty$, $C < \infty$, and $\delta > 0$,*

1. $\mathbb{E}\left[\boldsymbol{X}_g'\boldsymbol{e}_g\right] = 0$
2. $\max_{g \le G} \dfrac{n_g^2}{n} \to 0$

3. $n^{-1}\left(\sum_{g=1}^{G} n_g^r\right)^{2/r} \leq C$

4. $\mathbb{E}\left|Y_{ig}\right|^{2s} \leq C$

5. $\mathbb{E}\left\|X_{ig}\right\|^{2s} \leq C$

6. $\lambda_{\min}\left(\boldsymbol{Q}_n\right) \geq \delta$

7. $\lambda_{\min}\left(\boldsymbol{\Omega}_n\right) \geq \delta$.

Assumption 5.1 states that the model is a linear projection. Assumption 5.2 and 5.3 regulate the cluster sizes n_g. The assumptions allow n_g to be heterogeneous and increase with n, but do not allow any individual cluster to dominate the full sample asymptotically. Assumption 5.2 specifies that the largest cluster size must increase at a slower rate than the square root of the total sample size. Assumption 5.3 is non-intuitive, but is an additional restriction on the allowable heterogeneity in the cluster sample sizes. The parameter r involves a trade-off with the moment conditions of Assumptions 5.4-5.5. Assumption 5.3 is less restrictive for large r, and more restrictive for small r. (At $r = 2$ it requires the cluster sizes n_g to be bounded. At $r = \infty$ it states $n^{-1}\max_{g \leq G} n_g^2 \leq C$, which is implied by Assumption 5.3 so is redundant.) Assumptions 5.4-5.5 are moment bounds, where the number of required finite moments is $2s$ for some $s > r$. For bounded observations we can set $r = s = \infty$, eliminating the need for Assumption 5.3. The least restrictive moment condition sets $r = 2$, which requires just over four finite moments (as is conventional for regression asymptotic theory) but requires that the cluster sizes n_g are bounded. Assumptions 5.6-5.7 state that the covariance matrix components $\boldsymbol{Q}_n$ and $\boldsymbol{\Omega}_n$ are uniformly full rank.

We now establish that the linear functions of the least squares estimator $\widehat{\beta}$ are asymptotically normal when standardized by the jackknife covariance matrix estimator.

Theorem 8 *Take model (102) under Assumption 5. Let $\widehat{\beta}$ be the least squares estimator of β, and let $\widehat{\boldsymbol{V}}_5$ be the jackknife variance estimator (10). For any sequence of $k \times q$ full rank matrices $\boldsymbol{R}_n$,*

$$\left(\boldsymbol{R}_n' \widehat{\boldsymbol{V}}_5 \boldsymbol{R}_n\right)^{-1/2} \boldsymbol{R}_n' \left(\widehat{\beta} - \beta\right) \underset{d}{\longrightarrow} \mathrm{N}\left(0, \boldsymbol{I}_q\right) \tag{103}$$

as $n \to \infty$. For the case $q = 1$ this implies

$$\frac{\widehat{\theta} - \theta}{\widehat{\nu}_5} \underset{d}{\longrightarrow} \mathrm{N}(0,1). \tag{104}$$

For the Satterthwaite coefficients a and K (equations (34) and (35)) calculated with $\boldsymbol{\Sigma}_g^0 = \boldsymbol{I}_{n_g}$,

$$a \underset{p}{\longrightarrow} 1 \tag{105}$$

$$K \underset{p}{\longrightarrow} \infty. \tag{106}$$

For the Satterthwaite interval and p-value (equastions (30) and (31))

$$\mathbb{P}\left[\theta \in \widetilde{C}_5\right] \longrightarrow 1 - \alpha \tag{107}$$

$$p \underset{d}{\longrightarrow} U[0,1], \tag{108}$$

the latter under $\theta = \theta_0$.

Equation (104) shows that the jackknife t-ratios have standard asymptotic normal distributions under the same conditions as for CRVE t-ratios. Equation (103) is a multivariate generalization, showing that sets of coefficient estimates are asymptotically normal, and immediately implies that jackknife Wald statistics have standard asymptotic chi-square distributions.

