cemmap

centre for microdata methods and practice

Inference in linear regression models with many covariates and heteroskedasticity

Matias D. Cattaneo
Michael Jansson
Whitney K. Newey

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP03/17

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

An ESRC Research Centre

Inference in Linear Regression Models with Many Covariates and Heteroskedasticity*

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

January 12, 2017

Abstract

The linear regression model is widely used in empirical work in Economics, Statistics, and many other disciplines. 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 included 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: parametric linear models with many covariates, linear panel models with many fixed effects, and semiparametric semi-linear models with many technical regressors. Simulation evidence consistent with our theoretical results is also provided. The proposed methods are also illustrated with an empirical application.

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

*We thank Xinwei Ma, Ulrich Müller and Andres Santos for very thoughtful discussions regarding this project. We also thank Silvia Gonçalvez, Pat Kline and James MacKinnon. In addition, an Associate Editor and three reviewers offered excellent recommendations that improved this paper. 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 and Department of Statistics, 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.

Not only it is often the case that substantive discipline-specific theory (or intuition) will suggest a large set of variables that might be important, but also researchers usually prefer to include additional "technical" controls constructed using indicator variables, interactions, and other non-linear transformations of those variables. Therefore, many empirical studies include very many covariates in order to control for as broad array of confounders as possible. For example, it is common practice 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 (Belloni, Chernozhukov, and Hansen (2014)), many controls usually may remain in the final model specification. For example, Angrist and Hahn (2004) discuss when to include many covariates in treatment effect models.

We present valid inference methods that explicitly account for the presence of possibly many controls in linear regression models under (conditional) heteroskedasticity. We consider the setting where the object of interest is $\boldsymbol{\beta}$ in a model of the form

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

where $y_{i,n}$ is a scalar outcome variable, $\mathbf{x}_{i,n}$ is a regressor of small (i.e., fixed) dimension d,

$\mathbf{w}_{i,n}$ is a vector of covariates of possibly "large" 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 multi-way fixed effects and interactions thereof in panel data models. In both cases conducting OLS-based inference on $\boldsymbol{\beta}$ 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, is inappropriate in applications with many dummy variables and does not deliver a good distributional 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 $\mathbf{w}_{i,n}$ 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 $\boldsymbol{\beta}$ in (1) without (necessarily) assuming any special structure on the part of the covariate $\mathbf{w}_{i,n}$. 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 $\boldsymbol{\beta}$, allowing for the dimension of the nuisance covariates to be "large" relative to the sample size (i.e., $K_n/n \not\to 0$). 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. As a by-product, we also find that among the popular HCk class of standard errors estimators for linear models, a variant of the HC3 estimator delivers standard errors that are asymptotically upward biased in general. Thus, standard OLS inference employing HC3 standard errors will be asymptotically valid, albeit conservative, even in high-dimensional settings where the number of covariate $\mathbf{w}_{i,n}$ is large relative to the sample size, i.e., when $K_n/n \not\to 0$.

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), Kauermann and Carroll (2001), Bera, Suprayitno, and Premaratne (2002), Stock and Watson (2008), Cribari-Neto and da Gloria A. Lima (2011), Müller (2013), and Abadie, Imbens, and Zheng (2014). 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. They use unbiasedness or minimum norm quadratic unbiasedness to motivate a variance estimator that is similar in structure to ours, but their results are obtained for fixed K_n and n and is silent about the extent to which consistent variance estimation is even possible when $K_n/n \not\to 0$.

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), Zheng, Jiang, Bai, and He (2014), Li and Müller (2017), 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 in general high-dimensional linear models 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. In particular, we do not require the coefficients $\boldsymbol{\gamma}_n$ to be consistently estimated; in fact, they will not be in most of our examples discussed 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 $\mathbf{w}_{i,n}$ covers several special cases of interest for empirical work. For example, our results encompass (and weakens in certain sense) those reported in Stock and Watson (2008), who investigated the one-way fixed effects panel data regression model and showed that the conventional Eicker-White heteroskedasticity-robust variance estimator is inconsistent, 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 estimator enables an explicit characterization of this bias, and also leads to a direct plug-in consistent bias-corrected version of the Eicker-White variance estimator. The generic variance estimator proposed herein essentially reduces to this 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 and generalize to other settings.

