HETEROSKEDASTICITY-ROBUST INFERENCE IN LINEAR REGRESSION MODELS WITH MANY COVARIATES

KOEN JOCHMANS*
UNIVERSITY OF CAMBRIDGE

Abstract

We consider inference in linear regression models that is robust to heteroskedasticity and the presence of many control variables. When the number of control variables increases at the same rate as the sample size the usual heteroskedasticity-robust estimators of the covariance matrix are inconsistent. Hence, tests based on these estimators are size distorted even in large samples. An alternative covariance-matrix estimator that remains consistent is presented. Unlike the estimator of Cattaneo, Jansson and Newey (2018) this estimator remains well-defined in the presence of highly-influential observations. Simulation results and an empirical illustration are also provided.

Keywords: heteroskedasticity, inference, many regressors, statistical leverage.

1 Introduction

When performing inference in linear regression models it is common practice to safeguard against (conditional) heteroskedasticity of unknown form. The standard estimator of the covariance matrix of the least-squares estimator proposed by Eicker (1963, 1967) and White (1980) is known to be biased. The bias can be severe if the regression design contains observations with high leverage (Chesher and Jewitt, 1987). A necessary condition for the least-squares estimator to be consistent is that maximal leverage vanishes in large

*Address: University of Cambridge, Faculty of Economics, Austin Robinson Building, Sidgwick Avenue, Cambridge CB3 9DD, United Kingdom. E-mail: kj345@cam.ac.uk. I am grateful to Matias Cattaneo for generously sharing and discussing his replication material, and to an Associate Editor and three referees for very constructive feedback on an earlier version of this paper. Financial support from the European Research Council through grant n° 715787 (MiMo) is gratefully acknowledged. The first version of this paper dates from September 5, 2018. This version is dated August 28, 2019.

samples (Huber, 1973). This then also implies consistency of the robust covariance-matrix estimator.

The condition that maximal leverage vanishes is problematic when the regressors include a large set of control variables. In such a setting asymptotics where the number of regressors is treated as fixed are inappropriate. The slope coefficients on the control variables are nuisance parameters. Under asymptotics where their number, q_n, grows with the sample size, n, the robust covariance-matrix estimator will be inconsistent unless $q_n/n \to 0$, as shown by Cattaneo, Jansson and Newey (2018). This result is a manifestation of the incidental-parameter problem (Neyman and Scott, 1948) and the intuition behind it is easily grasped. While the control variables can be partialled-out for the purpose of point estimation, an estimator of the associated regression slopes is still needed to form the residuals that serve to construct the covariance-matrix estimator. The squared residuals, being nonlinear transformations of the nuisance parameters, are biased and inconsistent unless the sampling noise in the latter vanishes.

The problem just described is highly relevant for applied work. Angrist and Hahn (2004) discuss how many control variables arise in program evaluation. Another important example are models for grouped data. There, (possibly multi-way) fixed effects are routinely included to capture unobserved confounding factors at the group level. While dealing with fixed effects in the linear regression model is well understood the failure of the robust covariance-matrix estimator was only noted recently by Stock and Watson (2008) in the context of one-way regression models for short panel data. Although more difficult to analyze, the problem is equally present in the multi-way setting where the number of observations per group is bounded. Important examples include regressions of test scores on student-, teacher-, and classroom effects (Rockoff 2004, Verdier 2018) as well as the many variations of such regressions to problems with a similar structure.

Cattaneo, Jansson and Newey (2018) showed that (asymptotically) valid inference can be performed by using a bias-corrected covariance-matrix estimator in the spirit of Hartley, Rao and Kiefer (1969) and Bera, Suprayitno and Premaratne (2002). Their result hinges on a condition that rules out observations with moderate to high leverage, which is required

for their estimator of the covariance matrix to exist. One implication of their condition is that $\limsup_n q_n/n < \frac{1}{2}$, effectively putting an upper bound on the number of covariates that can be accommodated. On the other hand, when it exists, their estimator is attractive as it is asymptotically equivalent to a minimum-norm unbiased estimator in the sense of Rao (1970).

In this paper we present an alternative estimator of the covariance matrix that can deal with settings where observations have high leverage and remains consistent under an asymptotic scheme where $\limsup_n q_n/n < 1$. As mentioned in Young (2016) such cases arise frequently in the social sciences. A typical example concerns grouped data with a relatively small number of observations per group. In such cases, the conditions underlying the consistency result of Cattaneo, Jansson and Newey (2018) typically fail; see Verdier (2018) for an example on value-added models for student-achievement and Kline, Saggio and Sølvsten (2018) for an illustration concerning labor-market sorting with matched employer-employee data.

Our covariance-matrix estimator is similar to the one suggested in Kline, Saggio and Sølvsten (2018, Remark 4), although we provide conditions for it to be applicable in more general settings. We work in the framework of Cattaneo, Jansson and Newey (2018). It applies to an array of different settings and covers models with (vanishing) misspecification error. As such it includes, for example, models for grouped data with many fixed effects as well as the partially-linear model estimated by series approximation; see, for example, Newey (1997, 2009) for theory and applications.

In Section 2 we introduce the framework and present our covariance-matrix estimator. We establish the consistency of this estimator under a set of high-level conditions. Primitive conditions for the partially-linear model, the one-way model for panel data, and a generic regression model with increasing dimension are also given. In Section 3 we present and discuss the results from Monte Carlo experiments and apply our variance estimator to perform inference on the union wage premium. The appendix contains technical details and additional simulation results.

2 Inference with many regressors

2.1 Framework

Consider the linear model

$$y_{i,n} = \boldsymbol{x}_{i,n}'\boldsymbol{\beta}_n + \boldsymbol{w}_{i,n}'\boldsymbol{\gamma}_n + u_{i,n}, \qquad i = 1, \ldots, n, \tag{2.1}$$

where $y_{i,n}$ is a scalar outcome, $\boldsymbol{x}_{i,n}$ is a vector of regressors of fixed dimension r, $\boldsymbol{w}_{i,n}$ is a vector of covariates whose dimension, q_n, may grow with n, and $u_{i,n}$ is an unobserved error term. Our aim is to perform asymptotically-valid inference on $\boldsymbol{\beta}_n$ that is robust to (conditional) heteroskedasticity, when $\boldsymbol{\gamma}_n$ is high-dimensional, in the sense that q_n is not a vanishing fraction of the sample size. In such a case, the nuisance parameter $\boldsymbol{\gamma}_n$ is not consistently estimable (Huber, 1973).

The (ordinary) least-squares estimator of $\boldsymbol{\beta}_n$ is

$$\hat{\boldsymbol{\beta}}_n := \left(\sum_{i=1}^n \hat{\boldsymbol{v}}_{i,n}\hat{\boldsymbol{v}}_{i,n}'\right)^{-1}\left(\sum_{i=1}^n \hat{\boldsymbol{v}}_{i,n}y_{i,n}\right),$$

where

$$\hat{\boldsymbol{v}}_{i,n} := \sum_{j=1}^n (\boldsymbol{M}_n)_{i,j}\,\boldsymbol{x}_{j,n}, \qquad (\boldsymbol{M}_n)_{i,j} := \{i = j\} - \boldsymbol{w}_{i,n}'\left(\sum_{k=1}^n \boldsymbol{w}_{k,n}\boldsymbol{w}_{k,n}'\right)^{-1}\boldsymbol{w}_{j,n},$$

and $\{\cdot\}$ denotes the indicator function. We will provide an inference result based on the limit distribution of $\hat{\boldsymbol{\beta}}_n$. We begin by stating a set of high-level conditions that guarantee this distribution to be Gaussian that cover the case where $q_n/n \nrightarrow 0$ as $n \to \infty$. Our point of departure for this is Cattaneo, Jansson and Newey (2018, Theorem 1).

Let $\mathcal{X}_n := (\boldsymbol{x}_{1,n}, \ldots, \boldsymbol{x}_{n,n})$ and let $\mathcal{W}_n$ denote a collection of random variables such that $\mathbb{E}(\boldsymbol{w}_{i,n}|\mathcal{W}_n) = \boldsymbol{w}_{i,n}$. We introduce

$$\varepsilon_{i,n} := u_{i,n} - e_{i,n}, \qquad \boldsymbol{V}_{i,n} := \boldsymbol{x}_{i,n} - \mathbb{E}(\boldsymbol{x}_{i,n}|\mathcal{W}_n),$$

where $e_{i,n} := \mathbb{E}(u_{i,n}|\mathcal{X}_n, \mathcal{W}_n)$, to state our first assumption. We use $|\cdot|$ to denote the cardinality of a set.

Assumption 1 (Sampling)**.** *The errors $\varepsilon_{i,n}$ are independent across i conditional on $\mathcal{X}_n$ and $\mathcal{W}_n$, and the collections $\{\varepsilon_{i,n}, \boldsymbol{V}_{i,n} : i \in N_g\}$ are independent across g conditional on $\mathcal{W}_n$, where $\{N_1, \ldots, N_{G_n}\}$ represents a partition of $\{1, \ldots, n\}$ into G_n sets such that $\max_g |N_g| = O(1)$.*

This assumption covers random sampling as well as stratified random sampling, where dependence within the strata is allowed. for example. The latter is useful as it covers short panel data.

The second assumption contains regularity conditions. We let

$$\sigma_{i,n}^2 := \mathbb{E}(\varepsilon_{i,n}^2 | \mathcal{X}_n, \mathcal{W}_n), \qquad \tilde{\boldsymbol{V}}_{i,n} := \sum_{j=1}^{n} (\boldsymbol{M}_n)_{i,j} \boldsymbol{V}_{j,n},$$

denote the Euclidean and Frobenius norms by $\|\cdot\|$, and write $\lambda_{\min}(\cdot)$ for the minimum eigenvalue of its argument.

Assumption 2 (Design)**.** *With probability approaching one $\sum_{i=1}^n \boldsymbol{w}_{i,n} \boldsymbol{w}_{i,n}'$ has full rank,*

$$\max_i \left(\mathbb{E}(\varepsilon_{i,n}^4 | \mathcal{X}_n, \mathcal{W}_n) + \frac{1}{\sigma_{i,n}^2} + \mathbb{E}(\|\boldsymbol{V}_{i,n}\|^4 | \mathcal{W}_n) \right) + \frac{1}{\lambda_{\min}\left(\frac{\sum_{i=1}^n \mathbb{E}(\tilde{\boldsymbol{V}}_{i,n} \tilde{\boldsymbol{V}}_{i,n}' | \mathcal{W}_n)}{n} \right)} = O_p(1),$$

and $\limsup_n q_n/n < 1$.

The rank condition on the design matrix $\sum_{i=1}^n \boldsymbol{w}_{i,n} \boldsymbol{w}_{i,n}'$ is standard. Furthermore, given that the slope coefficients on $\boldsymbol{w}_{i,n}$ are not of direct interest to us, dropping any covariates that are (perfectly) collinear is not an issue. The second condition contains conventional moment conditions. The third condition, finally, allows for q_n to grow at the same rate as the sample size.

Our setting covers the setting where the regression in (2.1) is a linear-in-parameters mean-square approximation to the conditional expectation $\mu_{i,n} := \mathbb{E}(y_{i,n} | \mathcal{X}_n, \mathcal{W}_n)$, in the sense that we allow that $e_{i,n} \neq 0$. The third assumption contains conditions on how fast such an approximation should imprrove. They are expressed in terms of the two constants

$$\varrho_n := \frac{\sum_{i=1}^n \mathbb{E}(e_{i,n}^2)}{n}, \qquad \rho_n := \frac{\sum_{i=1}^n \mathbb{E}(\mathbb{E}(e_{i,n} | \mathcal{W}_n)^2)}{n}.$$

The assumption also contains a similar restriction on how well $\boldsymbol{V}_{i,n}$ can be approximated by

$$\boldsymbol{v}_{i,n} := \boldsymbol{x}_{i,n} - \left(\sum_{j=1}^{n} \mathbb{E}(\boldsymbol{x}_{j,n}\boldsymbol{w}_{j,n}')\right)\left(\sum_{j=1}^{n} \mathbb{E}(\boldsymbol{w}_{j,n}\boldsymbol{w}_{j,n}')\right)^{-1} \boldsymbol{w}_{i,n},$$

the deviation of $\boldsymbol{x}_{i,n}$ from its population linear projection. This is expressed using the constant

$$\chi_n := \frac{\sum_{i=1}^{n} \mathbb{E}(\|\boldsymbol{Q}_{i,n}\|^2)}{n},$$

where $\boldsymbol{Q}_{i,n} := \mathbb{E}(\boldsymbol{v}_{i,n}|\mathcal{W}_n)$.

