Standard Errors for Difference-in-Difference Regression

Bruce E. Hansen[*]
University of Wisconsin[†]

This version: November 2024

Abstract

This paper makes a case for the use of jackknife methods for standard error, p-value, and confidence interval construction for difference-in-difference (DiD) regression. We review cluster-robust, bootstrap, and jackknife standard error methods, and show that standard methods can substantially underperform in conventional settings. In contrast, our proposed jackknife inference methods work well in broad contexts. We illustrate the relevance by replicating several influential DiD applications, and showing how inferential results can change if jackknife standard error and inference methods are used.

[*] Research support from the Phipps Chair is gratefully acknowledged. Thanks to the Co-Editor and Associate Editor for helpful comments, to Matthew Webb for his help and suggestions with the empirical application, and to James MacKinnon for suggesting algebraic simplifications for the computational formulae.

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

1 Introduction

Difference-in-difference (DiD) regression is one of the most common empirical tools in current applied economic practice. The vast majority of applications report standard errors clustered at the level of treatment. These standard errors, however, are biased towards zero, and the magnitude of bias can be arbitrarily severe. As a consequence, conventionally reported standard errors, p-values, and confidence intervals are unreliable.

In this paper, we argue that two simple changes can greatly alleviate these problems. First, standard error calculation should be made by the jackknife. If the jackknife is implemented as proposed, the variance estimator is guaranteed to be never downward biased. Jackknife variance estimation is simple to implement, and is computationally efficient when there are a moderate number of clusters, which is typical in applications.

The second change we recommend is the use of adjusted student t p-values and confidence intervals based on a finite-sample distributional approximation. These p-values and confidence intervals are typically more conservative than conventional methods, and provide more accurate inferences in simulations. The adjusted student t approximation is computationally simple to implement, allowing for routine default use.

To illustrate the methods, we investigate a set of results from four influential DiD applications: Card and Krueger (1994), Bailey (2010), MacKinnon and Webb (2020), and Rao (2019). Using the original data from these papers, we calculate standard errors, p-values, and confidence intervals both by conventional cluster-robust and our proposed jackknife methods. We find that some results change considerably, while other results are unaffected. These examples illustrate the magnitude of the changes due to our proposed changes in relevant applications.

Heteroskedasticity-robust covariance matrix estimation was introduced to econometrics by White (1980), building on the work of Eicker (1963) and Huber (1967). This family of estimators is often abbreviated as HC (for heteroskedasticity-consistent). This class of estimators includes HC_0 (White, 1980), HC_1 (Hinkley, 1977), HC_2 (MacKinnon and White, 1985), and HC_3 (MacKinnon and White, 1985). (For definitions of these estimators, see Section 8.1.)

In the context of heteroskedasticity-robust variance estimation, a substantial literature has developed investigating the poor performance of HC_0 and HC_1. This literature includes MacKinnon and White (1985), Chesher and Jewitt (1987), Chesher (1989), Chesher and Austin (1991), Long and Ervin (2000), and Young (2019). This literature has coalesced on the recommendation to switch to HC_3/jackknife standard errors, which are simple to calculate, never-downward-biased, and robust to a variety of regressor settings.

There is also a literature exploring unbiased or approximately unbiased variance estimators, including Bera, Suprayitno, and Premaratne (2002), Cattaneo, Jansson, and Newey (2018), and Kline, Saggio, and Solvsten (2020). These estimators can be computationally prohibitive in large samples, are not necessarily non-negative, and have not yet been generalized to cluster-robust estimation.

Cluster-robust variance estimation was introduced by Liang and Zeger (1986) and Arellano (1987) as a natural extension of the heteroskedasticity-robust variance estimator. The common implementation codified by the Stata `cluster` variance option adds an ad hoc degree-of-freedom correction as an analog to the HC_1 estimator. Since the influential work of Bertrand, Duflo, and Mullainathan (2004), this estimator has become the ubiquitous approach for standard error construction for DiD regression.

An analog of HC_2 was proposed by Bell and McCaffrey (2002), endorsed by Imbens and Kolesár (2016), and codified in Stata 18. An analog of HC_3 was proposed and evaluated by MacKinnon, Nielsen, and Webb (2023abc). MacKinnon, Nielsen, and Webb (2023b) develop an efficient jackknife computational implementation. Hansen (2024) analyzed the statistical properties of this estimator with some modifications, and showed that this is the only known cluster-robust variance estimator which is never downward biased.

A number of papers investigate the poor performance of cluster-robust methods in regressions with a small number of clusters and/or a small number of treated clusters. This includes Ibragimov and Müller (2016), Rokicki, Cohen, Fink, Salomon, and Landrum (2018), Ferman and Pinto (2019), Hagemann (2019), and Niccodemi and Wansbeek (2022).

The jackknife estimator of variance was introduced by Tukey (1958) and was developed in the monographs of Efron (1982) and Shao and Tu (1995). Efron and Stein (1981) examined its statistical properties, and showed that a version of the jackknife estimator is never downward biased in certain settings.

A modified student t distributional approximation to t-ratios constructed with the Bell-McCaffrey standard error was proposed by Bell and McCaffrey (2002), Imbens and Kolesár (2016), and Pustejovsky and Tipton (2018), and a related method for the conventional cluster-robust t-ratio was proposed by Young (2016). Inference based on the wild bootstrap was proposed by Cameron, Gelbach, and Miller (2008), and its statistical properties investigated by Djogbenou, MacKinnon, and Nielsen (2019) and Canay, Santos, and Shaikh (2021). Randomization inference was proposed by MacKinnon and Webb (2020).

The performance of cluster-robust methods deteriorates when there are a small number of treated clusters. In the extreme case of one treated cluster, conventional inference methods fail. In contrast, as shown by Hansen (2024), a properly-constructed jackknife variance estimator remains never-downward-biased in this context, resulting in conservative inference (100% coverage). Other methods have been developed for inference with a single treated cluster under somewhat stronger assumptions by Conley and Taber (2011) and Hagemann (2023).

A Stata and R program `jregress` which calculates our recommended jackknife method is available on the author's website `users.ssc.wisc.edu/~bhansen/`, in addition to data and code for full replication of all numerical results reported in this paper.

2 Framework

The ubiquitous difference-in-difference equation (DiD) is the clustered twoway fixed effect regression

$$Y_{igt} = \theta D_{igt} + \gamma' Z_{igt} + \alpha_g + \phi_t + e_{igt} \tag{1}$$

where $g = 1, ..., G$ denotes group/cluster, i denotes individual, and t denotes the time period. The variable Y is the outcome, the binary variable D is treatment status, the vector Z contains a set of possible controls, α_g is a group-level fixed effect, ϕ_t is a time-level fixed effect, and e is a regression error. Typically, the treatment D applies to a subset of groups (the treated groups) for a subset of time periods (the treatment period). The coefficient θ is often the primary parameter of interest, and equals the Average Treatment Effect on the Treated (ATT) under a set of widely-studied conditions[1]. The observations are typically assumed to be cluster dependent within each group and independent across groups.

In the model (1), it is possible for the same individuals to be observed each time period, or for different individuals to be observed in each period. The "time" index can also be a stand-in for other subgroupings, such as different classrooms within a school.

While the classic DiD model (1) specifies fixed effects coincident with the level of clustering, many applications deviate from this structure. Some applications, for example, include a smaller set of fixed effects, presumbly for small sample considerations. Other applications may include more extensive interactive fixed effects. To allow these possibilities, we generalize (1) by allowing general fixed effects specifications, absorbing all fixed effects into the control vector Z_{igt} by inclusion of a suitable set of fixed effect dummy variables, and notationally omitting α_g and ϕ_t from the regression.

Notationally, let $X_{igt} = (D_{igt}, Z'_{igt})'$ denote the k full set of regressors including all fixed effects, and let β be the full set of coefficients. Stacking the observations by cluster, this regression model can be written at the cluster level as

$$\boldsymbol{Y}_g = \boldsymbol{X}_g \beta + \boldsymbol{e}_g. \tag{2}$$

The coefficients of (2) are typically estimated by least squares. This equals

$$\widehat{\beta} = \left(\sum_{g=1}^{G} \boldsymbol{X}'_g \boldsymbol{X}_g \right)^{-1} \left(\sum_{g=1}^{G} \boldsymbol{X}'_g \boldsymbol{Y}_g \right). \tag{3}$$

In the classic DiD model (1) with group and time fixed effects, this corresponds to the twoway fixed effects estimator. The least squares estimator (3) is the dominant estimator of DiD regressions in empirical applications, and therefore is our focus. However, the general ideas should be generalizable to other estimators.

We are interested in standard error construction and inference on the coefficients in (2).

We illustrate our goals with a well-known application. Card and Krueger (1994) estimated the effect of the 1992 increase of the New Jersey minimum wage on worker hours, by surveying fastfood restaurant employee hours both before the wage increase (February-March 1992) and after the wage increase (November-December 1992) in a sample of restaurants in New Jersey and eastern Pennsylvania. Their estimate can be calculated by a linear regression of restaurant hours on three variables: (1) *treatment* (a binary indicator for New Jersey after the wage increase); (2) *state* (a binary indicator for New Jersey); and (3) *time* (a binary indicator for the post-increase period). We calculate and report these regression

[1] This paper is not concerned with identification; there is a large literature focusing on the conditions under which θ equals the ATT, conditions under which this equality fails, and alternative estimation strategies which can be employed in such contexts.

estimates in Table 1 below, along with conventional clustered standard errors.

Table 1: Card and Krueger (1994)
Effect of Minimum Wage on Employment

	Coefficient	Std Err	t	pv	95% interval
Treatment	2.75	1.34	2.05	.041	[0.12, 5.38]
State	−2.95	1.48	−1.99	.047	[−5.86, −0.04]
Time	−2.28	1.25	−1.83	.068	[−4.74, 0.17]
Intercept	23.38	1.38	16.92	.000	[20.66, 26.10]
Clusters	Store (384)				
Observations	768				

We present the output as commonly displayed by regression packages. This is a list of all variables included in the regression. For each variable is displayed its coefficient estimate, standard error, t-ratio, p-value (for the test of the hypothesis that the coefficient equals zero), and a 95% confidence interval. Each of these pieces is useful to the researcher in their evaluation of the regression estimates, even though only a subset of this information is typically reported in a research paper.

After the coefficient estimate itself, the second most important statistic reported is the standard error. It is a direct measure of precision, and is also the foundation for the reported t-ratio, p-value, and confidence interval.

Our contention is that all statistics displayed in this table are important, as all are examined by an empirical researcher in the course of their investigation. It is desirable for all default reported statistics to be accurate in broad settings without user intervention. There should be default choices for their calculation which are reasonably accurate in any regression setting. It is important that these default methods apply to all coefficient estimates (not just a single estimate of interest), as the full regression output is often studied by researchers, even if the full model is not reported in their paper. Finally, it is important that default methods are computationally efficient, as users require quick results for routine calculations. These goals motivate our proposals.

