Queen's Economics Department Working Paper No. 1516

Jackknife Inference with Two-Way Clustering

James G. MacKinnon
Queen's University

Morten Ørregaard Nielsen
Aarhus University

Matthew D. Webb
Carleton University

Department of Economics
Queen's University
94 University Avenue
Kingston, Ontario, Canada
K7L 3N6

5-2024

Jackknife Inference with Two-Way Clustering *

James G. MacKinnon†
Queen's University
mackinno@queensu.ca

Morten Ørregaard Nielsen
Aarhus University
mon@econ.au.dk

Matthew D. Webb
Carleton University
matt.webb@carleton.ca

June 13, 2024

Abstract

For linear regression models with cross-section or panel data, it is natural to assume that the disturbances are clustered in two dimensions. However, the finite-sample properties of two-way cluster-robust tests and confidence intervals are often poor. We discuss several ways to improve inference with two-way clustering. Two of these are existing methods for avoiding, or at least ameliorating, the problem of undefined standard errors when a cluster-robust variance matrix estimator (CRVE) is not positive definite. One is a new method that always avoids the problem. More importantly, we propose a family of new two-way CRVEs based on the cluster jackknife. Simulations for models with two-way fixed effects suggest that, in many cases, the cluster-jackknife CRVE combined with our new method yields surprisingly accurate inferences. We provide a simple software package, `twowayjack` for `Stata`, that implements our recommended variance estimator.

Keywords: cluster jackknife, cluster sizes, clustered data, cluster-robust variance estimator, CRVE, grouped data, two-way fixed effects.

JEL Codes: C10, C12, C21, C23.

*We are grateful to conference and seminar participants at the Canadian Econometric Study Group, University of Graz, UCLA, and the Aarhus Workshop in Econometrics. MacKinnon and Webb thank the Social Sciences and Humanities Research Council of Canada (SSHRC grant 435-2021-0396) for financial support. Nielsen thanks the Danish National Research Foundation for financial support (DNRF Chair grant number DNRF154). Code and data files may be found at http://qed.econ.queensu.ca/pub/faculty/mackinnon/twowayjack/

†Corresponding author.

1 Introduction

The use of two-way cluster-robust variance estimators for linear regression models was independently proposed by Miglioretti and Heagerty (2006), Cameron, Gelbach, and Miller (2011), and Thompson (2011). Although it has been widely used in empirical work, the asymptotic theory to justify two-way clustering is quite recent. See, among others, Davezies, D'Haultfœuille, and Guyonvarch (2021), MacKinnon, Nielsen, and Webb (2021), Menzel (2021), Chiang, Kato, Ma, and Sasaki (2022), Chiang, Kato, and Sasaki (2023), Chiang, Hansen, and Sasaki (2023), and Yap (2024). The finite-sample properties of statistical inference are much less well understood for two-way clustering than for one-way clustering. For an up-to-date discussion of the latter, with recommendations for empirical practice, see MacKinnon, Nielsen, and Webb (2023a).

The jackknife variance estimator has been around for a very long time (Tukey 1958; Efron 1981; Efron and Stein 1981). The cluster jackknife CRVE (sometimes called the CV_3 estimator) for linear regression models with one-way clustering was proposed in Bell and McCaffrey (2002) and has been available in `Stata` for many years. Nevertheless, it has not been studied or applied much until very recently. In part, this is because Bell and McCaffrey (2002) followed MacKinnon and White (1985) by computing the CV_3 estimator in a way that is efficient when all clusters are very small but extremely inefficient when any clusters are large; see MacKinnon, Nielsen, and Webb (2023b). This seems to have given many investigators the erroneous impression that CV_3 is very expensive to compute, even though the `Stata` implementation uses a method that is reasonably efficient when the number of clusters is not too large. However, it is not as efficient as the method discussed in MacKinnon, Nielsen, and Webb (2023c) and implemented in the `Stata` package `summclust` (MacKinnon, Nielsen, and Webb 2023d).

In Section 2, we discuss the linear regression model with two-way clustering. Two existing CRVEs are discussed, along with their theoretical and practical deficiencies. For the CRVE that is theoretically soundest, the chief deficiency is that it may not be positive definite in finite samples. We discuss two ways to overcome this problem. One is the eigen-decomposition method suggested in Cameron, Gelbach, and Miller (2011). The other is a new and extremely simple procedure which can readily be implemented using existing software.

In Section 3, we show how to extend the cluster jackknife CRVEs discussed in MacKinnon, Nielsen, and Webb (2023b) and Hansen (2023) to two-way clustering. Based on what is known about the finite-sample performance of cluster jackknife CRVEs for one-way clustering, it seems very likely that inference based on them will be more conservative, and usually more reliable, than conventional inference in the two-way case as well. Some theoretical arguments to support this conjecture are provided in Section 4.

In Section 5, we use simulation experiments to study the finite-sample performance of several procedures for inference. Using the cluster jackknife methods of Section 3 in combination with either of the procedures discussed in Section 2 often performs much better than existing methods

for cluster-robust inference. In Section 6, we apply several methods to two empirical examples. The results that we obtain are entirely consistent with the simulations in Section 5. We conclude that, while conventional methods probably do not yield reliable inferences for these examples, our preferred methods based on the cluster jackknife probably do. Finally, Section 7 concludes.

2 Cluster-Robust Variance Estimation in Two Dimensions

Consider the linear regression model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}, \tag{1}$$

where $\boldsymbol{y}$ and $\boldsymbol{u}$ are $N \times 1$ vectors of observations and disturbances, $\boldsymbol{X}$ is an $N \times k$ matrix of covariates, and $\boldsymbol{\beta}$ is a $k \times 1$ parameter vector. The model is assumed to have two dimensions of clustering, where the numbers of clusters in the two dimensions are G and H, respectively. It is illuminating to rewrite (1) in terms of the intersections of the two clustering dimensions:

$$\boldsymbol{y}_{gh} = \boldsymbol{X}_{gh}\boldsymbol{\beta} + \boldsymbol{u}_{gh}, \quad g = 1, \dots, G, \quad h = 1, \dots, H. \tag{2}$$

Here the vectors $\boldsymbol{y}_{gh}$ and $\boldsymbol{u}_{gh}$ and the matrix $\boldsymbol{X}_{gh}$ contain, respectively, the rows of $\boldsymbol{y}$, $\boldsymbol{u}$, and $\boldsymbol{X}$ that correspond to both the g^{th} cluster in the first clustering dimension and the h^{th} cluster in the second clustering dimension. Similarly, we use $\boldsymbol{y}_g$, $\boldsymbol{X}_g$, and $\boldsymbol{u}_g$ to denote vectors that contain the rows of $\boldsymbol{y}$, $\boldsymbol{X}$, and $\boldsymbol{u}$ for the g^{th} cluster in the first dimension, and $\boldsymbol{y}_h$, $\boldsymbol{X}_h$, and $\boldsymbol{u}_h$ to denote the corresponding rows for the h^{th} cluster in the second dimension. The vector $\boldsymbol{y}_g$ contains the subvectors $\boldsymbol{y}_{g1}$ through $\boldsymbol{y}_{gH}$.

We use N_g to denote the number of observations in cluster g for the first dimension, N_h to denote the number of observations in cluster h for the second dimension, and N_{gh} to denote the number of observations in the intersection of cluster g in the first dimension with cluster h in the second dimension. We assume that $N_g \geq 1$ and $N_h \geq 1$. Thus, the number of observations in the entire sample is

$$N = \sum_{g=1}^{G} N_g = \sum_{h=1}^{H} N_h = \sum_{g=1}^{G}\sum_{h=1}^{H} N_{gh}.$$

Note that some of the intersections may be empty, so that N_{gh} might well equal 0 for some values of g and h. The number of non-empty intersections is $I \leq GH$.

Various score vectors play key roles in cluster-robust inference. The score vector for the entire sample is $\boldsymbol{s} = \boldsymbol{X}^\top\boldsymbol{u}$. The score subvector for cluster g in the first dimension is $\boldsymbol{s}_g = \boldsymbol{X}_g^\top\boldsymbol{u}_g$, and the score subvector for cluster h in the second dimension is $\boldsymbol{s}_h = \boldsymbol{X}_h^\top\boldsymbol{u}_h$. Thus there are G score vectors $\boldsymbol{s}_g$ and H score vectors $\boldsymbol{s}_h$. The score subvector for intersection gh is $\boldsymbol{s}_{gh} = \boldsymbol{X}_{gh}^\top\boldsymbol{u}_{gh}$.

The variance matrix of the scores can always be written as

$$\boldsymbol{\Sigma} = \mathrm{E}(\boldsymbol{X}^\top\boldsymbol{u}\boldsymbol{u}^\top\boldsymbol{X}) = \sum_{g,g'=1}^{G}\sum_{h,h'=1}^{H} \mathrm{E}\Big(\boldsymbol{s}_{gh}\boldsymbol{s}_{g'h'}^\top\Big). \tag{3}$$

Under two-way clustering, it must be the case that

$$\mathrm{E}(\boldsymbol{s}_{gh}\boldsymbol{s}_{g'h'}^\top) = \mathbf{0} \quad \text{if} \quad g' \neq g \text{ and } h' \neq h, \tag{4}$$

but the covariances may be arbitrary when either $g = g'$ or $h = h'$. The variance matrices for the score subvectors $\boldsymbol{s}_g$, $\boldsymbol{s}_h$, and $\boldsymbol{s}_{gh}$ are respectively denoted

$$\boldsymbol{\Sigma}_g = \mathrm{E}(\boldsymbol{s}_g\boldsymbol{s}_g^\top), \quad \boldsymbol{\Sigma}_h = \mathrm{E}(\boldsymbol{s}_h\boldsymbol{s}_h^\top), \quad \text{and} \quad \boldsymbol{\Sigma}_{gh} = \mathrm{E}(\boldsymbol{s}_{gh}\boldsymbol{s}_{gh}^\top). \tag{5}$$

From (4) and (5), it is evident that

$$\boldsymbol{\Sigma} = \sum_{g=1}^{G} \boldsymbol{\Sigma}_g + \sum_{h=1}^{H} \boldsymbol{\Sigma}_h - \sum_{g=1}^{G}\sum_{h=1}^{H} \boldsymbol{\Sigma}_{gh}. \tag{6}$$

This follows from the inclusion-exclusion principle. The third term in (6) is essential to avoid double-counting, but, as we shall see, it causes practical difficulties for estimating $\boldsymbol{\Sigma}$.

As usual, the OLS estimator of $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}$, and the OLS residual vector is $\hat{\boldsymbol{u}}$. The subvectors of $\hat{\boldsymbol{u}}$ for cluster g, cluster h, and the intersection gh are denoted $\hat{\boldsymbol{u}}_g$, $\hat{\boldsymbol{u}}_h$, and $\hat{\boldsymbol{u}}_{gh}$, respectively. From standard arguments for sandwich variance matrices,

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{\Sigma}(\boldsymbol{X}^\top\boldsymbol{X})^{-1} = \boldsymbol{V}_G + \boldsymbol{V}_H - \boldsymbol{V}_I, \tag{7}$$

where the component matrices are

$$\boldsymbol{V}_G = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\left(\sum_{g=1}^{G} \boldsymbol{\Sigma}_g\right)(\boldsymbol{X}^\top\boldsymbol{X})^{-1} \tag{8}$$

$$\boldsymbol{V}_H = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\left(\sum_{h=1}^{H} \boldsymbol{\Sigma}_h\right)(\boldsymbol{X}^\top\boldsymbol{X})^{-1}, \text{ and} \tag{9}$$

$$\boldsymbol{V}_I = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\left(\sum_{g=1}^{G}\sum_{h=1}^{H} \boldsymbol{\Sigma}_{gh}\right)(\boldsymbol{X}^\top\boldsymbol{X})^{-1}. \tag{10}$$

The empirical analog of (7) is the three-term two-way CRVE

$$\hat{\boldsymbol{V}}_1^{(3)} = \hat{\boldsymbol{V}}_G + \hat{\boldsymbol{V}}_H - \hat{\boldsymbol{V}}_I, \tag{11}$$

where the subscript "1" identifies this as a CV_1 estimator, by analogy with the HC_1 estimator of MacKinnon and White (1985). The three variance estimators on the right-hand side of (11) correspond to (8), (9), and (10). The natural way to estimate them is to use the empirical score subvectors $\hat{\boldsymbol{s}}_g$, $\hat{\boldsymbol{s}}_h$, and $\hat{\boldsymbol{s}}_{gh}$, which take the same form as the actual score subvectors, but with $\hat{\boldsymbol{u}}$

replacing $\boldsymbol{u}$. We obtain the CV_1 estimators

$$\hat{\boldsymbol{V}}_G = \frac{G(N-1)}{(G-1)(N-k)}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\left(\sum_{g=1}^{G}\hat{\boldsymbol{s}}_g\hat{\boldsymbol{s}}_g^\top\right)(\boldsymbol{X}^\top\boldsymbol{X})^{-1}, \tag{12}$$

$$\hat{\boldsymbol{V}}_H = \frac{H(N-1)}{(H-1)(N-k)}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\left(\sum_{h=1}^{H}\hat{\boldsymbol{s}}_h\hat{\boldsymbol{s}}_h^\top\right)(\boldsymbol{X}^\top\boldsymbol{X})^{-1}, \quad \text{and} \tag{13}$$

$$\hat{\boldsymbol{V}}_I = \frac{I(N-1)}{(I-1)(N-k)}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\left(\sum_{g=1}^{G}\sum_{h=1}^{H}\hat{\boldsymbol{s}}_{gh}\hat{\boldsymbol{s}}_{gh}^\top\right)(\boldsymbol{X}^\top\boldsymbol{X})^{-1}. \tag{14}$$

The leading scalar factors here are analogous to the scalar factor for the usual one-way CRVE. Since some of the intersections may contain no observations, some of the $\hat{\boldsymbol{s}}_{gh}$ may not exist. In practice, it may therefore be advisable to replace the double summation in (14) with a single summation over all non-empty intersections.