Assumption 3 (Approximations)**.** *$\chi_n = O(1)$, $\varrho_n + n(\varrho_n - \rho_n) + n\chi_n\varrho_n = o(1)$, and $\max_i \|\hat{\boldsymbol{v}}_{i,n}\|/\sqrt{n} = o_p(1)$.*

The last part of this assumption is a high-level negligibility condition on the residuals from the auxiliary regression of $\boldsymbol{x}_{i,n}$ on the covariates $\boldsymbol{w}_{i,n}$. Given that maximal leverage in that regression satisfies

$$\max_i \left(\hat{\boldsymbol{v}}_{i,n}' \left(\sum_{j=1}^{n} \hat{\boldsymbol{v}}_{j,n}\hat{\boldsymbol{v}}_{j,n}'\right)^{-1} \hat{\boldsymbol{v}}_{i,n}\right) \leq \max_i \frac{\|\hat{\boldsymbol{v}}_{i,n}\|^2}{n} \left\| \left(\frac{\sum_{j=1}^{n} \hat{\boldsymbol{v}}_{j,n}\hat{\boldsymbol{v}}_{j,n}'}{n}\right)^{-1} \right\|^2$$

and should vanish in large sample for $\hat{\boldsymbol{\beta}}_n$ to be consistent (Huber, 1973), this requirement appears close to minimal.

Apart from the first requirement in Assumption 1, where they require only the absence of (conditional) correlation and we impose full independence, Assumptions 1–3 co-incide with Assumptions 1–3 in Cattaneo, Jansson and Newey (2018). Consequently, an application of their Theorem 1 yields

$$\boldsymbol{\Omega}_n^{-1/2}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n) \xrightarrow{d} \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{I}_r) \tag{2.2}$$

as $n \to \infty$, where

$$\boldsymbol{\Omega}_n := \left(\sum_{i=1}^{n} \hat{\boldsymbol{v}}_{i,n}\hat{\boldsymbol{v}}_{i,n}'\right)^{-1} \left(\sum_{i=1}^{n} \hat{\boldsymbol{v}}_{i,n}\hat{\boldsymbol{v}}_{i,n}'\,\sigma_{i,n}^2\right) \left(\sum_{i=1}^{n} \hat{\boldsymbol{v}}_{i,n}\hat{\boldsymbol{v}}_{i,n}'\right)^{-1},$$

and $\boldsymbol{I}_r$ denotes the $r \times r$ identity matrix.

2.2 Variance estimation

Constructing confidence intervals and test statistics based on (2.2) requires a consistent estimator of $\boldsymbol{\Omega}_n$, and thus of

$$\boldsymbol{\Sigma}_n := \sum_{i=1}^{n} \hat{\boldsymbol{v}}_{i,n}\hat{\boldsymbol{v}}_{i,n}' \, \sigma_{i,n}^2.$$

When $q_n/n \nrightarrow 0$ and the errors are permitted to be heteroskedastic, this is non-trivial. To appreciate the problem, consider the conventional covariance-matrix estimator as proposed by Eicker (1963, 1967) and White (1980). This estimator uses

$$\hat{\boldsymbol{\Sigma}}_n := \sum_{i=1}^{n} \hat{\boldsymbol{v}}_{i,n}\hat{\boldsymbol{v}}_{i,n}' \, \hat{u}_{i,n}^2,$$

where $\hat{u}_{i,n} := \sum_{j=1}^{n} (\boldsymbol{M}_n)_{i,j} \, (y_{i,n} - \boldsymbol{x}_{i,n}'\hat{\boldsymbol{\beta}}_n)$ are the least-squares residuals. The estimator $\hat{\boldsymbol{\Sigma}}_n$ is well known to be (conditionally) biased. The bias arises from the sampling noise in the least-squares estimator and can be severe, especially when the regression design contains observations with high leverage (Chesher and Jewitt, 1987). Unless $q_n/n \to 0$, some observations remain influential as $n \to \infty$, in the sense that their leverage does not vanish. This causes the bias in $\hat{\boldsymbol{\Sigma}}_n$ to persist in large samples, and renders it inconsistent. The same is true for the alternatives to $\hat{\boldsymbol{\Sigma}}_n$ that have been proposed in the literature (as reviewed in Long and Ervin 2000) to alleviate its small-sample bias (see Cattaneo, Jansson and Newey 2018, Theorem 3).

Cattaneo, Jansson and Newey (2018) proposed the estimator

$$\grave{\boldsymbol{\Sigma}}_n := \sum_{i=1}^{n} \hat{\boldsymbol{v}}_{i,n}\hat{\boldsymbol{v}}_{i,n}' \left(\sum_{j=1}^{n} ((\boldsymbol{M}_n * \boldsymbol{M}_n)^{-1})_{i,j} \, \hat{u}_{j,n}^2 \right),$$

where $\boldsymbol{M}_n * \boldsymbol{M}_n$ denotes the elementwise product of the matrix $\boldsymbol{M}_n$. Their proposal has its origins in work by Hartley, Rao and Kiefer (1969) and Rao (1970) and can be motivated through an (asymptotic) bias calculation (see also Bera, Suprayitno and Premaratne 2002 and Anatolyev 2018). For $\grave{\boldsymbol{\Sigma}}_n$ to be well defined, $\boldsymbol{M}_n * \boldsymbol{M}_n$ needs to be positive definite. Necessary and sufficient conditions for this are stated in Mallela (1972) but these are neither simple nor intuitive (Horn, Horn and Duncan, 1975). As already noted by Horn and Horn

(1975), a sufficient condition is that

$$\min_i (\boldsymbol{M}_n)_{i,i} > \frac{1}{2}.$$

Depending on the problem at hand, this condition may also be necessary; an example is the one-way panel model. Cattaneo, Jansson and Newey (2018, Theorem 4) show that $\grave{\boldsymbol{\Sigma}}_n$ is consistent if

$$\Pr\left(\min_i (\boldsymbol{M}_n)_{i,i} > \frac{1}{2}\right) \to 1, \qquad \frac{1}{\min_i (\boldsymbol{M}_n)_{i,i} - \frac{1}{2}} = O_p(1). \tag{2.3}$$

Note that, because $\sum_{i=1}^n (\boldsymbol{M}_n)_{i,i} = q_n$, we have $\min_i (\boldsymbol{M}_n)_{i,i} \leq q_n/n$, and so these conditions require that $\limsup q_n/n < \frac{1}{2}$.

The low-leverage requirement can be problematic. As observed by Young (2016), many applications in the social sciences feature a highly unbalanced regressor design and values of $\min_i (\boldsymbol{M}_n)_{i,i}$ that fall far below the threshold of Horn and Horn (1975). A useful alternative estimator is

$$\acute{\boldsymbol{\Sigma}}_n := \sum_{i=1}^n \hat{\boldsymbol{v}}_{i,n} \hat{\boldsymbol{v}}_{i,n}' \left(y_{i,n} \acute{u}_{i,n}\right), \qquad \acute{u}_{i,n} := \frac{\hat{u}_{i,n}}{(\boldsymbol{M}_n)_{i,i}}.$$

As stated, this estimator is well defined provided that

$$\min_i (\boldsymbol{M}_n)_{i,i} > 0.$$

Notice that $(\boldsymbol{M}_n)_{i,i} = 0$ means that the model reserves a parameter for this observation. This implies that the auxiliary regression of the regressors of interest on the other covariates yields a perfect prediction, in the sense that $\hat{\boldsymbol{v}}_{i,n} = \boldsymbol{0}$. Consequently, such an observation does not carry information on $\boldsymbol{\beta}_n$ and can be dropped. It does not affect the least-squares estimator $\hat{\boldsymbol{\beta}}_n$ and does not contribute to its covariance matrix $\boldsymbol{\Omega}_n$. This is important as perfect prediction of this form arises frequently in empirical work when many dummy variables are included.

To explain the form of $\acute{\boldsymbol{\Sigma}}_n$ it is useful to consider the case where $e_{i,n} = 0$, so that their is no misspecification bias in (2.1). A straightforward calculation shows that, in that case,

$$\mathbb{E}(y_{i,n} \hat{u}_{i,n} | \mathcal{X}_n, \mathcal{W}_n) = (\boldsymbol{M}_n)_{i,i}\, \sigma_{i,n}^2 + o_p(1),$$

and so $\mathbb{E}(\acute{\boldsymbol{\Sigma}}_n|\mathcal{X}_n, \mathcal{W}_n) = \boldsymbol{\Sigma}_n + o_p(n)$. When $e_{i,n} \neq 0$, our estimator of $\sigma_{i,n}^2$ has an additional bias term but under regularity conditions this term, too, will be asymptotically negligible. Note that $\acute{\boldsymbol{\Sigma}}_n$ is similar to the leave-one-out covariance estimator suggested independently by Kline, Saggio and Sølvsten (2018, Remark 4), although it is targeted to the setting with many nuisance parameters. They provide a consistency result for their estimator under substantially stronger restrictions than we do here. An interesting feature of their estimator is that it has an interpretation as an estimator based on sample splitting. Such estimators are a useful device for bias correction in a variety of context; see, for example, Chernozhukov et al. (2018) and Newey and Robins (2018).

Additional conditions are needed to show that $\acute{\boldsymbol{\Sigma}}_n$ is consistent. We let

$$\tilde{\boldsymbol{Q}}_{i,n} := \sum_{j=1}^{n} (\boldsymbol{M}_n)_{i,j} \boldsymbol{Q}_{i,n}.$$

in the following assumption.

Assumption 4 (Variance estimation)**.** $n\varrho_n = O(1)$, $\Pr(\min_i (\boldsymbol{M}_n)_{i,i} > 0) \to 1$,

$$\frac{1}{\min_i (\boldsymbol{M}_n)_{i,i}} = O_p(1), \qquad \frac{\sum_{i=1}^n \|\tilde{\boldsymbol{Q}}_{i,n}\|^4}{n} = O_p(1),$$

and $\max_i \|\mu_{i,n}\|/\sqrt{n} = o_p(1)$.

The first part of Assumption 4 is relevant only when (2.1) is misspecified, in the sense that $e_{i,n} \neq 0$. In that case it is a strengthening of Assumption 3 only when $\chi_n = o(1)$. The conditions on the diagonal entries of the projection matrix are a considerable relaxation of the requirements in (2.3). Providing primitive conditions for them in great generality is difficult. However, when $\boldsymbol{w}_{i,n} \sim \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{I}_{q_n})$ they follow under $\limsup_n q_n/n < 1$ as stated in Assumption 2 in the same way as in Cattaneo, Jansson and Newey (2018). In the one-way panel model they hold automatically while Verdier (2018) gives sufficient conditions for them to be satisfied in the two-way model. To understand why the last part of Assumption 4 is needed, note that the (conditional) variance of $\acute{u}_{i,n}$ depends on $\mu_{i,n}^2$. The requirement that $\max_i \mu_{i,n}^2 = o_p(n)$ allows to control the variance of $\acute{\boldsymbol{\Sigma}}_n$ and is needed to show consistency. Weak moment requirements typically suffice for this condition

to be satisfied. The condition on $\tilde{\boldsymbol{Q}}_{i,n}$ is used in concordance with the condition on $\mu_{i,n}$. One simple sufficient condition for it is that $n\chi_n = O(1)$ but it can also be satisfied when $\chi_n = O(1)$. The condition can be dispensed with altogether if we are willing to impose that $\max_i \mu_{i,n}^2 = O_p(1)$, which can be a plausible restriction in settings where all regressors are discrete, for example.