Furthermore, our general inference results can be used when many multi-way fixed effects and similar discrete covariates are introduced in a linear regression model, as it is usually the case in social interaction and network settings. For example, in a very recent contribution, Verdier (2017) develops new results for two-way fixed effect design and projection matrices, and use them to verify our high-level conditions in linear models with two-way unobserved heterogeneity and sparsely matched data (which can also be interpreted as a network setting). These results provide another interesting and empirically relevant illustration of our generic theory. Verdier (2017) also develops inference results able to handle time series dependence in his specific context, which are not covered by our assumptions because we impose independence in the cross-sectional dimension of the (possibly grouped) data.

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 our general framework, discusses high-level assumptions and illustrates the applicability of our methods using three leading examples. Section 4 gives the main results of the paper. Section 5 reports the results of a Monte Carlo experiment, while 6 illustrates our methods using an empirical application. Section 7 concludes. Proofs and additional methodological and numerical results are reported in the online supplemental appendix.

2 Overview of Results

For the purposes of discussing distribution theory and variance estimators associated with the OLS estimator $\hat{\boldsymbol{\beta}}_n$ of $\boldsymbol{\beta}$ in (1), when possibly the K_n-dimensional nuisance covariates $\mathbf{w}_{i,n}$ is of "large" dimension and/or the parameters $\boldsymbol{\gamma}_n$ cannot be estimated consistently, it is convenient to write the estimator in "partialled out" form as

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

where $M_{ij,n} = \mathbb{1}(i=j) - \mathbf{w}_{i,n}'(\sum_{k=1}^{n} \mathbf{w}_{k,n}\mathbf{w}_{k,n}')^{-1}\mathbf{w}_{j,n}$, $\mathbb{1}(\cdot)$ denotes the indicator function, and the relevant inverses are assumed to exist. Defining $\hat{\boldsymbol{\Gamma}}_n = \sum_{i=1}^{n} \hat{\mathbf{v}}_{i,n}\hat{\mathbf{v}}_{i,n}'/n$, the objective is to establish a valid Gaussian distributional approximation of the finite sample distribution of the OLS estimator $\hat{\boldsymbol{\beta}}_n$, and then find an estimator $\hat{\boldsymbol{\Sigma}}_n$ of the variance of $\sum_{i=1}^{n} \hat{\mathbf{v}}_{i,n}u_{i,n}/\sqrt{n}$ such that

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

in which case asymptotic valid inference on $\boldsymbol{\beta}$ can be conducted in the usual way by employing the distributional approximation $\hat{\boldsymbol{\beta}}_n \overset{a}{\sim} \mathcal{N}(\boldsymbol{\beta}, \hat{\boldsymbol{\Omega}}_n/n)$. Our assumptions below will ensure that

$\hat{\boldsymbol{\beta}}_n$ remains $\sqrt{n}$-consistent because we show in the supplemental appendix that $\hat{\boldsymbol{\Omega}}_n^{-1} = O_p(1)$ even when $K_n/n \not\to 0$.

Our first result, Theorem 1 below, gives sufficient conditions for a valid Gaussian approximation of the distribution of the infeasible statistic $\boldsymbol{\Omega}_n^{-1/2}\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})$, where $\boldsymbol{\Omega}_n = \hat{\boldsymbol{\Gamma}}_n^{-1}\boldsymbol{\Sigma}_n\hat{\boldsymbol{\Gamma}}_n^{-1}$ and $\boldsymbol{\Sigma}_n$ denotes the variance of $\sum_{i=1}^n \hat{\mathbf{v}}_{i,n}u_{i,n}/\sqrt{n}$, even when possibly $K_n/n \not\to 0$ and the linear regression model exhibits conditional heteroskedasticity. This result, in turn, gives the basic ingredient for discussing valid variance estimation in high-dimensional linear regression models. Defining $\hat{u}_{i,n} = \sum_{j=1}^n M_{ij,n}(y_{j,n} - \hat{\boldsymbol{\beta}}_n'\mathbf{x}_{j,n})$, standard choices of $\hat{\boldsymbol{\Sigma}}_n$ in the fixed-K_n case include the homoskedasticity-only estimator