The superscript "(3)" on $\hat{\boldsymbol{V}}_1^{(3)}$ in (11) emphasizes that this estimator has three terms, which correspond to the three terms in (6). Because $\hat{\boldsymbol{V}}_I$ is subtracted from the sum of $\hat{\boldsymbol{V}}_G$ and $\hat{\boldsymbol{V}}_H$, the matrix $\hat{\boldsymbol{V}}_1^{(3)}$ is not necessarily positive definite in finite samples. This problem is not trivial, and there is more than one way to deal with it.

One approach, suggested in Cameron, Gelbach, and Miller (2011) and implemented in `Stata 18`, is to compute the eigenvalues of $\hat{\boldsymbol{V}}_1^{(3)}$, say $\lambda_1, \ldots, \lambda_k$. When any of them is not positive, $\hat{\boldsymbol{V}}_1^{(3)}$ is replaced by the eigen-decomposition $\hat{\boldsymbol{V}}_1^{(3+)} = \boldsymbol{U}\boldsymbol{\Lambda}^+\boldsymbol{U}^\top$, where $\boldsymbol{U}$ is the $k \times k$ matrix of eigenvectors and $\boldsymbol{\Lambda}^+$ is a diagonal matrix with typical diagonal element $\lambda_j^+ = \max\{\lambda_j, 0\}$. In practice, it may be numerically safer to compare the eigenvalues with a very small positive number, say η, and define λ_j^+ as $\max\{\lambda_j, \eta\}$. In our programs, we use $\eta = 10^{-12}$. Doing this ensures that $\hat{\boldsymbol{V}}_1^{(3+)}$ is positive definite, albeit just barely so.

This approach is not entirely satisfactory. Wald statistics and t-statistics based on $\hat{\boldsymbol{V}}_1^{(3+)}$ are computable, but they may be extremely large. Even when this does not happen, and all quantities of interest can be computed using $\hat{\boldsymbol{V}}_1^{(3)}$, replacing $\hat{\boldsymbol{V}}_1^{(3)}$ by $\hat{\boldsymbol{V}}_1^{(3+)}$ can change all the standard errors. Moreover, the standard error of any element of $\hat{\boldsymbol{\beta}}$, say $\hat{\beta}_j$, is not invariant to nonsingular transformations of the remaining columns of the matrix $\boldsymbol{X}$. Thus, for example, precisely how fixed effects or other dummy variables are specified may affect the standard error of $\hat{\beta}_j$, even though $\hat{\beta}_j$ itself is invariant to such reparametrizations.

A simpler way to avoid the problem that $\hat{\boldsymbol{V}}_1^{(3)}$ may not be positive definite is to replace it by the two-term estimator

$$\hat{\boldsymbol{V}}_1^{(2)} = \hat{\boldsymbol{V}}_G + \hat{\boldsymbol{V}}_H. \tag{15}$$

This estimator has been studied in Davezies, D'Haultfœuille, and Guyonvarch (2021). It omits the third term in (11) and therefore involves double-counting. The justification for omitting $\hat{\boldsymbol{V}}_I$ is that, under a strong regularity condition, it becomes asymptotically negligible as both G and

H tend to infinity. Because

$$\hat{\boldsymbol{V}}_1^{(3)} - \hat{\boldsymbol{V}}_1^{(2)} = \hat{\boldsymbol{V}}_I \tag{16}$$

is positive definite, it follows that a Wald statistic or t-statistic based on $\hat{\boldsymbol{V}}_1^{(2)}$ will always be larger than the same statistic based on $\hat{\boldsymbol{V}}_1^{(3)}$.

Unfortunately, the conditions for $\hat{\boldsymbol{V}}_1^{(2)}$ to yield asymptotically valid inferences are much stronger than the ones needed for $\hat{\boldsymbol{V}}_1^{(3)}$ to do so. MacKinnon, Nielsen, and Webb (2021) considers three cases and shows that $\hat{\boldsymbol{V}}_1^{(2)}$ yields asymptotically valid inferences in only one of them. Whenever the scores are actually independent, or whenever they are only correlated at the intersection level, $\hat{\boldsymbol{V}}_1^{(2)}$ yields test statistics that are asymptotically too small. In this case, $\hat{\boldsymbol{V}}_G \approx \hat{\boldsymbol{V}}_H \approx \hat{\boldsymbol{V}}_I$. Therefore,

$$\hat{\boldsymbol{V}}_1^{(2)} = \hat{\boldsymbol{V}}_G + \hat{\boldsymbol{V}}_H \approx 2\hat{\boldsymbol{V}}_I, \tag{17}$$

whereas

$$\hat{\boldsymbol{V}}_1^{(3)} = \hat{\boldsymbol{V}}_G + \hat{\boldsymbol{V}}_H - \hat{\boldsymbol{V}}_I \approx \hat{\boldsymbol{V}}_I. \tag{18}$$

Thus, in this case, $\hat{\boldsymbol{V}}_1^{(2)}$ is approximately twice as large as $\hat{\boldsymbol{V}}_1^{(3)}$, and twice as large as it should be. The use of "$\approx$" in (17) and (18) is deliberately informal, since we did not take limits or introduce any factors of the sample size in (7). For a rigorous treatment, see MacKinnon, Nielsen, and Webb (2021, Theorem 1). The result (17) suggests that $\hat{\boldsymbol{V}}_1^{(2)}$ is also likely to perform poorly in finite samples when much of the intra-cluster correlation is at the intersection level.

We now propose a third way to avoid cases in which test statistics based on the three-term estimator $\hat{\boldsymbol{V}}_1^{(3)}$ are not positive. Our proposal is simply to compute three test statistics and use the one that takes the smallest positive value. For the hypothesis that $\boldsymbol{R\beta} = \boldsymbol{r}$, the three Wald statistics are

$$\begin{aligned} W_3 &= (\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r})^\top (\boldsymbol{R}\hat{\boldsymbol{V}}_1^{(3)}\boldsymbol{R}^\top)^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r}), \\ W_G &= (\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r})^\top (\boldsymbol{R}\hat{\boldsymbol{V}}_G\boldsymbol{R}^\top)^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r}), \text{ and} \\ W_H &= (\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r})^\top (\boldsymbol{R}\hat{\boldsymbol{V}}_H\boldsymbol{R}^\top)^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r}). \end{aligned} \tag{19}$$

The statistic we propose to use is

$$W_{\min} = \min\Big\{\max\{W_3, 0\}, W_G, W_H\Big\}, \tag{20}$$

where $\max\{W_3, 0\}$ equals 0 whenever W_3 is either negative or undefined, as it can be when $\hat{\boldsymbol{V}}_1^{(3)}$ is not positive definite. By using $W_{\min}$ defined in (20), we not only avoid Wald statistics that are not positive numbers but also Wald statistics that are misleadingly large. In a particular sample, one or more diagonal elements of $\hat{\boldsymbol{V}}_I$ may randomly happen to be just a little smaller than the sum of the corresponding elements of $\hat{\boldsymbol{V}}_G$ and $\hat{\boldsymbol{V}}_H$. Thus $\hat{\boldsymbol{V}}_1^{(3)}$ can yield extremely large test statistics which are completely misleading.

In most cases, it is not necessary to calculate the entire $\hat{\boldsymbol{V}}_1^{(3)}$ matrix. Only the rows and columns needed for the Wald statistic have to be calculated. Of course, when there is only one

restriction, we can use a t-statistic instead of a Wald statistic. In this case, we just need to find the largest of the three standard errors and calculate a t-statistic using that standard error. Since this is by far the most common case, we will refer to our procedure as the "max-se" procedure.

Henceforth, we denote the variance and standard error estimators based on $\hat{\boldsymbol{V}}_1^{(2)}$ and $\hat{\boldsymbol{V}}_1^{(3)}$ as $\text{CV}_1^{(2)}$ and $\text{CV}_1^{(3)}$ estimators, respectively, the ones based on $\hat{\boldsymbol{V}}_1^{(3+)}$ as $\text{CV}_1^{(3+)}$ estimators, and the ones implicit in (20) as $\text{CV}_1^{(\max)}$ estimators. In the scalar case, $\text{CV}_1^{(\max)} = \max\{\hat{V}_1^{(3)}, \hat{V}_G, \hat{V}_H\}$. This explains the "(max)" superscript and also makes it clear that, asymptotically, the $\text{CV}_1^{(3)}$, $\text{CV}_1^{(3+)}$, and $\text{CV}_1^{(\max)}$ estimators must be identical whenever the scores are positively correlated in either or both of the G and H dimensions.

In most cases where it makes sense to specify $\boldsymbol{\Sigma}$ as in (3), the $\text{CV}_1^{(3)}$, $\text{CV}_1^{(3+)}$, and $\text{CV}_1^{(\max)}$ estimators will have exactly the same asymptotic properties. They may or may not be identical in practice. In fact, there are cases where they may differ greatly. This seems to be most common when there is very little intra-cluster correlation and/or the number of clusters is small, and/or the number of regressors is large, as we shall see in Section 5.

3 Two-Way Cluster Jackknife CRVEs

The component CRVEs defined in (12), (13), and (14) all have the form of the widely-used CV_1 estimator. However, recent work by MacKinnon, Nielsen, and Webb (2023b) and Hansen (2023) strongly suggests that, in the one-way case, it is better to use a CRVE based on the cluster jackknife, which is analogous to the HC_3 estimator of MacKinnon and White (1985). The key idea of the cluster jackknife is to compute G (or H or I) sets of parameter estimates, each of which omits one cluster at a time, and then compute a CRVE using the variation among these estimates.

Let $J \in \{G, H, I\}$, and let j denote the corresponding lower-case letter. The OLS estimates of $\boldsymbol{\beta}$ when each cluster in the J dimension is omitted in turn are

$$\hat{\boldsymbol{\beta}}^{(j)} = (\boldsymbol{X}^\top\boldsymbol{X} - \boldsymbol{X}_j^\top\boldsymbol{X}_j)^{-1}(\boldsymbol{X}^\top\boldsymbol{y} - \boldsymbol{X}_j^\top\boldsymbol{y}_j), \quad j = 1, \ldots, J. \tag{21}$$

Then the component cluster jackknife variance matrix estimators are

$$\hat{\boldsymbol{V}}_J^{\text{JK}} = \frac{J-1}{J}\sum_{j=1}^{J}(\hat{\boldsymbol{\beta}}^{(j)} - \hat{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}}^{(j)} - \hat{\boldsymbol{\beta}})^\top \quad \text{for } \{j, J\} = \{g, G\}, \{h, H\}, \{i, I\}. \tag{22}$$

Thus the three-term jackknife CRVE is

$$\hat{\boldsymbol{V}}_3^{(3)} = \hat{\boldsymbol{V}}_G^{\text{JK}} + \hat{\boldsymbol{V}}_H^{\text{JK}} - \hat{\boldsymbol{V}}_I^{\text{JK}}, \tag{23}$$

which is analogous to (11). The subscript "3" here follows the usual notation for jackknife variance matrices; see MacKinnon, Nielsen, and Webb (2023b). There is also a two-term jackknife CRVE and, more interestingly, one that is analogous to the $\text{CV}_1^{(\max)}$ estimator. We refer to the three CRVEs based on the cluster jackknife as $\text{CV}_3^{(2)}$, $\text{CV}_3^{(3)}$, and $\text{CV}_3^{(\max)}$.

The CRVEs defined in (22) are not the only cluster jackknife variance matrix estimators. Instead of computing variances around $\hat{\boldsymbol{\beta}}$, one can instead compute them around (in the two-way case) the three sample averages, $\bar{\boldsymbol{\beta}}^J = J^{-1}\sum_{j=1}^J \hat{\boldsymbol{\beta}}^{(j)}$. This makes the alternative CRVEs a little smaller than the ones given in (22). Because simulation experiments in Bell and McCaffrey (2002) and MacKinnon, Nielsen, and Webb (2023b) suggest that, in the one-way case with G clusters, inferences based on the alternative jackknife CRVE are almost identical to ones based on $\hat{\boldsymbol{V}}_G^{\text{JK}}$, we do not study the former in this paper.