Equations (105)-(108) describe the properties of the default Satterthwaite approximations under the same conditions. Equation (105) shows that the scale adjustment converges in probability to one, and (106) shows that the Satterthwaite degree-of-freedom K diverges to infinity, meaning that asymptotically the adjustment becomes negligible, and adjusted inference reduces to conventional inference. Equations (107)-(108) show that this implies that the Satterthwaite inference procedures produce correct inferences. Equation (107) shows that the recommended default Satterthwaite confidence interval has asymptotically correct coverage for any regression model satisfying Assumption 5. Similarly, equation (108) shows that the recommend default Satterthwaite p-values have asymptotically correct $U[0,1]$ null distributions.

Proof of Theorem 8: We start with some preliminary results. Define $\widehat{\boldsymbol{Q}}_n = \frac{1}{n}\boldsymbol{X}'\boldsymbol{X}$ and recall the definition $\boldsymbol{M}_g = \boldsymbol{I}_{n_g} - \boldsymbol{X}_g\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}_g'$. We first establish that

$$\left\|\widehat{\boldsymbol{Q}}_n^{-1}\right\| \leq O_p(1) \tag{109}$$

$$\frac{1}{n}\max_{g\leq G}\left\|\boldsymbol{X}_g\right\|^2 \leq o_p(1) \tag{110}$$

$$\max_{g\leq G}\left\|\boldsymbol{I}_{n_g} - \boldsymbol{M}_g\right\| = \max_{g\leq G}\left\|\boldsymbol{X}_g\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}_g'\right\| = o_p(1) \tag{111}$$

$$\max_{g\leq G}\left\|\boldsymbol{M}_g^{-1} - \boldsymbol{I}_{n_g}\right\| = o_p(1). \tag{112}$$

By the Schwarz matrix inequality

$$\left\|\boldsymbol{Q}_n^{-1/2}\widehat{\boldsymbol{Q}}_n\boldsymbol{Q}_n^{-1/2}-\boldsymbol{I}_k\right\| \leq \left\|\boldsymbol{Q}_n^{-1}\right\|\left\|\widehat{\boldsymbol{Q}}_n-\boldsymbol{Q}_n\right\| \leq o_p(1).$$

The final inequality holds because Assumption 5.6 implies $\left\|\boldsymbol{Q}_n^{-1}\right\| \leq \delta^{-1} < \infty$ and Theorem 1 of Hansen and Lee (2019) established $\left\|\widehat{\boldsymbol{Q}}_n-\boldsymbol{Q}_n\right\| = o_p(1)$. Equivalently, $\boldsymbol{Q}_n^{-1/2}\widehat{\boldsymbol{Q}}_n\boldsymbol{Q}_n^{-1/2} \xrightarrow[p]{} \boldsymbol{I}_k$. By the continuous mapping theorem we deduce $\boldsymbol{Q}_n^{1/2}\widehat{\boldsymbol{Q}}_n^{-1}\boldsymbol{Q}_n^{1/2} \xrightarrow[p]{} \boldsymbol{I}_k$. Using the triangle inequality, the Schwarz matrix inequality, and $\left\|\boldsymbol{Q}_n^{-1}\right\| \leq \delta^{-1}$,

$$\begin{aligned}
\left\|\widehat{\boldsymbol{Q}}_n^{-1}\right\| &= \left\|\widehat{\boldsymbol{Q}}_n^{-1}-\boldsymbol{Q}_n^{-1}+\boldsymbol{Q}_n^{-1}\right\| \\
&\leq \left\|\widehat{\boldsymbol{Q}}_n^{-1}-\boldsymbol{Q}_n^{-1}\right\| + \left\|\boldsymbol{Q}_n^{-1}\right\| \\
&\leq \left\|\boldsymbol{Q}_n^{-1}\right\|\left\|\boldsymbol{Q}_n^{1/2}\widehat{\boldsymbol{Q}}_n^{-1}\boldsymbol{Q}_n^{1/2}-\boldsymbol{I}_k\right\| + \left\|\boldsymbol{Q}_n^{-1}\right\| \\
&\leq O_p(1).
\end{aligned}$$

This is (109).