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

and the Eicker-White-type estimator

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

Perhaps not too surprisingly, in Theorem 2 below, we find that consistency of $\hat{\boldsymbol{\Sigma}}_n^{\mathtt{HO}}$ under homoskedasticity holds quite generally even for models with many covariates. In contrast, construction of a heteroskedasticity-robust estimator of $\boldsymbol{\Sigma}_n$ is more challenging, as it turns out that consistency of $\hat{\boldsymbol{\Sigma}}_n^{\mathtt{EW}}$ generally requires K_n to be a vanishing fraction of n.

To fix ideas, suppose $(y_{i,n}, \mathbf{x}_{i,n}', \mathbf{w}_{i,n}')$ are i.i.d. over i. It turns out that, under certain regularity conditions,

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

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

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

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 $\hat{\boldsymbol{\Sigma}}_n^{\mathtt{EW}}$, we study the more general class of estimators of the form

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

where $\kappa_{ij,n}$ denotes element (i,j) of a symmetric matrix $\boldsymbol{\kappa}_n = \boldsymbol{\kappa}_n(\mathbf{w}_{1,n},\ldots,\mathbf{w}_{n,n})$. Estimators that can be written in this fashion include $\hat{\boldsymbol{\Sigma}}_n^{\mathtt{EW}}$ (which corresponds to $\boldsymbol{\kappa}_n = \mathbf{I}_n$) as well as variants of the so-called HCk estimators, $k \in \{1,2,3,4\}$, reviewed by Long and Ervin (2000) and MacKinnon (2012), among many others. 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 $(\Upsilon_{i,n},\xi_{i,n}) = (1,0)$ for HC0 (and corresponding to $\hat{\boldsymbol{\Sigma}}_n^{\mathtt{EW}}$), $(\Upsilon_{i,n},\xi_{i,n}) = (n/(n-K_n),0)$ 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. See Sections 4.3 for more details.

In Theorem 3 below, we show that all of the HCk-type estimators, which correspond to a diagonal choice of $\boldsymbol{\kappa}_n$, have 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 $\boldsymbol{\kappa}_n$ 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 $\kappa_{ij,n}$, $\hat{\boldsymbol{\Sigma}}_n(\boldsymbol{\kappa}_n)$ satisfies

$$\hat{\boldsymbol{\Sigma}}_n(\boldsymbol{\kappa}_n) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\kappa_{ik,n}M_{kj,n}^2\hat{\mathbf{v}}_{i,n}\hat{\mathbf{v}}_{i,n}'\mathbb{E}[u_{j,n}^2|\mathbf{x}_{j,n},\mathbf{w}_{j,n}] + o_p(1), \tag{4}$$

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

$$\sum_{k=1}^{n}\kappa_{ik,n}M_{kj,n}^2 = \mathbb{1}(i=j), \qquad 1 \leq i,j \leq n. \tag{5}$$

Accordingly, we define

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

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

$$\boldsymbol{\kappa}_n^{\texttt{HC}} = \begin{pmatrix} \kappa_{11,n}^{\texttt{HC}} & \cdots & \kappa_{1n,n}^{\texttt{HC}} \\ \vdots & \ddots & \vdots \\ \kappa_{n1,n}^{\texttt{HC}} & \cdots & \kappa_{nn,n}^{\texttt{HC}} \end{pmatrix} = \begin{pmatrix} M_{11,n}^2 & \cdots & M_{1n,n}^2 \\ \vdots & \ddots & \vdots \\ M_{n1,n}^2 & \cdots & M_{nn,n}^2 \end{pmatrix}^{-1} = (\mathbf{M}_n \odot \mathbf{M}_n)^{-1}.$$

