Unbiased estimation of the OLS covariance matrix when the errors are clustered

Boot, Tom; Wansbeek, Thomas; Niccodemi, Gianmaria

Published in:
Empirical Economics

DOI:
10.1007/s00181-023-02379-w

Publication date:
2023

Empirical Economics (2023) 64:2511–2533
https://doi.org/10.1007/s00181-023-02379-w

Unbiased estimation of the OLS covariance matrix when the errors are clustered

Tom Boot[1] · **Gianmaria Niccodemi**[1] · **Tom Wansbeek**[1]

Received: 15 June 2022 / Accepted: 25 January 2023 / Published online: 17 March 2023

Abstract
When data are clustered, common practice has become to do OLS and use an estimator of the covariance matrix of the OLS estimator that comes close to unbiasedness. In this paper, we derive an estimator that is unbiased when the random-effects model holds. We do the same for two more general structures. We study the usefulness of these estimators against others by simulation, the size of the t-test being the criterion. Our findings suggest that the choice of estimator hardly matters when the regressor has the same distribution over the clusters. But when the regressor is a cluster-specific treatment variable, the choice does matter and the unbiased estimator we propose for the random-effects model shows excellent performance, even when the clusters are highly unbalanced.

Keywords Clustered errors · Degrees-of-freedom correction · Placebo regression · Treatment effect · Unbiased estimator

1 Introduction

Within-cluster dependence presents a considerable challenge for reliable inference. Even with large data sets, a small number of clusters induces substantial finite-sample bias in the estimated variance of the regression coefficients. Several options are available to mitigate this bias. Stata uses a scalar correction to the Liang and Zeger (1986) cluster-robust variance estimator, while Bell and McCaffrey (2002) develop cluster extensions of the MacKinnon and White (1985) heteroskedasticity-robust variance estimators. See Cameron and Miller (2015) and MacKinnon et al. (2023) for recent

The authors are grateful for the incisive and most useful comments from two referees. Tom Boot acknowledges financial support from the Dutch Research Council (NWO) under research grant No. 201E.011.

✉ Tom Wansbeek
t.j.wansbeek@rug.nl

[1] University of Groningen, Groningen, The Netherlands

surveys on the topic. However, with the exception of some special cases, none of these variance adjustments completely eliminates the bias.

In this paper, we develop variance estimators that are unbiased under progressively more complicated dependence structures. Our aim is to investigate whether removing the bias in the variance estimators leads to improved inference, in particular by delivering hypothesis tests with more accurate size control. The key idea underlying the unbiased variance estimator is a cluster extension of the variance estimator by Hartley et al. (1969), which is unbiased under heteroskedasicity. In its original form, this variance estimator has the drawback that it requires inverting a matrix that grows quadratically with the sample size. We show how the underlying structure of this matrix can be exploited to make the computation feasible even with large microeconometric data sets.

With a large number of clusters, test statistics based on cluster-robust variance estimators have a standard normal distribution, see for instance Hansen and Lee (2019). With a small number of clusters, the use of the normal distribution to obtain confidence intervals and critical values can lead to substantial size distortions as discussed in Cameron and Miller (2015), Sect. VI.D, unless the within-cluster dependence is restricted as in Ibragimov and Müller (2016). The use of a t-distribution reduces the size distortion, but this requires selecting the appropriate degrees of freedom (d.f.). For our proposed variance estimators, we derive a data-driven estimator for the d.f. following the approach based on an independence assumption on the errors as in Bell and McCaffrey (2002), as well as the generalization to a random-effects structure studied in Imbens and Kolesár (2016).

We focus on three dependence structures of increasing generality. First, we assume that in each cluster, the errors follow the same random-effects structure. In this case, the covariance structure depends on two (unknown) parameters. Second, we extend this setting by allowing the RE parameters to be cluster dependent, increasing the number of parameters to two times the number of clusters. Finally, we consider a fully unrestricted setting where each cluster has an arbitrary covariance matrix. This captures for example a setting with conditional heteroskedasticity where the covariance matrix depends via an unknown functional form on a set of continuous regressors. In practice, leaving the correlation structure completely undetermined is generally preferred. However, tighter parametrizations can be useful to reduce estimation uncertainty and improve the behavior of tests in settings where the number of clusters is small and the parametrization is only mildly misspecified.

As said, the first two structures contain random effects, and one might consider treating the effects as fixed, that is, adding cluster-level fixed effects to the model. This will greatly reduce the intra-cluster correlation and might be considered a simple alternative. However, the drawbacks outweigh the benefits. Just as in the closely connected case of panel data analysis, fixed effects spawn the within-transformation, which often eliminates most variation in the data while wiping out regressors that are the same for all observations in a cluster, like a cluster-specific treatment dummy, a case of great empirical relevance. At the same time, the main advantage of fixed over random effects in a panel data context, controlling for endogeneity, is not a particular issue in the current context.

For each of the three dependence structures, we numerically evaluate the size properties of hypothesis tests based on the unbiased variance estimators. We compare their performance with the default Stata option, as well as the HC2 variance estimator by Bell and McCaffrey (2002) with d.f. as in Imbens and Kolesár (2016). The model we consider includes a treatment dummy and a continuous variable. For each covariance structure, we vary the number of treated clusters and consider both a balanced design, where each cluster has the same number of observations, as well as an unbalanced design.

Under the specification where the random-effects structure is the same across clusters, we find that the corresponding unbiased variance estimator performs remarkably well. Even with only a single treated cluster, hypothesis tests provide accurate size control on both the treatment dummy and the continuous variable. When the number of observations differs between clusters, we find that the d.f. calculated under the more general RE assumption improve substantially over those calculated under independence assumptions. In a more general setting where the RE structure is cluster dependent, we find that using the corresponding variance estimator improves over the benchmarks particularly when the design is unbalanced. Finally, we consider a setting where there is conditional heteroskedasticity that depends on the continuous variable. The most general unbiased variance estimator continues to control size in this setup.

After these simulations with fully artificial data, we compare methods using real-life data with an artificial element added. That is, we estimate a wage equation on the basis of US data, clustered by state. To the real-life data, we added an artificial state-wide policy dummy variable. We study the size of an hypothesis test on the effect of this dummy variable by sampling subsets of states either at random or based on their number of observations.

Finally, we remark that we develop our variance estimators in what Abadie et al. (2020) refer to as a sampling-based framework, where we condition on the available regressors and the cluster structure is determined by the covariance structure of the regression errors. When the regressors are random as in a design-based framework, the relevant cluster structure is instead determined by the clustering in both the regressors and the regression errors. For instance, when the regressor is a treatment dummy that is randomized at the unit level, there is no need to account for clustering at all. From this perspective, we expect that in a design-based framework, the tighter cluster parametrizations we propose can be useful when these correspond to the cluster structure in the assignment mechanism for the treatment dummy.

The paper is organized as follows. In Sect. 2, we start by deriving the general form of unbiased estimators for error covariance matrices with a linear structure. We then specify this for clusters in Sect. 3. We first consider in Sect. 3.1 a simple structure with just two parameters, one for the overall error and one for the within-cluster error. In Sect. 3.2, we generalize this and make these parameters specific per cluster. In Sect. 3.3, we generalize this one more step and allow all covariances within clusters to vary freely. We proceed to compare the performance of the various unbiased variance estimators, first by simulation and then through an application to real-life data. Our performance measure is the size of the t-test. The d.f. of the t-tests play an important role, and in Sect. 4, we discuss how we set them. Section 5 describes the setup of the

simulations, while the results are presented in Sect. 6. The results for the real-life data are given in Sect. 7. Section 8 concludes.

Most derivations are given in "Online Appendix A, B and C," contained in the Supplementary Information available online. The MATLAB code for the computations reported in this paper is available from https://sites.google.com/view/tomboot/.

2 Unbiased variance estimation

We consider the linear regression model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, with $\mathbf{X}$ exogenous of order $n \times k$. We follow the usual notation $\mathbf{M} \equiv \mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ and $\mathbf{P} \equiv \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. The errors are distributed according to $\boldsymbol{\varepsilon} \sim (\mathbf{0}, \boldsymbol{\Sigma})$ and we consider the case where $\boldsymbol{\Sigma}$ is linear in parameters,

$$\mathrm{vec}\boldsymbol{\Sigma} = \mathbf{D}\boldsymbol{\pi},$$

with $\boldsymbol{\pi}$ of order $r \times 1$ and the design matrix $\mathbf{D}$ of order $n^2 \times r$. We are interested in unbiased estimation of the covariance matrix $\mathbf{V}$ of the OLS estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$,

$$\mathbf{V} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}.$$

As will become clear below, our analyses involving $\mathbf{V}$ require us to consider it in stacked form, $\mathbf{v} \equiv \mathrm{vec}\mathbf{V}$. With

$$\mathbf{R}' \equiv \left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \otimes (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right)\mathbf{D},$$

we have in stacked form

$$\begin{aligned}\mathbf{v} &= \left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \otimes (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right)\mathrm{vec}\boldsymbol{\Sigma} \\ &= \mathbf{R}'\boldsymbol{\pi}.\end{aligned}$$

We base our estimator on a function of the residuals $\hat{\boldsymbol{\varepsilon}} \equiv \mathbf{M}\boldsymbol{\varepsilon}$ that is aligned with the structure of $\boldsymbol{\Sigma}$. We hence project the squared residuals on the space spanned by $\mathbf{D}$, so we use $\mathbf{D}(\mathbf{D}'\mathbf{D})^{-1}\mathbf{D}'(\hat{\boldsymbol{\varepsilon}} \otimes \hat{\boldsymbol{\varepsilon}})$, leading to the estimator

$$\begin{aligned}\tilde{\mathbf{v}} &= \left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \otimes (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right)\mathbf{D}(\mathbf{D}'\mathbf{D})^{-1}\mathbf{D}'(\hat{\boldsymbol{\varepsilon}} \otimes \hat{\boldsymbol{\varepsilon}}) \\ &= \mathbf{R}'(\mathbf{D}'\mathbf{D})^{-1}\mathbf{D}'(\hat{\boldsymbol{\varepsilon}} \otimes \hat{\boldsymbol{\varepsilon}}). \qquad (1)\end{aligned}$$

