

# Cambridge Working Papers in Economics

Cambridge Working Papers in Economics: 1957

## HETEROSKEDASTICITY-ROBUST INFERENCE IN LINEAR REGRESSION MODELS

Koen Jochmans

25 June 2019

This paper considers inference in heteroskedastic linear regression models with many control variables. The slope coefficients on these variables are nuisance parameters. Our setting allows their number to grow with the sample size, possibly at the same rate, in which case they are not consistently estimable. A prime example of this setting are models with many (possibly multi-way) fixed effects. The presence of many nuisance parameters introduces an incidental-parameter problem in the usual heteroskedasticity-robust estimators of the covariance matrix, rendering them biased and inconsistent. Hence, tests based on these estimators are size distorted even in large samples. An alternative covariance-matrix estimator that is conditionally unbiased and remains consistent is presented and supporting simulation results are provided.

# HETEROSKEDASTICITY-ROBUST INFERENCE IN LINEAR REGRESSION MODELS

KOEN JOCHMANS\*  
UNIVERSITY OF CAMBRIDGE

## Abstract

This paper considers inference in heteroskedastic linear regression models with many control variables. The slope coefficients on these variables are nuisance parameters. Our setting allows their number to grow with the sample size, possibly at the same rate, in which case they are not consistently estimable. A prime example of this setting are models with many (possibly multi-way) fixed effects. The presence of many nuisance parameters introduces an incidental-parameter problem in the usual heteroskedasticity-robust estimators of the covariance matrix, rendering them biased and inconsistent. Hence, tests based on these estimators are size distorted even in large samples. An alternative covariance-matrix estimator that is conditionally unbiased and remains consistent is presented and supporting simulation results are provided.

**Keywords:** bias, fixed effects, heteroskedasticity, inference, leave-one-out estimator, many regressors, unbalanced regressor design, robust covariance matrix, size control, statistical leverage.

---

\*Address: University of Cambridge, Faculty of Economics, Austin Robinson Building, Sidgwick Avenue, Cambridge CB3 9DD, U.K. E-mail: [kj345@cam.ac.uk](mailto:kj345@cam.ac.uk). Support of the European Research Council through grant n° 715787 (MiMo) is gratefully acknowledged.

[OLD:]

# 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—even when the errors are, in fact, homoskedastic—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 samples ([Huber, 1981](#)). 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,  $q$ , grows with the sample size,  $n$ , the robust covariance-matrix estimator will be inconsistent unless  $q/n \rightarrow 0$ , as formally 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](#), [Chetty, Friedman and Rockoff 2014](#), [Verdier 2018](#)) as well as the many variations of such regressions to problems with a similar structure.

A solution is to construct a covariance-matrix estimator that uses a (conditionally) unbiased estimator of the observation-specific error variances. As shown below this can be achieved by using a leave-one-out estimator of the slope coefficients. Under regularity conditions, the resulting covariance-matrix estimator will be consistent under asymptotics where  $q/n \rightarrow c$  as  $n \rightarrow \infty$  for any  $c < 1$ . Subsample estimators have a long history, originating with the jackknife ([Quenouille 1956](#), [Tukey 1958](#)), and have been found useful in many settings. They have been used in the estimation of covariance matrices, but not quite in the form considered here. Indeed, both the ‘almost-unbiased’ estimator of [Horn, Horn and Duncan \(1975\)](#) and the jackknife-type estimator of [MacKinnon and White \(1985\)](#) make use of them. However, as discussed in more detail below, both these estimators are biased, in general, and inconsistent unless  $q/n \rightarrow 0$ .

Our results build on and extend [Cattaneo, Jansson and Newey \(2018\)](#). They provided a rigorous derivation of the limit distribution of the least-squares estimator allowing for  $q/n \rightarrow 0$  and formally showed the inconsistency of the Eicker-White covariance estimator as well as of various alternatives available in the literature. They also showed consistency of a bias-corrected covariance-matrix estimator in the spirit of [Hartley, Rao and Kiefer \(1969\)](#) and [Bera, Suprayitno and Premaratne \(2002\)](#) under the requirement that  $q/n \rightarrow c$  for some  $c < \frac{1}{2}$ . This estimator can be seen as a generalization of the one proposed by [Stock and Watson \(2008\)](#) for one-way regression models. The low-leverage requirement is substantially stronger than our condition that  $q/n \rightarrow c$  for some  $c < 1$  and may be problematic. It does not hold for one-way models for two-wave data, for example, and it will typically not hold in matched data sets, such as the student-teacher setting and its

many variants mentioned above (Jochmans and Weidner 2018, Verdier 2018). Moreover, the simulations reported on below, shows that the performance of their estimator breaks down when  $q/n > \frac{1}{2}$  while our proposal continues to perform well, even when  $q/n$  is close to one.

## 2 Problem statement

Consider the linear regression model

$$y_i = \mathbf{x}_i' \boldsymbol{\beta} + \varepsilon_i, \quad i = 1, \dots, n,$$

where  $y_i$  is a scalar outcome,  $\mathbf{x}_i$  is an  $r$ -vector of covariates, and  $\varepsilon_i$  is the regression error. The  $n \times r$  design matrix  $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)'$  is taken to have rank  $r < n$ , so that the (ordinary) least-squares estimator of  $\boldsymbol{\beta}$ , say  $\hat{\boldsymbol{\beta}}$ , is well defined and the residuals are not all zero. Let  $E_{\mathbf{X}}$  denote expectations taken conditional on  $\mathbf{X}$ . We suppose that  $E_{\mathbf{X}}(\varepsilon_i) = 0$  for all  $i$  and allow for the regression errors to exhibit (conditional) heteroskedasticity, that is,

$$E_{\mathbf{X}}(\varepsilon_i \varepsilon_j) = \begin{cases} \sigma_i^2 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}.$$

Write  $\mathbf{X} = (\mathbf{A}, \mathbf{B})$ , where  $\mathbf{A}$  is  $n \times p$  and  $\mathbf{B}$  is  $n \times q$ , and partition  $\boldsymbol{\beta} = (\boldsymbol{\alpha}', \boldsymbol{\eta}')'$  accordingly. Our aim is to perform inference on  $\boldsymbol{\alpha}$  treating  $\boldsymbol{\eta}$  as a nuisance parameter. This reflects a setup where  $\mathbf{A}$  are the variables whose slope coefficients are of interest and the columns of  $\mathbf{B}$  serve as control variables. While  $p$  is treated as fixed we will allow  $q$  to grow with  $n$ , possibly at the same rate. This accommodates models with a large set of control variables whose coefficients may not be consistently estimable. This happens, for example, when the data have a group structure. There, (possibly multi-way) fixed effects are included almost by default to control for unobserved confounding factors at the group level. Of course, our results will also be useful in situations where  $q = 0$  and the regression design is highly unbalanced.