Since the spectral norm is less than the Frobenius norm and $n_g \leq \sqrt{n}$ for n sufficiently large by Assumption 5.2,

$$\frac{1}{n}\left\|\boldsymbol{X}_g\right\|^2 \leq \frac{1}{n}\sum_{i=1}^{n_g}\left\|X_{ig}\right\|^2 \leq \frac{1}{n^{1/2}}\left(\frac{1}{n_g}\sum_{i=1}^{n_g}\left\|X_{ig}\right\|^2\right). \tag{113}$$

Assumption 5.5 implies that $\left\|X_{ig}\right\|^{2r}$ is uniformly integrable. Lemma 1 of Hansen and Lee (2019) shows that this implies that $\left(n_g^{-1}\sum_{i=1}^{n_g}\left\|X_{ig}\right\|^2\right)^r$ is uniformly integrable. Theorem 9.7 of Hansen (2022) shows that this implies $\max_{g\leq G}\left(n_g^{-1}\sum_{i=1}^{n_g}\left\|X_{ig}\right\|^2\right) = o_p\left(G^{1/r}\right)$. We find that (113) is uniformly bounded by $o_p\left(n^{-1/2}G^{1/r}\right) \leq o_p(1)$, since $G \leq n$ and $r \geq 2$. This establishes (110).

Using the Schwarz matrix inequality, (109), and (110),

$$\max_{g\leq G}\left\|\boldsymbol{I}_{n_g}-\boldsymbol{M}_g\right\| = \max_{g\leq G}\left\|\boldsymbol{X}_g\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}_g'\right\| \leq \left\|\widehat{\boldsymbol{Q}}_n^{-1}\right\|\frac{1}{n}\max_{g\leq G}\left\|\boldsymbol{X}_g\right\|^2 \leq o_p(1).$$

This is (111). As $\boldsymbol{I}_{n_g}-\boldsymbol{M}_g$ is positive semi-definite,

$$\max_{g\leq G}\left\|\boldsymbol{I}_{n_g}-\boldsymbol{M}_g\right\| = \max_{g\leq G}\lambda_{\max}\left(\boldsymbol{I}_{n_g}-\boldsymbol{M}_g\right) = 1-\min_{g\leq G}\lambda_{\min}(\boldsymbol{M}_g).$$

Thus (111) implies that $\min_{g\leq G}\lambda_{\min}(\boldsymbol{M}_g) \xrightarrow[p]{} 1$. One implication is that the matrices $\boldsymbol{M}_g$ are asymptotically invertible, which means that the regresssion is clusterwise invertible. For the remainder of the proof we assume that the sample size is sufficiently large so that this holds.

Equation (92) also implies that

$$\begin{aligned}\max_{g\leq G}\left\|\boldsymbol{M}_g^{-1}-\boldsymbol{I}_{n_g}\right\| &= \max_{g\leq G}\lambda_{\max}(\boldsymbol{M}_g^{-1}-\boldsymbol{I}_{n_g}) \\ &= \max_{g\leq G}\lambda_{\max}(\boldsymbol{M}_g^{-1})-1 \\ &= \frac{1}{\min_{g\leq G}\lambda_{\min}(\boldsymbol{M}_g)}-1 \\ &\underset{p}{\longrightarrow} 0\end{aligned}$$

This is (112).

Hansen and Lee (2019, Theorem 9) proved (103) for the CRVE variance estimator under Assumption 5. We establish (103) for the jackknife variance estimator by showing that the replacement of the variance estimators is asymptotically negligible.

It will be useful to define the central component of the CRVE estimator

$$\widehat{\boldsymbol{\Omega}}_n = \frac{1}{n}\sum_{g=1}^{G}\boldsymbol{X}_g'\widehat{\boldsymbol{e}}_g\widehat{\boldsymbol{e}}_g'\boldsymbol{X}_g.$$

We define the analog for the jackknife estimator:

$$\widetilde{\boldsymbol{\Omega}}_n = \frac{1}{n}\sum_{g=1}^{G}\boldsymbol{X}_g'\widehat{\boldsymbol{e}}_{-g}\widehat{\boldsymbol{e}}_{-g}'\boldsymbol{X}_g = \frac{1}{n}\sum_{g=1}^{G}\boldsymbol{X}_g'\boldsymbol{M}_g^{-1}\widehat{\boldsymbol{e}}_g\widehat{\boldsymbol{e}}_g'\boldsymbol{M}_g^{-1}\boldsymbol{X}_g.$$

The second equality holds because MacKinnon, Nielsen, and Webb (2023) established that under clusterwise invertibility, $\widehat{\boldsymbol{e}}_{-g} = \boldsymbol{M}_g^{-1}\widehat{\boldsymbol{e}}_g$.