3 Variance Matrix Estimation

The most common method for variance matrix estimation for (3) is the cluster-robust variance estimator (CRVE) of Liang and Zeger (1986) and Arellano (1987) with a degree of freedom correction. This equals

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

where $\widehat{\boldsymbol{e}}_g = \boldsymbol{Y}_g - \boldsymbol{X}_g\widehat{\beta}$ is the least squares residual vector for the gth cluster, n is the total number of observations, and k is the total number of regressors. We call the estimator (4) CRVE$_1$. The CRVE$_1$ estimator (4) is the natural cluster-dependence generalization of the heteroskedasticity-robust estimator HC$_1$ (see equation (16)).

The CRVE$_1$ estimator is simple and intuitive. However, it can be highly downward biased. Indeed, Hansen (2024) shows that the downward bias of $\widehat{\boldsymbol{V}}_1$ can be arbitrarily large. One consequence of this downward bias is that confidence intervals constructed using CRVE$_1$ standard errors can have coverage rates arbitrarily close to zero.

An alternative is the variance estimator of Bell and McCaffrey (2002), promoted by Imbens and Kolesár (2016). It is motivated as an unbiased estimator under the auxiliary assumption that the errors e_{igt} are i.i.d. Define the partial projection matrices

$$\boldsymbol{M}_g = \boldsymbol{I}_{n_g} - \boldsymbol{X}_g \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{X}'_g, \tag{5}$$

let $\boldsymbol{A}^{1/2}$ denote the symmetric square root of the matrix $\boldsymbol{A}$, and let $\boldsymbol{A}^{+}$ denote the Moore-Penrose generalized inverse of $\boldsymbol{A}$. Their estimator is

$$\widehat{\boldsymbol{V}}_2 = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \left(\sum_{g=1}^{G} \boldsymbol{X}'_g \boldsymbol{M}_g^{+1/2} \widehat{\boldsymbol{e}}_g \widehat{\boldsymbol{e}}'_g \boldsymbol{M}_g^{+1/2} \boldsymbol{X}_g \right) \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}. \tag{6}$$

We call the estimator (6) CRVE$_2$. The use of the generalized inverse in (6) was introduced by Kolesár (2023) so that CRVE$_2$ is defined even when $\boldsymbol{M}_g$ is non-invertible. This is a potentially important generalization, as the matrix $\boldsymbol{M}_g$ is not invertible in many important contexts, including when treatment is applied to only a single cluster. The CRVE$_2$ estimator is available in Stata 18 through its `vce(hc2 clustvar)` option. The CRVE$_2$ estimator (6) is the natural cluster-dependence generalization of the estimator HC$_2$ (see equation (17)).

As mentioned above, the CRVE$_2$ estimator has the attractive feature that it is unbiased when the errors are i.i.d. However, unbiasedness can fail when the errors have within-cluster correlation, are conditionally heteroskedastic, or one of the $\boldsymbol{M}_g$ matrices is non-invertible. Indeed, as shown by Hansen (2024), the downward bias of $\widehat{\boldsymbol{V}}_2$ can be arbitrarily large. This implies that confidence intervals constructed using CRVE$_2$ standard errors can have coverage rates arbitrarily close to zero.

A bootstrap variance estimator can be obtained by nonparametric pairs clustered resampling. Each bootstrap sample is constructed by resampling G clusters $(\boldsymbol{Y}_g, \boldsymbol{X}_g)$ with replacement from the original sample of clusters. Least squares estimation is applied to the bootstrap sample, producing the bootstrap estimator $\widehat{\beta}^*$. This is repeated B times, yielding the bootstrap replications $\{\widehat{\beta}_1^*, ..., \widehat{\beta}_B^*\}$. The bootstrap variance estimator is their empirical covariance matrix

$$\widehat{\boldsymbol{V}}_{\text{boot}} = \frac{1}{B-1} \sum_{b=1}^{B} \left(\widehat{\beta}_b^* - \overline{\widehat{\beta}}^*\right) \left(\widehat{\beta}_b^* - \overline{\widehat{\beta}}^*\right)'. \tag{7}$$

A complication is that it is possible that in some bootstrap samples the regressor matrix will not be full rank, implying that the bootstrap least squares estimator will not be uniquely defined. (This will occur with high probability if the number of treated clusters is small, for then it is possible to draw an entire bootstrap sample with no treated clusters.) It is typical (e.g., the Stata implementation) to discard these bootstrap samples and calculate the bootstrap variance only on the subset of bootstrap samples which have full rank regressor matrices. This seemingly technical workaround may be inconsequential if the frequency of discarded bootstrap samples is small, but if the frequency is high then this implementation induces selection bias. Consequently, we should not expect bootstrap variance estimation to be generically well-behaved.

The final variance matrix estimator we consider is the jackknife. There are several implementations; our recommendation is

$$\widehat{\boldsymbol{V}}_{\text{jack}} = \sum_{g=1}^{G} \left(\widehat{\beta}_{-g} - \widehat{\beta}\right)\left(\widehat{\beta}_{-g} - \widehat{\beta}\right)' \tag{8}$$

where

$$\widehat{\beta}_{-g} = \left(\boldsymbol{X}'\boldsymbol{X} - \boldsymbol{X}_g'\boldsymbol{X}_g\right)^{+}\left(\boldsymbol{X}'\boldsymbol{Y} - \boldsymbol{X}_g'\boldsymbol{Y}_g\right) \tag{9}$$

is a generalized delete-one-cluster estimator. By defining the jackknife variance estimator this way the estimator (9) is uniquely defined[2] and the sum (8) includes all clusters. In contrast, the most common implementation of the jackknife discards clusters from the sum (8) if the delete-one-cluster least squares estimator is not uniquely defined, which occurs, for example, when treatment is applied to a single cluster. This can severely downward bias the variance estimator. Two other differences between the definition (8) and some other definitions of the jackknife are that (8) does not use a degree-of-freedom correction[3], and (8) centers the delete-one-cluster estimators at the full-sample estimator $\widehat{\beta}$ rather than at the mean of $\widehat{\beta}_{-g}$. We do not use either modification as either leads to violation of the "never-downward-biased" property of (8) discussed below. The jackknife estimator (8) is the natural cluster-dependence generalization of the estimator HC$_3$ (see equations (18)-(19)).

Hansen (2024) established two important properties of the jackknife estimator (8). First, $\widehat{\boldsymbol{V}}_{\text{jack}}$ is never downward biased, in the sense that the expected value of $\widehat{\boldsymbol{V}}_{\text{jack}}$ is never less than (in a positive definite sense) the true variance matrix. This holds under broad conditions, including arbitrary cluster sizes, number of treated clusters, regressor leverage, within-cluster correlation, and heteroskedastisticity. Second, if the errors are normally distributed (but potentially heteroskedastic and within-cluster correlated) and the matrices $\boldsymbol{X}'\boldsymbol{X} - \boldsymbol{X}_g'\boldsymbol{X}_g$ are all invertible, then the finite sample distribution of a t-ratio constructed with the jackknife standard error is bounded by the Cauchy distribution. This implies that confidence intervals constructed with jackknife standard errors have guaranteed coverage rates, unlike intervals constructed with CRVE$_1$ and CRVE$_2$ standard errors.

The most common purpose of covariance matrix estimation is standard error construction. Let R be the $k \times 1$ vector which selects the coefficient of interest, e.g. for θ, $R = (1, 0, ..., 0)'$. Then a standard error for $\widehat{\theta} = R'\widehat{\beta}$ based on the covariance matrix estimator $\widehat{\boldsymbol{V}}$ is $\widehat{v} = \sqrt{R'\widehat{\boldsymbol{V}}R}$. Let $\widehat{v}_1$, $\widehat{v}_2$, $\widehat{v}_{\text{boot}}$, and $\widehat{v}_{\text{jack}}$ denote the standard errors constructed using (4), (6), (7), and (8), respectively.

Calculation of (8) is somewhat more computationally demanding than computation of (4) due to the need to calculate the G estimators (9). In Appendix 8.3 we present numerical evidence that this computational cost is minor in a variety of sample sizes and regressor dimensions.

[2] The theoretical properties of the jackknife variance estimator (8) described in this paper hold if (9) is constructed with any generalized inverse. We recommend the Moore-Penrose inverse as it is the unique minimum-length minimizer of the least-squares criterion, and thus tends to produce variance estimators (8) which are less excessively conservative, relative to estimates constructed with other generalized inverse formulae.

[3] In contrast, a common degree-of-freedom correction is to multiply (4) by $(G-1)/G$.

4 Adjusted P-Values and Confidence Intervals

Current empirical practice, as exemplified by the output displayed in Table 1, is to construct p-values and confidence intervals for individual coefficients based on the student t_{G-1} distribution, or the t_{n-k} distribution in the absence of clustering. These approximations can be very poor in practice as cluster-robust t-ratios do not in general have these distributions. An alternative simple student t approximation was introduced by Bell and McCaffrey (2002) for the HC$_2$ and CRVE$_2$ t-ratios, extended to CRVE$_1$ standard errors by Young (2016), and to jackknife t-ratios by Hansen (2024). This approximation can be used to produce adjusted p-values and confidence intervals which are simple to calculate and, in general, have excellent finite sample coverage. We now describe this approximation and adjusted inference methods.

Consider the t-ratio for θ constructed with the jackknife standard error,

$$T = \frac{\widehat{\theta} - \theta}{\widehat{\nu}_{\text{jack}}}.$$

Under the assumption that the regression error vector $\boldsymbol{e} \sim N(0, \boldsymbol{\Sigma})$ is jointly normally distributed (allowing for heteroskedasticity and arbitrary correlation), the coefficient estimator satisfies $\widehat{\theta} - \theta \sim N(0, \nu^2)$ where ν^2 is the finite-sample variance of $\widehat{\theta}$. Furthermore, with a little algebra, the variance estimator can be written as a quadratic function in the regression errors, $\widehat{\nu}^2_{\text{jack}} = \boldsymbol{e}'\boldsymbol{B}\boldsymbol{e}$, where $\boldsymbol{B}$ is a known (function of the regressors $\boldsymbol{X}$) positive-semi-definite matrix of rank at most G. It follows that $\widehat{\nu}^2_{\text{jack}}$ has the exact finite-sample distribution $\widehat{\nu}^2_{\text{jack}}/\nu^2 \sim \sum_{j=1}^{G} \lambda_j \chi^2_j$ where χ^2_j are independent chi-square random variables with one degree of freedom and $\lambda_j \geq 0$ are the eigenvalues of $\boldsymbol{B\Sigma}/\nu^2$. The widely-studied Satterthwaite (1946) approximation states that this weighted sum of chi-squares can be reasonably approximated by a single scaled chi-square, where the scale and degree-of-freedom are selected to match the first two moments. This approximation is

$$\sum_{j=1}^{G} \lambda_j \chi^2_j \approx a^2 \frac{\chi^2_K}{K}$$

where

$$a = \sqrt{\sum_{j=1}^{G} \lambda_j} \tag{10}$$

$$K = \frac{\left(\sum_{j=1}^{G} \lambda_j\right)^2}{\sum_{j=1}^{G} \lambda_j^2}. \tag{11}$$

Substituting this approximation into the expression for the t-ratio, we obtain the distributional approximation

$$T \approx \frac{N(0,1)}{a\sqrt{\frac{\chi^2_K}{K}}} \approx \frac{t_K}{a} \tag{12}$$

where t_K is distributed student t with K degrees of freedom. The second approximation in (12) holds with equality when the numerator and denominator are independent, which holds when $\boldsymbol{\Sigma} = \boldsymbol{I}_n\sigma^2$. The approximation (12) leads to the suggestion to use the scaled student t distribution t_K/a in place of the conventional t_{G-1} distribution for p-value calculation and confidence interval construction. The approximation is not exact, but it is much improved relative to the conventional t_{G-1} distribution.