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

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

The fact that $\mathcal{M}_n < 1/2$ implies invertibility of $\mathbf{M}_n \odot \mathbf{M}_n$ is a consequence of the Gershgorin circle theorem. For details, see Section 3 in the supplemental appendix. 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 $\hat{\boldsymbol{\Sigma}}_n = \hat{\boldsymbol{\Sigma}}_n^{\texttt{HC}}$. Our final result, Theorem 4 below, formalizes this finding (see also the supplemental appendix for further intuition underlying this result).

The key intuition underlying our variance estimation result is that, even though each conditional variance $\mathbb{E}[u_{i,n}^2|\mathbf{x}_{i,n}, \mathbf{w}_{i,n}]$ cannot be well estimated due to the curse of dimensionality, an averaged version such as the leading term in (3) can be estimated consistently. Thus, taking $\widehat{\mathbb{E}}[u_{i,n}^2|\mathbf{x}_{i,n}, \mathbf{w}_{i,n}] = \sum_{k=1}^n \kappa_{ik,n} \hat{u}_{k,n}^2$ as an estimator of $\mathbb{E}[u_{i,n}^2|\mathbf{x}_{i,n}, \mathbf{w}_{i,n}]$, plugging into the leading term in (3), and computing conditional expectations, we obtain the leading term in (4). To make this leading term equal to the desired target $\sum_{i=1}^n \hat{\mathbf{v}}_{i,n}\hat{\mathbf{v}}_{i,n}'\mathbb{E}[u_{i,n}^2|\mathbf{x}_{i,n}, \mathbf{w}_{i,n}]$, it is natural to require

$$\sum_{j=1}^n \sum_{k=1}^n \kappa_{ik,n} M_{kj,n}^2 \mathbb{E}[u_{j,n}^2|\mathbf{x}_{j,n}, \mathbf{w}_{j,n}] = \mathbb{E}[u_{i,n}^2|\mathbf{x}_{i,n}, \mathbf{w}_{i,n}] \qquad 1 \le i \le n.$$

Since $\mathbb{E}[u_{i,n}^2|\mathbf{x}_{i,n}, \mathbf{w}_{i,n}]$ are unknown, our variance estimator solves (5), which generates enough equations to solve for all $n(n-1)/2$ possibly distinct elements in $\boldsymbol{\kappa}_n^{\texttt{HC}}$.

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

3 Setup

This section introduces a general framework encompassing several special cases of linear-in-parameters regression models of the form (1). We first present generic high-level assumptions, and then discuss their implications as well as some easier to verify sufficient conditions. Finally, to close this setup section, we briefly discuss three motivating leading examples: linear regression models with increasing dimension, muti-way fixed effect linear models, and semi-parametric semi-linear regression. Technical details and related results for these examples are given in the supplemental appendix.

3.1 Framework

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

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

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

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

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

We impose the following three high-level conditions. Let $\lambda_{\min}(\cdot)$ denote the minimum eigenvalue of its argument, and $\overline{\lim}_{n\to\infty} a_n = \limsup_{n\to\infty} a_n$ for any sequence a_n.

Assumption 1 (Sampling) *$\mathbb{C}[U_{i,n}, U_{j,n}|\mathcal{X}_n, \mathcal{W}_n] = 0$ 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 (Design) *$\mathbb{P}[\lambda_{\min}(\sum_{i=1}^n \mathbf{w}_{i,n}\mathbf{w}'_{i,n}) > 0] \to 1$, $\overline{\lim}_{n\to\infty} K_n/n < 1$, and $\mathcal{C}_n = O_p(1)$.*

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

3.2 Discussion of Assumptions

Assumptions 1–3 are meant to be high-level and general, allowing for different linear-in-parameters regression models. We now discuss the main restrictions imposed by these assumptions. We further illustrate them in the following subsection using more specific examples.

