Jackknife Standard Errors for Clustered Regression

Bruce E. Hansen*
University of Wisconsin†

This version: December 2023

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 our jackknife-based confidence interval has coverage probability bounded by the Cauchy distribution. 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 minimal 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. 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, 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 and standard errors 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 regresson 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 the following new theoretical insights. First, under standard assumptions, the CRVE variance estimator has full worst-case downward estimation bias, while our jackknife variance estimator is conservative (its expectation is larger than the exact variance). Second, under the assumption of cluster-dependent normality, 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 possible regressor and covariance matrix configurations.

Of the above described results, the most important contribution of this paper is the demonstration that the finite sample coverage probability of our jackknife confidence interval is uniformly bounded by the Cauchy distribution. This is not an elementary result, but requires a combination of finite sample analysis, convex analysis, and numerical computation.

All of our theory holds under minimal assumptions, allowing arbitrary cluster sizes, regressor leverage, within-cluster correlation, heteroskedasticity, and delete-cluster invertibility failures. Concerning the latter – delete-cluster invertibility failures – our results allow 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 saturated regressions with sparse cell proportions. 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 theoretical results are backed by the existing extensive simulation literature investigating robust variance estimation which has demonstrated that jackknife standard errors provide dramatically improved finite sample inference in a wide variety of settings.

Furthermore, we investigate the exact distribution of jackknife confidence intervals. We provide a new exact characterization of the finite sample distribution, and provide a practical approximation to this distribution based on a student t distribution.

The family of heteroskedasticity-consistent variance estimators are often written by the monikers HC_0, HC_1, HC_2, and HC_3. The HC_0 version was introduced by Eicker (1963), Huber (1967), and White (1980). The degree-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 monikers 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 Solvsten (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 the conditional critical values of Pötscher and Preinerstorfer (2023).

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 Erwin (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, 2023), MacKinnon and Webb (2020), Canay, Santos, and Shaikh (2021), and Niccodemi and Wansbeek (2022). These applications will also benefit from jackknife standard errors and our adjusted degree of freedom approximation, as their application will reduce size distortions.

The organization of the paper is as follows. Section 2 introduces clustered regression and the cluster-robust variance estimator. Section 3 introduces jackknife variance estimation. Section 4 presents results on variance estimation bias. Section 5 presents the core results on the coverage rates of confidence

intervals. Section 6 presents results on the distribution of the jackknife t-statistic. Section 7 proposes an adjusted degree-of-freedom for a student t distributional approximation. Section 8 discusses models with cluster-level fixed effects. Section 9 presents simulation evidence. Section 10 presents an empirical application. A conclusion is presented in Section 11. The technical proofs are presented in Section 12.

The R code which generates the numerical calculations presented in the paper, as well as R and Stata code to implement the proposed variance estimator and degree-of-freedom correction, is posted on the author's webpage `users.ssc.wisc.edu/~bhansen/`.

2 Clustered Regression and the CRVE

The model is a standard clustered sampling regression. The observations are separated into G unbalanced mutually independent clusters. Notationally, we write the observations on the ith individual in the gth cluster as (Y_{ig}, X_{ig}), for $i = 1, ..., n_g$ and $g = 1 = 1, ..., G$. The variable Y_{ig} is scalar while X_{ig} is a k-vector. It is useful to stack the observations by cluster, so that $\boldsymbol{Y}_g = (Y_{1g}, ..., Y_{n_g g})'$ is an $n_g \times 1$ vector and $\boldsymbol{X}_g = (X_{1g}, ..., X_{n_g g})'$ is an $n_g \times k$ matrix. 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 satisfy the standard 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. 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 $\boldsymbol{\Sigma}$ has finite elements.*

We focus on the least squares estimator

$$\widehat{\beta} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{Y}\right) = \left(\sum_{g=1}^{G}\boldsymbol{X}_g'\boldsymbol{X}_g\right)^{-1}\left(\sum_{g=1}^{G}\boldsymbol{X}_g'\boldsymbol{Y}_g\right).$$

It is well known that in 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

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

where $\widehat{\boldsymbol{e}}_g = \boldsymbol{Y}_g - \boldsymbol{X}_g\widehat{\beta}$ denotes the least squares residual vector for the gth cluster. CRVE_1 was introduced by Liang and Zeger (1986) and Arellano (1987). The constant appearing in (4) is an *ad hoc* degree-of-freedom correction, apparently introduced by the Stata "`cluster`" covariance matrix option.

In the absence of clustering, (4) specializes to the HC_1 "heteroskedasticity-robust" variance estimator, which is the Eicker-Huber-White (EHW) estimator of Eicker (1963), Huber (1967), and White (1980), multiplied by an $n/(n-k)$ degree-of-freedom correction as suggested by Hickley (1977). HC_1 dominates empirical practice due to its encoding as the "`r`" covariance matrix option in Stata.

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 generalized inverse of the symmetric 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 in MacKinnon and White (1985) and CRVE_2 in Bell and McCaffrey (2002) 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.

The motivation for the CRVE_2 estimator (5) that it is unbiased when the regression errors are i.i.d. (so that $\boldsymbol{\Sigma}_g = \boldsymbol{I}_{n_g}\sigma^2$) and all $\boldsymbol{M}_g$ are invertible.

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

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

and

$$\bar{\beta} = \frac{1}{G}\sum_{g=1}^{G}\widehat{\beta}_{-g}. \tag{9}$$

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

$$\widehat{\boldsymbol{V}}_4 = \frac{G-1}{G}\sum_{g=1}^{G}\left(\widehat{\beta}_{-g}-\widehat{\beta}\right)\left(\widehat{\beta}_{-g}-\widehat{\beta}\right)', \tag{10}$$

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 (10) as written are undefined if there is a cluster g for which $\boldsymbol{X}'\boldsymbol{X}-\boldsymbol{X}_g'\boldsymbol{X}_g$ is noninvertible[1]. The implementation in Stata 18 circumvents this difficulty by excluding from the sums in (7)-(10) any cluster where (8) is undefined[2]. We follow this interpretation; henceforth, we assume that (7)-(10) are implemented with this modification. We describe (7)-(10) 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 define the following jackknife estimator:

$$\widehat{\boldsymbol{V}}_5 = \sum_{g=1}^{G}\left(\widetilde{\beta}_{-g}-\widehat{\beta}\right)\left(\widetilde{\beta}_{-g}-\widehat{\beta}\right)' \tag{11}$$

where

$$\widetilde{\beta}_{-g} = \left(\boldsymbol{X}'\boldsymbol{X}-\boldsymbol{X}_g'\boldsymbol{X}_g\right)^{+}\left(\boldsymbol{X}'\boldsymbol{Y}-\boldsymbol{X}_g'\boldsymbol{Y}_g\right) \tag{12}$$

and $\boldsymbol{A}^{+}$ denotes the Moore-Penrose generalized inverse of $\boldsymbol{A}$. $\widetilde{\beta}_{-g}$ is a generalized delete-one-cluster estimator. It is defined for all g and therefore (11) includes all clusters. $\widetilde{\beta}_{-g}$ is a minimizer of the delete-one-cluster least squares criterion, and is therefore a valid delete-one-cluster version of the full-sample

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

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

estimator $\widehat{\beta}$.

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 $\widehat{\boldsymbol{V}}_5 > \widehat{\boldsymbol{V}}_4 > \widehat{\boldsymbol{V}}_3$. 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 degree-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 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 occurance, which is presumably the motivation for the "drop noninvertible clusters" modification described above. However, the properties of any such modification 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 degree-of-freedom correction $(G-1)/G$, which is typically inconsequential. They can differ more substantially, however, under clusterwise noninvertibility.

The jackknife estimators $\widehat{\boldsymbol{V}}_3$ and $\widehat{\boldsymbol{V}}_4$ for clustered samples were developed by Cochran (1977), Rust and Rao (1996), and Bell and McCaffrey (2002). See MacKinnon, Nielsen, and Webb (2023abc) for detailed discussions. Stata has codified the modification to delete noninvertible clusters. MacKinnon, Nielsen, and Webb (2023c) propose and implement a generalized delete-one-cluster estimator similar to (12), though they do not investigate the statistical properties of the resulting jackknife variance estimator.

Alternative algebraic representations of the jackknife estimator are available; see MacKinnon, Nielsen, and Webb (2023b). One we will find useful is based on the delete-one-cluster prediction errors

$$\widetilde{\boldsymbol{e}}_g = \boldsymbol{Y}_g - \boldsymbol{X}_g\widetilde{\beta}_{-g}. \tag{13}$$

In the proof of Theorem 1 in Section 12 we show that

$$\widetilde{\beta}_{-g} - \widehat{\beta} = -\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}_g'\widetilde{\boldsymbol{e}}_g, \tag{14}$$

and interestingly, this equality holds even allowing for clusterwise noninvertibility. Given (14), we can write (11) as

$$\widehat{\boldsymbol{V}}_5 = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\sum_{g=1}^{G}\boldsymbol{X}_g'\widetilde{\boldsymbol{e}}_g\widetilde{\boldsymbol{e}}_g'\boldsymbol{X}_g\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}. \tag{15}$$

The jackknife estimators (7)-(11) simplify in the absence of clustering, though existing proposals and

analysis all assume that the $\boldsymbol{M}_g$ are all invertible[3]. Under these conditions the estimator (7) corresponds to that proposed by MacKinnon and White (1985) and the estimator (11) 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 invertibility fails, the estimators HC_2 and HC_3 are undefined[4]. For more details, see the monographs of Efron (1982) and Shao and Tu (1995).

4 Biased vs Conservative Variance Estimation

In the classical regression model, the classical variance estimator is unbiased for the exact finite sample variance. Unsurprisingly, 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,*

$$\mathbb{E}\left[\widehat{\boldsymbol{V}}_5\right] \geq \boldsymbol{V}. \tag{16}$$

Theorem 1 shows that our recommended jackknife estimator $\widehat{\boldsymbol{V}}_5$ is *never downward biased* in a positive semi-definite sense. A never-downward-biased variance estimator can be described as *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 (16) 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 $\mathrm{var}\left[\bar{\beta}\right]$, not as an estimator of $\mathrm{var}\left[\widehat{\beta}\right]$. Furthermore, Theorem 2 below shows $\widehat{\boldsymbol{V}}_3$ does not satisfy the never-downward-biased property (16).

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 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

$$v^2 = \mathrm{var}\left[\widehat{\theta}\right] = R'\boldsymbol{V}R.$$

[3]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 $h_{ii} < 1$.

[4]A word of caution: Stata 18 reports HC_2 and HC_3 standard errors even when the regressor matrix is clusterwise noninvertible, and does not document how they are calculated in this situation.

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 $\widehat{v}_j = \sqrt{R'\widehat{\boldsymbol{V}}_j R}$.

It will be convenient to define sets of models. For fixed k and G let $\mathscr{F}$ be the class of all regressor and covariance matrices $(\boldsymbol{X}, \boldsymbol{\Sigma})$ such that $\boldsymbol{X}$ is full rank, $\boldsymbol{\Sigma}$ has finite elements, and $v^2 > 0$. Let $\mathscr{F}^* \subset \mathscr{F}$ be the subset where $\boldsymbol{X}$ satisfies clusterwise invertibility. Let $\mathscr{F}_0^* \subset \mathscr{F}^*$ and $\mathscr{F}_0 \subset \mathscr{F}$ be the further subsets where $\boldsymbol{\Sigma} = \boldsymbol{I}_n \sigma^2$.

The following result shows that the variance estimators other than $\widehat{\boldsymbol{V}}_5$ can be severely biased.

Theorem 2 *Suppose Assumption 1 holds. For* $(\boldsymbol{X}, \boldsymbol{\Sigma}) \in \mathscr{F}_0^*$,

$$\inf_{(\boldsymbol{X},\boldsymbol{\Sigma})\in\mathscr{F}_0^*} \frac{\mathbb{E}\left[\widehat{v}_1^2\right]}{v^2} = 0. \tag{17}$$

For $(\boldsymbol{X}, \boldsymbol{\Sigma}) \in \mathscr{F}^*$,

$$\inf_{(\boldsymbol{X},\boldsymbol{\Sigma})\in\mathscr{F}^*} \frac{\mathbb{E}\left[\widehat{v}_2^2\right]}{v^2} = 0, \tag{18}$$

$$\inf_{(\boldsymbol{X},\boldsymbol{\Sigma})\in\mathscr{F}^*} \frac{\mathbb{E}\left[\widehat{v}_3^2\right]}{v^2} = \left(\frac{G-1}{G}\right)^2 < 1, \tag{19}$$

and

$$\inf_{(\boldsymbol{X},\boldsymbol{\Sigma})\in\mathscr{F}^*} \frac{\mathbb{E}\left[\widehat{v}_4^2\right]}{v^2} = \frac{G-1}{G} < 1. \tag{20}$$

For $(\boldsymbol{X}, \boldsymbol{\Sigma}) \in \mathscr{F}_0$,

$$\inf_{(\boldsymbol{X},\boldsymbol{\Sigma})\in\mathscr{F}_0} \frac{\mathbb{E}\left[\widehat{v}_3^2\right]}{v^2} = 0 \tag{21}$$

and

$$\inf_{(\boldsymbol{X},\boldsymbol{\Sigma})\in\mathscr{F}_0} \frac{\mathbb{E}\left[\widehat{v}_4^2\right]}{v^2} = 0. \tag{22}$$

For $(\boldsymbol{X}, \boldsymbol{\Sigma}) \in \mathscr{F}$,

$$\inf_{(\boldsymbol{X},\boldsymbol{\Sigma})\in\mathscr{F}} \frac{\mathbb{E}\left[\widehat{v}_5^2\right]}{v^2} = 1. \tag{23}$$

Equation (17) 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 set of models considered are those satisfying $\boldsymbol{\Sigma} = \boldsymbol{I}_n \sigma^2$ and clusterwise invertibility, so the result (17) has nothing to do with heteroskedasticity, correlated errors, or invertibility failure. Rather, it is a consequence of extreme regressor leverage (unbalanced regressors and/or cluster sizes).

Equation (18) shows a similar result for CRVE_2 in the class of models which allows general covariance matrices (heteroskedasticity and correlation) but still requires clusterwise invertibility. The difference between (17) and (18) is that 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 (19) and (20) show related results for the conventional jackknife estimators. They show that $\widehat{v}_3^2$ and $\widehat{v}_4^2$ violate the never-downward-biased property, even under clusterwise invertibility. The

magnitude of the violation, however, is small when G is large. Regardless, (19) and (20) show that the conventional interpretation[5] 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 (21) and (22) extend the analysis of the conventional jackknife estimators to the class of models allowing clusterwise noninvertibility. The results show that under clusterwise noninvertibility the conventional estimators $\widehat{v}_3^2$ and $\widehat{v}_4^2$ are fully downward biased. This arises even though (21)-(22) restrict the set of models to satisfy $\mathbf{\Sigma} = \boldsymbol{I}_n\sigma^2$. 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 (23) examines our recommended jackknife estimator $\widehat{v}_5^2$ in the broadest context, allowing for general heteroskedasticity, correlated errors, and clusterwise noninvertibility. As implied by Theorem 1, $\widehat{v}_5^2$ is never downward biased. Equation (23) extends this further and demonstrates that the infimum equals one, meaning that the inequality of Theorem 1 is sharp.

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 $\mathbf{\Sigma}$ to vary freely. It is important to understand that the statements of Theorem 2 hold for all G, from the very small to very large. Thus the bias of CRVE_1, CRVE_2, and the conventional jackknife can be arbitrarily bad in both very small and very large samples.

The worstcase downward bias in (17)-(22) 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 degree-of-freedom adjustment of CRVE_1, are insufficient to counter the bias.

5 Unbounded vs Bounded Inference

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

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

For a nominal $100(1-\alpha)\%$ interval the conventional[6] choice for the critical value is $c = t_{G-1}^{1-\alpha/2}$, the $1-\alpha/2$ quantile of the student t distribution with $G-1$ degrees of freedom.

For the results of this section, we require that the errors are normally distributed.

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

[6] For example, that used in Stata.

Assumption 2 $\boldsymbol{e}_g$ *is distributed* $\mathrm{N}\left(0, \boldsymbol{\Sigma}_g\right)$.

Assumption 2 states that the cluster error vectors $\boldsymbol{e}_g$ are normally distributed, but their within-cluster covariance structure is left unrestricted. 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.

We now evaluate the worst-case coverage probabilities of CRVE confidence intervals.

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 most important contribution of the paper, which provides a lower bound on the coverage rate of our recommended jackknife confidence interval. Let $F(x; k_1, k_2)$ denote the F distribution function with degrees of freedom k_1 and k_2.