This suggestion requires the calculation of the adjustment coefficients a and K, which are functions of the eigenvalues of the matrix $\boldsymbol{B}\boldsymbol{\Sigma}/v^2$. While $\boldsymbol{B}$ is known, the covariance matrix $\boldsymbol{\Sigma}$ is unknown, so the true values of a and K cannot be calculated. Bell and McCaffrey (2002) suggested to use a reference model (akin to a rule-of-thumb), in particular $\boldsymbol{\Sigma} = \boldsymbol{I}_n\sigma^2$. Using this reference model the coefficients a and K are straightforward functions of the regressor matrix $\boldsymbol{X}$. Explicit expressions are provided in equations (20) and (21) of Appendix 8.2, and computation is discussed in Appendix 8.3. The expressions depend on the specific coefficient (or, more generally, the specific linear combination R) and therefore need to be calculated separately for each coefficient. However, as documented in Appendix 8.3, these calculations are not computationally demanding.

Based on the distributional approximation (12) using (20) and (21), we propose adjusted confidence intervals and p-values for θ. The adjusted $1-\alpha$ confidence interval for θ is

$$\text{Jack}^* = \widehat{\theta} \pm \frac{t_K^{1-\alpha/2}}{a}\widehat{v}_{\text{jack}} \tag{13}$$

where $t_K^{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the student t distribution with K degress of freedom. The difference with the standard confidence interval is that (13) calculates the critical value using K degrees of freedom instead of $G-1$, and scales down the critical value by a.

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

$$p^* = 1 - F\left(a^2\left(\frac{\widehat{\theta}-\theta_0}{\widehat{v}_{\text{jack}}}\right)^2; 1, K\right) \tag{14}$$

where $F(x;1,K)$ is the F distribution with degrees of freedom $(1,K)$. The difference with the standard p-value is that (14) scales the t-statistic by a, and calculates significance using K degrees of freedom instead of $G-1$.

The adjusted degree-of-freedom K satisfies $1 \le K \le G$. Its value will reflect the degree of leverage and nonhomogeneity among the regressors and cluster sizes, with K equaling 1 in the most unbalanced cases. Small values of K are most likely to occur when the regressor of interest has high leverage, meaning that there are a small number of observations or clusters which dominate the variance of $\widehat{\theta}$. Common contexts where this occurs include models with cluster-level fixed effects, treatment indicators when there are a small number of treated clusters, and/or dummy variables which are non-zero for only a small number of observations or clusters. Small values of the degree-of-freedom K can also occur when regressors are leptokurtotic or when cluster sizes are highly unbalanced. These are the contexts (as discussed by MacKinnon, Nielsen, and Webb (2023a), Section 4.1) where conventional cluster-robust inference is known to be highly unreliable. Thus, if we see a small value of K for an estimated coefficient of interest, this is not only a signal to adjust the degree-of-freedom for jackknife standard errors, but it is a signal to avoid CRVE_1-based inference.

The scale a satisfies $a \geq 1$ and reflects the proportional bias of the jackknife standard error, calculated under the assumption of the reference model. Since the jackknife estimator is never downward biased, this constant satisfies $a \geq 1$.

The adjusted confidence interval (13) and p-value (14) will typically be more conservative than the intervals and p-values calculated with the conventional t_{G-1} distribution, but they are not necessarily so, as the adjustments K and a work in opposite directions. If desired, more conservative inference can be achieved by two possible modifications. First, the adjustment a could be omitted from (13) and (14), meaning that inference would be based on the jackknife t-ratio with the adjusted degree-of-freedom K. I do not recommend this modification as it appears to lead to excessively conservative inference under high leverage. Second, the confidence interval and p-value can be calculated in two ways, by (13)-(14), and by using the t_{G-1} distribution (or t_{n-k} distribution for non-clustered observations) conventionally, and reporting the more conservative of the two. This latter modification is ad hoc, but ensures that the adjusted intervals are always more conservative than conventional intervals. The impact of this modification, however, appears to be minor in practice. For our reported simulations, empirical applications, and programs, we use (13)-(14) without modification.

5 Simulation

5.1 Potential Outcome Framework

We investigate the proposed methods in a simple simulation experiment.

The observations are $\{Y_{igt}, D_{gt}, Z_{jigt}\}$ for $g = 1, ..., G$, $t = 1, 2$, $j = 1, ..., J$, and $i = 1, ..., n_g$, where n_g is cluster size. The observations are generated from potential outcomes $Y_{igt}(D_{gt})$ where $D_{gt} \in \{0,1\}$ is treatment status. The clusters are divided into G_0 untreated clusters and G_1 treated clusters, with $G_0 + G_1 = G$. Treatment ($D_{gt} = 1$) is applied only in period $t = 2$ to the treated clusters. We vary the number of clusters among $G \in \{10, 20, 50, 200\}$ and the number of treated clusters among $G_1 \in \{4, 3, 2\}$. In our baseline model the cluster sizes are homogeneous, $n_g = 10$ for all g.

We generate the potential outcomes using a cluster-dependent framework. In our baseline model they are generated as:

$$\begin{aligned}
Y_{igt}(0) &= e_{igt} + u_g + h_{ig} v_g \\
e_{igt} &\sim N(0,1) \\
u_g &\sim N(0,1) \\
v_g &\sim N(0,1) \\
Y_{igt}(1) &= Y_{igt}(0) + \theta_{ig} \\
\theta_{ig} &\sim N(\theta, \sigma_\theta^2).
\end{aligned}$$

The coefficient h_{ig} is set to equal $+1$ for one-half of the individuals i in each cluster, and to equal -1 for the others. This specification creates cluster-level dependence in $Y_{igt}(0)$ which is not fully eliminated by the within transformation.

Notice that the model specifies that the treatment effect θ_{ig} is heterogeneous with ATT θ. We vary treatment effect heterogeneity by varying $\sigma_\theta \in \{1, 10\}$.

The variables Z_{jigt} are auxiliary regressors, generated as i.i.d. $Z_{jigt} \sim N(D_{gt}, 1)$ with $J = 2$ in the baseline model.

For each simulation replication we estimate the coefficients of the regression model (1) by least squares. This is a least squares regression[4] of the observed outcome Y_{igt} on treatment D_{gt}, the regressors Z_{jigt}, a time dummy, and group fixed effect dummies. The coefficient $\widehat{\theta}$ on D_{gt} is the estimated ATT. We calculate the four standard errors $\widehat{v}_1$, $\widehat{v}_2$, $\widehat{v}_{\text{boot}}$, and $\widehat{v}_{\text{jack}}$ discussed in Section 3, the bootstrap using $B = 999$ replications.

We evaluate eight confidence intervals for the ATT θ. The first four intervals combine the four standard errors with conventional student t critical values. Thus, given a standard error $\widehat{v}$ we form the interval $\widehat{\theta} \pm t_{G-1}^{0.975}\widehat{v}$ where $t_{G-1}^{0.975}$ is the 0.975 quantile of the t_{G-1} distribution. We use the $t_{G-1}^{0.975}$ critical value as this is the current implementation in Stata for cluster-robust inference.

The fifth and sixth intervals are the wild cluster bootstrap symmetric percentile-t interval calculated with the CRVE$_1$ and jackknife standard errors and 999 bootstrap replications. This (using CRVE$_1$) is the method proposed by Cameron, Gelbach, and Miller (2008) for hypothesis testing[5], and in principle could be used to construct confidence intervals by test inversion. First[6], the coefficients are re-estimated imposing the hypothesized value of θ to obtain restricted estimates $\widetilde{\beta}$ and residuals $\widetilde{\boldsymbol{e}}_g = \boldsymbol{Y}_g - \boldsymbol{X}_g\widetilde{\beta}$. Next, the clusters, regressors $\boldsymbol{X}_g$, and restricted residuals $\widetilde{\boldsymbol{e}}_g$ are held fixed. The bootstrap samples are generated as $\boldsymbol{Y}_g^* = \xi_g\widetilde{\boldsymbol{e}}_g$ where ξ_g is an independent Rademacher variable (equals +1 and −1 each with probability 1/2). The bootstrap sample then consists of the observations $(\boldsymbol{Y}_g^*, \boldsymbol{X}_g)$. On each bootstrap sample we calculate the least squares estimate $\widehat{\theta}^*$ and its CRVE$_1$ and jackknife standard errors $\widehat{v}_1^*$ and $\widehat{v}_{\text{jack}}^*$. From the 999 bootstrap samples we calculate the 95% quantiles $\widehat{c}_1^*(\theta)$ and $\widehat{c}_{\text{jack}}^*(\theta)$ of the statistics $\left|\widehat{\theta}^*\right|/\widehat{v}_1^*$ and $\left|\widehat{\theta}^*\right|/\widehat{v}_{\text{jack}}^*$. The wild bootstrap confidence intervals[7] equal $\text{Wild}_1 = \left\{\theta : \left|\widehat{\theta} - \theta\right|/\widehat{v}_1 \le \widehat{c}_1^*(\theta)\right\}$ and $\text{Wild}_\text{J} = \left\{\theta : \left|\widehat{\theta} - \theta\right|/\widehat{v}_{\text{jack}} \le \widehat{c}_{\text{jack}}^*(\theta)\right\}$.

Our seventh confidence interval is the adjusted CRVE$_2$ interval proposed by Bell and McCaffrey (2002). This is $\text{BM} = \widehat{\theta} \pm t_K^{0.975}\widehat{v}_2$ where $\widehat{v}_2$ is the CRVE$_2$ standard error and K is a non-standard degree-of-freedom[8] calculated similar to (11).

Our final confidence interval Jack* is our proposed adjusted jackknife interval (13).

By simulation with 20,000 replications, we compute the empirical coverage probability of these nominal 95% intervals.

[4] Computationally we use the within estimator to eliminate the group-level fixed effects, as this is algebraically equivalent to the full least squares regression yet computationally more efficient.