Computing the component CRVEs in (22) that are needed for $\text{CV}_3^{(3)}$, $\text{CV}_3^{(3+)}$, and $\text{CV}_3^{(\max)}$ is more work than computing the ones in (12), (13), and (14) that are needed for $\text{CV}_1^{(3)}$, $\text{CV}_1^{(3+)}$, and $\text{CV}_1^{(\max)}$, especially when the number of non-empty intersections, I, is large. Nevertheless, it should be manageable in most cases. The first thing is to calculate the cluster-level matrices and vectors

$$\boldsymbol{X}_j^\top \boldsymbol{X}_j \text{ and } \boldsymbol{X}_j^\top \boldsymbol{y}_j, \quad j = 1, \ldots, J, \quad \text{for } \{j, J\} = \{g, G\}, \{h, H\}, \{i, I\}. \tag{24}$$

These quantities can be computed for the intersections with a single pass over the N observations. The ones for the G and H dimensions are just summations of the ones for the appropriate intersections. The three sets of $\hat{\boldsymbol{\beta}}^{(j)}$ can then be computed using (21) for the three clustering dimensions. Unfortunately, this may be expensive when both k and I are large, because computing the delete-one-cluster estimates for the intersections involves inverting I different $k \times k$ matrices.

In many cases, the regression model (1) will include fixed effects in the G and H dimensions; that is, two-way fixed effects. If so, it may be rewritten as

$$\boldsymbol{y} = \boldsymbol{Z}\boldsymbol{\beta}_p + \boldsymbol{D}^G\boldsymbol{\gamma} + \boldsymbol{D}^H\boldsymbol{\delta} + \boldsymbol{u}. \tag{25}$$

Here the matrix $\boldsymbol{Z}$, which has p columns, corresponds to the actual explanatory variables (which should not include a constant term), and $\boldsymbol{\beta}_p$ contains the elements of $\boldsymbol{\beta}$ for those variables. The matrices $\boldsymbol{D}^G$ and $\boldsymbol{D}^H$ contain dummy variables for the fixed effects in dimensions G and H, respectively. Collectively, these have $G + H - 1$ columns, say G for $\boldsymbol{D}^G$ and $H - 1$ for $\boldsymbol{D}^H$. Thus $\boldsymbol{X} = [\boldsymbol{Z} \;\; \boldsymbol{D}^G \;\; \boldsymbol{D}^H]$, and $k = p + G + H - 1$.

For the model (25), there is an important computational issue. It is impossible to invert the matrices $\boldsymbol{X}^\top\boldsymbol{X} - \boldsymbol{X}_g^\top\boldsymbol{X}_g$ and $\boldsymbol{X}^\top\boldsymbol{X} - \boldsymbol{X}_h^\top\boldsymbol{X}_h$ in (24), because for each of them the row and column corresponding to the fixed effect for cluster g or cluster h contains only zeros. There are two ways to deal with this issue. The simplest is just to replace the inverse in (21) by a generalized inverse. Then all of the coefficients except the fixed effect for the omitted cluster can be computed, and the latter is set to zero. Thus, whenever there are two-way fixed effects, $\hat{\boldsymbol{V}}_{\text{JK}}^{(3)}$ in (23) can only be calculated as a $p \times p$ matrix instead of a $k \times k$ matrix.

Instead of using a generalized inverse, it seems natural to partial out the cluster fixed effects before computing the one-way CRVEs. However, this must be done with great care. It is valid

to partial out cluster fixed effects in the G dimension when calculating $\hat{\boldsymbol{V}}_G^{\text{JK}}$, but it is invalid to partial them out when calculating either $\hat{\boldsymbol{V}}_H^{\text{JK}}$ or $\hat{\boldsymbol{V}}_I^{\text{JK}}$. The problem is that, after the cluster fixed effects in the G dimension have been partialed out, the observations for every cluster in the H and I dimensions generally depend on observations in some or all of the other clusters in those dimensions. Thus $\hat{\boldsymbol{\beta}}^{(h)}$ and $\hat{\boldsymbol{\beta}}^{(i)}$ would not actually be vectors of delete-one-cluster estimates. Similarly, it is invalid to partial out fixed effects in the H dimension when calculating either $\hat{\boldsymbol{V}}_G^{\text{JK}}$ or $\hat{\boldsymbol{V}}_I^{\text{JK}}$.

The I dimension is always the most expensive one to deal with, because it involves the largest number of clusters, and it is not valid to partial out fixed effects in either the G or H dimensions when calculating $\hat{\boldsymbol{V}}_I^{\text{JK}}$. Thus, even though it would be possible to partial out the fixed effects for dimension G when computing the $\hat{\boldsymbol{\beta}}^{(g)}$ and the fixed effects for dimension H when computing the $\hat{\boldsymbol{\beta}}^{(h)}$, it is probably not worth the additional programming complexity.

It is conventional to employ the Student's t distribution with $\min\{G, H\} - 1$ degrees of freedom to obtain P values or critical values for t-statistics based on $\text{CV}_1^{(3)}$. As in the one-way case, it seems reasonable to use the same distribution for t-statistics based on $\text{CV}_3^{(3)}$ as well, and this is the approach that we take.

However, at least two other approaches could in principle be used. For one-way clustering, Bell and McCaffrey (2002) proposed a way to obtain approximate critical values for CV_1-based t-tests using a t distribution with an estimated degrees-of-freedom parameter; see also Imbens and Kolesár (2016). For the two-way case, one could in principle use the same sort of approximate critical value. However, we are not aware of any method for obtaining such a critical value for t-statistics based on two-way clustering. This is an area for future research.

The wild cluster bootstrap (Cameron, Gelbach, and Miller 2008; Djogbenou, MacKinnon, and Nielsen 2019) has been widely used for inference with one-way clustering, and MacKinnon, Nielsen, and Webb (2021) suggested using it for two-way clustering as well. That paper simply uses the usual wild cluster bootstrap for one of the G, H, or I dimensions to generate the bootstrap samples and computes $\text{CV}_1^{(3)}$-based t-tests for both the actual and bootstrap samples. More recently, Hounyo and Lin (2024) proposes a wild bootstrap DGP that gives positive weight to both dimensions. All these bootstrap methods usually lead to more reliable inferences than simply using Student's t critical values, but they do not always perform particularly well, and none of them is theoretically satisfactory. No existing wild bootstrap DGP for models with two-way clustering can actually replicate the intra-cluster covariances in the residuals, because it appears to be impossible to do so.

In the absence of any satisfactory alternative, we currently recommend using the cluster jackknife together with critical values based on the Student's t distribution with $\min\{G, H\} - 1$ degrees of freedom. As we shall see in Section 5, this approach often works remarkably well. Whether combining the jackknife with a bootstrap procedure would perform even better is a topic

for future research; see MacKinnon et al. (2023b) for evidence on this with one-way clustering.

Computing the three-term cluster-jackknife estimator for the two-way fixed-effects model (25) can be costly when G and H are not fairly small. The cost of forming the $\boldsymbol{X}_j^\top \boldsymbol{X}_j$ matrices and the $\boldsymbol{X}_j^\top \boldsymbol{y}_j$ vectors is roughly $O(Nk^2) = O(N(G+H+p-1)^2)$, because $\boldsymbol{X}$ has $k = p+G+H-1$ columns. Since (21) has to be computed $G+H+I \approx G+H+GH$ times, the cost of computing the cluster-jackknife estimates after the $\boldsymbol{X}_j^\top \boldsymbol{X}_j$ matrices and $\boldsymbol{X}_j^\top \boldsymbol{y}_j$ vectors have been formed is roughly $O(GHk^2) = O(GH(G+H+p-1)^2) = O(G^4)$ if $G \approx H$. A few timing experiments suggest that this approximation may provide a reasonable guide in practice.

Much of the computational cost of the two-way cluster jackknife arises from the fact that there are GH intersections. When $I << GH$, the cost can be greatly reduced if the empty intersections are skipped when calculating the delete-one-cluster estimates using (21).

The largest values of G, H, and GH in the experiments of Section 5 are 45, 36, and 1620, respectively. With 90,000 observations and $p = 10$, each replication took about 1.7 seconds on one core of a 13^{th} generation Intel i9 processor using Fortran. The vast majority of the CPU time was used to compute the delete-one-cluster estimates, mainly for the intersections. Similar computations might take either more or less time using R or Stata, depending on how many cores could be effectively utilized. These timings, together with the result that computer time is approximately $O(G^4)$, suggest that it may not be feasible to compute three-term cluster-jackknife variance matrices for the model (25) when $\max\{G, H\}$ exceeds 300 or so, unless $\min\{G, H\}$ is a much smaller number or $I << GH$. Of course, $\text{CV}_1^{(\max)}$ will probably work well enough in such cases that more sophisticated methods are not needed.

Computing the three-term cluster-jackknife estimators may also be infeasible if p, the number of explanatory variables in (25), is too large. Even when G and H are not large, the $\boldsymbol{X}_j^\top \boldsymbol{X}_j$ matrices are $k \times k$, with $k = p + G + H - 1$. For large enough p, and hence k, the repeated computations in (21) can become burdensome. Since there is no reason to expect $\text{CV}_1^{(\max)}$ to work well in such cases, this is a more serious problem than having a large number of intersections.

4 Robustness of the Cluster Jackknife CRVE

In this section, we discuss why two-way cluster jackknife CRVEs perform better in finite samples than conventional CRVEs. The key reason seems to be that the former handle cluster size variation, and heterogeneity more generally, better than the latter do. This is particularly important for three-term estimators, as we explain.

Properties of classic jackknife variance estimators are well known. However, for the cluster jackknife, the only analysis of theoretical properties that we are aware of is Hansen (2023). In the context of the linear regression model with one-way clustering, it shows that a certain cluster jackknife variance estimator (which is not quite the same as $\widehat{\boldsymbol{V}}_3$, but should usually be very similar) is never downward biased. Moreover, the associated t-tests and confidence intervals

have worst-case size, or coverage, that is controlled by the Cauchy distribution. In contrast, variance estimators based on CV_1 can be severely downward biased, the associated t-tests have worst-case size of one, and the associated confidence intervals have worst-case coverage of zero.

It has been known for some time that the downward bias of one-way CV_1-based variance estimators is negatively associated with the number of clusters and positively associated with the extent to which they are heterogeneous in size and leverage (Djogbenou, MacKinnon, and Nielsen 2019; Boot, Niccodemi, and Wansbeek 2023). In many two-way designs, clusters vary greatly in size and/or leverage in one or both dimensions. Thus, when $\hat{\boldsymbol{V}}_G$ and $\hat{\boldsymbol{V}}_H$ are based on CV_1, one or both of them is likely to be seriously downward biased. However, because $\hat{\boldsymbol{V}}_I$ is normally based on a much larger number of clusters, its downward bias is likely to be comparatively moderate. In consequence, when $\hat{\boldsymbol{V}}_I$ is subtracted from the sum of $\hat{\boldsymbol{V}}_G$ and $\hat{\boldsymbol{V}}_H$ to form $\hat{\boldsymbol{V}}_1^{(3)}$, there is a good chance that the latter will be very severely biased.

In contrast, when $\hat{\boldsymbol{V}}_G$ and $\hat{\boldsymbol{V}}_H$ are based on CV_3, the results of Hansen (2023) suggest that neither of them is likely to suffer from much downward bias, although either or both may be upward biased. It is possible that $\hat{\boldsymbol{V}}_I$ may be upward biased in this case, but since it is normally based on a much larger number of clusters, any such bias is likely to be modest, and subtracting it is not likely to cause downward bias in $\hat{\boldsymbol{V}}_3^{(3)}$ itself. These arguments suggest that $\hat{\boldsymbol{V}}_3^{(3)}$ is more likely to be positive definite than $\hat{\boldsymbol{V}}_1^{(3)}$ and that tests based on $\hat{\boldsymbol{V}}_3^{(3)}$ should be more reliable than ones based on $\hat{\boldsymbol{V}}_1^{(3)}$.

The above arguments suggest that, if the sample is heterogeneous in only one dimension, so that only one of $\hat{\boldsymbol{V}}_G$ and $\hat{\boldsymbol{V}}_H$ is severely downward biased, then the downward bias in $\hat{\boldsymbol{V}}_1^{(3)}$ is likely to be relatively moderate. This case probably occurs quite often in panel settings, where samples (and cluster sizes) are often heterogeneous across cross-sectional units but homogeneous across time periods. In such cases, we would still expect CV_3-based estimators to be more accurate than CV_1-based ones, but probably by a smaller margin than in cases with double heterogeneity.