Theorem 3 *Under Assumptions 1-2, for any* $1 \leq c < \infty$,

$$\inf_{(\boldsymbol{X},\boldsymbol{\Sigma})\in\mathscr{F}} \mathbb{P}\left[\theta \in \widehat{C}_5(c)\right] \geq F(c; 1, 1). \tag{24}$$

Equation (24) shows that the interval $\widehat{C}_5(c)$ has coverage probability which is uniformly bounded[7] away from zero in the broad model class $\mathscr{F}$. The lower bound is the F distribution with (1,1) degrees of freedom, which is the square of 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. This result holds over the broadest model class $\mathscr{F}$. This includes all regression models, including those with the most extreme leverage, within-cluster correlation, heteroskedasticity, and clusterwise noninvertibility. Thus the jackknife interval $\widehat{C}_5(c)$ is robust to these contexts.

We contrast (24) with the worst-case coverage of confidence intervals constructed with the other standard errors.

Theorem 4 *Under Assumptions 1-2, for any* $0 \leq c < \infty$,

$$\inf_{(\boldsymbol{X},\boldsymbol{\Sigma})\in\mathscr{F}_0^*} \mathbb{P}\left[\theta \in \widehat{C}_1(c)\right] = 0, \tag{25}$$

$$\inf_{(\boldsymbol{X},\boldsymbol{\Sigma})\in\mathscr{F}^*} \mathbb{P}\left[\theta \in \widehat{C}_2(c)\right] = 0, \tag{26}$$

$$\inf_{(\boldsymbol{X},\boldsymbol{\Sigma})\in\mathscr{F}_0} \mathbb{P}\left[\theta \in \widehat{C}_3(c)\right] = 0, \tag{27}$$

[7]The bound (28) requires $c \geq 1$. This is not an important restriction for inference as all conventional critical values exceed 1.

and

$$\inf_{(\boldsymbol{X},\boldsymbol{\Sigma})\in\mathscr{F}_0} \mathbb{P}\left[\theta \in \widehat{C}_4(c)\right] = 0. \tag{28}$$

Equation (25) shows that the worst-case coverage of the CRVE_1 confidence interval equals 0. This demonstrates that coverage can be arbitrarily distorted from the nominal level, and this holds for *any* critical value c. This means that it is impossible to uniformly[8] achieve any desired coverage probability. The set of models considered are those satisfying $\boldsymbol{\Sigma} = \boldsymbol{I}_n\sigma^2$ and clusterwise invertibility, so the result (25) 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).

Equation (26) shows a similar result for the CRVE_2 confidence interval in the class of models which allows general covariance matrices: that 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 (25) is that (26) requires the model class to included non-i.i.d. errors.

Equations (27) and (28) 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.

The results (25)-(28) 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. Indeed, the proof of Theorem 4 is a simple manipulation of Theorem 2. 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. It is also important to understand that the zero coverage rates of (25)-(28) 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 with normal errors. For example, with the conventional $c = 1.96$ critical value, the bound (24) is 0.70. Thus, the finite sample coverage of $\widehat{C}_5(1.96)$ 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. 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, the practical message of Theorems 3 and 4 is that conventional EHW and CRVE confidence intervals can have arbitrary coverage distortion, while the jackknife interval has bounded distortion.

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

This is a strong motivation for replacement of the EHW and CRVE standard errors by simple-to-calculate jackknife standard errors.

While the bound (24) may appear simple, its derivation is not. Our derivation relies on a characterization of the exact distribution of the t-ratio, the variance bound (23), convex optimization, and numerical calculation. The use of numerical calculation in the proof is theoretically inelegant; it is only needed for the clusterwise noninvertible case. Under the added assumption of clusterwise invertibility, the proof of Theorem 3 does not require a numerical argument.

6 Distribution of Jackknife t statistic

In this section we present a characterization of the finite sample distribution of the squared jackknife t-statistic

$$T_5^2 = \frac{\left(\widehat{\theta} - \theta\right)^2}{\widehat{v}_5^2}. \tag{29}$$

Let $F(x; k_1, k_2)$ denote the F distribution function with degrees of freedom k_1 and k_2.

Theorem 5 *Under Assumptions 1-2, $v^2 > 0$, and any $0 \leq x < \infty$,*

$$\mathbb{P}\left[T_5^2 \leq x\right] \geq \mathbb{P}\left[\frac{v^2 \xi_0^2}{\sum_{j=1}^G \phi_j \xi_j^2} \leq x\right] \simeq F\left(a^2 x; 1, K\right) \tag{30}$$

where ξ_j^2 are mutually independent χ_1^2 random variables, ϕ_j are the eigenvalues of the $G \times G$ matrix $\boldsymbol{L}$ defined in (94),

$$a = \sqrt{\frac{\mathrm{tr}\left[\boldsymbol{L}\right]}{v^2}}, \tag{31}$$

and

$$K = \frac{\left(\mathrm{tr}\left[\boldsymbol{L}\right]\right)^2}{\mathrm{tr}\left[\boldsymbol{L}\boldsymbol{L}\right]}. \tag{32}$$

The approximation in (30) is the Satterthwaite (1946) approximation to the weighted sum of chi-squares $\sum_{j=1}^G \phi_j \xi_j^2$.

Theorem 5 provides two distributional characterizations in (30). The first shows that the exact distribution of T_5^2 is bounded below by that of a ratio of independent weighted sums of chi-squares. The second shows that the latter is approximately F distributed with a non-standard scale and degree-of-freedom. Equivalently, this shows that the distribution of the t-statistic $T_5 = \left(\widehat{\theta} - \theta\right)/\widehat{v}_5$ is bounded by an approximate student t with scale a and degree-of-freedom K.

One new feature of Theorem 5 is the bound for the exact distribution by a ratio of *independent* weighted sums of chi-squares. It is well known that this representation holds for common t-statistics under i.i.d. errors, and is also widely known that a similar representation, but with correlated weighted sums of chi-squares, holds for non-i.i.d. errors. The fact that the representation (30) holds for independent χ_1^2 variables is new.

The constants a and K in (31)-(32) are written as functions of the eigenvalues of the $G \times G$ matrix $\boldsymbol{L}$ defined in the proof of Theorem 5. Of particular importance is the non-standard degree-of-freedom K. Examining (32) and observing that the matrix $\boldsymbol{L}$ is $G \times G$, we can deduce that K takes values in $[1, G]$. The case $K = G$ occurs when the matrix $\boldsymbol{L}$ is perfectly balanced so that all its eigenvalues are equal. The case $K = 1$ occurs when the matrix $\boldsymbol{L}$ is highly unbalanced with only one non-zero eigenvalue. Most applications will lie between these two extremes. Note in contrast that conventional software uses $G - 1$ degrees of freedom.

The constants a and K which determine the F distribution of Theorem 5 depend on the unknown variance matrices $\boldsymbol{\Sigma}_g$. If the $\boldsymbol{\Sigma}_g$ are known then a and K can be calculated and this F distribution used. However, in practice these variance matrices are unknown. Bell and McCaffery suggested calculating a and K based on a reference model, in particular, $\boldsymbol{\Sigma}_g = \boldsymbol{I}_{n_g}$ (which holds under i.i.d. errors). Other plausible reference models are $\boldsymbol{\Sigma}_g = \mathbf{1}_{n_g}\mathbf{1}_{n_g}'$ (holds under perfect within-cluster correlation) and $\boldsymbol{\Sigma}_g = \boldsymbol{X}_g\boldsymbol{X}_g'$ (strong conditional heteroskedasticity). We follow Bell and McCaffery and recommend calculation of a and K based on the simple reference model $\boldsymbol{\Sigma}_g = \boldsymbol{I}_{n_g}$. If desired, other reference models for $\boldsymbol{\Sigma}_g$ could be used.

We now present explicit expressions for the constants a and K under this reference model which do not rely on explicit computation of the matrix $\boldsymbol{L}$ nor its eigenvalues.

Theorem 6 *Computationally convenient expressions for the constants appearing in (31)-(32) for the case* $\boldsymbol{\Sigma}_g = \boldsymbol{I}_{n_g}$ *are*

$$\nu^2 = R'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}R, \tag{33}$$

$$\operatorname{tr}[\boldsymbol{L}] = \sum_{g=1}^{G}\boldsymbol{S}_g'\boldsymbol{S}_g - \frac{\boldsymbol{T}'\boldsymbol{T}}{\boldsymbol{Z}'\boldsymbol{Z}} + \operatorname{tr}\left[\boldsymbol{U}'\boldsymbol{U}\widehat{\boldsymbol{X}}'\widehat{\boldsymbol{X}}\right] - 2\operatorname{tr}\left[\boldsymbol{U}'\boldsymbol{V}\right], \tag{34}$$

and

$$\begin{aligned}\operatorname{tr}[\boldsymbol{L}\boldsymbol{L}] = {} & \sum_{g=1}^{G}\left(\boldsymbol{S}_g'\boldsymbol{S}_g\right)^2 - 2\frac{\sum_{g=1}^{G}\left(\boldsymbol{S}_g'\boldsymbol{S}_g\right)T_g^2}{\boldsymbol{Z}'\boldsymbol{Z}} + \left(\frac{\boldsymbol{T}'\boldsymbol{T}}{\boldsymbol{Z}'\boldsymbol{Z}}\right)^2 - 2\frac{\boldsymbol{T}'\boldsymbol{U}\widehat{\boldsymbol{X}}'\widehat{\boldsymbol{X}}\boldsymbol{U}'\boldsymbol{T}}{\boldsymbol{Z}'\boldsymbol{Z}} \\ & + \operatorname{tr}\left[\widehat{\boldsymbol{X}}'\widehat{\boldsymbol{X}}\boldsymbol{U}'\boldsymbol{U}\widehat{\boldsymbol{X}}'\widehat{\boldsymbol{X}}\boldsymbol{U}'\boldsymbol{U}\right] + 2\operatorname{tr}\left[\boldsymbol{V}'\boldsymbol{U}\boldsymbol{V}'\boldsymbol{U}\right] + 2\operatorname{tr}\left[\boldsymbol{U}'\boldsymbol{W}\widehat{\boldsymbol{X}}'\widehat{\boldsymbol{X}}\right] \\ & - 4\operatorname{tr}\left[\boldsymbol{V}'\boldsymbol{W}\right] + 4\frac{\boldsymbol{T}'\boldsymbol{V}\boldsymbol{U}'\boldsymbol{T}}{\boldsymbol{Z}'\boldsymbol{Z}} - 4\operatorname{tr}\left[\boldsymbol{U}'\boldsymbol{U}\widehat{\boldsymbol{X}}'\widehat{\boldsymbol{X}}\boldsymbol{U}'\boldsymbol{V}\right] + 2\operatorname{tr}\left[\boldsymbol{U}'\boldsymbol{U}\boldsymbol{V}'\boldsymbol{V}\right],\end{aligned} \tag{35}$$

where

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

$$
\boldsymbol{U} = \begin{bmatrix} \boldsymbol{U}_1' \\ \vdots \\ \boldsymbol{U}_G' \end{bmatrix}, \qquad \boldsymbol{V} = \begin{bmatrix} \boldsymbol{V}_1' \\ \vdots \\ \boldsymbol{V}_G' \end{bmatrix}, \qquad \boldsymbol{W} = \begin{bmatrix} \boldsymbol{W}_1' \\ \vdots \\ \boldsymbol{W}_G' \end{bmatrix}, \qquad \boldsymbol{S} = \begin{bmatrix} \boldsymbol{S}_1 \\ \vdots \\ \boldsymbol{S}_G \end{bmatrix}, \qquad \boldsymbol{T} = \begin{bmatrix} T_1 \\ \vdots \\ T_G \end{bmatrix}.
$$

The expression (34) and (35) in Theorem 6 look imposing, but they are computationally simple to implement. All of the matrices for which the trace operator are applied are $k \times k$. The expressions (34)-(35) are functions of (and only of) the regressors $\boldsymbol{X}$ and vector R.

7 Adjusted Confidence Intervals

Simulation evidence shows that while our jackknife confidence interval using conventional critical values has improved coverage rates relative to the CRVE_1 and CRVE_2 intervals, it can still under-cover. The source of the problem is the conventional student t distributional approximation.

A promising class of distributional adjustments was introduced by Bell and McCaffery (2002) for HC_2 and CRVE_2 t-ratios. They are simple to calculate and produce greatly improved coverage rates in finite samples. These methods have been endorsed by other authors, including Imbens and Kolesár (2016). In this section, we extend the Bell-McCaffery adjustment to jackknife t tests and confidence intervals. Our extension involves both scale and degree-of-freedom adjustments.

Our Bell-McCaffery adjustment is based on the approximation of Theorem 5, which states that the finite sample distribution of the squared t-statistic is approximately F distributed (or the non-squared t-statistic is approximately student t distributed) with a scale adjustment and a non-standard degree-of-freedom. The original Bell-McCaffery adjustment (for HC_2 and CRVE_2 t-ratios) does not have a scale adjustment, because the CRVE_2 variance estimator is unbiased under i.i.d. errors. In contrast, our adjustment for the jackknife statistic has a scale adjustment to account for jackknife variance estimation bias.

As discussed in the previous section, these approximations need to be calculated under a reference model for the unknown variances, and following Bell and McCaffery we recommend $\boldsymbol{\Sigma}_g = \boldsymbol{I}_{n_g}$. In this context, efficient formula for the inputs for the constants a and K are given in expressions (33)-(35).

The adjusted $100(1-\alpha)\%$ confidence interval for θ is

$$\widetilde{C}_5 = \widehat{\theta} \pm \frac{t_K^{1-\alpha/2}\widehat{v}_5}{a} \tag{36}$$

where $t_K^{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the student t distribution with K degrees of freedom and $\widehat{v}_5$ is our recommended jackknife standard error. Equivalently, $t_K^{1-\alpha/2} = \sqrt{q_{1,K}^{1-\alpha}}$, where $q_{1,K}^{1-\alpha}$ is the $1-\alpha$ quantile of the $F(x;1,K)$ distribution.

Specifically, the confidence interval is formed by the following steps. Take any linear coefficient $\theta = R'\beta$.

1. Calculate the jackknife variance $\widehat{\boldsymbol{V}}_5$ from (11) and standard error $\widehat{v}_5 = \sqrt{R'\widehat{\boldsymbol{V}}_5 R}$.
2. Calculate the adjustment parameters K and a from (34) and (35).
3. Calculate the critical value $t_K^{1-\alpha/2}$ from the student t distribution with the degree-of-freedom K.
4. Set $\widehat{\theta} = R'\widehat{\beta}$ and $\widetilde{C}_5$ as in (36).

The adjustment parameters K and a are coefficient-specific, so the above steps need to be repeated for each element of β in order to calculate confidence intervals for all individual coefficients.

The adjusted degree-of-freedom K will typically be smaller than the conventional degree-of-freedom, thereby increasing the critical value and the width of the confidence interval. This adjustment accounts for the non-standard distribution of the variance estimator. In contrast, the adjustment factor a is typically greater than 1, so the scale adjustment decreases the width of the confidence interval. This counters the bias of the jackknife variance estimator.

Similarly, the adjusted p-value for a test of $\theta = \theta_0$ is

$$p = 1 - F\left(a^2\left(\frac{\widehat{\theta}-\theta_0}{\widehat{v}_5}\right)^2; 1, K\right) \tag{37}$$

where $F(x;1,K)$ is the F distribution with degrees of freedom $(1,K)$. As for the confidence intervals, this makes both a degree-of-freedom adjustment (through K) and a scale adjustment (through a).

The theoretical justification for the confidence interval (36) is as follows. Using the equalities $\widetilde{C}_5 = \widehat{C}_5\left(t_K^{1-\alpha/2}/a\right)$, $t_K^{1-\alpha/2} = \sqrt{q_{1,K}^{1-\alpha}}$, then Theorem 5, we find

$$\begin{aligned}
\mathbb{P}\left[\theta \in \widetilde{C}_5\right] &= \mathbb{P}\left[\theta \in \widehat{C}_5\left(t_K^{1-\alpha/2}/a\right)\right] \\
&= \mathbb{P}\left[T_5^2 \le q_{1,K}^{1-\alpha}/a^2\right] \\
&\gtrsim F\left(q_{1,K}^{1-\alpha}; 1, K\right) \\
&= 1-\alpha.
\end{aligned} \tag{38}$$

The inequality (38) combines the inequality and approximation of Theorem 5, and holds under the reference model $\boldsymbol{\Sigma} = \boldsymbol{I}_n\sigma^2$.

This shows that, approximately, $\widetilde{C}_5$ will have coverage close to or exceeding the desired level. There are three sources of approximation error: (1) the inequality bound in (30); (2) the Satterthwaite approximation in (30); (3) the reference model approximation. Each may contribute to a distortion of actual from nominal coverage.

8 Fixed Effects

It is common in panel/clustering contexts to include cluster-specific fixed effects[9]. 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 fixed effect dummy variables are present in the regressor matrix $\boldsymbol{X}$, the traditional jackknife variances estimators $\widehat{\boldsymbol{V}}_3$ and $\widehat{\boldsymbol{V}}_4$ are undefined. In contrast, our jackknife variance estimator $\widehat{\boldsymbol{V}}_5$ is well-defined, and conservative by Theorem 1 for all the coefficients including the estimated fixed effects.

