UNIVERSITY OF CAMBRIDGE

Faculty of Economics

Cambridge Working Papers in Economics

HETEROSKEDASTICITY-ROBUST INFERENCE IN LINEAR REGRESSION MODELS WITH MANY COVARIATES

Koen Jochmans

28 April 2020

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 for such a setting is presented that complements recent work by Cattaneo, Jansson and Newey (2018). We provide high-level conditions for our approach to deliver (asymptotically) size-correct inference as well as more primitive conditions for three special cases. Simulation results and an empirical illustration to inference on the union premium are also provided.

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 for such a setting is presented that complements recent work by Cattaneo, Jansson and Newey (2018). We provide high-level conditions for our approach to deliver (asymptotically) size-correct inference as well as more primitive conditions for three special cases. Simulation results and an empirical illustration to inference on the union premium 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 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 regressor design contains observations with high leverage (Chesher and Jewitt, 1987).[1] A necessary condition for the least-squares estimator to be asymptotically normal is that maximal leverage vanishes as the sample size grows large (Huber, 1973). This condition, then, also implies consistency of the robust covariance-matrix estimator (under regularity conditions).

The requirement that maximal leverage vanishes is problematic when the regressors include a large set of control variables. 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 demonstrated by Cattaneo, Jansson and Newey (2018). They obtained the same result for the other members of the so-called HC-class of covariance-matrix estimators (see, e.g., Long and Ervin 2000 and MacKinnon 2012 for reviews) and showed that the jackknife variance estimator of MacKinnon and White (1985), although inconsistent, can be used to perform asymptotically-conservative inference under asymptotics where $q_n/n \nrightarrow 0$. On the other hand, a bias-corrected covariance-matrix estimator in the spirit of Hartley, Rao and Kiefer (1969) and Bera, Suprayitno and Premaratne (2002) is shown to be consistent under conditions that bound maximal leverage below $\frac{1}{2}$. One implication of these conditions is that $\limsup_n q_n/n < \frac{1}{2}$. Although not guaranteed to be positive semi-definite, this estimator is attractive as it is asymptotically equivalent to a minimum-norm unbiased estimator in the sense of Rao (1970).

In this paper we discuss an alternative estimator of the covariance matrix that can deal with designs where maximal leverage is bounded away from 1. As such it remains consistent

---

[1]The leverage of an observation i is defined as the ith diagonal element of the hat matrix, i.e., the matrix that transforms the observed outcomes into fitted values. It is bounded between zero and one and measures the influence of the observation on its own fitted value; a larger value reflects a higher influence (see, e.g., Hoaglin and Welsch 1978).

when $\limsup_n q_n/n < 1$. To show this we need to impose additional conditions relative to Cattaneo, Jansson and Newey (2018). A consistency result is first provided under high-level conditions. Primitive conditions are then given for three special cases; the partially-linear regression model, the one-way model for short panel data, and the generic linear model with increasing dimension. Again, our covariance-matrix estimator need not be positive semi-definite in small samples. It is further not invariant to changes in the scale of the regression slopes. Achieving such invariance under asymptotics where $\limsup_n q_n/n > \frac{1}{2}$ appears difficult, as we discuss below.

The idea underlying our variance estimator can be traced back to work by Kline, Saggio and Sølvsten (2019, Remark 4 and Lemma 5); see below. The chief difference lies in the conditions under which the consistency result is obtained. They considered settings where the observations are independent and the regressors are fixed, and entertained models that are correctly specified with regression functions that are uniformly bounded. This is reasonable in analysis-of-variance problems, which are the main focus of application in their work. Here, we maintain the framework of Cattaneo, Jansson and Newey (2018). Some dependence between observations is allowed, regressors are stochastic, and the model can feature (vanishing) misspecification bias. Our consistency result can be understood to be an extension of Kline, Saggio and Sølvsten (2019) to such settings. It allows for regressors to have unbounded support and for observations to depend on a growing number of parameters. Our conditions do rule out dynamic models, for example. Verdier (2020) provides estimation and inference results for two-way models estimated from short panel data that allow for dynamics over time.

In Section 2 we introduce the framework, present our covariance-matrix estimator, and provide a consistency result under a set of high-level conditions. In Section 3 we connect our work to the literature and, notably, to Cattaneo, Jansson and Newey (2018) and Kline, Saggio and Sølvsten (2019). In Section 4 we provide primitive conditions for three special cases of our setup. In Section 5 we present and discuss the results from Monte Carlo experiments and apply our variance estimator to perform inference on the union wage premium. A short conclusion ends the paper. The supplemental appendix contains

technical details and additional simulation results.

2 Inference with many regressors

2.1 Framework

Consider the linear model

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

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 $\boldsymbol{\beta}$ 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, in general.

The (ordinary) least-squares estimator of $\boldsymbol{\beta}$ 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) and we closely follow their notation.

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

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

4

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 uncorrelated 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 standard random sampling but also repeated-measurement data (such as short panel data, for example) where strata are independent but dependence between observations within the strata is allowed, for example.

The second assumption contains regularity conditions. We let

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

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,*

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

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 situations 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 improve. 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 $\boldsymbol{w}_{i,n}$. Given that

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

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

Assumptions 1–3 co-incide with Assumptions 1–3 in Cattaneo, Jansson and Newey (2018). Consequently, by their Theorem 1,

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

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 an estimator of $\boldsymbol{\Omega}_n$, and thus of

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

When $q_n/n \nrightarrow 0$ and the errors are permitted to be heteroskedastic, the construction of a consistent estimator is non-trivial. To appreciate the problem, consider the estimator of Eicker (1963, 1967) and White (1980), which uses

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

where $\hat{u}_{i,n} := \sum_{j=1}^n (\boldsymbol{M}_n)_{i,j} \, (y_{j,n} - \boldsymbol{x}'_{j,n} \hat{\boldsymbol{\beta}}_n)$ are the least-squares residuals. This estimator is well known to be (conditionally) biased. The bias arises from the sampling noise in the least-squares estimator and can be severe (Chesher and Jewitt, 1987). Unless $q_n/n \to 0$ as $n \to \infty$, some observations will remain influential, in the sense that maximal leverage does not vanish. This causes the bias in $\hat{\boldsymbol{\Sigma}}_n$ to persist in large samples, implying that it is inconsistent.

The alternative to $\hat{\boldsymbol{\Sigma}}_n$ that we consider in this paper is

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

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 $\boldsymbol{\beta}$ 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.