3.2.1 Assumption 1

This assumption concerns the sampling properties of the observed data. It generalizes classical i.i.d. sampling by allowing for groups or "clusters" of finite but possibly heterogeneous size with arbitrary intra-group dependence, which is very common in the context of fixed effects linear regression models. As currently stated, this assumption does not allow for dependence in the error terms across units, and therefore excludes clustered, spacial or time series dependence in the sample. We conjecture our main results extend to the latter cases, though here we focus on i.n.i.d. (conditionally) heteroskedastic models only, and hence relegate the extension to errors exhibiting clustered, spacial or time series dependence for future work. Assumption 1 reduces to classical i.i.d. sampling when $N_n = n$, $\mathcal{T}_{i,n} = \{i\}$ [implying $\max_{1\leq i\leq N_n} \#\mathcal{T}_{i,n} = 1$], and all observations have the same distribution.

3.2.2 Assumption 2

This assumption concerns basic design features of the linear regression model. The first two restrictions are mild and reflect the main goal of this paper, that is, analyzing linear regression models with many nuisance covariates $\mathbf{w}_{i,n}$. In practice, the first restriction regarding the minimum eigenvalue of the design matrix $\sum_{i=1}^{n} \mathbf{w}_{i,n}\mathbf{w}_{i,n}'$ is always imposed by removing redundant (i.e., linearly dependent) covariates; from a theoretical perspective this condition requires either restrictions on the distributional relationship of such covariates or some form of trimming leading to selection of included covariates (e.g., most software packages remove covariates leading to "too" small eigenvalues of the design matrix by means of some hard-thresholding rule).

On the other hand, the last condition, $\mathcal{C}_n = O_p(1)$, may be restrictive in some settings: for example, if the covariates have unbounded support (e.g., they are normally distributed) and heteroskedasticity is unbounded (e.g., unbounded multiplicative heteroskedasticity), then the assumption may fail. Simple sufficient conditions for $\mathcal{C}_n = O_p(1)$ can be formulated when the covariates have compact support, or the heteroskedasticity is multiplicative and bounded, because in these cases it is easy to bound the conditional moments of the error terms. It would be useful to know whether the condition $\mathcal{C}_n = O_p(1)$ can be relaxed to a version involving only unconditional moments, though we conjecture this weaker assumption will require a different method of proof (see the supplemental appendix for details).

3.2.3 Assumption 3

This assumption requires two basic approximations to hold. First, concerning bias, conditions on ϱ_n are related to the approximation quality of the linear-in-parameters model (1) for the "long" conditional expectation $\mathbb{E}[y_{i,n}|\mathcal{X}_n, \mathcal{W}_n]$. Similarly, conditions on ρ_n and χ_n are related to linear-in-parameters approximations for the "short" conditional expectations $\mathbb{E}[y_{i,n}|\mathcal{W}_n]$ and $\mathbb{E}[\mathbf{x}_{i,n}|\mathcal{W}_n]$, respectively. All these approximations are measured in terms of population mean square error, and are at the heart of empirical work employing linear-in-parameters regression models. Depending on the model of interest, different sufficient conditions can be given for these assumptions. Here we briefly mention the most simple one: (a) if $\mathbb{E}[u_{i,n}|\mathcal{X}_n, \mathcal{W}_n] = 0$ for all i and n, which can be interpreted as exogeneity (e.g., no misspecification bias), then $0 = \rho_n = n(\varrho_n - \rho_n) + n\chi_n\varrho_n$ for all n; and (b) if $\mathbb{E}[\|\mathbf{x}_{i,n}\|^2] < \infty$ for all i and n, then $\chi_n = O(1)$. Other sufficient conditions are discussed below.