Regardless, our recommendation[10] is to first apply the within transformation and then apply least squares estimation to the within transformed variables, to obtain the coefficient estimates and jackknife standard errors. This is both computationally and theoretically preferred. It is theoretically preferred because the model after the within transformation is clusterwise invertible, so has reduced distributional distortions. In contrast, the model with the included fixed effect dummy variables is clusterwise noninvertible, and while our finite sample distribution theory applies, the inequalities suggest that the confidence intervals will be conservative.

9 Simulation Evidence

We present a simulation experiment to investigate the performance of the methods. 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 $\boldsymbol{Y}_g = \alpha + \boldsymbol{X}_g\beta + \boldsymbol{e}_g$ with $\boldsymbol{X}_g$ a single $n_g \times 1$ stochastic regressor. Within this model we consider six designs which vary the distributions of the regressors $\boldsymbol{X}_g$, the heteroskedasticity of the equation errors, and the cluster size heterogeneity. Let $\boldsymbol{I}_g$ denote the $n_g \times n_g$ identity matrix, $\mathbf{1}_g = (1, 1, , ...)$ 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: $\boldsymbol{X}_g \sim \mathrm{N}\left(\mathbf{1}_g, \boldsymbol{I}_g + \mathbf{1}_g\mathbf{1}_g'\right)$.

[9] We focus on the case where the fixed effects are included for the same clusters as used

[10] These comments apply to the case where the fixed effects are applied at the same level as clustering, and not when fixed effects are applied at a different or more aggregate level.

2. LogNormal with Clustered Dependence: $\boldsymbol{X}_g \sim \exp\left(\mathrm{N}\left(0, \boldsymbol{I}_g + \mathbf{1}_g\mathbf{1}_g'\right)\right)$, recentered and rescaled so the elements have a mean of 1 and variance of 2.

In the first design, the regressors are normally distributed with 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 both models the regressors are centered and scaled so to have the same first and second moments.

The designs for the errors are:

1. Homoskedastic with Clustered Dependence: $\boldsymbol{e}_g \sim \mathrm{N}\left(0, \boldsymbol{I}_g + \mathbf{1}_g\mathbf{1}_g' + \boldsymbol{h}_g\boldsymbol{h}_g'\right)$.
2. Heteroskedastic: $\boldsymbol{e}_g \sim \mathrm{N}\left(0, \boldsymbol{I}_g + \left(\boldsymbol{X}_g - \mathbf{1}_g\right)\left(\boldsymbol{X}_g - \mathbf{1}_g\right)'\right)$.

In the first design, the errors are normally distributed with a cluster-level factor covariance matrix. In the second design, they are conditionally heteroskedastic. The scalings are set so that the unconditional variances are the same in both designs.

The designs for the cluster sizes are

1. Homogeneous: $n_g = 10$ for $g = 1, ..., G$.
2. Heterogeneous: $n_1 = n_2 = 5G/2 + 5$, and $n_g = 5$ for $g = 3, ..., G$.

The homogeneous cluster size design sets all clusters equal to 10. The heterogeneous cluster size design puts between 1/4 and 1/3 of the observations in each of the first and second clusters, with the remaining spread among the other clusters. There are an average of 10 observations per cluster to match the homogeneous cluster size design.

The six designs combine these elements as described in Table 1. For each design, the number of clusters G is varied among $\{6, 12, 40, 100\}$. As there are an average of 10 observations per cluster, the associated total sample sizes are $\{60, 120, 400, 1000\}$.

Table 1: Simulation Designs

	Regressor	Error	Cluster Size
Design 1	Normal	Homoskedastic	Homogeneous
Design 2	LogNormal	Homoskedastic	Homogeneous
Design 3	LogNormal	Homoskedastic	Heterogeneous
Design 4	Normal	Heteroskedastic	Homogeneous
Design 5	LogNormal	Heteroskedastic	Homogeneous
Design 6	LogNormal	Heteroskedastic	Heterogeneous

For each simulation replication we estimate the coefficients by least squares. We use six methods to calculate standard errors: the five variances estimators described in the text: CRVE_1 ($\widehat{v}_1$), CRVE_2 ($\widehat{v}_2$),

the two conventional jackknife estimators $\widehat{v}_3$ and $\widehat{v}_4$, our recommended jackknife estimator $\widehat{v}_5$, and the nonparametric pairs cluster bootstrap[11] which we denote as $\widehat{v}_6$.

We calculate confidence intervals for β using twelve methods.

The first six are conventional, based on the six 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 two intervals are adjusted t intervals. The first is that recommended by Bell and McCaffrey (2002) and equals $\widehat{\beta} \pm t_K^{0.975}\widehat{v}_2$ where K is calculated[12] similar to (32) under the reference model $\boldsymbol{\Sigma}_g = \boldsymbol{I}_{n_g}$. The second interval is our adjusted interval (36) using the jackknife standard error.

The next two intervals are nonparametric pairs cluster bootstrap symmetric percentile-t intervals, using the standard errors $\widehat{v}_1$ and $\widehat{v}_5$, and 1000 bootstrap replications. The bootstrap samples are constructed by nonparametric pairs resampling, as described earlier. On each bootstrap sample we calculate the least squares estimate $\widehat{\beta}^*$ and the CRVE_1 and jackknife standard errors $\widehat{v}_1^*$ and $\widehat{v}_5^*$. From the 1000 bootstrap samples we calculate the 95% quantiles $\widehat{c}_1^*$ and $\widehat{c}_5^*$ of the statistics $\left|\widehat{\beta}^* - \widehat{\beta}\right| / \widehat{v}_1^*$ and $\left|\widehat{\beta}^* - \widehat{\beta}\right| / \widehat{v}_5^*$. The confidence intervals are $\widehat{\beta} \pm \widehat{c}_1^* \widehat{v}_1$ and $\widehat{\beta} \pm \widehat{c}_5^* \widehat{v}_5$.

The final two intervals are wild cluster bootstrap symmetric percentile-t intervals, using the standard errors $\widehat{v}_1$ and $\widehat{v}_5$, and 1000 bootstrap replications. This is the method proposed by Cameron, Gelbach, and Miller (2008), Djogbenou, MacKinnon, and Nielsen (2019), and Canay, Santos, and Shaikh (2021) for hypothesis testing, and can be used to construct a confidence interval by test inversion[13]. First, the coefficients are re-estimated imposing known β to obtain restricted estimates $\left(\widetilde{\alpha}, \widetilde{\beta} = \beta\right)$ and residuals $\widetilde{\boldsymbol{e}}_g$. Next, the clusters, regressors $\boldsymbol{X}_g$, and restricted residuals $\widetilde{\boldsymbol{e}}_g$ are held fixed. The bootstrap errors are generated as $\boldsymbol{e}_g^* = \zeta_g \widetilde{\boldsymbol{e}}_g$ where ζ_g is an independent Rademacher variable (equals +1 and −1 each with probability 1/2), and $\boldsymbol{Y}_g^* = \widetilde{\alpha} + \boldsymbol{X}_g\widetilde{\beta} + \boldsymbol{e}_g^*$. The bootstrap sample then consists of the observations $(\boldsymbol{Y}_g^*, \boldsymbol{X}_g)$. On each bootstrap sample we calculate the least squares estimate $\widehat{\beta}^*$ and the standard errors $\widehat{v}_1^*$ and $\widehat{v}_5^*$. From the 1000 bootstrap samples we calculate the 95% quantiles $\widehat{c}_1^*(\beta)$ and $\widehat{c}_5^*(\beta)$ of the the statistics $\left|\widehat{\beta}^* - \beta\right| / \widehat{v}_1^*$ and $\left|\widehat{\beta}^* - \beta\right| / \widehat{v}_5^*$. The confidence intervals[14] consist of all β such that $\left|\widehat{\beta} - \beta\right| / \widehat{v}_1 \leq \widehat{c}_1^*(\beta)$ and $\left|\widehat{\beta} - \beta\right| / \widehat{v}_5 \leq \widehat{c}_5^*(\beta)$, respectively.

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 report the results for the baseline regression model in Table 2. The top block reports the results for $G = 6$. The first six columns are for the conventional confidence intervals. We can see that the conventional CRVE_1 confidence interval has substantial under-coverage in most designs. The worst-case is

[11] This is the standard implementation of the bootstrap for clustered observations. Each bootstrap sample is constructed by resampling G clusters $(\boldsymbol{Y}_g, \boldsymbol{X}_g)$ with replacement from the original sample of clusters. Least squares estimation is applied to the bootstrap sample. The bootstrap standard error is the empirical standard deviation of the bootstrap least squares estimates. 1000 bootstrap replications were made in each simulation replication. For details see Cameron, Gelbach, and Miller (2008).

[12] The CRVE_2 standard error $\widehat{v}_2$ and Bell-McCaffrey adjusted interval are calculated using the `dfadjust` R package of Kolesár (2023), which is identical to the implementation in Stata 18.

[13] MacKinnon, Nielsen and Webb (2023b) review several variants of the wild cluster bootstrap. Our implementation corresponds to their WCR-V method.

[14] To assess the coverage rate, it is sufficient to do the calculation for the true value of β.

Table 2: Baseline Regression Model. Coverage of Nominal 95% Confidence Intervals for β

Cr. Value	Conventional t_{G-1}						Adjusted t_K		Pairs Boot		Wild Boot	
St. Error	$\widehat{v}_1$	$\widehat{v}_2$	$\widehat{v}_3$	$\widehat{v}_4$	$\widehat{v}_5$	$\widehat{v}_6$	$\widehat{v}_2$	$\widehat{v}_5$	$\widehat{v}_1$	$\widehat{v}_5$	$\widehat{v}_1$	$\widehat{v}_5$
$G = 6$												
Design 1	0.91	0.93	0.95	0.95	0.96	0.94	0.95	0.96	0.96	0.96	0.93	0.94
Design 2	0.85	0.91	0.95	0.95	0.96	0.98	0.99	0.99	0.93	0.96	0.96	0.95
Design 3	0.83	0.90	0.95	0.95	0.96	0.99	0.99	0.99	0.93	0.97	0.95	0.95
Design 4	0.89	0.91	0.93	0.93	0.95	0.91	0.94	0.95	0.95	0.96	0.93	0.94
Design 5	0.64	0.74	0.88	0.89	0.90	0.92	0.93	0.95	0.80	0.92	0.91	0.94
Design 6	0.61	0.72	0.88	0.88	0.90	0.92	0.92	0.94	0.81	0.93	0.91	0.94
$G = 12$												
Design 1	0.92	0.93	0.95	0.95	0.95	0.93	0.95	0.95	0.95	0.95	0.95	0.95
Design 2	0.83	0.89	0.94	0.94	0.94	0.97	0.98	0.99	0.89	0.93	0.96	0.96
Design 3	0.80	0.88	0.94	0.94	0.95	0.98	0.99	0.99	0.90	0.95	0.95	0.95
Design 4	0.91	0.92	0.93	0.93	0.94	0.91	0.94	0.94	0.95	0.95	0.94	0.95
Design 5	0.63	0.74	0.87	0.87	0.88	0.86	0.93	0.95	0.82	0.92	0.92	0.93
Design 6	0.59	0.71	0.87	0.87	0.88	0.86	0.93	0.95	0.81	0.92	0.92	0.94
$G = 40$												
Design 1	0.94	0.94	0.95	0.95	0.95	0.94	0.95	0.95	0.95	0.95	0.95	0.95
Design 2	0.86	0.90	0.93	0.93	0.93	0.95	0.97	0.98	0.89	0.92	0.95	0.96
Design 3	0.81	0.88	0.94	0.94	0.94	0.95	0.98	0.99	0.90	0.95	0.94	0.96
Design 4	0.93	0.94	0.94	0.94	0.95	0.93	0.94	0.94	0.95	0.95	0.95	0.95
Design 5	0.70	0.79	0.88	0.88	0.89	0.84	0.93	0.95	0.90	0.94	0.94	0.94
Design 6	0.64	0.75	0.87	0.87	0.88	0.83	0.93	0.95	0.87	0.94	0.93	0.94
$G = 100$												
Design 1	0.95	0.95	0.95	0.95	0.95	0.94	0.95	0.95	0.95	0.95	0.95	0.95
Design 2	0.89	0.91	0.93	0.93	0.93	0.95	0.97	0.97	0.91	0.93	0.95	0.95
Design 3	0.81	0.88	0.94	0.94	0.95	0.92	0.97	0.98	0.92	0.96	0.94	0.96
Design 4	0.94	0.94	0.95	0.95	0.95	0.94	0.95	0.95	0.95	0.95	0.95	0.95
Design 5	0.77	0.84	0.90	0.90	0.90	0.85	0.93	0.95	0.92	0.95	0.95	0.95
Design 6	0.69	0.78	0.89	0.89	0.89	0.83	0.93	0.95	0.89	0.94	0.94	0.95

Design 6, where the interval has only 61% coverage. The CRVE_2 confidence interval has slightly better coverage, but still substantially under-covers in most designs. The jackknife and bootstrap confidence intervals have better coverge, but undercover in Designs 5 & 6.

The next two columns are for the confidence intervals using adjusted critical values. The adjusted CRVE_2 interval has improved coverage over the conventional CRVE_2 interval, but slightly undercovers (92%) for Designs 5 and 6. The adjusted jackknife interval has excellent coverage, with the coverage probability exceeding 94% in all designs.

The results for the pairs bootstrap-t methods are reported in the next two columns. The intervals based on the CRVE_1 standard error generally undercover (and severely for Designs 5 & 6). The intervals based on the jackknife standard errors have much better coverage, with coverage exceeding 92% in all designs.

Table 3: Regression with Treatment Dummy. Coverage of Nominal 95% Confidence Intervals for β

Cr. Value	Conventional t_{G-1}						Adjusted t_K		Pairs Boot		Wild Boot	
St. Error	$\widehat{v}_1$	$\widehat{v}_2$	$\widehat{v}_3$	$\widehat{v}_4$	$\widehat{v}_5$	$\widehat{v}_6$	$\widehat{v}_2$	$\widehat{v}_5$	$\widehat{v}_1$	$\widehat{v}_5$	$\widehat{v}_1$	$\widehat{v}_5$
$G = 6$												
Design 1	0.91	0.93	0.94	0.94	0.96	0.94	0.95	0.95	0.95	0.96	0.93	0.94
Design 2	0.85	0.91	0.92	0.93	0.97	0.93	0.99	0.99	0.92	0.96	0.96	0.96
Design 3	0.83	0.90	0.87	0.89	0.96	0.87	0.99	0.98	0.90	0.95	0.95	0.95
Design 4	0.89	0.91	0.91	0.92	0.95	0.90	0.94	0.93	0.93	0.94	0.93	0.94
Design 5	0.64	0.74	0.81	0.83	0.90	0.84	0.93	0.94	0.81	0.92	0.92	0.94
Design 6	0.61	0.72	0.74	0.76	0.89	0.74	0.92	0.93	0.79	0.91	0.92	0.94
$G = 12$												
Design 1	0.92	0.93	0.94	0.94	0.96	0.93	0.95	0.95	0.95	0.95	0.95	0.95
Design 2	0.83	0.89	0.92	0.93	0.94	0.95	0.99	0.99	0.89	0.93	0.96	0.96
Design 3	0.82	0.89	0.89	0.90	0.95	0.89	0.99	0.99	0.88	0.93	0.95	0.96
Design 4	0.91	0.92	0.93	0.93	0.94	0.91	0.94	0.94	0.94	0.95	0.94	0.94
Design 5	0.63	0.74	0.84	0.85	0.88	0.83	0.93	0.95	0.81	0.91	0.92	0.93
Design 6	0.59	0.71	0.76	0.77	0.87	0.73	0.93	0.94	0.79	0.90	0.92	0.94
$G = 40$												
Design 1	0.94	0.94	0.95	0.95	0.95	0.94	0.95	0.95	0.95	0.95	0.95	0.95
Design 2	0.86	0.90	0.93	0.93	0.93	0.95	0.97	0.98	0.89	0.92	0.95	0.96
Design 3	0.82	0.89	0.91	0.91	0.94	0.89	0.98	0.98	0.88	0.93	0.94	0.96
Design 4	0.93	0.94	0.94	0.94	0.95	0.93	0.94	0.94	0.95	0.95	0.95	0.95
Design 5	0.70	0.79	0.88	0.88	0.89	0.83	0.93	0.95	0.90	0.94	0.94	0.94
Design 6	0.64	0.75	0.79	0.79	0.88	0.73	0.93	0.95	0.84	0.92	0.93	0.94
$G = 100$												
Design 1	0.95	0.95	0.95	0.95	0.95	0.94	0.95	0.95	0.95	0.95	0.95	0.95
Design 2	0.89	0.91	0.93	0.93	0.94	0.95	0.97	0.97	0.91	0.93	0.95	0.95
Design 3	0.83	0.89	0.92	0.92	0.94	0.89	0.97	0.98	0.89	0.94	0.93	0.96
Design 4	0.94	0.94	0.95	0.95	0.95	0.94	0.95	0.95	0.95	0.95	0.95	0.95
Design 5	0.77	0.84	0.90	0.90	0.90	0.85	0.93	0.95	0.92	0.95	0.95	0.95
Design 6	0.69	0.78	0.80	0.80	0.89	0.74	0.93	0.95	0.86	0.92	0.94	0.95

The results for the wild bootstrap-t are presented in the final two columns. The coverage rates are excellent when the jackknife standard errors are used.