A restriction on the magnitude of $\mu_{i,n}$ is not needed for $\grave{\boldsymbol{\Sigma}}_n$. The interpretation of $\acute{\boldsymbol{\Sigma}}_n$ as an (approximate) cross-fit estimator as in Newey and Robins (2018) and Kline, Saggio and Sølvsten (2018) is useful to highlight an apparent tension between invariance of a covariance-matrix estimator to changes in $\mu_{i,n}$ and the possibility for it to yield valid inference when $\limsup_n q_n/n > \frac{1}{2}$. Moreover, again assuming for simplicity that $e_{i,n} = 0$, a (conditionally) unbiased cross-fit estimator of $\sigma_{i,n}^2$ that is invariant is necessarily of the form $\dot{u}_{i,n}\ddot{u}_{i,n}$, where $\dot{u}_{i,n}$ and $\ddot{u}_{i,n}$ are least-squares residuals obtained from two (conditionally) independent subsamples, each excluding the ith observation. Of course, each of these auxiliary least-squares regression can be based on at most $\lfloor (n-1)/2 \rfloor$ observations, implying that such an approach cannot accommodate settings where $q_n/n > \frac{1}{2}$. Our approach circumvents the need for two independent estimators of $u_{i,n}$ by using the level of the outcome variable, $y_{i,n}$ as a proxy for $u_{i,n}$. This allows to deal with cases where q_n/n is large but makes the variance estimator sensitive to $\mu_{i,n}$.

The following theorem, when combined with (2.2), permits valid inference when the regressor design contains observations with high leverage.

Theorem 1 (Inference)**.** *Let Assumptions 1–4 hold. Then*

$$\boldsymbol{\Sigma}_n^{-1}\acute{\boldsymbol{\Sigma}}_n \xrightarrow{p} \boldsymbol{I}_r$$

as $n \to \infty$.

Proof. The proof is given in the appendix. □

2.3 Examples

We next provide more primitive conditions in three special cases that fit out general setup. We focus on sufficient conditions for Assumption 4. Cattaneo, Jansson and Newey (2018)

give such conditions for the other assumptions— and, notably, for Assumption 3—to hold. The calculations underlying this section are collected in the appendix.

Partially-linear model Suppose that observations on $(y_i, \boldsymbol{x}_i, \boldsymbol{z}_i)$ are independent and identically distributed. The partially-linear regression model states that

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \varphi(\boldsymbol{z}_i) + \varepsilon_i, \qquad \mathbb{E}(\varepsilon_i|\boldsymbol{x}, \boldsymbol{z}_i) = 0,$$

for an unknown function φ. A series approximation of order κ_n of $\varphi(\boldsymbol{z}_i)$ takes the form $\boldsymbol{w}_{i,n}'\boldsymbol{\gamma}_n$ for $\boldsymbol{w}_{i,n} = (p_1(\boldsymbol{z}_i), \ldots, p_{\kappa_n}(\boldsymbol{z}_i))'$ and $\{p_1, \ldots, p_{\kappa_n}\}$ a collection of basis functions such as orthogonal polynomials. Our estimator of $\boldsymbol{\beta}$ is the least-squares estimator of $\boldsymbol{\beta}_n$ in

$$y_{i,n} = \boldsymbol{x}_i'\boldsymbol{\beta}_n + \boldsymbol{w}_{i,n}'\boldsymbol{\gamma}_n + u_{i,n}, \qquad u_{i,n} = \varepsilon_i + \varphi(\boldsymbol{z}_i) - \boldsymbol{w}_{i,n}'\boldsymbol{\gamma}_n.$$

Note that $\mathbb{E}(u_{i,n}|\boldsymbol{x}_i, \boldsymbol{z}_i) \neq 0$, in general. Here,

$$\varrho_n = \min_{\boldsymbol{\gamma}} \mathbb{E}\left(\|\varphi(\boldsymbol{z}_i) - \boldsymbol{w}_{i,n}'\boldsymbol{\gamma}\|^2\right) \qquad \chi_n = \min_{\boldsymbol{\delta}} \mathbb{E}\left(\|\mathbb{E}(\boldsymbol{x}_{i,n}|\boldsymbol{w}_{i,n}) - \boldsymbol{\delta}'\boldsymbol{w}_{i,n}\|^2\right),$$

and standard smoothness conditions on the functions being approximated (Newey, 1997) imply that $n\varrho_n = o(1)$ and that $n\chi_n = O(1)$, which yield the first and the fourth condition of Assumption 4, respectively. The requirement that $\max_i \|\mu_{i,n}\|/\sqrt{n} = o_p(1)$ is satisfied if the moment conditions $\mathbb{E}(\|\boldsymbol{x}_i\|^{2+\theta}) = O(1)$ and $\mathbb{E}(\|\varphi(\boldsymbol{z}_i)\|^{2+\theta}) = O(1)$ hold. These conditions are only slightly stronger than the ones imposed in Cattaneo, Jansson and Newey (2018), but can deal with series approximations of a much larger dimension. This can be useful when the dimension of $\boldsymbol{z}_i$ is large, so that many terms are included in the approximation even for small κ_n, or when the underlying functions are not very smooth, so that a large κ_n needs to be used to control bias.

One-way model for panel data For double-indexed data $(y_{(g,m)}, \boldsymbol{x}_{(g,m)})$, the fixed-effect model is

$$y_{(g,m)} = \boldsymbol{x}_{(g,m)}'\boldsymbol{\beta} + \alpha_g + \varepsilon_{(g,m)}, \qquad g = 1, \ldots, G, \quad m = 1 \ldots, M,$$

where α_g is a group-specific intercept. Observations are assumed independent across groups but may be dependent within each group. We assume $\mathbb{E}(\varepsilon_{(g,m)}|\boldsymbol{x}_{(g,1)},\ldots,\boldsymbol{x}_{(g,M)})=0$. The usual asymptotic approximation here has $G\to\infty$ with $M=O(1)$, which fits Assumption 1. In that case, the number of fixed effects grows at the same rate as the sample size. Moreover, $n=G\times M$ and $q_n=G$ so that $q_n/n=1/M$, which does not vanish. The fixed-effect estimator equals the ordinary least-squares estimator of $y_{(g,m)}$ on $\boldsymbol{x}_{(g,m)}$ and G dummy variables that indicate group membership. Here, our Assumption 4 holds provided that

$$\frac{\sum_{g=1}^{G}\alpha_g^{2+\theta}}{G}=O(1)$$

and $\max_g\max_m\mathbb{E}(\|\boldsymbol{x}_{(g,m)}\|^{2+\theta})=O(1)$ for some $\theta>0$. Given that Cattaneo, Jansson and Newey (2018) impose that $\max_g\max_m\mathbb{E}(\|\boldsymbol{x}_{(g,m)}\|^4)=O(1)$ to validate Assumption 2, the moment condition on the fixed effects is the only additional, and arguably quite weak, requirement needed for our variance estimator to work.

Linear model with increasing dimension. Finally, consider the regression model that takes (2.1) as the data generating process for independent and identically distributed observations $(y_{i,n},\boldsymbol{x}_{i,n},\boldsymbol{w}_{i,n})$, i.e.,

$$y_{i,n}=\boldsymbol{x}_{i,n}'\boldsymbol{\beta}_n+\boldsymbol{w}_{i,n}'\boldsymbol{\gamma}_n+u_{i,n},\qquad n\,\mathbb{E}(\|\mathbb{E}(u_{i,n}|\boldsymbol{x}_{i,n},\boldsymbol{w}_{i,n})\|^2)=n\varrho_n=o(1),$$

as in Cattaneo, Jansson and Newey (2018). The generic nature of this model makes it difficult to specify a single set of simple sufficient conditions for our Assumption 4 to hold. First consider the requirement that $\sum_{i=1}^{n}\|\tilde{\boldsymbol{Q}}_{i,n}\|^4=O_p(n)$. This is satisfied provided that

$$n\,\chi_n=n\min_{\boldsymbol{\delta}}\mathbb{E}\left(\|\mathbb{E}(\boldsymbol{x}_{i,n}|\boldsymbol{w}_{i,n})-\boldsymbol{\delta}'\boldsymbol{w}_{i,n}\|^2\right)=o(1).$$

This will be the case, for example, when $\boldsymbol{w}_{i,n}$ is discrete and a saturated regression model is used. It will also hold under smoothness conditions on the function $\mathbb{E}(\boldsymbol{x}_{i,n}|\boldsymbol{w}_{i,n})$ when the $\boldsymbol{w}_{i,n}$ are approximating functions, as discussed above. Alternatively, in cases where $\chi_n=O(1)$, the condition is satisfied when a condition is imposed on the growth rate of

$n_{i,n} := \sum_{j=1}^{n} \{(\boldsymbol{M}_n)_{i,j} \neq 0\}$. One version of this is the requirement that

$$\max_i n_{i,n} = O_p\left(n^{\frac{1}{4}\left(1-\frac{4}{\theta-1}\right)}\right)$$

together with $\min_{\boldsymbol{\delta}} \mathbb{E}\left(\|\mathbb{E}(\boldsymbol{x}_{i,n}|\boldsymbol{w}_{i,n}) - \boldsymbol{\delta}'\boldsymbol{w}_{i,n}\|^{\theta}\right) = O(1)$, where $\theta \geq 5$. As more moments exist, the rate approaches $n^{1/4}$. Cattaneo, Jansson and Newey (2018) use a similar sparsity condition to validate Assumption 3 when χ_n does not vanish. This rate is the faster $n^{1/3}$, however, and only requires that $\min_{\boldsymbol{\delta}} \mathbb{E}\left(\|\mathbb{E}(\boldsymbol{x}_{i,n}|\boldsymbol{w}_{i,n}) - \boldsymbol{\delta}'\boldsymbol{w}_{i,n}\|^{4}\right) = O(1)$. Finally, for $\max_i \|\mu_{i,n}\|/\sqrt{n} = o_p(1)$ to hold we again impose that $\mathbb{E}(\|\boldsymbol{x}_{i,n}\|^{2+\theta}) = O(1)$ for some $\theta > 0$. When $\boldsymbol{w}_{i,n}$ are approximating functions a similar moment condition on the function being approximated will again suffice. In the case where $\boldsymbol{w}_{i,n}$ are just many control variables included in the regression we can again use a sparsity condition. Let $\kappa_{i,n}$ denote the number of nuisance parameters on which $y_{i,n}$ depends. Then one alternative sufficient condition is that

$$\max_i \kappa_{i,n} = O_p(n^{\frac{1}{2}\frac{\theta}{2+\theta}}),$$

for some $\theta > 0$, together with the assumption that the entries of $\boldsymbol{w}_{i,n}$ have $2+\theta$ moments. A condition on $\kappa_{i,n}$ is different from a condition on $n_{i,n}$, as it only pertains the regression of $y_{i,n}$ on $\boldsymbol{x}_{i,n}$ and $\boldsymbol{w}_{i,n}$ and does not restrict the auxiliary regression of $\boldsymbol{x}_{i,n}$ on $\boldsymbol{w}_{i,n}$. The rate on $\max_i \kappa_{i,n}$ is also generally faster than the one imposed on $\max_i n_{i,n}$ above. When all the covariates are normally distributed or have bounded support, for example, we can allow for $\max_i \kappa_{i,n} = O_p(\sqrt{n})$.

3 Numerical illustrations

3.1 Simulations

We present numerical results for a setup taken from Cattaneo, Jansson and Newey (2018). Data are generated as

$$y_{i,n} = x_{i,n}\beta + \boldsymbol{w}_{i,n}'\boldsymbol{\gamma}_n + \varepsilon_{i,n},$$

where $x_{i,n} \sim$ i.i.d. $\boldsymbol{N}(0,1)$, $\boldsymbol{w}_{i,n}$ contains a constant term and a collection of $q_n - 1$ zero/one dummy variables, and $\varepsilon_{i,n} \sim$ i.i.d. $\boldsymbol{N}(0,1)$. The dummy variables are drawn independently with success probability π and $\boldsymbol{\gamma}_n = \mathbf{0}$. The sample size was fixed to $n = 700$ throughout and we considered $q_n \in \{1, 71, 141, 211, 281, 351, 421, 491, 561, 631\}$. All statistics reported below were computed over $10,000$ Monte Carlo replications and all the variables were redrawn in each replication.