The reason why CV_3 estimators are less biased than CV_1 estimators can be seen as follows. Observe that the (one-way) cluster-jackknife CRVEs in (22) can be rewritten as

$$\hat{\boldsymbol{V}}_J^{\text{JK}} = \frac{J-1}{J}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\left(\sum_{j=1}^{J}\ddot{\boldsymbol{s}}_j\ddot{\boldsymbol{s}}_j^\top\right)(\boldsymbol{X}^\top\boldsymbol{X})^{-1} \quad \text{for } \{j,J\} = \{g,G\},\{h,H\},\{i,I\}, \tag{26}$$

where the modified score vectors $\ddot{\boldsymbol{s}}_j$ are defined as

$$\ddot{\boldsymbol{s}}_j = \boldsymbol{X}^\top\boldsymbol{M}_{jj}^{-1}\hat{\boldsymbol{u}}_j, \tag{27}$$

and $\boldsymbol{M}_{jj}$ denotes the $(j,j)^{\text{th}}$ block of $\boldsymbol{M}_{\boldsymbol{X}} = \mathbf{I}_N - \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top$. For a proof of equality of (22) and (26), see MacKinnon, Nielsen, and Webb (2023b, pp. 675–676). The modified score vectors in (27) are normalized by the factor $\boldsymbol{M}_{jj}^{-1}$ in order to undo some of the shrinkage caused by least squares. Since the $\boldsymbol{M}_{jj}$ are inversely related to cluster leverage (MacKinnon, Nielsen,

and Webb 2023c), the cluster jackknife CRVE puts more weight on clusters with high leverage compared with the CV_1 estimator. This accounts for the smaller bias of the former relative to the latter, because high-leverage clusters are relatively more important in determining the actual variance of the estimator.

In empirical research, it is very commonly found that some intersections of the two clustering dimensions contain no observations. The possibility of empty intersections can be important for two-way clustering, but it cannot arise for one-way clustering. To examine the importance of empty intersections, consider two hypothetical samples, each with $G = H = 10$. Thus there are 100 intersections. For one sample, no intersections are empty, but 70 of them contain just 1 observation. For the other sample, there are 70 empty intersections. Now consider the cluster jackknife estimator, $\hat{\boldsymbol{V}}_I^{\text{JK}}$. In the first sample, it is based on 100 terms. Since dropping just one observation should not change $\hat{\boldsymbol{\beta}}^{(i)}$ very much, the terms in the summation in (22) corresponding to the tiny intersections must all be very small. In the second sample, the cluster jackknife estimate is based on just 30 terms. The terms that were small in the first sample have vanished, which seems to be a small difference. The only other difference between the two samples is that the leading factor in $\hat{\boldsymbol{V}}_I^{\text{JK}}$ will be 99/100 in the first sample and 29/30 in the second, which seems inconsequential. Thus the cluster jackknife estimator handles empty intersections in a reasonable fashion.

5 Simulation Experiments

Almost all of our experiments deal with the two-way fixed-effects model (25). The number of coefficients is $k = p + G + H - 1$, but we focus on tests of a single coefficient, say β_1. Although (25) is very widely used, many existing simulation experiments for two-way clustering do not include cluster fixed effects. This is probably because, when the intra-cluster correlations are generated by a random-effects model, cluster fixed effects absorb all of them. For example, the experiments in Cameron, Gelbach, and Miller (2011, Section 3.1) and MacKinnon, Nielsen, and Webb (2021) do not include fixed effects. In contrast, the placebo-regression experiments in Section 3.2 of the former paper use actual data instead of a random-effects model, and they do include two-way fixed effects.

In order to generate data for the model (25), the disturbances must be generated in a way that allows for two-way intra-cluster correlation that is not removed by cluster fixed effects. We use factor models of the form

$$\begin{aligned} z_{ghi} &= \sigma_g \xi_g^1 + \sigma_h \xi_h^1 + \sigma_\epsilon \zeta_{ghi} \quad \text{if } i \text{ is odd}, \\ z_{ghi} &= \sigma_g \xi_g^2 + \sigma_h \xi_h^2 + \sigma_\epsilon \zeta_{ghi} \quad \text{if } i \text{ is even}. \end{aligned} \tag{28}$$

Here ξ_g^1 and ξ_g^2 are random effects, distributed as $\text{N}(0,1)$, which apply respectively to the odd-numbered and even-numbered observations within the g^{th} cluster in the G dimension. Similarly,

ξ_h^1 and ξ_h^2 are random effects which apply to the odd-numbered and even-numbered observations within the h^{th} cluster in the H dimension. The ζ_{ghi} are independent standard normals.

The values of σ_g, σ_h, and σ_ϵ determine the amount of correlation for the odd-numbered and even-numbered observations within each cluster, and hence the correlations within and across the clusters in the G, H, and I dimensions. There will be no correlation for observations that belong to different clusters in the G and H dimensions. Instead of specifying σ_g and σ_h directly, we specify them as functions of correlations ρ_g and ρ_h, with $\sigma_j = (\rho_j/(1-\rho_j))^{1/2}$ for $j = g, h$. To ensure that the z_{ghi} have variance unity, the value of σ_ϵ is $(1 - \sigma_g^2 - \sigma_h^2)^{1/2}$. This constrains ρ_g and ρ_h not to be too large.

The factor model (28) provides a simple way to generate data for a model with two-way fixed effects. It is based on a one-way DGP used in MacKinnon, Nielsen, and Webb (2023e) and can be interpreted in a variety of ways, depending on the nature of the data. The idea is that there are two types of observations within each cluster in each dimension, and all the intra-cluster correlation is within each type. For example, with clustering at the geographical level, there might be two sub-regions. With clustering at the industry level, there might be two types of firm. The key assumption is that the researcher knows which cluster an observation belongs to in each dimension, but not which type. Including cluster fixed effects explains some of the intra-cluster correlation by estimating averages of ξ_g^1 and ξ_g^2 for each G cluster and averages of ξ_h^1 and ξ_h^2 for each H cluster, but it does not explain all of it. Thus cluster-robust inference is still needed.

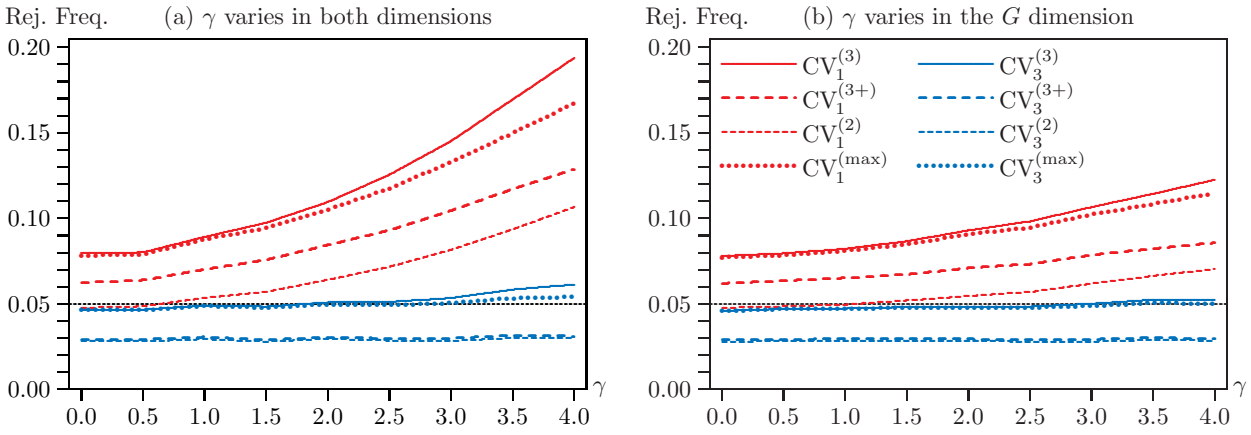
In several of the experiments, we focus on cluster size variation. Following MacKinnon and Webb (2017) and Djogbenou, MacKinnon, and Nielsen (2019), the cluster sizes in the G dimension are given by

$$N_g = \left[N \frac{\exp(\gamma g/G)}{\sum_{j=1}^{G} \exp(\gamma j/G)} \right], \quad g = 1, \ldots, G-1, \tag{29}$$

where $[x]$ denotes the integer part of x. The value of N_G is then set to $N - \sum_{g=1}^{G-1} N_g$. The formula (29), perhaps with a different value of γ, is also used in the H dimension. Assuming that the distributions are independent, $N_{gh} \approx N_g N_h / N$. In a final step, the cluster sizes are adjusted to ensure that they are all integers with $N = \sum_{g=1}^{G} N_g = \sum_{h=1}^{H} N_h = \sum_{g=1}^{G} \sum_{h=1}^{H} N_{gh}$.

The way in which the regressors are generated inevitably affects the finite-sample properties of every cluster-robust test statistic. The differences between asymptotic and finite-sample distributions arise mainly from the discrepancies between the disturbance vector $\boldsymbol{u}$ and the residual vector $\hat{\boldsymbol{u}} = \boldsymbol{M}_{\boldsymbol{X}} \boldsymbol{u}$, where $\boldsymbol{M}_{\boldsymbol{X}} = \mathbf{I}_N - \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top$. We use (28) to generate the regressor matrix $\boldsymbol{Z}$ in (25) as well as the disturbances. In most experiments, we set $\rho_g^x = \rho_h^x = 0.2$ for the regressors and $\rho_g = \rho_h = 0.1$ for the disturbances. We use these values because, in practice, regressors often display more intra-cluster correlation than residuals. For this base case, we deliberately avoid situations, to be discussed below, in which the amount of intra-cluster

Figure 1: Rejection frequencies as functions of how cluster sizes vary



Notes: There are $N = 10{,}000$ observations, with $G = 15$, $H = 12$, and $I = 180$. The regressors and disturbances are generated using (28), with $\rho_g^x = \rho_h^x = 0.2$ for the regressors and $\rho_g = \rho_h = 0.1$ for the disturbances. The regressand is generated using (25) with all coefficients equal to 0. The vertical axis shows rejection frequencies for t-tests at the .05 level based on the $t(\min\{G, H\} - 1)$ distribution. In Panel (a), γ is the same in both dimensions. In Panel (b), the value of γ is 0 for the H dimension and varies for the G dimension. The values of p and k are 10 and 36. There are 100,000 replications.

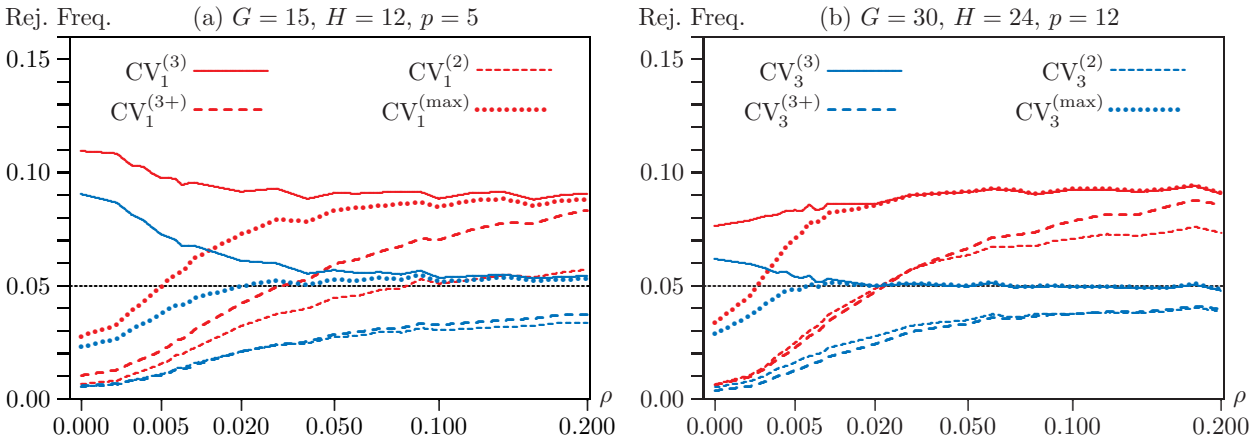
correlation is very small.

The first set of experiments focuses on cluster size variation, determined by the parameter γ in (29). Figure 1 shows rejection frequencies for eight different t-tests from 18 experiments with $G = 15$, $H = 12$, and $N = 10{,}000$. In Panel (a), the value of γ is varied simultaneously from 0.0 to 4.0 in both dimensions. In Panel (b), $\gamma = 0$ for the H dimension, and γ varies from 0.0 to 4.0 for the G dimension. In these experiments, the number of regressors is $p = 10$. All tests would have performed better if this had been a smaller number. The effects of varying p will be investigated below.

In Panel (a), when all clusters are (approximately) the same size (the leftmost point on the horizontal axes), t-tests based on the classic $\mathrm{CV}_1^{(3)}$ variance matrix estimator over-reject noticeably, as do those based on $\mathrm{CV}_1^{(\max)}$. Rejection frequencies are considerably lower for $\mathrm{CV}_1^{(3+)}$, and lower still for $\mathrm{CV}_1^{(2)}$. In contrast, t-tests based on the $\mathrm{CV}_3^{(3)}$ and $\mathrm{CV}_3^{(\max)}$ estimators are very close to nominal size, while those based on $\mathrm{CV}_3^{(3+)}$ and $\mathrm{CV}_3^{(2)}$ under-reject substantially. As the value of γ increases, all the CV_1 rejection frequencies rise sharply, while those for the CV_3 tests hardly change. Using the max-se procedure has almost no effect when cluster sizes vary little, but it modestly reduces rejection frequencies for CV_1 tests when they vary a lot.