[5] MacKinnon, Nielsen and Webb (2023b) review several variants of the wild cluster bootstrap. Our implementation correspond to their WCR-C and WCR-V methods.

[6] We describe here a conceptual implementation of the wild bootstrap algorithm. For our actual calculation we use the fast computation algorithm described in MacKinnon (2023).

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

[8] See Kolesár (2023) for efficient computation.

Table 2: Baseline Model: Coverage of Nominal 95% Confidence Intervals

G	σ_θ	G_1	CRVE$_1$	CRVE$_2$	Boot	Jack	Wild$_1$	Wild$_J$	BM	Jack*
10	1	4	0.93	0.94	0.93	0.96	0.93	0.93	0.95	0.95
10	1	3	0.91	0.92	0.91	0.95	0.93	0.93	0.96	0.96
10	1	2	0.85	0.88	0.84	0.92	0.96	0.96	0.99	0.99
10	10	4	0.89	0.90	0.89	0.93	0.89	0.90	0.91	0.91
10	10	3	0.83	0.86	0.84	0.90	0.83	0.84	0.90	0.91
10	10	2	0.70	0.76	0.69	0.83	0.69	0.69	0.91	0.94
20	1	4	0.91	0.92	0.91	0.94	0.94	0.94	0.96	0.96
20	1	3	0.87	0.89	0.88	0.92	0.95	0.94	0.96	0.97
20	1	2	0.79	0.82	0.78	0.87	0.99	0.99	1.00	1.00
20	10	4	0.85	0.88	0.87	0.91	0.90	0.90	0.92	0.93
20	10	3	0.79	0.83	0.81	0.88	0.81	0.81	0.92	0.93
20	10	2	0.65	0.73	0.66	0.80	0.69	0.70	0.93	0.95
50	1	4	0.87	0.89	0.89	0.92	0.94	0.93	0.96	0.96
50	1	3	0.82	0.85	0.84	0.89	0.98	0.98	0.97	0.97
50	1	2	0.70	0.77	0.71	0.83	1.00	1.00	1.00	0.99
50	10	4	0.84	0.87	0.86	0.90	0.89	0.89	0.94	0.94
50	10	3	0.78	0.82	0.80	0.87	0.80	0.81	0.94	0.94
50	10	2	0.63	0.71	0.64	0.79	0.76	0.77	0.95	0.95
200	1	4	0.83	0.87	0.86	0.90	0.94	0.94	0.95	0.95
200	1	3	0.78	0.82	0.80	0.87	1.00	1.00	0.96	0.96
200	1	2	0.64	0.72	0.65	0.79	1.00	1.00	0.99	0.98
200	10	4	0.83	0.86	0.86	0.89	0.89	0.89	0.95	0.95
200	10	3	0.76	0.81	0.79	0.86	0.84	0.84	0.95	0.95
200	10	2	0.61	0.70	0.63	0.78	0.93	0.93	0.95	0.95

Table 3: Asymmetric Cluster Sizes: Coverage of Nominal 95% Confidence Intervals

G	σ_θ	G_1	CRVE$_1$	CRVE$_2$	Boot	Jack	Wild$_1$	Wild$_J$	BM	Jack*
10	1	4	0.87	0.93	1.00	0.98	0.93	0.94	0.98	1.00
10	1	3	0.85	0.93	0.99	0.98	0.95	0.94	0.99	1.00
10	1	2	0.79	0.90	0.95	0.98	0.97	0.97	0.99	0.99
10	10	4	0.63	0.81	0.99	0.94	0.68	0.85	0.89	0.97
10	10	3	0.59	0.79	0.97	0.94	0.68	0.81	0.90	0.97
10	10	2	0.50	0.76	0.87	0.94	0.65	0.72	0.90	0.96
20	1	4	0.81	0.91	1.00	0.97	0.97	0.95	0.99	1.00
20	1	3	0.78	0.90	0.99	0.97	0.98	0.97	1.00	1.00
20	1	2	0.69	0.87	0.94	0.97	0.99	0.99	1.00	0.99
20	10	4	0.56	0.78	0.99	0.93	0.68	0.87	0.94	0.97
20	10	3	0.52	0.76	0.97	0.93	0.68	0.82	0.94	0.97
20	10	2	0.42	0.73	0.85	0.93	0.69	0.74	0.93	0.95
50	1	4	0.72	0.88	1.00	0.97	0.99	0.96	1.00	0.99
50	1	3	0.69	0.87	0.99	0.97	1.00	0.99	1.00	0.99
50	1	2	0.58	0.83	0.91	0.96	1.00	1.00	1.00	0.98
50	10	4	0.51	0.77	0.99	0.93	0.75	0.90	0.97	0.97
50	10	3	0.47	0.75	0.97	0.93	0.77	0.85	0.97	0.97
50	10	2	0.39	0.72	0.85	0.93	0.78	0.81	0.95	0.95
200	1	4	0.65	0.85	1.00	0.96	1.00	0.97	1.00	0.99
200	1	3	0.62	0.84	0.98	0.96	1.00	1.00	1.00	0.98
200	1	2	0.50	0.80	0.89	0.95	1.00	1.00	0.99	0.97
200	10	4	0.49	0.76	0.99	0.93	0.88	0.91	0.98	0.97
200	10	3	0.45	0.74	0.96	0.93	0.88	0.92	0.98	0.96
200	10	2	0.36	0.70	0.84	0.93	0.94	0.95	0.95	0.95

Table 4: Geometrically Distributed Cluster Sizes: Coverage of Nominal 95% Confidence Intervals

G	σ_θ	G_1	CRVE$_1$	CRVE$_2$	Boot	Jack	Wild$_1$	Wild$_J$	BM	Jack*
10	1	4	0.91	0.93	0.97	0.96	0.94	0.94	0.96	0.97
10	1	3	0.89	0.92	0.95	0.95	0.94	0.94	0.97	0.98
10	1	2	0.82	0.87	0.88	0.94	0.96	0.96	0.99	0.99
10	10	4	0.85	0.88	0.93	0.92	0.87	0.86	0.92	0.93
10	10	3	0.79	0.84	0.90	0.91	0.82	0.83	0.92	0.94
10	10	2	0.65	0.76	0.76	0.87	0.73	0.75	0.91	0.95
20	1	4	0.88	0.90	0.95	0.93	0.94	0.94	0.97	0.98
20	1	3	0.84	0.88	0.92	0.92	0.96	0.96	0.98	0.99
20	1	2	0.74	0.82	0.81	0.90	0.99	0.99	0.99	0.99
20	10	4	0.80	0.85	0.91	0.90	0.86	0.85	0.93	0.95
20	10	3	0.74	0.81	0.87	0.88	0.79	0.80	0.93	0.96
20	10	2	0.60	0.73	0.72	0.85	0.74	0.75	0.94	0.95
50	1	4	0.83	0.87	0.93	0.91	0.95	0.96	0.98	0.98
50	1	3	0.78	0.84	0.89	0.90	0.99	0.99	0.98	0.99
50	1	2	0.66	0.78	0.76	0.87	1.00	1.00	0.99	0.99
50	10	4	0.78	0.83	0.91	0.89	0.87	0.86	0.95	0.96
50	10	3	0.71	0.79	0.86	0.87	0.81	0.82	0.95	0.97
50	10	2	0.56	0.71	0.70	0.84	0.81	0.82	0.95	0.95
200	1	4	0.80	0.85	0.92	0.90	0.96	0.97	0.97	0.98
200	1	3	0.74	0.82	0.88	0.88	1.00	1.00	0.98	0.98
200	1	2	0.61	0.74	0.73	0.85	1.00	1.00	0.99	0.97
200	10	4	0.78	0.83	0.91	0.88	0.88	0.88	0.96	0.97
200	10	3	0.70	0.79	0.86	0.86	0.87	0.88	0.96	0.97
200	10	2	0.55	0.71	0.69	0.83	0.94	0.94	0.95	0.95

5.2 Baseline Model

We report the results for the baseline model in Table 2. Ideally, all entries should equal 0.95. However, many of the actual entries are far from this ideal. The CRVE$_1$ interval undercovers in all designs, and in many settings quite severely, with a worst-case coverage of 61%. Undercoverage is increasing as the asymmetry in the number of treated clusters and/or treatment effect heterogeneity is increased. Undercoverage is also increasing as the number of clusters increases, because this increases the asymmetry between treated and untreated clusters.

The CRVE$_2$ interval has improved coverage relative to CRVE$_1$, but also undercovers in all designs. As for CRVE$_1$, undercoverage is increasing in treatment asymmetry, treatment effect heterogeneity, and as the number of clusters increases. Its worst-case coverage is 70%.

The bootstrap interval has similar coverage to CRVE$_1$ and thus severely undercovers. Its worst-case coverage is 63%.

The jackknife interval with conventional critical values has better coverage relative to CRVE$_1$, CRVE$_2$, and the bootstrap, but undercovers under asymmetry in the number of treated clusters and under treatment effect heterogeneity. Its worst-case coverage is 78%.

The Wild bootstrap confidence intervals have mixed results. First, we observe that in this baseline specification, the Wild_1 and Wild_J intervals have essentially identical results. Their coverage rates are not strictly ranked relative to CRVE_1, CRVE_2, the bootstrap, or the jackknife. Their coverage rates generally improve as G increases. They have excellent coverage when treatment effect heterogeneity is mild, but have poor coverage when treatment effect heterogeneity is large. Their worst-case coverage is 69%.

The Bell-McCaffrey and adjusted jackknife confidence intervals both have generally good coverage, and both dominate the other six intervals. In most cases the two have similar coverage rates, but in some designs the adjusted jackknife interval has better coverage. In some cases they are conservative with coverage rates as high as 100%. Their worst-case coverage rates are 90% (Bell-McCaffrey) and 91% (adjusted jackknife).

5.3 Non-Homogeneous Cluster Sizes

We next investigate the impact of non-homogeneous cluster sizes. We modify the treated clusters only, by setting one treated cluster to have size $n_1 = 1+9G_1$ with the remaining treated clusters with size $n_g = 1$. All untreated cluster sizes are set at $n_g = 10$. This design maximizes nonhomogeneity among treated cluster sizes while maintaining the same number $(10G_1)$ of treated observations. The simulation estimates of the coverage rates are presented in Table 3. We find that the coverage rates of CRVE_1 and CRVE_2 are uniformly worse than in the baseline model, with worst-case coverage of 36% (CRVE_1) and 70% (CRVE_2). The bootstrap performs better than in the baseline model, and performs better than CRVE_1 and CRVE_2, but undercovers in some designs, with a worst-case coverage of 84%. The jackknife interval with conventional critical values also performs better than in the baseline model, with very good coverage rates, and worst-case coverage of 93%. In this specification the two wild bootstrap methods have significantly different performance under treatment effect heterogeneity, with Wild_J generally performing much better than Wild_1. However, both methods still under-cover, with worst-case coverage rates of 65% (Wild_1) and 72% (Wild_J). The Bell-McCaffrey interval has mixed performance, with worst-case coverage of 89%. The adjusted jackknife interval has excellent coverage, uniformly 95% or higher.