The baseline design has $\beta = 1$ and $\pi = .02$. This corresponds to the design in Cattaneo, Jansson and Newey (2018) (The description of the simulation design in Cattaneo, Jansson and Newey 2018, p. 1358 erroneously states that $\pi = .0062$.) Here, each dummy variable takes on the value one for about 14 observations. We also provide results for two deviations from this design. The first deviation is a more sparse design where $\beta = 1$ is maintained and π is reduced to .01, leading to the dummy switching on for only 7 observations in each replication, on average. The second deviation from the baseline maintains $\pi = .02$ but sets $\beta = 2$.

Table 1 provides the results for the baseline design. It contains the mean and standard deviation of the (estimated) standard error of $\hat{\beta}_n$ as computed using $\hat{\Sigma}_n$, $\grave{\Sigma}_n$, and $\acute{\Sigma}_n$—and indicated as $\sqrt{\hat{\Omega}_n}$, $\sqrt{\grave{\Omega}_n}$, and $\sqrt{\acute{\Omega}_n}$—together with the actual standard deviation of $\hat{\beta}_n$ as computed over the Monte Carlo replications, $\sqrt{\Omega_n}$. For each of the three standard errors, the table also gives the rejection frequency of a two-sided t-test with nominal size .05 and the average length of the implied 95% confidence interval. Both $\hat{\Sigma}_n$ and $\acute{\Sigma}_n$ always exist. In replications where $\grave{\Sigma}_n$ did not exist it was replaced with $\hat{\Sigma}_n$.

The simulation results confirm all of our theoretical findings. They verify the poor performance of the conventional standard error when q_n/n is not small. In such cases it substantially underestimates the true sampling variability in the least-squares estimator, leading to large over-rejection under the null of the associated t-test. Using $\grave{\Sigma}_n$ substantially alleviates the bias and leads to much more reliable inference when $q_n/n < \frac{1}{2}$. As q_n/n increases beyond that cut off the performance of $\sqrt{\grave{\Omega}_n}$ worsens due to $\grave{\Sigma}_n$ failing to exist. On the other hand, $\acute{\Sigma}_n$ provides a nearly unbiased estimator of $\sqrt{\Omega_n}$ for all values of q_n considered. Consequently, it remains a valid tool for inference in settings where $q_n/n > \frac{1}{2}$.

Table 1: $\beta = 1$ and $\pi = .02$

q_n	1	71	141	211	281	351	421	491	561	631
q_n/n	0.0014	0.1014	0.2014	0.3014	0.4014	0.5014	0.6014	0.7014	0.8014	0.9014
					Mean					
$\sqrt{\Omega}_n$	0.0387	0.0401	0.0422	0.0455	0.0489	0.0538	0.0598	0.0696	0.0850	0.1223
$\sqrt{\hat{\Omega}}_n$	0.0377	0.0378	0.0380	0.0381	0.0383	0.0385	0.0386	0.0389	0.0391	0.0392
$\sqrt{\grave{\Omega}}_n$	0.0377	0.0397	0.0421	0.0449	0.0483	0.0526	0.0583	0.0663	0.0784	0.0891
$\sqrt{\acute{\Omega}}_n$	0.0376	0.0397	0.0421	0.045	0.0486	0.0533	0.0594	0.0032	0.0837	0.1175
					Standard deviation					
$\sqrt{\hat{\Omega}}_n$	0.0020	0.0021	0.0022	0.0023	0.0024	0.0026	0.0028	0.0032	0.0038	0.0052
$\sqrt{\grave{\Omega}}_n$	0.0020	0.0023	0.0027	0.0033	0.0041	0.0055	0.0074	0.0113	0.0195	0.0494
$\sqrt{\acute{\Omega}}_n$	0.0027	0.0029	0.0032	0.0036	0.0040	0.0048	0.0057	0.0074	0.0110	0.0212
					Rejection frequency (nominal size is 5%)					
$\sqrt{\hat{\Omega}}_n$	0.0528	0.0626	0.0791	0.0994	0.1241	0.1597	0.2063	0.2786	0.3741	0.5224
$\sqrt{\grave{\Omega}}_n$	0.0525	0.0499	0.0528	0.0546	0.0525	0.0577	0.0641	0.0721	0.1014	0.1983
$\sqrt{\acute{\Omega}}_n$	0.0539	0.0515	0.0525	0.0546	0.0520	0.0536	0.0572	0.0544	0.0598	0.0701
					Average width of confidence interval					
$\sqrt{\hat{\Omega}}_n$	0.1477	0.1483	0.1489	0.1495	0.1500	0.1509	0.1514	0.1525	0.1531	0.1539
$\sqrt{\grave{\Omega}}_n$	0.1478	0.1557	0.1650	0.1760	0.1893	0.2064	0.2284	0.2600	0.3073	0.3493
$\sqrt{\acute{\Omega}}_n$	0.1476	0.1557	0.1651	0.1765	0.1906	0.2088	0.2329	0.2685	0.3282	0.4605

The ability to perform heteroskedasticity-robust inference in the high-dimensional setting comes at the cost of using a more noisy variance estimator. Indeed, both $\grave{\Sigma}_n$ and $\acute{\Sigma}_n$ are more variable than $\hat{\Sigma}_n$. When $\sqrt{\grave{\Omega}_n}$ exists it shows slightly less volatility than does $\sqrt{\acute{\Omega}_n}$. This is reflected in the reported standard deviations and in the width of the associated confidence intervals. The observation that the standard deviation of $\sqrt{\grave{\Omega}_n}$ is larger than that of $\sqrt{\acute{\Omega}_n}$ for large values of q_n/n is an artifact of the former defaulting to the much smaller $\sqrt{\hat{\Omega}_n}$ in many replications.

Table 2: $\beta = 1$ and $\pi = .01$

q_n	1	71	141	211	281	351	421	491	561	631
q_n/n	0.0014	0.1014	0.2014	0.3014	0.4014	0.5014	0.6014	0.7014	0.8014	0.9014
					Mean					
$\sqrt{\Omega_n}$	0.0397	0.0393	0.0422	0.0446	0.0490	0.0535	0.0603	0.0687	0.0857	0.1218
$\sqrt{\hat{\Omega}_n}$	0.0377	0.0381	0.0384	0.0388	0.0393	0.0399	0.0406	0.0414	0.0427	0.0452
$\sqrt{\grave{\Omega}_n}$	0.0377	0.0383	0.0385	0.0388	0.0393	0.0399	0.0406	0.0414	0.0427	0.0452
$\sqrt{\acute{\Omega}_n}$	0.0377	0.0398	0.0422	0.0450	0.0486	0.0533	0.0595	0.0686	0.0838	0.1174
					Standard deviation					
$\sqrt{\hat{\Omega}_n}$	0.0020	0.0021	0.0022	0.0024	0.0026	0.0029	0.0033	0.0038	0.0049	0.0081
$\sqrt{\grave{\Omega}_n}$	0.0020	0.0022	0.0023	0.0024	0.0026	0.0029	0.0033	0.0038	0.0049	0.0081
$\sqrt{\acute{\Omega}_n}$	0.0026	0.0029	0.0032	0.0036	0.0041	0.0048	0.0058	0.0075	0.0110	0.0215
					Rejection frequency (nominal size is 5%)					
$\sqrt{\hat{\Omega}_n}$	0.0514	0.0592	0.0748	0.0910	0.1181	0.1432	0.1867	0.2389	0.3354	0.4689
$\sqrt{\grave{\Omega}_n}$	0.0513	0.0577	0.0739	0.0909	0.1181	0.1432	0.1867	0.2389	0.3354	0.4689
$\sqrt{\acute{\Omega}_n}$	0.0517	0.0503	0.0520	0.0538	0.0520	0.0514	0.0562	0.0540	0.0617	0.0713
					Average width of confidence interval					
$\sqrt{\hat{\Omega}_n}$	0.1478	0.1492	0.1506	0.1522	0.1539	0.1562	0.1590	0.1623	0.1675	0.1773
$\sqrt{\grave{\Omega}_n}$	0.1479	0.1502	0.1509	0.1523	0.1539	0.1562	0.1590	0.1623	0.1675	0.1773
$\sqrt{\acute{\Omega}_n}$	0.1479	0.1559	0.1652	0.1765	0.1904	0.2087	0.2331	0.2691	0.3284	0.4600

Table 2 reports results for the more sparse regressor design. This is an interesting deviation from the baseline specification because, here, the leverage of the observations is larger. The chief consequence is that $\grave{\Sigma}_n$ fails to exists very frequently, even when q_n/n

is small. This is apparent on inspecting the differences between the rows for $\sqrt{\grave{\Omega}_n}$ and $\sqrt{\hat{\Omega}_n}$. Moreover, $\grave{\Sigma}_n$ does not exist in any of the Monte Carlo replications for all $q_n > 141$. Consequently, in this sparse setting the only available estimator is $\acute{\Sigma}_n$. Its performance is virtually unaffected by the increased sparsity of the regressor design, continuing to provide reliable inference for all values of q_n/n.

Table 3: $\beta = 2$ and $\pi = .02$

q_n	1	71	141	211	281	351	421	491	561	631
q_n/n	0.0014	0.1014	0.2014	0.3014	0.4014	0.5014	0.6014	0.7014	0.8014	0.9014
					Mean					
$\sqrt{\Omega}_n$	0.0379	0.0400	0.0424	0.0449	0.0488	0.0536	0.0597	0.0695	0.0850	0.1217
$\sqrt{\hat{\Omega}}_n$	0.0377	0.0379	0.0380	0.0381	0.0383	0.0385	0.0387	0.0389	0.0391	0.0394
$\sqrt{\grave{\Omega}}_n$	0.0377	0.0398	0.0421	0.0449	0.0483	0.0526	0.0582	0.0663	0.0781	0.0890
$\sqrt{\acute{\Omega}}_n$	0.0375	0.0396	0.0420	0.0449	0.0485	0.0530	0.0593	0.0681	0.0831	0.1158
					Standard deviation					
$\sqrt{\hat{\Omega}}_n$	0.0020	0.0021	0.0022	0.0023	0.0024	0.0026	0.0029	0.0032	0.0037	0.0052
$\sqrt{\grave{\Omega}}_n$	0.0020	0.0023	0.0027	0.0033	0.0041	0.0055	0.0076	0.0112	0.0194	0.0496
$\sqrt{\acute{\Omega}}_n$	0.0040	0.0044	0.0048	0.0054	0.0060	0.0070	0.0085	0.0109	0.0154	0.0310
					Rejection frequency (nominal size is 5%)					
$\sqrt{\hat{\Omega}}_n$	0.0515	0.0660	0.0780	0.0957	0.1271	0.1582	0.2079	0.2737	0.3699	0.5279
$\sqrt{\grave{\Omega}}_n$	0.0514	0.0532	0.0520	0.0513	0.0543	0.0579	0.0639	0.0721	0.1010	0.1974
$\sqrt{\acute{\Omega}}_n$	0.0574	0.0567	0.0562	0.0572	0.0593	0.0590	0.0613	0.0656	0.0699	0.0907
					Average width of confidence interval					
$\sqrt{\hat{\Omega}}_n$	0.1478	0.1484	0.1489	0.1494	0.1502	0.1508	0.1515	0.1523	0.1532	0.1545
$\sqrt{\grave{\Omega}}_n$	0.1479	0.1559	0.1651	0.1759	0.1895	0.2064	0.2283	0.2600	0.3063	0.3490
$\sqrt{\acute{\Omega}}_n$	0.1470	0.1553	0.1645	0.1758	0.1900	0.2078	0.2324	0.2669	0.3256	0.4539

Table 3 contains results for the design where we increase the value of β. This is useful to highlight that, while $\grave{\Sigma}_n$ is invariant to this change, $\acute{\Sigma}_n$ is not. The latter continues to provide a nearly-unbiased covariance estimator but is now more volatile. Nonetheless, the associated test statistic continues to be approximately size correct for all the values of q_n/n considered.

The appendix contains simulation results for a partially-linear regression model and for a one-way panel model.