As our focus is on a subset of the slope coefficients it is useful to work with formulae from which the control variables have been partialled-out. For any  $n \times k$  matrix  $\mathbf{Q}$  of

rank  $k$  we will write  $\mathbf{H}_Q = \mathbf{Q}(\mathbf{Q}'\mathbf{Q})^{-1}\mathbf{Q}'$  and  $\mathbf{M}_Q = \mathbf{I}_n - \mathbf{H}_Q$ , where  $\mathbf{I}_n$  is the  $n \times n$  identity matrix. Then the columns of the  $n \times p$  matrix  $\hat{\mathbf{V}} = \mathbf{M}_B\mathbf{A}$  are the residuals from a regression of the columns of  $\mathbf{A}$  on  $\mathbf{B}$ . With  $\hat{v}_i$  the  $p$ -vector of residuals for observation  $i$  the least-squares estimator of  $\boldsymbol{\alpha}$  is

$$\hat{\boldsymbol{\alpha}} = \left( \sum_{i=1}^n \hat{v}_i \hat{v}_i' \right)^{-1} \left( \sum_{i=1}^n \hat{v}_i y_i \right),$$

and its (conditional) covariance matrix is

$$\boldsymbol{\Omega} = E_{\mathbf{X}}((\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha})(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha})') = \left( \sum_{i=1}^n \hat{v}_i \hat{v}_i' \right)^{-1} \left( \sum_{i=1}^n \hat{v}_i \hat{v}_i' \sigma_i^2 \right) \left( \sum_{i=1}^n \hat{v}_i \hat{v}_i' \right)^{-1}.$$

Under regularity conditions (stated below) we have

$$\boldsymbol{\Omega}^{-1/2}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}) \xrightarrow{d} N(\mathbf{0}, \mathbf{I}_p)$$

as  $n \rightarrow \infty$ . To make inference based on this result operational an estimator of  $\boldsymbol{\Omega}$  is needed.

The covariance-matrix estimator proposed by [Eicker \(1963, 1967\)](#) and [White \(1980\)](#) is

$$\hat{\boldsymbol{\Omega}} = \left( \sum_{i=1}^n \hat{v}_i \hat{v}_i' \right)^{-1} \left( \sum_{i=1}^n \hat{v}_i \hat{v}_i' \hat{\varepsilon}_i^2 \right) \left( \sum_{i=1}^n \hat{v}_i \hat{v}_i' \right)^{-1},$$

where  $\hat{\varepsilon}_i = y_i - \mathbf{x}_i' \hat{\boldsymbol{\beta}}$ , and is well known to be biased. The bias arises from the sampling noise in  $\hat{\boldsymbol{\beta}}$  and can be severe, especially when the regression design contains observations with high leverage, i.e., some of the diagonal entries of  $\mathbf{H}_X$  are large ([Chesher and Jewitt, 1987](#)). The bias may persist in large samples, rendering the covariance estimator inconsistent, if some observations remain influential, in the sense that their leverage does not approach zero as the sample size grows. The requirement that maximal leverage vanishes can be problematic in settings with many control variables. As is well known,  $(\mathbf{H}_X)_{i,i} \in [0, 1]$  and

$$\sum_{i=1}^n \frac{(\mathbf{H}_X)_{i,i}}{n} = \frac{r}{n} = \frac{p}{n} + \frac{q}{n}.$$

Hence, a necessary condition for  $\hat{\boldsymbol{\Omega}}$  to be consistent will be that  $q/n$  approaches zero as  $n$  grows. In models with group fixed effects, for example, this requires the size of the groups

to grow with the sample size. This is essentially a manifestation of the incidental-parameter problem of [Neyman and Scott \(1948\)](#). An asymptotic framework where  $q/n \rightarrow 0$  may not be suitable. In the standard one-way model for  $N \times T$  panel data, for example, we have  $N$  fixed effects to estimate from  $n = NT$  observations. Hence,  $q/n = 1/T$  and so we would require that  $T \rightarrow \infty$ .

### 3 Leave-one-out variance estimation

To motivate an alternative estimator of  $\boldsymbol{\Omega}$  let  $\hat{\boldsymbol{\beta}}_{-i}$  be the least-squares estimator obtained on dropping the  $i^{\text{th}}$  observation. Then a one-line calculation reveals that the cross-fit estimator

$$y_i \check{\varepsilon}_i, \quad \check{\varepsilon}_i = y_i - \mathbf{x}'_i \hat{\boldsymbol{\beta}}_{-i},$$

satisfies  $E_{\mathbf{X}}(y_i \check{\varepsilon}_i) = \sigma_i^2$ . This simple observation suggests the alternative covariance matrix estimator

$$\check{\boldsymbol{\Omega}} = \left( \sum_{i=1}^n \hat{\mathbf{v}}_i \hat{\mathbf{v}}'_i \right)^{-1} \left( \sum_{i=1}^n \hat{\mathbf{v}}_i \hat{\mathbf{v}}'_i (y_i \check{\varepsilon}_i) \right) \left( \sum_{i=1}^n \hat{\mathbf{v}}_i \hat{\mathbf{v}}'_i \right)^{-1},$$

which is (conditionally) unbiased. This paves the way for asymptotically-valid inference under asymptotics where  $q/n \rightarrow 0$  as  $n \rightarrow \infty$ .