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. A notable exception is the conventional interval with bootstrap standard errors whose coverage probabilities worsen as G increases in Designs 3, 5, and 6. In general, the best performance is obtained by our adjusted jackknife interval and the wild bootstrap using jackknife standard errors. A close competitor is the pairs bootstrap percentile-t using jackknife standard errors. No other method has reliable coverage across the designs and the four sample sizes.

If we compare performance by standard error method, within each type of inference method (conventional, adjusted, pairs bootstrap-t, and wild bootstrap-t), we systematically see that methods based

Table 4: Regression with Treatment Dummy. Coverage of Nominal 95% Confidence Intervals for γ

Cr. Value	Conventional t_{G-1}						Adjusted t_K		Pairs Boot		Wild Boot	
St. Error	$\widehat{\nu}_1$	$\widehat{\nu}_2$	$\widehat{\nu}_3$	$\widehat{\nu}_4$	$\widehat{\nu}_5$	$\widehat{\nu}_6$	$\widehat{\nu}_2$	$\widehat{\nu}_5$	$\widehat{\nu}_1$	$\widehat{\nu}_5$	$\widehat{\nu}_1$	$\widehat{\nu}_5$
$G = 6$												
Design 1	0.64	0.66	0.68	0.68	1.00	0.66	0.73	1.00	0.83	1.00	1.00	1.00
Design 2	0.64	0.66	0.66	0.67	1.00	0.64	0.72	1.00	0.85	1.00	1.00	1.00
Design 3	0.59	0.64	0.67	0.68	1.00	0.66	0.76	1.00	0.82	1.00	0.98	0.98
Design 4	0.71	0.75	0.77	0.77	1.00	0.75	0.82	1.00	0.83	0.99	1.00	1.00
Design 5	0.69	0.72	0.74	0.75	1.00	0.72	0.76	1.00	0.87	1.00	1.00	1.00
Design 6	0.72	0.77	0.79	0.80	1.00	0.79	0.86	1.00	0.87	1.00	0.99	0.99
$G = 12$												
Design 1	0.49	0.50	0.51	0.51	1.00	0.49	0.52	1.00	0.57	1.00	1.00	1.00
Design 2	0.47	0.48	0.48	0.49	1.00	0.47	0.51	1.00	0.53	1.00	1.00	1.00
Design 3	0.47	0.51	0.55	0.56	1.00	0.50	0.59	1.00	0.59	1.00	0.99	0.98
Design 4	0.61	0.64	0.66	0.66	1.00	0.61	0.66	1.00	0.73	1.00	1.00	1.00
Design 5	0.54	0.58	0.62	0.62	1.00	0.56	0.60	1.00	0.66	1.00	1.00	1.00
Design 6	0.64	0.70	0.74	0.75	1.00	0.70	0.77	1.00	0.76	1.00	0.99	0.99
$G = 40$												
Design 1	0.28	0.28	0.29	0.29	1.00	0.28	0.29	1.00	0.29	1.00	1.00	1.00
Design 2	0.26	0.27	0.27	0.27	1.00	0.26	0.27	1.00	0.27	1.00	1.00	1.00
Design 3	0.33	0.38	0.43	0.43	1.00	0.35	0.43	1.00	0.48	1.00	0.99	0.99
Design 4	0.44	0.45	0.45	0.45	1.00	0.43	0.45	1.00	0.48	1.00	1.00	1.00
Design 5	0.36	0.40	0.44	0.44	1.00	0.39	0.41	1.00	0.47	1.00	1.00	1.00
Design 6	0.56	0.62	0.67	0.68	1.00	0.60	0.68	1.00	0.69	1.00	1.00	1.00
$G = 100$												
Design 1	0.17	0.17	0.17	0.17	1.00	0.17	0.17	1.00	0.18	1.00	1.00	1.00
Design 2	0.16	0.16	0.16	0.16	1.00	0.16	0.16	1.00	0.16	1.00	1.00	1.00
Design 3	0.28	0.33	0.38	0.38	1.00	0.29	0.37	1.00	0.47	1.00	0.99	0.98
Design 4	0.30	0.31	0.31	0.31	1.00	0.30	0.31	1.00	0.32	1.00	1.00	1.00
Design 5	0.29	0.32	0.36	0.36	1.00	0.31	0.32	1.00	0.39	1.00	1.00	1.00
Design 6	0.54	0.60	0.64	0.64	1.00	0.54	0.65	1.00	0.69	1.00	1.00	1.00

on jackknife standard errors perform better than any other standard error method. Specifically, this holds whether inference uses conventional student t critical values, adjusted critical values, pairs bootstrap-t critical values, or wild bootstrap-t critical values. This is strong evidence favoring jackknife standard errors, regardless of the inference method.

We expand the analysis by examining a model with clusterwise noninvertibility. This is $\boldsymbol{Y}_g = \alpha + \boldsymbol{X}_g\beta + \boldsymbol{D}_g\gamma + \boldsymbol{e}_g$ 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, errors, and cluster sizes.

At this point we need to discuss our implementation of the pairs bootstrap. As $\boldsymbol{D}_g$ is non-zero only

Table 5: Regression with Strongly Skewed Errors. Coverage of Nominal 95% Confidence Intervals for β

Cr. Value	Conventional t_{G-1}						Adjusted t_K		Pairs Boot		Wild Boot	
St. Error	$\widehat{v}_1$	$\widehat{v}_2$	$\widehat{v}_3$	$\widehat{v}_4$	$\widehat{v}_5$	$\widehat{v}_6$	$\widehat{v}_2$	$\widehat{v}_5$	$\widehat{v}_1$	$\widehat{v}_5$	$\widehat{v}_1$	$\widehat{v}_5$
$G = 6$												
Design 1	0.92	0.94	0.96	0.96	0.97	0.95	0.96	0.97	0.96	0.97	0.94	0.94
Design 2	0.86	0.92	0.96	0.96	0.97	0.99	0.99	0.99	0.93	0.97	0.96	0.95
Design 3	0.84	0.90	0.96	0.96	0.97	0.99	0.99	0.99	0.93	0.97	0.94	0.95
Design 4	0.86	0.88	0.90	0.90	0.92	0.88	0.91	0.92	0.93	0.94	0.91	0.91
Design 5	0.64	0.74	0.87	0.87	0.89	0.91	0.92	0.94	0.79	0.91	0.91	0.93
Design 6	0.61	0.72	0.87	0.88	0.89	0.93	0.92	0.93	0.80	0.92	0.91	0.94
$G = 12$												
Design 1	0.93	0.94	0.95	0.95	0.96	0.94	0.95	0.96	0.96	0.96	0.95	0.95
Design 2	0.85	0.91	0.94	0.94	0.95	0.97	0.99	0.99	0.90	0.94	0.96	0.96
Design 3	0.81	0.89	0.94	0.95	0.95	0.98	0.99	0.99	0.90	0.95	0.95	0.95
Design 4	0.88	0.90	0.91	0.91	0.92	0.89	0.91	0.92	0.93	0.93	0.92	0.92
Design 5	0.62	0.72	0.85	0.85	0.86	0.85	0.92	0.94	0.80	0.90	0.91	0.93
Design 6	0.58	0.69	0.84	0.85	0.86	0.85	0.91	0.94	0.79	0.90	0.91	0.93
$G = 40$												
Design 1	0.94	0.95	0.95	0.95	0.95	0.94	0.95	0.95	0.96	0.96	0.95	0.95
Design 2	0.87	0.91	0.94	0.94	0.94	0.96	0.98	0.98	0.90	0.93	0.96	0.96
Design 3	0.82	0.89	0.94	0.94	0.94	0.95	0.98	0.99	0.91	0.95	0.95	0.96
Design 4	0.92	0.92	0.93	0.93	0.93	0.91	0.93	0.93	0.94	0.94	0.93	0.93
Design 5	0.68	0.77	0.85	0.85	0.86	0.81	0.91	0.93	0.86	0.91	0.92	0.92
Design 6	0.62	0.72	0.84	0.84	0.85	0.80	0.90	0.93	0.84	0.91	0.91	0.92
$G = 100$												
Design 1	0.94	0.95	0.95	0.95	0.95	0.94	0.95	0.95	0.95	0.95	0.95	0.95
Design 2	0.89	0.92	0.93	0.93	0.94	0.95	0.97	0.97	0.91	0.93	0.95	0.95
Design 3	0.82	0.89	0.94	0.94	0.94	0.93	0.97	0.98	0.91	0.95	0.94	0.95
Design 4	0.94	0.94	0.94	0.94	0.94	0.93	0.94	0.94	0.95	0.95	0.94	0.94
Design 5	0.74	0.81	0.87	0.87	0.87	0.83	0.91	0.93	0.89	0.93	0.92	0.92
Design 6	0.66	0.75	0.86	0.86	0.86	0.80	0.90	0.93	0.86	0.92	0.91	0.92

for one cluster, a high percentage (about 37%) of bootstrap samples have a singular regressor matrix. In this context it is not clear how the bootstrap should treat these sample draws. We follow the recommendation of Shao and Tu (1995) and implementation in Stata which discards these bootstrap samples. Thus bootstrap standard errors and critical values are calculated from the bootstrap samples which have nonsingular regressor matrices.

Table 3 presents the coverage rates for confidence intervals for β. In general, the results are similar to those of Table 2 but with two important differences. First, in Table 3 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 errors $\widehat{v}_5$. In Table 2, these three methods were nearly identical; in Table 3 we can see that the two conventional intervals exhibit substantial undercoverage in most designs (which diminish as G increases). This difference is due to the treatment of noninvert-

Table 6: Baseline Regression. Average Interval Length

Cr. Value	Conventional t_{G-1}						Adjusted t_K		Pairs Boot	
St. Error	$\widehat{v}_1$	$\widehat{v}_2$	$\widehat{v}_3$	$\widehat{v}_4$	$\widehat{v}_5$	$\widehat{v}_6$	$\widehat{v}_2$	$\widehat{v}_5$	$\widehat{v}_1$	$\widehat{v}_5$
$G = 6$										
Design 1	1.00	1.12	1.28	1.29	1.41	1.05	1.26	1.43	1.85	2.08
Design 2	1.69	2.09	3.17	3.28	3.60	3.32	3.96	5.30	3.15	5.65
Design 3	1.71	2.18	3.51	3.66	4.01	3.92	4.46	5.87	3.77	7.24
Design 4	0.94	1.04	1.17	1.17	1.29	0.97	1.18	1.31	1.91	2.25
Design 5	0.83	1.08	1.71	1.78	1.95	1.46	2.15	2.96	2.94	6.26
Design 6	0.83	1.10	1.83	1.91	2.09	1.65	2.35	3.15	3.50	7.60
$G = 12$										
Design 1	0.67	0.71	0.76	0.76	0.80	0.67	0.76	0.80	0.85	0.90
Design 2	0.86	1.02	1.37	1.39	1.45	1.40	1.84	2.39	1.22	1.91
Design 3	0.93	1.15	1.67	1.70	1.77	1.67	2.26	2.97	1.58	2.70
Design 4	0.66	0.70	0.76	0.76	0.79	0.66	0.76	0.79	0.92	1.01
Design 5	0.63	0.82	1.23	1.25	1.31	0.93	1.64	2.30	2.29	4.65
Design 6	0.63	0.85	1.33	1.35	1.42	1.01	1.81	2.50	2.70	5.62
$G = 40$										
Design 1	0.36	0.37	0.38	0.38	0.38	0.36	0.38	0.38	0.39	0.39
Design 2	0.37	0.41	0.49	0.49	0.50	0.48	0.61	0.75	0.47	0.61
Design 3	0.45	0.55	0.73	0.73	0.74	0.64	0.91	1.19	0.75	1.12
Design 4	0.38	0.39	0.40	0.40	0.41	0.38	0.40	0.41	0.43	0.44
Design 5	0.55	0.69	0.93	0.93	0.94	0.69	1.15	1.56	1.69	2.80
Design 6	0.56	0.72	1.04	1.05	1.06	0.76	1.32	1.85	1.98	3.66
$G = 100$										
Design 1	0.23	0.23	0.24	0.24	0.24	0.23	0.24	0.24	0.24	0.24
Design 2	0.22	0.23	0.26	0.26	0.26	0.26	0.30	0.35	0.26	0.30
Design 3	0.31	0.37	0.47	0.47	0.48	0.39	0.54	0.71	0.55	0.80
Design 4	0.25	0.25	0.25	0.25	0.26	0.25	0.25	0.26	0.26	0.26
Design 5	0.51	0.60	0.75	0.75	0.76	0.58	0.87	1.12	1.23	1.85
Design 6	0.51	0.65	0.89	0.89	0.90	0.65	1.07	1.48	1.57	2.79

ible clusters. Second, the conventional interval using bootstrap standard errors exhibits a substantial deterioration in coverage relative to the clusterwise invertible case of Table 2.

Next, we examine Table 4, which presents coverage rates for the dummy variable coefficient γ in the regression model $\boldsymbol{Y}_g = \alpha + \boldsymbol{X}_g\beta + \boldsymbol{D}_g\gamma + \boldsymbol{e}_g$. This is a treacherous context, as we are essentially making inference for a coefficient based on a single cluster. The results in Table 4 reveal that the conventional inference methods fail[15], and worsen as the sample size increases. The Bell-McCaffrey adjusted interval, conventional bootstrap interval, and pairs bootstrap percentile-t using CRVE_1 standard errors similarly fail. The exceptions (which generally produce coverage rates of 100%) are any intervals (conventional, adjusted, bootstrap percentile-t, and wild bootstrap-t) using the jackknife $\widehat{v}_5$ standard error, and both wild bootstrap percentile-t intervals. The reason why the intervals based on the jackknife standard error

[15]The failure of conventional methods in this context is well known.

$\widehat{v}_5$ 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. The message of Table 4 is that our recommended jackknife standard errors and confidence intervals will not produce misleading significance, even in the most extreme context of inference on a dummy indicator for a single treated cluster.

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 skewed heavy-tailed distribution. Specifically, we use the "strongly skewed" distribution displayed in Figure 3.7(b) of Hansen (2022) which is a 9-component normal mixture distribution with a skew of 1.34 and kurtosis of 6.7. 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 2 and 5, we can see that the results are qualitatively similar, with reduced coverage accuracy of several methods in Table 5. The relative rankings of the methods, however, are unchanged.

We finally address the accuracy of the confidence intervals by calculating their average length in the baseline model. We do so for all methods except for the wild bootstrap. This is because calculation of the full wild bootstrap confidence interval must be done by numerical test statistic inversion, which is computationally demanding[16]. We report the results in Table 6. The results are a bit difficult to compare across methods, as most methods have substantial undercoverage. However, we can make the following observations. First, if we compare the conventional methods (the first six), we can see that while our proposed jackknife-based intervals have the longest length, the six are of rough similar length, and this holds despite the fact that the other intervals have substantial under-coverage. Second, our proposed adjusted confidence interval has longer average length than the conventional methods, but again of rough similarity. Third, the relative differences decrease as the sample size increases. Fourth, the pairs bootstrap-t intervals can be substantially wider than our proposed intervals, despite having similar coverage probabilities. In general, the results of Table 6 suggest that the proposed methods have reasonable accuracy.

In summary, we are able to draw the following conclusions from the simulation evidence. First, the standard error which produces confidence intervals with the best coverage rates is $\widehat{v}_5$. Second, the simple confidence interval $\widehat{\beta} \pm t_{G-1}^{1-\alpha/2}\widehat{v}_5$ has good coverage rates in many contexts, but can under-cover in extreme designs. Third, the adjusted confidence interval $\widehat{\beta} \pm t_{K}^{1-\alpha/2}\widehat{v}_5/a$ has excellent (but conservative) coverage in all contexts examined. Fourth, excellent coverage is also attained by the pairs and wild bootstrap-t intervals, but only if combined with the jackknife standard error $\widehat{v}_5$. 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 implememented without the discarding of iterations with noninvertible design matrices.

[16]Our calculation of Table 6 using parellel processing took 3 days. Adding the wild bootstrap would increase the computation time by an order of magnitude.

10 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 provinces), 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 and confidence interval construction, in addition to 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 7 we report the least squares estimates $\widehat{\beta}_1 + \widehat{\beta}_2$ plus three standard errors: CRVE_1, CRVE_2, and our recommended jackknife standard errors. 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 7. The values for the adjusted degree-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 $0.141 \pm 2.73 \times 0.066/1.21 = [-.01, .29]$, as reported.

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 $1 - F\left(1.21^2 \times 2.16^2; 1, 4.18\right) = 0.058$, as reported. 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). 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.

Table 7: 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

Notes: All regressions include log total population, log urban population, and year fixed effects. CRVE_1, CRVE_2, and jackknife standard errors for $\widehat{\beta}_1+\widehat{\beta}_2$, clustered by province, in parenthesis, square brackets, and angle brackets, respectively. K and a are the adjustment parameters for the confidence interval for $\widehat{\beta}_1+\widehat{\beta}_2$. The p-value is for the t-test of the hypothesis $\beta_1+\beta_2=0$.