3.2 Empirical example

We next use our variance estimator to infer the union membership premium. The data are a balanced panel on 545 working individuals and span 8 years, giving a total of 4,360 observations. They are taken from Wooldridge (2002) and are available in Stata format as `wagepan.dta` through the publisher's website `http://mitpress.mit.edu/books/`. We estimate the union premium as the coefficient from a least-squares regression of log wages on a dummy for union membership, after partialling-out a set of control variables. The control variables are a full set of individual fixed effects, time dummies, and occupational dummies, together with a set of covariates that are included in levels and as interactions with the time and occupation dummies. The covariates in question are average hours worked per week, dummies for marital status and for poor health, together with quadratic terms in years of experience and age. The resulting equation allows for the impact of these characteristics to vary both over time and across occupation. Here, $r = 1$ and $q_n = 924$. Although $q_n/n < \frac{1}{2}$, the variance estimator of Cattaneo, Jansson and Newey (2018) does not exist in these data. Our point estimate of the union premium is 6.8%. The conventional heteroskedasticity-robust standard error on this point estimate is .0175. Our standard error is .0179 which, in line with our Monte Carlo results, is somewhat higher. This translates into 95% confidence intervals that are equal to $[.0338, .1022]$ and $[.0330, .1030]$, respectively.

4 Conclusion

We have introduced a new heteroskedasticity-robust covariance matrix estimator for linear regression models. The estimator remains consistent under an asymptotic scheme where the number of covariates grows at the same rate as the sample size. Moreover, it can deal with highly unbalanced regressor designs where the alternative estimator of Cattaneo, Jansson and Newey (2018) does not exist. Our estimator is similar to the cross-fit estimator

suggested in Kline, Saggio and Sølvsten (2018, Remark 4) and can be useful as a device to correct for bias more generally. One example here would be Cattaneo, Jansson and Ma (2018), who characterize the bias in two-step estimators that rely on a high-dimensional linear regression in the first step.

A Appendix

A.1 Proof of Theorem 1

We need to show that

$$\frac{\sum_{i=1}^{n} \hat{\boldsymbol{v}}_{i,n}\hat{\boldsymbol{v}}_{i,n}' \left(y_{i,n}\acute{u}_{i,n}\right)}{n} = \frac{\sum_{i=1}^{n} \hat{\boldsymbol{v}}_{i,n}\hat{\boldsymbol{v}}_{i,n}' \, \sigma_{i,n}^2}{n} + o_p(1).$$

As Cattaneo, Jansson and Newey (2018), to ease notation, we set $r = 1$ without loss of generality.

Add and subtract $\varepsilon_{i,n}$ to get

$$\frac{\sum_{i=1}^{n} \hat{v}_{i,n}^2 \left(y_{i,n}\acute{u}_{i,n} - \sigma_{i,n}^2\right)}{n} = \frac{\sum_{i=1}^{n} \hat{v}_{i,n}^2 \left(\varepsilon_{i,n}^2 - \sigma_{i,n}^2\right)}{n} + \frac{\sum_{i=1}^{n} \hat{v}_{i,n}^2 \left(y_{i,n}\acute{u}_{i,n} - \varepsilon_{i,n}^2\right)}{n}. \tag{A.1}$$

Consider the first term on the right-hand side. Because $\sigma_{i,n}^2 = \mathbb{E}(\varepsilon_{i,n}^2 | \mathcal{X}_n, \mathcal{W}_n)$ by definition,

$$\mathbb{E}\left(\frac{\sum_{i=1}^{n} \hat{v}_{i,n}^2 \left(\varepsilon_{i,n}^2 - \sigma_{i,n}^2\right)}{n} \middle| \mathcal{X}_n, \mathcal{W}_n \right) = 0.$$

Next,

$$\begin{aligned}
\mathbb{E}\left(\left(\frac{\sum_{i=1}^{n} \hat{v}_{i,n}^2 \left(\varepsilon_{i,n}^2 - \sigma_{i,n}^2\right)}{n} \right)^2 \middle| \mathcal{X}_n, \mathcal{W}_n \right) &= \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} \hat{v}_{i,n}^2 \, \mathbb{E}(\varepsilon_{i,n}\varepsilon_{j,n} | \mathcal{X}_n, \mathcal{W}_n) \, \hat{v}_{j,n}^2}{n^2} \\
&= \frac{\sum_{i=1}^{n} \hat{v}_{i,n}^4 \, \sigma_{i,n}^2}{n^2},
\end{aligned}$$

where we use the fact that the $\varepsilon_{i,n}$ are conditionally uncorrelated, which follows from Assumption 1. Further, by Assumptions 2 and 3,

$$\frac{\sum_{i=1}^{n} \hat{v}_{i,n}^4 \, \sigma_{i,n}^2}{n^2} \leq (\max_i \sigma_{i,n}^2) \left(\max_i \frac{\|\hat{v}_{i,n}\|}{\sqrt{n}} \right)^2 \frac{\sum_{i=1}^{n} \hat{v}_{i,n}^2}{n} = o_p(1), \tag{A.2}$$

using

$$\frac{\sum_{i=1}^n \hat{v}_{i,n}^2}{n} \leq \frac{\sum_{i=1}^n v_{i,n}^2}{n} \leq 2\frac{\sum_{i=1}^n Q_{i,n}^2}{n} + 2\frac{\sum_{i=1}^n V_{i,n}^2}{n} = O_p(\chi_n) + O_p(1);$$

here, the first inequality follows from the fact that $\hat{v}_{i,n}$ is a least-squares residual—and thus has minimal variance—and the second is an application of the well-known inequality $\frac{1}{2}(a_1 + a_2) \leq \sqrt{\frac{1}{2}(a_1^2 + a_2^2)}$. Consequently,

$$\frac{\sum_{i=1}^n \hat{v}_{i,n}^2 \left(\varepsilon_{i,n}^2 - \sigma_{i,n}^2\right)}{n} = o_p(1),$$

and (A.1) reduces to

$$\frac{\sum_{i=1}^n \hat{v}_{i,n}^2 \left(y_{i,n}\acute{u}_{i,n} - \sigma_{i,n}^2\right)}{n} = \frac{\sum_{i=1}^n \hat{v}_{i,n}^2 \left(y_{i,n}\acute{u}_{i,n} - \varepsilon_{i,n}^2\right)}{n} + o_p(1). \tag{A.3}$$

We turn to the sample average on the right-hand side of this expression next.

To do so it is useful to work with the decomposition

$$\begin{aligned} y_{i,n}\acute{u}_{i,n} - \varepsilon_{i,n}^2 = &\sum_{j\neq i} \varepsilon_{i,n}(\boldsymbol{A}_n)_{i,j}\, \varepsilon_{j,n} + \left((\boldsymbol{A}_n)_{i,i} - 1\right) \varepsilon_{i,n}^2 \\ &+ \sum_{j=1}^n \mu_{i,n}(\boldsymbol{A}_n)_{i,j}\, \varepsilon_{j,n} + \sum_{j=1}^n \varepsilon_{i,n}(\boldsymbol{A}_n)_{i,j}\, e_{j,n} \\ &+ \sum_{j=1}^n \mu_{i,n}(\boldsymbol{A}_n)_{i,j}\, e_{j,n}, \end{aligned} \tag{A.4}$$

where

$$(\boldsymbol{A}_n)_{i,j} := \frac{(\boldsymbol{H}_n)_{i,j}}{(\boldsymbol{M}_n)_{i,i}}, \qquad (\boldsymbol{H}_n)_{i,j} := (\boldsymbol{M}_n)_{i,j} - \left(\frac{\sum_{k=1}^n \hat{v}_{k,n}^2}{n}\right)^{-1} \frac{\hat{v}_{i,n}\, \hat{v}_{j,n}}{n}.$$

Using standard formulae for partitioned regression, $\boldsymbol{H}_n$ can be seen to be the annihilator matrix of a regression on both $x_{i,n}$ and $\boldsymbol{w}_{i,n}$, whereas $\boldsymbol{M}_n$ follows from a projection on $\boldsymbol{w}_{i,n}$ alone. Observe that $(\boldsymbol{A}_n)_{i,j} \neq (\boldsymbol{A}_n)_{j,i}$. Equation (A.4) follows from observing that $y_{i,n} = \mu_{i,n} + \varepsilon_{i,n}$ and that

$$\acute{u}_{i,n} = \frac{\hat{u}_{i,n}}{(\boldsymbol{M}_n)_{i,i}} = \sum_{j=1}^n \frac{(\boldsymbol{H}_n)_{i,j}}{(\boldsymbol{M}_n)_{i,i}}\, u_{j,n} = \sum_{j=1}^n (\boldsymbol{A}_n)_{i,j}\, \varepsilon_{j,n} + \sum_{j=1}^n (\boldsymbol{A}_n)_{i,j}\, e_{j,n},$$

which itself is a consequence of $u_{i,n} = \varepsilon_{i,n} + e_{i,n}$ and the fact that $\hat{u}_{i,n} = \sum_{j=1}^{n} (\boldsymbol{H}_n)_{i,j}\, u_{j,n}$. Using (A.4) we have

$$\begin{aligned}
\frac{\sum_{i=1}^{n} \hat{v}_{i,n}^2 \left(y_{i,n}\acute{u}_{i,n} - \varepsilon_{i,n}^2\right)}{n} &= \frac{\sum_{i=1}^{n} \sum_{j\neq i} \hat{v}_{i,n}^2\, \varepsilon_{i,n} (\boldsymbol{A}_n)_{i,j}\, \varepsilon_{j,n}}{n} + \frac{\sum_{i=1}^{n} \hat{v}_{i,n}^2 \left((\boldsymbol{A}_n)_{i,i} - 1\right) \varepsilon_{i,n}^2}{n} \\
&+ \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} \hat{v}_{i,n}^2\, \mu_{i,n} (\boldsymbol{A}_n)_{i,j}\, \varepsilon_{j,n}}{n} \\
&+ \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} \hat{v}_{i,n}^2\, \varepsilon_{i,n} (\boldsymbol{A}_n)_{i,j}\, e_{j,n}}{n} \\
&+ \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} \hat{v}_{i,n}^2\, \mu_{i,n} (\boldsymbol{A}_n)_{i,j}\, e_{j,n}}{n}.
\end{aligned} \tag{A.5}$$

We will handle each of these five terms in turn.

For the first right-hand side term in (A.5),

$$\mathbb{E}\left(\frac{\sum_{i=1}^{n} \sum_{j\neq i} \hat{v}_{i,n}^2 \varepsilon_{i,n}\, (\boldsymbol{A}_n)_{i,j}\, \varepsilon_{j,n}}{n} \middle| \mathcal{X}_n, \mathcal{W}_n \right) = 0$$

because the $\varepsilon_{i,n}$ are (conditionally) uncorrelated by the assumption of independence under Assumption 1. To see that

$$\mathbb{E}\left(\left(\frac{\sum_{i=1}^{n} \sum_{j\neq i} \hat{v}_{i,n}^2 \varepsilon_{i,n}\, (\boldsymbol{A}_n)_{i,j}\, \varepsilon_{j,n}}{n} \right)^2 \middle| \mathcal{X}_n, \mathcal{W}_n \right) = o_p(1), \tag{A.6}$$

expand the square to get

$$\frac{\sum_{i_1=1}^{n} \sum_{i_2=1}^{n} \sum_{j_1\neq i_1} \sum_{j_2\neq i_2} \hat{v}_{i_1,n}^2\, (\boldsymbol{A}_n)_{i_1,j_1}\, \mathbb{E}(\varepsilon_{i_1,n}\varepsilon_{j_1,n}, \varepsilon_{i_2,n}\varepsilon_{j_2,n} | \mathcal{X}_n, \mathcal{W}_n)\, (\boldsymbol{A}_n)_{i_2,j_2}\, \hat{v}_{i_2,n}^2}{n^2}.$$

