Treatment Effects with Many Covariates and Heteroskedasticity*

Matias D. Cattaneo† Michael Jansson‡ Whitney K. Newey§

July 9, 2015

Abstract

The linear regression model is widely used in empirical work in Economics. Researchers often include many covariates in their linear model specification in an attempt to control for confounders. We give inference methods that allow for many covariates and heteroskedasticity. Our results are obtained using high-dimensional approximations, where the number of covariates are allowed to grow as fast as the sample size. We find that all of the usual versions of Eicker-White heteroskedasticity consistent standard error estimators for linear models are inconsistent under this asymptotics. We then propose a new heteroskedasticity consistent standard error formula that is fully automatic and robust to both (conditional) heteroskedasticity of unknown form and the inclusion of possibly many covariates. We apply our findings to three settings: (i) parametric linear models with many covariates, (ii) semiparametric semi-linear models with many technical regressors, and (iii) linear panel models with many fixed effects.

Keywords: high-dimensional models, linear regression, many regressors, heteroskedasticity, standard errors.

JEL: C12, C14, C21.

*We specially thank Urlich Müller and Andres Santos for very thoughtful discussions regarding this project. We also thank Silvia Gonçalvez, Pat Kline, James MacKinnon and seminar participants at Duke University, Princeton University, UCSD, University of Miami, UNC-Chapel Hill, and University of Montreal for their comments. The first author gratefully acknowledges financial support from the National Science Foundation (SES 1459931). The second author gratefully acknowledges financial support from the National Science Foundation (SES 1459967) and the research support of CREATES (funded by the Danish National Research Foundation under grant no. DNRF78).

†Department of Economics, University of Michigan.

‡Department of Economics, UC Berkeley and *CREATES*.

§Department of Economics, MIT.

1 Introduction

A key goal in empirical work is to estimate the structural, causal, or treatment effect of some variable on an outcome of interest, such as the impact of a labor market policy on outcomes like earnings or employment. Since many variables measuring policies or interventions are not exogenous, researchers often employ observational methods to estimate their effects. One important method is based on assuming that the variable of interest can be taken as exogenous after controlling for a sufficiently large set of other factors or covariates. A major problem that empirical researchers face when employing selection-on-observables methods to estimate structural effects is the availability of many potential covariates. This problem has become even more pronounced in recent years because of the widespread availability of large (or high-dimensional) new data sets.

While it is often the case that economic theory (or intuition) will suggest a large set of variables that might be important, researchers prefer to also include additional "technical" controls constructed using indicator variables, interactions and other non-linear transformations of those variables. Therefore, many economic studies include very many covariates in order to control for as broad array of confounders as possible. For example, it is common practice in microeconometrics to include dummy variables for many potentially overlapping groups based on age, cohort, geographic location, etc. Even when some controls are dropped after valid covariate selection, as was recently developed by Belloni, Chernozhukov, and Hansen (2014b), many controls usually may remain in the final model specification.

We present valid inference methods that explicitly account for the presence of possibly many controls in linear regression models with unrestricted (conditional) heteroskedasticity. Specifically, we consider the setting where the object of interest is β in a model of the form

$$y_{i,n} = \beta' x_{i,n} + \gamma_n' w_{i,n} + u_{i,n}, \qquad i = 1, \ldots, n, \tag{1}$$

where $y_{i,n}$ is a scalar outcome variable, $x_{i,n}$ is a regressor of small (i.e., fixed) dimension d, $w_{i,n}$ is a vector of covariates of possibly large (i.e., growing) dimension K_n, and $u_{i,n}$ is an unobserved error term. Two important cases discussed in more detail below, are "flexible" parametric modeling of controls via basis expansions such as higher-order powers and interactions (i.e., a series-based formulation of the partially linear regression model), and models with many dummy variables such as fixed effects and interactions thereof in panel data. In both cases conducting OLS-based inference on β in (1) is straightforward when the error $u_{i,n}$ is homoskedastic and/or the dimension K_n of the nuisance covariates is modeled as a vanishing fraction of the sample size. The latter modeling assumption, however, seems inappropriate in applications with many dummy variables model and does not deliver the best approximation when many covariates are included.

Motivated by the above observations, this paper studies the consequences of allowing the error $u_{i,n}$ in (1) to be (conditionally) heteroskedastic in a setting where the covariate $w_{i,n}$ is permitted to be high-dimensional in the sense that K_n is allowed, but not required, to be a non-vanishing fraction of the sample size. Our main purpose is to investigate the possibility of constructing heteroskedasticity-consistent variance estimators for the OLS estimator of β in (1) without (necessarily) assuming any special structure on the part of the covariate $w_{i,n}$. We present two main results. First, we provide high-level sufficient conditions guaranteeing a valid Gaussian distributional approximation to the finite sample distribution of the OLS estimator of β, allowing for the dimension of the nuisance covariates to be proportional to the sample size ($K_n \propto n$). Second, we characterize the large sample properties of a class of variance estimators and use this characterization to obtain both negative and positive results. The negative finding is that the Eicker-White estimator is inconsistent in general, as are popular variants of this estimator. The positive result gives conditions under which an alternative heteroskedasticity-robust variance estimator (described in more detail below) is consistent. The main condition needed for our constructive results is a high-level assumption on the nuisance covariates requiring in particular that their number be strictly less than half of the sample size.

Our results contribute to the already sizeable literature on heteroskedasticity-robust variance estimators for linear regression models, a recent review of which is given by MacKinnon (2012). Important papers whose results are related to ours include White (1980), MacKinnon and White (1985), Wu (1986), Chesher and Jewitt (1987), Shao and Wu (1987), Chesher (1989), Cribari-Neto, Ferrari, and Cordeiro (2000), Bera, Suprayitno, and Premaratne (2002), Stock and Watson (2008), Cribari-Neto and da Gloria A. Lima (2011), and Müller (2013). In particular, Bera, Suprayitno, and Premaratne (2002) analyze some finite sample properties of a variance estimator similar to the one whose asymptotic properties are studied herein.

This paper also adds to the literature on high-dimensional linear regression where the number of regressors grow with the sample size; see, e.g., Huber (1973), Koenker (1988), Mammen (1993), El Karoui, Bean, Bickel, Lim, and Yu (2013) and references therein. In particular, Huber (1973) showed that fitted regression values are not asymptotically normal when the number of regressors grows as fast as sample size, while Mammen (1993) obtained asymptotic normality for arbitrary contrasts of OLS estimators in linear regression models where the dimension of the covariates is at most a vanishing fraction of the sample size. More recently, El Karoui, Bean, Bickel, Lim, and Yu (2013) showed that, if a Gaussian distributional assumption on regressors and homoskedasticity is assumed, then certain estimated coefficients and contrasts in linear models are asymptotically normal when the number of regressors grow as fast as sample size, but do not discuss inference results (even under homoskedasticity). Our result in Theorem 1 below shows that certain contrasts of OLS estimators in high-dimensional linear models are asymptotically normal under fairly general regularity conditions. Intuitively, we circumvent the problems associated with the lack of asymptotic Gaussianity by focusing exclusively on a small subset of regressors when the number of covariates gets large. We give inference results by constructing heteroskedasticity consistent standard errors without imposing any distributional assumption or other very specific restrictions on the regressors.