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

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

and collect one such set of conditions 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 a small-bias condition; it 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 very weak. Providing primitive conditions for them in great generality appears to be difficult. However, when $\boldsymbol{w}_{i,n}$ is multivariate normal 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 (2020) 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 $y_{i,n} \acute{u}_{i,n}$ depends on $\mu_{i,n}^2$. The requirement that $\max_i \mu_{i,n}^2 = o_p(n)$ allows to control the variance of $\acute{\boldsymbol{\Sigma}}_n$. 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 $\chi_n = O(1)$. Primitive conditions for Assumption 4 in three special cases are given below.

We can now state our consistency result.

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

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

as $n \to \infty$.

Theorem 1, combined with the limit result in (2.2), implies that

$$\acute{\boldsymbol{\Omega}}_n^{-1/2}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \overset{d}{\to} \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{I}_r)$$

as $n \to \infty$, where

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

This result permits the construction of test statistics that (in large samples) will have correct size and of confidence regions that will exhibit correct coverage.

3   Connections to the literature

HC-class estimators.   The bias in the Eicker (1963, 1967) and White (1980) estimator has led to a variety of modifications to it being proposed that, following MacKinnon and White (1985), are often referred to as the HC-class of covariance-matrix estimators. These estimators are reviewed in Long and Ervin (2000) and MacKinnon (2012). Unfortunately, as shown by Cattaneo, Jansson and Newey (2018, Theorem 3), none of these alternatives is consistent, in general, when $q_n/n \not\to 0$. We briefly review their main findings on these estimators here.

The first variance estimator, HC1, differs from the conventional estimator, HC0, in that it performs a degrees-of-freedom correction (Hinkley, 1977, Eq. 2.11). This estimator will be consistent in the special case where errors are homoskedastic and the covariate design is balanced, i.e., when $(\boldsymbol{M}_n)_{1,1} = \ldots = (\boldsymbol{M}_n)_{n,n}$.

The second variance estimator, HC2, uses[2]

$$\sum_{i=1}^n \hat{\boldsymbol{v}}_{i,n}\hat{\boldsymbol{v}}'_{i,n}(\hat{u}_{i,n}\,\acute{u}_{i,n})$$

---

[2]We follow Cattaneo, Jansson and Newey (2018) and construct the HC-class variance estimators using the projection matrix $\boldsymbol{M}_n$. The original proposals were made in a context where q_n is treated as fixed and did not differentiate between $\boldsymbol{x}_{i,n}$ and $\boldsymbol{w}_{i,n}$; they used the annihilator matrix that projects out both sets of variables. The difference is asymptotically negligible under our assumptions.

as an estimator of $\boldsymbol{\Sigma}_n$ (Horn, Horn and Duncan, 1975). This estimator will be consistent under homoskedasticity. Because

$$y_{i,n}\acute{u}_{i,n} = \hat{u}_{i,n}\,\acute{u}_{i,n} + \hat{y}_{i,n}\,\acute{u}_{i,n},$$

where $\hat{y}_{i,n} := y_{i,n} - \hat{u}_{i,n}$ are fitted values, $\acute{\boldsymbol{\Sigma}}_n$ can be interpreted as a bias-corrected version of the HC2 estimator.

The third variance estimator, HC3, is constructed with

$$\sum_{i=1}^{n}\hat{\boldsymbol{v}}_{i,n}\hat{\boldsymbol{v}}'_{i,n}(\acute{u}_{i,n}\,\acute{u}_{i,n})$$

(MacKinnon and White 1985). While this estimator is inconsistent, its probability limit exceeds $\boldsymbol{\Sigma}_n$ (in the matrix sense). It follows that (under Assumptions 1–3) test procedures based on HC3 will be asymptotically conservative when $\limsup_n q/n \in (0,1)$, both under homoskedasticity and heteroskedasticity.

Bias-corrected estimation. Cattaneo, Jansson and Newey (2018) also considered the estimator

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

where $\boldsymbol{M}_n * \boldsymbol{M}_n$ denotes the elementwise product of the matrix $\boldsymbol{M}_n$. This estimator 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 a refinement under homoskedasticity. For $\dot{\boldsymbol{\Sigma}}_n$ to be well defined, $\boldsymbol{M}_n * \boldsymbol{M}_n$ needs to be positive semi-definite. Necessary and sufficient conditions for this to be the case are stated in Mallela (1972) but these are neither simple nor intuitive (Horn, Horn and Duncan, 1975). As noted by Horn and Horn (1975), a simple sufficient condition is that

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

Depending on the problem at hand it may also be necessary; an example is the one-way panel model.

10

Cattaneo, Jansson and Newey (2018, Theorem 4) show that $\grave{\boldsymbol{\Sigma}}_n$ is consistent for $\boldsymbol{\Sigma}_n$ if

$$\Pr\left(\min_i(\boldsymbol{M}_n)_{i,i} > \frac{1}{2}\right) \to 1, \qquad \frac{1}{\min_i(\boldsymbol{M}_n)_{i,i} - \frac{1}{2}} = O_p(1), \tag{3.3}$$

are added to Assumptions 1–3. Because $\sum_{i=1}^n (\boldsymbol{M}_n)_{i,i} = n - q_n$, $\min_i(\boldsymbol{M}_n)_{i,i} \le 1 - q_n/n$, and so

$$\limsup_n q_n/n < \frac{1}{2}$$

is required for (3.3) to be satisfied. This, in turn, is a strengthening of the condition that $\limsup_n q_n/n < 1$ in Assumption 2.

Stock and Watson (2008) proposed a covariance-matrix estimator for linear fixed-effect models that is applicable to short panel data. It is based on an explicit calculation of the probability limit of $\hat{\boldsymbol{\Sigma}}_n - \boldsymbol{\Sigma}_n$, adjusting the inconsistent $\hat{\boldsymbol{\Sigma}}_n$ by subtracting from it a plug-in estimator of this limit quantity. As discussed in Cattaneo, Jansson and Newey (2018), $\grave{\boldsymbol{\Sigma}}_n$ can be understood to be a generalization of this approach to the generic setting where $q_n/n \nrightarrow 0$.

Like $\hat{\boldsymbol{\Sigma}}_n$, the estimator $\grave{\boldsymbol{\Sigma}}_n$ need not be positive semi-definite in small samples. Being based on estimators of the individual $\sigma_{i,n}^2$ that are linear combinations of $\hat{u}_{1,n}^2, \ldots, \hat{u}_{n,n}^2$ it does, however, retain the invariance to changes in $\mu_{i,n}$ that is inherent in the HC-class estimators. This is a desirable feature and explains why a restriction on the magnitude of $\mu_{i,n}$ is not needed for this estimator to be consistent. Furthermore, jointly estimating $\sigma_{1,n}^2, \ldots, \sigma_{n,n}^2$ in this way is attractive from an efficiency point of view, as it is asymptotically equivalent to a minimum-norm unbiased estimator (see Hartley, Rao and Kiefer 1969 and Rao 1970 for details).