Examining the proof of Theorem 9 of Hansen and Lee (2019), the key is equation (89) in their supplemental appendix:

$$\left\|\boldsymbol{\Omega}_n^{-1/2}\left(\widehat{\boldsymbol{\Omega}}_n-\boldsymbol{I}_k\right)\boldsymbol{\Omega}_n^{-1/2}\right\| \underset{p}{\longrightarrow} 0. \tag{114}$$

The analog needed for the jackknife estimator is to demonstrate that

$$\left\|\boldsymbol{\Omega}_n^{-1/2}\left(\widetilde{\boldsymbol{\Omega}}_n-\boldsymbol{I}_k\right)\boldsymbol{\Omega}_n^{-1/2}\right\| \underset{p}{\longrightarrow} 0. \tag{115}$$

Our goal is therefore to demonstrate (115), which is sufficient to establish (103).

Using the triangle inequality

$$\begin{aligned}\left\|\boldsymbol{\Omega}_n^{-1/2}\left(\widetilde{\boldsymbol{\Omega}}_n-\boldsymbol{I}_k\right)\boldsymbol{\Omega}_n^{-1/2}\right\| &\leq \left\|\boldsymbol{\Omega}_n^{-1/2}\left(\widetilde{\boldsymbol{\Omega}}_n-\widehat{\boldsymbol{\Omega}}_n\right)\boldsymbol{\Omega}_n^{-1/2}\right\| + \left\|\boldsymbol{\Omega}_n^{-1/2}\left(\widehat{\boldsymbol{\Omega}}_n-\boldsymbol{I}_k\right)\boldsymbol{\Omega}_n^{-1/2}\right\| \\ &= \left\|\boldsymbol{\Omega}_n^{-1/2}\left(\widetilde{\boldsymbol{\Omega}}_n-\widehat{\boldsymbol{\Omega}}_n\right)\boldsymbol{\Omega}_n^{-1/2}\right\| + o_p(1)\end{aligned}$$

where the final equality is (114). It is therefore sufficient to show that

$$\left\|\boldsymbol{\Omega}_n^{-1/2}\left(\widetilde{\boldsymbol{\Omega}}_n-\widehat{\boldsymbol{\Omega}}_n\right)\boldsymbol{\Omega}_n^{-1/2}\right\| \xrightarrow[p]{} 0. \tag{116}$$

Define $\boldsymbol{P}_g = \boldsymbol{M}_g^{-1} - \boldsymbol{I}_{n_g}$, which satisfies $\max_{g\leq G}\left\|\boldsymbol{P}_g\right\| = o_p(1)$ by (112). Using the Triangle and Schwarz matrix inequalities,

$$\begin{aligned}
&\left\|\boldsymbol{\Omega}_n^{-1/2}\left(\widetilde{\boldsymbol{\Omega}}_n-\widehat{\boldsymbol{\Omega}}_n\right)\boldsymbol{\Omega}_n^{-1/2}\right\| \\
&= \left\|\boldsymbol{\Omega}_n^{-1/2}\left(\frac{1}{n}\sum_{g=1}^{G}\boldsymbol{X}_g'\boldsymbol{P}_g\widehat{\boldsymbol{e}}_g\widehat{\boldsymbol{e}}_g'\boldsymbol{P}_g\boldsymbol{X}_g+\frac{1}{n}\sum_{g=1}^{G}\boldsymbol{X}_g'\boldsymbol{P}_g\widehat{\boldsymbol{e}}_g\widehat{\boldsymbol{e}}_g'\boldsymbol{X}_g+\frac{1}{n}\sum_{g=1}^{G}\boldsymbol{X}_g'\widehat{\boldsymbol{e}}_g\widehat{\boldsymbol{e}}_g'\boldsymbol{P}_g\boldsymbol{X}_g\right)\boldsymbol{\Omega}_n^{-1/2}\right\| \\
&\leq \left\|\frac{1}{n}\sum_{g=1}^{G}\boldsymbol{\Omega}_n^{-1/2}\boldsymbol{X}_g'\boldsymbol{P}_g\widehat{\boldsymbol{e}}_g\widehat{\boldsymbol{e}}_g'\boldsymbol{P}_g\boldsymbol{X}_g\boldsymbol{\Omega}_n^{-1/2}\right\| + 2\left\|\frac{1}{n}\sum_{g=1}^{G}\boldsymbol{\Omega}_n^{-1/2}\boldsymbol{X}_g'\boldsymbol{P}_g\widehat{\boldsymbol{e}}_g\widehat{\boldsymbol{e}}_g'\boldsymbol{X}_g\boldsymbol{\Omega}_n^{-1/2}\right\| \\
&\leq \frac{1}{n}\sum_{g=1}^{G}\left\|\boldsymbol{\Omega}_n^{-1/2}\boldsymbol{X}_g'\widehat{\boldsymbol{e}}_g\right\|^2 o_p(1) \\
&\leq o_p(1).
\end{aligned}$$