As discussed in more detailed below, our high-level conditions allow for $K_n \propto n$ and restrict the data generating process in fairly general and intuitive ways. In particular, our generic sufficient condition on the nuisance covariates $w_{i,n}$ covers several special cases of interest for empirical work. For example, our results encompass (and weakens in some sense; see Remark 2 below) those reported in Stock and Watson (2008), who investigated the one-way fixed effects panel data regression model in detail and showed that the conventional Eicker-White heteroskedasticity-robust variance estimator is inconsistent in that model, being plagued by a non-negligible bias problem attributable to the presence of many covariates (i.e., the fixed effects). The very special structure of the covariates in the one-way fixed effects model estimator enabled Stock and Watson (2008) to give an explicit characterization of this bias and to demonstrate consistency of a bias-corrected version of the Eicker-White variance estimator. The generic variance estimator proposed herein essentially reduces to their bias corrected variance estimator in the special case of the one-way fixed effects model, even though our results are derived from a different perspective.

The rest of this paper is organized as follows. Section 2 presents the variance estimators we study and gives a heuristic description of their main properties. Section 3 introduces the three leading examples covered by our results. Section 4 introduces a general framework that unifies the examples, gives the main results of the paper, and discusses their implications for the three examples we consider. Section 5 reports the results of a Monte Carlo experiment, while Section 6 concludes. To conserve space, proofs of all results are reported in a supplemental appendix.

2 Variance Estimators

For the purposes of discussing variance estimators associated with the OLS estimator $\hat{\beta}_n$ of β in (1) it is convenient to write the estimator in "partialled out" form as

$$\hat{\beta}_n = (\sum_{i=1}^{n} \hat{v}_{i,n}\hat{v}_{i,n}')^{-1}(\sum_{i=1}^{n} \hat{v}_{i,n}y_{i,n}), \qquad \hat{v}_{i,n} = \sum_{j=1}^{n} M_{ij,n}x_{j,n},$$

where $M_{ij,n} = \mathbb{1}(i=j) - w_{i,n}'(\sum_{k=1}^{n} w_{k,n}w_{k,n}')^{-1}w_{j,n}$, $\mathbb{1}(\cdot)$ denotes the indicator function, and the relevant inverses are assumed to exist. Defining $\hat{\Gamma}_n = \sum_{i=1}^{n} \hat{v}_{i,n}\hat{v}_{i,n}'/n$, the objective is to find an estimator $\hat{\Sigma}_n$ of the variance of $\sum_{i=1}^{n} \hat{v}_{i,n}u_{i,n}/\sqrt{n}$ such that

$$\hat{\Omega}_n^{-1/2}\sqrt{n}(\hat{\beta}_n - \beta) \to_d \mathcal{N}(0, I_d), \qquad \hat{\Omega}_n = \hat{\Gamma}_n^{-1}\hat{\Sigma}_n\hat{\Gamma}_n^{-1}, \tag{2}$$

in which case asymptotically valid inference on β can be conducted in the usual way by employing the distributional approximation $\hat{\beta}_n \overset{a}{\sim} \mathcal{N}(\beta, \hat{\Omega}_n/n)$.

Defining $\hat{u}_{i,n} = \sum_{j=1}^{n} M_{ij,n}(y_{j,n} - \hat{\beta}_n' x_{j,n})$, standard choices of $\hat{\Sigma}_n$ in the fixed-K_n case include the homoskedasticity-only estimator

$$\hat{\Sigma}_n^{\mathtt{HO}} = \hat{\sigma}_n^2\hat{\Gamma}_n, \qquad \hat{\sigma}_n^2 = \frac{1}{n-d-K_n}\sum_{i=1}^{n} \hat{u}_{i,n}^2,$$

and the Eicker-White-type estimator

$$\hat{\Sigma}_n^{\mathtt{EW}} = \frac{1}{n}\sum_{i=1}^{n} \hat{v}_{i,n}\hat{v}_{i,n}'\hat{u}_{i,n}^2.$$

Perhaps not too surprisingly, we find that consistency of $\hat{\Sigma}_n^{\mathtt{HO}}$ under homoskedasticity holds quite generally even for models with many covariates. In contrast, construction of a heteroskedasticity-robust estimator of Σ_n is more challenging, as it turns out that consistency of $\hat{\Sigma}_n^{\mathtt{EW}}$ generally requires K_n to be a vanishing fraction of n.

To fix ideas, suppose $(y_{i,n}, x_{i,n}', w_{i,n}')$ are i.i.d. over i. It turns out that, under certain regularity conditions,

$$\hat{\Sigma}_n^{\mathtt{EW}} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} M_{ij,n}^2 \hat{v}_{i,n}\hat{v}_{i,n}'\mathbb{E}[u_{j,n}^2|x_{j,n}, w_{j,n}] + o_p(1),$$

whereas a requirement for (2) to hold is that the estimator $\hat{\Sigma}_n$ satisfies

$$\hat{\Sigma}_n = \frac{1}{n}\sum_{i=1}^{n} \hat{v}_{i,n}\hat{v}_{i,n}'\mathbb{E}[u_{i,n}^2|x_{i,n}, w_{i,n}] + o_p(1). \tag{3}$$

The difference between the leading terms in the expansions is non-negligible in general unless $K_n/n \to 0$. In recognition of this problem with $\hat{\Sigma}_n^{\mathtt{EW}}$, we study the more general class of estimators of the form

$$\hat{\Sigma}_n(\kappa_n) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} \kappa_{ij,n}\hat{v}_i\hat{v}_i'\hat{u}_j^2,$$

where $\kappa_{ij,n}$ denotes element (i,j) of a symmetric matrix $\kappa_n = \kappa_n(w_{1,n}, \ldots, w_{n,n})$. Estimators that can be written in this fashion include $\hat{\Sigma}_n^{\mathtt{EW}}$ (which corresponds to $\kappa_n = I_n$) as well as variants of the so-called HCk estimators, $k \in \{1, 2, 3, 4\}$, discussed by MacKinnon (2012), among others.[1]

All of the HCk-type estimators (correspond to a diagonal choice of κ_n and) share with $\hat{\Sigma}_n^{\mathtt{EW}}$ the shortcoming that they do not satisfy (3) when $K_n/n \nrightarrow 0$. On the other hand, it turns out that a certain non-diagonal choice of κ_n makes it possible to satisfy (3) even if K_n is a non-vanishing fraction of n. To be specific, it turns out that (under regularity conditions and) under mild conditions under the weights $\kappa_{ij,n}$, $\hat{\Sigma}_n(\kappa_n)$ satisfies

$$\hat{\Sigma}_n(\kappa_n) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n} \kappa_{ik,n}M_{kj,n}^2\hat{v}_{i,n}\hat{v}_{i,n}'\mathbb{E}[u_{j,n}^2|x_{j,n}, w_{j,n}] + o_p(1),$$

[1] To be specific, a natural variant of HCk is obtained by choosing κ_n to be diagonal with $\kappa_{ii,n} = \Upsilon_{i,n}M_{ii,n}^{-\xi_{i,n}}$, where $(\Upsilon_{i,n}, \xi_{i,n}) = (n/(n-K_n), 0)$ for HC1, $(\Upsilon_{i,n}, \xi_{i,n}) = (1, 1)$ for HC2, $(\Upsilon_{i,n}, \xi_{i,n}) = (1, 2)$ for HC3, and $(\Upsilon_{i,n}, \xi_{i,n}) = (1, \min(4, nM_{ii,n}/K_n))$ for HC4.

suggesting that (3) holds with $\hat{\Sigma}_n = \hat{\Sigma}_n(\kappa_n)$ provided κ_n is chosen in such a way that

$$\sum_{k=1}^{n} \kappa_{ik,n} M_{kj,n}^2 = \mathbb{1}(i=j), \qquad 1 \leq i, j \leq n.$$