Inference on variance components. The variance estimator $\acute{\boldsymbol{\Sigma}}_n$ is closely related to the work of Kline, Saggio and Sølvsten (2019). In our context, their proposal is to estimate each $\sigma_{i,n}^2$ by the cross-fit (Newey and Robins, 2018) estimator

$$y_{i,n}\,\breve{u}_{i,n},$$

where $\breve{u}_{i,n}$ is the residual for observation i when the slope coefficients are estimated from the sample from which the ith observation has been omitted. When the regression model is

11

correctly specified—i.e., when $e_{i,n} = 0$—then $y_{i,n}\,\breve{u}_{i,n}$ is (conditionally) unbiased, provided that the leave-own-out regression slopes are well defined. This will be the case if maximal leverage is bounded away from one.[3] In this case, from the Sherman-Morrison formula (as in, e.g., Miller 1974),

$$y_{i,n}\,\breve{u}_{i,n} = y_{i,n}\acute{u}_{i,n} + o_p(1),$$

under our assumptions, connecting the cross-fit estimator of Kline, Saggio and Sølvsten (2019) to our approach.

Kline, Saggio and Sølvsten (2019) use this device to construct unbiased estimators of quadratic forms and present several applications. One of these (Kline, Saggio and Sølvsten 2019, Remark 4 and Lemma 5) is a consistency result for the implied covariance-matrix estimator

$$\sum_{i=1}^{n} \hat{\boldsymbol{v}}_{i,n}\hat{\boldsymbol{v}}'_{i,n}\left(y_{i,n}\,\breve{u}_{i,n}\right)$$

in designs with fixed regressors where errors are independent and maximal leverage is bounded away from unity. This result is established under the assumption that $e_{i,n} = 0$ and that $\max_i \|\mu_{i,n}\|^2 = O(1)$, together with standard regularity conditions as those in Assumption 2. It follows that our variance estimator can be seen as a modification of theirs that is targeted to a setting with many control variables. Theorem 1 may then be understood to be an extension of their Lemma 5 to settings with stochastic regressors and (vanishing) misspecification bias, where the regressors can have unbounded support and observations may depend on a growing number of parameters.

The implied interpretation of $\acute{\boldsymbol{\Sigma}}_n$ as an (approximate) cross-fit estimator is also 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 be consistent when we have $\limsup_n q_n/n > \frac{1}{2}$. A 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 two least-squares residuals, each one obtained from a (conditionally) independent

---

[3]This is a slight strengthening of our requirement in Assumption 4 as we project out only $\boldsymbol{w}_{i,n}$ to obtain $\acute{u}_{i,n}$ while Kline, Saggio and Sølvsten (2019) project out both the regressors $\boldsymbol{x}_{i,n}$ and the covariates $\boldsymbol{w}_{i,n}$ to obtain $\breve{u}_{i,n}$.

subsample that excludes the ith observation. Because these auxiliary regressions can be based on at most $\lfloor (n-1)/2 \rfloor$ and $\lceil (n-1)/2 \rceil$ observations such an approach cannot accommodate situations where $q_n/n > \frac{1}{2}$. The alternative estimators $y_{i,n} \, \check{u}_{i,n}$ and $y_{i,n} \, \acute{u}_{i,n}$ circumvent 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 $\limsup_n q_n/n > \frac{1}{2}$ but makes the variance estimator sensitive to $\mu_{i,n}$.

4   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) already gave such conditions for the other assumptions—and, notably, for Assumption 3—to hold.

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 $\hat{\boldsymbol{\beta}}_n$, then, is the least-squares estimator of $\boldsymbol{\beta}$ in

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

Note that $\mathbb{E}(u_{i,n} | \boldsymbol{x}_i, \boldsymbol{z}_i) \neq 0$, in general. Consistency of $\hat{\boldsymbol{\beta}}_n$ requires that $\kappa_n \to \infty$ (and, thus, that $q_n \to \infty$) as $n \to \infty$. Here,

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

In this example the number of covariates can be large 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

13

underlying functions are not (assumed to be) very smooth, so that a large κ_n needs to be used to control bias.

Standard smoothness conditions on φ imply that $n\varrho_n = O(1)$ (Newey, 1997), yielding the first condition of Assumption 4. The fourth condition can be validated in the same way, by imposing sufficient smoothness on $\mathbb{E}(\boldsymbol{x}_{i,n}|\boldsymbol{w}_{i,n})$ to get $n\chi_n = O(1)$, which implies that $\sum_{i=1}^n \|\tilde{\boldsymbol{Q}}_{i,n}\|^4 = O_p(n)$. This is different from (and stronger than) the first set of primitive conditions discussed by Cattaneo, Jansson and Newey (2018) to validate Assumption 3, where more smoothness in $\mathbb{E}(\boldsymbol{x}_{i,n}|\boldsymbol{w}_{i,n})$ can be used to compensate for less smoothness in φ. Alternatively, if $\chi_n \leq \mathbb{E}(\|\mathbb{E}(\boldsymbol{x}_{i,n}|\boldsymbol{w}_{i,n})\|^2) = O(1)$ and a partitioning estimator (Cattaneo and Farrell 2013) is used to approximate φ then $\boldsymbol{M}_n$ is a band matrix. This is also sufficient to reach the desired result. Moreover, in this case, the first and fourth condition of Assumption 4 are implied by the rate requirements in Assumption 3, and by the second set of restrictions for this example given in Cattaneo, Jansson and Newey (2018). Next, simple sufficient conditions for the requirement that $\max_i\|\mu_{i,n}\|/\sqrt{n} = o_p(1)$ are the moment conditions $\mathbb{E}(\|\boldsymbol{x}_i\|^{2+\theta}) = O(1)$ and $\mathbb{E}(\|\varphi(\boldsymbol{z}_i)\|^{2+\theta}) = O(1)$ for some $\theta > 0$. While these two conditions are not imposed in Cattaneo, Jansson and Newey (2018), they would not appear to be overly strong.