5.4 Random Cluster Sizes

To explore the impact of varied cluster sizes, for our next experiment we use a random cluster size design. We generate the cluster sizes as 1 plus an i.i.d. draw from the geometric distribution with parameter 0.1. This process implies that the average cluster size is 10 with a standard deviation of about 9.5. This sampling framework technically lies outside the "fixed cluster size" distributional framework, though the latter obtains by conditioning on the cluster sizes, similar to a regression model with exogenous regressors. The simulation estimates of the coverage rates are presented in Table 4. The results are similar to those obtained in the baseline model, with worst-case coverage rates of 55% (CRVE_1), 71% (CRVE_2), 69% (bootstrap), 83% (jackknife with conventional critical values), 73% (wild bootstrap), 91% (Bell-McCaffrey), and 93% (adjusted jackknife). Again, the adjusted jackknife has the best performance.

Table 5: Skewed Heavy-Tailed Errors: Coverage of Nominal 95% Confidence Intervals

G	σ_θ	G_1	CRVE$_1$	CRVE$_2$	Boot	Jack	Wild$_1$	Wild$_J$	BM	Jack*
10	1	4	0.94	0.94	0.94	0.96	0.94	0.94	0.95	0.95
10	1	3	0.92	0.93	0.92	0.95	0.93	0.93	0.96	0.96
10	1	2	0.86	0.88	0.85	0.92	0.96	0.96	0.99	0.99
10	10	4	0.88	0.90	0.89	0.92	0.89	0.89	0.91	0.91
10	10	3	0.83	0.86	0.84	0.90	0.83	0.83	0.90	0.91
10	10	2	0.70	0.76	0.70	0.82	0.69	0.70	0.90	0.93
20	1	4	0.91	0.92	0.92	0.94	0.94	0.94	0.96	0.96
20	1	3	0.87	0.89	0.88	0.92	0.95	0.95	0.97	0.97
20	1	2	0.78	0.82	0.78	0.87	0.99	0.99	1.00	1.00
20	10	4	0.85	0.87	0.87	0.91	0.89	0.89	0.92	0.93
20	10	3	0.79	0.83	0.81	0.88	0.81	0.81	0.91	0.93
20	10	2	0.65	0.72	0.65	0.80	0.69	0.70	0.93	0.95
50	1	4	0.87	0.89	0.89	0.92	0.93	0.93	0.96	0.96
50	1	3	0.82	0.85	0.84	0.89	0.98	0.98	0.97	0.96
50	1	2	0.70	0.77	0.71	0.83	1.00	1.00	1.00	0.99
50	10	4	0.83	0.86	0.86	0.89	0.88	0.88	0.93	0.94
50	10	3	0.77	0.82	0.80	0.86	0.80	0.80	0.93	0.94
50	10	2	0.62	0.71	0.63	0.79	0.75	0.76	0.95	0.95
200	1	4	0.84	0.87	0.86	0.90	0.94	0.94	0.95	0.95
200	1	3	0.78	0.83	0.81	0.88	1.00	1.00	0.96	0.96
200	1	2	0.65	0.73	0.66	0.80	1.00	1.00	0.99	0.98
200	10	4	0.82	0.85	0.85	0.89	0.88	0.88	0.94	0.94
200	10	3	0.75	0.81	0.79	0.86	0.84	0.84	0.94	0.95
200	10	2	0.61	0.70	0.62	0.78	0.93	0.93	0.95	0.95

5.5 Non-Normal Errors

We next investigate the robustness of the results to the assumption of normal errors. For this investigation we draw the errors for $Y_{igt}(0)$ and θ_{ig} from a skewed heavy-tailed distribution[9]. The simulation estimates of the coverage rates are presented in Table 5. The results are almost identical to those under normal errors.

5.6 Binary Dependent Variable

Many difference-in-difference applications concern binary dependent variables in a linear probability model. Our third model for potential outcomes treats this case directly with a probit generating process. The potential outcomes are generated as follows. For some $\alpha \geq 0$,

$$\begin{aligned} Y_{igt}(0) &= \mathbf{1}\{e_{igt} + u_g + h_{ig} v_g > \alpha\} \\ Y_{igt}(1) &= \mathbf{1}\{e_{igt} + u_g + h_{ig} v_g > 0\} \end{aligned}$$

where e_{igt}, u_g, v_g, and h_{ig} are generated as in the baseline model. In this model the treatment effect is $\theta_{ig} = \mathbf{1}\{0 < e_{igt} + u_g + h_{ig} v_g \leq \alpha\}$ with ATT $\theta = \Phi(\alpha/\sqrt{3}) - \Phi(0)$. Treatment effect heterogeneity is increasing in α. We vary $\alpha \in \{0.1, 3\}$.

The simulation estimates of the coverage rates are presented in Table 6. For most of the designs and methods, the results are quite similar to those obtained under normal errors. The adjusted jackknife has worst-case coverage of 92%.

5.7 Fixed Effects

While the classic difference-in-difference framework includes group fixed effects at the same level as clustering, in many applications (including those presented in the following section) there is a divergence between the fixed effect and clustering level. The typical deviation is that there are fewer included fixed effects than the level of clustering; or, equivalently, clustering is done at a finer level than the fixed effects. This is done, typically, to conserve estimation degrees-of-freedom. As an example, the Card and Krueger estimates of Table 1 include state-level fixed effects but cluster at the restaurant level.

To explore the impact of differential fixed effect inclusion, we group our clusters into $N = G/5$ large groups, and replace the cluster-level fixed effects with these N large-group fixed effects. Specifically, as we vary cluster sizes as $G = \{10, 20, 50, 200\}$ we include $N = \{2, 4, 10, 40\}$ large-group fixed effects.

We present the results in Table 7. The methods perform qualitatively similarly as in the baseline model.

[9] We use the "strongly skewed" distribution displayed in Figure 3.7(b) of Hansen (2022), which is a 9-component normal mixture distribution with a skew of 1.34 and kurtosis of 6.7.

Table 6: Binary Dependent Variable: Coverage of Nominal 95% Confidence Intervals

G	α	G_1	CRVE$_1$	CRVE$_2$	Boot	Jack	Wild$_1$	Wild$_J$	BM	Jack*
10	0.1	4	0.94	0.95	0.94	0.97	0.94	0.94	0.95	0.95
10	0.1	3	0.93	0.94	0.93	0.96	0.94	0.94	0.97	0.97
10	0.1	2	0.88	0.90	0.88	0.94	0.97	0.97	0.99	1.00
10	3	4	0.89	0.90	0.89	0.93	0.90	0.91	0.91	0.92
10	3	3	0.84	0.86	0.85	0.90	0.85	0.85	0.90	0.92
10	3	2	0.72	0.77	0.72	0.84	0.73	0.74	0.92	0.94
20	0.1	4	0.92	0.93	0.92	0.95	0.95	0.94	0.96	0.97
20	0.1	3	0.89	0.91	0.89	0.94	0.96	0.96	0.98	0.98
20	0.1	2	0.82	0.86	0.81	0.90	1.00	1.00	1.00	1.00
20	3	4	0.85	0.88	0.87	0.91	0.90	0.90	0.92	0.93
20	3	3	0.80	0.83	0.82	0.88	0.82	0.82	0.93	0.94
20	3	2	0.67	0.74	0.67	0.81	0.78	0.79	0.94	0.95
50	0.1	4	0.88	0.90	0.90	0.92	0.95	0.94	0.97	0.97
50	0.1	3	0.84	0.87	0.85	0.91	0.99	0.99	0.98	0.98
50	0.1	2	0.74	0.80	0.74	0.85	1.00	1.00	1.00	1.00
50	3	4	0.84	0.87	0.86	0.90	0.89	0.89	0.94	0.94
50	3	3	0.78	0.81	0.80	0.87	0.83	0.83	0.94	0.94
50	3	2	0.63	0.71	0.64	0.79	0.90	0.91	0.95	0.94
200	0.1	4	0.84	0.87	0.87	0.91	0.95	0.95	0.96	0.96
200	0.1	3	0.79	0.84	0.82	0.89	1.00	1.00	0.97	0.97
200	0.1	2	0.68	0.76	0.69	0.82	1.00	1.00	0.98	0.97
200	3	4	0.83	0.85	0.85	0.89	0.88	0.88	0.94	0.94
200	3	3	0.77	0.80	0.79	0.85	0.93	0.93	0.94	0.94
200	3	2	0.60	0.69	0.61	0.78	1.00	1.00	0.94	0.93

Table 7: Large-Group Fixed Effects: Coverage of Nominal 95% Confidence Intervals

G	σ_θ	G_1	CRVE$_1$	CRVE$_2$	Boot	Jack	Wild$_1$	Wild$_J$	BM	Jack*
10	1	4	0.93	0.96	0.96	0.99	0.94	0.94	0.97	0.98
10	1	3	0.90	0.93	0.93	0.97	0.92	0.92	0.96	0.97
10	1	2	0.85	0.90	0.87	0.94	0.91	0.91	0.98	0.99
10	10	4	0.88	0.91	0.90	0.94	0.91	0.91	0.92	0.93
10	10	3	0.85	0.88	0.86	0.92	0.87	0.88	0.90	0.92
10	10	2	0.73	0.79	0.74	0.86	0.77	0.77	0.90	0.93
20	1	4	0.92	0.96	0.95	0.98	0.94	0.94	0.98	0.98
20	1	3	0.88	0.93	0.92	0.97	0.92	0.92	0.98	0.98
20	1	2	0.82	0.88	0.84	0.93	0.92	0.91	0.99	1.00
20	10	4	0.85	0.89	0.88	0.92	0.92	0.92	0.92	0.93
20	10	3	0.80	0.85	0.83	0.90	0.86	0.87	0.91	0.92
20	10	2	0.69	0.76	0.70	0.84	0.78	0.78	0.93	0.95
50	1	4	0.91	0.95	0.94	0.97	0.94	0.94	0.98	0.99
50	1	3	0.86	0.92	0.91	0.96	0.92	0.92	0.99	0.99
50	1	2	0.80	0.87	0.82	0.92	0.92	0.92	1.00	1.00
50	10	4	0.84	0.88	0.87	0.91	0.91	0.92	0.93	0.94
50	10	3	0.79	0.84	0.82	0.89	0.87	0.87	0.92	0.93
50	10	2	0.66	0.75	0.68	0.82	0.81	0.80	0.95	0.96
200	1	4	0.90	0.94	0.94	0.97	0.94	0.94	0.98	0.99
200	1	3	0.86	0.92	0.91	0.96	0.92	0.92	0.99	0.99
200	1	2	0.78	0.86	0.81	0.92	0.92	0.92	1.00	1.00
200	10	4	0.83	0.87	0.87	0.91	0.91	0.92	0.94	0.94
200	10	3	0.77	0.83	0.81	0.88	0.86	0.87	0.93	0.93
200	10	2	0.65	0.74	0.67	0.81	0.84	0.83	0.95	0.96