Accordingly, we define

$$\hat{\Sigma}_n^{\mathtt{HC}} = \hat{\Sigma}_n(\kappa_n^{\mathtt{HC}}) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} \kappa_{ij,n}^{\mathtt{HC}} \hat{v}_{i,n} \hat{v}_{i,n}' \hat{u}_{j,n}^2,$$

where, with M_n denoting the matrix with element (i,j) given by $M_{ij,n}$ and $\odot$ denoting the Hadamard product,

$$\kappa_n^{\mathtt{HC}} = \begin{pmatrix} \kappa_{11,n}^{\mathtt{HC}} & \cdots & \kappa_{1n,n}^{\mathtt{HC}} \\ \vdots & \ddots & \vdots \\ \kappa_{n1,n}^{\mathtt{HC}} & \cdots & \kappa_{nn,n}^{\mathtt{HC}} \end{pmatrix} = \begin{pmatrix} M_{11,n}^2 & \cdots & M_{1n,n}^2 \\ \vdots & \ddots & \vdots \\ M_{n1,n}^2 & \cdots & M_{nn,n}^2 \end{pmatrix}^{-1} = (M_n \odot M_n)^{-1}.$$

The estimator $\hat{\Sigma}_n^{\mathtt{HC}}$ is well defined whenever $M_n \odot M_n$ is invertible, a simple sufficient condition for which is that $\mathcal{M}_n < 1/2$, where[2]

$$\mathcal{M}_n = 1 - \min_{1 \leq i \leq n} M_{ii,n}.$$

More importantly, a slight strengthening of the condition $\mathcal{M}_n < 1/2$ will be shown to be sufficient for (2) and (3) to hold with $\hat{\Sigma}_n = \hat{\Sigma}_n^{\mathtt{HC}}$.

Remark 1. The estimator $\hat{\Sigma}_n^{\mathtt{HC}}$ can be written as $n^{-1} \sum_{i=1}^{n} \hat{v}_{i,n} \hat{v}_{i,n}' \tilde{u}_{i,n}^2$, where $\tilde{u}_{i,n}^2 = \sum_{j=1}^{n} \kappa_{ij,n}^{\mathtt{HC}} \hat{u}_{j,n}^2$ can be interpreted as a bias-corrected "estimator" of (the conditional expectation of) $u_{i,n}^2$.

3 Examples

The heuristics of the preceding section will be made precise in the next section. Before doing so, we present three leading examples, all of which are covered by the results developed in Section 4: (i) linear regression models with increasing dimension, (ii) semiparametric partially linear models, and (iii) fixed effects panel data regression models.

Let $\lambda_{\min}(\cdot)$ denote the minimum eigenvalue of its argument and let $\|\cdot\|$ denote the Euclidean norm.

3.1 Linear Regression Model with Increasing Dimension

The model of main interest is the linear regression model characterized by (1) and the following assumptions.

Assumption LR1 $\{(y_{i,n}, x_{i,n}', w_{i,n}') : 1 \leq i \leq n\}$ are i.i.d. over i.

[2] The fact that $\mathcal{M}_n < 1/2$ implies invertibility of $M_n \odot M_n$ is a consequence of the Gershgorin circle theorem. For details, see the supplemental appendix.

Assumption LR2 $\mathbb{E}[\|x_{i,n}\|^2] = O(1)$, $\mathbb{E}[u_{i,n}|x_{i,n}, w_{i,n}] = 0$, and $\max_{1\leq i\leq n} |\hat{v}_{i,n}|/\sqrt{n} = o_p(1)$.

Assumption LR3 $\mathbb{P}[\lambda_{\min}(\sum_{i=1}^n w_{i,n}w_{i,n}') > 0] \to 1$, $\overline{\lim}_{n\to\infty} K_n/n < 1$, and $\mathcal{C}_n^{\texttt{LR}} = O_p(1)$, where

$$\begin{aligned}\mathcal{C}_n^{\texttt{LR}} &= \max_{1\leq i\leq n}\{\mathbb{E}[u_{i,n}^4|x_{i,n}, w_{i,n}] + \mathbb{E}[\|V_{i,n}\|^4|w_{i,n}]\} \\ &\quad + \max_{1\leq i\leq n}\{1/\mathbb{E}[u_{i,n}^2|x_{i,n}, w_{i,n}] + 1/\lambda_{\min}(\mathbb{E}[V_{i,n}V_{i,n}'|w_{i,n}])\},\end{aligned}$$

with $V_{i,n} = x_{i,n} - \mathbb{E}[x_{i,n}|w_{i,n}]$.

We shall consider this model in some detail because it is important in its own right and because the insights obtained for it can be used constructively in other cases, including the partially linear model (4) and the fixed effects panel data regression model (5) presented below. Linear regression models with (possibly) increasing dimension have a long tradition in econometrics and statistics, and we consider them here as a theoretical device to obtain asymptotic approximations that better represent the finite-sample behavior of the statistics of interest.

The main difference between Assumptions LR1-LR3 and those familiar from the fixed-K_n case is the presence of the condition $\max_{1\leq i\leq n} |\hat{v}_{i,n}|/\sqrt{n} = o_p(1)$ in Assumption LR2. At the present level of generality it seems difficult to formulate primitive sufficient conditions for this condition that cover all cases of interest, but for completeness we mention that under mild moment conditions it suffices to require that one of the following conditions hold (see the supplemental appendix for details):

(i) $\mathcal{M}_n \to_p 0$, or

(ii) $\chi_n^{\texttt{LR}} = \min_{\delta\in\mathbb{R}^{K_n\times d}} \mathbb{E}[\|\mathbb{E}(x_{i,n}|w_{i,n}) - \delta' w_{i,n}\|^2] \to 0$, or

(iii) $\max_{1\leq i\leq n}\sum_{j=1}^n \mathbb{1}(M_{ij,n} \neq 0) = o_p(n^{1/3})$.

Each of these conditions is interpretable. First, $\mathcal{M}_n \geq K_n/n$ because $\sum_{i=1}^n M_{ii,n} = n - K_n$ and a necessary condition for **(i)** is therefore that $K_n/n \to 0$. Conversely, because

$$\mathcal{M}_n \leq \frac{K_n}{n}\frac{1-\min_{1\leq i\leq n} M_{ii,n}}{1-\max_{1\leq i\leq n} M_{ii,n}},$$

the condition $K_n/n \to 0$ is sufficient for **(i)** whenever the design is "approximately balanced" in the sense that $(1-\min_{1\leq i\leq n} M_{ii,n})/(1-\max_{1\leq i\leq n} M_{ii,n}) = O_p(1)$. In other words, **(i)** requires and effectively covers the case where it is assumed that K_n is a vanishing fraction of n. In contrast, conditions **(ii)** and **(iii)** can hold also when K_n is a non-vanishing fraction of n, which is the case of primary interest in this paper.

Because **(ii)** is a requirement on the accuracy of the approximation

$$\mathbb{E}[x_{i,n}|w_{i,n}] \approx \delta_n' w_{i,n}, \qquad \delta_n = \mathbb{E}[w_{i,n}w_{i,n}']^{-1}\mathbb{E}[w_{i,n}x_{i,n}'],$$

primitive conditions for it are available when the elements of $w_{i,n}$ are approximating functions, as in the partially linear model (4) discussed next. Indeed, in such cases one typically has $\chi_n^{\texttt{LR}} = O(K_n^{-\alpha})$

for some $\alpha > 0$, so condition **(ii)** not only accommodates $K_n/n \nrightarrow 0$, but actually places no upper bound on the magnitude of K_n in important special cases.

Finally, condition **(iii)**, and its underlying higher-level condition described in the supplemental appendix, is useful to handle cases where $w_{i,n}$ can not be interpreted as approximating functions, but rather just many different covariates included in the linear model specification. This condition is a "sparsity" condition on the matrix M_n, which allows for $K_n/n \nrightarrow 0$. Although somewhat stronger than needed, the condition is easy to verify in certain cases, including the panel data model (5) discussed below.