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

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

where α_g is a group-specific intercept and we assume that $\mathbb{E}(\varepsilon_{(g,m)}|\boldsymbol{x}_{(g,1)}, \ldots, \boldsymbol{x}_{(g,M)}) = 0$. The regressors $\boldsymbol{x}_{(g,m)}$ are assumed independent between groups but may be dependent within each group. The errors $\varepsilon_{(g,m)}$ are independent between groups and (conditionally) uncorrelated within groups. The usual asymptotic embedding here has $G_n \to \infty$ with $M = O(1)$. The number of fixed effects grows at the same rate as the sample size; we have $n = G_n \times M$ and $q_n = G_n$ so that $q_n/n = 1/M$, which does not vanish. These conditions fit Assumption 1.

The fixed-effect estimator is the least-squares estimator of $y_{(g,m)}$ on $\boldsymbol{x}_{(g,m)}$ and G_n dummy variables that indicate group membership. The estimated coefficients on these dummies are computed from M observations and are not consistent under our asymptotic approximation. In this example $\boldsymbol{M}_n = \boldsymbol{I}_{G_n} \otimes \boldsymbol{T}_M$, where $(\boldsymbol{T}_M)_{m,m'} := \{m = m'\} - M^{-1}$ is the $M \times M$ matrix that transforms observations into deviations from their within-group mean. Consequently,

$$\acute{\boldsymbol{\Sigma}}_n = \frac{M}{M-1} \sum_{g=1}^{G_n} \sum_{m=1}^{M} \tilde{\boldsymbol{x}}_{(g,m)} \, \tilde{\boldsymbol{x}}'_{(g,m)} \, y_{(g,m)} \, (\tilde{y}_{(g,m)} - \tilde{\boldsymbol{x}}'_{(g,m)} \hat{\boldsymbol{\beta}}_n),$$

where $\tilde{y}_{(g,m)} := \sum_{m'=1}^{M} (\boldsymbol{T}_M)_{m,m'} \, y_{(g,m')}$, and $\tilde{\boldsymbol{x}}_{(g,m)}$ and $\tilde{\varepsilon}_{(g,m)}$ are defined in the same way. Further, using that $\hat{\boldsymbol{\beta}}_n = \boldsymbol{\beta} + o_p(1)$,

$$\acute{\boldsymbol{\Sigma}}_n = \frac{M}{M-1} \sum_{g=1}^{G_n} \sum_{m=1}^{M} \tilde{\boldsymbol{x}}_{(g,m)} \, \tilde{\boldsymbol{x}}'_{(g,m)} \, \left(\varepsilon_{(g,m)} \tilde{\varepsilon}_{(g,m)} + \left(\boldsymbol{x}'_{(g,m)} \boldsymbol{\beta} + \alpha_g \right) \tilde{\varepsilon}_{(g,m)} \right) + o_p(G_n).$$

Because $\mathbb{E}(\varepsilon_{(g,m)} \tilde{\varepsilon}_{(g,m)} | \boldsymbol{x}_{(g,1)}, \ldots, \boldsymbol{x}_{(g,M)}) = \sigma^2_{(g,m)} (M-1)/M$ the first term constitutes an unbiased estimator of $\boldsymbol{\Sigma}_n$. The second term on the right-hand side is mean zero because the errors are mean-independent of the regressors. Its variance, however, depends on the $\alpha_1, \ldots, \alpha_{G_n}$.

In a two-wave panel,

$$\acute{\boldsymbol{\Sigma}}_n = \frac{1}{4} \sum_{g=1}^{G_n} \Delta \boldsymbol{x}_g \, \Delta \boldsymbol{x}'_g \, (\Delta y_g - \Delta \boldsymbol{x}'_g \hat{\boldsymbol{\beta}}_n) \, \Delta y_g,$$

where Δ denotes the first-difference operator; so, e.g., $\Delta y_g := y_{(g,2)} - y_{(g,1)}$. In this case the variance estimator does not depend on the fixed effects. The factor $\frac{1}{4}$ appears because of the de-meaning and is inconsequential. The implied estimator of the covariance matrix $\boldsymbol{\Omega}_n$ is

$$\left(\sum_{g=1}^{G_n} \Delta \boldsymbol{x}_g \, \Delta \boldsymbol{x}'_g \right)^{-1} \left(\sum_{g=1}^{G_n} \Delta \boldsymbol{x}_g \, \Delta \boldsymbol{x}'_g \, (\Delta y_g - \Delta \boldsymbol{x}'_g \hat{\boldsymbol{\beta}}_n) \, \Delta y_g \right) \left(\sum_{g=1}^{G_n} \Delta \boldsymbol{x}_g \, \Delta \boldsymbol{x}'_g \right)^{-1}.$$

The same estimator would be obtained if our procedure would be applied directly to the first-differenced model $\Delta y_g = \Delta \boldsymbol{x}'_g \boldsymbol{\beta} + \Delta \varepsilon_g$. The standard estimator of Eicker (1963, 1967)

and White (1980) applied to this model is

$$\left(\sum_{g=1}^{G_n} \Delta\boldsymbol{x}_g\,\Delta\boldsymbol{x}_g'\right)^{-1}\left(\sum_{g=1}^{G_n} \Delta\boldsymbol{x}_g\,\Delta\boldsymbol{x}_g'\,(\Delta y_g - \Delta\boldsymbol{x}_g'\hat{\boldsymbol{\beta}}_n)^2\right)\left(\sum_{g=1}^{G_n} \Delta\boldsymbol{x}_g\,\Delta\boldsymbol{x}_g'\right)^{-1},$$

and is known to be consistent here.

In the one-way panel model our Assumption 4 holds provided that

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

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 condition on the fixed effects is the only additional requirement needed for our variance estimator to be consistent in this model.

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.,

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

as in Cattaneo, Jansson and Newey (2018). The generic nature of this model makes it difficult to specify a single set of simple primitive conditions for our Assumption 4 to hold. First consider the requirement that $\sum_{i=1}^{n}\|\tilde{\boldsymbol{Q}}_{i,n}\|^4 = O_p(n)$. A sufficient condition here is 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, if $\chi_n = O(1)$, a sparsity condition on the projection matrix $\boldsymbol{M}_n$ can be used. One such condition is $\max_i n_{i,n} = O_p(1)$, together with $\sum_{i=1}^{n}\mathbb{E}(\|\boldsymbol{Q}_{i,n}\|^4) = O(n)$, where we have introduced $n_{i,n} := \sum_{j=1}^{n}\{(\boldsymbol{M}_n)_{i,j} \neq 0\}$. Cattaneo, Jansson and Newey (2018) showed that the weaker

condition $\max_i n_{i,n} = o_p(n^{1/3})$ can be used to support Assumption 3 under the same moment condition on the $\boldsymbol{Q}_{i,n}$. Such a rate would appear difficult to obtain here. 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 that is 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

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

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 to 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 $\boldsymbol{w}_{i,n}$. When the covariates are normally distributed or have bounded support, for example, we can allow for $\max_i \kappa_{i,n} = O(\sqrt{n})$.