Table 8: Ten Auxillary Regressors: Coverage of Nominal 95% Confidence Intervals

G	σ_θ	G_1	CRVE$_1$	CRVE$_2$	Boot	Jack	Wild$_1$	Wild$_J$	BM	Jack*
10	1	4	0.95	0.95	0.95	0.97	0.94	0.94	0.95	0.95
10	1	3	0.94	0.94	0.94	0.96	0.94	0.94	0.96	0.96
10	1	2	0.91	0.92	0.91	0.95	0.95	0.95	0.98	0.99
10	10	4	0.91	0.91	0.92	0.94	0.90	0.91	0.92	0.92
10	10	3	0.86	0.88	0.88	0.92	0.86	0.87	0.90	0.91
10	10	2	0.76	0.80	0.77	0.86	0.77	0.78	0.89	0.93
20	1	4	0.93	0.94	0.94	0.95	0.94	0.94	0.96	0.96
20	1	3	0.91	0.92	0.92	0.94	0.95	0.94	0.97	0.97
20	1	2	0.86	0.89	0.86	0.92	0.98	0.98	0.99	1.00
20	10	4	0.86	0.88	0.88	0.91	0.90	0.90	0.91	0.92
20	10	3	0.81	0.84	0.83	0.89	0.84	0.84	0.90	0.92
20	10	2	0.68	0.75	0.69	0.82	0.72	0.74	0.91	0.94
50	1	4	0.90	0.91	0.91	0.93	0.95	0.94	0.96	0.96
50	1	3	0.86	0.89	0.88	0.92	0.96	0.96	0.98	0.97
50	1	2	0.78	0.82	0.78	0.87	1.00	1.00	1.00	1.00
50	10	4	0.84	0.87	0.86	0.90	0.90	0.90	0.92	0.93
50	10	3	0.78	0.82	0.80	0.87	0.82	0.82	0.92	0.93
50	10	2	0.63	0.72	0.64	0.79	0.74	0.75	0.94	0.95
200	1	4	0.86	0.88	0.88	0.91	0.94	0.94	0.96	0.96
200	1	3	0.80	0.84	0.83	0.89	1.00	1.00	0.97	0.97
200	1	2	0.68	0.75	0.69	0.82	1.00	1.00	1.00	0.99
200	10	4	0.82	0.85	0.85	0.89	0.88	0.88	0.94	0.94
200	10	3	0.76	0.81	0.79	0.86	0.82	0.83	0.94	0.94
200	10	2	0.62	0.70	0.63	0.78	0.89	0.89	0.95	0.95

5.8 Many Auxiliary Regressors

Our baseline regression included two auxiliary regressors ($J = 2$). To explore the impact of varying this specification we repeat the exercise including ten auxiliary regressors ($J = 10$).

The results are reported in Table 8. The coverage rates are qualitatively similar to those in the baseline model. The difference is that many of the methods have somewhat improved coverage rates for small G. Overall, the impact of varying the number of auxiliary regressors is minor.

5.9 One Treated Cluster

For our final simulation we investigate performance in a model with one treated cluster ($G_1 = 1$). It should be emphasized that this is a treacherous context where it is well known that standard methods fail. Regardless, we believe that investigating performance in this context sheds insight concerning robustness to extreme situations. We repeat our analysis using the baseline model with normal innovations as in Table 2, but now set $G_1 = 1$. We report the results in Table 9.

As might be expected, the confidence interval methods have poor performance. The CRVE$_1$, CRVE$_2$, bootstrap, and BM methods have similar dramatic undercoverage. All have worst-case coverage of 2%-

Table 9: One Treated Cluster: Coverage of Nominal 95% Confidence Intervals

G	σ_θ	CRVE$_1$	CRVE$_2$	Boot	Jack	Wild$_1$	Wild$_J$	BM	Jack*
10	1	0.59	0.58	0.55	1.00	1.00	0.99	0.58	1.00
20	1	0.43	0.42	0.41	1.00	1.00	1.00	0.42	1.00
50	1	0.28	0.27	0.27	1.00	1.00	1.00	0.27	1.00
200	1	0.14	0.13	0.13	1.00	1.00	1.00	0.13	1.00
10	10	0.17	0.17	0.15	1.00	0.60	0.59	0.17	1.00
20	10	0.10	0.10	0.09	1.00	0.69	0.68	0.10	1.00
50	10	0.05	0.05	0.05	1.00	0.85	0.85	0.05	1.00
200	10	0.03	0.02	0.02	1.00	0.99	0.99	0.02	1.00

3%. The Wild bootstrap displays undercoverage when there is high treatment effect heterogeneity, with worst-case coverage of 58%. Essentially, all of these methods produce confidence intervals which are much too small.

In contrast, the jackknife and adjusted jackknife intervals are conservative, with 100% coverage. What happens is that when there is one treated cluster we find that $\widehat{v}_{\text{jack}} \simeq \left|\widehat{\theta}\right|$, the jackknife standard error approximately equals the coefficient estimate $\widehat{\theta}$, and thus its t-ratio is always close to 1 and never “significant”. Essentially, robust inference on the treatment effect when there is one treated cluster is similar to inference on the mean when there is a single observation with an unknown variance. The jackknife interval is not informative about the treatment effect, but is also not misleading regarding significance.

5.10 Summary of Simulation Evidence

Comparing the eight feasible confidence interval methods across Tables 2-9, the only method with reasonable coverage control in all contexts is the adjusted jackknife. The simulation evidence strongly supports our recommended procedure: use jackknife standard errors and base inference on the adjusted student t distribution.

These results are largely consistent with previous simulation studies, including that of MacKinnon, Nielsen, and Webb (2023b). One difference between our simulation and theirs is our investigation of the impact of treatment effect heterogeneity (which induces conditional heteroskedasticity), as MacKinnon, Nielsen, and Webb (2023b) only investigate homoskedastic designs. This explains the divergence between our findings and theirs concerning the Wild bootstrap. our results show that the Wild bootstrap performs well under low treatment effect heterogeneity (e.g., homoskedasticity) but not under high treatment effect heterogeneity (e.g., heteroskedasticity).

It is worthwhile to discuss in greater detail the contrast between the performance of the Bell-McCaffrey and adjusted jackknife intervals. Why should we prefer one over the other? The Jack* interval has three distinct advantages. First, it is robust to the context of a single treated cluster, while BM is not. In this context, the matrix $\boldsymbol{M}_g$ is not invertible for the treated cluster, and the CRVE$_2$ estimator uses its generalized inverse as an ad hoc workaround. A consequence is that the CRVE$_2$ variance estimator is downward biased. This problem extends to inference on any regression coefficient which suffers from “delete-one-

Table 10: Card and Krueger (1994)
Effect of Minimum Wage on Employment

	Coefficient	Std Err	t	pv	95% interval	K	a
CRVE_1	2.75	1.34	2.05	.041	[0.12, 5.38]		
Jackknife	2.75	1.35	2.04	.043	[0.89, 5.41]	112	1.01
Fixed Effects:	State (2), Time (2)						
Clusters	Store (384)						
Observations	768						

cluster" invertibility failure, which arises frequently in applications. In these contexts, the CRVE_2 standard errors and BM intervals will be misleadingly small. The second advantage of the Jack* interval is that it is built from the t-ratio with the jackknife standard error, which by itself produces confidence intervals with better coverage than t-ratios with CRVE_2 standard errors. Therefore, the joint display of $\widehat{v}_{\mathrm{jack}}$ with the adjusted p-values and confidence intervals is more internally consistent than the joint display of $\widehat{v}_2$ with the BM p-values and confidence intervals. Third, the simulation results explored show that Jack* has uniformly better coverage control than BM.

6 Illustrations

We illustrate the application of the jackknife standard errors and adjusted inference methods by application to multiple datasets. Our purpose is to demonstrate how inferences can meaningfully change in some contexts, while being unaltered in others.

6.1 Card and Krueger (1994)

For our first application we return to the Card and Krueger (1994) investigation of the impact of the minimum wage on employment hours. In the first line of Table 10 we repeat the estimated treatment effect coefficient from Table 1, and in the second line of Table 10 present the analogous result computed with jackknife standard errors, together with p-values and confidence intervals calculated using the student *t* adjustment. What we can see in this case is that there are only very minor changes in the standard errors, p-values, and confidence intervals.

We also display the data-based degree-of-freedom ($K = 112$) and scale adjustment ($a = 1.01$) for the jackknife inference adjustment. We can see that their values are consistent with essentially no meaningful adjustment being made. The reason, in this case, is because of the large number of clusters ($G = 384$) with a high degree of homogeneity.

To illustrate the fragility of inference we change the clustering level. In most current applications, clustering is done at a broad level of aggregation; indeed, most applications cluster at the level of treatment. In this example this would implying clustering by state, but this is infeasible as there are only two states in the sample. However, there is an intermediate case. The dataset includes an indicator for *region*, separating the New Jersey and eastern Pennslyvanian stores into three and two regions, respectively, for a

Table 11: Card and Krueger (1994)
Effect of Minimum Wage on Employment

	Coefficient	Std Err	t	pv	95% interval	K	a
$\mathtt{CRVE}_1$	2.75	1.17	2.35	.079	[−0.51, 6.01]		
Jackknife	2.75	2.09	1.31	.255	[−6.98, 12.48]	1.42	1.41
Fixed Effects:	State (2), Time (2)						
Clusters	Region (5)						
Observations	768						

total of five regions. We repeat the analysis, clustering by region. While this is a small number of clusters, it is not unusual in reported applications.

We report the results in Table 11. The first line reports the estimated treatment effect using CRVE_1 standard errors; the second line reports jackknife standard errors with adjusted p-values and confidence intervals. Examining the first line and comparing with Table 10, the changes are minimal, with the standard error decreasing somewhat. A researcher may be lulled into the false sense that "the results are robust to clustering by region". However, this interpretation vanishes when we examine the second line of Table 11. The jackknife standard error is nearly twice the magnitude of the CRVE_1 standard error, its p-value far from significant, and its 95% confidence interval extremely wide. The results are qualitatively different.

We also report (for the jackknife estimates) the degree-of-freedom K and scale a adjustment parameters for the distribution of the treatment effect t-ratio. In this setting we see that the degree-of-freedom equals $K = 1.4$, which is considerably smaller than the conventional degree-of-freedom $G - 1 = 4$. This is a signal that the conventional student t distribution approximation is poor, as the sample exhibits regressor leverage and heterogeneity. We also see that the scale adjustment $a = 1.41$ is considerably above 1, indicating that the jackknife standard error is likely biased upwards. Examining these two adjustment coefficients can be used to signal that conventional inference is unreliable.