3.2 Semiparametric Partially Linear Model

Another econometric model covered by our results is the partially linear model

$$y_i = \beta' x_i + g(z_i) + \varepsilon_i, \qquad i = 1, \ldots, n, \tag{4}$$

where x_i and z_i are explanatory variables, ε_i is an error term, and the function $g(z)$ is unknown. Suppose $\{p^k(z) : k = 1, 2, \cdots, K_n\}$ are functions having the property that linear combinations can approximate square-integrable functions of z well, in which case $g(z_i) \approx \gamma_n' p_n(z_i)$ for some γ_n, where $p_n(z) = (p^1(z), \ldots, p^{K_n}(z))'$. Defining $y_{i,n} = y_i$, $x_{i,n} = x_i$, $w_{i,n} = p_n(z_i)$, and $u_{i,n} = \varepsilon_i + g(z_i) - \gamma_n' w_{i,n}$, the model (4) is of the form (1), and $\hat{\beta}_n$ is the series estimator of β previously studied by Donald and Newey (1994) and Cattaneo, Jansson, and Newey (2015). In this case, our analysis of $\hat{\beta}_n$ will proceed under the following assumptions.

Assumption PL1 $\{(y_i, x_i', z_i') : 1 \leq i \leq n\}$ are i.i.d. over i.

Assumption PL2 $\mathbb{E}[\varepsilon_i | x_i, z_i] = 0$, $\varrho_n^{\mathtt{PL}} \to 0$, $\chi_n^{\mathtt{PL}} \to 0$, and $n \varrho_n^{\mathtt{PL}} \chi_n^{\mathtt{PL}} \to 0$, where

$$\varrho_n^{\mathtt{PL}} = \min_{\gamma \in \mathbb{R}^{K_n}} \mathbb{E}[|\mathbb{E}[y_i - \beta' x_i | x_i, z_i] - \gamma' p_n(z_i)|^2], \qquad \chi_n^{\mathtt{PL}} = \min_{\delta \in \mathbb{R}^{K_n \times d}} \mathbb{E}[\|\mathbb{E}[x_i | z_i] - \delta' p_n(z_i)\|^2].$$

Assumption PL3 $\mathbb{P}[\lambda_{\min}(\sum_{i=1}^n p_n(z_i) p_n(z_i)') > 0] \to 1$, $\overline{\lim}_{n \to \infty} K_n/n < 1$, and $\mathcal{C}_n^{\mathtt{PL}} = O_p(1)$, where

$$\mathcal{C}_n^{\mathtt{PL}} = \max_{1 \leq i \leq n} \{\mathbb{E}[\varepsilon_i^4 | x_i, z_i] + \mathbb{E}[\|\nu_i\|^4 | z_i] + 1/\mathbb{E}[\varepsilon_i^2 | x_i, z_i] + 1/\lambda_{\min}(\mathbb{E}[\nu_i \nu_i' | z_i])\},$$

with $\nu_i = x_i - \mathbb{E}[x_i | z_i]$.

Because $g(z_i) \neq \gamma_n' p_n(z_i)$ in general, the partially linear model does not (necessarily) satisfy $\mathbb{E}[u_{i,n} | x_{i,n}, w_{i,n}] = 0$. To accommodate this failure a relaxation of Assumption LR2 is needed. The approach taken here, made precise in Assumption PL2, is motivated by the fact that linear combinations of $\{p^k(z)\}$ are assumed to be able to approximate the functions $g(z)$ and $h(z)$ well, where $h(z_i) = \mathbb{E}[x_i | z_i]$. Under standard smoothness conditions, and for standard choices of basis functions, we have $\varrho_n^{\mathtt{PL}} = O(K_n^{-\alpha_g})$ and $\chi_n^{\mathtt{PL}} = O(K_n^{-\alpha_h})$ for some pair (α_g, α_h) of positive constants, in which case Assumption PL2 holds provided $K_n^{\alpha_g+\alpha_h}/n \to \infty$. For further technical details see, for example, Newey (1997), Chen (2007), Cattaneo and Farrell (2013), and Belloni, Chernozhukov, Chetverikov, and Kato (2015).

3.3 Fixed Effects Panel Data Regression Model

Stock and Watson (2008) consider heteroskedasticity-robust inference for the panel data regression model

$$Y_{it} = \alpha_i + \beta' X_{it} + U_{it}, \qquad i = 1, \ldots, N, \qquad t = 1, \ldots, T, \tag{5}$$

where $\alpha_i \in \mathbb{R}$ is an individual-specific intercept, $X_{it} \in \mathbb{R}^d$ is a regressor of dimension d, $U_{it} \in \mathbb{R}$ is an error term, and the following assumptions are satisfied.

Assumption FE1 $\{(U_{i1}, \ldots, U_{iT}, X'_{i1} \ldots, X'_{iT}) : 1 \leq i \leq n\}$ are independent over i, $T \geq 3$ is fixed, and $\mathbb{E}[U_{it}U_{is}|X_{i1} \ldots, X_{iT}] = 0$ for $t \neq s$.

Assumption FE2 $\mathbb{E}[U_{it}|X_{i1} \ldots, X_{iT}] = 0$.

Assumption FE3 $\mathcal{C}_N^{\mathtt{FE}} = O_p(1)$, where

$$\begin{aligned} \mathcal{C}_N^{\mathtt{FE}} &= \max_{1 \leq i \leq N, 1 \leq t \leq T} \{\mathbb{E}[U_{it}^4|X_{i1} \ldots, X_{iT}] + \mathbb{E}[\|X_{it}\|^4]\} \\ &\quad + \max_{1 \leq i \leq N, 1 \leq t \leq T} \{1/\mathbb{E}[U_{it}^2|X_{i1} \ldots, X_{iT}] + 1/\lambda_{\min}(\mathbb{E}[\tilde{V}_{it}\tilde{V}'_{it}])\}, \end{aligned}$$

with $\tilde{V}_{it} = X_{it} - \mathbb{E}[X_{it}] - T^{-1}\sum_{s=1}^{T}(X_{is} - \mathbb{E}[X_{is}])$.

Defining $n = NT$, $K_n = N$, $\gamma_n = (\alpha_1, \ldots, \alpha_N)'$, and

$$(y_{(i-1)T+t,n}, x'_{(i-1)T+t,n}, u_{(i-1)T+t,n}, w'_{(i-1)T+t,n}) = (Y_{it}, X'_{it}, U_{it}, e'_{i,N}), \qquad 1 \leq i \leq N,\ 1 \leq t \leq T,$$

where $e_{i,N} \in \mathbb{R}^N$ is the i-th unit vector of dimension N, the model (5) is also of the form (1) and $\hat{\beta}_n$ is the fixed effects estimator of β. In general, this model does not satisfy Assumption LR1, but Assumption FE1 enables us to employ results for independent random variables when developing asymptotics. In other respects this model is in fact more tractable than the previous models due to the special nature of the covariates $w_{i,n}$.