5 Numerical illustrations

5.1 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 $\boldsymbol{\gamma}_n = \boldsymbol{0}$. 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.

We consider three designs that vary in β and π. Design A is the design of Cattaneo, Jansson and Newey (2018). It has $\beta = 1$ and $\pi = .02$.[4] Each dummy variable takes on the value one for about 14 out of 700 observations, on average. Design B is a more sparse design where $\beta = 1$ is maintained and π is reduced to .01, leading to each dummy switching on for only 7 observations in each replication, on average. Design C, in turn, sets $\beta = 2$ and maintains $\pi = .02$. This implies that the conditional variance of $y_{i,n}\, \acute{u}_{i,n}$ increases by a factor of four.

The results for the three design variations are presented in Tables 1–3. Empirical size of the two-sided t-test of the null that $\beta = 1$ and the average width of the corresponding confidence interval are given for all variance estimators discussed; HC0, its modifications HC1, HC2, and HC3, the bias-corrected estimator of Cattaneo, Jansson and Newey (2018) (HCK) and the estimator presented here (HCA).

As the setup features homoskedastic errors, inference based on HC0 will be liberal for large n when q_n/n is not small (Chesher and Jewitt 1987, Cattaneo, Jansson and Newey 2018). This is apparent from inspection of the tables. The degrees-of-freedom correction performed by HC1 alleviates most of this concern here. HC2, which is consistent under homoskedasticity, performs quite similarly to HC1. Both corrections do come with (on average) wider confidence intervals. The simulations also illustrate that HC3 yields conservative inference. The rejection frequency (under the null) approaches zero as q_n/n increases. The confidence intervals are also wide. This implies that tests based on HC3 will suffer from low power. The relative inefficiency compared to HC1 and HC2 grows as q_n/n increases.

In Design A HCK gives close to correct size and confidence intervals of a comparable length as HC1 and HC2 for most of the values of q_n/n. As q_n/n increases above $\frac{1}{2}$ this variance estimator does not always exist in each replication. When this happens in our Monte Carlo, HCK defaults to HC0. Consequently, for the larger values of q_n/n in Table 1, the size of the HCK method increases above its nominal size and the average length of

---

[4]The description of the simulation design in Cattaneo, Jansson and Newey (2018, p. 1358) contains a typo that would imply that $\pi = .0062$.

the confidence interval equally shrinks relative to HC1 and HC2. In the sparser Design B the non-existence of the HCK estimator is more frequent and also arises for smaller values of q_n/n. This explains the large overrejection rates observed in Table 2. The results for Design C in Table 3 are similar to those in Table 1 as the performance of HCK is invariant to the scale of the regression slopes.

The simulation results also confirm the theory behind our variance estimator. It yields close to correct inference in all three designs and for all values of q_n/n considered, although some more overrejection is observed for the top-end values of this ratio. The average length of its confidence interval is comparable to those obtained for HC1, HC2, and HCK and is substantially smaller than those for HC3 for large values of q_n/n. These results show that this estimator is useful in problems with many covariates when HCK is not available, for example due to sparseness of the regressor design or because of the presence of high-leverage observations more generally.

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

5.2   Empirical example

We next use the different variance estimators available to infer the union membership premium. The data are a balanced panel on 545 working individuals and span 8 years (1980–1987), giving a total of 4,360 observations. They are taken from Vella and Verbeek (1998) and are available from the data archive of the *Journal of Applied Econometrics* (http://qed.econ.queensu.ca/jae/1998-v13.2/vella-verbeek/).

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. This set contains average hours worked per week, dummies for marital status and for poor health, a quadratic term in years of experience, and a large collection of dummies, as follows. First, as the data are a panel, both individual fixed effects and year fixed effects are included. Second, as the data contain information on the type of occupation

(out of 9 categories) and the industry (out of a total of 12) in which the job is located we also control for these by including sector and industry dummies. Third, we allow for interaction effects between these categorical variables by including occupation-by-year, sector-by-year, and occupation-by-sector dummies as well as occupation-by-sector-by-year dummies. Baseline categories for the year-, sector-, and occupational dummies are chosen and their corresponding dummies are dropped as to avoid a dummy-variable trap. Certain interactions of the occupation and sector dummies never take on the value one and so are equally removed from the analysis. This gives a total of 1,086 control variables relative to 4,360 observations. Our point estimate of the union premium is 7.43%. This is in line with the literature (see, e.g., Jakubson 1991).

The standard errors on this point estimate, obtained by the various methods discussed, together with the implied 95% confidence intervals for the union premium are collected in Table 4. A total of 330 observations have leverage that exceeds the one-half threshold. 36 of these observations can be perfectly explained by the control variables because of the inclusion of the occupation-by-sector and the occupation-by-sector-by-year dummies and so contain no information on the union premium. The HCK variance estimator could not be computed here. (We note that the invertibility condition fails also after dropping the 36 non-informative observations.) This explains why an entry for this estimator is not available. HC0 yields the smallest standard error on our point estimate (1.72%). HC3 yields the largest (2.32%). The standard error of HCA (1.93%) is roughly in the middle of these two bounds. HC1 and HC2 give similar, slightly larger, standard errors as HCA (1.99% and 1.98%).

6 Conclusion

This paper has presented a heteroskedasticity-robust covariance-matrix estimator for linear regression models that is consistent under an asymptotic scheme where the number of control variables, q_n, grows at the same rate as the sample size, n. The estimator is similar to the proposal of Kline, Saggio and Sølvsten (2019) but our consistency result

covers more general settings. The estimator complements work by Cattaneo, Jansson and Newey (2018), who derived inconsistency results for members of the HC-class of variance estimators, proved asymptotic-conservativeness of the HC3 estimator, and presented an alternative variance estimator (that is based on Hartley, Rao and Kiefer 1969) that remains consistent when $\limsup_n q_n/n < \frac{1}{2}$. Under a set of additional high-level conditions, our estimator allows to weaken this restriction to $\limsup_n q_n/n < 1$. Primitive conditions for these where given for partially-linear models, fixed-effect panel data models, and generic regression models with increasing dimension. Simulation verify the theoretical properties. An empirical application to estimation of the union premium from panel data was also presented.