It is well known ([Miller 1974](#)) that

$$\check{\varepsilon}_i = \frac{\hat{\varepsilon}_i}{(\mathbf{M}_{\mathbf{X}})_{i,i}}.$$

Hence, the leave-one-out estimator  $\hat{\boldsymbol{\beta}}_{-i}$  exists as long as  $(\mathbf{M}_{\mathbf{X}})_{i,i} > 0$ . Furthermore,  $\check{\boldsymbol{\Omega}}$ , as stated above, will be well-defined provided that

$$\min_i (\mathbf{M}_{\mathbf{X}})_{i,i} > 0,$$

that is, as long as there is no observation for which  $(\mathbf{H}_{\mathbf{X}})_{i,i} = 1$ . If such observations exist, dropping them would make  $\hat{\boldsymbol{\beta}}$  break down. The fact that existence of the least-squares estimator depends on a single observation is typically seen as poor regressor design and such observations are often treated as outliers. Furthermore, the Eicker-White covariance-matrix

estimator of  $\hat{\boldsymbol{\beta}}$  is singular if  $\max_i(\mathbf{H}_{\mathbf{X}})_{i,i} = 1$ , implying that there are linear combinations of  $\hat{\boldsymbol{\beta}}$  whose robust covariance estimator is exactly zero (Chesher and Jewitt, 1987). As such, our leverage condition seems close to necessary for inference robust to heteroskedasticity to be possible.

A leading case where an observation may have unit leverage is when  $(\mathbf{H}_{\mathbf{B}})_{i,i} = 1$ , in which case the control variables yield perfect prediction. Such observations carry no information on  $\boldsymbol{\alpha}$ , however, and so dropping them does not affect its least-squares estimator. Furthermore, if  $(\mathbf{H}_{\mathbf{B}})_{i,i} = 1$ , we necessarily have that  $\hat{\mathbf{v}}_i = \mathbf{0}$ , and so the observation does not contribute to our covariance-matrix estimator either. Hence, we are free to ignore such observations.

Inference based on  $\check{\boldsymbol{\Omega}}$  will be asymptotically valid under the following conditions.

**Assumption 1.**

(i) Let the set  $\{N_1, \dots, N_{G_n}\}$  be a partition of the set  $\{1, \dots, n\}$  into  $G_n$  groups such that  $\max_g |N_g| = O(1)$ . Then, conditional on  $\mathbf{B}$ , the collections  $\{\mathbf{a}_i : i \in N_g\}$  are independent across  $g$ , where we write  $\mathbf{A} = (\mathbf{a}_1, \dots, \mathbf{a}_n)'$ .

(ii) The errors  $\varepsilon_i$  are independent across  $i$  conditional on  $\mathbf{X}$ .

(iii) The matrix  $\mathbf{B}'\mathbf{B}$  has full rank with probability approaching one and  $\limsup_n q/n < 1$ .

(iv) Let  $\mathbf{V} = (\mathbf{v}_1, \dots, \mathbf{v}_n)' = \mathbf{A} - E_{\mathbf{B}}(\mathbf{A})$  and  $\dot{\mathbf{V}} = \mathbf{M}_{\mathbf{B}}\mathbf{V}$  and write  $\lambda$  for the smallest eigenvalue of the matrix  $E_{\mathbf{B}}(\dot{\mathbf{V}}'\dot{\mathbf{V}})$ . Then

$$\max_i \left( E_{\mathbf{X}}(\varepsilon_i^4) + E_{\mathbf{B}}(\|\mathbf{v}_i\|^4) + \frac{1}{\sigma_i^2} + \frac{n}{\lambda} \right) = O_p(1).$$

(v)  $E(\|\mathbf{a}_i\|^2) = O(1)$  (all  $i$ ),  $\max_i \|\hat{\mathbf{v}}_i\|/\sqrt{n} = o_p(1)$ ,  $\max_i \sigma_i^2 = O_p(1)$ ,  $\max_i (\mathbf{x}'_i \boldsymbol{\beta})^2 = O_p(1)$ .

Assumption 1 is adapted from Cattaneo, Jansson and Newey (2018) and we refer to them for detailed discussion. Here we only note that the setup allows data structures with stratified observations and for many nuisance parameters, in the sense that  $q/n$  is allowed to be close to unity, even in large samples.



Assumption 1 is both sufficient for  $\mathbf{\Omega}^{-1/2}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha})$  to be asymptotically (multivariate) standard normal and for  $\check{\mathbf{\Omega}}$  to be consistent for  $\mathbf{\Omega}$ . Slutsky’s theorem then yields our main result.

**Theorem 1.** *Let Assumption 1 hold. Then*

$$\check{\mathbf{\Omega}}^{-1/2}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}) \xrightarrow{d} N(\mathbf{0}, \mathbf{I}_p)$$

as  $n \rightarrow \infty$ .

The proof is in the Appendix.

## 4 Related work

There is a substantial literature on alternative estimators of  $\mathbf{\Omega}$ . The most popular such estimators are reviewed in [Long and Ervin \(2000\)](#) and [MacKinnon \(2012\)](#). The general idea underlying these suggestions is to modify  $\hat{\mathbf{\Omega}}$  by giving larger weight to residuals with higher leverage. As such, they are related to our proposal. However, unlike  $\check{\mathbf{\Omega}}$ , all these estimators are biased and inconsistent unless  $q/n \rightarrow 0$  as  $n \rightarrow \infty$ , as formally shown by [Cattaneo, Jansson and Newey \(2018\)](#). For example, the ‘almost-unbiased’ estimator of [Horn, Horn and Duncan \(1975\)](#) estimates  $\sigma_i^2$  by  $\hat{\varepsilon}_i \check{\varepsilon}_i$ , which is unbiased under homoskedasticity but not more generally. Furthermore, as  $y_i \check{\varepsilon}_i = \hat{\varepsilon}_i \check{\varepsilon}_i + \mathbf{x}'_i \hat{\boldsymbol{\beta}} \check{\varepsilon}_i$  our proposal can be understood to be a bias-corrected version of this estimator. The (uncentered) jackknife estimator of [MacKinnon and White \(1985\)](#) uses  $\check{\varepsilon}_i^2 = (y_i - \mathbf{x}'_i \hat{\boldsymbol{\beta}}_{-i})(y_i - \mathbf{x}'_i \hat{\boldsymbol{\beta}}_{-i})$  as an estimator of  $\sigma_i^2$ . This estimator suffers from bias because the error is estimated using a single leave-out estimator. Following the intuition behind our leave-one-out procedure, an unbiased variant would instead be  $(y_i - \mathbf{x}'_i \dot{\boldsymbol{\beta}}_{-i})(y_i - \mathbf{x}'_i \ddot{\boldsymbol{\beta}}_{-i})$ , where  $\dot{\boldsymbol{\beta}}_{-i}$  and  $\ddot{\boldsymbol{\beta}}_{-i}$  are least-squares estimators constructed from non-overlapping subsets of the data (from which the  $i^{\text{th}}$  observation has been removed); see also [Newey and Robins \(2018\)](#). In the current context, however, such an approach does not seem preferable over our proposal as it requires a more severe form of sample splitting.