In Panel (b), the overall patterns are similar. However, as predicted in Section 4, rejection frequencies increase less rapidly when γ just increases in the G dimension than when it increases in both dimensions. In both panels, as must be the case, t-tests based on two-term variance estimators always reject less often than t-tests based on three-term ones. This is a good thing for the CV_1 tests, but not for the CV_3 ones.

Figure 2: Rejection frequencies as functions of disturbance correlations



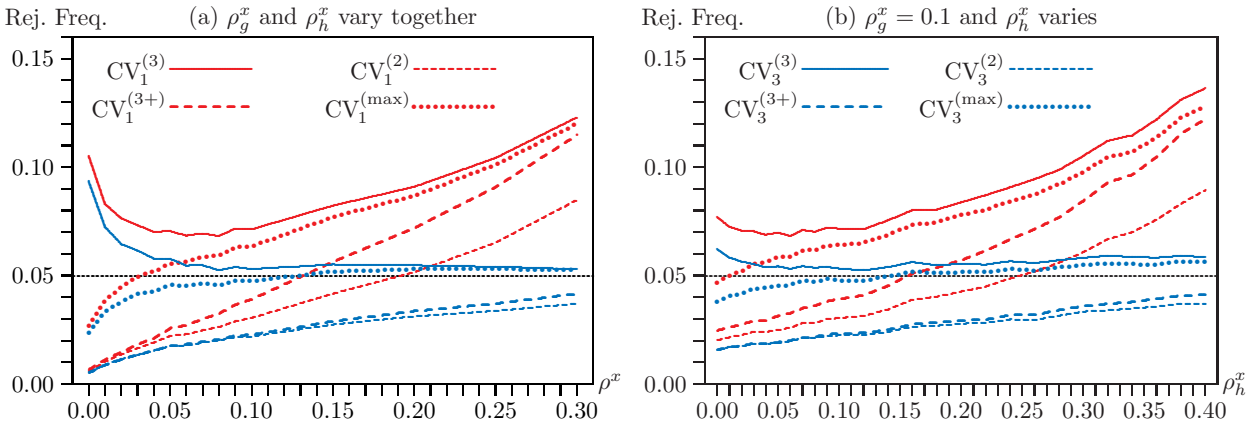
Notes: In Panel (a), $N = 10{,}000$, with $G = 15$, $H = 12$, $I = 180$, and $p = 5$. In Panel (b), $N = 40{,}000$, with $G = 30$, $H = 24$, $I = 720$, and $p = 15$. In both panels, $\gamma = 2$ in both dimensions. The p regressors are generated using (28) with $\rho_g^x = \rho_h^x = 0.2$. The disturbances are generated in the same way, but with $\rho_g = \rho_h = \rho$, which varies from 0.000 to 0.010 by 0.001, from 0.010 to 0.100 by 0.010, and from 0.120 to 0.200 by 0.020. The regressand is generated using (25) with all coefficients equal to 0. The vertical axis shows rejection frequencies for t-tests at the .05 level based on the $t(\min\{G, H\} - 1)$ distribution. The horizontal axis shows ρ, which is graphed on a square-root scale. There are 100,000 replications.

One possibly surprising feature of Figure 1 is how much the 3+ tests based on the eigendecomposition differ from the ordinary three-term tests. This happens because, with 36 coefficients to estimate (26 of them fixed effects), the three-term variance matrices are always singular. This will inevitably happen for models with cluster fixed effects or with large numbers of regressors for other reasons. For CV_1, the 3+ variant performs better than the usual three-term test, but for CV_3, it performs worse, under-rejecting about as much as the two-term test.

Except for quite small values of γ, the intersections in these experiments vary greatly in size. For example, when both values of γ equal 2, which is the base case for many of our subsequent experiments, the smallest intersection contains 6 observations, and the largest contains 253. The sizes of the intersections vary much more than those of the G clusters, which range from 223 to 1443, or the H clusters, which range from 282 to 1769. Although these numbers depend on the way in which we generate cluster sizes, it is inevitable that, when the cluster sizes vary in both dimensions, the sizes of the intersections vary more dramatically.

As MacKinnon, Nielsen, and Webb (2021) shows, test statistics based on the two-term variance estimator are asymptotically too small whenever the scores are asymptotically uncorrelated beyond the intersection level. This suggests that they are likely to under-reject severely when the amount of intra-cluster correlation is very small. In Figure 2, we vary both values of ρ for the disturbances from 0.000 to 0.200. For clarity, the horizontal axis uses a square root transformation. The numbers of clusters, observations, and regressors are larger in Panel (b) than in Panel (a); see the notes to the figure.

Figure 3: Rejection frequencies as functions of regressor correlations



Notes: In both panels, $N = 10{,}000$, with $G = 15$, $H = 12$, $I = 180$, $p = 5$, and $\gamma = 2$ in both dimensions. The disturbances are generated using (28) with $\rho_g = \rho_h = 0.1$. The regressors are also generated using (28), but the ρ^x parameters vary. In Panel (a), they both vary together between 0.00 and 0.30. In Panel (b), $\rho_g^x = 0.10$, and ρ_h^x varies between 0.00 and 0.40. The regressand is generated using (25) with all coefficients equal to 0. The vertical axis shows rejection frequencies for t-tests at the .05 level based on the $t(\min\{G, H\} - 1)$ distribution. There are 100,000 replications.

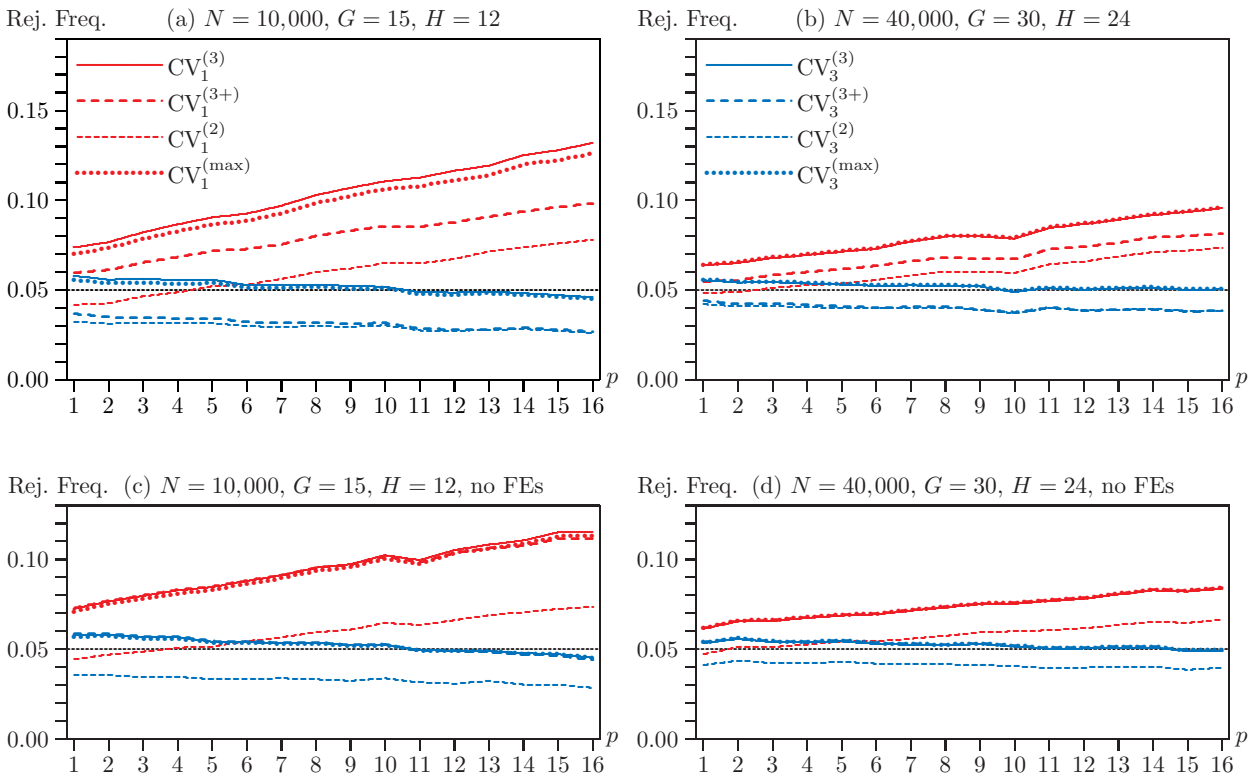
Several results stand out in Figure 2. For small values of ρ, the rejection frequencies of the three-term tests are much higher than those of the corresponding max-se tests. This is particularly true for the CV_1 tests. For the smallest values of ρ, the two-term tests under-reject to an extreme extent, as the theory in MacKinnon, Nielsen, and Webb (2021) predicts. Interestingly, so do the eigen-decomposition tests. In fact, in both panels, the two-term and 3+ tests perform very similarly. In both panels and for all values of ρ, the CV_3-based tests reject less than the corresponding CV_1-based tests. Except for the smallest values of ρ, the $CV_3^{(3)}$ and $CV_3^{(max)}$ tests are very similar in Panel (a) and identical in Panel (b), and they perform very well.

For the smallest values of ρ in these experiments, there were a number of replications for which the three-term variance of $\hat{\beta}_1$ was negative. This happened more often for $G = 15$ than for $G = 30$, and more often for CV_1 than for CV_3. Since we could not calculate the t-statistic for these replications, we classified them as rejections. In the most extreme case, when $\rho = 0.000$ for $G = 15$ ($G = 30$), this happened 2.60% (0.24%) of the time for $CV_1^{(3)}$ and 2.1% (0.16%) for $CV_3^{(3)}$. These numbers declined sharply as the value of ρ increased.

It is not only the correlations of the disturbances that matter. In Figure 3, we vary the correlations of the regressors, either in both dimensions, in Panel (a), or just in the H dimension, in Panel (b). Although they may seem small, the largest values of the ρ^x parameters here are not far short of the largest possible values; see the discussion below (28). The horizontal axis does not use a square-root scale as Figure 2 did, because the dependence on ρ^x for small values is not as extreme as the dependence on ρ in that figure.

It is clear from Figure 3 that the way in which the regressors are distributed can have

Figure 4: Rejection frequencies as functions of number of regressors



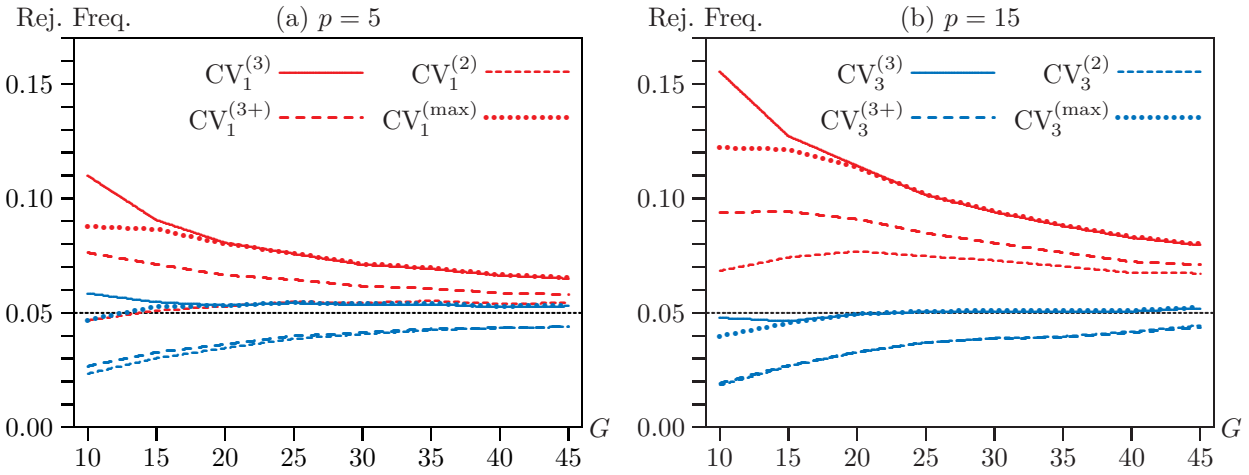
Notes: There are 10,000 observations in Panels (a) and (c) and 40,000 in Panels (b) and (d). The number of ordinary regressors (p) varies from 1 to 16. They are generated using (28) with $\gamma = 2$ and $\rho_g^x = \rho_h^x = 0.2$. The disturbances are generated in the same way, but with $\rho_g = \rho_h = 0.1$. In Panels (a) and (b), the regressand is generated using (25) with all coefficients equal to 0. In Panels (c) and (d), all the fixed effects are replaced by a constant term. The vertical axis shows rejection frequencies for t-tests at the .05 level based on the $t(\min\{G, H\} - 1)$ distribution. There are 100,000 replications.

substantial effects on rejection frequencies. Every test except $\mathrm{CV}_1^{(3)}$ and $\mathrm{CV}_3^{(3)}$ can either over-reject or under-reject, depending on the values of the two ρ^x parameters. The two three-term tests always over-reject, although only very slightly for $\mathrm{CV}_3^{(3)}$ in Panel (a) for $\rho^x > 0.05$. The most reliable tests are the ones based on $\mathrm{CV}_3^{(3)}$ and, especially, $\mathrm{CV}_3^{(\max)}$. This is particularly the case for larger values of the ρ^x parameters, where all the CV_1-based tests over-reject substantially. Panels (a) and (b) are quite similar when both ρ^x parameters, or just ρ_h^x, are large, but the two panels differ substantially when the intra-cluster correlations are small.