11 Conclusion

Heteroskedasticity-consistent and cluster-robust standard errors are routinely reported in applied econometric practice. It is prudent for the profession 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 rates. They are also computationally simple to implement.

12 Technical Proofs

Proof of Theorem 1: We first show that (14) holds for definitions (12) and (13). The estimator (12) with any generalized inverse is a minimizer of the least-squares criterion, so solves the first order condition

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

Pre-multiplying by $\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$, rearranging, and using (13), we obtain

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

which is (14) as claimed.

By definition (13) and model (1), the prediction errors equal

$$\widetilde{\boldsymbol{e}}_g = \boldsymbol{Y}_g - \boldsymbol{X}_g\widetilde{\beta}_{-g} = \boldsymbol{e}_g - \boldsymbol{X}_g\left(\widetilde{\beta}_{-g} - \beta\right). \tag{39}$$

Squaring and expanding,

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

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

The inequality holds because the final term in (40) takes the form $\boldsymbol{A}\boldsymbol{A}'$ and is thus positive semi-definite.

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

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

Using expression (15) and inequality (42),

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

This is (16). ■

Proof of Theorem 2, equation (17): Without loss of generality, normalize $\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{I}_k$ and $R'R = 1$. Set $\boldsymbol{Z}_g = \boldsymbol{X}_g R$ and $c_g = \boldsymbol{Z}_g'\boldsymbol{Z}_g$, and observe that $\sum_{g=1}^{G} c_g = 1$.

The CRVE_1 estimator can be written as

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

where $d = G(n-1)/(G-1)(n-k)$. The restriction $(\boldsymbol{X}, \boldsymbol{\Sigma}) \in \mathscr{F}_0^*$ imposes $\boldsymbol{\Sigma}_g = \sigma^2\boldsymbol{I}_{n_g}$, under which we can calculate that $v^2 = \sigma^2$ and

$$\mathbb{E}\left[\widehat{\boldsymbol{e}}_g\widehat{\boldsymbol{e}}_g'\right] = \sigma^2\boldsymbol{M}_g \tag{44}$$

where

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

The inequality in (45) can be shown by the following argument. Let $\boldsymbol{R}_\perp$ be $k \times (k-1)$ such that $[R, \boldsymbol{R}_\perp]$ is

orthonormal. Set $\boldsymbol{Z}_{\perp g} = \boldsymbol{X}_g \boldsymbol{R}_\perp$. Then

$$\boldsymbol{X}_g \boldsymbol{X}_g' = \boldsymbol{X}_g \left[R, \boldsymbol{R}_\perp\right] \begin{bmatrix} R' \\ \boldsymbol{R}_\perp' \end{bmatrix} \boldsymbol{X}_g' = \left[\boldsymbol{Z}_g, \boldsymbol{Z}_{\perp g}\right] \begin{bmatrix} \boldsymbol{Z}_g' \\ \boldsymbol{Z}_{\perp g}' \end{bmatrix} = \boldsymbol{Z}_g \boldsymbol{Z}_g' + \boldsymbol{Z}_{\perp g} \boldsymbol{Z}_{\perp g}' \geq \boldsymbol{Z}_g \boldsymbol{Z}_g',$$

which implies (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

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

Together with $v^2 = \sigma^2$, we find that

$$\frac{\mathbb{E}\left[\widehat{v}_1^2\right]}{v^2} \leq d\left(1 - \sum_{g=1}^G c_g^2\right). \tag{46}$$

Inequality (46) implies that the left side of (17) is weakly smaller than the infimum of the right-hand side of (46) over c_g. Thus

$$\inf_{(\boldsymbol{X},\boldsymbol{\Sigma}) \in \mathscr{F}_0^*} \frac{\mathbb{E}\left[\widehat{v}_1^2\right]}{v^2} \leq d \inf_{\sum_{g=1}^G c_g = 1} \left(1 - \sum_{g=1}^G c_g^2\right) = 0. \tag{47}$$

The rightmost equality in (47) is attained as $(c_1, ..., c_G) \to (1, 0, ..., 0)$. Since the left side of (47) is non-negative, the equation hold as an equality. This verifies (17). ■

Proof of Theorem 2, equation (18): We use the same normalizations and notation as in the proof of (17), add the assumption that $\lambda_{\min}\left(\boldsymbol{M}_g\right) \geq \delta > 0$ for all $g > 1$, which means that all clusters other than $g = 1$ are uniformly clusterwise invertible, and assume that $\boldsymbol{\Sigma}_1 = \boldsymbol{I}_{n_1}$ and $\boldsymbol{\Sigma}_g = 0$, which is extreme heteroskedasticity. Under these conditions, you can calculate that $\boldsymbol{V} = \boldsymbol{X}_1' \boldsymbol{X}_1$ and $v^2 = c_1$.

The CRVE_2 estimator for v^2 can be written as

$$\widehat{v}_2^2 = \sum_{g=1}^G \boldsymbol{Z}_g' \boldsymbol{M}_g^{-1/2} \widehat{\boldsymbol{e}}_g \widehat{\boldsymbol{e}}_g' \boldsymbol{M}_g^{-1/2} \boldsymbol{Z}_g. \tag{48}$$

We can calculate that for $g = 1$,

$$\mathbb{E}\left[\widehat{\boldsymbol{e}}_1 \widehat{\boldsymbol{e}}_1'\right] = \boldsymbol{I}_{n_1} - 2\boldsymbol{X}_1 \boldsymbol{X}_1' + \boldsymbol{X}_1 \boldsymbol{X}_1' \boldsymbol{X}_1 \boldsymbol{X}_1' = \boldsymbol{M}_1 \boldsymbol{M}_1, \tag{49}$$

and for $g > 1$,

$$\mathbb{E}\left[\widehat{\boldsymbol{e}}_g \widehat{\boldsymbol{e}}_g'\right] = \boldsymbol{X}_g \boldsymbol{X}_1' \boldsymbol{X}_1 \boldsymbol{X}_g'. \tag{50}$$

Taking expectations of (48) and using (49)-(50)

$$\begin{aligned}\mathbb{E}\left[\widehat{v}_2^2\right] &= \sum_{g=1}^{G} \boldsymbol{Z}_g' \boldsymbol{M}_g^{-1/2} \mathbb{E}\left[\widehat{\boldsymbol{e}}_g \widehat{\boldsymbol{e}}_g'\right] \boldsymbol{M}_g^{-1/2} \boldsymbol{Z}_g \\ &= \boldsymbol{Z}_1' \boldsymbol{M}_1^{-1/2} \boldsymbol{M}_1 \boldsymbol{M}_1 \boldsymbol{M}_1^{-1/2} \boldsymbol{Z}_1 + \sum_{g=2}^{G} \boldsymbol{Z}_g' \boldsymbol{M}_g^{-1/2} \boldsymbol{X}_g \boldsymbol{X}_1' \boldsymbol{X}_1 \boldsymbol{X}_g' \boldsymbol{M}_g^{-1/2} \boldsymbol{Z}_g. \end{aligned} \tag{51}$$

The first term on the right side of (51) equals

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

where the inequality is (45). Since

$$\lambda_{\max}\left(\boldsymbol{M}_g^{-1/2} \boldsymbol{X}_g \boldsymbol{X}_1' \boldsymbol{X}_1 \boldsymbol{X}_g' \boldsymbol{M}_g^{-1/2}\right) \leq \lambda_{\max}\left(\boldsymbol{M}_g^{-1}\right) \lambda_{\max}\left(\boldsymbol{X}_1' \boldsymbol{X}_1\right) \lambda_{\max}\left(\boldsymbol{X}_g' \boldsymbol{X}_g\right) \leq \frac{1}{\delta}$$

(using the assumption $\lambda_{\min}\left(\boldsymbol{M}_g\right) \geq \delta$ for any $g \geq 2$ and the fact $\boldsymbol{X}_g' \boldsymbol{X}_g \leq \boldsymbol{X}' \boldsymbol{X} = \boldsymbol{I}_k$ for any $g \geq 1$), the second term on the right side of (51) satisfies

$$\sum_{g=2}^{G} \boldsymbol{Z}_g' \boldsymbol{M}_g^{-1/2} \boldsymbol{X}_g \boldsymbol{X}_1' \boldsymbol{X}_1 \boldsymbol{X}_g' \boldsymbol{M}_g^{-1/2} \boldsymbol{Z}_g \leq \sum_{g=2}^{G} \frac{\boldsymbol{Z}_g' \boldsymbol{Z}_g}{\delta} = \frac{1 - c_1}{\delta}. \tag{52}$$

Together with $\nu^2 = 1$, we find that

$$\frac{\mathbb{E}\left[\widehat{v}_2^2\right]}{\nu^2} \leq 1 - c_1 + \frac{1 - c_1}{\delta c_1}. \tag{53}$$

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 (53) over c_1. Hence

$$\inf_{(\boldsymbol{X}, \boldsymbol{\Sigma}) \in \mathscr{F}^*} \frac{\mathbb{E}\left[\widehat{v}_2^2\right]}{\nu^2} \leq \inf_{0 < c_1 < 1} \left(1 - c_1 + \frac{1 - c_1}{\delta c_1}\right) = 0. \tag{54}$$

The rightmost equality in (54) is attained as $c_1 \to 1$. Equation (54) implies (18), as claimed. ■

Proof of Theorem 2, equation (19): We first show that when $\boldsymbol{X}$ is clusterwise invertible,

$$\mathbb{E}\left[\widehat{\boldsymbol{V}}_3\right] \geq \left(\frac{G-1}{G}\right)^2 \boldsymbol{V}. \tag{55}$$

The proof of (55) is analogous to that of Theorem 1. Using (9) under clusterwise invertibility, collecting

terms, (14), and then (39), we calculate that

$$\begin{aligned}\widetilde{\beta}_{-g} - \bar{\beta} &= \left(\widetilde{\beta}_{-g} - \widehat{\beta}\right) - \frac{1}{G}\sum_{h=1}^{G}\left(\widetilde{\beta}_{-h} - \widehat{\beta}\right) \\ &= \left(\frac{G-1}{G}\right)\left(\widetilde{\beta}_{-g} - \widehat{\beta}\right) - \frac{1}{G}\sum_{h\neq g}\left(\widetilde{\beta}_{-h} - \widehat{\beta}\right) \\ &= -\left(\frac{G-1}{G}\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}_g'\widetilde{\boldsymbol{e}}_g + \frac{1}{G}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\sum_{h\neq g}\boldsymbol{X}_h'\widetilde{\boldsymbol{e}}_h \\ &= -\left(\frac{G-1}{G}\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}_g'\boldsymbol{e}_g - \boldsymbol{S}_g \end{aligned} \tag{56}$$

where

$$\boldsymbol{S}_g = -\left(\frac{G-1}{G}\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}_g'\boldsymbol{X}_g\left(\widetilde{\beta}_{-g} - \beta\right) - \frac{1}{G}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\sum_{h\neq g}\boldsymbol{X}_h'\boldsymbol{e}_h + \frac{1}{G}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\sum_{h\neq g}\boldsymbol{X}_h'\boldsymbol{X}_h\left(\widetilde{\beta}_{-h} - \beta\right).$$

Notice that the first two components of $\boldsymbol{S}_g$ are uncorrelated with $\boldsymbol{e}_g$. Thus

$$\begin{aligned}\mathbb{E}\left[\left(\frac{G-1}{G}\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}_g'\boldsymbol{e}_g\boldsymbol{S}_g'\right] &= \frac{G-1}{G^2}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\sum_{h\neq g}\boldsymbol{X}_g'\mathbb{E}\left[\boldsymbol{e}_g\left(\widetilde{\beta}_{-h} - \beta\right)'\right]\boldsymbol{X}_h'\boldsymbol{X}_h\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \\ &= \frac{G-1}{G^2}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\sum_{h\neq g}\boldsymbol{X}_g'\boldsymbol{\Sigma}_g\boldsymbol{X}_g\left(\boldsymbol{X}'\boldsymbol{X} - \boldsymbol{X}_h'\boldsymbol{X}_h\right)^{-1}\boldsymbol{X}_h'\boldsymbol{X}_h\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \\ &= \boldsymbol{A}_g,\end{aligned}$$

say, where the second equality uses the relationship (under clusterwise invertibility)

$$\widetilde{\beta}_{-h} - \beta = \left(\boldsymbol{X}'\boldsymbol{X} - \boldsymbol{X}_h'\boldsymbol{X}_h\right)^{-1}\left(\sum_{\ell\neq h}\boldsymbol{X}_\ell'\boldsymbol{e}_\ell\right).$$

Using (56), it follows that

$$\begin{aligned}\mathbb{E}\left[\widehat{\boldsymbol{V}}_3\right] &= \left(\frac{G-1}{G}\right)\sum_{g=1}^{G}\mathbb{E}\left[\left(\widetilde{\beta}_{-g} - \bar{\beta}\right)\left(\widetilde{\beta}_{-g} - \bar{\beta}\right)'\right] \\ &= \left(\frac{G-1}{G}\right)^2\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\sum_{g=1}^{G}\boldsymbol{X}_g'\mathbb{E}\left[\boldsymbol{e}_g\boldsymbol{e}_g'\right]\boldsymbol{X}_g\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} + \sum_{g=1}^{G}\left(\mathbb{E}\left[\boldsymbol{S}_g\boldsymbol{S}_g'\right] + \boldsymbol{A}_g + \boldsymbol{A}_g'\right) \\ &\geq \left(\frac{G-1}{G}\right)^2\boldsymbol{V}.\end{aligned} \tag{57}$$

The positive semi-definite (PSD) inequality (57) holds because $\mathbb{E}\left[\boldsymbol{S}_g\boldsymbol{S}_g'\right]$ and $\boldsymbol{A}_g + \boldsymbol{A}_g'$ are PSD. The sum $\boldsymbol{A}_g + \boldsymbol{A}_g'$ is PSD because $\boldsymbol{A}_g$ is the matrix product of PSD matrices, and thus has non-negative eigenvalues (see Zhang and Zhang (2006, Corollary 11)). This establishes (55).

Equation (55) implies

$$\inf_{(\boldsymbol{X},\boldsymbol{\Sigma})\in\mathscr{F}^*} \frac{\mathbb{E}\left[\widehat{v}_3^2\right]}{v^2} = \inf_{(\boldsymbol{X},\boldsymbol{\Sigma})\in\mathscr{F}^*} \frac{R'\mathbb{E}\left[\widehat{\boldsymbol{V}}_3\right]R}{R'\boldsymbol{V}R} \geq \left(\frac{G-1}{G}\right)^2. \tag{58}$$

To show that (58) holds as an equality we calculate an upper bound for the left side of (58) in the context of the example from the proof of (18). 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 and $\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{I}_k$ implies

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

Consequently,

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

Using (49), the expectation of (59) equals $((G-1)/G)^2$ times

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

Using (50) the expectation of (60) equals $((G-1)/G)^2$ times

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

where the inequality follows by similar same steps as for (52).

We calculate that for $g \neq 1$

$$\mathbb{E}\left[\widehat{\boldsymbol{e}}_g\widehat{\boldsymbol{e}}_1'\right] = -\boldsymbol{X}_g\left(\boldsymbol{I}_k - \boldsymbol{X}_1'\boldsymbol{X}_1\right)\boldsymbol{X}_1'. \tag{63}$$

The assumption $\lambda_{\min}\left(\boldsymbol{M}_g\right) \geq \delta$ implies that

$$\sum_{g=2}^{G} \boldsymbol{X}_g' \boldsymbol{M}_g^{-1} \boldsymbol{X}_g \geq \frac{1}{\delta} \sum_{g=2}^{G} \boldsymbol{X}_g' \boldsymbol{X}_g = \frac{1}{\delta}\left(\boldsymbol{I}_k - \boldsymbol{X}_1' \boldsymbol{X}_1\right). \tag{64}$$

Using the Woodbury identity and the fact $\boldsymbol{X}_1' \boldsymbol{X}_1 \leq \boldsymbol{X}' \boldsymbol{X} = \boldsymbol{I}_k$,

$$\begin{aligned} \boldsymbol{X}_1' \boldsymbol{M}_1^{-1} \boldsymbol{X}_1 &= \boldsymbol{X}_1' \boldsymbol{X}_1 + \boldsymbol{X}_1' \boldsymbol{X}_1 \left(\boldsymbol{I}_k - \boldsymbol{X}_1' \boldsymbol{X}_1\right)^{-1} \boldsymbol{X}_1' \boldsymbol{X}_1 \\ &\leq \boldsymbol{I}_k + \left(\boldsymbol{I}_k - \boldsymbol{X}_1' \boldsymbol{X}_1\right)^{-1} \\ &\leq 2\left(\boldsymbol{I}_k - \boldsymbol{X}_1' \boldsymbol{X}_1\right)^{-1}. \end{aligned} \tag{65}$$

Combining (63), (64), and (65), the expectation of (61) equals $2(G-1)/G^2$ times

$$\begin{aligned} R' \sum_{g=2}^{G} \boldsymbol{X}_g' \boldsymbol{M}_g^{-1} \boldsymbol{X}_g \left(\boldsymbol{I}_k - \boldsymbol{X}_1' \boldsymbol{X}_1\right) \boldsymbol{X}_1' \boldsymbol{M}_1^{-1} \boldsymbol{X}_1 R &\leq \frac{2}{\delta} R' \left(\boldsymbol{I}_k - \boldsymbol{X}_1' \boldsymbol{X}_1\right) \left(\boldsymbol{I}_k - \boldsymbol{X}_1' \boldsymbol{X}_1\right) \left(\boldsymbol{I}_k - \boldsymbol{X}_1' \boldsymbol{X}_1\right)^{-1} R \\ &= \frac{2}{\delta} R' \left(\boldsymbol{I}_k - \boldsymbol{X}_1' \boldsymbol{X}_1\right) R \\ &= \frac{2\left(1 - c_1\right)}{\delta}. \end{aligned}$$

We calculate that for $h \neq g \neq 1$

$$\mathbb{E}\left[\widehat{\boldsymbol{e}}_g \widehat{\boldsymbol{e}}_h'\right] = \boldsymbol{X}_g \boldsymbol{X}_1' \boldsymbol{X}_1 \boldsymbol{X}_h',$$

so the expectation of (62) equals $(G-1)/G^2$ times

$$-R' \sum_{h \neq g \neq 1} \boldsymbol{X}_g' \boldsymbol{M}_g^{-1} \boldsymbol{X}_g \boldsymbol{X}_1' \boldsymbol{X}_1 \boldsymbol{X}_h' \boldsymbol{M}_h^{-1} \boldsymbol{X}_h R \leq 0,$$

where the inequality holds since it is a quadratic form of a sum of positive semi-definite matrices.