Remark 2. One implication of Assumptions FE1 and FE2 is that $\mathbb{E}[Y_{it}|X_{i1}, \ldots, X_{iT}] = \alpha_i + \beta' X_{it}$, where α_i can depend on i and the conditioning variables $(X_{i1}, \ldots, X_{iT})$ in an arbitrary way. In the spirit of "fixed effects" (as opposed to "correlated random effects") Assumptions FE1-FE3 further allow $\mathbb{V}[Y_{it}|X_{i1}, \ldots, X_{iT}]$ to depend not only on $(X_{i1}, \ldots, X_{iT})$, but also on i. In particular, unlike Stock and Watson (2008), we do not require $(U_{i1}, \ldots, U_{iT}, X'_{i1} \ldots, X'_{iT})$ to be i.i.d. over i. In addition, we do not require any kind of stationarity on the part of (U_{it}, X'_{it}). The amount of variance heterogeneity permitted is quite large, as Assumption FE3 basically only requires $\mathbb{V}[Y_{it}|X_{i1}, \ldots, X_{iT}] = \mathbb{E}[U_{it}^2|X_{i1}, \ldots, X_{iT}]$ to be bounded and bounded away from zero. (On the other hand, serial correlation is assumed away because Assumptions FE1 and FE2 imply that $\mathbb{C}[Y_{it}, Y_{is}|X_{i1}, \ldots, X_{iT}] = 0$ for $t \neq s$.)

4 Results

The three models presented in the previous section are non-nested, but may be treated in a unified way by embedding them in a general framework. This general framework, which accommodates our motivating examples as well as others, is presented next.

4.1 General Framework

Suppose $\{(y_{i,n}, x'_{i,n}, w'_{i,n}) : 1 \leq i \leq n\}$ is generated by (1). Let $\mathcal{X}_n = (x_{1,n}, \ldots, x_{n,n})$ and for a set $\mathcal{W}_n$ of random variables satisfying $\mathbb{E}[w_{i,n}|\mathcal{W}_n] = w_{i,n}$, define the constants

$$\begin{aligned} \varrho_n &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[R_{i,n}^2], \qquad R_{i,n} = \mathbb{E}[u_{i,n}|\mathcal{X}_n, \mathcal{W}_n], \\ \rho_n &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[r_{i,n}^2], \qquad r_{i,n} = \mathbb{E}[u_{i,n}|\mathcal{W}_n], \\ \chi_n &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[\|Q_{i,n}\|^2], \qquad Q_{i,n} = \mathbb{E}[v_{i,n}|\mathcal{W}_n], \end{aligned}$$

where $v_{i,n} = x_{i,n} - (\sum_{j=1}^{n} \mathbb{E}[x_{j,n}w'_{j,n}])(\sum_{j=1}^{n} \mathbb{E}[w_{j,n}w'_{j,n}])^{-1} w_{i,n}$ is the population counterpart of $\hat{v}_{i,n}$. Also, define

$$\mathcal{C}_n = \max_{1 \leq i \leq n} \{\mathbb{E}[U_{i,n}^4|\mathcal{X}_n, \mathcal{W}_n] + \mathbb{E}[\|V_{i,n}\|^4|\mathcal{W}_n] + 1/\mathbb{E}[U_{i,n}^2|\mathcal{X}_n, \mathcal{W}_n]\} + 1/\lambda_{\min}(\mathbb{E}[\tilde{\Gamma}_n|\mathcal{W}_n])\},$$

where $U_{i,n} = y_{i,n} - \mathbb{E}[y_{i,n}|\mathcal{X}_n, \mathcal{W}_n]$, $V_{i,n} = x_{i,n} - \mathbb{E}[x_{i,n}|\mathcal{W}_n]$, $\tilde{\Gamma}_n = \sum_{i=1}^{n} \tilde{V}_{i,n}\tilde{V}'_{i,n}/n$, and $\tilde{V}_{i,n} = \sum_{j=1}^{n} M_{ij,n} V_{j,n}$.

In the supplemental appendix we show how the three examples fit in this general framework and verify that Assumptions LR1–LR3, PL1–PL3 and FE1–FE3, respectively, imply the following three assumptions.

Assumption 1 *$\mathbb{C}[U_{i,n}, U_{j,n}|\mathcal{X}_n, \mathcal{W}_n] = 0$ for $i \neq j$ and $\max_{1 \leq i \leq N_n} \#\mathcal{T}_{i,n} = O(1)$, where $\#\mathcal{T}_{i,n}$ is the cardinality of $\mathcal{T}_{i,n}$ and where $\{\mathcal{T}_{i,n} : 1 \leq i \leq N_n\}$ is a partition of $\{1, \ldots, n\}$ such that $\{(U_{t,n}, V_{t,n}) : t \in \mathcal{T}_{i,n}\}$ are independent over i conditional on $\mathcal{W}_n$.*

Assumption 2 *$\chi_n = O(1)$, $\varrho_n + n(\varrho_n - \rho_n) + n\chi_n\varrho_n = o(1)$, and $\max_{1 \leq i \leq n} |\hat{v}_{i,n}|/\sqrt{n} = o_p(1)$.*

Assumption 3 *$\mathbb{P}[\lambda_{\min}(\sum_{i=1}^{n} w_{i,n}w'_{i,n}) > 0] \to 1$, $\overline{\lim}_{n\to\infty} K_n/n < 1$, and $\mathcal{C}_n = O_p(1)$.*

4.2 General Results

As a means to the end of establishing (2), we give an asymptotic normality result for $\hat{\beta}_n$ which may be of interest in its own right.

Theorem 1 *Suppose Assumptions 1–3 hold. Then*

$$\Omega_n^{-1/2}\sqrt{n}(\hat{\beta}_n - \beta) \to_d \mathcal{N}(0, I_d), \qquad \Omega_n = \hat{\Gamma}_n^{-1}\Sigma_n\hat{\Gamma}_n^{-1}, \tag{6}$$

where $\Sigma_n = \sum_{i=1}^n \hat{v}_{i,n}\hat{v}_{i,n}'\mathbb{E}[U_{i,n}^2|\mathcal{X}_n, \mathcal{W}_n]/n$.

In the literature on high-dimensional linear models, Mammen (1993) obtains a similar asymptotic normality result as in Theorem 1 but under the condition $K_n^{1+\delta}/n \to 0$ for $\delta > 0$ restricted by certain moment condition on the covariates. In contrast, our result only requires $\overline{\lim}_{n\to\infty} K_n/n < 1$ but imposes a different restriction on the high-dimensional covariates (e.g., condition **(i)**, **(ii)** or **(iii)** discussed previously), and exploits the partially linear structure of the model (i.e., in Mammen (1993) notation, it considers the case $c = (\iota', 0')'$ with ι denoting a d-dimensional vector of ones and 0 denoting a K_n-dimensional vector of zeros). In addition, Theorem 1 is a substantial improvement over Cattaneo, Jansson, and Newey (2015, Theorem 1) because here it is not required that $K_n \to \infty$ nor $\chi_n = o(1)$, thereby allowing for quite general form of nuisance covariate $w_{i,n}$ beyond specific approximating basis functions (and thus the corresponding smoothness assumptions).

Achieving (2), the counterpart of (6) in which the unknown matrix Σ_n is replaced by the estimator $\hat{\Sigma}_n$, requires additional assumptions. One possibility is to impose homoskedasticity.

Theorem 2 *Suppose the assumptions of Theorem 1 hold. If* $\mathbb{E}[U_{i,n}^2|\mathcal{X}_n, \mathcal{W}_n] = \sigma_n^2$*, then (2) holds with* $\hat{\Sigma}_n = \hat{\Sigma}_n^{\texttt{HO}}$.

This result shows in quite some generality that homoskedastic inference in linear models remains valid even when K_n is proportional to n, provided the variance estimator incorporates a degrees-of-freedom correction, as $\hat{\Sigma}_n^{\texttt{HO}}$ does.

Establishing (2) is also possible when K_n is assumed to be a vanishing fraction of n, as is of course the case in the usual fixed-K_n linear regression model setup. The following theorem establishes consistency of the conventional standard error estimator $\hat{\Sigma}_n^{\texttt{EW}}$ under the assumption $\mathcal{M}_n \to_p 0$, and also derives an asymptotic representation for estimators of the form $\hat{\Sigma}_n(\kappa_n)$ without imposing this assumption.