It is not my purpose to take a stand on the level of clustering. Rather, my goal is for regression packages to report valid measures of precision for any regression a researcher might estimate. In the present application, it is my contention that the CRVE_1 standard error, p-value, and confidence interval presented in the first line of Table 11 are misleading, while the jackknife analogs in the second line are more reliable.

6.2 Bailey (2010)

Our second illustration is taken from Bailey (2010), who estimates the effect of sales bans on birth control use from surveys[10] of married women in 1965 and 1970, exploiting the 1965 U.S. Supreme Court *Griswold* decision which legalized contraceptives in the United States. I focus on her baseline regression, reported in her Table 2 column (1). A replication[11] of her regression (with CRVE_1 standard errors,

[10] This is an example where different individuals are sampled in the two time periods.

[11] Our results are slightly different from those reported in Bailey (2010) for two reasons. First, her replication dataset has 21 fewer observations than the one used in her published paper. Second, Bailey reports average marginal effects from probit

Table 12: Bailey (2010) Table 2, Column (1)
Effect of Sales Ban on Birth Control Use

	Coefficient	Std Err	t	pv	95% interval	K	a
CRVE_1							
Sales Ban	−.055	.020	−2.71	.010	[−.095, −.014]		
Sales Ban×1970	.039	.029	1.37	.177	[−.018, .097]		
Jackknife							
Sales Ban	−.055	.028	−1.98	.046	[−.108, −.001]	7.95	1.19
Sales Ban×1970	.039	.035	1.13	.214	[−.027, .105]	10.1	1.17
Fixed Effects:	Region×Year (8)						
Clusters	State (47)						
Observations	6929						

clustered by state) is reported in the top panel of Table 12. We follow Bailey (2020) and report only two coefficients, that for the indicator for the Sales Ban, and that for its interaction with an indicator for 1970. In addition, the regression includes indicators for states with physician exceptions and its interaction with 1970, as well as census region-by-year fixed effects. Of these estimates, Bailey (2010) paid particular attention to the coefficient on the Sales Ban, which is negative and significant at the 1% level, arguing that this means that "women in states with sales bans were significantly less likely to have used oral contraception before the 1965 *Griswold* decision".

We repeat the estimation in the bottom panel of Table 12 using our jackknife methods. Both standard errors increase significantly; that for the key Sales Ban variable by 40%. Its p-value increases from 1% to 4.6%. This change arises despite the fact that there are a reasonably large ($G = 47$) number of clusters and a very large ($n = 6929$) number of observations. While the jackknife methods do not reverse Bailey's conclusions, they moderate their significance.

It is also useful to examine the reported degree-of-freedom K and scale adjustment a coefficients. In this application the value of K for the two reported coefficients ($K = 8$ and $K = 10$) are moderately small, and lower than the conventional $G - 1 = 46$. The scale adjustments $a = 1.2$ are also moderate. These values indicate that we should expect only minor distributional deviation from conventional.

6.3 MacKinnon and Webb (2020)

Our investigation next follows in the footsteps of MacKinnon and Webb (2020)[12]. We augment the regression of Table 12 with a dummy variable indicating if a state repealed their sales ban in 1961, four years before the *Griswold* decision. There are two such states (Illinois and Colorado). We repeat an analog[13] of their regression in the top panel of Table 13, and then repeat the analysis using our jackknife methods in the bottom panel.

regression, while Table 12, following MacKinnon and Webb (2020), reports linear probability estimates.

[12] Their purpose was to illustrate inference based on randomization methods.

[13] In Table 1 of MacKinnon and Webb (2020) they add two dummy variables rather than just one, interacting the "Repeal in 1961" indicator with year dummies. We do not do so as this regression suffers from poor identification (the coefficients are not identified if Illinois is omitted, as there are no observations for Colorado in 1970.) This is a "one treated cluster" context. While our inference methods are valid in this case, we did not want this to be the focus of this illustration.

Table 13: MacKinnon and Webb (2020), Table 1
Effect of Early Repeal on Birth Control Use

	Coefficient	Std Err	t	pv	95% interval	K	a
CRVE_1							
Sales Ban	−.046	.016	−2.81	.007	[−.079, −.013]		
Sales Ban×1970	.036	.028	1.30	.200	[−.020, .092]		
Repeal in 1961	−.082	.019	−4.23	.000	[−.121, −.043]		
Jackknife							
Sales Ban	−.046	.023	−2.06	.039	[−.090, −.003]	8.29	1.19
Sales Ban×1970	.036	.033	1.09	.230	[−.027, .099]	10.1	1.17
Repeal in 1961	−.082	.106	−0.77	.178	[−.373, .209]	1.02	4.38
Fixed Effects:	Region×Year (8)						
Clusters	State (47)						
Observations	6929						

The results in the top panel indicate that the coefficient on “Repeal in 1961” is negative and statistically significant, with a p-value of 0.000. This appears to suggest the counter-intuitive finding that the early repeal resulted in a lower probability of birth control use. However, if we examine the bottom panel we find that the standard error for “Repeal in 1961” increases fivefold when the jackknife is used, and the reported p-value increases to 0.178. The “significance” of the result disappears.

It is instructive to examine the degree-of-freedom K and scale adjustment a. We see that for the “Repeal in 1961” coefficient, $K = 1$ and $a = 4.4$, which are extreme values. Seeing this, we should investigate the cause, and uncover that this regression coefficient is poorly identified. The value of K indicates that conventional inference will be invalid, and the conventional CRVE_1 t-ratio unreliable. This helps explain why the conventional t-ratio spuriously indicates a “significant” effect.

Our message is that a researcher who uses conventional CRVE_1 methods could easily be misled by regressions such as that in the top panel of Table 13, but will not be as easily misled if they use jackknife methods as presented in the bottom panel. As shown by MacKinnon and Webb (2020), similar inferences can be obtained by randomization methods. An important difference is that the jackknife can be a computationally simple *default* method for calculation of standard errors, p-values, and confidence intervals, not just as a specialized robustness check.

6.4 Rao (2019)

Our third and fourth illustrations are from Rao (2019). He investigates the impact of the integration of poor children into elite private schools on the social behaviors of rich students, using a combination of administrative data and field experiments. His paper reports many regressions; I report two. I start with the first reported in his paper, from column 1 of his Table 2, which measures the effect of integration on whether a rich student volunteers for charity. I repeat his regression in the top panel of Table 14, which reports a linear regression of an indicator for volunteering on treatment (the presence of poor children in a student’s classroom), four demographic controls, and school and grade fixed effects. Clustering is

Table 14: Rao (2019), Table 2, Column 1
Effect of Integration on Volunteering for Charity

	Coefficient	Std Err	t	pv	95% interval	K	a
CRVE$_1$							
Treated classroom	.130	.026	5.05	.000	[.079, .182]		
Age	.029	.035	0.84	.407	[−.041, .010]		
Male	.010	.018	0.56	.576	[−.026, .046]		
Family Owns Car	.038	.026	1.47	.146	[−.014, .100]		
Private Driver	.015	.025	0.61	.541	[−.034, .065]		
Jackknife							
Treated classroom	.130	.038	3.43	.000	[.066, .195]	20.1	1.23
Age	.029	.036	0.82	.407	[−.041, .010]	58.8	1.01
Male	.010	.018	0.55	.577	[−.026, .046]	61.1	1.02
Family Owns Car	.038	.026	1.45	.146	[−.014, .091]	48.9	1.02
Private Driver	.015	.025	0.61	.539	[−.034, .065]	56.6	1.01
Fixed Effects:	School (17), Grade (4)						
Clusters	School×Grade (68)						
Observations	2364						

done at the school-by-grade level, so there are $G = 68$ clusters and $n = 2304$ observations. The coefficient of interest is that for treatment.

We repeat the analysis using our jackknife methods in the bottom panel. The standard error on treatment increases by 46%, while the standard errors on the other estimates do not change. The p-value for treatment in both regressions is highly significant, so the conclusion that integration affects behavior is not altered, but the fact that the standard error increases by nearly 50% illustrates how conventional inference is potentially fragile.

As a second example I take Rao's regression reported in column 2 of his Table 6, which measures the effect of integration on a discriminatory behavior (choosing a lower-ability wealthy student over a higher-ability poor student as a teammate in an athletic contest). In this regression, in addition to the primary treatment indicator there are four other coefficients of interest (two indicators of higher prize money, and interactions of these indicators with the treatment indicator) as well as school and grade fixed effects. In this example there are $G = 8$ clusters and $n = 342$ observations.

We repeat Rao's results in the top panel of Table 15 and present the jackknife results in the bottom panel. Rao's results appear to show that treatment has a significant negative effect on discriminatory behavior, and so does the offer of higher prize money. The jackknife results, however, moderate these inferences. The standard error on treatment triples, and its p-value increases from 0.006 to 0.121. The impact of integration no longer appears to have a statistically significant impact on behavior. The standard errors and p-values for the prize levels, in contrast, increase more moderately.

Again it is useful to examine the degree-of-freedom coefficients K. In Table 14 they are all very large ($K \geq 20$ for all coefficients), raising no concerns. However, in Table 15 the values of K for the treatment coefficients (especially the interaction effects) are very low. This should be taken as a signal that conven-

Table 15: Rao (2019), Table 6, Column 2
Effect of Integration on Discriminatory Behavior

	Coefficient	Std Err	t	pv	95% interval	K	a
CRVE$_1$							
Treated classroom	−.256	.065	−3.91	.006	[−.411, −.101]		
Prize = Rs 200	−.137	.054	−2.54	.039	[−.265, −.009]		
Prize = Rs 500	−.314	.050	−6.32	.000	[−.432, −.197]		
Treated×Prize=200	.085	.067	1.28	.242	[−.072, .243]		
Treated×Prize=500	.186	.094	1.99	.087	[−.035, .408]		
Jackknife							
Treated classroom	−.256	.194	−1.32	.121	[−.655, .143]	2.42	1.78
Prize = Rs 200	−.137	.061	−2.26	.056	[−.279, .005]	4.98	1.10
Prize = Rs 500	−.314	.055	−5.69	.002	[−.445, −.184]	4.81	1.10
Treated×Prize=200	.085	.094	0.90	.377	[−.299, .470]	1.63	1.32
Treated×Prize=500	.186	.157	1.19	.280	[−.427, .800]	1.69	1.32
Fixed Effects:	School (2), Grade (4)						
Clusters	School×Grade (8)						
Observations	342						

tional inference methods are misleading.

My view is that if results such as the bottom panel of Table 13 were routinely displayed, rather than the results from the top panel, researchers would make more informed decisions.

7 Conclusion

Difference-in-difference regression is a standard tool in contemporary economics. The vast majority of applications report cluster-robust standard errors, but the conventional formula produces estimates which can be highly biased towards zero, resulting in spurious levels of statistical significance. Two simple changes can alleviate this problem: the use of jackknife variance estimation, and adjusted student t critical values. These alternatives are computationally efficient, and could be set for default use.