Together, with $\nu^2 = c_1$, we have

$$\frac{\mathbb{E}\left[\widehat{v}_3^2\right]}{\nu^2} \leq \left(\frac{G-1}{G}\right)^2 + \left(\frac{G-1}{G}\right)^2 \frac{1-c_1}{\delta^2 c_1} + 4\left(\frac{G-1}{G^2}\right) \frac{1-c_1}{\delta}. \tag{66}$$

As in the proof of (18), the left side of (19) is weakly smaller than the infimum of (66) over c_1. Hence

$$\inf_{(\boldsymbol{X},\boldsymbol{\Sigma}) \in \mathscr{F}^*} \frac{\mathbb{E}\left[\widehat{v}_3^2\right]}{\nu^2} \leq \inf_{0 < c_1 < 1} \left(\left(\frac{G-1}{G}\right)^2 + \left(\frac{G-1}{G}\right)^2 \frac{1-c_1}{\delta^2 c_1} + 4\left(\frac{G-1}{G^2}\right) \frac{1-c_1}{\delta}\right) = \left(\frac{G-1}{G}\right)^2.$$

The rightmost equality in (54) is attained as $c_1 \to 1$. Combined with (58) this establishes that (58) holds as an equality, which is (19), as claimed. ■

Proof of Theorem 2, equation (20): Since the model is clusterwise invertible,

$$\widehat{v}_4^2 = \left(\frac{G-1}{G}\right) \widehat{v}_5^2.$$

The result follows from (23), which we establish below. ■

Proof of Theorem 2, equations (21) and (22): We calculate an upper bound for the left side of (21)-(22) assuming one noninvertible cluster. Without loss of generality, assume $\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{I}_k$ and assume the noninvertible cluster is $g = 1$. This means that there is a linear combination $\boldsymbol{X}_g\gamma$ of the regressors which is identically zero for $g \neq 1$. Without loss of generality, assume $\gamma = R$. Since the regressor matrix is singular when cluster 1 is omitted, this cluster is discarded from the calculation of $\widehat{\nu}_3^2$ and $\widehat{\nu}_4^2$. Therefore the latter equals

$$\begin{aligned}\widehat{\nu}_4^2 &= \left(\frac{G-2}{G-1}\right)\sum_{g=2}^{G} R'\left(\widetilde{\beta}_{-g} - \widehat{\beta}\right)\left(\widetilde{\beta}_{-g} - \widehat{\beta}\right)' R \\ &= \left(\frac{G-2}{G-1}\right)\sum_{g=2}^{G} R'\boldsymbol{X}_g'\boldsymbol{M}_g^{-1}\widehat{\boldsymbol{e}}_g\widehat{\boldsymbol{e}}_g'\boldsymbol{M}_g^{-1}\boldsymbol{X}_g R \\ &= 0\end{aligned}$$

since $\boldsymbol{X}_g R = 0$ for $g \neq 1$. Thus $\widehat{\nu}_4^2$ is identically zero. Since $\widehat{\nu}_3^2 \leq \widehat{\nu}_4^2$ we find that $\widehat{\nu}_3^2 = 0$ as well. This implies (21)-(22) as stated. ■

Proof of Theorem 2, equation (23): Equation (16) implies

$$\inf_{(\boldsymbol{X},\boldsymbol{\Sigma})\in\mathscr{F}} \frac{\mathbb{E}\left[\widehat{\nu}_5^2\right]}{\nu^2} = \inf_{(\boldsymbol{X},\boldsymbol{\Sigma})\in\mathscr{F}} \frac{R'\mathbb{E}\left[\widehat{\boldsymbol{V}}_5\right]R}{R'\boldsymbol{V}R} \geq \inf_{(\boldsymbol{X},\boldsymbol{\Sigma})\in\mathscr{F}} \frac{R'\boldsymbol{V}R}{R'\boldsymbol{V}R} = 1. \tag{67}$$

To show that this is a strict equality we calculate an upper bound for the left side of (67) in the context of the example from the proof of (19). By the calculations from that proof,

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

which satisfies

$$\frac{\mathbb{E}\left[\widehat{\nu}_5^2\right]}{\nu^2} \leq 1 + \frac{1-c_1}{\delta^2 c_1}. \tag{68}$$

This implies

$$\inf_{(\boldsymbol{X},\boldsymbol{\Sigma})\in\mathscr{F}} \frac{\mathbb{E}\left[\widehat{\nu}_5^2\right]}{\nu^2} \leq \inf_{0<c_1<1}\left[1 + \frac{1-c_1}{\delta^2 c_1}\right] = 1. \tag{69}$$

Combined with (67) this yields (23) as stated. ■

For the proof of Theorem 3 we use the following distributional lower bound.

Theorem 7 *For any $n \geq 1$ let $\boldsymbol{\xi} = (\xi_0, \xi_1, ..., \xi_n) \sim \mathrm{N}(0, \boldsymbol{\Omega})$ where*

$$\boldsymbol{\Omega} = \begin{bmatrix} 1 & \rho_1 & \cdots & \rho_n \\ \rho_1 & 1 & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ \rho_n & 0 & \cdots & 1 \end{bmatrix}.$$

Then for any $w_j \geq 0$ satisfying $\sum_{j=1}^n w_j \geq 1$, any δ_j, and any $1 \leq x < \infty$,

$$\mathbb{P}\left[\frac{\xi_0^2}{\sum_{j=1}^n w_j \left(\xi_j + \delta_j\right)^2} \leq x\right] \geq F(x; 1, 1), \tag{70}$$

where $F(x; k_1, k_2)$ denotes the F distribution function with degrees of freedom k_1 and k_2.

Proof of Theorem 7: Our proof examines the inequality in four cases: (a) $\delta_j = 0$ and $\rho_j = 0$ for all j; (b) $\rho_j = 0$ for all j but $\delta_j \neq 0$; (c) $\delta_j = 0$ for all j but $\rho_j \neq 0$; (d) general ρ_j and δ_j.

Case (a): $\delta_j = 0$ and $\rho_j = 0$ for all j. The asssumption that $\rho_j = 0$ implies that the ξ_j are mutually independent. Set $\bar{w} = \sum_{j=1}^n w_j \geq 1$, $\alpha_j = w_j/\bar{w}$ for $j \leq n-1$, $\alpha_n = 1 - \sum_{j=1}^{n-1} \alpha_j$, $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_{n-1})'$, $\boldsymbol{U} = (\xi_1^2 - \xi_n^2, ..., \xi_{n-1}^2 - \xi_n^2)$, and let $H(x) = \mathbb{P}\left[\xi_0^2 \leq x\right]$ denote the χ_1^2 cumulative distribution function. Using the law of iterated expectations, the monotonicity of $H(x)$, and $\bar{w} \geq 1$, the left side of (70) equals

$$\begin{aligned} \mathbb{E}\left[\mathbb{P}\left[\xi_1^2 \leq x \sum_{j=1}^n w_j \xi_j^2 \;\middle|\; \xi_1^2, ..., \xi_n^2\right]\right] &= \mathbb{E}\left[H\left(\bar{w} x \sum_{j=1}^n \alpha_j \xi_j^2\right)\right] \\ &\geq \mathbb{E}\left[H\left(x \sum_{j=1}^n \alpha_j \xi_j^2\right)\right] \\ &= \mathbb{E}\left[H\left(x\boldsymbol{\alpha}'\boldsymbol{U} + x\xi_n^2\right)\right] \\ &\equiv f(\boldsymbol{\alpha}). \end{aligned} \tag{71}$$

We calculate that

$$\frac{\partial^2}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\alpha}'} f(\boldsymbol{\alpha}) = x^2 \mathbb{E}\left[\boldsymbol{U}\boldsymbol{U}' H''\left(x\boldsymbol{\alpha}'\boldsymbol{U} + x\xi_n^2\right)\right].$$

This matrix is negative semi-definite since the χ_1^2 distribution function $H(t)$ is globally concave. Hence, the function $f(\boldsymbol{\alpha})$ is concave in $\boldsymbol{\alpha}$. It follows that the minimum of $f(\boldsymbol{\alpha})$ over $\boldsymbol{\alpha}$ is obtained at a corner. By symmetry we can take any corner, e.g. $\boldsymbol{\alpha}_{\min} = (1, 0, ..., 0)$. Hence, the minimum of (71) over $\boldsymbol{\alpha}$ equals

$$\mathbb{E}\left[H\left(x\boldsymbol{\alpha}_{\min}'\boldsymbol{U} + x\xi_n^2\right)\right] = \mathbb{E}\left[H\left(x\xi_1^2\right)\right] = \mathbb{P}\left[\frac{\xi_0^2}{\xi_1^2} \leq x\right] = F(x; 1, 1),$$

since ξ_0^2/ξ_1^2 is distributed F with degrees of freedom $(1, 1)$. This establishes (70).

Case (b): $\rho_j = 0$ for all j but $\delta_j \neq 0$. In this case, $\sum_{j=1}^n w_j \left(\xi_j + \delta_j\right)^2$ is independent of ξ_0^2. By Theorem 3 of Mathew and Nordström (1997), $\sum_{j=1}^n w_j \xi_j^2$ is stochastically dominated by $\sum_{j=1}^n w_j \left(\xi_j + \delta_j\right)^2$. This

implies

$$\mathbb{P}\left[\frac{\xi_0^2}{\sum_{j=1}^n w_j\left(\xi_j+\delta_j\right)^2}\leq x\right]\geq\mathbb{P}\left[\frac{\xi_0^2}{\sum_{j=1}^n w_j\xi_j^2}\leq x\right]. \tag{72}$$

The right-side corresponds to Case (a), which we have already shown is bounded below by $F(x;1,1)$. This establishes (70).

Case (c): $\delta_j=0$ for all j but $\rho_j\neq 0$. Define the $(n+1)\times(n+1)$ positive semi-definite matrices $\boldsymbol{A}=\text{diag}\{1,0,...,0\}$, $\boldsymbol{B}=\text{diag}\{0,w_1,...,w_n\}$, and $\boldsymbol{C}=\boldsymbol{\Omega}^{1/2}(\boldsymbol{A}-x\boldsymbol{B})\boldsymbol{\Omega}^{1/2}$. By the spectral decomposition, $\boldsymbol{C}=\boldsymbol{H}\boldsymbol{\Lambda}\boldsymbol{H}'$, where $\boldsymbol{H}'\boldsymbol{H}=\boldsymbol{I}_{n+1}$, $\boldsymbol{\Lambda}=\text{diag}\{\lambda_0,...,\lambda_n\}$, and $\lambda_0\geq\lambda_1\geq\cdots\geq\lambda_n$ are the ordered eigenvalues of $\boldsymbol{C}$. Set $\boldsymbol{u}=\boldsymbol{\Omega}^{-1/2}\boldsymbol{\xi}\sim\text{N}(0,\boldsymbol{I}_{n+1})$ and $\boldsymbol{\zeta}=\boldsymbol{H}'\boldsymbol{u}\sim\text{N}(0,\boldsymbol{I}_{n+1})$. Partition $\boldsymbol{\zeta}=(\zeta_0,...,\zeta_n)'$ in conformity with $\boldsymbol{\Lambda}$.

The left side of (70) equals

$$\begin{aligned}\mathbb{P}\left[\frac{\xi_0^2}{\sum_{j=1}^n w_j\xi_j^2}\leq x\right]&=\mathbb{P}\left[\xi_0^2\leq x\sum_{j=1}^n w_j\xi_j^2\right]\\&=\mathbb{P}\left[\boldsymbol{\xi}'(\boldsymbol{A}-x\boldsymbol{B})\boldsymbol{\xi}\leq 0\right]\\&=\mathbb{P}\left[\boldsymbol{u}'\boldsymbol{C}\boldsymbol{u}\leq 0\right]\\&=\mathbb{P}\left[\boldsymbol{\zeta}'\boldsymbol{\Lambda}\boldsymbol{\zeta}\leq 0\right]\\&=\mathbb{P}\left[\sum_{j=0}^n\lambda_j\zeta_j^2\leq 0\right].\end{aligned} \tag{73}$$

We next establish two bounds on the eigenvalues λ_j. First,

$$\begin{aligned}\lambda_0&=\lambda_{\max}(\boldsymbol{C})\\&\leq\lambda_{\max}\left(\boldsymbol{\Omega}^{1/2}\boldsymbol{A}\boldsymbol{\Omega}^{1/2}\right)+\lambda_{\max}\left(-x\boldsymbol{\Omega}^{1/2}\boldsymbol{B}\boldsymbol{\Omega}^{1/2}\right)\\&=\lambda_{\max}(\boldsymbol{A}\boldsymbol{\Omega})-x\lambda_{\min}(\boldsymbol{B}\boldsymbol{\Omega})\\&=1,\end{aligned} \tag{74}$$

since $\lambda_{\max}(\boldsymbol{A}\boldsymbol{\Omega})=1$ by direct calculation, and $\lambda_{\min}(\boldsymbol{B}\boldsymbol{\Omega})=0$ since $\boldsymbol{B}$ has deficient rank.

Second, by a corollary of the Weyl eigenvalue inequality for Hermitian matrices (Corollary 4.3.15 of Horn and Johnson (2013)), for each $j\geq 1$,

$$\begin{aligned}\lambda_j&=\lambda_{j+1}(\boldsymbol{C})\\&\leq\lambda_{j+1}\left(\boldsymbol{\Omega}^{1/2}\boldsymbol{A}\boldsymbol{\Omega}^{1/2}\right)+\lambda_{\max}\left(-x\boldsymbol{\Omega}^{1/2}\boldsymbol{B}\boldsymbol{\Omega}^{1/2}\right)\\&=\lambda_{j+1}(\boldsymbol{A}\boldsymbol{\Omega})-x\lambda_{\min}(\boldsymbol{B}\boldsymbol{\Omega})\\&=0,\end{aligned} \tag{75}$$

the final equality since both $\boldsymbol{A}$ and $\boldsymbol{B}$ have deficient rank. Together, (74) and (75) show that the largest eigenvalue λ_0 of $\boldsymbol{C}$ is less than one, and the remaining are non-positive.

If $\lambda_0\leq 0$ then $\lambda_j\leq 0$ for all j by (75), so (73) equals 1. This implies (70). We thus assume $\lambda_0>0$ for the

remainder of the proof. For $j = 1, ..., n$ set $\alpha_j = -\lambda_j / (\lambda_0 x) \geq 0$. We see that (73) equals

$$\mathbb{P}\left[\zeta_0^2 \leq x \sum_{j=1}^{n} \alpha_j \zeta_j^2\right] = \mathbb{P}\left[\frac{\zeta_0^2}{\sum_{j=1}^{n} \alpha_j \zeta_j^2} \leq x\right]. \tag{76}$$

We next establish

$$\sum_{j=1}^{n} \alpha_j \geq 1. \tag{77}$$

Using the property $\operatorname{tr}(\boldsymbol{C}) = \sum_{j=0}^{n} \lambda_j$ and the direct calculations $\operatorname{tr}(\boldsymbol{A\Omega}) = 1$ and $\operatorname{tr}(\boldsymbol{B\Omega}) = \sum_{j=1}^{n} w_j$,

$$\begin{aligned}
\sum_{j=1}^{n} \alpha_j &= \frac{-\sum_{j=1}^{n} \lambda_j}{\lambda_0 x} \\
&= \frac{\lambda_0 - \operatorname{tr}(\boldsymbol{C})}{\lambda_0 x} \\
&= \frac{\lambda_0 - \operatorname{tr}(\boldsymbol{A\Omega}) + x \operatorname{tr}(\boldsymbol{B\Omega})}{\lambda_0 x} \\
&= \frac{\lambda_0 - 1 + x \sum_{j=1}^{n} w_j}{\lambda_0 x} \\
&\geq \frac{\lambda_0 - 1 + x}{\lambda_0 x} \\
&\geq \frac{\lambda_0 + (x-1)\lambda_0}{\lambda_0 x} \\
&= 1
\end{aligned}$$

the first inequality using $\sum_{j=1}^{n} w_j \geq 1$ and the second using $x \geq 1$ and (74). This establishes (77). We have shown that the left side of (70) equals (76) where $\zeta_j \sim \mathrm{N}(0,1)$ are mutually independent and $\sum_{j=1}^{n} \alpha_j \geq 1$. This corresponds to Case (a), which we have already shown is bounded below by $F(x;1,1)$. This establishes (70).