The idea underlying our variance estimator can be useful as a device to correct for bias more generally. Kline, Saggio and Sølvsten (2019) use it to bias-correct quadratic forms in fixed-effect estimators. Chernozhukov et al. (2018) and Newey and Robins (2018) use related cross-fitting techniques to reduce bias in high-dimensional estimation problems that feature machine-learning estimators. Cattaneo, Jansson and Ma (2019) characterize the bias in (nonlinear) two-step estimators when the first step features a high-dimensional linear regression.

References

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

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. and M. H. Farrell (2013). Optimal convergence rates, Bahadur representation, and asymptotic normality of partitioning estimators. *Journal of Econometrics 174*, 127–143.

Cattaneo, M. D., M. Jansson, and X. Ma (2019). Two-step estimation and inference with possibly many included covariates. *Review of Economic Studies 86*, 1095–1122.

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.

Hinkley, D. V. (1977). Jackknifing in unbalanced situations. *Technometrics 19*, 285–292.

Hoaglin, D. C. and R. E. Welsch (1978). The hat matrix in regression and ANOVA. *The American Statistician 32*, 17–22.

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.

Jakubson, G. (1991). Estimation and testing of the union wage effect using panel data. *Review of Economic Studies 58*, 971–991.

Kline, P., R. Saggio, and M. Sølvsten (2019). 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.

MacKinnon, J. G. (2012). Thirty years of heteroscedasticity-robust inference. In X. Chen and N. R. Swanson (Eds.), *Recent Advances and Future Directions in Causality, Prediction, and Specification Analysis*, pp. 437–461. Springer: New York, NY.

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

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.

Miller, R. G. (1974). An unbalanced jackknife. *Annals of Statistics 2*, 880–891.

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

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

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

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

Vella, F. and M. Verbeek (1998). Whose wages do unions raise? A dynamic model of unionism and wage rate determination for young man. *Journal of Applied Econometrics 13*, 163–183.

Verdier, V. (2020). Estimation and inference for linear models with two-way fixed effects and sparsely matched data. *Review of Economics and Statistics 102*, 1–16.

White, H. (1980). A heteroskedasticity-consistent covariance matrix and a direct test for

heteroskedasticity. *Econometrica 48*, 817–838.

Table 1: Design A ($\beta = 1$ and $\pi = .02$)

q_n	1	71	141	211	281	351	421	491	561	631
q_n/n	.0014	.1014	.2014	.3014	.4014	.5014	.6014	.7014	.8014	.9014
Rejection frequency of 5%-level test										
HC0	.0502	.0671	.0771	.1050	.1287	.1605	.2054	.2746	.3691	.5309
HC1	.0495	.0518	.0485	.0534	.0531	.0459	.0426	.0476	.0438	.0446
HC2	.0498	.0529	.0517	.0567	.0574	.0505	.0495	.0556	.0530	.0589
HC3	.0495	.0416	.0280	.0241	.0131	.0058	.0022	.0005	.0001	.0000
HCK	.0498	.0530	.0520	.0588	.0593	.0563	.0598	.0770	.0976	.2688
HCA	.0519	.0535	.0529	.0570	.0566	.0524	.0521	.0605	.0567	.0674
Average width of 95% confidence interval										
HC0	.1477	.1484	.1490	.1494	.1502	.1510	.1517	.1525	.1533	.1541
HC1	.1479	.1567	.1669	.1789	.1944	.2141	.2408	.2797	.3452	.4944
HC2	.1478	.1560	.1654	.1766	.1909	.2093	.2340	.2698	.3300	.4646
HC3	.1479	.1645	.1851	.2113	.2468	.2964	.3707	.4937	.7405	1.4803
HCK	.1478	.1559	.1651	.1758	.1894	.2065	.2292	.2606	.3064	.3482
HCA	.1477	.1557	.1652	.1763	.1907	.2087	.2334	.2688	.3284	.4611

Table 2: Design B ($\beta = 1$ and $\pi = .01$)

q_n	1	71	141	211	281	351	421	491	561	631
q_n/n	.0014	.1014	.2014	.3014	.4014	.5014	.6014	.7014	.8014	.9014
Rejection frequency of 5%-level test										
HC0	.0479	.0606	.0767	.0963	.1137	.1457	.1907	.2450	.3353	.4639
HC1	.0474	.0462	.0479	.0471	.0410	.0432	.0390	.0340	.0291	.0256
HC2	.0477	.0477	.0526	.0523	.0496	.0543	.0553	.0544	.0555	.0584
HC3	.0474	.0376	.0301	.0227	.0121	.0050	.0026	.0005	.0000	.0000
HCK	.0477	.0582	.0765	.0963	.1136	.1457	.1907	.2450	.3353	.4639
HCA	.0492	.0512	.0529	.0540	.0523	.0566	.0571	.0553	.0602	.0705
Average width of 95% confidence interval										
HC0	.1479	.1492	.1504	.1520	.1537	.1561	.1588	.1621	.1671	.1768
HC1	.1481	.1575	.1685	.1820	.1989	.2212	.2518	.2971	.3757	.5648
HC2	.1480	.1560	.1653	.1765	.1905	.2089	.2335	.2691	.3291	.4633
HC3	.1481	.1645	.1848	.2109	.2459	.2953	.3688	.4908	.7352	1.4661
HCK	.1480	.1501	.1507	.1520	.1537	.1561	.1588	.1621	.1671	.1768
HCA	.1479	.1557	.1649	.1763	.1903	.2085	.2333	.2686	.3278	.4590

Table 3: Design C ($\beta = 2$ and $\pi = .02$)

q_n	1	71	141	211	281	351	421	491	561	631
q_n/n	.0014	.1014	.2014	.3014	.4014	.5014	.6014	.7014	.8014	.9014
Rejection frequency of 5%-level test										
HC0	.0505	.0667	.0800	.1052	.1273	.1611	.2045	.2653	.3708	.5273
HC1	.0501	.0535	.0495	.0508	.0490	.0466	.0464	.0488	.0462	.0453
HC2	.0504	.0542	.0519	.0536	.0529	.0518	.0521	.0563	.0545	.0583
HC3	.0501	.0420	.0310	.0209	.0117	.0062	.0019	.0006	.0000	.0000
HCK	.0504	.0545	.0525	.0557	.0553	.0577	.0631	.0763	.1026	.2635
HCA	.0555	.0580	.0584	.0612	.0588	.0567	.0624	.0673	.0710	.0935
Average width of 95% confidence interval										
HC0	.1480	.1483	.1489	.1496	.1502	.1508	.1515	.1523	.1535	.1543
HC1	.1482	.1566	.1668	.1792	.1944	.2139	.2405	.2794	.3458	.4950
HC2	.1481	.1559	.1654	.1769	.1909	.2091	.2336	.2696	.3304	.4652
HC3	.1482	.1644	.1851	.2116	.2468	.2961	.3698	.4933	.7411	1.4817
HCK	.1481	.1558	.1650	.1761	.1894	.2064	.2286	.2601	.3066	.3491
HCA	.1472	.1552	.1644	.1761	.1900	.2080	.2320	.2673	.3266	.4535