However, this estimator is biased; with $\mathrm{E}\left(\mathbf{D}'(\hat{\boldsymbol{\varepsilon}} \otimes \hat{\boldsymbol{\varepsilon}})\right) = \mathbf{D}'(\mathbf{M} \otimes \mathbf{M})\mathbf{D}\boldsymbol{\pi}$ there holds

$$\mathrm{E}(\tilde{\mathbf{v}}) = \mathbf{R}'(\mathbf{D}'\mathbf{D})^{-1}[\mathbf{D}'(\mathbf{M} \otimes \mathbf{M})\mathbf{D}]\boldsymbol{\pi} \neq \mathbf{R}'\boldsymbol{\pi} = \mathbf{v}.$$

The bias is easily removed by replacing the term $(\mathbf{D}'\mathbf{D})^{-1}$ by $[\mathbf{D}'(\mathbf{M} \otimes \mathbf{M})\mathbf{D}]^{-1}$. For the special case of heteroskedasticity, this idea is due to Hartley et al. (1969). The adapted, unbiased estimator of $\mathbf{v}$ then is

$$\hat{\mathbf{v}} \equiv \mathbf{R}'[\mathbf{D}'(\mathbf{M} \otimes \mathbf{M})\mathbf{D}]^{-1}\mathbf{D}'(\hat{\boldsymbol{\varepsilon}} \otimes \hat{\boldsymbol{\varepsilon}}). \tag{2}$$

For computational purposes, (2) is unattractive as the matrix $\mathbf{M} \otimes \mathbf{M}$ is huge with large data sets. However, we show below how the simple structure of $\mathbf{M}$, being the sum of the unit matrix and a matrix of low rank, can be exploited to avoid computational difficulties. A relatively common issue with unbiased estimation of variance components, see for instance Kline et al. (2020), is that the estimator is not guaranteed to be positive definite. However, corrections that make the estimator positively biased are readily available and avoid overrejection.

Below we will consider three cases, with different design matrices $\mathbf{D}$. In the third case, the number of columns of $\mathbf{D}$ can be very large. Then, we can use an adapted version of (2). Let

$$\begin{aligned}\mathbf{A} &\equiv \mathbf{D}'\mathbf{D} - \mathbf{D}'(\mathbf{I}_n \otimes \mathbf{P})\mathbf{D} - \mathbf{D}'(\mathbf{P} \otimes \mathbf{I}_n)\mathbf{D}\\ \mathbf{W} &\equiv \mathbf{X}'\mathbf{X} \otimes \mathbf{X}'\mathbf{X}\\ \mathbf{F} &\equiv \mathbf{D}'(\mathbf{X} \otimes \mathbf{X}).\end{aligned}$$

Then,

$$\begin{aligned}\mathbf{R}' &= \mathbf{W}^{-1}\mathbf{F}'\\ \mathbf{D}'\big(\mathbf{P} \otimes \mathbf{P}\big)\mathbf{D} &= \mathbf{F}\mathbf{W}^{-1}\mathbf{F}'\\ \mathbf{D}'\big(\mathbf{M} \otimes \mathbf{M}\big)\mathbf{D} &= \mathbf{D}'\mathbf{D} - \mathbf{D}'\big(\mathbf{I}_n \otimes \mathbf{P}\big)\mathbf{D} - \mathbf{D}'\big(\mathbf{P} \otimes \mathbf{I}_n\big)\mathbf{D} + \mathbf{D}'\big(\mathbf{P} \otimes \mathbf{P}\big)\mathbf{D}\\ &= \mathbf{A} + \mathbf{F}\mathbf{W}^{-1}\mathbf{F}'\end{aligned}$$

Since

$$\big(\mathbf{W} + \mathbf{F}'\mathbf{A}^{-1}\mathbf{F}\big)\mathbf{W}^{-1}\mathbf{F}' = \mathbf{F}'\mathbf{A}^{-1}\big(\mathbf{A} + \mathbf{F}\mathbf{W}^{-1}\mathbf{F}'\big)$$

there holds

$$\mathbf{W}^{-1}\mathbf{F}'\big(\mathbf{A} + \mathbf{F}\mathbf{W}^{-1}\mathbf{F}'\big)^{-1} = \big(\mathbf{W} + \mathbf{F}'\mathbf{A}^{-1}\mathbf{F}\big)^{-1}\mathbf{F}'\mathbf{A}^{-1}.$$

Substitution in (2) yields

$$\begin{aligned}\hat{\mathbf{v}} &= \mathbf{W}^{-1}\mathbf{F}'\big(\mathbf{A} + \mathbf{F}\mathbf{W}^{-1}\mathbf{F}'\big)^{-1}\mathbf{D}'\big(\hat{\boldsymbol{\varepsilon}} \otimes \hat{\boldsymbol{\varepsilon}}\big)\\ &= \big(\mathbf{W} + \mathbf{F}'\mathbf{A}^{-1}\mathbf{F}\big)^{-1}\mathbf{F}'\mathbf{A}^{-1}\mathbf{D}'\big(\hat{\boldsymbol{\varepsilon}} \otimes \hat{\boldsymbol{\varepsilon}}\big).\end{aligned} \tag{3}$$

This expression still contains the inverse of the matrix $\mathbf{A}$, which has the same number of columns as $\mathbf{D}$. It will appear, though, that $\mathbf{A}^{-1}$ occurs only in the form $\mathbf{F}'\mathbf{A}^{-1}$, which appears to have a simple expression in this third case.

We now turn to the cluster structure. We denote the number of clusters by C and index them by $c = 1, \ldots, C$. Cluster c has n_c observations, so $\sum_c n_c = n$. We let

$$\begin{aligned}\ddot{n} &\equiv \sum_c n_c^2\\ \boldsymbol{\Delta}_n &\equiv \operatorname{diag} n_c.\end{aligned}$$

Let $\mathbf{i}_c$ an n_c-vector of ones. With a slight abuse of notation, we will write $\mathbf{I}_c$ for $\mathbf{I}_{n_c}$ and let

$$\mathbf{G}_c \equiv \begin{pmatrix} \mathbf{O} \\ \vdots \\ \mathbf{I}_c \\ \vdots \\ \mathbf{O} \end{pmatrix} \qquad \mathbf{b}_c \equiv \begin{pmatrix} \mathbf{0} \\ \vdots \\ \mathbf{i}_c \\ \vdots \\ \mathbf{0} \end{pmatrix} \qquad \mathbf{B} \equiv (\mathbf{b}_1, \ldots, \mathbf{b}_c, \ldots, \mathbf{b}_C). \tag{4}$$

The regressors for cluster c are collected in $\mathbf{X}_c \equiv \mathbf{G}_c'\mathbf{X}$ and their sum over the cluster in the row vector $\tilde{\mathbf{x}}_c' \equiv \mathbf{b}_c'\mathbf{X}$. The $\tilde{\mathbf{x}}_c'$s are collected in the $C \times k$ matrix $\tilde{\mathbf{X}} \equiv \mathbf{B}'\mathbf{X}$. Likewise, $\hat{\boldsymbol{\varepsilon}}_c \equiv \mathbf{G}_c'\hat{\boldsymbol{\varepsilon}}$ and $\tilde{\hat{\boldsymbol{\varepsilon}}}_c \equiv \mathbf{b}_c'\hat{\boldsymbol{\varepsilon}}$ so $\tilde{\hat{\boldsymbol{\varepsilon}}} = \mathbf{B}'\hat{\boldsymbol{\varepsilon}}$.

Below we will frequently perform matrix operations using

$$\begin{aligned} \text{vec}(\mathbf{ABC}) &= (\mathbf{C}' \otimes \mathbf{A})\text{vec}\mathbf{B} \\ \text{tr}(\mathbf{ABCD}) &= \text{vec}(\mathbf{A}')'(\mathbf{D}' \otimes \mathbf{B})\text{vec}\mathbf{C}, \end{aligned}$$

for conformable generic $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ and $\mathbf{D}$. A piece of notation that is useful in the third case that we will study is the Kronecker product with a dot on top. With $\mathbf{e}_c$ be the cth unit vector, we write

$$\sum_c \mathbf{e}_c' \dot{\otimes} \mathbf{A}_c = (\mathbf{A}_1, \ldots, \mathbf{A}_C)$$

for matrices $\mathbf{A}_1, \ldots, \mathbf{A}_C$ with the same number of rows but possibly different number of columns. The use of $\dot{\otimes}$ is as straightforward as the use of $\otimes$.

3 Application to three forms of clustering

In this section, we consider three, increasingly general structures for $\boldsymbol{\Sigma}$ and present the variance estimator (2) for each case. The results are in stacked form, $\hat{\mathbf{v}}$. We also present the simpler expressions that are obtained when $\boldsymbol{\varepsilon}$ would be observable and $\hat{\boldsymbol{\varepsilon}}$ is substituted for $\boldsymbol{\varepsilon}$ afterward, that is, the results that we obtain when we neglect the presence of the regressors. These simpler expressions can be put in the usual, "unstacked" form, that is, as $\hat{\mathbf{V}}$ rather than $\hat{\mathbf{v}}$. Derivations are relegated to "Online Appendix A".