Case (d): general ρ_j and δ_j. Use the same notation as in the proof of Case (c), plus $\boldsymbol{\delta} = (0, \delta_1 ..., \delta_n)'$ and $\boldsymbol{\mu} = \boldsymbol{H}' \boldsymbol{\Omega}^{-1/2} \boldsymbol{\delta}$. By similar manipulations, the left side of (70) equals

$$\mathbb{P}\left[\sum_{j=0}^{n} \lambda_j \left(\zeta_j + \mu_j\right)^2 \leq 0\right] = \mathbb{P}\left[\sum_{j=0}^{n} \lambda_j \chi_1^2\left(\mu_j^2\right) \leq 0\right] \tag{78}$$

where $\chi_1^2\left(\mu_j^2\right)$ are mutually independent non-central chi-square random variables with one degree of freedom and non-centrality parameter μ_j^2, and λ_j are the ordered eigenvalues of $\boldsymbol{C}$. For these calcula-

tions we use the following explicit formulae. Standard manipulations reveal that

$$\boldsymbol{C} = \begin{bmatrix} \sigma^2 & \sigma\rho_1 & \cdots & \sigma\rho_n \\ \sigma\rho_1 & \rho_1^2 - xw_1 & & \rho_n\rho_1 \\ \vdots & & \ddots & \vdots \\ \sigma\rho_n & \rho_n\rho_1 & \cdots & \rho_n^2 - xw_2 \end{bmatrix}, \qquad \boldsymbol{\mu} = \boldsymbol{H}' \begin{bmatrix} -\sum_{j=1}^n \rho_j\delta_j/\sigma \\ \delta_1 \\ \vdots \\ \delta_n \end{bmatrix}$$

where $\sigma^2 = 1 - \sum_{j=1}^n \rho_j^2$. Given the inputs $(x, \rho_1, ..., \rho_n, \delta_1, ..., \delta_n, w_1, ..., w_n)$, the matrix $\boldsymbol{C}$ can be numerically calculated, from $\boldsymbol{C}$ the eigenvalues λ_j and eigenvector matrix $\boldsymbol{H}$ can be calculated, and from these the non-centrality coefficients μ_j. Given the inputs $(\lambda_1, ..., \lambda_n, \mu_1, ..., \mu_n)$ the probability (78) can be calculated using the method of Imhof (1961). We use R package `CompQuadForm` of Duchesne and Lafaye de Micheaux (2010). The results can be compared to verify that (78) exceeds $F(x; 1, 1)$.

The most important cases to investigate are $n = 1$ and $n = 2$, so for these our calculations were most thorough. For $n = 1$ we calculated (78) for each $\rho_1 \in (-1, 1)$ and $\mu_1 \in [-3, 3]$ on a double grid with increments of 0.01, and for each $x \geq 1$ corresponding to increments of $F(x; 1, 1)$ of 0.01. Uniformly, the inequality (70) held.

For $n = 2$ we generated 100,000,000 draws of parameters, calculated (78) for all x as for the case $n = 1$, and compared the result with $F(x; 1, 1)$. Uniformly across the 100,000,000 parameter draws and all values of x, the inequality (70) held. The parameters were drawn as follows. The correlations (ρ_1, ρ_2) were drawn from a uniform distribution on the unit disk. The coefficients $(\delta_1\delta_2)$ were drawn from the $U[-2, 2]^2$ distribution. The weights (w_1, w_2) were drawn from a flat (uniform) Dirichlet distribution.

For the cases $3 \leq n \leq 10$ our calculations were similar to the $n = 2$ case, but using 10,000,000 draws of parameters for each n. The correlations $(\rho_1, ..., \rho_n)$ were drawn from a uniform distribution on the unit disk. The coefficients $(\delta_1, ..., \delta_n)$ were drawn from the $U[-2, 2]^n$ distribution. The weights $(w_1, ..., w_n)$ were drawn from a flat (uniform) Dirichlet distribution. For each parameter draw and all values of x, the inequality (70) held.

We conclude that the inequality (70) holds as stated. ■

Proof of Theorem 3: Take any $(\boldsymbol{X}, \boldsymbol{\Sigma}) \in \mathscr{F}$ and $1 \leq c < \infty$. Define

$$\xi_0 = \frac{\widehat{\theta} - \theta}{\nu} \sim \mathrm{N}(0, 1). \tag{79}$$

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

$$\widetilde{\boldsymbol{M}}_{gj} = \begin{cases} \boldsymbol{I}_{n_g} & g = j \\ -\boldsymbol{X}_g \left(\boldsymbol{X}'\boldsymbol{X} - \boldsymbol{X}_g'\boldsymbol{X}_g \right)^+ \boldsymbol{X}_j' & g \neq j. \end{cases} \tag{80}$$

This operator has the algebraic property that it creates delete-one-cluster prediction errors, $\widetilde{\boldsymbol{M}}\boldsymbol{Y} = \widetilde{\boldsymbol{e}}$, and is the jackknife analog of the least squares annihilation matrix. It has the property that $\widetilde{\boldsymbol{M}}\boldsymbol{X} = 0$ under

clusterwise invertibility, but not under clusterwise noninvertibility. Define

$$\boldsymbol{Z} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}R, \tag{81}$$

partition by cluster as $\boldsymbol{Z} = \left(\boldsymbol{Z}_1', ...\boldsymbol{Z}_G'\right)'$, and define

$$\bar{\boldsymbol{Z}} = \text{diag}\{\boldsymbol{Z}_1, ..., \boldsymbol{Z}_G\}. \tag{82}$$

Using (15), these definitions, and the equations $\widetilde{\boldsymbol{e}} = \widetilde{\boldsymbol{M}}\boldsymbol{Y}$ and $\boldsymbol{Y} = \boldsymbol{X}\beta + \boldsymbol{e}$ we find

$$\begin{aligned}
\widehat{v}_5^2 &= R'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\sum_{g=1}^{G}\boldsymbol{X}_g'\widetilde{\boldsymbol{e}}_g\widetilde{\boldsymbol{e}}_g'\boldsymbol{X}_g\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}R \\
&= \sum_{g=1}^{G}\boldsymbol{Z}_g'\widetilde{\boldsymbol{e}}_g\widetilde{\boldsymbol{e}}_g'\boldsymbol{Z}_g \\
&= \widetilde{\boldsymbol{e}}'\bar{\boldsymbol{Z}}\bar{\boldsymbol{Z}}'\widetilde{\boldsymbol{e}} \\
&= \boldsymbol{Y}'\widetilde{\boldsymbol{M}}\bar{\boldsymbol{Z}}\bar{\boldsymbol{Z}}'\widetilde{\boldsymbol{M}}\boldsymbol{Y} \\
&= \left(\boldsymbol{X}\beta + \boldsymbol{e}\right)'\widetilde{\boldsymbol{M}}\bar{\boldsymbol{Z}}\bar{\boldsymbol{Z}}'\widetilde{\boldsymbol{M}}\left(\boldsymbol{X}\beta + \boldsymbol{e}\right) \qquad (83) \\
&= \left(\boldsymbol{\psi} + \boldsymbol{U}\right)'\left(\boldsymbol{\psi} + \boldsymbol{U}\right) \qquad (84)
\end{aligned}$$

where $\boldsymbol{\psi} = \bar{\boldsymbol{Z}}'\widetilde{\boldsymbol{M}}\boldsymbol{X}\beta$ and $\boldsymbol{U} = \bar{\boldsymbol{Z}}'\widetilde{\boldsymbol{M}}\boldsymbol{e}$.

Note that $\boldsymbol{U} \sim \text{N}(0, \boldsymbol{A})$ where $\boldsymbol{A} = \bar{\boldsymbol{Z}}'\widetilde{\boldsymbol{M}}\boldsymbol{\Sigma}\widetilde{\boldsymbol{M}}'\bar{\boldsymbol{Z}}$. Let $r = \text{rank}(\boldsymbol{A})$. By the spectral decomposition,

$$\boldsymbol{A} = \boldsymbol{H}\begin{bmatrix} \boldsymbol{\Lambda} & 0 \\ 0 & 0 \end{bmatrix}\boldsymbol{H}'$$

where $\boldsymbol{\Lambda}$ has the r non-zero eigenvalues of $\boldsymbol{A}$ on the diagonal and $\boldsymbol{H}\boldsymbol{H}' = \boldsymbol{I}_n$. Partition $\boldsymbol{H} = [\boldsymbol{H}_1, \boldsymbol{H}_2]$ conformably and define $\boldsymbol{\delta}_1 = \boldsymbol{\Lambda}^{-1/2}\boldsymbol{H}_1'\boldsymbol{\psi}$, $\boldsymbol{\delta}_2 = \boldsymbol{H}_2'\boldsymbol{\psi}$, and $\boldsymbol{\xi}_1 = \boldsymbol{\Lambda}^{-1/2}\boldsymbol{H}_1'\boldsymbol{U} \sim \text{N}(0, \boldsymbol{I}_r)$. Observe that $\boldsymbol{H}_2'\bar{\boldsymbol{Z}}'\widetilde{\boldsymbol{M}}\boldsymbol{e} = 0$ almost surely. Since $\boldsymbol{I}_n = \boldsymbol{H}\boldsymbol{H}'$, (84) equals

$$\begin{aligned}
\left(\boldsymbol{H}'\boldsymbol{\psi} + \boldsymbol{H}'\boldsymbol{U}\right)'\left(\boldsymbol{H}'\boldsymbol{\psi} + \boldsymbol{H}'\boldsymbol{U}\right) &= \left(\begin{pmatrix} \boldsymbol{\Lambda}^{1/2}\boldsymbol{\delta}_1 \\ \boldsymbol{\delta}_2 \end{pmatrix} + \begin{pmatrix} \boldsymbol{\Lambda}^{1/2}\boldsymbol{\xi}_1 \\ 0 \end{pmatrix}\right)'\left(\begin{pmatrix} \boldsymbol{\Lambda}^{1/2}\boldsymbol{\delta}_1 \\ \boldsymbol{\delta}_2 \end{pmatrix} + \begin{pmatrix} \boldsymbol{\Lambda}^{1/2}\boldsymbol{\xi}_1 \\ 0 \end{pmatrix}\right) \\
&= \begin{pmatrix} \boldsymbol{\delta}_1 + \boldsymbol{\xi}_1 \\ \boldsymbol{\delta}_2 \end{pmatrix}'\begin{bmatrix} \boldsymbol{\Lambda} & 0 \\ 0 & \boldsymbol{I}_{n-r} \end{bmatrix}\begin{pmatrix} \boldsymbol{\delta}_1 + \boldsymbol{\xi}_1 \\ \boldsymbol{\delta}_2 \end{pmatrix} \\
&= \left(\boldsymbol{\delta}_1 + \boldsymbol{\xi}_1\right)'\boldsymbol{\Lambda}\left(\boldsymbol{\delta}_1 + \boldsymbol{\xi}_1\right) + \boldsymbol{\delta}_2'\boldsymbol{\delta}_2 \\
&\geq \left(\boldsymbol{\delta}_1 + \boldsymbol{\xi}_1\right)'\boldsymbol{\Lambda}\left(\boldsymbol{\delta}_1 + \boldsymbol{\xi}_1\right) \\
&= \sum_{j=1}^{r}\lambda_j\left(\xi_j + \delta_j\right)^2 \qquad (85)
\end{aligned}$$

where $\boldsymbol{\xi}_1 = (\xi_1, ..., \xi_r)$ and $\boldsymbol{\delta}_1 = (\delta_1, ..., \delta_r)$.

We can write $\widetilde{\boldsymbol{M}} = \boldsymbol{I}_n - \widetilde{\boldsymbol{P}}$, where $\widetilde{\boldsymbol{P}}$ is a delete-one-cluster projection operator which has the property

that its diagonal blocks are zeros. The eigenvalues λ_j satisfy

$$\begin{aligned}\sum_{j=1}^{r}\lambda_j &= \operatorname{tr}(\boldsymbol{A})\\ &= \operatorname{tr}\left(\bar{\boldsymbol{Z}}'\widetilde{\boldsymbol{M}}\boldsymbol{\Sigma}\widetilde{\boldsymbol{M}}'\bar{\boldsymbol{Z}}\right)\\ &= \operatorname{tr}\left(\bar{\boldsymbol{Z}}'\boldsymbol{\Sigma}\bar{\boldsymbol{Z}}\right) - 2\operatorname{tr}\left(\bar{\boldsymbol{Z}}'\widetilde{\boldsymbol{P}}\boldsymbol{\Sigma}\bar{\boldsymbol{Z}}\right) + \operatorname{tr}\left(\bar{\boldsymbol{Z}}'\widetilde{\boldsymbol{P}}\boldsymbol{\Sigma}\widetilde{\boldsymbol{P}}'\bar{\boldsymbol{Z}}\right)\\ &\geq \nu^2 \end{aligned} \tag{86}$$

because

$$\operatorname{tr}\left(\bar{\boldsymbol{Z}}'\boldsymbol{\Sigma}\bar{\boldsymbol{Z}}\right) = \sum_{g=1}^{G} Z_g'\boldsymbol{\Sigma}_g Z_g = \nu^2$$

and $\operatorname{tr}\left(\bar{\boldsymbol{Z}}'\widetilde{\boldsymbol{P}}\boldsymbol{\Sigma}\bar{\boldsymbol{Z}}\right) = 0$, the latter because the matrices $\bar{\boldsymbol{Z}}'$ and $\boldsymbol{\Sigma}\bar{\boldsymbol{Z}}$ are block diagonal, while $\widetilde{\boldsymbol{P}}$ has diagonal blocks equalling zero.

Using $\widehat{C}_5(c) = \widehat{\theta} \pm \widehat{v}_5 c$, the definition $\xi_0 = \left(\widehat{\theta} - \theta\right)/\nu$, and the derivations (84)-(85), we find that

$$\mathbb{P}\left[\theta \in \widehat{C}_5(c)\right] = \mathbb{P}\left[\frac{\left(\widehat{\theta}-\theta\right)^2}{\widehat{v}_5^2} \leq c^2\right] = \mathbb{P}\left[\frac{\xi_0^2}{\sum_{j=1}^{r} w_j\left(\xi_j + \delta_j\right)^2} \leq c^2\right], \tag{87}$$

where $w_j = \lambda_j/\nu^2$ satisfy $\sum_{j=1}^{r} w_j \geq 1$ by (86), and the ξ_j are normal random variables satisfying the conditions of Theorem 7 (they are mean zero, unit variance, and ξ_j are mutually uncorrelated for $j \geq 1$). Hence by Theorem 7, (87) is bounded below by $F(c;1,1)$.

If the model is assumed to satisfy clusterwise invertibility, then $\widetilde{\boldsymbol{M}}\boldsymbol{X} = 0$, which implies $\boldsymbol{\psi} = 0$, $\boldsymbol{\delta}_1 = 0$, and $\delta_j = 1$ for $j = 1, ..., r$. In this case, the application of Theorem 7 to bound (87) below by $F(c;1,1)$ does not rely on the numerical argument, justifying the claim in the text.

This holds for all models in $\mathscr{F}$ and thus establishes (24), completing the proof. ■

For the proof of Theorem 4 we use an intermediate result which connects full downward bias with zero coverage.