Figure 4 shows that the number of regressors, p, also matters. In Panels (a) and (c), $G = 15$, $H = 12$, and $N = 10{,}000$. In Panels (b) and (d), $G = 30$, $H = 24$, and $N = 40{,}000$. In all panels, p varies from 1 to 16. As it increases, the rejection rates for the CV_1 tests increase, but those for the CV_3 tests decrease slightly. In all panels, the max-se tests perform nearly the same as the three-term tests. Throughout Figure 4, the $\mathrm{CV}_3^{(3)}$ and $\mathrm{CV}_3^{(\max)}$ tests perform very well.

In the lower two panels, the 26 or 53 fixed effects are replaced by a constant term. Rejection

Figure 5: Rejection frequencies as functions of number of clusters



Notes: The value of G varies from 10 to 45 by 5, with $H = 4G/5$ and γ the same in both dimensions. There are $10{,}000GH/180$ observations, so that N varies from 1,111 to 90,000. The regressors are generated using (28) with $\rho_g^x = \rho_h^x = 0.2$. The disturbances are generated in the same way, but with $\rho_g = \rho_h = 0.1$. The regressand is generated using (25) with all coefficients equal to 0. The vertical axis shows rejection frequencies for t-tests at the .05 level based on the $t(\min\{G, H\} - 1)$ distribution. There are 100,000 replications.

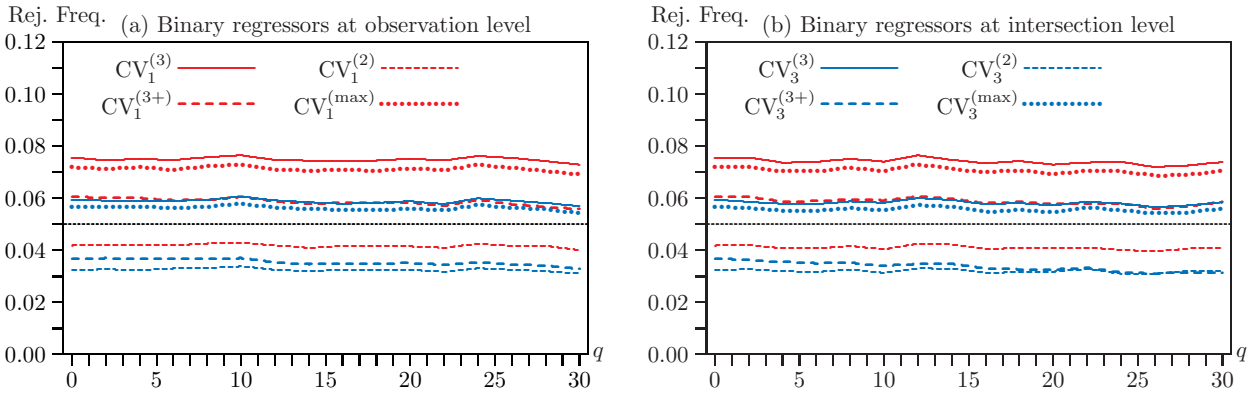
frequencies for the CV_1 tests still increase with p, but more slowly, while those for the CV_3 tests still decrease, at about the same slow rate. In these two panels, the 3+ tests are nearly identical to the ordinary three-term tests. These are the only experiments in which we omit the fixed effects. Their presence evidently has a large impact on the performance of the 3+ tests but a fairly modest effect on that of the other tests.

We would expect all the results to improve as the number of clusters increases, as happens in Figure 4. Therefore, in Figure 5, G varies from 10 to 45 by 5, H is always equal to $4G/5$, and N is proportional to GH, so that the sizes of the intersections are roughly constant. In Panel (a), $p = 5$. Here, $\mathrm{CV}_1^{(3)}$ and $\mathrm{CV}_1^{(\max)}$, and to a lesser extent $\mathrm{CV}_3^{(3+)}$, always over-reject, but they improve steadily as G (and H) increase. $\mathrm{CV}_1^{(\max)}$ over-rejects less severely than $\mathrm{CV}_1^{(3)}$ for the smallest values of G, but the former is almost indistinguishable from the latter for $G \geq 15$. In contrast, $\mathrm{CV}_3^{(3)}$ always works almost perfectly, with $\mathrm{CV}_3^{(\max)}$ yielding virtually identical results for $G \geq 15$. By what seems to be coincidence, $\mathrm{CV}_1^{(2)}$ also works well.

In Panel (b) of Figure 5, p is increased to 15. The CV_1-based tests now over-reject much more severely, but tests based on $\mathrm{CV}_3^{(3)}$ and $\mathrm{CV}_3^{(\max)}$ perform extremely well. In contrast, tests based on $\mathrm{CV}_3^{(2)}$ and $\mathrm{CV}_3^{(3+)}$ are almost identical and always under-reject. Clearly, omitting the intersection term or using the eigen-decomposition is helpful for CV_1, because the three-term tests over-reject, but harmful for CV_3, because the three-term tests are approximately sized correctly.

Figure 4 suggests that rejection frequencies for all the CV_1 tests increase fairly rapidly with p, the number of regressors that are not fixed effects, while those for all the CV_3 tests decrease quite slowly. We conjecture that this is happening because all the regressors are correlated

Figure 6: Rejection frequencies as functions of number of extra binary regressors



Notes: In both panels, $N = 10{,}000$, $G = 15$, $H = 12$, and $\gamma = 2$ in both dimensions. There is one continuous regressor, like the ones in Figures 1–5. In addition, there are q binary regressors, which equal 1 with probability 0.5. These vary independently at the observation level in Panel (a) and at the intersection level in Panel (b). The vertical axis shows rejection frequencies for t-tests at the .05 level based on the $t(\min\{G, H\} - 1)$ distribution. There are 100,000 replications.

within clusters in one or both dimensions. Thus, as the number of regressors increases, more and more of the intra-cluster correlation in the disturbances is explained by the regressors, so that less of it remains in the residuals.

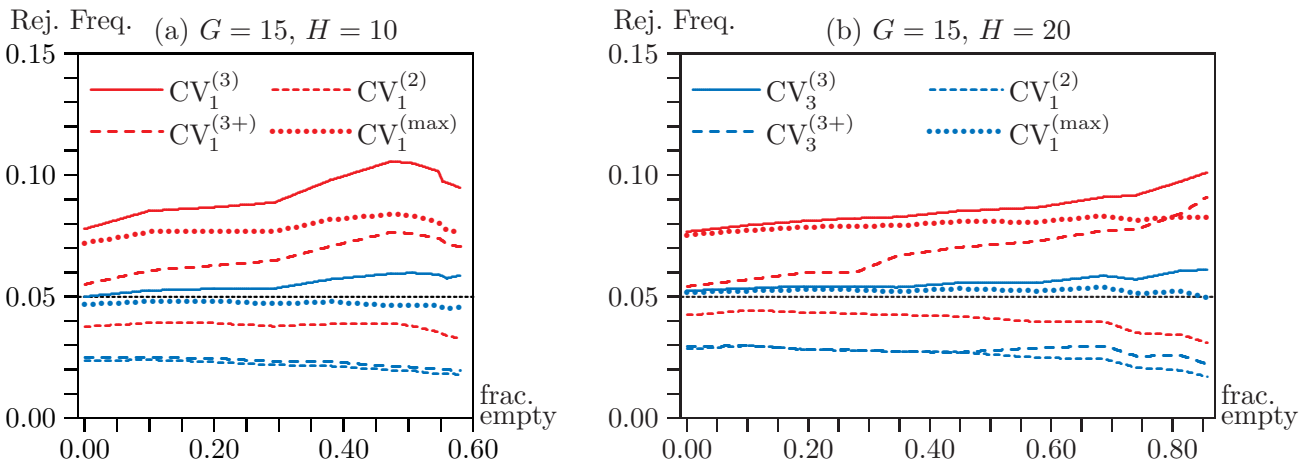
In order to investigate this conjecture, we modify the way in which we generate the regressors. The first one (the test regressor) is generated as before, but then we generate an additional q binary regressors which equal 0 or 1 with probability 0.5. In one set of simulations, they are completely independent across observations. In a second set, they are generated at the intersection level, identical within each intersection and independent across intersections.

Figure 6 shows rejection frequencies as a function of q, which varies from 0 to 30. In both panels, the test based on $\text{CV}_3^{(\text{max})}$ performs best, over-rejecting slightly for all values of q. The tests based on $\text{CV}_3^{(3)}$, $\text{CV}_1^{(3+)}$, and $\text{CV}_1^{(2)}$ perform nearly as well, with the former two over-rejecting slightly and the latter under-rejecting slightly. The value of q has very little effect on most of the tests.

The differences between Figure 4 and Figure 6 are striking. In the former, all the regressors are correlated within both the G and H clusters. We saw there that adding more regressors with this property can substantially increase rejection frequencies for CV_1 tests and slightly decrease them for CV_3 tests. In contrast, adding more regressors that are uncorrelated across observations or across intersections has almost no effect on rejection frequencies. Empirical applications of two-way clustering often involve many controls. Whether or not these controls exhibit substantial correlation in either dimension can evidently be important.

Up to this point, the data for all of our experiments have been generated in such a way that $I = GH$. In other words, there have been no datasets with empty intersections. But

Figure 7: Rejection frequencies as functions of fraction of empty intersections



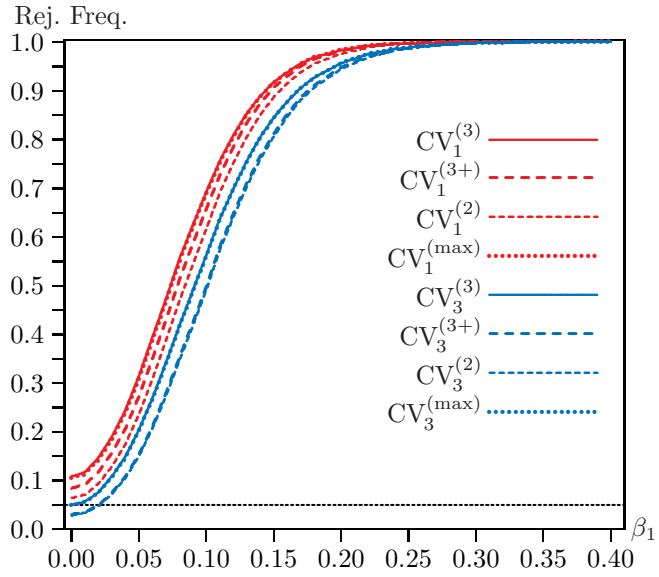
Notes: In Panel (a), $N = 6{,}000$. In Panel (b), $N = 12{,}000$. The disturbances are generated using (28) with $\rho_g = \rho_h = 0.1$. There are 5 continuous regressors, which are generated using (28) with $\rho_g^x = \rho_h^x = 0.2$, along with 5 binary regressors, which vary at the intersection level and equal 1 with probability 0.25. The fraction of empty intersections varies on the horizontal axis. The vertical axis shows rejection frequencies for t-tests at the .05 level based on the $t(\min\{G, H\} - 1)$ distribution. There are 100,000 replications.

empirical examples with two-way clustering often involve empty intersections. In the next set of experiments, we therefore change the DGP so that intersections can be empty. The details are somewhat complicated and not of much interest. There is a parameter which, as it increases, causes the fraction of empty intersections to increase from zero to an upper limit that depends on G, H, and how the clusters vary in size. We perform two sets of experiments. In both of them, the 15 clusters in the G dimension are generated from (29) with $\gamma = 2$. In the first set $H = 10$, and in the second set $H = 20$. The maximum observed number of empty intersections is 87 (out of 150) in the first set and 257 (out of 300) in the second set.

Figure 7 shows rejection frequencies as functions of the fraction of empty intersections. This fraction evidently matters, especially for the CV_1-based tests, although not dramatically so in these experiments. As usual, t-tests based on $\mathrm{CV}_3^{(\max)}$ always perform best, and in fact they perform extremely well. Some of the other tests perform quite poorly. As a rule, tests that over-reject or under-reject when there are no empty intersections do the same thing to a greater extent when there are many empty intersections.

Since some of the tests tend to over-reject and others tend to under-reject under the null hypothesis, it is inevitable that the former will appear to have more power than the latter. Figure 8 shows power functions for all eight tests. The functions never cross, so there is nothing surprising here. For every alternative, the ranking of the tests by power is identical to their ranking by rejection frequencies under the null hypothesis. Thus the fact that all of the CV_1 tests appear to be more powerful than any of the CV_3 tests simply reflects that the former are over-sized under the null. In this experiment, the power functions for $\mathrm{CV}_1^{(\max)}$ and $\mathrm{CV}_3^{(\max)}$ are

Figure 8: Power functions for eight tests



Notes: There are 10,000 observations, with $G = 15$, $H = 12$, $p = 10$, and $\gamma = 2$ in both dimensions. The regressors are generated using (28) with $\rho_g^x = \rho_h^x = 0.2$. The disturbances are generated in the same way, but with $\rho_g = \rho_h = 0.1$. The regressand is generated using (25) with $\beta_1 \geq 0$ and all other coefficients equal to 0. The vertical axis shows rejection frequencies for t-tests at the .05 level based on the $t(\min\{G, H\} - 1)$ distribution. There are 100,000 replications.

indistinguishable from those for $CV_1^{(3)}$ and $CV_3^{(3)}$, respectively. All the tests evidently reject with probability one when β_1 is sufficiently large.