3.1 Equicorrelated errors

We first consider the case where the errors are equicorrelated within clusters, so

$$\boldsymbol{\Sigma} = \sigma^2\mathbf{I}_n + \tau^2\mathbf{B}\mathbf{B}',$$

with $\mathbf{B}$ as given in (4). Let

$$\boldsymbol{\Psi} = \begin{pmatrix} n-k & n-s \\ n-s & \ddot{n} - 2\breve{s} + \dot{s} \end{pmatrix},$$

with

$$\begin{aligned} s &\equiv \operatorname{tr}(\mathbf{X}'\mathbf{X})^{-1}\tilde{\mathbf{X}}'\tilde{\mathbf{X}} \\ \dot{s} &\equiv \operatorname{tr}(\mathbf{X}'\mathbf{X})^{-1}\tilde{\mathbf{X}}'\tilde{\mathbf{X}}(\mathbf{X}'\mathbf{X})^{-1}\tilde{\mathbf{X}}'\tilde{\mathbf{X}} \\ \breve{s} &\equiv \operatorname{tr}(\mathbf{X}'\mathbf{X})^{-1}\tilde{\mathbf{X}}'\boldsymbol{\Delta}_n\tilde{\mathbf{X}} \end{aligned}$$

Then,

$$\hat{\mathbf{v}} = (\mathbf{X}'\mathbf{X} \otimes \mathbf{X}'\mathbf{X})^{-1} \left(\operatorname{vec} \mathbf{X}'\mathbf{X}, \operatorname{vec} \tilde{\mathbf{X}}'\tilde{\mathbf{X}}\right) \boldsymbol{\Psi}^{-1} (\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}, \tilde{\hat{\boldsymbol{\varepsilon}}}'\tilde{\hat{\boldsymbol{\varepsilon}}})' \tag{5}$$

is an unbiased estimator of $\mathbf{v}$.

Two remarks are in order here. The first one concerns symmetry. The $k \times k$ covariance matrix $\hat{\mathbf{V}}$, obtained by rearranging $\hat{\mathbf{v}}$ into a matrix, should be symmetric. The derivation of (5) did not take this requirement into consideration. However, it is easy to show that $\hat{\mathbf{V}}$ is symmetric, by employing the commutation matrix $\mathbf{K}_k$, with properties $\mathbf{K}_k(\mathbf{A} \otimes \mathbf{B}) = (\mathbf{B} \otimes \mathbf{A})\mathbf{K}_k$ for any $k \times k$ matrices $\mathbf{A}$ and $\mathbf{B}$ and $\mathbf{K}_k \operatorname{vec}\mathbf{C} = \operatorname{vec}\mathbf{C}$ for any symmetric $k \times k$ matrix $\mathbf{C}$. Symmetry of $\hat{\mathbf{V}}$ is equivalent to $\mathbf{K}_k \operatorname{vec}\hat{\mathbf{v}} = \operatorname{vec}\hat{\mathbf{v}}$. By using $\mathbf{K}_k = \mathbf{K}_k^{-1}$, this readily follows. The same holds for the other two variance estimators derived below.

The second remark concerns the role played by the regressors. When they would have been neglected in the derivation, that is, estimating $\mathbf{v}$ by (1) instead of by (2), we would have obtained

$$\boldsymbol{\Psi} = \begin{pmatrix} n & n \\ n & \ddot{n} \end{pmatrix} \quad \text{so} \quad \boldsymbol{\Psi}^{-1} = \frac{1}{n(\ddot{n} - n)} \begin{pmatrix} \ddot{n} & -n \\ -n & n \end{pmatrix}. \tag{6}$$

We can then write

$$\hat{\mathbf{v}} = (\mathbf{X}'\mathbf{X} \otimes \mathbf{X}'\mathbf{X})^{-1} \left(\operatorname{vec} \mathbf{X}'\mathbf{X}, \operatorname{vec} \tilde{\mathbf{X}}'\tilde{\mathbf{X}}\right) (\hat{\sigma}^2, \hat{\tau}^2)'$$

or

$$\hat{\mathbf{V}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\hat{\boldsymbol{\Sigma}}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}, \tag{7}$$

with $\hat{\boldsymbol{\Sigma}} = \hat{\sigma}^2\mathbf{I}_n + \hat{\tau}^2\mathbf{B}\mathbf{B}'$, where

$$\hat{\sigma}^2 = \frac{1}{n}\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} - \hat{\tau}^2 \tag{8}$$

$$\hat{\tau}^2 = \frac{1}{\ddot{n} - n}(\tilde{\hat{\boldsymbol{\varepsilon}}}'\tilde{\hat{\boldsymbol{\varepsilon}}} - \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}). \tag{9}$$

In this form, $\hat{\mathbf{\Sigma}}$ is the estimator for $\mathbf{\Sigma}$ used by Imbens and Kolesár (2016) in their d.f. derivation, to be discussed in the following Sect. 4.

3.2 Cluster-specific parameters

We next let σ^2 and τ^2 vary over clusters, so now

$$\mathbf{\Sigma} = \sum_c (\sigma_c^2 \mathbf{G}_c \mathbf{G}_c' + \tau_c^2 \mathbf{b}_c \mathbf{b}_c').$$

Let

$$\mathbf{\Phi} = \begin{pmatrix} \mathbf{\Delta}_n - 2\mathbf{\Delta}_s + \mathbf{A} & \mathbf{\Delta}_n - 2\mathbf{\Delta}_{\tilde{s}} + \mathbf{L} \\ \mathbf{\Delta}_n - 2\mathbf{\Delta}_{\tilde{s}} + \mathbf{L}' & \mathbf{\Delta}_n^2 - 2\mathbf{\Delta}_n \mathbf{\Delta}_{\tilde{s}} + \mathbf{Q} \end{pmatrix},$$

with

$$\begin{aligned} \mathbf{\Delta}_s &= \text{diag tr}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_c'\mathbf{X}_c \\ \mathbf{\Delta}_{\tilde{s}} &= \text{diag } \tilde{\mathbf{x}}_c'(\mathbf{X}'\mathbf{X})^{-1}\tilde{\mathbf{x}}_c \end{aligned}$$

while $\mathbf{A}$, $\mathbf{L}$ and $\mathbf{Q}$ are matrices of order $C \times C$ with typical elements

$$\begin{aligned} a_{cd} &\equiv \text{tr}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_c'\mathbf{X}_c(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_d'\mathbf{X}_d \\ \ell_{cd} &\equiv \tilde{\mathbf{x}}_d'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_c'\mathbf{X}_c(\mathbf{X}'\mathbf{X})^{-1}\tilde{\mathbf{x}}_d \\ q_{cd} &\equiv \left(\tilde{\mathbf{x}}_c'(\mathbf{X}'\mathbf{X})^{-1}\tilde{\mathbf{x}}_d\right)^2. \end{aligned}$$

Then,

$$\hat{\mathbf{v}} = (\mathbf{X}'\mathbf{X} \otimes \mathbf{X}'\mathbf{X})^{-1} \sum_c \left[(\text{vec}\mathbf{X}_c'\mathbf{X}_c)\mathbf{e}_c', (\tilde{\mathbf{x}}_c \otimes \tilde{\mathbf{x}}_c)\mathbf{e}_c'\right] \mathbf{\Phi}^{-1} \sum_c \begin{pmatrix} \hat{\boldsymbol{\varepsilon}}_c' \hat{\boldsymbol{\varepsilon}}_c \mathbf{e}_c \\ \tilde{\hat{\boldsymbol{\varepsilon}}}_c^2 \mathbf{e}_c \end{pmatrix} \quad (10)$$

is the unbiased estimator for the variance $\mathbf{v}$.

Also, here, we present the simpler result when the regressors are neglected. Then,

$$\mathbf{\Phi} = \begin{pmatrix} \mathbf{\Delta}_n & \mathbf{\Delta}_n \\ \mathbf{\Delta}_n & \mathbf{\Delta}_n^2 \end{pmatrix} \quad \text{so} \quad \mathbf{\Phi}^{-1} = \begin{pmatrix} \mathbf{\Delta}_n \mathbf{W} & -\mathbf{W} \\ -\mathbf{W} & \mathbf{W} \end{pmatrix},$$

with $\mathbf{W} \equiv (\mathbf{\Delta}_n^2 - \mathbf{\Delta}_n)^{-1}$. Then,

$$\mathbf{\Phi}^{-1} \sum_c \begin{pmatrix} \hat{\boldsymbol{\varepsilon}}_c' \hat{\boldsymbol{\varepsilon}}_c \mathbf{e}_c \\ \tilde{\hat{\boldsymbol{\varepsilon}}}_c^2 \mathbf{e}_c \end{pmatrix} = \sum_c \frac{1}{n_c(n_c-1)} \begin{pmatrix} (n_c \hat{\boldsymbol{\varepsilon}}_c' \hat{\boldsymbol{\varepsilon}}_c - \tilde{\hat{\boldsymbol{\varepsilon}}}_c^2)\mathbf{e}_c \\ (\tilde{\hat{\boldsymbol{\varepsilon}}}_c^2 - \hat{\boldsymbol{\varepsilon}}_c' \hat{\boldsymbol{\varepsilon}}_c)\mathbf{e}_c \end{pmatrix} \equiv \sum_c \begin{pmatrix} \hat{\sigma}_c^2 \mathbf{e}_c \\ \hat{\tau}_c^2 \mathbf{e}_c \end{pmatrix},$$

with $\hat{o}_c^2$ and $\hat{\tau}_c^2$ implicitly defined. Then,

$$\hat{\mathbf{V}} = (\mathbf{X}'\mathbf{X})^{-1}\left[\sum_c \left(\hat{o}_c^2\mathbf{X}_c'\mathbf{X}_c + \hat{\tau}_c^2\tilde{\mathbf{x}}_c\tilde{\mathbf{x}}_c'\right)\right](\mathbf{X}'\mathbf{X})^{-1},$$

which is the obvious extension to the case of cluster-specific parameters from the one where the parameters are the same over clusters, as discussed in Sect. 3.1.