The bootstrap, while often a powerful alternative to inference based on asymptotic approximations, too, fails when  $q/n \not\rightarrow 0$  as  $n \rightarrow \infty$  (Bickel and Freedman, 1983). Of course, its invalidity should not be too surprising, as  $\hat{\beta}$  is not asymptotically linear in this case.

An alternative approach is to bias-correct  $\hat{\Omega}$ . To describe it, write  $\hat{\varepsilon} = (\hat{\varepsilon}_1, \dots, \hat{\varepsilon}_n)'$  and  $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)'$ . Then

$$E_{\mathbf{X}}(\hat{\varepsilon} * \hat{\varepsilon}) = (\mathbf{M}_{\mathbf{X}} * \mathbf{M}_{\mathbf{X}}) E_{\mathbf{X}}(\varepsilon * \varepsilon),$$

where  $*$  denotes the elementwise (Shur or Hadamard) product. Consequently, as observed by Hartley, Rao and Kiefer (1969),

$$(\mathbf{M}_{\mathbf{X}} * \mathbf{M}_{\mathbf{X}})^{-1} (\hat{\varepsilon} * \hat{\varepsilon})$$

is an unbiased estimator of the vector of error variances. Moreover, the  $i^{\text{th}}$  such estimator is

$$\hat{\varepsilon}_i^2 = \sum_{j=1}^n ((\mathbf{M}_{\mathbf{X}} * \mathbf{M}_{\mathbf{X}})^{-1})_{i,j} \hat{\varepsilon}_j^2$$

and involves all least-squares residuals. It then immediately follows that

$$\hat{\Omega} = \left( \sum_{i=1}^n \hat{\mathbf{v}}_i \hat{\mathbf{v}}_i' \right)^{-1} \left( \sum_{i=1}^n \hat{\mathbf{v}}_i \hat{\mathbf{v}}_i' \hat{\varepsilon}_i^2 \right) \left( \sum_{i=1}^n \hat{\mathbf{v}}_i \hat{\mathbf{v}}_i' \right)^{-1}$$

is (conditionally) unbiased (see Bera, Suprayitno and Premaratne 2002). However,  $\hat{\Omega}$  appears to have gone mostly unnoticed (for example, it is not mentioned in the reviews of Long and Ervin 2000 and MacKinnon 2012). The chief reason appears to be that  $(\mathbf{M}_{\mathbf{X}} * \mathbf{M}_{\mathbf{X}})$  may not be invertible, in which case the estimator does not exist. Necessary and sufficient conditions are stated in Mallela (1972) but these are neither simple nor intuitive (Horn, Horn and Duncan, 1975). A sufficient condition (Horn and Horn, 1975) is

$$\min_i (\mathbf{M}_{\mathbf{X}})_{i,i} > \frac{1}{2},$$

which is substantially stronger than the requirement that  $\min_i (\mathbf{M}_{\mathbf{X}})_{i,i} > 0$  above.<sup>1</sup> In recent work, Cattaneo, Jansson and Newey (2018) considered a similar estimator, say  $\hat{\Omega}'$ ,

---

<sup>1</sup>There are cases where the condition that  $\min_i (\mathbf{M}_{\mathbf{X}})_{i,i} > \frac{1}{2}$  is also necessary. An example is the one-way regression model for panel data.

using the entries of  $(\mathbf{M}_B * \mathbf{M}_B)$  rather than those of  $(\mathbf{M}_X * \mathbf{M}_X)$  to estimate the error variances.<sup>2</sup> They provide conditions under which  $\hat{\boldsymbol{\Omega}}$  is consistent, allowing for  $q$  to grow with  $n$  so that  $\limsup_n q/n < \frac{1}{2}$ . This condition ensures that  $\min_i (\mathbf{M}_B)_{i,i} > \frac{1}{2}$  as  $n$  grows.  $\hat{\boldsymbol{\Omega}}$  is a substantial generalization of the bias-corrected covariance-matrix estimator of [Stock and Watson \(2008\)](#) for one-way fixed-effect regressions.

The small-leverage requirement can be problematic in settings with many fixed effects. For example it will typically not hold when fitting a two-way regression model to matched employer-employee or student-teacher data sets (as in [Abowd, Kramarz and Margolis 1999](#) and [Rockoff 2004](#)) as the informational content of such data is plagued by issues of limited mobility ([Jochmans and Weidner, 2018](#)). [Verdier \(2018\)](#) provides explicit conditions on the data structure for the leverage condition to hold in this context.

## 5 Simulations

Some numerical results are provided to compare the performance of our variance estimator to existing alternatives. All statistics reported below were calculated over 10,000 Monte Carlo replications.

The first set of simulations uses a design from [Cattaneo, Jansson and Newey \(2018\)](#). The design features the classical linear regression model with standard-normal errors, a univariate regressor of interest,  $a_i \sim N(0, 1)$ , and a  $q$ -vector of binary control variables,  $\mathbf{b}_i = (b_{i,1}, \dots, b_{i,q})'$ , where  $b_{i,j} = \{w_{i,j} > 2\}$  for  $w_{i,j} \sim \text{i.i.d. } N(0, 1)$ . We set  $\alpha = 1$ ,  $\boldsymbol{\eta} = \mathbf{0}$ , and generate  $n = 500$  observations with  $q \in \{10, 100, 250, 400, 450\}$ . This yields  $q/n \in \{.02, .20, .50, .90, .95\}$ . On average, each dummy variable will be one for about 10 observations. This is a rather sparse design that mimics a fixed-effect setting. [Table 1](#) contains the bias and root mean-squared error (RMSE) of various estimators of  $\boldsymbol{\Omega}$ , relative to the oracle variance estimator that assumes  $\varepsilon_i^2$  to be observed ( $\tilde{\boldsymbol{\Omega}}$ ). The table also provides

---

<sup>2</sup>This modification introduces bias but has negligible impact in large samples as long as  $\hat{\boldsymbol{\alpha}}$  is consistent. Indeed, using a formula for partitioned-matrix inversion allows to write  $\mathbf{M}_X = \mathbf{M}_B - \mathbf{H}_{M_B A}$ . The entries of the matrix  $\mathbf{H}_{M_B A}$  all vanish as  $n \rightarrow \infty$  if  $\hat{\boldsymbol{\alpha}}$  is consistent.