Table 4: Inference on the union premium

	std. error	95% conf. int.
$\hat{\beta}_n = .0743$		
HC0	.0172	[.0406, .1081]
HC1	.0199	[.0354, .1133]
HC2	.0198	[.0356, .1131]
HC3	.0232	[.0288, .1199]
HCA	.0193	[.0365, .1122]

HETEROSKEDASTICITY-ROBUST INFERENCE IN LINEAR REGRESSION MODELS WITH MANY COVARIATES

SUPPLEMENTAL APPENDIX

KOEN JOCHMANS[*]

UNIVERSITY OF CAMBRIDGE

This supplemental appendix contains the proof of Theorem 1 (in Section A.1), derivations underlying the primitive conditions in Section 4 (in Section A.2), and additional simulations results (in Section B).

A   Technical details

A.1   Proof of Theorem 1

We need to show that

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

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

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

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

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

[*]Address: University of Cambridge, Faculty of Economics, Austin Robinson Building, Sidgwick Avenue, Cambridge CB3 9DD, United Kingdom. E-mail: `kj345@cam.ac.uk`.

Next, letting $c_{i,j} := \mathbb{E}(\varepsilon_{i,n}^2 \varepsilon_{j,n}^2 | \mathcal{X}_n, \mathcal{W}_n) - \sigma_{i,n}^2 \sigma_{j,n}^2$ and noting that

$$\|c_{i,j}\| \leq \max_i \mathbb{E}(\varepsilon_{i,n}^4 | \mathcal{X}_n, \mathcal{W}_n) + \max_i \sigma_{i,n}^4 =: c_n,$$

Assumptions 1–3 imply that

$$\mathbb{E}\left(\left. \left(\frac{\sum_{i=1}^n \hat{v}_{i,n}^2 (\varepsilon_{i,n}^2 - \sigma_{i,n}^2)}{n} \right)^2 \right| \mathcal{X}_n, \mathcal{W}_n \right) = \frac{\sum_{g=1}^{G_n} \sum_{i \in N_g} \sum_{j \in N_g} \hat{v}_{i,n}^2 \hat{v}_{j,n}^2 \, c_{i,j}}{n^2}$$

$$\leq c_n \left(\max_g |N_g| \right) \left(\max_i \frac{\|\hat{v}_{i,n}\|}{\sqrt{n}} \right)^2 \frac{\sum_{i=1}^n \hat{v}_{i,n}^2}{n}$$

$$= o_p(1)$$

because $c_n = O_p(1)$, $\max_g |N_g| = O(1)$, $\max_i \|\hat{v}_{i,n}\|/\sqrt{n} = o_p(1)$, and

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

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,

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

and (A.1) reduces to

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

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

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

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 recalling 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

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

We will handle each of these five terms in turn.

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

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

because the $\varepsilon_{i,n}$ are (conditionally) uncorrelated by Assumption 1. Next, to see that also

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

first expand the square to get

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

3

Conditional independence of the $\varepsilon_{i,n}$ across groups g, g' in the partition of the observations and zero correlation within each group g implies that the summand will be zero unless either (i) $i_1 = i_2$ and $j_1 = j_2$, with $i_1 \in N_g$ and $j_1 \in N_{g'}$; or (ii) $i_1 = j_2$ and $i_2 = j_1$, with $i_1 \in N_g$ and $i_2 \in N_{g'}$; or (iii) $(i_1, i_2, j_1, j_2) \in N_g$, where $g \neq g'$. The contribution of Case (i) terms equals

$$
\frac{\sum_{g=1}^{G_n} \sum_{g' \neq g} \sum_{i \in N_g} \sum_{j \in N_{g'}} \hat{v}_{i,n}^4 \, \sigma_{i,n}^2 (\boldsymbol{A}_n)_{i,j}^2 \, \sigma_{j,n}^2}{n^2} \leq \frac{(\max_i \sigma_{i,n}^2)^2 \sum_{i=1}^n \hat{v}_{i,n}^4 \sum_{j \neq i} (\boldsymbol{A}_n)_{i,j}^2}{n^2}
$$

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

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

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

$$
= o_p(1),
$$

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 $(\max_i \sigma_{i,n}^2)(\min_i (\boldsymbol{M}_n)_{i,i})^{-1} = O_p(1)$. Similarly, the contribution of Case (ii) terms equals

$$
\frac{\sum_{g=1}^{G_n} \sum_{g' \neq g} \sum_{i \in N_g} \sum_{j \in N_{g'}} \hat{v}_{i,n}^2 \, \sigma_{i,n}^2 (\boldsymbol{A}_n)_{i,j} (\boldsymbol{A}_n)_{j,i} \, \hat{v}_{j,n}^2 \sigma_{j,n}^2}{n^2}
$$

and is bounded by

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

where, now, we use that $(\boldsymbol{A}_n)_{i,j}(\boldsymbol{A}_n)_{j,i} \geq 0$ to validate the first upper bound, 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 upper bound. Finally, the contribution of Case (iii) terms equals

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

and, using the Cauchy-Schwarz inequality, is bounded by

$$\frac{\left(\max_i \mathbb{E}(\varepsilon_{i,n}^4 | \mathcal{X}_n, \mathcal{W}_n)\right)\left(\max_g |N_g|\right)^3}{\left(\min_i (\boldsymbol{M}_n)_{i,i}\right)^2} \left(\max_i \frac{\|\hat{v}_{i,n}\|}{\sqrt{n}}\right)^2 \frac{\sum_{i=1}^n \hat{v}_{i,n}^2}{n} = o_p(1),$$

which follows by the same arguments. Equation (A.6) has been shown.