3.3 Unrestricted error correlation within clusters

The third case we consider has errors that correlate freely within clusters, in a way that differs over clusters. Thus,

$$\mathbf{\Sigma} = \text{diag } \mathbf{\Lambda}_c, \tag{11}$$

where the $\mathbf{\Lambda}_c$ are $n_c \times n_c$ matrices of parameters. With

$$\mathbf{S}_c \equiv \mathbf{I}_{k^2} - \mathbf{I}_k \otimes \mathbf{X}_c'\mathbf{X}_c(\mathbf{X}'\mathbf{X})^{-1} - \mathbf{X}_c'\mathbf{X}_c(\mathbf{X}'\mathbf{X})^{-1} \otimes \mathbf{I}_k,$$

we now obtain

$$\hat{\mathbf{v}} = \left(\mathbf{X}'\mathbf{X} \otimes \mathbf{X}'\mathbf{X} + \sum_c \mathbf{S}_c^{-1}\left(\mathbf{X}_c'\mathbf{X}_c \otimes \mathbf{X}_c'\mathbf{X}_c\right)\right)^{-1} \sum_c \mathbf{S}_c^{-1}\left(\mathbf{X}_c'\hat{\boldsymbol{\varepsilon}}_c \otimes \mathbf{X}_c'\hat{\boldsymbol{\varepsilon}}_c\right) \tag{12}$$

as the unbiased estimator of $\mathbf{v}$ for this case.

As regards the computability of $\hat{\mathbf{v}}$, notice the expression includes the inverse of a $k^2 \times k^2$ matrix with no exploitable structure. However, its inversion should not be problematic computationally for a typical value of k. When cluster-specific dummies are added to the model, the matrix to be inverted increases in size to $(k+C-1)^2 \times (k+C-1)^2$. Any computational problem that might arise is easily averted by eliminating the dummies through the within-transformation (subtract the cluster mean), which can be performed in $O(n)$. The results remain the same, but now in transformed variables, and without the intercept, which becomes zero after the within-transformation.

Also, here, we consider the version of $\hat{\mathbf{v}}$ that neglects the regressors. Rearranged into matrix format, it appears to be

$$\hat{\mathbf{V}} = (\mathbf{X}'\mathbf{X})^{-1}\sum_c \mathbf{X}_c'\hat{\boldsymbol{\varepsilon}}_c\hat{\boldsymbol{\varepsilon}}_c'\mathbf{X}_c(\mathbf{X}'\mathbf{X})^{-1}. \tag{13}$$

This estimator directly generalizes the White (1980) for cross-sections to clusters and was introduced in the context of panel data analysis by Liang and Zeger (1986), where it underlies the widely used panel-robust standard errors allowing for both heteroskedasticity and correlation over time, see, e.g., Cameron and Trivedi (2005).

4 Degrees of freedom

The various expressions for $\hat{\mathbf{V}}$ or $\hat{\mathbf{v}}$ may be of interest by themselves but their main use will be in inference on one particular regression coefficient, β_ℓ, say. For large C, the critical values from a standard normal distribution can be used. However, in practice, C is often small, and using a t-distribution is to be preferred. For instance, Stata uses a $t(C-1)$-distribution after the command REGRESS Y X, VCE(CLUSTER *clustvar*).

Following Satterthwaite (1946), Bell and McCaffrey (2002) proposed a refinement by making the d.f. in the t-distribution data-dependent. The idea is as follows. Let v_ℓ^2 be the variance of the OLS estimator $\hat{\beta}_\ell$ and $\hat{v}_\ell^2$ an estimator of v_ℓ^2. Let

$$T = \frac{\hat{\beta}_\ell}{v_\ell} \Big/ \frac{\hat{v}_\ell}{v_\ell}.$$

Under normality of the regression errors, the numerator is $N(0,1)$ when $\beta_\ell = 0$. Letting $\hat{v}_\ell^2$ be the usual OLS-based estimator of v_ℓ^2, the denominator is distributed according to

$$(n-k)\frac{\hat{v}_\ell^2}{v_\ell^2} \sim \chi^2_{n-k}, \tag{14}$$

leading to the $t(n-k)$-distribution for T. This classical result gets lost when we employ another estimator $\hat{v}_\ell^2$ than the usual one, like one of the cluster-robust estimators discussed in Sect. 3. The proposal of Bell and McCaffrey (2002) is to stay close to (14), by setting the d.f. d_ℓ such that

$$d_\ell \frac{\hat{v}_\ell^2}{v_\ell^2} \overset{\text{app}}{\sim} \chi^2_{d_\ell},$$

where "app" stands for "approximately" in the sense that the first two moments of $d_\ell \hat{v}_\ell^2 / v_\ell^2$ match those of a χ^2-distribution with d_ℓ d.f. Using unbiased estimators of the variance as derived in the preceding section proves its usefulness here since then the first moments left and right match. Letting the second moments match means $\text{var}(d_\ell \hat{v}_\ell^2 / v_\ell^2) = 2d_\ell$ or

$$d_\ell = 2\frac{(v_\ell^2)^2}{\text{var}(\hat{v}_\ell^2)}. \tag{15}$$

Obviously, d_ℓ is not known and needs to be estimated. There are two issues with this. One is that d_ℓ may depend on parameters, which have to be estimated. A second issue is that evaluating v_ℓ^2 and $\text{var}(\hat{v}_\ell^2)$ requires the distribution of $\boldsymbol{\varepsilon}$. As a practical solution to obtain a reasonable value of $\hat{d}_\ell$, Bell and McCaffrey (2002) propose to simply take $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_n)$ as the "reference distribution." Imbens and Kolesár (2016) suggested to take the RE model as the reference distribution, $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_n + \tau^2 \mathbf{BB})$, with $\mathbf{B}$ as defined in (4). We will now derive expressions for d_ℓ for both cases. Given our focus

on unbiased estimation, we extend previous results by using an unbiased estimator for $\text{var}(\hat{v}_\ell^2)$ and by using an unbiased estimator of any parameter that we meet in d_ℓ.

So, first following Bell and McCaffrey (2002), we let $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I}_n)$. As $\hat{v}_\ell^2$ is quadratic in $\hat{\boldsymbol{\varepsilon}}$, we can write $\hat{v}_\ell^2 = \hat{\boldsymbol{\varepsilon}}'\mathbf{A}_\ell\hat{\boldsymbol{\varepsilon}}$ for some symmetric $n \times n$ matrix $\mathbf{A}_\ell$ whose particular form follows from (5), (10) or (12), depending on the case under consideration. For notational simplicity, we will omit the subscript ℓ to $\mathbf{A}$ from now on and denote $\mathbf{a} \equiv \text{vec}\mathbf{A}$, so

$$\begin{aligned}\hat{v}_\ell^2 &= \hat{\boldsymbol{\varepsilon}}'\mathbf{A}\hat{\boldsymbol{\varepsilon}} \\ &= \mathbf{a}'(\hat{\boldsymbol{\varepsilon}} \otimes \hat{\boldsymbol{\varepsilon}}) \\ &= \mathbf{a}'(\mathbf{M} \otimes \mathbf{M})(\boldsymbol{\varepsilon} \otimes \boldsymbol{\varepsilon})\end{aligned}$$

hence,

$$\begin{aligned}\text{var}(\hat{v}_\ell^2) &= 2\sigma^4\mathbf{a}'(\mathbf{M} \otimes \mathbf{M})\mathbf{a} \\ &= 2\sigma^4\text{tr}\mathbf{AMAM}. \end{aligned} \tag{16}$$

From (5), (10) and (12), $\mathbf{A}$ readily appears to be block-diagonal, with cth block $\mathbf{A}_c$ given by

$$\mathbf{A}_c = r_1\mathbf{I}_c + r_2\mathbf{i}_c\mathbf{i}_c', \qquad (r_1, r_2) = \mathbf{f}_\ell'(\mathbf{X}'\mathbf{X} \otimes \mathbf{X}'\mathbf{X})^{-1}(\text{vec}\mathbf{X}'\mathbf{X}, \text{vec}\tilde{\mathbf{X}}'\tilde{\mathbf{X}})\boldsymbol{\Psi}^{-1}$$

$$\mathbf{A}_c = r_{1c}\mathbf{I}_c + r_{2c}\mathbf{i}_c\mathbf{i}_c', \qquad (\mathbf{r}_1', \mathbf{r}_2') = \mathbf{f}_\ell'(\mathbf{X}'\mathbf{X} \otimes \mathbf{X}'\mathbf{X})^{-1}\sum_c \left((\text{vec}\mathbf{X}_c'\mathbf{X}_c)\mathbf{e}_c', (\tilde{\mathbf{x}}_c \otimes \tilde{\mathbf{x}}_c)\mathbf{e}_c'\right)\boldsymbol{\Phi}^{-1}$$

$$\mathbf{A}_c = \mathbf{X}_c\mathbf{Q}_c\mathbf{X}_c', \qquad (\text{vec}\mathbf{Q}_c)' = \mathbf{f}_\ell'\left(\mathbf{X}'\mathbf{X} \otimes \mathbf{X}'\mathbf{X} + \sum_c \mathbf{S}_c^{-1}(\mathbf{X}_c'\mathbf{X}_c \otimes \mathbf{X}_c'\mathbf{X}_c)\right)^{-1}\mathbf{S}_c^{-1},$$

respectively, with $\mathbf{f}_\ell \equiv \mathbf{e}_\ell \otimes \mathbf{e}_\ell$ and $\mathbf{r}_1 \equiv (r_{11}, \ldots, r_{1C})'$ and likewise for $\mathbf{r}_2$. Since $v_\ell^2 = \sigma^2\mathbf{e}_\ell'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{e}_\ell$, we obtain

$$d_\ell = \frac{\left(\mathbf{e}_\ell'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{e}_\ell\right)^2}{\text{tr}\mathbf{AMAM}}, \tag{17}$$

with

$$\begin{aligned}\text{tr}\mathbf{AMAM} &= \text{tr}\sum_{c,d}\mathbf{G}_c\mathbf{A}_c\mathbf{G}_c'(\mathbf{I} - \mathbf{P})\mathbf{G}_d\mathbf{A}_d\mathbf{G}_d'(\mathbf{I} - \mathbf{P}) \\ &= \text{tr}\sum_c \mathbf{A}_c^2 - 2\text{tr}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{A}^2\mathbf{X} + \text{tr}((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{A}\mathbf{X})^2. \end{aligned} \tag{18}$$

Computational gains can be had by exploiting the structure of $\mathbf{A}_c$. Notice that the expression for d_ℓ does not depend on unknown parameters since the factors σ^4 in the numerator and the denominator cancel out.