rejection frequencies of the two-sided  $t$ -test for the null that  $\alpha = 1$  (at the 5% level) using each of these variance estimators. The variance estimators considered are the Eicker-White estimator ( $\hat{\Omega}$ ), the ‘almost-unbiased’ estimator of [Horn, Horn and Duncan \(1975\)](#) ( $\hat{\Omega}_{\text{AU}}$ ), the (uncentered) jackknife estimator of [MacKinnon and White \(1985\)](#) ( $\hat{\Omega}_{\text{JK}}$ ), the estimators  $\check{\Omega}$  and  $\hat{\Omega}$  that bias-correct the squared residual vector, and the leave-one-out estimator  $\tilde{\Omega}$ . In replications where  $\check{\Omega}$  or  $\hat{\Omega}$  did not exist they were replaced by  $\tilde{\Omega}$ . While this is not an issue for small values of  $q/n$  it does happen very frequently for the larger values of  $q/n$  considered.<sup>3</sup>

Table 1: Design from [Cattaneo, Jansson and Newey \(2018\)](#)

$q$	$\tilde{\Omega}$	$\check{\Omega}$	$\hat{\Omega}_{\text{AU}}$	$\hat{\Omega}_{\text{JK}}$	$\hat{\Omega}$	$\hat{\Omega}$	$\tilde{\Omega}$
Relative bias							
10	0.000	-0.048	0.000	0.000	0.000	-0.048	0.000
100	0.000	-0.200	0.000	0.280	0.000	0.000	0.000
250	0.000	-0.475	0.000	1.050	-0.075	-0.100	0.000
400	0.000	-0.784	0.000	4.157	-0.216	-0.294	0.000
450	0.000	-0.889	-0.010	9.534	-0.293	-0.567	-0.005
Relative RMSE							
10	1.000	0.996	1.004	1.055	1.017	1.005	1.337
100	1.000	1.643	1.018	2.327	1.202	1.187	1.369
250	1.000	3.167	1.000	7.167	2.000	2.000	1.500
400	1.000	4.211	1.211	23.211	3.263	3.053	1.684
450	1.000	3.680	1.280	41.940	4.960	3.440	1.820
Empirical size (5% level)							
10	0.049	0.053	0.051	0.049	0.053	0.054	0.053
100	0.048	0.079	0.051	0.029	0.056	0.056	0.056
250	0.047	0.153	0.049	0.006	0.078	0.080	0.053
400	0.049	0.362	0.054	0.000	0.170	0.177	0.061
450	0.052	0.522	0.060	0.000	0.244	0.302	0.078

<sup>3</sup>In a previous version of this paper, in the case of non-existence of the inverses  $(\mathbf{M}_{\mathbf{X}} * \mathbf{M}_{\mathbf{X}})^{-1}$  and  $(\mathbf{M}_{\mathbf{B}} * \mathbf{M}_{\mathbf{B}})^{-1}$  they were replaced by Moore-Penrose pseudo inverses in the calculation of  $\check{\Omega}$  or  $\hat{\Omega}$ . This gave similar results to those reported here.

The table shows the poor performance of the standard variance estimator  $\tilde{\Omega}$  when  $q$  is not very small compared to  $n$ . It suffers from large (downward) bias which leads to severe overrejection under the null. While  $\hat{\Omega}_{\text{AU}}$  does well here (being conditionally unbiased under homoskedasticity), the jackknife estimator  $\hat{\Omega}_{\text{JK}}$  performs poorly with many regressors. It has both large (positive) bias and large variance and yields test statistics that are very conservative. The bias-corrected estimators,  $\hat{\Omega}$  and  $\hat{\Omega}'$ , are more variable than  $\hat{\Omega}_{\text{AU}}$ . They yield test statistics with comparable size when  $q/n \leq \frac{1}{2}$  but overreject quite severely when  $q/n$  takes on larger values. The leave-one-out estimator, while also more variable than  $\hat{\Omega}_{\text{AU}}$ , does well in terms of size for all ratios of  $q/n$ . These findings verify our theoretical results.

We next consider a one-way fixed-effect regression for  $N \times T$  panel data. Here,  $n = NT$ , there are  $q = N$  nuisance parameters (the unit-specific intercepts), and the panel is typically short (i.e.,  $T/N$  is close to zero). Here, the control variables are the fixed effects and the fact that they are not well estimable in short panels is well understood. We use designs from [Stock and Watson \(2008\)](#). The two designs have a scalar regressor of interest, which is again standard normal and has a slope parameter that equals one, and all the fixed effects set to zero. The designs differ only in the form of the conditional error variance. We have  $\varepsilon_i \sim N(0, \sigma_i^2)$  with

$$\text{Design (A): } \sigma_i^2 = \frac{\lambda}{(.10 + x_i^2)}; \quad \text{Design (B): } \sigma_i^2 = \lambda(.10 + x_i^2),$$

where  $\lambda$  is chosen so that the error has (unconditional) unit variance. [Tables 2 and 3](#) provide the results for Design (A) and (B), respectively. The same variance estimators as in the previous illustration were considered and the tables have the same lay-out as before. We consider sample sizes with  $N \in \{100, 250\}$  and  $T \in \{2, 3, 4\}$ . Note that  $\hat{\Omega}$  and  $\hat{\Omega}'$  do not exist for  $T = 2$ .