Theorem 8 *Under Assumptions 1-2, if for some variance estimator $\widehat{\nu}^2$ and model class $\mathscr{F}_a \subset \mathscr{F}$,*

$$\inf_{(\boldsymbol{X},\boldsymbol{\Sigma})\in\mathscr{F}_a} \frac{\mathbb{E}\left[\widehat{\nu}^2\right]}{\nu^2} = 0, \tag{88}$$

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

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

Proof of Theorem 8: Set $\varepsilon > 0$. Let $Q = \left(\widehat{\theta} - \theta\right)^2/\nu^2 \sim \chi_1^2$. Let q be the εth quantile of the χ_1^2 distribution and set $\eta = q/c^2$. Define the events $A = \left\{\widehat{\nu}^2/\nu^2 \leq \eta\right\}$ and $B = \left\{\left(\widehat{\theta} - \theta\right)^2/\widehat{\nu}^2 \leq c^2\right\}$. They jointly imply the

event $\{Q \leq c^2\eta\}$. Thus

$$\mathbb{P}[B \cap A] \leq \mathbb{P}\left[Q \leq c^2\eta\right] = \varepsilon. \tag{89}$$

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

$$\frac{\mathbb{E}\left[\widehat{v}^2\right]}{v^2} \leq \eta\varepsilon, \tag{90}$$

which is feasible by (88). By Markov's inequality and (90),

$$\mathbb{P}\left[B \cap A^c\right] \leq \mathbb{P}\left[A^c\right] = \mathbb{P}\left[\frac{\widehat{v}^2}{v^2} > \eta\right] \leq \frac{\mathbb{E}\left[\widehat{v}^2\right]}{\eta v^2} \leq \varepsilon. \tag{91}$$

Equations (89) and (91) imply that

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

As ε is arbitrary this establishes the stated result. ■

Proof of Theorem 4: Results (25), (26), (27), and (28) follow from Theorem 2, equations (17), (18), (21), and (22), combined with Theorem 8. ■

Proof of Theorem 5: Define ξ_0, $\boldsymbol{Z}$, and $\bar{\boldsymbol{Z}}$ as in (79), (81), and (82). Consider the multivariate regression of $\boldsymbol{e}$ on ξ_0:

$$\boldsymbol{e} = v\boldsymbol{\Sigma}\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{\Sigma}\boldsymbol{Z}\right)^{-1}\xi_0 + \boldsymbol{\varepsilon} \tag{92}$$

where

$$\boldsymbol{\varepsilon} = \left(\boldsymbol{I}_n - \boldsymbol{\Sigma}\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{\Sigma}\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\right)\boldsymbol{e}.$$

The variables $\boldsymbol{\varepsilon}$ and ξ_0 are jointly normal, uncorrelated, and therefore independent. The variance matrix of $\boldsymbol{\varepsilon}$ is

$$\boldsymbol{\Omega} = \boldsymbol{\Sigma} - \boldsymbol{\Sigma}\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{\Sigma}\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{\Sigma}. \tag{93}$$

Combining equation (83) from the proof of Theorem 4 with (92), we find

$$\widehat{v}_5^2 = \left(\boldsymbol{q} + \boldsymbol{\varepsilon}\right)'\widetilde{\boldsymbol{M}}\bar{\boldsymbol{Z}}\bar{\boldsymbol{Z}}'\widetilde{\boldsymbol{M}}\left(\boldsymbol{q} + \boldsymbol{\varepsilon}\right)$$

where

$$\boldsymbol{q} = \boldsymbol{X}\beta + v\boldsymbol{\Sigma}\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{\Sigma}\boldsymbol{Z}\right)^{-1}\xi_0.$$

Following similar steps as in the proof of Theorem 4, we find that the squared t-statistic satisfies

$$T_5^2 = \frac{\left(\widehat{\theta} - \theta\right)^2}{\widehat{v}_5^2} = \frac{v^2\xi_0^2}{\left(\boldsymbol{t}_1 + \boldsymbol{\xi}_1\right)'\boldsymbol{\Phi}\left(\boldsymbol{t}_1 + \boldsymbol{\xi}_1\right) + \boldsymbol{t}_2'\boldsymbol{t}_2}$$

where $\boldsymbol{\Phi}$ are the non-zero eigenvalues of

$$\boldsymbol{L} = \bar{\boldsymbol{Z}}'\widetilde{\boldsymbol{M}}\boldsymbol{\Sigma}_\varepsilon\widetilde{\boldsymbol{M}}'\bar{\boldsymbol{Z}}, \tag{94}$$

$\boldsymbol{H}^* = \left[\boldsymbol{H}_1^*, \boldsymbol{H}_2^*\right]$ are the associated eigenvectors, $\boldsymbol{t}_1 = \boldsymbol{\Phi}^{-1/2}\boldsymbol{H}_1^{*\prime}\tilde{\boldsymbol{Z}}'\widetilde{\boldsymbol{M}}\boldsymbol{q}$, $\boldsymbol{t}_2 = \boldsymbol{H}_2^{*\prime}\tilde{\boldsymbol{Z}}'\widetilde{\boldsymbol{M}}\boldsymbol{q}$, and $\boldsymbol{\xi}_1 = \boldsymbol{\Phi}^{-1/2}\boldsymbol{H}_1^{*\prime}\tilde{\boldsymbol{Z}}'\widetilde{\boldsymbol{M}}\boldsymbol{\varepsilon} \sim \mathrm{N}(0, \boldsymbol{I}_r)$. Thus

$$T_5^2 \geq \frac{\nu^2\xi_0^2}{\sum_{j=1}^r \phi_j\left(t_j + \xi_j\right)^2}$$

where ϕ_j, t_j, and ξ_j are the components of $\boldsymbol{\Phi}$, $\boldsymbol{t}_1$, and $\boldsymbol{\xi}_1$.

Notice that $\boldsymbol{\xi}_1$ is independent of $(\xi_0, \boldsymbol{t}_1)$. Thus, conditioning on $(\xi_0, \boldsymbol{t}_1)$, by Theorem 3 of Mathew and Nordström (1997) (see equation (72)), $\sum_{j=1}^r \phi_j\left(t_j + \xi_j\right)^2$ stochastically dominates $\sum_{j=1}^r \phi_j\xi_j^2$, so

$$\mathbb{P}\left[T_5^2 \leq x\right] \geq \mathbb{P}\left[\frac{\nu^2\xi_0^2}{\sum_{j=1}^r \phi_j\left(t_j + \xi_j\right)^2} \leq x\right] \geq \mathbb{P}\left[\frac{\nu^2\xi_0^2}{\sum_{j=1}^r \phi_j\xi_j^2} \leq x\right]. \tag{95}$$

The random variables $(\xi_0, \xi_1, ..., \xi_r)$ are independent $\mathrm{N}(0,1)$. This is the inequality in (30).

The Satterthwaite approximation states that since $Q = \sum_{j=1}^r \phi_j\xi_j^2$ is a weighted sum of independent chi-squares with non-negative weights, then $Q \simeq bQ_K/K$, with

$$b = \sum_{j=1}^G \phi_j = \mathrm{tr}\left(\boldsymbol{L}\right),$$

$$K = \frac{\left(\sum_{j=1}^r \phi_j\right)^2}{\sum_{j=1}^r \phi_j^2} = \frac{(\mathrm{tr}\left(\boldsymbol{L}\right))^2}{\mathrm{tr}\left(\boldsymbol{L}\boldsymbol{L}\right)},$$

and $Q_K \sim \chi_K^2$. It follows that $\xi_0^2/\left(Q_K/K\right)$ has distribution function $F\left(x; 1, K\right)$. Then the right side of (95) equals

$$\mathbb{P}\left[\frac{\nu^2\xi_0^2}{Q} \leq x\right] \simeq \mathbb{P}\left[\frac{\nu^2\xi_0^2}{bQ_K/K} \leq x\right] = \mathbb{P}\left[\frac{\xi_0^2}{Q_K/K} \leq \frac{bx}{\nu^2}\right] = F\left(a^2x; 1, K\right). \tag{96}$$

with $a = \sqrt{b/\nu^2}$. This is the approximation in (30) as stated. ■

Proof of Theorem 6: Recalling definition (82), with some algebra we can write

$$\widetilde{\boldsymbol{M}}'\tilde{\boldsymbol{Z}} = \begin{bmatrix} \boldsymbol{Z}_1 & -\boldsymbol{X}_1\boldsymbol{U}_2 & \cdots & -\boldsymbol{X}_1\boldsymbol{U}_G \\ -\boldsymbol{X}_2\boldsymbol{U}_1 & \boldsymbol{Z}_2 & \cdots & -\boldsymbol{X}_2\boldsymbol{U}_G \\ \vdots & \vdots & \ddots & \vdots \\ -\boldsymbol{X}_G\boldsymbol{U}_1 & -\boldsymbol{X}_G\boldsymbol{U}_2 & \cdots & \boldsymbol{Z}_G \end{bmatrix} = \boldsymbol{S} - \boldsymbol{X}\boldsymbol{U}'.$$

Under $\boldsymbol{\Sigma} = \boldsymbol{I}_n$, (93) equals $\boldsymbol{\Omega} = \boldsymbol{I}_n - \boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'$. Then (94) equals

$$\boldsymbol{L} = \tilde{\boldsymbol{Z}}'\widetilde{\boldsymbol{M}}\boldsymbol{\Omega}\widetilde{\boldsymbol{M}}'\tilde{\boldsymbol{Z}} = \left(\boldsymbol{S}' - \boldsymbol{U}\boldsymbol{X}'\right)\boldsymbol{\Omega}\left(\boldsymbol{S} - \boldsymbol{X}\boldsymbol{U}'\right) = \boldsymbol{S}'\boldsymbol{\Omega}\boldsymbol{S} + \boldsymbol{U}\widehat{\boldsymbol{X}}'\widehat{\boldsymbol{X}}\boldsymbol{U}' - \boldsymbol{V}\boldsymbol{U}' - \boldsymbol{U}\boldsymbol{V}'$$

where we use $\widehat{\boldsymbol{X}} = \boldsymbol{\Omega}\boldsymbol{X}$ and $\boldsymbol{V} = \boldsymbol{S}'\widehat{\boldsymbol{X}}$.

Using (93) and the relationship $\boldsymbol{T} = \boldsymbol{S}'\boldsymbol{Z}$ we find

$$\begin{aligned}\operatorname{tr}[\boldsymbol{L}] &= \operatorname{tr}\left[\boldsymbol{S}'\boldsymbol{S}\right] - \operatorname{tr}\left[\boldsymbol{S}'\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{S}\right] + \operatorname{tr}\left[\boldsymbol{U}\widehat{\boldsymbol{X}}'\widehat{\boldsymbol{X}}\boldsymbol{U}'\right] - \operatorname{tr}\left[\boldsymbol{V}\boldsymbol{U}'\right] - \operatorname{tr}\left[\boldsymbol{U}\boldsymbol{V}'\right] \\ &= \sum_{g=1}^{G}\boldsymbol{S}_g'\boldsymbol{S}_g - \frac{\boldsymbol{T}'\boldsymbol{T}}{\boldsymbol{Z}'\boldsymbol{Z}} + \operatorname{tr}\left[\boldsymbol{U}'\boldsymbol{U}\widehat{\boldsymbol{X}}'\widehat{\boldsymbol{X}}\right] - 2\operatorname{tr}\left[\boldsymbol{U}'\boldsymbol{V}\right]\end{aligned}$$

which is (34).

Similarly,

$$\begin{aligned}\operatorname{tr}[\boldsymbol{L}\boldsymbol{L}] &= \operatorname{tr}\left[\left(\boldsymbol{S}'\boldsymbol{\Omega}\boldsymbol{S} + \boldsymbol{U}\widehat{\boldsymbol{X}}'\widehat{\boldsymbol{X}}\boldsymbol{U}' - \boldsymbol{V}\boldsymbol{U}' - \boldsymbol{U}\boldsymbol{V}'\right)\left(\boldsymbol{S}'\boldsymbol{\Omega}\boldsymbol{S} + \boldsymbol{U}\widehat{\boldsymbol{X}}'\widehat{\boldsymbol{X}}\boldsymbol{U}' - \boldsymbol{V}\boldsymbol{U}' - \boldsymbol{U}\boldsymbol{V}'\right)\right] \\ &= \operatorname{tr}\left[\left(\boldsymbol{S}'\boldsymbol{\Omega}\boldsymbol{S}\right)\left(\boldsymbol{S}'\boldsymbol{\Omega}\boldsymbol{S}\right)\right] + \operatorname{tr}\left[\boldsymbol{U}\widehat{\boldsymbol{X}}'\widehat{\boldsymbol{X}}\boldsymbol{U}'\boldsymbol{U}\widehat{\boldsymbol{X}}'\widehat{\boldsymbol{X}}\boldsymbol{U}'\right] + 2\operatorname{tr}\left[\boldsymbol{V}\boldsymbol{U}'\boldsymbol{V}\boldsymbol{U}'\right] \\ &+ 2\operatorname{tr}\left[\boldsymbol{S}'\boldsymbol{\Omega}\boldsymbol{S}\boldsymbol{U}\widehat{\boldsymbol{X}}'\widehat{\boldsymbol{X}}\boldsymbol{U}'\right] - 4\operatorname{tr}\left[\boldsymbol{S}'\boldsymbol{\Omega}\boldsymbol{S}\boldsymbol{V}\boldsymbol{U}'\right] - 4\operatorname{tr}\left[\boldsymbol{U}\widehat{\boldsymbol{X}}'\widehat{\boldsymbol{X}}\boldsymbol{U}'\boldsymbol{V}\boldsymbol{U}'\right] + 2\operatorname{tr}\left[\boldsymbol{V}\boldsymbol{U}'\boldsymbol{U}\boldsymbol{V}'\right].\end{aligned}$$

Using (93) and the relationship $\boldsymbol{W} = \boldsymbol{S}'\boldsymbol{S}\boldsymbol{U}$ we find that this equals

$$\begin{aligned}&= \operatorname{tr}\left[\left(\boldsymbol{S}'\boldsymbol{S}\right)\left(\boldsymbol{S}'\boldsymbol{S}\right)\right] - 2\frac{\boldsymbol{T}'\boldsymbol{S}'\boldsymbol{S}\boldsymbol{T}}{\boldsymbol{Z}'\boldsymbol{Z}} + \left(\frac{\boldsymbol{T}'\boldsymbol{T}}{\boldsymbol{Z}'\boldsymbol{Z}}\right)^2 \\ &+ \operatorname{tr}\left[\widehat{\boldsymbol{X}}'\widehat{\boldsymbol{X}}\boldsymbol{U}'\boldsymbol{U}\widehat{\boldsymbol{X}}'\widehat{\boldsymbol{X}}\boldsymbol{U}'\boldsymbol{U}\right] + 2\operatorname{tr}\left[\boldsymbol{V}'\boldsymbol{U}\boldsymbol{V}'\boldsymbol{U}\right] + 2\operatorname{tr}\left[\boldsymbol{U}'\boldsymbol{W}\widehat{\boldsymbol{X}}'\widehat{\boldsymbol{X}}\right] - 2\frac{\boldsymbol{T}'\boldsymbol{U}\widehat{\boldsymbol{X}}'\widehat{\boldsymbol{X}}\boldsymbol{U}'\boldsymbol{T}}{\boldsymbol{Z}'\boldsymbol{Z}} \\ &- 4\operatorname{tr}\left[\boldsymbol{W}'\boldsymbol{V}\right] + 4\frac{\boldsymbol{T}'\boldsymbol{V}\boldsymbol{U}'\boldsymbol{T}}{\boldsymbol{Z}'\boldsymbol{Z}} - 4\operatorname{tr}\left[\boldsymbol{U}'\boldsymbol{U}\widehat{\boldsymbol{X}}'\widehat{\boldsymbol{X}}\boldsymbol{U}'\boldsymbol{V}\right] + 2\operatorname{tr}\left[\boldsymbol{U}'\boldsymbol{U}\boldsymbol{V}'\boldsymbol{V}\right]\end{aligned}$$

which equals (35). ■

References

[1] Andrews, Donald W. K. (1991): "Asymptotic normality of series estimators for nonparameric 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] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

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

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

[10] Chesher, Andrew D. and Ian D. Jewitt (1987): "The bias of the heteroskedasticity consistent covariance matrix estimator," *Econometrica*, 55, l217-1272.

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

[12] 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.

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

[14] 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.

[15] Duchesne, Pierre and Pierre Lafaye de Micheaux (2010): "Computing the distribution of quadratic forms: Further comparisons between the Liu-Tang-Zhang approximation and exact methods," *Computational Statistics and Data Analysis*, 54, 858-862.

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

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

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

[19] 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.

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

[21] Hagemann, Andreas (2023): "Inference with a single treated cluster," working paper.

[22] Hansen, Bruce E. (2022): *Probability and Statistics for Economists*, Princeton University Press.

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

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

[25] 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.

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

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

[28] Imhof, J. P. (1961): "Computing the distribution of quadratic forms in normal variables," *Biometrika*, 48, 419-426.

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

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

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

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

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

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

[35] 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.

[36] 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*.

[37] 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*, forthcoming.

[38] Mathew, Thomas, and Kenneth Nordström (1997): "Inequalities for the probability content of a rotated ellipse and related stochastic domination results," *The Annals of Applied Probability*, 7, 1106-1117.

[39] 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.

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

[41] Pötscher, Benedikt M. and David Preinerstorfer (2023): "Valid heteroskedasticity robust testing," *Econometric Theory*.

[42] 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.

[43] 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.

[44] 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.

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

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

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

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

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

[50] Zhang, Fuzhen and Qingling Zhang (2006): "Eigenvalue inequalities for matrix product," *IEEE Transactions on Automatic Control*, 51, 1506-1509.