Theorem 3 *Suppose the assumptions of Theorem 1 hold.*
(a) If $\mathcal{M}_n \to_p 0$*, then (2) holds with* $\hat{\Sigma}_n = \hat{\Sigma}_n^{\texttt{EW}}$.
(b) If $\|\kappa_n\|_\infty = \max_{1\le i\le n}\sum_{j=1}^n |\kappa_{ij,n}| = O_p(1)$*, then*

$$\hat{\Sigma}_n(\kappa_n) = \frac{1}{n}\sum_{i=1}^n\sum_{j=1}^n\sum_{k=1}^n \kappa_{ik,n}M_{kj,n}^2\hat{v}_{i,n}\hat{v}_{i,n}'\mathbb{E}[U_{j,n}^2|\mathcal{X}_n, \mathcal{W}_n] + o_p(1).$$

The conclusion of part (a) typically fails when the condition $K_n/n \to 0$ is dropped. For example, when specialized to $\kappa_n = I_n$ part (b) implies that in the homoskedastic case (i.e., when the assumptions of Theorem 2 are satisfied)

$$\hat{\Sigma}_n^{\texttt{EW}} = \Sigma_n - \frac{\sigma_n^2}{n}\sum_{i=1}^{n}(1 - M_{ii,n})\hat{v}_{i,n}\hat{v}_{i,n}' + o_p(1),$$

where $\sum_{i=1}^{n}(1 - M_{ii,n})\hat{v}_{i,n}\hat{v}_{i,n}'/n \neq o_p(1)$ in general (unless $K_n/n \to 0$). Similar remarks apply to the variants of the HCk estimators mentioned above; see the supplemental appendix for details. On the other hand, because $\sum_{1\leq k\leq n}\kappa_{ik,n}^{\texttt{HC}}M_{kj,n}^2 = \mathbb{1}(i = j)$ by construction, part (b) implies that $\hat{\Sigma}_n^{\texttt{HC}}$ is consistent provided $\|\kappa_n^{\texttt{HC}}\|_\infty = O_p(1)$. A simple condition for this to occur can be stated in terms of $\mathcal{M}_n$. Indeed, if $\mathcal{M}_n < 1/2$, then $\kappa_n^{\texttt{HC}}$ is diagonally dominant and it follows from Theorem 1 of Varah (1975) that

$$\|\kappa_n^{\texttt{HC}}\|_\infty \leq \frac{1}{1/2 - \mathcal{M}_n}.$$

As a consequence, we obtain the following theorem, whose conditions can hold even if $K_n/n \nrightarrow 0$.

Theorem 4 *Suppose the assumptions of Theorem 1 hold.*
If $\mathbb{P}[\mathcal{M}_n < 1/2] \to 1$ and if $1/(1/2 - \mathcal{M}_n) = O_p(1)$, then (2) holds with $\hat{\Sigma}_n = \hat{\Sigma}_n^{\texttt{HC}}$.

Because $\mathcal{M}_n \geq K_n/n$, a necessary condition for Theorem 4 to be applicable is that $\overline{\lim}_{n\to\infty}K_n/n < 1/2$. When the design is balanced, that is, when $M_{11,n} = \ldots = M_{nn,n}$ (as occurs in the panel data model (5)), the condition $\overline{\lim}_{n\to\infty}K_n/n < 1/2$ is also sufficient, but in general it seems difficult to formulate primitive sufficient conditions for the assumption made about $\mathcal{M}_n$ in Theorem 4. In practice, the fact that $\mathcal{M}_n$ is observed means that the condition $\mathcal{M}_n < 1/2$ is verifiable, and therefore unless $\mathcal{M}_n$ is found to be "close" to 1/2 there is reason to expect $\hat{\Sigma}_n^{\texttt{HC}}$ to perform well.

4.3 Examples

4.3.1 Linear Regression Model with Increasing Dimension

Specializing Theorems 2–4 to the linear regression model, we obtain the following result.

Theorem LR Suppose Assumptions LR1–LR3 hold.

(a) If $\mathbb{E}[u_{i,n}^2|x_{i,n}, z_{i,n}] = \sigma_n^2$, then (2) holds with $\hat{\Sigma}_n = \hat{\Sigma}_n^{\texttt{HO}}$.

(b) If $\mathcal{M}_n \to_p 0$, then (2) holds with $\hat{\Sigma}_n = \hat{\Sigma}_n^{\texttt{EW}}$.

(c) If $\mathbb{P}[\mathcal{M}_n < 1/2] \to 1$ and if $1/(1/2 - \mathcal{M}_n) = O_p(1)$, then (2) holds with $\hat{\Sigma}_n = \hat{\Sigma}_n^{\texttt{HC}}$.

This theorem gives a formal justification for employing $\hat{\Sigma}_n^{\texttt{HC}}$ as the variance estimator when forming confidence intervals for β in linear models with possibly many nuisance covariates and heteroskedasticity. The resulting confidence intervals for β will remain consistent even when K_n is proportional to n, provided the technical conditions given in part (c) are satisfied.

Remark 3. Our main results for linear models concern large-sample approximations for the finite-sample distribution of the usual t-statistics. An alternative, equally automatic approach is to employ the bootstrap and closely related resampling procedures (see, among others, Freedman (1981), Mammen (1993), Gonçalvez and White (2005), Kline and Santos (2012)). Assuming $K_n/n \nrightarrow 0$, Bickel and Freedman (1983) demonstrated an invalidity result for the bootstrap. We conjecture that similar results can be obtained for other resampling procedures. Furthermore, we also conjecture that employing appropriate resampling methods on the "bias-corrected" residuals $\tilde{u}_{i,n}^2$ (Remark 1) can lead to valid inference procedures. Investigating these conjectures, however, is beyond the scope of this paper.

4.3.2 Semiparametric Partially Linear Model

The results for the partially linear model (4) are in perfect analogy with those for the linear regression model.

Theorem PL Suppose Assumptions PL1–PL3 hold.

(a) If $\mathbb{E}[\varepsilon_i^2|x_i, z_i] = \sigma^2$, then (2) holds with $\hat{\Sigma}_n = \hat{\Sigma}_n^{\texttt{HO}}$.

(b) If $\mathcal{M}_n \to_p 0$, then (2) holds with $\hat{\Sigma}_n = \hat{\Sigma}_n^{\texttt{EW}}$.

(c) If $\mathbb{P}[\mathcal{M}_n < 1/2] \to 1$ and if $1/(1/2 - \mathcal{M}_n) = O_p(1)$, then (2) holds with $\hat{\Sigma}_n = \hat{\Sigma}_n^{\texttt{HC}}$.

A result similar to Theorem PL(a) was previously reported in Cattaneo, Jansson, and Newey (2015), but parts (b) and (c) of Theorem PL are new.

4.3.3 Fixed Effects Panel Data Regression Model

Finally, consider the panel data model (5). Because $K_n/n = 1/T$ is fixed this model does not admit an analog of Theorem 3. On the other hand, it does admit an analog of Theorems 2 and 4.

Theorem FE Suppose Assumptions FE1–FE3 hold. Then (2) holds with $\hat{\Sigma}_n = \hat{\Sigma}_n^{\texttt{HC}}$. If also $\mathbb{E}[U_{it}^2|X_{i1}, \ldots, X_{iT}] = \sigma^2$, then (2) holds with $\hat{\Sigma}_n = \hat{\Sigma}_n^{\texttt{HO}}$.