The simulation results again demonstrate the inadequacy of the usual covariance-matrix estimators. All of  $\hat{\Omega}$ ,  $\hat{\Omega}_{\text{AU}}$ , and  $\hat{\Omega}_{\text{JK}}$  suffer from bias and yield size-distorted test statistics. In accordance with the theory, their performance does not improve as  $T$  remains fixed while  $N$  grows large. The estimators  $\hat{\Omega}$  and  $\hat{\Omega}'$  do not exist when  $T = 2$  and are biased when

Table 2: Design (A) from [Stock and Watson \(2008\)](#)

$N$	$T$	$\tilde{\Omega}$	$\hat{\Omega}$	$\hat{\Omega}_{\text{AU}}$	$\hat{\Omega}_{\text{JK}}$	$\dot{\Omega}$	$\acute{\Omega}$	$\check{\Omega}$
Relative bias								
100	2	0.000	-0.500	0.000	1.031	—	—	-0.015
250	2	0.000	-0.520	0.000	1.000	—	—	0.000
100	3	0.000	-0.160	0.240	0.880	-0.280	-0.280	0.000
250	3	0.000	-0.200	0.200	0.900	-0.300	-0.300	0.000
100	4	0.000	0.000	0.357	0.786	-0.071	-0.071	0.000
250	4	0.000	-0.016	0.315	0.756	-0.097	-0.098	-0.001
Relative RMSE								
100	2	1.000	1.619	1.048	3.810	—	—	1.286
250	2	1.000	2.600	1.000	5.400	—	—	1.400
100	3	1.000	1.000	1.667	4.167	1.833	1.833	1.333
250	3	1.000	1.256	2.092	6.122	2.238	2.238	1.361
100	4	1.000	1.000	2.000	4.000	1.333	1.333	1.333
250	4	1.000	0.913	2.774	6.211	1.462	1.462	1.447
Empirical size (5% level)								
100	2	0.0507	0.171	0.0508	0.006	—	—	0.072
250	2	0.0548	0.173	0.0556	0.005	—	—	0.061
100	3	0.0486	0.076	0.0292	0.007	0.128	0.128	0.067
250	3	0.0526	0.076	0.0273	0.006	0.109	0.109	0.056
100	4	0.0495	0.053	0.0247	0.010	0.065	0.066	0.060
250	4	0.0503	0.051	0.0239	0.009	0.064	0.064	0.053

$T = 3$  (and when  $T = 4$ , but much less so). The bias is negative in Design (A) and positive in Design (B). It leads to large RMSE relative to  $\hat{\Omega}_{\text{AU}}$  and to, respectively, over- and underrejection of the null. Their empirical size is particularly poor for Design (A) in this case.  $\check{\Omega}$  performs well across the board, both in terms of precision as in terms of size, and dominates the bias-corrected estimators both in terms of bias and RMSE as in rejection frequencies.

Table 3: Design (B) from [Stock and Watson \(2008\)](#)

$N$	$T$	$\tilde{\Omega}$	$\hat{\Omega}$	$\hat{\Omega}_{\text{AU}}$	$\hat{\Omega}_{\text{JK}}$	$\dot{\Omega}$	$\acute{\Omega}$	$\check{\Omega}$
Relative bias								
100	2	0.000	-0.526	-0.037	0.963	—	—	-0.021
250	2	0.000	-0.513	-0.013	0.987	—	—	-0.013
100	3	0.000	-0.436	-0.145	0.300	0.127	0.100	-0.000
250	3	0.000	-0.432	-0.136	0.295	0.136	0.114	0.000
100	4	0.000	-0.359	-0.128	0.167	0.038	0.026	0.000
250	4	0.000	-0.355	-0.129	0.194	0.065	0.032	0.000
Relative RMSE								
100	2	1.000	1.559	1.088	3.500	—	—	1.294
250	2	1.000	2.353	1.176	5.000	—	—	1.353
100	3	1.000	1.625	1.000	1.688	1.438	1.344	1.156
250	3	1.000	2.500	1.250	2.125	1.625	1.500	1.250
100	4	1.000	1.550	1.000	1.350	1.150	1.100	1.100
250	4	1.000	2.400	1.200	1.600	1.200	1.200	1.200
Empirical size (5% level)								
100	2	0.049	0.198	0.067	0.009	—	—	0.068
250	2	0.051	0.177	0.060	0.009	—	—	0.059
100	3	0.049	0.152	0.078	0.032	0.048	0.047	0.056
250	3	0.051	0.146	0.074	0.027	0.042	0.043	0.055
100	4	0.047	0.121	0.071	0.037	0.051	0.052	0.054
250	4	0.050	0.113	0.070	0.034	0.047	0.047	0.052

## Appendix: Proof of Theorem 1

Write

$$\check{\Sigma} = n^{-1} \sum_{i=1}^n \hat{\mathbf{v}}_i \hat{\mathbf{v}}_i' (y_i \check{\varepsilon}_i), \quad \tilde{\Sigma} = n^{-1} \sum_{i=1}^n \hat{\mathbf{v}}_i \hat{\mathbf{v}}_i' \varepsilon_i^2,$$

and

$$\Sigma = n^{-1} \sum_{i=1}^n \hat{\mathbf{v}}_i \hat{\mathbf{v}}_i' \sigma_i^2.$$

We need to show that  $\check{\Sigma} - \Sigma \xrightarrow{p} \mathbf{0}$  and that  $\check{\Sigma} - \tilde{\Sigma} \xrightarrow{p} \mathbf{0}$  as  $n \rightarrow \infty$ . It will then follow that

$$\check{\Omega}^{-1/2}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}) \xrightarrow{d} N(\mathbf{0}, \mathbf{I}_p)$$

holds. As [Cattaneo, Jansson and Newey \(2018\)](#), to ease notation, we set  $p = 1$  without loss of generality.

To show the first result note that

$$\check{\Sigma} - \Sigma = \sum_{i=1}^n \hat{v}_i^2 (\varepsilon_i^2 - \sigma_i^2)$$

is a zero-mean random variable as  $E_{\mathbf{X}}(\varepsilon_i^2) = \sigma_i^2$  by definition. Its variance is

$$\text{var}_{\mathbf{X}} \left( n^{-1} \sum_{i=1}^n \hat{v}_i^2 (\varepsilon_i^2 - \sigma_i^2) \right) = \frac{\sum_{i=1}^n \sum_{j=1}^n \hat{v}_i^2 E_{\mathbf{X}}(\varepsilon_i \varepsilon_j) \hat{v}_j^2}{n^2} = \frac{\sum_{i=1}^n \hat{v}_i^4 \sigma_i^2}{n^2}.$$

We have

$$\frac{\sum_{i=1}^n \hat{v}_i^4 \sigma_i^2}{n^2} \leq \max_i \sigma_i^2 \left( \max_i \frac{\|\hat{\mathbf{v}}_i\|}{\sqrt{n}} \right)^2 \frac{\sum_{i=1}^n \hat{v}_i^2}{n} = o_p(1),$$

where we use that  $n^{-1} \sum_{i=1}^n \hat{v}_i^2 = O_p(1)$ , which follows from the arguments in the proof of Lemma SA-2 in [Cattaneo, Jansson and Newey \(2018\)](#). Hence,  $\check{\Sigma} - \Sigma \xrightarrow{p} \mathbf{0}$  has been shown.