6 Empirical Examples

In this section, we study two empirical examples that apply very different types of two-way clustering with clusters that behave in different ways across the two clustering dimensions.

6.1 The Tsetse Fly in African Development

In a fascinating paper, Alsan (2015) studies the effects of the tsetse fly on African development. The key explanatory variable is the "tsetse suitability index," or TSI, which measures the extent to which climate (temperature and humidity) is suitable for the tsetse fly to thrive. There are seven dependent variables, which measure various aspects of economic and political development. Each of these is regressed on the TSI, whose coefficient is denoted β, and on eleven other variables in the columns labeled "(4)" in Table 1 and "(8)" in Table 3 of Alsan (2015). The former uses one-way clustering by "cultural province" and the latter uses two-way clustering by cultural province and country. There are 44 countries and either 43 or 44 cultural provinces, depending on the regressand. Since the total number of observations varies between 315 and 485, most clusters are quite small, and there are many empty intersections. The number of non-empty intersections varies between 112 and 142.

Table 1: Empirical results for tsetse fly example

Panel A: Conventional CRVEs		P values						
Dependent variable	$\hat{\beta}$	HC_1	CV_1-I	CV_1-G	CV_1-H	$\text{CV}_1^{(2)}$	$\text{CV}_1^{(3)}$	$\text{CV}_1^{(\max)}$
Large animals	−0.2310	0.0000	0.0000	0.0000	0.0000	0.0002	0.0000	0.0000
Intensive agriculture	−0.0905	0.0080	0.0087	0.0020	0.0045	0.0321	0.0003	0.0045
Plow use	−0.0572	0.0096	0.0171	0.0149	0.0791	0.1496	0.0715	0.0791
Female participation	0.2057	0.0001	0.0001	0.0014	0.0001	0.0099	0.0009	0.0014
Log pop. density	−0.7446	0.0000	0.0002	0.0023	0.0005	0.0182	0.0028	0.0028
Indigenous slavery	0.1013	0.0060	0.0137	0.0160	0.0162	0.0834	0.0157	0.0162
Centralization	−0.0746	0.0727	0.0460	0.0395	0.0471	0.1482	0.0365	0.0471

Panel B: Jackknife CRVEs		P values						
Dependent variable	$\hat{\beta}$	HC_3	CV_3-I	CV_3-G	CV_3-H	$\text{CV}_3^{(2)}$	$\text{CV}_3^{(3)}$	$\text{CV}_3^{(\max)}$
Large animals	−0.2310	0.0000	0.0000	0.0000	0.0000	0.0007	0.0000	0.0000
Intensive agriculture	−0.0905	0.0092	0.0130	0.0044	0.0123	0.0549	0.0027	0.0123
Plow use	−0.0572	0.0124	0.0220	0.0207	0.1422	0.2112	0.1365	0.1422
Female participation	0.2057	0.0001	0.0002	0.0036	0.0008	0.0239	0.0051	0.0051
Log pop. density	−0.7446	0.0000	0.0005	0.0060	0.0021	0.0357	0.0090	0.0090
Indigenous slavery	0.1013	0.0077	0.0247	0.0353	0.0496	0.1462	0.0575	0.0575
Centralization	−0.0746	0.0798	0.0564	0.0545	0.0817	0.1925	0.0751	0.0817

Panel C: Cluster diagnostics	Coefficients of variation							
	partial leverage			$\hat{\beta}^{(g)}$			G^*	
Dependent variable	culture	country	intersect	culture	country	intersect	culture	country
Large animals	1.2877	1.1571	1.7094	0.0323	0.0271	0.0173	23.18	16.25
Intensive agriculture	1.2875	1.1572	1.7095	0.0513	0.0591	0.0336	23.20	16.27
Plow use	1.2877	1.1571	1.7094	0.0643	0.1035	0.0365	23.18	16.25
Female participation	1.6073	1.4962	2.0302	0.0507	0.0444	0.0248	20.84	13.71
Log pop. density	1.3241	1.2277	1.8453	0.0540	0.0481	0.0258	20.88	15.51
Indigenous slavery	1.3307	1.2431	1.7912	0.0710	0.0774	0.0376	22.71	15.84
Centralization	1.3189	1.1751	1.7403	0.0780	0.0861	0.0442	22.76	16.20

Notes: Estimates correspond to Table 1 (4) and Table 3 (8) from Alsan (2015). Panels A and B show coefficient estimates and P values. CV-G is clustered by cultural province, CV-H is clustered by country, and CV-I is clustered by intersection. P values for HC standard errors are based on the $t(N-13)$ distribution. P values for CV-G, CV-H, and CV-I are based on the $t(G-1)$, $t(H-1)$, or $t(I-1)$ distributions, respectively. P values for two-way clustering are based on the $t(\min\{G,H\})$ distribution. In all cases, the P values based on $\text{CV}_3^{(3+)}$ are identical to those based on $\text{CV}_3^{(3)}$ to the number of digits reported. Panel C shows coefficients of variation for partial leverage and omit-one-cluster coefficients by both clustering dimensions and their intersection, as well as the effective number of clusters, $G^* = G^*(0)$, computed by `twowayjack`.

In Table 1, we report P values based on sixteen different standard error estimates, eight using conventional standard errors (Panel A) and eight using jackknife ones (Panel B). Because the ordinary three-term and eigen-decomposition three-term standard errors are identical in all cases (to the number of digits reported), we only report the former.

Even though the clusters are quite small (the largest is 63, which is for clustering by country when the dependent variable is the log of population density), the way in which we cluster often makes a substantial difference. Not clustering at all sometimes leads to extremely small P values, as does one-way clustering by intersection. Clustering in two dimensions often, but not always, leads to larger P values than clustering in just one dimension. For two-way clustering, the cluster jackknife P values are never smaller than the conventional ones, and they are mostly considerably larger.

Panel C of Table 1 presents a number of the summary statistics calculated by `twowayjack` for this example. Specifically, it presents coefficients of variation for the partial leverages of the TSI variable and for the $\hat{\beta}^{(g)}$ for clustering by cultural province, country, and intersection. It also displays the effective number of clusters $G^* = G^*(0)$ for the two primary dimensions (Carter et al. 2017; MacKinnon et al. 2023c). These diagnostics can help to explain why some of the P values in Panels A and B differ by more than others. The notable P value differences between $\text{CV}_3^{(\max)}$ and $\text{CV}_1^{(\max)}$ occur for 'plow use,' 'indigenous slavery,' and 'centralization.' For these three variables, we see the largest coefficients of variation for the omit-one-country estimates, and, to a slightly lesser extent, for the omit-one-culture ones.

The results of Section 5 suggest that $\text{CV}_3^{(\max)}$ yields the most reliable P values. The $\text{CV}_3^{(\max)}$ P value for TSI is less than 0.05 for four of the seven dependent variables. In contrast, the CV_1 P values for one-way clustering by cultural province used in Alsan (2015) are less than 0.05 for all seven variables in Table 1 (4), and the ones for two-way $\text{CV}_1^{(3)}$ clustering are less than 0.05 for six of them in Table 3 (8). Thus, although there is still a good deal of evidence that the TSI matters for a variety of outcomes, the evidence is not quite as strong as it originally seemed to be.

6.2 Minimum Wages in Canada

Our second example examines the relationship between minimum wages in Canada and the log of hourly earnings. We focus attention on men between 18 and 24 years of age who immigrated to Canada less than ten years ago. Our sample contains 28,599 observations for the years 2008 to 2019. Except for a few federally-regulated industries, minimum wages in Canada are set at the provincial level. They tend to change infrequently, and they never go down. In fact, although our sample contains observations for 1440 province-month pairs, the minimum wage variable takes on only 63 unique values.

The equation we estimate is

$$\text{logearn}_{ipmt} = \alpha + \beta\,\text{logmw}_{pmt} + \gamma\,\text{bigcity}_{ipmt} + \delta\,\text{age}_{ipmt} + \text{year}_t + \text{month}_m + \text{prov}_p + \epsilon_{ipmt}, \quad (30)$$

Table 2: Empirical results for minimum wage example

Panel A: Conventional CRVEs	HC_1	CV_1-I	CV_1-G	CV_1-H	$CV_1^{(2)}$	$CV_1^{(3)}$	$CV_1^{(3+)}$	$CV_1^{(max)}$
P values	0.0000	0.0000	0.0041	0.0034	0.0261	0.0140	0.0141	0.0140
Placebo rej. freq.	0.8947	0.6301	0.5958	0.3018	0.1454	0.2431	0.2431	0.2319
Panel B: Jackknife CRVEs	HC_3	CV_3-I	CV_3-G	CV_3-H	$CV_3^{(2)}$	$CV_3^{(3)}$	$CV_3^{(3+)}$	$CV_3^{(max)}$
P values	0.0000	0.0001	0.0125	0.0565	0.1116	0.0808	0.0810	0.0808
Placebo rej. freq.	0.8947	0.5725	0.5432	0.0896	0.0254	0.0649	0.0649	0.0572

Panel C: Cluster diagnostics	Coefficients of variation							
	partial leverage			$\hat{\beta}^{(g)}$			G^*	
	year	province	intersect	year	province	intersect	year	province
Log earnings	0.0607	1.1909	1.1794	0.1061	0.1577	0.0232	6.51	4.49

Notes: There are 28,599 observations, 12 year (G) clusters, 10 province (H) clusters, and 120 intersection (I) clusters. The coefficient estimate for the log minimum wage is $\hat{\beta} = 0.2934$. The first row in each of Panels A and B reports P values using HC standard errors based on the $t(28{,}564)$ distribution, P values for one-way clustering based on the $t(11)$, $t(9)$, or $t(119)$ distributions, as appropriate, and P values for two-way clustering based on the $t(9)$ distribution. The second row reports rejection frequencies for 100,000 placebo regressions. Panel C reports coefficients of variation for partial leverage and omit-one-cluster coefficients by both clustering dimensions and their intersection, as well as the effective number of clusters, $G^* = G^*(0)$, computed by `twowayjack`.

where logearn_{ipmt} is the log of hourly earnings for individual i in province p in month m of year t, logmw_{pmt} is the log of the minimum wage, bigcity_{ipmt} is a dummy for being in one of nine large cities, age_{ipmt} is a dummy for being 22 to 24, and the remaining regressors are year fixed effects, month-within-year fixed effects, and province fixed effects. The total number of regressors, including the constant term, is 35.

This example is one for which reliable cluster-robust inference is likely to be difficult. We cluster by year and province, but there are only 12 years and 10 provinces. The year clusters are reasonably homogeneous in size; they vary from 2051 to 2723 observations. But the province clusters are very heterogeneous; they vary from 163 (P.E.I.) to 6554 (Ontario). Although there are no empty intersections, the smallest contains just 3 observations, and the largest contains 710.

Table 2 contains three panels. Panel C presents some cluster diagnostics, calculated using `twowayjack`. The coefficients of variation are quite revealing. For partial leverage, there is considerable variation across provinces and intersections, but very little across years. For the $\hat{\beta}^{(g)}$, there is modest variation when leaving out a province or a year, but very little when leaving out an intersection cluster.

These features of the sample suggest that many methods, perhaps all methods, will not yield reliable inferences. In order to investigate this conjecture, we employ placebo-regression

simulations as advocated by MacKinnon, Nielsen, and Webb (2023a, Section 3.5). These are similar in spirit to the "placebo laws" simulations of Bertrand, Duflo, and Mullainathan (2004). For each of 100,000 simulations and each province, we generate a sequence of values of a placebo regressor that resembles the actual minimum wage sequences: The value tends to stay constant for a while and then rise by a random amount from time to time in a fashion that is correlated across provinces. This placebo regressor is then added to regression (30), and we calculate sixteen P values for its coefficient based on all sixteen standard errors used for the actual regression. If regression (30) is correctly specified and any particular way of obtaining P values is valid for our sample, then the fraction of the time that the placebo-regression P value is less than 0.05 should be very close to 0.05, subject to experimental error.

Panels A and B of Table 2 show both the actual P values and rejection frequencies for the placebo regressions for all sixteen methods. The conventional P values in Panel A imply that the minimum wage is significant at the 0.01 level for all the one-way clustering methods and at the 0.05 level for all the two-way methods. However, the placebo-regression rejection frequencies vary from 15% to 89%, suggesting that none of the conventional P values should be believed.