To see the connection between our results and those in Stock and Watson (2008), observe that $M_n = I_N \otimes [I_T - \iota_T\iota_T'/T]$ for $\iota_T \in \mathbb{R}^T$ a $T \times 1$ vector of ones. We then obtain $M_{ii,n} = 1 - 1/T$ (for $i = 1, \ldots, n$) and therefore $\mathcal{M}_n \leq 1/3$ because $T \geq 3$. More importantly, perhaps, we obtain a closed-form expression for $\kappa_n^{\texttt{HC}}$ given by

$$\kappa_n^{\texttt{HC}} = I_N \otimes \frac{T}{T-2}\left[I_T - \frac{1}{(T-1)^2}\iota_T\iota_T'\right].$$

As a consequence,

$$\hat{\Sigma}_n^{\texttt{HC}} = \frac{1}{N(T-2)}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{X}_{it}\tilde{X}_{it}'\hat{U}_{it}^2 - \frac{1}{N(T-2)}\sum_{i=1}^{N}\left(\frac{1}{T-1}\sum_{t=1}^{T}\tilde{X}_{it}\tilde{X}_{it}'\right)\left(\frac{1}{T-1}\sum_{t=1}^{T}\hat{U}_{it}^2\right),$$

where $\tilde{X}_{it} = X_{it} - T^{-1}\sum_{s=1}^{T} X_{is}$ and $\hat{U}_{it} = Y_{it} - T^{-1}\sum_{s=1}^{T} Y_{is} - \hat{\beta}_n' \tilde{X}_{it}$. Apart from an asymptotically negligible degrees of freedom correction, this estimator coincides with the estimator $\hat{\Sigma}^{HR-FE}$ of Stock and Watson (2008, Eq. (6), p. 156).

Remark 4. The result above not only highlights a tight connection between our general standard error estimator and the one in Stock and Watson (2008), but also indicates that our general formula $\hat{\Sigma}_n^{\mathtt{HC}}$ could be used to derive explicit, simple expressions in other contexts where multi-way fixed effects or similar discrete regressors are included.

5 Simulations

We report the results from a small Monte Carlo experiment aimed to capture the extent to which our main theoretical findings are present in samples of moderate size. To facilitate comparability with other studies, we employ a data generating process (DGP) that is as similar as possible to those employed in the literature before. In particular, we consider the following model:

$$\begin{aligned} y_i &= \beta x_i + \gamma' w_i + u_i, \qquad & u_i|(x_i, w_i) &\sim \text{i.i.d. } \mathcal{N}(0, \sigma_{ui}^2), \qquad & \sigma_{ui}^2 &= \varkappa_u (1 + (x_i + \iota' w_i)^2)^{\vartheta}, \\ x_i &= v_i, & v_i|w_i &\sim \text{i.i.d. } \mathcal{N}(0, \sigma_{vi}^2), & \sigma_{vi}^2 &= \varkappa_v (1 + (\iota' w_i)^2)^{\vartheta}, \end{aligned}$$

where $\iota = (1, 1, \cdots, 1)'$, $\beta = 0$ and $\gamma = 0$, and the constants $\varkappa_u$ and $\varkappa_v$ are chosen so that $\mathbb{V}[u_i] = \mathbb{V}[v_i] = 1$. In the absence of the additional covariates w_i, this design coincides with the one in Stock and Watson (2008), and is very similar to the one considered in MacKinnon (2012).

The simulation study employs $5,000$ replications, sets the sample size to $n = 1,000$, and considers models with $K_n/n \in \{0.1, 0.2, 0.3, 0.4\}$. The two main parameters varying in the Monte Carlo experiments are: the constant ϑ and the distribution of the covariates w_i. The first parameter controls the degree of heteroskedasticity: $\vartheta = 0$ corresponds to homoskedasticity, and $\vartheta = 1$ corresponds to moderate heteroskedasticity, as classified by MacKinnon (2012). For the distribution of the covariates we consider the following cases: independent standard $\mathcal{N}(0, 1)$ (Model 1), independent $\mathcal{U}(-1, 1)$ (Model 2), independent discrete covariates constructed as $\mathbf{1}(\mathcal{N}(0, 1) \geq 2.33)$.

The results are given in Table 1. Following MacKinnon (2012), these tables report empirical coverage rates for eight distinct nominal 95% confidence intervals for β, across the range of K_n and values of ϑ. Each confidence interval considered employs a different standard error formula: HO_0 uses homoskedastic standard errors without degrees of freedom correction, HO_1 uses homoskedastic standard errors with degrees of freedom correction, HC_0–HC_4 are described in footnote 1 (see also Section 4.5 in the supplemental appendix), and HC_K uses $\hat{\Sigma}_n^{\mathtt{HC}}$.

The main findings from the small simulation study are in line with our theoretical results. We find that the confidence interval estimators constructed our proposed standard errors formula $\hat{\Sigma}_n^{\mathtt{HC}}$, denoted HC_K, offer close-to-correct empirical coverage in all cases considered. The alternative heteroskedasticity consistent standard errors currently available in the literature lead to confidence intervals that could deliver substantial under or over coverage depending on the design and degree of heteroskedasticity considered. We also found that inference based on HC_3 standard errors is conservative, a general asymptotic result that is formally established in the supplemental appendix.

6 Conclusion

We established asymptotic normality of the OLS estimator of a subset of coefficients in high-dimensional linear regression models with many nuisance covariates, and investigated the properties of several popular heteroskedasticity-robust standard error estimators in this high-dimensional context. We showed that none of the usual formulas deliver consistent standard errors when the number of covariates is not a vanishing proportion of the sample size. We also proposed a new standard error formula that is consistent under (conditional) heteroskedasticity and many covariates, which is fully automatic and does not assume special, restrictive structure on the regressors.

Our results concern high-dimensional models where the number of covariates is at most a non-vanishing fraction of the sample size. A quite recent related literature concerns ultra-high-dimensional models where the number of covariates is much larger than the sample size, but some form of (approximate) sparsity is imposed in the model; see, e.g., Belloni, Chernozhukov, and Hansen (2014a,b), Belloni, Chernozhukov, Hansen, and Fernandez-Val (2014), Farrell (2015), and references therein. In that setting, inference is conducted after covariate selection, where the resulting number of selected covariates is at most a vanishing fraction of the sample size (usually much smaller). Thus, it would be of interest to investigate whether the methods proposed herein can be applied also for inference post covariate selection in ultra-high-dimensional settings, which would allow for weaker forms of sparsity because more covariates could be selected for inference.

References

Belloni, A., V. Chernozhukov, D. Chetverikov, and K. Kato (2015): "On the Asymptotic Theory for Least Squares Series: Pointwise and Uniform Results," *Journal of Econometrics*, 186(2), 345–366.

Belloni, A., V. Chernozhukov, and C. Hansen (2014a): "High-Dimensional Methods and Inference on Structural and Treatment Effects," *Journal of Economic Perspectives*, 28(2), 29–50.

——— (2014b): "Inference on Treatment Effects after Selection among High-Dimensional Controls," *Review of Economic Studies*, 81(2), 608–650.

Belloni, A., V. Chernozhukov, C. Hansen, and I. Fernandez-Val (2014): "Program Evaluation with High-Dimensional Data," Working paper, MIT.

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 *A Festschrift for Erich L. Lehmann*, ed. by P. Bickel, K. Doksum, and J. Hodges. Chapman and Hall.

Cattaneo, M. D., and M. H. Farrell (2013): "Optimal Convergence Rates, Bahadur Representation, and Asymptotic Normality of Partitioning Estimators," *Journal of Econometrics*, 174(2), 127–143.