To show the second result first note that

$$\check{\varepsilon}_i = \frac{\hat{\varepsilon}_i}{(\mathbf{M}_{\mathbf{X}})_{i,i}} = \frac{\sum_{j=1}^n (\mathbf{M}_{\mathbf{X}})_{i,j} \varepsilon_j}{(\mathbf{M}_{\mathbf{X}})_{i,i}} = \sum_{j=1}^n \frac{(\mathbf{M}_{\mathbf{X}})_{i,j}}{(\mathbf{M}_{\mathbf{X}})_{i,i}} \varepsilon_j = \sum_{j=1}^n m_{i,j} \varepsilon_j \quad (\text{say}),$$

where we use the well-known fact that  $\hat{\boldsymbol{\varepsilon}} = \mathbf{M}_{\mathbf{X}} \boldsymbol{\varepsilon}$ . Therefore, we can decompose the noise in the leave-one-out estimator as

$$y_i \check{\varepsilon}_i - \varepsilon_i^2 = (\mathbf{x}'_i \boldsymbol{\beta} + \varepsilon_i) \sum_{j=1}^n m_{i,j} \varepsilon_j - \varepsilon_i^2 = \sum_{j=1}^n \mathbf{x}'_i \boldsymbol{\beta} m_{i,j} \varepsilon_j + \sum_{j \neq i} \varepsilon_i m_{i,j} \varepsilon_j.$$



Hence,

$$\begin{aligned}\tilde{\Sigma} - \tilde{\Sigma} &= n^{-1} \sum_{i=1}^n \hat{v}_i^2 ((y_i \check{\varepsilon}_i - \varepsilon_i^2)) \\ &= n^{-1} \sum_{i=1}^n \sum_{j \neq i} \hat{v}_i^2 \varepsilon_i m_{i,j} \varepsilon_j + n^{-1} \sum_{i=1}^n \sum_{j=1}^n \hat{v}_i^2 \mathbf{x}'_i \boldsymbol{\beta} m_{i,j} \varepsilon_j\end{aligned}$$

Each of these right-hand side terms is a zero-mean random variable. We will calculate variance of each term, in turn.

For the first term, exploiting independence of the errors and symmetry of the summands,

$$\begin{aligned}\text{var}_{\mathbf{X}} \left( n^{-1} \sum_{i=1}^n \sum_{j \neq i} \hat{v}_i^2 \varepsilon_i m_{i,j} \varepsilon_j \right) &= \frac{\sum_{i=1}^n \sum_{k=1}^n \sum_{j \neq i} \sum_{\ell \neq k} \hat{v}_i^2 m_{i,j} \text{cov}_{\mathbf{X}}(\varepsilon_i \varepsilon_j, \varepsilon_k \varepsilon_\ell) m_{k,\ell} \hat{v}_k^2}{n^2} \\ &= 2 \frac{\sum_{i=1}^n \sum_{j \neq i} \hat{v}_i^2 m_{i,j} \text{cov}_{\mathbf{X}}(\varepsilon_i \varepsilon_j, \varepsilon_i \varepsilon_j) m_{i,j} \hat{v}_i^2}{n^2} \\ &\quad + 2 \frac{\sum_{i=1}^n \sum_{j \neq i} \hat{v}_i^2 m_{i,j} \text{cov}_{\mathbf{X}}(\varepsilon_i \varepsilon_j, \varepsilon_j \varepsilon_i) m_{j,i} \hat{v}_j^2}{n^2} \\ &= 2 \frac{\sum_{i=1}^n \sum_{j \neq i} \hat{v}_i^4 m_{i,j}^2 \sigma_i^2 \sigma_j^2}{n^2} \\ &\quad + 2 \frac{\sum_{i=1}^n \sum_{j \neq i} \hat{v}_i^2 \hat{v}_j^2 m_{i,j} m_{j,i} \sigma_i^2 \sigma_j^2}{n^2}.\end{aligned}$$

Now, using that

$$\sum_{j \neq i} m_{i,j}^2 = \sum_{j=1}^n m_{i,j}^2 - m_{i,i}^2 = \frac{\sum_{j=1}^n (\mathbf{M}_{\mathbf{X}})_{i,j}^2 - (\mathbf{M}_{\mathbf{X}})_{i,i}^2}{(\mathbf{M}_{\mathbf{X}})_{i,i}^2} = \frac{(\mathbf{M}_{\mathbf{X}})_{i,i} - (\mathbf{M}_{\mathbf{X}})_{i,i}^2}{(\mathbf{M}_{\mathbf{X}})_{i,i}^2} = \frac{(\mathbf{H}_{\mathbf{X}})_{i,i}}{(\mathbf{M}_{\mathbf{X}})_{i,i}},$$

we obtain

$$\begin{aligned}\frac{\sum_{i=1}^n \sum_{j \neq i} \hat{v}_i^4 m_{i,j}^2 \sigma_i^2 \sigma_j^2}{n^2} &\leq \max_i (\sigma_i^2)^2 \frac{\sum_{i=1}^n \hat{v}_i^4 \sum_{j \neq i} m_{i,j}^2}{n^2} \\ &= \max_i (\sigma_i^2)^2 \frac{\sum_{i=1}^n \hat{v}_i^4 (\mathbf{H}_{\mathbf{X}})_{i,i} / (\mathbf{M}_{\mathbf{X}})_{i,i}}{n^2} \\ &\leq \frac{\max_i (\sigma_i^2)^2}{\min_i (\mathbf{M}_{\mathbf{X}})_{i,i}} \frac{\sum_{i=1}^n \hat{v}_i^4}{n^2} \\ &\leq \frac{\max_i (\sigma_i^2)^2}{\min_i (\mathbf{M}_{\mathbf{X}})_{i,i}} \left( \max_i \frac{\|\hat{v}_i\|}{\sqrt{n}} \right)^2 \frac{\sum_{i=1}^n \hat{v}_i^2}{n} = o_p(1),\end{aligned}$$