Next, following Imbens and Kolesár (2016), we let $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}_n + \tau^2\mathbf{B}\mathbf{B}'$, with $\mathbf{B}$ as defined in (4). Instead of (16), we now have $\text{var}(\hat{v}_\ell^2) = 2\text{tr}\mathbf{AM}\boldsymbol{\Sigma}\mathbf{MAM}\boldsymbol{\Sigma}\mathbf{M}$, and (17) generalizes to

$$d_\ell = \frac{\left(\mathbf{e}_\ell'(\sigma^2(\mathbf{X}'\mathbf{X})^{-1} + \tau^2(\mathbf{X}'\mathbf{X})^{-1}\tilde{\mathbf{X}}'\tilde{\mathbf{X}}(\mathbf{X}'\mathbf{X})^{-1})\mathbf{e}_\ell\right)^2}{\text{tr}\mathbf{AM\Sigma MAM\Sigma M}}. \tag{19}$$

Here, both numerator and denominator depend on the parameters σ^4, τ^4 and $\sigma^2\tau^2$, which do not cancel out and hence have to be replaced by estimators. The lengthy expression in the denominator poses another complication. Both complications are addressed in "Online Appendix B". Our simulation results indicate that this more general procedure to estimate the degrees of freedom is particularly useful when the clusters are of unequal size.

5 Simulation design

We take the simulation design of MacKinnon and Webb (2018) as our point of departure. The data generating process includes a treatment dummy and a continuous variable. For $c = 1, \ldots, C$, it is

$$\mathbf{y}_c = \mathbf{i}_c\alpha + \mathbf{d}_c\beta + \mathbf{x}_c\gamma + \boldsymbol{\varepsilon}_c, \tag{20}$$

with $\mathbf{i}_c$ the intercept, $\mathbf{d}_c$ the treatment dummy equal to 1 in clusters $1, \ldots, C_1$, which we will vary from 1 to $C-1$, and $\mathbf{x}_c$ the continuous regressor, whose elements are independent $N(0, 1)$. The regression errors $\boldsymbol{\varepsilon}_c$ within cluster c are normally distributed with their covariance matrix $\boldsymbol{\Sigma}_c$ specified below. The errors are independent across clusters. We set the parameters $\alpha = \beta = \gamma = 0$, the number of clusters $C = 14$, and the total number of observations $n = 2800$. The results below are based on 200,000 draws of (20). We draw the continuous variable $\mathbf{x}_c$ only once.

Error covariance matrix To generate the data, we consider three increasingly complicated designs for the covariance matrix of the $\boldsymbol{\varepsilon}_c$.

1. Homogeneous design as Sect. 3.1,

$$\boldsymbol{\Sigma}_c = \sigma^2\mathbf{I}_c + \tau^2\mathbf{i}_c\mathbf{i}_c'. \tag{21}$$

   with $\sigma^2 = 1$ and $\tau^2 = 0.1$.
2. Restricted heterogeneous design as in Sect. 3.2,

$$\boldsymbol{\Sigma}_c = \sigma_c^2\mathbf{I}_c + \tau_c^2\mathbf{i}_c\mathbf{i}_c' \qquad \sigma_c^2 = \exp\left(2\delta\frac{C-c}{C-1}\right) \qquad \tau_c^2 = \rho\sigma_c^2. \tag{22}$$

   This way of including heterogeneity across clusters is borrowed from MacKinnon and Webb (2018). We set $\rho = 0.1$ and $\delta = \ln(2)/2$, which means that σ_c^2 ranges from 1 to 2.
3. Unrestricted heterogeneous design as in Sect. 3.3,

$$\boldsymbol{\Sigma}_c = \sigma^2\mathbf{I}_c + \tau^2\mathbf{i}_c\mathbf{i}_c' + \text{diag}(\mathbf{x}_c)^2/2, \tag{23}$$

   with σ^2 and τ^2 as in the homogeneous design.

Balance An important design choice is the number of observations per cluster. We first consider a balanced design, where the number of observations per cluster is equal to $n/C = 200$, and next an unbalanced design, where the number of observations depends on the cluster index according to

$$n_c = \text{int}\left(n\frac{\exp(\gamma c/C)}{\sum_c \exp(\gamma c/C)}\right), \quad c = 1, \ldots, C-1, \quad n_C = n - \sum_c n_c. \tag{24}$$

We set $\gamma = 2$, which implies cluster sizes ranging from 67 to 438 observations.
Variance estimators and reference distributions We consider the following methods to obtain t-values for the OLS estimate for β in (20).

1. The first benchmark t-values are based on the cluster extension of White's standard errors due to Liang and Zeger (1986) as already introduced in (13), but with a finite-sample correction as implemented in Stata,

$$\hat{\mathbf{V}}_{\text{LZ1}} = \frac{C}{C-1}\frac{n-1}{n-k}(\mathbf{X}'\mathbf{X})^{-1}\sum_c \mathbf{X}_c'\hat{\boldsymbol{\varepsilon}}_c\hat{\boldsymbol{\varepsilon}}_c'\mathbf{X}_c(\mathbf{X}'\mathbf{X})^{-1}.$$

   Following Stata, we compare the resulting t-statistic against the critical values of a $t(C-1)$ distribution. We denote this benchmark method by STATA.
2. The second benchmark t-values implement the Liang and Zeger (1986) standard errors with a HC2 correction as in Bell and McCaffrey (2002).

$$\hat{\mathbf{V}}_{\text{LZ2}} = (\mathbf{X}'\mathbf{X})^{-1}\sum_c \mathbf{X}_c'(\mathbf{I}_c - \mathbf{P}_{cc})^{-1/2}\hat{\boldsymbol{\varepsilon}}_c\hat{\boldsymbol{\varepsilon}}_c'(\mathbf{I}_c - \mathbf{P}_{cc})^{-1/2}\mathbf{X}_c(\mathbf{X}'\mathbf{X})^{-1},$$

   where $\mathbf{P}_{cc} \equiv \mathbf{X}_c(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_c'$. Computation of $\hat{\mathbf{V}}_{\text{LZ2}}$ involves the inverse of the square root of the $n_c \times n_c$ matrices $\mathbf{I}_c - \mathbf{P}_{cc}$, which can be problematic for large n_c. However, Niccodemi et al. (2020) and Kolesár (2022) show how efficient computation can be achieved, that is, in $O(n_c)$. We compare the t-statistic that follows from using $\hat{\mathbf{V}}_{\text{LZ2}}$ against the critical values of a $t(d_{\text{IK}})$ distribution, with d_{IK} the d.f. suggested by Imbens and Kolesár (2016). We denote this benchmark method by LZIK.
3. The third benchmark is the wild cluster bootstrap proposed by Cameron et al. (2008). Its asymptotic validity under a diverging number of clusters was shown by Djogbenou et al. (2019). We implement the restricted version as described in Sect. 3.2 of MacKinnon (2022), where we set the number of bootstrap draws at 999. This version calculates the distribution of t-statistics that are based on the variance estimator $\hat{\mathbf{V}}_{\text{LZ1}}$ defined above. Since our simulation setting is close to the one analyzed by MacKinnon and Webb (2018), the results for the wild bootstrap coincide with their findings.
4. We use the three unbiased variance estimators from Sects. 3.1–3.3, denoted by UV1, UV2, and UV3, respectively, and compare the resulting t-statistics against the critical values of a t-distribution for both reference distributions considered (indicated by RV0 and RV1, respectively), so with d.f. d_ℓ from (17) and from (19).

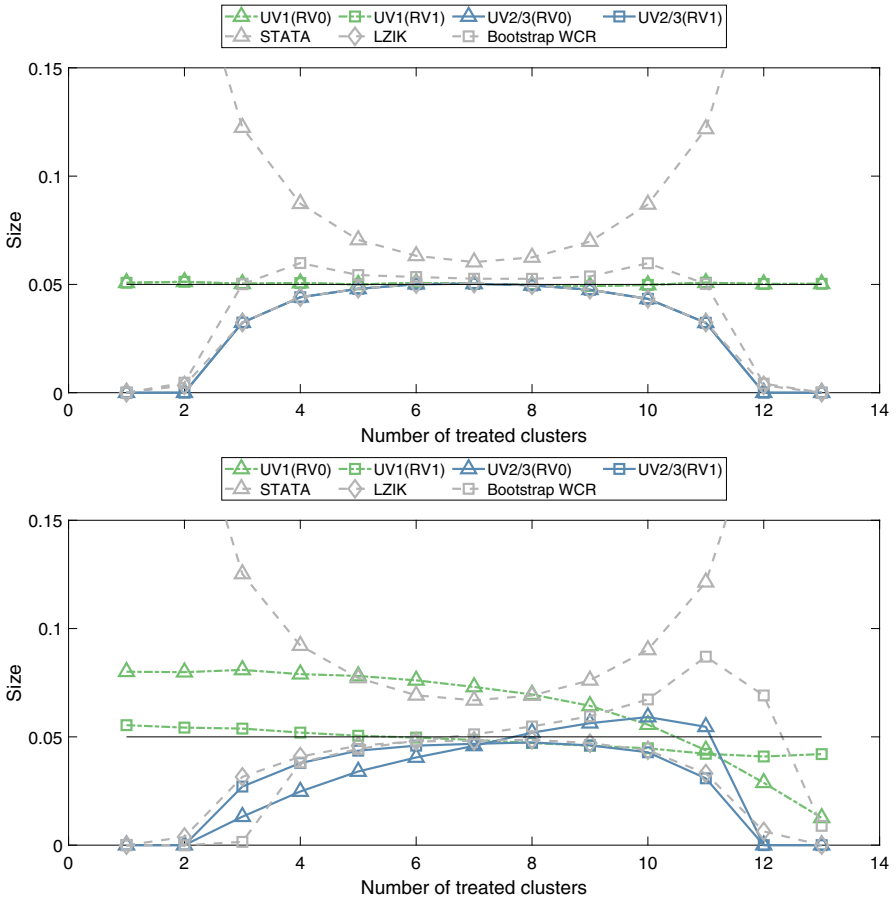

Fig. 1 Size of the t test for the treatment dummy, SV1

This yields six cases, UV1(RV0), UV1(RV1), UV2(RV0), UV2(RV1), UV3(RV0), and UV3(RV1).