Conditional independence of the $\varepsilon_{i,n}$ implies that the summand will be zero unless either (i) $i_1 = i_2$ and $j_1 = j_2$, or (ii) $i_1 = j_2$ and $i_2 = j_1$. Collecting terms, the variance is equal to

$$\frac{\sum_{i=1}^{n} \sum_{j\neq i} \hat{v}_i^4\, \sigma_i^2 (\boldsymbol{A}_n)_{i,j}^2\, \sigma_j^2}{n^2} + \frac{\sum_{i=1}^{n} \sum_{j\neq i} \hat{v}_i^2\, \sigma_i^2 (\boldsymbol{A}_n)_{i,j}\, (\boldsymbol{A}_n)_{j,i}\, \hat{v}_j^2\, \sigma_j^2}{n^2},$$

where the first term arises from quadruples of the form under (i) and the second term stems

from those under (ii). Now,

$$\begin{aligned}\frac{\sum_{i=1}^n \sum_{j\neq i} \hat{v}_{i,n}^4\, \sigma_{i,n}^2 (\boldsymbol{A}_n)_{i,j}^2\, \sigma_{j,n}^2}{n^2} &\leq \frac{(\max_i \sigma_{i,n}^2)^2 \sum_{i=1}^n \hat{v}_{i,n}^4 \sum_{j\neq i} (\boldsymbol{A}_n)_{i,j}^2}{n^2} \\ &\leq \frac{(\max_i \sigma_{i,n}^2)^2 \sum_{i=1}^n \hat{v}_{i,n}^4\, (\boldsymbol{M}_n)_{i,i}^{-2}}{n^2} \\ &\leq \frac{(\max_i \sigma_{i,n}^2)^2\, (\min_i (\boldsymbol{M}_n)_{i,i})^{-2} \sum_{i=1}^n \hat{v}_{i,n}^4}{n^2} \\ &\leq \frac{(\max_i \sigma_{i,n}^2)^2}{(\min_i (\boldsymbol{M}_n)_{i,i})^2} \left(\max_i \frac{\|\hat{v}_{i,n}\|}{\sqrt{n}}\right)^2 \frac{\sum_{i=1}^n \hat{v}_{i,n}^2}{n} = o_p(1),\end{aligned}$$

where we use that $\sum_{j\neq i} (\boldsymbol{A}_n)_{i,j}^2 \leq \sum_{j=1}^n (\boldsymbol{A}_n)_{i,j}^2 = (\boldsymbol{H}_n)_{i,i}\, (\boldsymbol{M}_n)_{i,i}^{-2}$—which follows from the fact that $\boldsymbol{H}_n$ is a projection matrix, and so $\sum_{j=1}^n (\boldsymbol{H}_n)_{i,j}^2 = (\boldsymbol{H}_n)_{i,i} \in [0,1]$ by idempotency—and envoke Assumptions 2 and 4 for $(\max_i \sigma_{i,n}^2)(\min_i (\boldsymbol{M}_n)_{i,i})^{-1} = O_p(1)$. The final equality then follows as in (A.2). Similarly,

$$\begin{aligned}\frac{\sum_{i=1}^n \sum_{j\neq i} \hat{v}_{i,n}^2\, \sigma_{i,n}^2\, (\boldsymbol{A}_n)_{i,j}\, (\boldsymbol{A}_n)_{j,i}\, \hat{v}_{j,n}^2 \sigma_{j,n}^2}{n^2} &\leq (\max_i \sigma_{i,n}^2)^2\, \frac{\sum_{i=1}^n \sum_{j\neq i} \hat{v}_i^2 \hat{v}_j^2\, (\boldsymbol{A}_n)_{i,j}\, (\boldsymbol{A}_n)_{j,i}}{n^2} \\ &\leq \frac{(\max_i \sigma_{i,n}^2)^2 \left(\max_i \frac{\|\hat{v}_{i,n}\|}{\sqrt{n}}\right)^2}{(\min_i (\boldsymbol{M}_n)_{i,i})^2}\, \frac{\sum_{i=1}^n \hat{v}_{i,n}^2}{n} = o_p(1),\end{aligned}$$

where, now, we use that $(\boldsymbol{A}_n)_{i,j}(\boldsymbol{A}_n)_{j,i} \geq 0$ to validate the first inequality, and that $\sum_{j\neq i} (\boldsymbol{A}_n)_{i,j}(\boldsymbol{A}_n)_{j,i} \leq \sum_{j=1}^n (\boldsymbol{A}_n)_{i,j}(\boldsymbol{A}_n)_{j,i} \leq (\min_i (\boldsymbol{M}_n)_{i,i})^{-2}$ for the second inequality. Equation (A.6) has been shown.

The second right-hand side term in (A.5) has mean

$$\mathbb{E}\left(\left\|\frac{\sum_{i=1}^n \hat{v}_{i,n}^2\, ((\boldsymbol{A}_n)_{i,i} - 1)\, \varepsilon_{i,n}^2}{n}\right\| \middle| \mathcal{X}_n, \mathcal{W}_n\right) = \left(\frac{\sum_{i=1}^n \hat{v}_{i,n}^2}{n}\right)^{-1} \frac{\sum_{i=1}^n \hat{v}_{i,n}^4 \sigma_{i,n}^2\, (\boldsymbol{M}_n)_{i,i}^{-1}}{n^2},$$

where we use that

$$((\boldsymbol{A}_n)_{i,i} - 1) = -\left(\frac{\sum_{i=1}^n \hat{v}_{i,n}^2}{n}\right)^{-1} \left(\frac{\hat{v}_{i,n}^2}{n}\, \frac{1}{(\boldsymbol{M}_n)_{i,i}}\right).$$

This vanishes because $\sum_{i=1}^n \hat{v}_{i,n}^2 = O_p(n)$ and

$$\frac{\sum_{i=1}^n \hat{v}_{i,n}^4 \sigma_{i,n}^2\, (\boldsymbol{M}_n)_{i,i}^{-1}}{n^2} \leq (\min_i (\boldsymbol{M}_n)_{i,i})^{-1}\, \frac{\sum_{i=1}^n \hat{v}_{i,n}^4 \sigma_{i,n}^2}{n^2} = o_p(1),$$

which follows from (A.2). Similarly, by an application of the Cauchy-Schwarz inequality, the second moment satisfies

$$\mathbb{E}\left(\left(\frac{\sum_{i=1}^n \hat{v}_{i,n}^2\left((\boldsymbol{A}_n)_{i,i}-1\right)\varepsilon_{i,n}^2}{n}\right)^2 \middle| \mathcal{X}_n, \mathcal{W}_n\right) \leq \left(\frac{\sum_{i=1}^n \hat{v}_{i,n}^4\,(\boldsymbol{M}_n)_{i,i}^{-1}\sqrt{\mathbb{E}(\varepsilon_{i,n}^4|\mathcal{X}_n,\mathcal{W}_n)}}{n^2}\right)^2,$$

which is $o_p(1)$ by the same argument as that used for the mean, only now utilizing that $\max_i \mathbb{E}(\varepsilon_{i,n}^4|\mathcal{X}_n,\mathcal{W}_n) = O_p(1)$ by Assumption 2.

The third right-hand side term in (A.5) is mean zero because $\mathbb{E}(\varepsilon_{i,n}|\mathcal{X}_n,\mathcal{W}_n) = 0$ by construction. Its variance is

$$\mathbb{E}\left(\left(\frac{\sum_{i=1}^n\sum_{j=1}^n \hat{v}_{i,n}^2\,\mu_{i,n}(\boldsymbol{A}_n)_{i,j}\,\varepsilon_{j,n}}{n}\right)^2 \middle| \mathcal{X}_n, \mathcal{W}_n\right) = \sum_{j=1}^n \sigma_{j,n}^2\left(\frac{\sum_{i=1}^n \hat{v}_{i,n}^2\,\mu_{i,n}(\boldsymbol{A}_n)_{i,j}}{n}\right)^2.$$

To show that this vanishes first use

$$\sum_{j=1}^n (\boldsymbol{A}_n)_{i,j}\,(\boldsymbol{A}_n)_{k,j} = \sum_{j=1}^n \frac{(\boldsymbol{H}_n)_{i,j}\,(\boldsymbol{H}_n)_{j,k}}{(\boldsymbol{M}_n)_{i,i}\,(\boldsymbol{M}_n)_{k,k}} = \frac{(\boldsymbol{H}_n)_{i,k}}{(\boldsymbol{M}_n)_{i,i}\,(\boldsymbol{M}_n)_{k,k}},$$

which follows from idempotency of $\boldsymbol{H}_n$, to see that

$$\sum_{j=1}^n\left(\frac{\sum_{i=1}^n \hat{v}_{i,n}^2\,\mu_{i,n}(\boldsymbol{A}_n)_{i,j}}{n}\right)^2 = \frac{\sum_{i=1}^n\sum_{k=1}^n \hat{v}_{i,n}^2\,\mu_{i,n}\left(\sum_{j=1}^n(\boldsymbol{A}_n)_{i,j}\,(\boldsymbol{A}_n)_{k,j}\right)\hat{v}_{k,n}^2\,\mu_{k,n}}{n^2}$$

is a quadratic form in the matrix $\boldsymbol{H}_n$. Because $\boldsymbol{H}_n$ is a projection matrix we have that $\sum_{i=1}^n\sum_{j=1}^n a_i(\boldsymbol{H}_n)_{i,j}a_j \leq \sum_{i=1}^n a_i^2$ for any $(a_1,\ldots,a_n)$. Consequently, the variance satisfies

$$\begin{aligned}
\sum_{j=1}^n \sigma_{j,n}^2\left(\frac{\sum_{i=1}^n \hat{v}_{i,n}^2\,\mu_{i,n}(\boldsymbol{A}_n)_{i,j}}{n}\right)^2 &\leq (\max_i \sigma_{i,n}^2)\sum_{j=1}^n\left(\frac{\sum_{i=1}^n \hat{v}_{i,n}^2\,\mu_{i,n}(\boldsymbol{A}_n)_{i,j}}{n}\right)^2 \\
&\leq (\max_i \sigma_{i,n}^2)\sum_{i=1}^n\left(\frac{\hat{v}_{i,n}^2\,\mu_{i,n}\,(\boldsymbol{M}_n)_{i,i}^{-1}}{n}\right)^2 \\
&\leq \frac{(\max_i \sigma_{i,n}^2)}{(\min_i(\boldsymbol{M}_n)_{i,i})^2}\frac{\sum_{i=1}^n \hat{v}_{i,n}^4\,\mu_{i,n}^2}{n^2} \\
&\leq \frac{(\max_i \sigma_{i,n}^2)}{(\min_i(\boldsymbol{M}_n)_{i,i})^2}\left(\max_i \frac{\|\mu_{i,n}\|}{\sqrt{n}}\right)^2\frac{\sum_{i=1}^n \hat{v}_{i,n}^4}{n} = o_p(1),
\end{aligned}$$

where the final transition uses the fact that $\hat{v}_{i,n} = \tilde{Q}_{i,n} + \tilde{V}_{i,n}$ together with the inequality

$$\frac{\sum_{i=1}^n \hat{v}_{i,n}^4}{n} \leq 4\frac{\sum_{i=1}^n \tilde{Q}_{i,n}^4}{n} + 4\frac{\sum_{i=1}^n \tilde{V}_{i,n}^4}{n} = O_p(1);$$

the last equality following from Assumption 4 and the fact that $\max_i \mathbb{E}(\tilde{V}_{i,n}^4|\mathcal{W}_n) = O_p(1)$ by Assumption 1 and Assumption 2, a detailed derivation of this last result is available in the proof of Lemma SA-7 in the supplementary material to Cattaneo, Jansson and Newey (2018).

The fourth right-hand side term in (A.5), is zero mean for the same reason as the third. Its variance is

$$\mathbb{E}\left(\left(\frac{\sum_{i=1}^n \sum_{j=1}^n \hat{v}_{i,n}^2\, \varepsilon_{i,n} (\boldsymbol{A}_n)_{i,j}\, e_{j,n}}{n}\right)^2 \middle| \mathcal{X}_n, \mathcal{W}_n\right) = \frac{\sum_{i=1}^n \sigma_{i,n}^2 \hat{v}_{i,n}^4 \left(\sum_{j=1}^n (\boldsymbol{A}_n)_{i,j}\, e_{j,n}\right)^2}{n^2}$$

and is bounded by

$$\frac{(\max_i \sigma_{i,n}^2)}{(\min_i (\boldsymbol{M}_n)_{i,i})^2} \left(\max_i \frac{\|\hat{v}_{i,n}\|}{\sqrt{n}}\right)^4 n \left(\frac{\sum_{i=1}^n e_{i,n}^2}{n}\right) = o_p(n\varrho_n) = o_p(1),$$

where we use $(\sum_{j=1}^n (\boldsymbol{A}_n)_{i,j}\, e_{j,n})^2 \leq (\sum_{j=1}^n (\boldsymbol{A}_n)_{i,j}^2)\,(\sum_{j=1}^n e_{j,n}^2) \leq (\boldsymbol{M}_n)_{i,i}^{-2}\,(\sum_{j=1}^n e_{j,n}^2)$ and rely on $\varrho_n = O(n^{-1})$ as stated in Assumption 4 to reach the desired conclusion.