and, similarly, using that

$$\sum_{j \neq i} m_{i,j} m_{j,i} = \sum_{j \neq i} \frac{(\mathbf{M}_{\mathbf{X}})_{i,j} (\mathbf{M}_{\mathbf{X}})_{j,i}}{(\mathbf{M}_{\mathbf{X}})_{i,i} (\mathbf{M}_{\mathbf{X}})_{j,j}} \leq \frac{(\mathbf{M}_{\mathbf{X}})_{i,i}}{\min_k (\mathbf{M}_{\mathbf{X}})_{k,k}} \sum_{j \neq i} m_{i,j}^2 = \frac{(\mathbf{H}_{\mathbf{X}})_{i,i}}{\min_k (\mathbf{M}_{\mathbf{X}})_{k,k}}$$

is bounded from above by  $1/\min_i(\mathbf{M}_\mathbf{X})_{i,i}$  we find

$$\begin{aligned}
\frac{\sum_{i=1}^n \sum_{j \neq i} \hat{v}_i^2 \hat{v}_j^2 m_{i,j} m_{j,i} \sigma_i^2 \sigma_j^2}{n^2} &\leq \max_i (\sigma_i^2)^2 \frac{\sum_{i=1}^n \sum_{j \neq i} \hat{v}_i^2 \hat{v}_j^2 m_{i,j} m_{j,i}}{n^2} \\
&\leq \max_i (\sigma_i^2)^2 \frac{\max_i \hat{v}_i^2}{n} \frac{\sum_{i=1}^n \hat{v}_i^2 \sum_{j \neq i} m_{i,j} m_{j,i}}{n} \\
&\leq \frac{\max_i (\sigma_i^2)^2}{\min_i (\mathbf{M}_\mathbf{X})_{i,i}} \frac{\max_i \hat{v}_i^2}{n} \frac{\sum_{i=1}^n \hat{v}_i^2}{n} \\
&\leq \frac{\max_i (\sigma_i^2)^2}{\min_i (\mathbf{M}_\mathbf{X})_{i,i}} \left( \max_i \frac{\|\hat{v}_i\|}{\sqrt{n}} \right)^2 \frac{\sum_{i=1}^n \hat{v}_i^2}{n} = o_p(1).
\end{aligned}$$

This yields the conclusion for the first term.

For the second term we may proceed similarly. We have

$$\begin{aligned}
\text{var}_\mathbf{X} \left( n^{-1} \sum_{i=1}^n \sum_{j=1}^n \hat{v}_i^2 \mathbf{x}'_i \boldsymbol{\beta} m_{i,j} \varepsilon_j \right) &= \frac{\sum_{j=1}^n \sigma_j^2 (\sum_{i=1}^n \hat{v}_i^2 (\mathbf{x}'_i \boldsymbol{\beta}) m_{i,j})^2}{n^2} \\
&= \frac{\sum_{j=1}^n \sigma_j^2 (\sum_{i=1}^n \sum_{k=1}^n \hat{v}_i^2 \hat{v}_k^2 (\mathbf{x}'_i \boldsymbol{\beta}) (\mathbf{x}'_k \boldsymbol{\beta}) m_{i,j} m_{k,j})^2}{n^2} \\
&\leq \max_i \sigma_i^2 \frac{\sum_{i=1}^n \sum_{k=1}^n \hat{v}_i^2 \hat{v}_k^2 (\mathbf{x}'_i \boldsymbol{\beta}) (\mathbf{x}'_k \boldsymbol{\beta}) \sum_{j=1}^n m_{i,j} m_{k,j}}{n^2} \\
&\leq \frac{\max_i \sigma_i^2}{\min_i (\mathbf{M}_\mathbf{X})_{i,i}^2} \max_i (\mathbf{x}'_i \boldsymbol{\beta})^2 \left( \max_i \frac{\|\hat{v}_i\|}{\sqrt{n}} \right)^2 \frac{\sum_{i=1}^n \hat{v}_i^2}{n} = o_p(1),
\end{aligned}$$

where we have used that

$$\sum_{j=1}^n m_{i,j} m_{k,j} = \frac{(\mathbf{M}_\mathbf{X})_{i,k}}{(\mathbf{M}_\mathbf{X})_{i,i} (\mathbf{M}_\mathbf{X})_{k,k}}, \quad \sum_{k=1}^n \sum_{j=1}^n m_{i,j} m_{k,j} \leq \frac{1}{\min_i (\mathbf{M}_\mathbf{X})_{i,i}^2}.$$

This yields the conclusion for the second term.

Taken together, these results imply that  $\tilde{\Sigma} - \tilde{\Sigma} \xrightarrow{p} \mathbf{0}$  which completes the proof of the theorem.

## References

- Abowd, J. M., F. Kramarz, and D. N. Margolis (1999). High wage workers and high wage firms. *Econometrica* 67, 251–333.
- 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.
- Bickel, P. J. and D. A. Freedman (1983). Bootstrapping regression models with many parameters. In P. Bickel, K. Doksum, and J. Hodges (Eds.), *A Festschrift for Erich L. Lehmann*, pp. 28–48. Chapman and Hall: Boca Raton, FL.
- Cattaneo, M. D., M. Jansson, and W. K. Newey (2018). Inference in linear regression models with many covariates and heteroskedasticity. Forthcoming in *Journal of the American Statistical Association*.
- Chesher, A. and I. Jewitt (1987). The bias of a heteroskedasticity consistent covariance matrix estimator. *Econometrica* 55, 1217–1222.
- Chetty, R., J. N. Friedman, and J. E. Rockoff (2014). Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates. *American Economic Review* 104, 2593–2632.
- 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. L. 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. (1981). *Robust Statistics*. New York: John Wiley.
- Jochmans, K. and M. Weidner (2018). Fixed-effect regressions on network data. Mimeo.
- 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. and J. M. Robins (2018). Cross-fitting and fast remainder rates for semiparametric estimation. Mimeo.
- Neyman, J. and E. L. Scott (1948). Consistent estimates based on partially consistent observations. *Econometrica* 16, 1–32.
- Quenouille, M. H. (1956). Notes on bias in estimation. *Biometrika* 43, 353–360.
- 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.
- Tukey, J. W. (1958). Bias and confidence in not-quite large samples (Abstract). *Annals of Mathematical Statistics* 29, 614.
- Verdier, V. (2018). Estimation and inference for linear models with two-way fixed effects and sparsely matched data. Mimeo.
- White, H. (1980). A heteroskedasticity-consistent covariance matrix and a direct test for heteroskedasticity. *Econometrica* 48, 817–838.