Notice that LZ2 does not exist when the number of (un)treated clusters is smaller than two, and that UV2(·), UV3(·) do not exist when the number of (un)treated clusters is smaller than three. We then set the size to zero.

6 Simulation results

The main results of the simulations are presented in Figs. 1, 2 and 3, based on data simulated with error covariance matrix as in (21), (22) and (23), respectively. They show the size of the t test for $H_0 : \beta = 0$, with β the coefficient of the dummy variable in (20). The number of treated clusters is on the horizontal axis. The upper panel of each figure is for the balanced case and the lower panel for the unbalanced

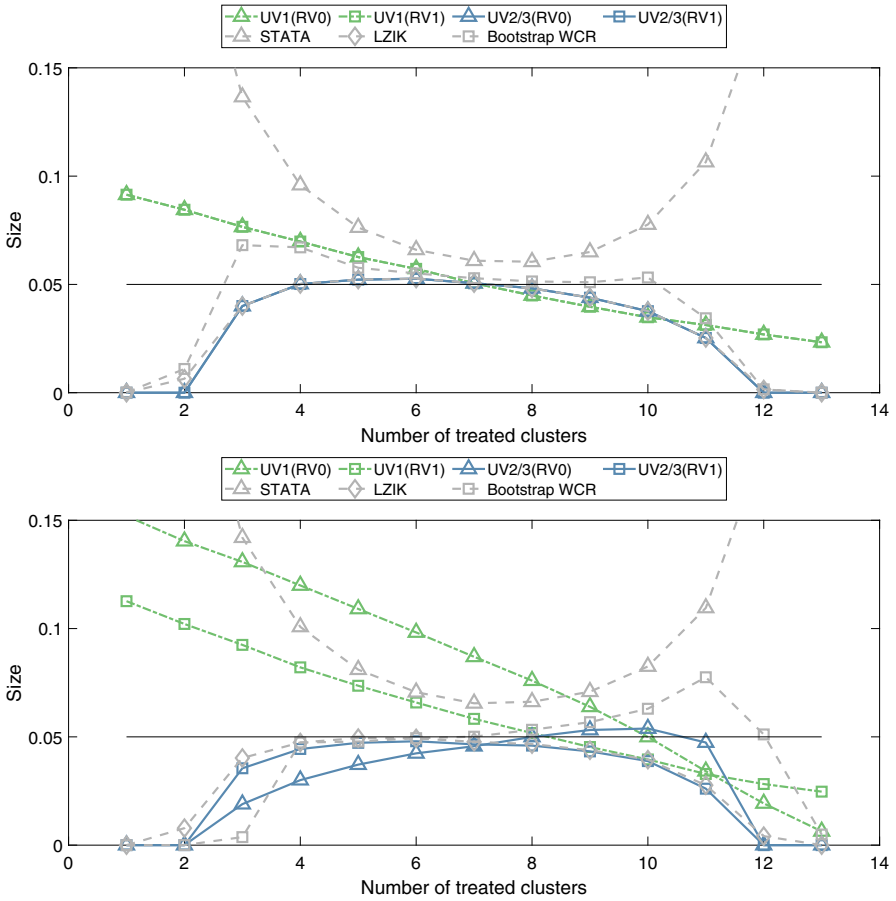

Fig. 2 Size of the t test for the treatment dummy, SV2

case as described in (24). Each figure shows seven curves. The first four are STATA, LZIK, UV1(RV0), UV1(RV1). When we analyze the t-test on the treatment variable, the differences between UV2(RV0) and UV3(RV0), as well as those between UV2(RV1) and UV3(RV1), are not visible, so we report those as UV2/3(RV0) and UV2/3(RV1). Finally, we report the results for the restricted wild bootstrap. Notice that three variances are involved: the reference variance to obtain d_ℓ; the variance whose unbiased estimator was used; and the variance used in the simulation. For clarity, Table 1 summarizes.

The most relevant curves in all three figures are the ones labeled UV1(RV1) in Fig. 1, upper and lower panels. The homogeneous RE design can be considered the more or less generic case in the clustered-error literature and, as is apparent from Table 1, this particular curve is maximally based on this design as it underlies the data generation SV1, the variance estimator UV1, and d_ℓ based on RV1.

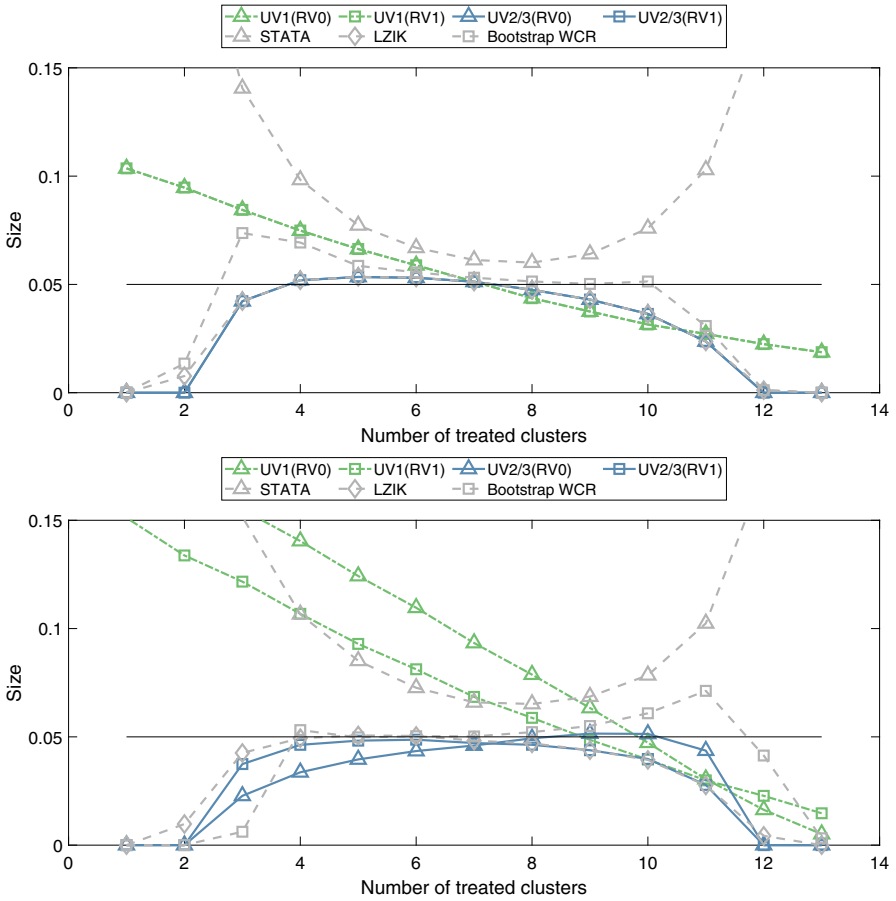

Fig. 3 Size of the t test for the treatment dummy, SV3

Table 1 Overview of the variances used

$\boldsymbol{\Sigma}_c$	Reference variance	Unbiased estimator	Simulation variance
$\sigma^2\mathbf{I}_c$	RV0		
$\sigma^2\mathbf{I}_c + \tau^2\mathbf{i}_c\mathbf{i}_c'$	RV1	UV1	SV1
$\sigma_c^2\mathbf{I}_c + \tau_c^2\mathbf{i}_c\mathbf{i}_c'$		UV2	SV2
$\boldsymbol{\Lambda}_c$		UV3	SV3

SV1 Inspecting Fig. 1 we see, for the balanced design in the upper panel, excellent size control for UV1(·). This holds even when there is only a single treated cluster. It does not appear to matter whether the d.f. are calculated under the more restrictive i.i.d. assumption, UV1(RV0), or the RE structure, UV1(RV1). By contrast, UV2(·), UV3(·) and LZIK are slightly conservative when we have a small or large number of treated clusters. The STATA variance estimator performs quite poorly, especially when the

number of treated clusters is small or large. The bootstrap with t-statistics based on the STATA variance estimator performs much better.

Moving to the unbalanced setup in the lower panel of Fig. 1, we see that UV1(RV0) no longer provides accurate size control. However, UV1(RV1), the most relevant case as argued above, still exhibits excellent performance. The additional computational complexity of this approach appears to pay off. We also see that, unlike in the balanced case, the results for UV2(RV0) and UV3(RV0) differ from the benchmark variance estimator LZIK. The unbiased variance estimators are more conservative for a small number of treated clusters, while becoming slightly oversized for 9–11 treated clusters. UV2(RV1) and UV3(RV1) are again very close to LZIK. The STATA variance estimator again is found not to accurately control size. For the bootstrap, we find that it is undersized for a small number of treated clusters and oversized for a large number of treated clusters.

SV2 In Fig. 2, we show the size for t tests based on the various variances estimators under the restricted heterogeneous design where each cluster has its own variance and covariance parameter. This setup is more general than the homogeneous design in which each cluster has the same variance and covariance parameter. As expected, the performance of UV1(RV0) and UV1(RV1) somewhat deteriorates in this setup, with size slightly below 0.10 for the case of a single treated cluster and balanced design. The same is observed for an unbalanced design, with the size obtained under UV1(RV1) being just over 0.10.