The fact $n^{-1}\sum_{g=1}^{G}\left\|\boldsymbol{\Omega}_n^{-1/2}\boldsymbol{X}_g'\widehat{\boldsymbol{e}}_g\right\|^2 = O_p(1)$ follows implicitly from Hansen and Lee's proof of (114). This is (116), which completes the proof of (103).

Theorem 7 in the main text established (105) and (106) under (109)-(110).

We next establish (107). Given the definition of $\widetilde{C}_5$ and results (104), (105), and (106),

$$\mathbb{P}\left[\theta\in\widetilde{C}_5\right] = \mathbb{P}\left[\left|\frac{\widehat{\theta}-\theta}{\widehat{v}_5}\right| \leq \frac{t_K^{1-\alpha/2}}{a}\right] \to \mathbb{P}\left[|\mathrm{N}(0,1)| \leq t_\infty^{1-\alpha/2}\right] = 1-\alpha$$

which is (107).

Similarly, given the definition of the p-value p,

$$p = 2\left(1-G\left(a\left|\frac{\widehat{\theta}-\theta}{\widehat{v}_5}\right|, K\right)\right) \to 2\left(1-\Phi\left(|\mathrm{N}(0,1)|\right)\right) \sim U[0,1]$$

which is (108).

This completes the proof. ■

Wild Bootstrap

We describe here the details of our implementation of the wild bootstrap with jackknife standard errors, as used in the numerical simulation. Our implementation is modeled on MacKinnon (2023) who describes a fast wild bootstrap implementation with CRVE_1 standard errors.

It is useful to describe the algorithm first for a fixed value of the parameter θ, and then discuss how this is used to (separately) calculate p-values and confidence intervals. Define $\nu_0^2 = R'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}R$ and $\boldsymbol{Z}_g = \boldsymbol{X}_g\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}R$. Let $\widetilde{\beta} = \widehat{\beta} - \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}R\widehat{\theta}/\nu_0^2$ denote the constrained least squares estimator subject to $\widetilde{\theta} = R'\widetilde{\beta} = 0$. Let $\widetilde{\boldsymbol{e}}_g = \boldsymbol{Y}_g - \boldsymbol{X}_g\widetilde{\beta} = \widehat{\boldsymbol{e}}_g + \boldsymbol{Z}_g\widehat{\theta}/\nu_0^2$ denote the associated cluster-level residual. For given θ, let $\widetilde{\beta}(\theta) = \widetilde{\beta} + \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}R\theta/\nu_0^2$ denote the constrained least squares estimator subject to $\widetilde{\theta}(\theta) = R'\widetilde{\beta} = \theta$. Let $\widetilde{\boldsymbol{e}}_g(\theta) = \boldsymbol{Y}_g - \boldsymbol{X}_g\widetilde{\beta}(\theta) = \widetilde{\boldsymbol{e}}_g - \boldsymbol{Z}_g\theta/\nu_0^2$ denote the associated cluster-level residual.