The fifth right-hand side term in (A.5), finally, is the bias term. The Cauchy-Schwarz inequality gives

$$\left(\frac{\sum_{i=1}^n \sum_{j=1}^n \hat{v}_{i,n}^2 \mu_{i,n}\, (\boldsymbol{A}_n)_{i,j}\, e_{j,n}}{n}\right)^2 \leq \frac{\sum_{i=1}^n \hat{v}_{i,n}^4\, (\boldsymbol{M}_n)_{i,i}^{-2}}{n}\, \frac{\sum_{i=1}^n \mu_{i,n}^2 (\sum_{j=1}^n (\boldsymbol{H}_n)_{i,j}\, e_{j,n})^2}{n},$$

where

$$\frac{\sum_{i=1}^n \hat{v}_{i,n}^4\, (\boldsymbol{M}_n)_{i,i}^{-2}}{n} \leq \frac{(\min_i (\boldsymbol{M}_n)_{i,i}^{-2})\, \sum_{i=1}^n \hat{v}_{i,n}^4}{n} = O_p(1),$$

and

$$\frac{\sum_{i=1}^n \mu_{i,n}^2 (\sum_{j=1}^n (\boldsymbol{H}_n)_{i,j}\, e_{j,n})^2}{n} \leq \left(\max_i \frac{\|\mu_{i,n}\|}{\sqrt{n}}\right)^2 \sum_{i=1}^n \left(\sum_{j=1}^n (\boldsymbol{H}_n)_{i,j}\, e_{j,n}\right)^2$$
$$= \left(\max_i \frac{\|\mu_{i,n}\|}{\sqrt{n}}\right)^2 \left(\sum_{j=1}^n \sum_{k=1}^n e_{j,n} \left(\sum_{i=1}^n (\boldsymbol{H}_n)_{i,j}\, (\boldsymbol{H}_n)_{i,k}\right) e_{k,n}\right)$$
$$= \left(\max_i \frac{\|\mu_{i,n}\|}{\sqrt{n}}\right)^2 \left(\sum_{j=1}^n \sum_{k=1}^n e_{j,n} (\boldsymbol{H}_n)_{j,k}\, e_{k,n}\right)$$
$$\leq \left(\max_i \frac{\|\mu_{i,n}\|}{\sqrt{n}}\right)^2 \left(\sum_{j=1}^n e_{j,n}^2\right) = o_p(n\varrho_n) = o_p(1)$$

by the same arguments as before.

Collecting results for the five right-hand side terms in (A.5) implies that (A.3) becomes

$$\frac{\sum_{i=1}^n \hat{v}_{i,n}^2\, (y_{i,n} \acute{u}_{i,n})}{n} = \frac{\sum_{i=1}^n \hat{v}_{i,n}^2\, \sigma_{i,n}^2}{n} + o_p(1),$$

which is what we wanted to show. □

A.2 Sufficient conditions for Assumption 4.

We continue to work with the case where $r = 1$. We provide primitive conditions for the requirements

$$\text{(i)} \quad \frac{\sum_{i=1}^n \tilde{Q}_{i,n}^4}{n} = O_p(1) \qquad \text{and} \qquad \text{(ii)} \quad \max_i \frac{\|\mu_{i,n}\|}{\sqrt{n}} = o_p(1),$$

in turn.

Condition (i). As shown in the proof of Lemma SA-7 in the supplementary material to Cattaneo, Jansson and Newey (2018), $\chi_n = o(1)$ implies that

$$\max_i \frac{\|\tilde{Q}_{i,n}\|}{\sqrt{n}} = o_p(1).$$

Consequently,

$$\frac{\sum_{i=1}^n \tilde{Q}_{i,n}^4}{n} \leq \left(\max_i \frac{\|\tilde{Q}_{i,n}\|}{\sqrt{n}}\right)^2 \sum_{i=1}^n \tilde{Q}_{i,n}^2 \leq \left(\max_i \frac{\|\tilde{Q}_{i,n}\|}{\sqrt{n}}\right)^2 n \left(\frac{\sum_{i=1}^n Q_{i,n}^2}{n}\right) = o_p(n\chi_n),$$

where

$$\sum_{i=1}^{n} \tilde{Q}_{i,n}^2 = \sum_{i=1}^{n} \left(\sum_{j=1}^{n} (\boldsymbol{M}_n)_{i,j} Q_{j,n} \right)^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} Q_{i,n} (\boldsymbol{M}_n)_{i,j} Q_{j,n} \leq \sum_{i=1}^{n} Q_{i,n}^2$$

was used. Hence, $n\chi_n = O(1)$ is sufficient for Condition (i) to hold.

Next, recall that

$$n_{i,n} = \sum_{j=1}^{n} \{(\boldsymbol{M}_n)_{i,j} \neq 0\},$$

and let $[i]_n := \{j : (\boldsymbol{M}_n)_{i,j} \neq 0\}$. If $\max_i n_{i,n} = O_p(1)$ and $\sum_{i=1}^{n} Q_{i,n}^4 = O_p(n)$, then we obtain

$$\begin{aligned}
\frac{\sum_{i=1}^{n} \tilde{Q}_{i,n}^4}{n} = \frac{\sum_{i=1}^{n} \left(\sum_{j=1}^{n} (\boldsymbol{M}_n)_{i,j} Q_{j,n} \right)^4}{n} &\leq \frac{\sum_{i=1}^{n} \left(\sum_{j=1}^{n} (\boldsymbol{M}_n)_{i,j}^{4/3} \right)^3 \left(\sum_{j \in [i]_n} Q_{j,n}^4 \right)}{n} \\
&\leq \frac{\sum_{i=1}^{n} n_{i,n}^3 \left(\sum_{j \in [i]_n} Q_{j,n}^4 \right)}{n} \\
&\leq (\max_i n_{i,n})^3 \frac{\sum_{i=1}^{n} \sum_{j \in [i]_n} Q_{j,n}^4}{n} \\
&\leq (\max_i n_{i,n})^4 \frac{\sum_{j=1}^{n} Q_{j,n}^4}{n} = O_p(1)
\end{aligned}$$

by an application of Hölder's inequality.

The requirement $\max_i n_{i,n} = O_p(1)$ can be relaxed if higher-order moments of $Q_{i,n}$ exist. By another application of Hölder's inequality, for some $\theta \geq 5$,

$$\begin{aligned}
\frac{\sum_{i=1}^{n} \tilde{Q}_{i,n}^4}{n} &\leq \left(\frac{\sum_{j=1}^{n} \|Q_{i,n}\|^\theta}{n} \right)^{4/\theta} \left(\frac{\sum_{i=1}^{n} \left(\sum_{j=1}^{n} (\boldsymbol{M}_n)_{i,j}^{\theta/(\theta-1)} \right)^{4(\theta-1)}}{n^{(\theta-4)}} \right)^{1/\theta} \\
&\leq \left(\frac{\sum_{j=1}^{n} \|Q_{i,n}\|^\theta}{n} \right)^{4/\theta} \left(\frac{\sum_{i=1}^{n} n_{i,n}^{4(\theta-1)}}{n^{(\theta-4)}} \right)^{1/\theta} \\
&\leq \left(\frac{\sum_{j=1}^{n} \|Q_{i,n}\|^\theta}{n} \right)^{4/\theta} \left(\frac{(\max_i n_{i,n})^{4(\theta-1)}}{n^{(\theta-5)}} \right)^{1/\theta}.
\end{aligned}$$

Consequently, Condition (i) is satisfied if

$$\frac{\sum_{j=1}^{n} \mathbb{E}(\|Q_{i,n}\|^\theta)}{n} = O(1), \qquad \text{and} \qquad \max_i n_{i,n} = O_p\left(n^{\frac{1}{4}\left(1 - \frac{4}{\theta-1}\right)} \right).$$

For example, if $\max_i \|Q_{i,n}\| = O_p(1)$, the rate requirement becomes $\max_i n_{i,n} = O_p(n^{1/4})$.

Condition (ii). We first note that

$$\max_i \|\mu_{i,n}\| \leq \max_i \|x_{i,n}\| \, \|\beta_n\| + \max_i \|\boldsymbol{w}'_{i,n}\boldsymbol{\gamma}_n + e_{i,n}\|,$$

and that the first term on the right-hand side is easily handled. For any $\epsilon > 0$ and $\theta > 0$, we have

$$\Pr\left(\max_i \frac{\|x_{i,n}\|}{\sqrt{n}} > \epsilon\right) \leq \sum_{i=1}^{n} \Pr\left(\|x_{i,n}\| > \epsilon\sqrt{n}\right) \leq \left(\frac{n^{-\theta/2}}{\epsilon^{2+\theta}}\right) \frac{\sum_{i=1}^{n} \mathbb{E}\left(\|x_{i,n}\|^{2+\theta}\right)}{n}.$$

Consequently, $\max_i \|x_{i,n}\|/\sqrt{n} = o_p(1)$ follows from $\sum_{i=1}^{n} \mathbb{E}(\|x_{i,n}\|^{2+\theta}) = O(n)$, which is a conventional requirement.

The same argument can be used for the second term in cases where $\boldsymbol{w}'_{i,n}\boldsymbol{\gamma}_n$ is a series approximation to a well-behaved function $\varphi(\boldsymbol{z}_{i,n})$. In that case, $\boldsymbol{w}'_{i,n}\boldsymbol{\gamma}_n + e_{i,n} = \varphi(\boldsymbol{z}_{i,n})$, and the requirement that $\sum_{i=1}^{n} \mathbb{E}(\|\varphi(\boldsymbol{z}_{i,n})\|^{2+\theta}) = O(n)$ again does not appear overly strong to impose.

We can also tackle the problem by first noting that

$$\max_i \|\boldsymbol{w}'_{i,n}\boldsymbol{\gamma}_n + e_{i,n}\| \leq \max_i \|\boldsymbol{w}'_{i,n}\boldsymbol{\gamma}_n\| + \max_i \|e_{i,n}\| \leq \max_i \|\boldsymbol{w}'_{i,n}\boldsymbol{\gamma}_n\| + o_p(\sqrt{n}),$$

using the fact that

$$\Pr\left(\max_i \frac{\|e_{i,n}\|}{\sqrt{n}} \geq \epsilon\right) \leq \frac{1}{\epsilon^2} \frac{\sum_{i=1}^{n} \mathbb{E}(\|e_{i,n}\|^2)}{n} = O(\varrho_n) = o(1)$$

by Assumption 3, and then imposing a growth rate on the number of parameters that affect each observation, together with a moment condition on $\boldsymbol{w}_{i,n}$. Moreover, writing $\boldsymbol{w}_{i,n} = (w_{i,n,1}, \ldots, w_{i,n,q_n})'$ and $\boldsymbol{\gamma}_n = (\gamma_{n,1}, \ldots, \gamma_{n,q_n})'$,

$$\boldsymbol{w}'_{i,n}\boldsymbol{\gamma}_n = \sum_{j=1}^{q_n} w_{i,n,j}\, \gamma_{n,j} = \sum_{j \in (i)_n} w_{i,n,j}\, \gamma_{n,j},$$

where the index set $(i)_n$ has cardinality $\kappa_{i,n} := |(i)_n|$. In grouped data with κ-way fixed effects, for example, $\kappa_{i,n} = \kappa$ for all i, but we will also cover the case where $\kappa_{i,n}$ grows with n. By Hölder's inequality,

$$\mathbb{E}(\|\boldsymbol{w}'_{i,n}\boldsymbol{\gamma}_n\|^{2+\theta}) \leq \left(\sum_{j \in (i)_n} \mathbb{E}(\|w_{i,n,j}\|^{2+\theta})\right) \left(\sum_{j \in (i)_n} \gamma_{n,j}^{(2+\theta)/(1+\theta)}\right)^{1+\theta}.$$

Consequently, if $\sum_{j\in(i)_n} \mathbb{E}(\|w_{i,n,j}\|^{2+\theta}) = O(\kappa_i)$ for all i and $\gamma_{n,i} = O(1)$ for all i we obtain that

$$\Pr\left(\max_i \frac{\|\boldsymbol{w}'_{i,n}\boldsymbol{\gamma}_n\|}{\sqrt{n}}\right) \leq \left(\frac{n^{-\theta/2}}{\epsilon^{2+\theta}}\right) \frac{\sum_{i=1}^n \mathbb{E}\left(\|w'_{i,n}\boldsymbol{\gamma}_n\|^{2+\theta}\right)}{n} = O\left(\frac{(\max_i \kappa_{i,n})^{2+\theta}}{n^{\theta/2}}\right),$$

which vanishes provided that $\max_i \kappa_{i,n} = O(n^{\frac{1}{2}\frac{\theta}{2+\theta}})$. The same rate requirement can equally be obtained under the alternative condition that

$$\frac{\sum_{i=1}^n \sum_{j\in(i)_n} \gamma_{n,j}^{2+\theta}}{n} = O(1),$$

together with the assumption that $\max_i \max_j \mathbb{E}(\|w_{i,n,j}\|^{2+\theta}) = O(1)$, by another application of Hölder's inequality.