For UV2 and UV3, under both d.f., and LZIK, we see the test slightly overrejects for a small number of treated clusters. When the number of treated clusters increases, the tests become progressively more conservative. Again, a difference emerges between UV2, UV3 and LZIK in the unbalanced case presented in the lower panel of Fig. 2. Here, size control is more accurate for UV2 and UV3 compared to LZIK. Especially UV1(RV0) and UV2(RV0) perform well in this setup, providing accurate size control up to roughly eight treated clusters. With more treated clusters, they tend to be conservative, although not as much as LZIK. The bootstrap performance is similar to that in SV1, although in a balanced design with a small number of treated clusters, it is slightly oversized.

SV3 The results for the unrestricted heterogeneous design are nearly identical to those in the homogenous design for the STATA variance, the bootstrap, LZIK and UV2 and UV3 under both d.f. corrections. For UV1, we find reasonable performance when clusters are balanced. When the clusters are unbalanced, UV1(RV0) becomes oversized for a small number of treated clusters and undersized when the number of treated clusters is large. The more general d.f. correction in UV1(RV1) partly corrects these size distortions.

So far for the test on β, the coefficient of the cluster-specific dummy variable. We can be much more concise as to γ, the coefficient of the continuous variable. For SV1 and SV2, the size control is almost perfect. This no longer holds for SV3, where the size is still almost perfect for STATA, LZIK, UV3$(\cdot)$ but appears to be double the nominal size for $UV1(\cdot)$ and UV2$(\cdot)$; the latter methods are apparently sensitive when the data are generated according to more general scheme SV3.

Degrees of freedom in SV1 Given the notable differences in performance when using degrees of freedom based on RV0 or RV1, we analyze the degrees of freedom under

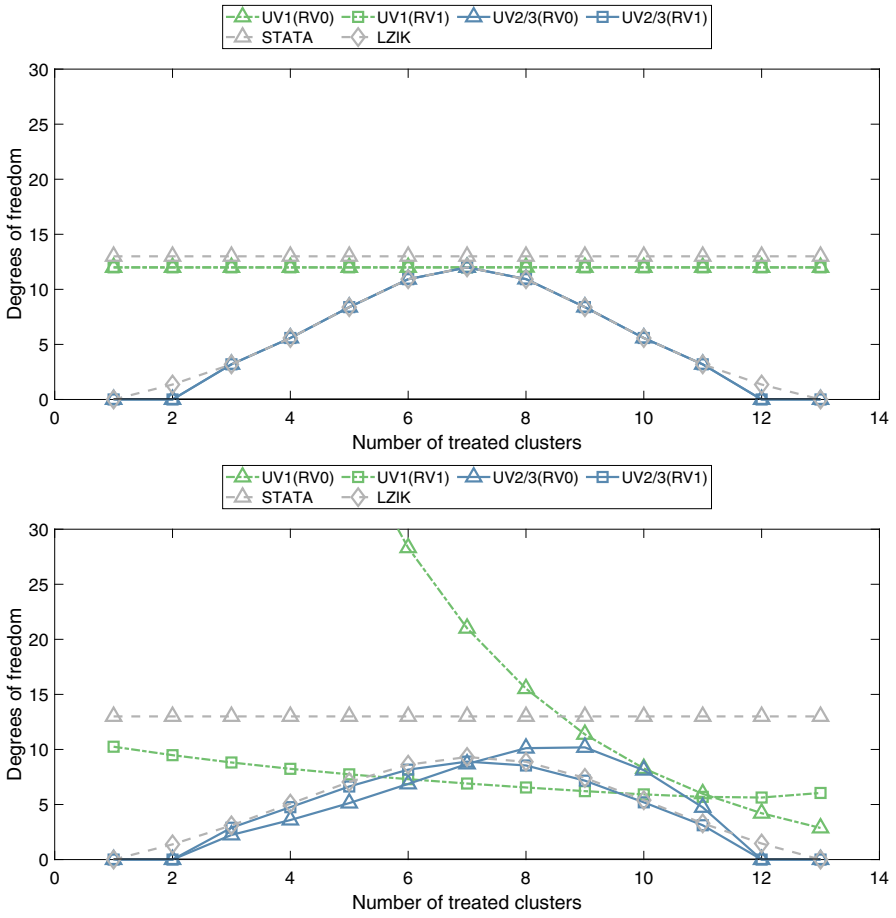

Fig. 4 Simulations: treatment dummy with homogeneous error covariance matrix. Degrees of freedom

SV1 in Fig. 4. For a balanced design, we see that the degrees of freedom for UV1 are equal to $C - 2$. Donald and Lang (2007) show that if the design is balanced and if all regressors are invariant within clusters, the t-statistic is $t(C - k)$ distributed, where k is the number of regressors in the model. We can expect the same result to apply here since the continuous variable is uncorrelated with the treatment dummy.

Under a balanced design, the degrees of freedom for the other methods are nearly identical. They are low when the number of treated clusters is low and increase to their maximum when half of the clusters is treated. This maximum appears to coincide numerically with $C - k$ as well.

When the design is unbalanced, we see a strong deviation from the degrees of freedom under RV0 to those under RV1. This is especially true for UV1 and a small number of treated clusters. For the remaining variance estimators, we see that under RV0 the degrees of freedom are asymmetric in the number of treated clusters, while those under RV1 are symmetric.

Theoretical explanation of the differences The simulations highlight that UV1 can offer accurate size control even with only a single treated cluster, while UV2 and UV3 require a somewhat larger number of treated clusters. To explain these results from a theoretical perspective, we derive in "Online Appendix C" the required conditions for the consistency of the variance estimators in a simple model. There is a single treatment variable that is equal to one in t_C out of C clusters. The design is balanced, so that each cluster has n/C observations. The errors are $N(\mathbf{0}, \mathbf{\Sigma})$ with $\mathbf{\Sigma}$ as in Sect. 3.1. "Online Appendix C" shows that UV1 is consistent if $n^2/C^3 \to 0$. This requires the number of clusters to grow sufficiently fast, but does not impose any restriction on the number of *treated* clusters. For UV2 and UV3 on the other hand, we find that the number of treated clusters should diverge sufficiently fast so that $n^2/(C^2 \cdot t_C) \to 0$. In contrast to UV1, we now only achieve consistency when the number of *treated* clusters goes to infinity. These results explain the difference in performance of the variance estimators when the number of treated clusters is small.

7 A placebo-regression experiment

To analyze the performance of the unbiased variance estimators in an empirical setting, we consider a placebo-regression experiment. Placebo regressions were originally proposed by Bertrand et al. (2004) to analyze the validity of commonly used standard errors for difference-in-difference estimators. We consider an application similar to that in Cameron and Miller (2015).

We use the Current Population Survey (CPS) 2012 data set that can be obtained from `https://cps.ipums.org/cps/`. The data consist of 51 clusters: the fifty American states and the District of Columbia. The number of observations in each cluster varies from 519 (Montana) to 5866 (California). For observation h in cluster $i = 1, \ldots, C$, we define the model

$$\ln(\texttt{wage})_{hi} = \beta_0 + \beta_1 \texttt{educ}_{hi} + \beta_2 \texttt{age}_{hi} + \beta_3 \texttt{age}_{hi}^2 + \beta_4 \texttt{policy}_i + \varepsilon_{hi}. \tag{25}$$

Here, `policy` is a fake policy variable that is randomly assigned to $C_1 = 1, \ldots, C-1$ sampled clusters and constant within each cluster. Since the policy variable is fake, we expect 5% rejections across the replications when we test the hypothesis $\mathrm{H}_0 : \beta_4 = 0$ at the 5% level.

In line with the simulations in the previous section, we sample a subset of $C = 14$ clusters from the 51 available clusters. We consider two different ways of sampling this subset. In the first, we randomly sample clusters with replacement. To test the methods in an unbalanced setup, we also consider using the 3 states with the most observations and the 11 states with the fewest observations. To preserve the relative share of observations in each cluster, we randomly sample with replacement 20% of the observations within each sampled cluster.

Figures 5 and 6 show the empirical size (upper panel) and the degrees of freedom (lower panel) averaged over 10,000 replications for the four different designs. The x-axis again depicts the number of treated clusters.