For each bootstrap draw we simulate a $G\times 1$ vector $\boldsymbol{\phi} \sim \mathrm{N}(0, \boldsymbol{I}_G)$ with gth element ϕ_g. The bootstrap dependent variable equals $\boldsymbol{Y}_g^* = \widetilde{\boldsymbol{e}}_g(\theta)\phi_g$. The bootstrap version of $\widehat{\theta}$ is

$$\widehat{\theta}^* = R'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{Y}^*\right) = \sum_{g=1}^{G}\boldsymbol{Z}_g'\widetilde{\boldsymbol{e}}_g(\theta)\phi_g = \boldsymbol{a}_0'\boldsymbol{\phi} - \boldsymbol{a}_1'\boldsymbol{\phi}\theta$$

where $\boldsymbol{a}_0$ and $\boldsymbol{a}_1$ are $G\times 1$ with gth elements $a_{0g} = \boldsymbol{Z}_g'\widetilde{\boldsymbol{e}}_g$ and $a_{1g} = \boldsymbol{Z}_g'\boldsymbol{Z}_g/\nu_0^2$, respectively. The bootstrap version of $\widehat{\theta}_{-g} - \widehat{\theta}$ is

$$\begin{aligned}
\widehat{\theta}_{-g}^* - \widehat{\theta}^* &= R'\left(\boldsymbol{X}'\boldsymbol{X} - \boldsymbol{X}_g'\boldsymbol{X}_g\right)^{-}\left(\sum_{h\neq g}\boldsymbol{X}_h'\boldsymbol{Y}_h^*\right) - \widehat{\theta}^* \\
&= R'\left(\boldsymbol{X}'\boldsymbol{X} - \boldsymbol{X}_g'\boldsymbol{X}_g\right)^{-}\left(\sum_{h\neq g}\boldsymbol{X}_h'\widetilde{\boldsymbol{e}}_h(\theta)\phi_h\right) - \left(\boldsymbol{a}_0'\boldsymbol{\phi} - \boldsymbol{a}_1'\boldsymbol{\phi}\theta\right) \\
&= \left(\boldsymbol{d}_{0g} - \boldsymbol{a}_0\right)'\boldsymbol{\phi} - \left(\boldsymbol{d}_{1g} - \boldsymbol{a}_1\right)'\boldsymbol{\phi}\theta
\end{aligned}$$

where $\boldsymbol{d}_{0g}$ and $\boldsymbol{d}_{1g}$ are $G\times 1$. Their gth elements are 0, and for $h\neq g$ their hth elements are $R'\left(\boldsymbol{X}'\boldsymbol{X} - \boldsymbol{X}_g'\boldsymbol{X}_g\right)^{-}\boldsymbol{X}_h'\widetilde{\boldsymbol{e}}_h$ and $R'\left(\boldsymbol{X}'\boldsymbol{X} - \boldsymbol{X}_g'\boldsymbol{X}_g\right)^{-}\boldsymbol{X}_h'\boldsymbol{Z}_h/\nu_0^2$, respectively. Stacking, we obtain the $G\times 1$ vector $\widehat{\theta}_{-}^* - \widehat{\theta}^* = \boldsymbol{C}_0'\boldsymbol{\phi} - \boldsymbol{C}_1'\boldsymbol{\phi}\theta$ where $\boldsymbol{C}_0$ and $\boldsymbol{C}_1$ are $G\times G$ with gth column $\boldsymbol{d}_{0g} - \boldsymbol{a}_0$ and $\boldsymbol{d}_{1g} - \boldsymbol{a}_1$, respectively.

The bootstrap version of $\widehat{\nu}_5^2$ is

$$\widehat{\nu}_5^{2*} = \sum_{g=1}^{G}\left(\widehat{\theta}_{-g}^* - \widehat{\theta}^*\right)^2 = \boldsymbol{\phi}'\boldsymbol{C}_0\boldsymbol{C}_0'\boldsymbol{\phi} - 2\boldsymbol{\phi}'\boldsymbol{C}_0\boldsymbol{C}_1'\boldsymbol{\phi}\theta + \boldsymbol{\phi}'\boldsymbol{C}_1\boldsymbol{C}_1'\boldsymbol{\phi}\theta^2.$$

The bootstrap version of the squared t-statistic $\left(\widehat{\theta} - \theta\right)^2/\widehat{\nu}_5^2$ is

$$T^* = \frac{\widehat{\theta}^{*2}}{\widehat{\nu}_5^{2*}} = \frac{\left(\boldsymbol{a}_0'\boldsymbol{\phi} - \boldsymbol{a}_1'\boldsymbol{\phi}\theta\right)^2}{\boldsymbol{\phi}'\boldsymbol{C}_0\boldsymbol{C}_0'\boldsymbol{\phi} - 2\boldsymbol{\phi}'\boldsymbol{C}_0\boldsymbol{C}_1'\boldsymbol{\phi}\theta + \boldsymbol{\phi}'\boldsymbol{C}_1\boldsymbol{C}_1'\boldsymbol{\phi}\theta^2}.$$