In contrast, the jackknife P values in Panel B are greater than 0.05 for one-way clustering by province and for the two-way clustering methods. The placebo-regression rejection frequencies for the one-way methods vary between 9% and 89%, suggesting that they should not be trusted. For the two-term two-way estimator $\text{CV}_3^{(2)}$, the placebo rejection frequency is just 2.5%, which is in line with existing theory (Section 2) and some of our simulations (e.g. Figure 2). For the three-term cluster jackknife estimators, the placebo rejection frequencies are between 5.7% and 6.5%, which is remarkably good in view of the small numbers of clusters and the cluster diagnostics. For these methods, the P value is 0.081. Thus the most reliable methods all suggest that the evidence for the effect of the minimum wage in this sample is fairly weak.

7 Conclusions

It is common to assume that the disturbances in linear regression models are clustered in two dimensions. Unless the regressor(s) of interest are uncorrelated in one or both dimensions, it is therefore necessary to employ a cluster-robust variance estimator that allows for two-way clustering. Unfortunately the most widely-used cluster-robust variance matrix estimator (CRVE), $\text{CV}_1^{(3)}$, due to Cameron et al. (2011), is not guaranteed to be positive definite. Inferences based on it are known to be seriously unreliable in finite samples (MacKinnon et al. 2021).

In Section 2, we discuss several ways to avoid, or at least ameliorate, the problem of undefined standard errors when a CRVE is not positive definite. Most importantly, we propose a new and simple solution to this problem. For tests of a single restriction, it just involves using whichever of three standard errors is the largest. Two of these are based on one-way clustering in each of the two dimensions, and the third is a three-term two-way standard error. Asymptotically, the

latter should always be the largest of the three when there really is two-way clustering, but it may not be the largest (and may indeed not be defined) in finite samples. In many cases, our so-called max-se procedure yields results identical to those from the corresponding three-term two-way CRVE, but it can yield substantially lower (and more accurate) rejection frequencies in some cases.

The second, and in our view more important, contribution of the paper is to propose and study two-way cluster-jackknife CRVEs. Recent work on the cluster-jackknife, or CV_3, CRVE for one-way clustering (Hansen 2023; MacKinnon et al. 2023b, c) suggests that it can perform much better in finite samples than the usual CV_1 CRVE. It therefore seems attractive to extend it to the two-way case. This is remarkably simple. We just need to perform three sets of cluster-jackknife calculations, one for each of the two dimensions, and then a third one for their intersections. In many cases, this is straightforward, although cluster fixed effects do raise some computational issues (Section 3), and the calculations can be costly when the number of intersections is large, especially when there are cluster fixed effects. We provide a `Stata` package called `twowayjack` that implements our methods and also calculates some cluster diagnostics (MacKinnon et al. 2023c); see Appendix A.

In Section 5, we study rejection frequencies for t-tests based on eight different cluster-robust standard errors. Four of them are of the usual CV_1 type, and the other four are of the CV_3 type. In most cases, tests based on the CV_3 max-se standard error yield the most reliable inferences. Even when they do not, they only perform slightly worse than whatever procedure(s) perform better, and they are usually much more reliable than all of the CV_1-based tests.

Because most of our simulations involve two-way cluster fixed effects, three-term variance matrices based on either CV_1 or CV_3 tend to be singular, so the versions that use an eigen-decomposition (Section 2) can differ greatly from the versions that do not. This always reduces rejection frequencies, which is a good thing for CV_1 tests but usually a bad thing for CV_3 tests. Tests based on two-term variance matrices usually reject even less frequently than tests based on three-term variance matrices with the eigen-decomposition. Thus we do not recommend using tests based on either $CV_3^{(2)}$ or $CV_3^{(3+)}$.

Our simulations show that precisely how the data are generated can have a large effect on finite-sample performance. All the tests are most likely to perform poorly when the number of clusters in either dimension is small, cluster sizes vary greatly, there are many empty intersections, the number of regressors that are clustered in one or both dimensions is large, or either the disturbances or the regressor(s) of interest are only weakly correlated in both dimensions. In many of these cases, alternative test statistics tend to perform quite differently.

In practice, it can often be illuminating to employ placebo regression simulations, as in Section 6.2. These will show how well alternative tests perform for the particular model and dataset under study. It is probably safe to rely on $CV_3^{(\max)}$-based tests if they perform well in

these simulations, or perhaps on some other tests if they perform better.

Appendix: The twowayjack Package

We have written a package `twowayjack` for `Stata` that implements the variance estimators discussed in this paper. The package relies on our earlier package `summclust` (MacKinnon et al. 2023c), and it calculates both $\text{CV}_3^{(\max)}$ and $\text{CV}_1^{(\max)}$ for the coefficient of interest, as well as confidence intervals and P values.

The package also provides coefficients of variation for several diagnostic measures as described in MacKinnon et al. (2023c). For the two primary clustering dimensions and their intersections, it calculates the coefficients of variation for the cluster sizes, leverage, partial leverage, and omit-one-cluster estimates, $\hat{\beta}^{(g)}$. In addition, it displays the number of clusters G and the effective number of clusters $G^* = G^*(0)$ from Carter et al. (2017). The latest version may be obtained from https://github.com/mattdwebb/twowayjack. The data and programs used in the paper may be found at http://qed.econ.queensu.ca/pub/faculty/mackinnon/twowayjack/.

A.1 Syntax

The syntax for `twowayjack` is

```
twowayjack varlist, cluster(varlist) [fevar(varlist) sample(string)]
```

Here `varlist` contains a list of variables. The first one is the dependent variable, the second is the regressor for which standard errors and P values are to be calculated, and the remaining ones are all the other continuous and binary regressors. Categorical variables to be treated as fixed effects should be listed using the `fevar` option.

`cluster(varlist)` is mandatory, where `varlist` contains the two variables by which observations are clustered. The program returns an error if exactly two variables are not specified.

`fevar(varlist)`. Categorical variables to be included in the model as fixed effects should be listed here. They are handled equivalently to `i.varlist` in a regression model. Since this option uses a generalized inverse, CV_3 can be calculated even when some of the omit-one-cluster subsamples are singular. This always happens with cluster-level fixed effects.

`sample(string)` limits the sample. Use the text you would enter after an "if" in a regression command. For instance, `sample(female==1)` is equivalent to "`if female==1`."

A.2 Illustration

In this section, we highlight how the package can be called using an example from the `webuse` dataset `nlswork`. The outcome of interest is hours worked. The independent variable of interest is `vismin`, which is set to 0 if a person is white, and 1 otherwise. In part because hours vary with age, and across industry, we cluster by both age and industry.

The first commands load and clean the dataset.

```
webuse nlswork, clear
keep if inrange(age,25,35)
gen vismin  = inrange(race,2,3)
```

For comparison purposes, the native `Stata` regression results with one-way clustering are obtained from the command:

```
reg hours vismin south i.age i.birth_yr i.year i.ind , cluster(ind)
```

This yields the results:

```
Linear regression                               Number of obs     =     13,754
                                                F(10, 11)         =          .
                                                Prob > F          =          .
                                                R-squared         =     0.0659
                                                Root MSE          =     9.6979
                                (Std. err. adjusted for 12 clusters in ind_code)
------------------------------------------------------------------------------
             |               Robust
       hours | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
      vismin |   1.054672   .4202197     2.51   0.029     .1297746    1.979569
[additional output truncated]
```

We can instead estimate the same model with two-way clustering using `twowayjack`:

```
twowayjack hours vismin south , fevar(age birth_yr year ind) cluster(age ind)
```

The output is:

```
TWOWAYJACK
Reference: James G. MacKinnon, Morten Ø. Nielsen, and Matthew D. Webb
           Jackknife inference with two-way clustering

Two-way cluster jackknife variance estimation.
Cluster summary statistics for vismin when clustered by age and ind_code.

Regression Output
    s.e. |      Coeff   Sd. Err.   t-stat  P value    CI-lower    CI-upper
---------+--------------------------------------------------------------
  CV1max |   1.054672   0.420220   2.5098   0.0309    0.223377    1.885967
  CV3max |   1.054672   0.521628   2.0219   0.0708   -0.107587    2.216931
------------------------------------------------------------------------------
```

```
Coefficients of Variation, G, and G*
 dimension |       Ng   Leverage  Partial L.   beta no g        G      Gstar
-----------+-------------------------------------------------------------
       age |   0.0987     0.1813      0.0927      0.0431       11      10.90
  ind_code |   1.1815     0.8823      1.1849      0.1565       12       5.21
 intersect |   1.1507     0.8925      1.1557      0.0173      132      56.26
-------------------------------------------------------------------------
```

References

Alsan, M. (2015). The effect of the tsetse fly on African development. *American Economic Review 105*, 382–410.

Bell, R. M. and D. F. McCaffrey (2002). Bias reduction in standard errors for linear regression with multi-stage samples. *Survey Methodology 28*, 169–181.

Bertrand, M., E. Duflo, and S. Mullainathan (2004). How much should we trust differences-in-differences estimates? *Quarterly Journal of Economics 119*, 249–275.

Boot, T., G. Niccodemi, and T. Wansbeek (2023). Unbiased estimation of the OLS covariance matrix when the errors are clustered. *Empirical Economics 64*, 2511–2533.

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

Cameron, A. C., J. B. Gelbach, and D. L. Miller (2011). Robust inference with multiway clustering. *Journal of Business & Economic Statistics 29*, 238–249.

Carter, A. V., K. T. Schnepel, and D. G. Steigerwald (2017). Asymptotic behavior of a t test robust to cluster heterogeneity. *Review of Economics and Statistics 99*, 698–709.

Chiang, H. D., B. E. Hansen, and Y. Sasaki (2023). Standard errors for two-way clustering with serially correlated time effects. ArXiv e-prints 2201.11304v4.

Chiang, H. D., K. Kato, Y. Ma, and Y. Sasaki (2022). Multiway cluster robust double/debiased machine learning. *Journal of Business & Economic Statistics 40*, 1046–1056.

Chiang, H. D., K. Kato, and Y. Sasaki (2023). Inference for high-dimensional exchangeable arrays. *Journal of the American Statistical Association 118*, 1595–1605.

Davezies, L., X. D'Haultfœuille, and Y. Guyonvarch (2021). Empirical process results for exchangeable arrays. *Annals of Statistics 49*, 845–862.

Djogbenou, A. A., J. G. MacKinnon, and M. Ø. Nielsen (2019). Asymptotic theory and wild bootstrap inference with clustered errors. *Journal of Econometrics 212*, 393–412.

Efron, B. (1981). Nonparametric estimates of standard error: The jackknife, the bootstrap and other methods. *Biometrika 68*, 589–599.

Efron, B. and C. Stein (1981). The jackknife estimate of variance. *Annals of Statistics 9*, 586–596.

Hansen, B. E. (2023). Jackknife standard errors for clustered regression. Working paper, University of Wisconsin.

Hounyo, U. and J. Lin (2024). Reliable wild bootstrap inference with multiway clustering. SSRN Working Paper 4701693.

Imbens, G. W. and M. Kolesár (2016). Robust standard errors in small samples: Some practical advice. *Review of Economics and Statistics 98*, 701–712.

MacKinnon, J. G., M. Ø. Nielsen, and M. D. Webb (2021). Wild bootstrap and asymptotic inference with multiway clustering. *Journal of Business & Economic Statistics 39*, 505–519.

MacKinnon, J. G., M. Ø. Nielsen, and M. D. Webb (2023a). Cluster-robust inference: A guide to empirical practice. *Journal of Econometrics 232*, 272–299.

MacKinnon, J. G., M. Ø. Nielsen, and M. D. Webb (2023b). Fast jackknife and bootstrap methods for cluster-robust inference. *Journal of Applied Econometrics 38*, 671–694.

MacKinnon, J. G., M. Ø. Nielsen, and M. D. Webb (2023c). Leverage, influence, and the jackknife in clustered regression models: Reliable inference using summclust. *Stata Journal 23*, 942–982.

MacKinnon, J. G., M. Ø. Nielsen, and M. D. Webb (2023d). summclust: Stata module to compute cluster level measures of leverage, influence, and a cluster jackknife variance estimator. https://ideas.repec.org/c/boc/bocode/s459072.html.

MacKinnon, J. G., M. Ø. Nielsen, and M. D. Webb (2023e). Testing for the appropriate level of clustering in linear regression models. *Journal of Econometrics 235*, 2027–2056.

MacKinnon, J. G. and M. D. Webb (2017). Wild bootstrap inference for wildly different cluster sizes. *Journal of Applied Econometrics 32*, 233–254.

MacKinnon, J. G. and H. White (1985). Some heteroskedasticity consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics 29*, 305–325.

Menzel, K. (2021). Bootstrap with cluster-dependence in two or more dimensions. *Econometrica 89*, 2143–2188.

Miglioretti, D. L. and P. J. Heagerty (2006). Marginal modeling of nonnested multilevel data using standard software. *American Journal of Epidemiology 165*, 453–463.

Thompson, S. B. (2011). Simple formulas for standard errors that cluster by both firm and time. *Journal of Financial Economics 99*, 1–10.

Tukey, J. W. (1958). Bias and confidence in not quite large samples. *Annals of Mathematical Statistics 29*, 614.

Yap, L. (2024). Asymptotic theory for two-way clustering. ArXiv e-prints 2301.03805v2.