Cattaneo, M. D., M. Jansson, and W. K. Newey (2015): "Alternative Asymptotics and the Partially Linear Model with Many Regressors," Working paper, University of Michigan.

Chen, X. (2007): "Large Sample Sieve Estimation of Semi-Nonparametric Models," in *Handbook of Econometrics, Volume VI*, ed. by J. J. Heckman, and E. E. Leamer, p. 5549Ű5632. Elsevier Science B.V.

Chesher, A. (1989): "Hájek Inequalities, Measures of Leverage and the Size of Heteroskedasticity Robust Wald Tests," *Econometrica*, 57(4), 971–977.

Chesher, A., and I. Jewitt (1987): "The Bias of a heteroskedasticity consistent covariance matrix estimator," *Econometrica*, 55(5), 1217–1222.

Cribari-Neto, F., and M. da Gloria A. Lima (2011): "A sequence of improved standard errors under heteroskedasticity of unknown form," *Journal of Statistical Planning and Inference*, 141(11), 3617–3627.

Cribari-Neto, F., S. L. P. Ferrari, and G. M. Cordeiro (2000): "Improved Heteroscedasticity-Consistent Covariance Matrix Estimators," *Biometrika*, 87(4), 907–918.

Donald, S. G., and W. K. Newey (1994): "Series Estimation of Semilinear Models," *Journal of Multivariate Analysis*, 50(1), 30–40.

El Karoui, N., D. Bean, P. J. Bickel, C. Lim, and B. Yu (2013): "On Robust Regression with High-Dimensional Predictors," *Proceedings of the National Academy of Sciences*, 110(36), 14557–14562.

Farrell, M. H. (2015): "Robust Inference on Average Treatment Effects with Possibly More Covariates than Observations," *Journal of Econometrics*, forthcoming.

Freedman, D. A. (1981): "Bootstrapping Regression Models," *Annals of Statistics*, 9(6), 1218–1228.

Gonçalvez, S., and H. White (2005): "Bootstrap Standard Error Estimates for Linear Regression," *Journal of the American Statistical Association*, 100(471), 970–979.

Huber, P. J. (1973): "Robust Regression: Asymptotics, Conjectures, and Monte Carlo," *Annals of Stastistics*, 1(5), 799–821.

KLINE, P., AND A. SANTOS (2012): "Higher order properties of the wild bootstrap under misspecification," *Journal of Econometrics*, 171(1), 54–70.

KOENKER, R. (1988): "Asymptotic Theory and Econometric Practice," *Journal of Applied Econometrics*, 3(2), 139–147.

MACKINNON, J., AND H. WHITE (1985): "Some Heteroskedasticity-consistent Covariance Matrix Estimators with Improved Finite Sample Properties," *Journal of Econometrics*, 29, 305–325.

MACKINNON, J. G. (2012): "Thirty years of heteroskedasticity-robust inference," in *Recent Advances and Future Directions in Causality, Prediction, and Specification Analysis*, ed. by X. Chen, and N. R. Swanson. Springer.

MAMMEN, E. (1993): "Bootstrap and Wild Bootstrap for High Dimensional Linear Models," *Annals of Statistics*, 21(1), 255–285.

MÜLLER, U. K. (2013): "Risk of Bayesian Inference in Misspecified Models, and the Sandwich Covariance Matrix," *Econometrica*, 81(5), 1805–1849.

NEWEY, W. K. (1997): "Convergence Rates and Asymptotic Normality for Series Estimators," *Journal of Econometrics*, 79, 147–168.

SHAO, J., AND C. F. J. WU (1987): "Heteroscedasticity-Robustness of Jackknife Variance Estimators in Linear Models," *Annals of Statistics*, 15(4), 1563–1579.

STOCK, J. H., AND M. W. WATSON (2008): "Heteroskedasticity-Robust Standard Errors for Fixed Effects Panel Data Regression," *Econometrica*, 76(1), 155–174.

VARAH, J. M. (1975): "A Lower Bound for the Smallest Singular Value of a Matrix," *Linear Algebra and its Applications*, 11(1), 3–5.

WHITE, H. (1980): "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," *Econometrica*, 48(4), 817–838.

WU, C. F. J. (1986): "Jackknife, Bootstrap and Other Resampling Methods in Regression Analysis," *Annals of Statistics*, 14(4), 1261–1295.

Table 1: Empirical Coverage of 95% Confidence Intervals.

(a) Model 1: Gaussian $w_{i,n}$ Regressors

ϑ	K_n/n	HO_0	HO_1	HC_0	HC_1	HC_2	HC_3	HC_4	HC_K
0	0.1	0.939	0.952	0.940	0.950	0.950	0.963	0.977	0.950
0	0.2	0.922	0.952	0.920	0.952	0.951	0.969	0.993	0.950
0	0.3	0.899	0.948	0.897	0.949	0.949	0.981	0.987	0.947
0	0.4	0.867	0.952	0.866	0.951	0.951	0.987	0.974	0.950
1	0.1	0.421	0.442	0.899	0.917	0.918	0.932	0.962	0.929
1	0.2	0.436	0.479	0.852	0.896	0.896	0.934	0.976	0.929
1	0.3	0.446	0.516	0.809	0.878	0.881	0.941	0.955	0.928
1	0.4	0.442	0.554	0.742	0.858	0.858	0.937	0.902	0.922

(b) Model 2: Uniform $w_{i,n}$ Regressors

ϑ	K_n/n	HO_0	HO_1	HC_0	HC_1	HC_2	HC_3	HC_4	HC_K
0	0.1	0.937	0.950	0.938	0.950	0.950	0.962	0.980	0.950
0	0.2	0.930	0.959	0.929	0.960	0.960	0.975	0.993	0.959
0	0.3	0.905	0.955	0.904	0.953	0.952	0.982	0.988	0.953
0	0.4	0.872	0.951	0.870	0.952	0.951	0.989	0.973	0.950
1	0.1	0.387	0.406	0.901	0.922	0.922	0.939	0.964	0.936
1	0.2	0.420	0.463	0.862	0.905	0.906	0.939	0.981	0.932
1	0.3	0.427	0.499	0.807	0.886	0.885	0.943	0.959	0.931
1	0.4	0.419	0.521	0.736	0.853	0.853	0.942	0.908	0.927

(c) Model 3: Discrete $w_{i,n}$ Regressors

ϑ	K_n/n	HO_0	HO_1	HC_0	HC_1	HC_2	HC_3	HC_4	HC_K
0	0.1	0.936	0.949	0.935	0.948	0.947	0.960	0.976	0.946
0	0.2	0.919	0.948	0.920	0.948	0.947	0.967	0.993	0.946
0	0.3	0.896	0.945	0.897	0.947	0.946	0.978	0.982	0.945
0	0.4	0.868	0.945	0.871	0.947	0.943	0.988	0.971	0.943
1	0.1	0.346	0.366	0.834	0.861	0.900	0.949	0.991	0.942
1	0.2	0.516	0.569	0.802	0.856	0.893	0.957	0.992	0.940
1	0.3	0.616	0.703	0.777	0.858	0.892	0.970	0.964	0.943
1	0.4	0.670	0.790	0.751	0.867	0.899	0.982	0.927	0.950

Notes:
(i) $\vartheta = 0$ and $\vartheta = 1$ correspond to homoskedastic and heteroskedastic models, respectively.
(ii) HO_0 and HO_1 employ homoskedastic consistent standard errors without and with degrees of freedom correction, respectively.
(iii) HC_0–HC_4 employ HC_k heteroskedastic consistent standard errors; see footnote 1 for details.
(iv) HC_K employs our proposed standard errors formula, denoted by $\hat{\Sigma}_n^{\mathtt{HC}}$.