A Stata and R program `jregress` which calculates the recommended methods is available on the author's website `users.ssc.wisc.edu/~bhansen/`.

8 Appendix

8.1 Heteroskedasticity-Robust Covariance Estimators

For reference, we list the common heteroskedasticity-robust covariance matrix estimators for the linear model $Y_i = \widehat{\beta}' X_i + \widehat{e}_i$ under assumed cross-sectional independence, n observations, and k regressors.

The HC$_0$ estimator of Eicker (1963), Huber (1967), and White (1980) is

$$\widehat{\boldsymbol{V}}_0 = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\sum_{i=1}^{n} X_i X_i' \widehat{e}_i^2\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}. \tag{15}$$

The HC$_1$ estimator of Hinkley (1977) is

$$\widehat{\boldsymbol{V}}_1 = \frac{n}{n-k}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\sum_{i=1}^{n} X_i X_i' \widehat{e}_i^2\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}. \tag{16}$$

The HC$_2$ estimator of MacKinnon and White (1985) is

$$\widehat{\boldsymbol{V}}_2 = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\sum_{i=1}^{n} X_i X_i' \frac{\widehat{e}_i^2}{1-h_i}\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}, \tag{17}$$

where $h_i = X_i'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} X_i$. The HC$_3$/jackknife estimator of MacKinnon and White (1985) is

$$\begin{aligned}\widehat{\boldsymbol{V}}_3 &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\sum_{i=1}^{n} X_i X_i' \frac{\widehat{e}_i^2}{(1-h_i)^2}\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} && (18)\\ &= \sum_{i=1}^{n}\left(\widehat{\beta}_{-i}-\widehat{\beta}\right)\left(\widehat{\beta}_{-i}-\widehat{\beta}\right)', && (19)\end{aligned}$$

where $\widehat{\beta}_{-i} = \left(\boldsymbol{X}'\boldsymbol{X} - X_i X_i'\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{Y} - X_i Y_i\right)$ is the leave-one-out estimator of β.

8.2 Adjusted Jackknife Inference Formula

The following formula for the constants K and a for the confidence interval (13) and p-value (14) are taken from Hansen (2024):

$$a = \sqrt{\frac{\operatorname{tr}[\boldsymbol{L}]}{R'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}R}}, \tag{20}$$

and

$$K = \frac{(\operatorname{tr}[\boldsymbol{L}])^2}{\operatorname{tr}[\boldsymbol{L}\boldsymbol{L}]}, \tag{21}$$

with

$$\operatorname{tr}[\boldsymbol{L}] = \sum_{g=1}^{G} S_g - \operatorname{tr}\left[\boldsymbol{U}'\boldsymbol{V}\right],$$

and

$$\begin{aligned}\operatorname{tr}[\boldsymbol{L}\boldsymbol{L}] = \sum_{g=1}^{G} S_g^2 &+ \operatorname{tr}\left[\boldsymbol{X}'\boldsymbol{X}\boldsymbol{U}'\boldsymbol{U}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{U}'\boldsymbol{U}\right] + 2\operatorname{tr}\left[\boldsymbol{V}'\boldsymbol{U}\boldsymbol{V}'\boldsymbol{U}\right]\\ &- 2\operatorname{tr}\left[\boldsymbol{V}'\boldsymbol{W}\right] - 4\operatorname{tr}\left[\boldsymbol{U}'\boldsymbol{U}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{U}'\boldsymbol{V}\right] + 2\operatorname{tr}\left[\boldsymbol{U}'\boldsymbol{U}\boldsymbol{V}'\boldsymbol{V}\right],\end{aligned}$$

where

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

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

8.3 Computational Considerations

Calculation of both the variance estimator (8) and the correction coefficients (20)-(21) requires looping over clusters, and the latter also require looping over individual coefficients (in order to calculate standard errors for each coefficient estimate). The major computational burden in each loop is the generalized inverse $\left(\boldsymbol{X}'\boldsymbol{X} - \boldsymbol{X}_g'\boldsymbol{X}_g\right)^+$. It is efficient if each of these is calculated just once and the inverse matrices stored.

In our R program we use the following method to compute the Moore-Penrose inverse. First, calculate the eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0$ and associated eigenvalues $h_1, ..., h_n$ of $\boldsymbol{X}'\boldsymbol{X} - \boldsymbol{X}_g'\boldsymbol{X}_g$, which satisfy the spectral decomposition $\boldsymbol{X}'\boldsymbol{X} - \boldsymbol{X}_g'\boldsymbol{X}_g = \boldsymbol{H}\boldsymbol{\Lambda}\boldsymbol{H}'$ where $\boldsymbol{\Lambda} = \text{diag}\{\lambda_1, ..., \lambda_n\}$ and $\boldsymbol{H} = [h_1, ..., h_n]$. For some threshold $\epsilon > 0$ (close to machine zero) calculate the trimmed eigenvalue inverses $\lambda_j^+ = \lambda_j^{-1}\mathbf{1}\{\lambda_j \geq \epsilon\}$ and set $\boldsymbol{\Lambda}^+ = \text{diag}\{\lambda_1^+, ..., \lambda_n^+\}$. The numerical Moore-Penrose inverse is then found as $\left(\boldsymbol{X}'\boldsymbol{X} - \boldsymbol{X}_g'\boldsymbol{X}_g\right)^+ = \boldsymbol{H}\boldsymbol{\Lambda}^+\boldsymbol{H}'$. For small to moderate k this is computationally reasonable. However, as the dimension k increases the eigenvalue calculation becomes computationally burdensome. Consequently, for computation with for very large k a faster implementation of the Moore-Penrose inverse would be desirable.

In the standard DiD regression (1), it is common that the regression includes a large number of group-level fixed effects dummies. When these fixed effects correspond to the level of clustering, these regressors can be eliminated by application of the within transformation to all regressors. Whether the full regression is estimated or the regression after the within transformation is applied, the remaining regression coefficient estimates, jackknife covariance matrix estimator, and correction coefficients K and a are all identical. This can dramatically reduce the number of effective regressors k, and this reduces the computation time. Therefore, if the fixed effect coefficients themselves are not of interest, it is computationally advised to first eliminate the fixed effect dummy variables by applying the within transformation. However, if the fixed effects are different than the level of clustering (which is true, for example, in the empirical examples of this paper), then this equivalence is not valid, and estimation and inference should be done by explicit inclusion of all fixed effects using dummy variables as regressors.

To investigate computation cost of our proposed jackknife methods and our specific `jregress` programs we report computation times on randomly generated data sets. All variables (the dependent vari-

Table 16: Computation Time (Seconds)

			R 4.4.1		Stata SE 18	
G	n_g	k	CRVE$_1$	Jackknife	CRVE$_1$	Jackknife
20	100	10	0.006	0.096	0.005	0.024
20	100	50	0.028	0.029	0.011	0.111
20	100	100	0.014	0.173	0.020	0.361
20	100	200	0.036	1.394	0.067	1.954
20	1000	10	0.023	0.013	0.018	0.036
20	1000	50	0.047	0.1000	0.038	0.222
20	1000	100	0.120	0.316	0.087	0.650
20	1000	200	0.375	1.782	0.273	2.912
200	100	10	0.019	0.031	0.011	0.043
200	100	50	0.040	0.216	0.046	0.392
200	100	100	0.135	1.053	0.090	1.700
200	100	200	0.320	6.591	0.268	10.59
200	1000	10	0.280	0.225	0.116	0.246
200	1000	50	0.646	0.823	0.330	1.361
200	1000	100	1.527	2.543	0.702	4.690
200	1000	200	4.574	11.23	2.101	20.15

Computation performed under Windows 11 on an i7-12700 processor with 32 GB of RAM.

able and k regressors) were generated as i.i.d. $N(0,1)$, and the observations organized into G clusters each with n_g observations. We vary $n_g \in \{100,1000\}$, $G = \{20,200\}$, and $k = \{10,50,100,200\}$. Notice that the total number of observations range among $n = \{2000,20000,200000\}$, so these computations are for large to very large samples. We do calculations in both R (version 4.4.1) and Stata SE 18, on a standard office PC. In R, the computation of CRVE$_1$ is done by the `lm_robust` application from the `estimatr` package. The computation of the jackknife is done with our `jregress` program. In Stata, the computation of CRVE$_1$ is done by `regress` with the `cluster` option, and the jackknife is done with our `jregress` program. The Stata calculations are done `quietly` to emphasize calculation rather than screen display. All calculate the least squares estimates, standard errors, p-values, and confidence intervals (with adjustments for the jackknife method). We perform each calculation once for each configuration.

In Table 16 we report the elapsed computation time in seconds. We make the following general observations:

1. In all contexts the jackknife calculation times are reasonable for default implementaion. In most cases, computation time is a fraction of a second. In models with a large number of observations and regressors, jackknife computation time can take multiple seconds, but this is also the case for CRVE$_1$ estimation.

2. While the jackknife is generally more computationally costly than CRVE$_1$, it is not uniformly more costly, and in most cases the differences are minor.

3. Computation speeds are generally similar between the R and Stata packages. Stata, however, has

a faster default implementation of CRVE_1 than the R `lm_robust` package; and our R `jregress` package is somewhat faster than our Stata `jregress` package.

4. Computation cost is increasing in k (the number of regressors), G (the number of clusters), and n_g (the number of observations per cluster). To get a rough understanding of these impacts, we fitted regressions of log computation time on log inputs, and found that, roughly, the computational cost of the jackknife methods is $O\left(G^{0.7} n_g^{0.3} k^{1.3}\right)$. Thus, computation time is most strongly affected by the number of regressors k, and secondly by the number of clusters G.

Overall, the calculations demonstrate that the proposed jackknife methods are computationally reasonable to implement on standard office computers, at least for data sets up to 200,000 observations and 200 regressors.

References

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

[2] Bailey, Martha J. (2010): "Momma's got the pill: How Anthony Comstock and *Griswold v. Connecticut* shaped U.S. childbearing," *American Economic Review*, 100, 98-129.

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

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

[5] Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan (2004): "How much should we trust difference-in-differences estimates?" *Quarterly Journal of Economics*, 119, 249-275.

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

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

[8] Card, David and Alan Krueger (1994): "Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania," *American Economic Review*, 84, 772-793.

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

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

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

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

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

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

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

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

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

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

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

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

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

[22] Hansen, Bruce E. (2024) "Jackknife standard errors for clustered regression," working paper.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[39] Rao, Gautam (2019): "Familiarity does not breed contempt: Generosity, discrimination, and diversity in Delhi schools," *American Economic Review*, 109, 774-809.

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

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

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

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

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

[45] Young, Alwyn (2016): "Improved, nearly exact, statistical inference with robust and clustered covariance matrices using effective degrees of freedom corrections," working paper.

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