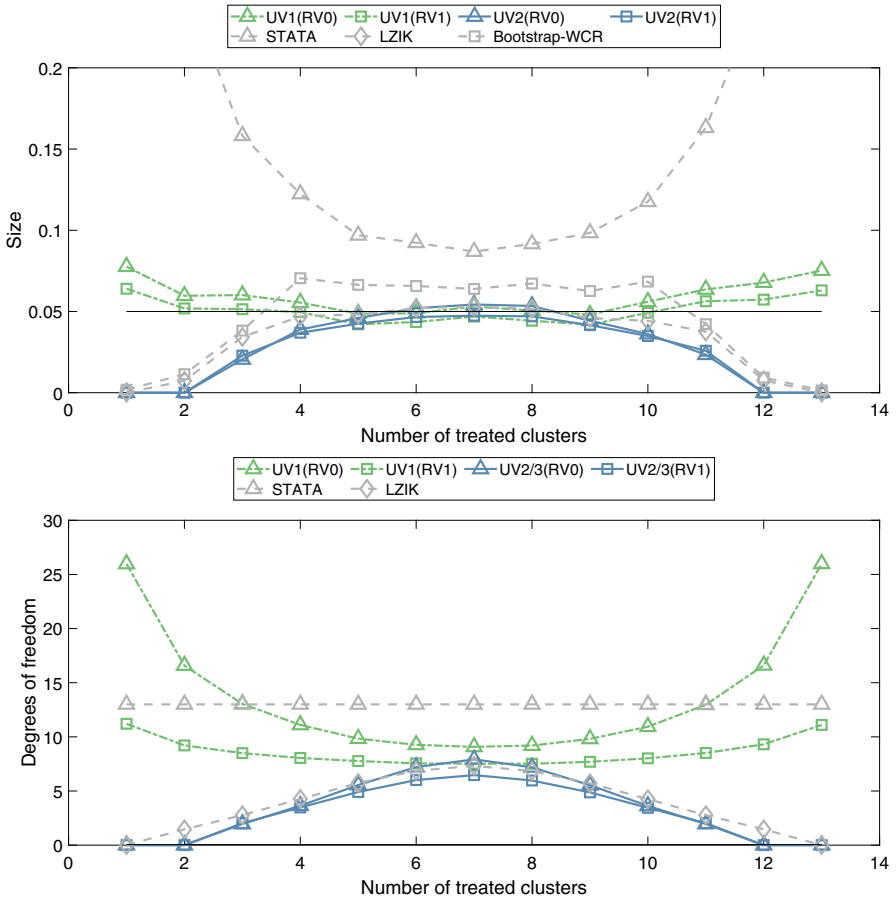

Fig. 5 Application: randomly drawn states. Size and degrees of freedom

In line with the Monte Carlo results from the previous section, we see that the Stata variance estimator with $C - 1$ degrees of freedom is severely oversized. This effect is largely mitigated by using the bootstrap, although it is consistently oversized for a moderate number of treated clusters. This is especially the case in the "3–11" setting. In contrast, we find remarkably good size control for UV1(RV1) across the designs. The degrees of freedom drop considerably when moving from RV0 to RV1. This shows that the use of RV1 is of empirical relevance, especially in the settings with higher imbalance and a small number of (un)treated clusters. The LZIK variance estimator also performs well, although it is oversized in the highly unbalanced "3–11" setting. There the unbiased variance matrix estimators control size more accurately.

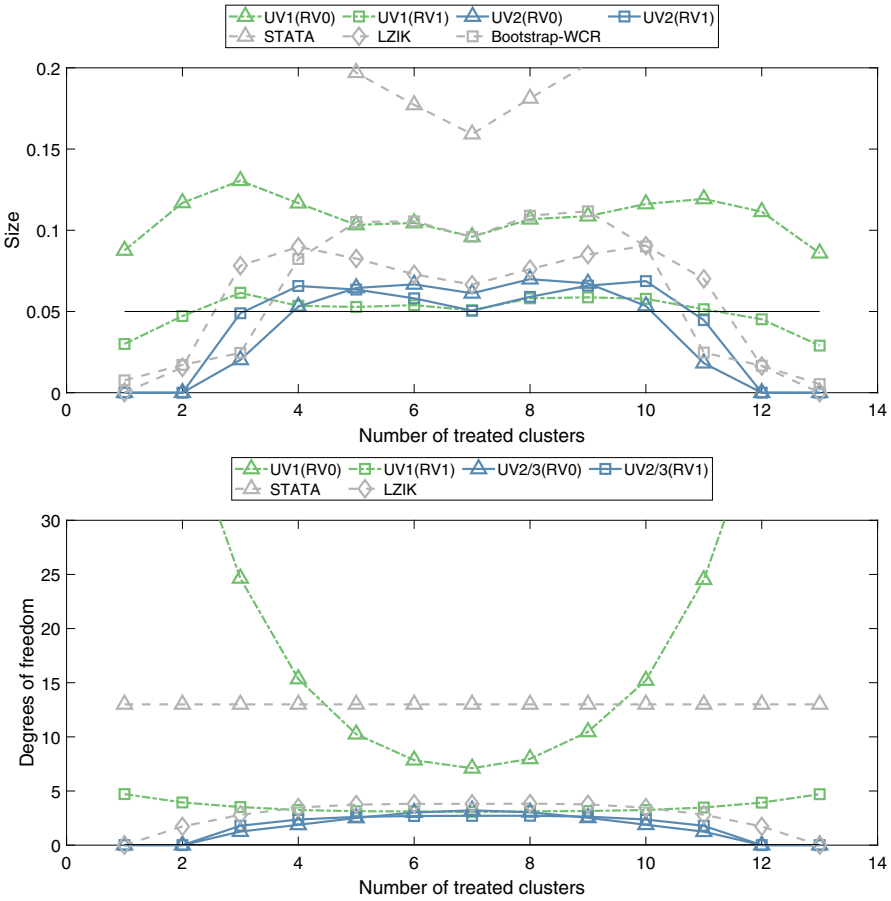

Fig. 6 Application: "3–11". Size and degrees of freedom

8 Concluding remarks

The point of departure in this paper has been to derive unbiased estimators of the covariance matrix of the OLS estimator when the data are clustered. We considered three cases, the leading one being the RE model. This led to our main research question, which is to assess the performance of these estimators in the t test for a particular regression coefficient, both among each other and vis-à-vis two oft-used alternatives.

We addressed this question by simulation, in a regression model with a two regressors, one being continuous and distributed equally in all clusters, while the other regressor represented a cluster-specific treatment dummy. The main finding of the simulation study was the excellent behavior of the t test based on the unbiased estimator for the RE model, for the case that the data actually have been generated according to this model and the degrees of freedom have been based on it. So the three variances that play a role are aligned. This result holds for the coefficient of the cluster-specific

dummy variable; there is hardly a noticeable difference in performance for between the other variance estimators underlying the t test.

The random-effects model considered in Sect. 3.1 suggests an issue worthy of investigation. Throughout the paper, we considered the OLS estimator and the t-values related to it under various specifications. However, we can also consider the feasible GLS estimator. When the random-effects specification would be the correct one (and the random-effects parameters would be known exactly), the model would have no clustered-error terms anymore, after the usual transformation well-known from the panel data literature. Unlike the transformation corresponding with fixed effects, the transformation for random effects keeps cluster-specific regressors in the model, although with little variation over time, so leading to large variances of the GLS estimators. It is interesting to know how this would work out in theory and practice.

In our analysis, we have restricted ourselves to the case of a cross-sectional model. An obvious topic for future research is an extension to case of panel data and difference-in-difference models.

A next step is to see if the excellent behavior mentioned above also shows up in the case where the three variances are still aligned but now pertain to the more flexible RE model where the two error-components parameters differ over clusters. While by itself this is eminently doable, the question arises to test this heterogeneous RE structure against the homogeneous one. An obvious starting point is the score test context proposed by Breusch and Pagan (1980). Deriving the relevant expression is straightforward but deriving the (limiting) distribution of the test statistic is not since the number of parameters grows with the number of clusters. We can have $n \to \infty$ when $C \to \infty$ but we can also consider keeping C fixed while letting the number of observations per cluster go to infinity, or any combination of the two.

The results in the paper on the quality of unbiased estimators in the t test are based on simulation only. We are not aware of any theory that might help giving these results a theoretical basis. There is certainly a research challenge here.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00181-023-02379-w.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Human and animal rights This article does not contain any studies with human participants or animals performed by any of the authors.

References

Abadie A, Athey S, Imbens GW, Wooldridge JM (2020) Sampling-based versus design-based uncertainty in regression analysis. Econometrica 88(1):265–296
Bell RM, McCaffrey DF (2002) Bias reduction in standard errors for linear regression with multi-stage samples. Surv Methodol 28:169–179
Bertrand M, Duflo E, Mullainathan S (2004) How much should we trust differences-in-differences estimates? Quart J Econ 119:249–275
Breusch TS, Pagan AR (1980) The Lagrange multiplier test and its applications to model specification in econometrics. Rev Econ Stud 47:239–253
Cameron AC, Miller DL (2015) A practitioner's guide to cluster-robust inference. J Hum Resour 50:317–372
Cameron AC, Trivedi PK (2005) Microeconometrics. Cambridge University Press, Cambridge
Cameron AC, Gelbach JB, Miller DL (2008) Bootstrap-based improvements for inference with clustered errors. Rev Econ Stat 90(3):414–427
Djogbenou AA, MacKinnon JG, Nielsen MØ (2019) Asymptotic theory and wild bootstrap inference with clustered errors. J Econom 212(2):393–412
Donald SG, Lang K (2007) Inference with difference-in-differences and other panel data. Rev Econ Stat 89(2):221–233
Hansen BE, Lee S (2019) Asymptotic theory for clustered samples. J Econom 210(2):268–290
Hartley H, Rao J, Kiefer G (1969) Variance estimation with one unit per stratum. J Am Stat Assoc 64:173–181
Ibragimov R, Müller UK (2016) Inference with few heterogeneous clusters. Rev Econ Stat 98(1):83–96
Imbens GW, Kolesár M (2016) Robust standard errors in small samples: some practical advice. Rev Econ Stat 98(4):701–712
Kline P, Saggio R, Sølvsten M (2020) Leave-out estimation of variance components. Econometrica 88(5):1859–1898
Kolesár M (2022) Robust standard errors in small samples. https://cran.r-project.org/web/packages/dfadjust/vignettes/dfadjust.pdf
Liang K-Y, Zeger SL (1986) Longitudinal data analysis using generalized linear models. Biometrika 73(1):13–22
MacKinnon JG (2022) Fast cluster bootstrap methods for linear regression models. Econom Stat 21
MacKinnon JG, Webb MD (2018) The wild bootstrap for few (treated) clusters. Econom J 21(2):114–135
MacKinnon JG, White HL (1985) Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. J Econom 29(3):305–325
MacKinnon JG, Nielsen MØ Webb MD (2023) Cluster-robust inference: a guide to empirical practice. J Econom 232(2):272–299
Niccodemi G, Alessie RJM, Angelini V, Mierau JO, Wansbeek T (2020) Refining clustered standard errors with few clusters. https://www.rug.nl/feb/research/som-research-reports-2012-2022/som-research-reports-2020/2020002-eef-def.pdf
Satterthwaite FE (1946) An approximate distribution of estimates of variance components. Biom Bull 2:110–114
White HL (1980) A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. Econometrica 48:817–838