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

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

where we use that

$$((\boldsymbol{A}_n)_{i,i} - 1) = -\left(\frac{\sum_{i=1}^n \hat{v}_{i,n}^2}{n}\right)^{-1}\left(\frac{\hat{v}_{i,n}^2}{n}\frac{1}{(\boldsymbol{M}_n)_{i,i}}\right). \qquad (A.7)$$

This vanishes because $\left(n^{-1}\sum_{i=1}^n \hat{v}_{i,n}^2\right)^{-1} = O_p(1)$ (Cattaneo, Jansson and Newey 2018, Lemma SA-1) and

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

To see that

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

first expand the square and once again use (A.7) to see that the second moment can be written as

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

Applying the Cauchy-Schwarz inequality to $\mathbb{E}(\varepsilon_{i,n}^2\varepsilon_{j,n}^2|\mathcal{X}_n, \mathcal{W}_n)$ then yields the upper bound

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

this follows by the same argument as that just used for the first moment, only now using the fact that the fourth moment is uniformly bounded.

5

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

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

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 $\sum_{i=1}^{n} \sum_{j=1}^{n} a_i (\boldsymbol{H}_n)_{i,j} a_j \le \sum_{i=1}^{n} a_i^2$ for any $(a_1, \ldots, a_n)$. Consequently, the variance satisfies

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

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

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

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

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

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

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

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

and is bounded by

$$\frac{(\max_{i}\sigma_{i,n}^{2})}{(\min_{i}(\boldsymbol{M}_{n})_{i,i})^{2}}\left(\max_{i}\frac{\|\hat{v}_{i,n}\|}{\sqrt{n}}\right)^{2}\left(\frac{\sum_{i=1}^{n}\hat{v}_{i,n}^{2}}{n}\right)n\left(\frac{\sum_{i=1}^{n}e_{i,n}^{2}}{n}\right)=o_{p}(n\varrho_{n})=o_{p}(1),$$

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 $n\varrho_{n}=O(1)$, 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

$$\begin{aligned}\frac{\sum_{i=1}^{n}\mu_{i,n}^{2}(\sum_{j=1}^{n}(\boldsymbol{H}_{n})_{i,j}\,e_{j,n})^{2}}{n}&\leq\left(\max_{i}\frac{\|\mu_{i,n}\|}{\sqrt{n}}\right)^{2}\sum_{i=1}^{n}\left(\sum_{j=1}^{n}(\boldsymbol{H}_{n})_{i,j}\,e_{j,n}\right)^{2}\\&=\left(\max_{i}\frac{\|\mu_{i,n}\|}{\sqrt{n}}\right)^{2}\left(\sum_{j=1}^{n}\sum_{k=1}^{n}e_{j,n}\left(\sum_{i=1}^{n}(\boldsymbol{H}_{n})_{i,j}\,(\boldsymbol{H}_{n})_{i,k}\right)e_{k,n}\right)\\&=\left(\max_{i}\frac{\|\mu_{i,n}\|}{\sqrt{n}}\right)^{2}\left(\sum_{j=1}^{n}\sum_{k=1}^{n}e_{j,n}(\boldsymbol{H}_{n})_{j,k}\,e_{k,n}\right)\\&\leq\left(\max_{i}\frac{\|\mu_{i,n}\|}{\sqrt{n}}\right)^{2}\left(\sum_{j=1}^{n}e_{j,n}^{2}\right)=o_{p}(n\varrho_{n})=o_{p}(1),\end{aligned}$$

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. $\qquad\square$

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.

8

Condition (ii). We first note that

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

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

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

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

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

using the fact that

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

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,

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

9

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.

B   Additional simulation results

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

B.1   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 $(y_{(g,m)}, x_{(g,m)})$, the model is

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

and α_g is a group-specific intercept. The within-group (fixed-effect) estimator equals the ordinary least-squares estimator of $y_{(g,m)}$ on $x_{(g,m)}$ and G dummy variables that capture group membership of the individual observations. We draw $x_{(g,m)} \sim$ i.i.d. $\boldsymbol{N}(0,1)$ and $\varepsilon_{(g,m)} \sim$ i.i.d. $\boldsymbol{N}(0,1)$, set $\beta = 1$ and $\alpha_g = 0$ for all groups g. The samples sizes considered have $G = \lfloor 700/M \rfloor$ for $M \in \{700, 10, 5, 4, 3, 2\}$, which yields a total sample size of 700

(except when $M = 3$, in which case the sample size is 699). This is the same total sample size as in our simulations in the main text.

Table A.1 reports the simulation results for this model. In the case $M = 2$ HCK does not exist and the table reports results for the standard HC0 estimator applied to the first-differenced model.

B.2   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.

References

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.

Table A.1: One-way panel model

G	1	70	140	175	233	350
M	700	10	5	4	3	2
q_n/n	.0014	.1000	.2000	.2500	.3333	.5000
Rejection frequency of 5%-level test						
HC0	.0529	.0667	.0835	.0971	.1155	.1665
HC1	.0526	.0518	.0519	.0552	.0539	.0525
HC2	.0527	.0519	.0519	.0555	.0541	.0526
HC3	.0526	.0419	.0315	.0286	.0187	.0069
HCK	.0527	.0527	.0533	.0560	.0555	*.0526*
HCA	.0526	.0544	.0540	.0576	.0557	.0550
Average width of 95% confidence interval						
HC0	.1478	.1478	.1478	.1477	.1479	.1479
HC1	.1480	.1559	.1654	.1708	.1813	.2095
HC2	.1479	.1558	.1653	.1706	.1811	.2092
HC3	.1480	.1642	.1848	.1970	.2218	.2958
HCK	.1479	.1557	.1651	.1704	.1808	*.2092*
HCA	.1478	.1556	.1651	.1704	.1809	.2089

Table A.2: Partially-linear model

κ_n	1	2	3	4	5
q_n	7	28	84	210	462
q_n/n	.01	.04	.12	.30	.66
Rejection frequency of 5%-level test					
HC0	.0649	.0536	.0650	.0927	.1823
HC1	.0634	.0493	.0497	.0451	.0218
HC2	.0636	.0494	.0511	.0521	.0508
HC3	.0630	.0449	.0366	.0197	.0012
HCK	.0636	.0493	.0512	.0523	.0605
HCA	.0678	.0571	.0568	.0603	.0700
Width of 95% confidence interval					
HC0	.1387	.1476	.1481	.1520	.1741
HC1	.1395	.1508	.1580	.1819	.2992
HC2	.1394	.1506	.1573	.1765	.2527
HC3	.1401	.1537	.1677	.2110	.4335
HCK	.1394	.1506	.1572	.1763	.2496
HCA	.1391	.1498	.1563	.1754	.2496