The bootstrap repeats this calculation for B independent draws of the random vector $\boldsymbol{\phi}$. As described by MacKinnon (2023), it is computationally efficient to calculate the vectors and

matrices $\boldsymbol{a}_0$, $\boldsymbol{a}_1$, $\boldsymbol{C}_0$, and $\boldsymbol{C}_1$ before making the draws $\boldsymbol{\phi}$ and calculating the bootstrap statistics T^*. This way, calculation of T^* only involves a small number of basic matrix operations. It is also useful to observe that given the statistics $\boldsymbol{a}_0'\boldsymbol{\phi}$, $\boldsymbol{a}_1'\boldsymbol{\phi}$, $\boldsymbol{\phi}'\boldsymbol{C}_0\boldsymbol{C}_0'\boldsymbol{\phi}$, $\boldsymbol{\phi}'\boldsymbol{C}_0\boldsymbol{C}_1'\boldsymbol{\phi}$, and $\boldsymbol{\phi}'\boldsymbol{C}_1\boldsymbol{C}_1'\boldsymbol{\phi}$ the bootstrap statistic T^* is a simple function of θ (a ratio of quadratics). This is a useful insight for confidence interval construction, where T^* will need to be iteratively re-calculated for many values of θ.

For any θ, the bootstrap $1-\alpha$ critical value $c(\theta)$ is the $1-\alpha$ quantile of the empirical distribution of the bootstrap statistics T^* across the bootstrap draws $\boldsymbol{\phi}$.

A hypothesis $\theta = \theta_0$ is accepted if $\left(\widehat{\theta}-\theta_0\right)^2/\widehat{\mathbb{V}}_5^2 \leq c(\theta_0)$ and rejected otherwise. This is how we calculate the coverage probabilities in the simulation.

The $1-\alpha$ level wild bootstrap confidence interval for θ is the set of values which are accepted by the bootstrap test; equivalently, the set of θ which satisfy $\left(\widehat{\theta}-\theta\right)^2/\widehat{\mathbb{V}}_5^2 \leq c(\theta)$. Let θ_j be the set of solutions to $\left(\widehat{\theta}-\theta\right)^2/\widehat{\mathbb{V}}_5^2 = c(\theta)$. There are at least two solutions, satisfying $\theta_1 \leq \widehat{\theta} \leq \theta_2$, but is possible that there are more. We define the confidence interval as $[\theta_L, \theta_U]$ with $\theta_L = \min\theta_j$ and $\theta_U = \max_j \theta$.

For the confidence interval length results presented in the simulation we calculate the confidence interval endpoints as follows. We take the endpoints of the Satterthwaite intervals as initial values, thus $\theta_L = \widehat{\theta} - t_K^{1-\alpha/2}\widehat{\mathbb{V}}_5/a$ and $\theta_U = \widehat{\theta} + t_K^{1-\alpha/2}\widehat{\mathbb{V}}_5/a$, then search for solutions to $\left(\widehat{\theta}-\theta\right)^2/\widehat{\mathbb{V}}_5^2 = c(\theta)$ using a local grid search to find values which bracket the solution, and then apply the bisection algorithm. This locates the unique endpoint solution when it exists, and otherwise locates the solution closest to the Satterthwaite interval.

References

[1] Hansen, Bruce E. (2022): *Econometrics*, Princeton University Press.

[2] Hansen, Bruce E. and Seojeong Lee (2019): "Asymptotic theory for clustered samples," *Journal of Econometrics*, 210, 268-290.

[3] MacKinnon, James G. (2023): "Fast cluster bootstrap methods for linear regression models," *Econometrics and Statistics*, 26, 52-71.

[4] MacKinnon, James G., Morten Orregaard Nielsen, and Matthew D. Webb (2023): "Fast and reliable jackknife and bootstrap methods for cluster-robust inference," *Journal of Applied Econometrics*, 38, 671-694.