A.3 Additional simulation results

We next present simulation results for the other models considered in the supplementary material to Cattaneo, Jansson and Newey (2018).

One-way panel model. The first model considered is the standard fixed-effect model for panel data. The design is similar to the design used in the main text, although here there is no randomness in the dummies and the groups do no overlap. For double-indexed data $y_{(g,i)}, x_{(g,i)}$, the model is

$$y_{(g,i)} = x_{(g,i)}\,\beta + \alpha_g + \varepsilon_{(g,i)}, \qquad g = 1,\ldots,N, \quad i = 1\ldots,T,$$

and α_g is a group-specific intercept. The within-group (fixed-effect) estimator equals the ordinary least-squares estimator of $y_{(g,i)}$ on $x_{(g,i)}$ and N dummy variables that capture the group membership. We draw $x_{(g,i)} \sim$ i.i.d. $\boldsymbol{N}(0,1)$ and $\varepsilon_{(g,i)} \sim$ i.i.d. $\boldsymbol{N}(0,1)$, set $\beta = 1$ and $\alpha_g = 0$ for all groups g, and consider samples with $N\lfloor 700/T\rfloor$ for $T \in \{700, 10, 5, 4, 3, 2\}$, which yields a total sample size of 700 (except when $T = 3$, in which case the sample size is 699), as in our simulations in the main text.

Table A.1 reports the results of our simulations. Note that, when $T = 2$, the variance estimator $\grave{\Sigma}_n$ does not exist. For that case, the least-squares estimator is equivalen to

a cross-sectional regression on first-differenced data, and the usual Eicker-White standard error is valid for that regression. Hence, the table reports results for that estimator instead. For clarity, these values are italicized in the table.

Table A.1: One-way panel model

N	1	70	140	175	233	350
T	700	10	5	4	3	2
q_n/n	0.0014	0.1000	0.2000	0.2500	0.3333	0.5000
			Mean			
$\sqrt{\Omega_n}$	0.0382	0.0399	0.0426	0.0442	0.0464	0.0539
$\sqrt{\hat{\Omega}_n}$	0.0377	0.0377	0.0377	0.0377	0.0377	0.0376
$\sqrt{\grave{\Omega}_n}$	0.0377	0.0397	0.0421	0.0434	0.0461	*0.0532*
$\sqrt{\acute{\Omega}_n}$	0.0377	0.0397	0.0421	0.0434	0.0461	0.0531
			Standard deviation			
$\sqrt{\hat{\Omega}_n}$	0.0020	0.0021	0.0021	0.0022	0.0023	0.0028
$\sqrt{\grave{\Omega}_n}$	0.0020	0.0023	0.0027	0.0029	0.0033	*0.0040*
$\sqrt{\acute{\Omega}_n}$	0.0026	0.0029	0.0032	0.0034	0.0038	0.0054
			Rejection frequency (nominal size is 5%)			
$\sqrt{\hat{\Omega}_n}$	0.0550	0.0663	0.0842	0.0931	0.1137	0.1733
$\sqrt{\grave{\Omega}_n}$	0.0547	0.0551	0.0560	0.0539	0.0555	*0.0541*
$\sqrt{\acute{\Omega}_n}$	0.0556	0.0575	0.0556	0.0545	0.0555	0.0577
			Average width of confidence interval			
$\sqrt{\hat{\Omega}_n}$	0.1478	0.1477	0.1477	0.1477	0.1477	0.1475
$\sqrt{\grave{\Omega}_n}$	0.1479	0.1556	0.1650	0.1703	0.1807	*0.2086*
$\sqrt{\acute{\Omega}_n}$	0.1477	0.1555	0.1650	0.1702	0.1806	0.2082

Partially-linear model. We next provide simulation results for a series estimator of the partially-linear model

$$y_i = x_i\,\beta + \exp(-\sqrt{\|\boldsymbol{z}_i\|}) + \varepsilon_i, \qquad x_i = \exp(\sqrt{\|\boldsymbol{z}_i\|}) + V_i,$$

where, again $x_i \sim$ i.i.d. $\boldsymbol{N}(0,1)$ with $\beta = 1$, and we draw $(\varepsilon_i, V_i) \sim$ i.i.d. $\boldsymbol{N}(\boldsymbol{0}, \boldsymbol{I}_2)$ and, for $\boldsymbol{z}_i := (z_{i,1}, \ldots, z_{i,6})'$, generate each $z_{i,j} \sim$ i.i.d. Uniform$[-1,1]$. We approximate the

function $\exp(-\sqrt{\|\boldsymbol{z}_i\|})$ by a power-series expansion of order κ_n. Moreover, for a given κ_n, let $\boldsymbol{w}_{i,n}$ denote the vector that collects all (unique) terms of the form $z_{i,1}^{k_1} \times z_{i,2}^{k_2} \times \cdots \times z_{i,6}^{k_6}$ with $k_1 + \cdots + k_6 = \kappa_n$. This yields $q_n = (6+\kappa_n)!/(6! \times \kappa_n!)$ as the dimension of the nuisance parameter. We then estimate β by the least-squares estimator of y_i on x_i and $\boldsymbol{w}_{i,n}$, again maintaining a sample size of 700. Note that, here, the vector $\boldsymbol{\gamma}_n$ is non-zero, but $\|\boldsymbol{\gamma}_n\| = O(1)$ because the approximation converges.

The simulation results for $\kappa_n \in \{1, 2, 3, 4, 5\}$ are collected in Table A.2.

Table A.2: Partially-linear model

κ_n	1	2	3	4	5
q_n	7	28	84	210	462
q_n/n	0.01	0.04	0.12	0.30	0.66
			Mean		
$\sqrt{\Omega}_n$	0.0359	0.0389	0.0405	0.0454	0.0654
$\sqrt{\hat{\Omega}}_n$	0.0354	0.0377	0.0378	0.0388	0.0444
$\sqrt{\grave{\Omega}}_n$	0.0355	0.0384	0.0401	0.0450	0.0636
$\sqrt{\acute{\Omega}}_n$	0.0354	0.0382	0.0398	0.0446	0.0633
			Standard deviation		
$\sqrt{\hat{\Omega}}_n$	0.0019	0.0020	0.0021	0.0024	0.0039
$\sqrt{\grave{\Omega}}_n$	0.0019	0.0021	0.0024	0.0031	0.0082
$\sqrt{\acute{\Omega}}_n$	0.0043	0.0048	0.0052	0.0063	0.0123
			Rejection frequency (nominal size is 5%)		
$\sqrt{\hat{\Omega}}_n$	0.0682	0.0576	0.0690	0.0937	0.1851
$\sqrt{\grave{\Omega}}_n$	0.0666	0.0519	0.0513	0.0536	0.0632
$\sqrt{\acute{\Omega}}_n$	0.0734	0.0596	0.0603	0.0600	0.0747
			Width of confidence interval		
$\sqrt{\hat{\Omega}}_n$	0.1387	0.1476	0.1482	0.1520	0.1740
$\sqrt{\grave{\Omega}}_n$	0.1393	0.1506	0.1573	0.1762	0.2493
$\sqrt{\acute{\Omega}}_n$	0.1389	0.1496	0.1562	0.1750	0.2481

References

Anatolyev, S. (2018). Almost unbiased variance estimation in linear regressions with many covariates. *Economics Letters 169*, 20–23.

Angrist, J. D. and J. Hahn (2004). When to control for covariates? Panel asymptotics for estimates of treatment effects. *Review of Economics and Statistics 86*, 58–72.

Bera, A. K., T. Suprayitno, and G. 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.

Cattaneo, M. D., M. Jansson, and X. Ma (2018). Two-step estimation and inference with possibly many included covariates. Forthcoming in *Review of Economic Studies*.

Cattaneo, M. D., M. Jansson, and W. K. Newey (2018). Inference in linear regression models with many covariates and heteroskedasticity. *Journal of the American Statistical Association 113*, 1350–1361.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2018). Double/debiased machine learning for treatment and structural parameters. *Econometrics Journal 21*, C1–C68.

Chesher, A. and I. Jewitt (1987). The bias of a heteroskedasticity consistent covariance matrix estimator. *Econometrica 55*, 1217–1222.

Eicker, F. (1963). Asymptotic normality and consistency of the least squares estimators for families of linear regressions. *Annals of Mathematical Statistics 34*, 447–456.

Eicker, F. (1967). Limit theorems for regressions with unequal and dependent errors. In L. Le Cam and J. Neyman (Eds.), *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Volume 1, pp. 59–82.

Hartley, H. O., J. N. K. Rao, and G. Kiefer (1969). Variance estimation with one unit per stratum. *Journal of the American Statistical Association 64*, 841–851.

Horn, S. D. and R. A. Horn (1975). Comparison of estimators of heteroscedastic variances in linear models. *Journal of the American Statistical Association 70*, 872–875.

Horn, S. D., R. A. Horn, and D. B. Duncan (1975). Estimating heteroskedastic variances in linear models. *Journal of the American Statistical Association 70*, 380–385.

Huber, P. J. (1973). Robust regression: Asymptotics, conjectures, and Monte Carlo. *Annals of Statistics 1*, 799–821.

Kline, P., R. Saggio, and M. Sølvsten (2018). Leave-out estimation of variance components. arXiv:1806.01494.

Long, J. S. and L. H. Ervin (2000). Using heteroscedasticity consistent standard errors in the linear regression model. *The American Statistician 54*, 217–224.

Mallela, P. (1972). Necessary and sufficient conditions for MINQU-Estimation of heteroskedastic variance in linear models. *Journal of the American Statistical Association 67*, 486–487.

Newey, W. K. (1997). Convergence rates and asymptotic normality for series estimators. *Journal of Econometrics 79*, 147–168.

Newey, W. K. (2009). Two-step series estimation of sample selection models. *Econometrics Journal 12*, S217–S229.

Newey, W. K. and J. M. Robins (2018). Cross-fitting and fast remainder rates for semiparametric estimation. arXiv:1801.09138.

Neyman, J. and E. L. Scott (1948). Consistent estimates based on partially consistent observations. *Econometrica 16*, 1–32.

Rao, C. R. (1970). Estimation of heteroscedastic variances in linear models. *Journal of the American Statistical Association 65*, 161–172.

Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review 94*, 247–252.

Stock, J. H. and M. W. Watson (2008). Heteroskedasticity-robust standard errors for fixed effects panel data regression. *Econometrica 76*, 155–174.

Verdier, V. (2018). Estimation and inference for linear models with two-way fixed effects and sparsely matched data. Forthcoming in *Review of Economics and Statistics*.

White, H. (1980). A heteroskedasticity-consistent covariance matrix and a direct test for heteroskedasticity. *Econometrica 48*, 817–838.

Wooldridge, J. M. (2002). *Econometric Analysis of Cross Section and Panel Data*. MIT Press.

Young, A. (2016). Improved, nearly exact, statistical inference with robust and clustered covariance matrices using effective degrees of freedom corrections. Mimeo, London School of Economics.