Second, the high-level condition $\max_{1\leq i\leq n} \|\hat{\mathbf{v}}_{i,n}\|/\sqrt{n} = o_p(1)$ restricts the distributional relationship between the finite dimensional covariate of interest $\mathbf{x}_{i,n}$ and the high-dimensional nuisance covariate $\mathbf{w}_{i,n}$. This condition can be interpreted as a negligibility condition and thus comes close to minimal for the central limit theorem to hold. At the present level of generality it seems difficult to formulate primitive sufficient conditions for this restriction 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 Lemma SA-7 in the supplemental appendix for details and weaker conditions):

(i) $\mathcal{M}_n = o_p(1)$, or

(ii) $\chi_n = o(1)$, 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 $\mathbb{E}[\mathbf{x}_{i,n}|\mathbf{w}_{i,n}] \approx \boldsymbol{\delta}_n'\mathbf{w}_{i,n}$ with $\boldsymbol{\delta}_n = \mathbb{E}[\mathbf{w}_{i,n}\mathbf{w}_{i,n}']^{-1}\mathbb{E}[\mathbf{w}_{i,n}\mathbf{x}_{i,n}']$, primitive conditions for it are available when, for example, the elements of $\mathbf{w}_{i,n}$ are approximating functions. Indeed, in such cases one typically has $\chi_n = O(K_n^{-\alpha})$ 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. This condition also holds when $\mathbf{w}_{i,n}$ are dummy variables or discrete covariates, as we discuss in more detail below.

Finally, condition **(iii)**, and its underlying higher-level condition described in the supplemental appendix, is useful to handle cases where $\mathbf{w}_{i,n}$ cannot 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 projection matrix $\mathbf{M}_n$, which allows for $K_n/n \nrightarrow 0$. The condition is easy to verify in certain cases, including those where "locally bounded" approximating functions or fixed effects are used (see below for concrete examples).

3.3 Motivating Examples

We briefly mention three motivating examples of linear-in-parameter regression models covered by our results. All technical details are given in the supplemental appendix.

3.3.1 Linear Regression Model with Increasing Dimension

This leading example has a long tradition in statistics and econometrics. The model takes (1) as the data generating process (DGP), typically with i.i.d. data and the exogeneity condition $\mathbb{E}[u_{i,n}|\mathbf{x}_{i,n}, \mathbf{w}_{i,n}] = 0$. However, our assumptions only require $n\mathbb{E}[(\mathbb{E}[u_{i,n}|\mathbf{x}_{i,n}, \mathbf{w}_{i,n}])^2] = o(1)$, and hence (1) can be interpreted as a linear-in-parameters mean-square approximation to the unknown conditional expectation $\mathbb{E}[y_{i,n}|\mathbf{x}_{i,n}, \mathbf{w}_{i,n}]$. Either way, $\hat{\boldsymbol{\beta}}_n$ is the standard OLS estimator.

Setting $\mathcal{W}_n = (\mathbf{w}_{1,n}, \ldots, \mathbf{w}_{n,n})$, $N_n = n$, $\mathcal{T}_{i,n} = \{i\}$ and $\max_{1 \le i \le N_n} \#\mathcal{T}_{i,n} = 1$, Assumptions 1–2 are standard, while Assumption 3 is satisfied provided that $\mathbb{E}[\|\mathbf{x}_{i,n}\|^2] < \infty$ [implying $\chi_n = O(1)$], $n\mathbb{E}[(\mathbb{E}[u_{i,n}|\mathbf{x}_{i,n}, \mathbf{w}_{i,n}])^2] = o(1)$ [implying $n(\varrho_n - \rho_n) + n\chi_n\varrho_n = o(1)$], and $\max_{1 \le i \le n} \|\hat{\mathbf{v}}_{i,n}\|/\sqrt{n} = o_p(1)$. Primitive sufficient conditions for the latter negligibility condition can be given as discussed above. For example, under regularity conditions, $\chi_n = o(1)$ if either (a) $\mathbb{E}[\mathbf{x}_{i,n}|\mathbf{w}_{i,n}] = \boldsymbol{\delta}'\mathbf{w}_{i,n}$, (b) the nuisance covariates are discrete and a saturated dummy variables model is used, or (c) $\mathbf{w}_{i,n}$ are constructed using sieve functions. Alternatively, $\max_{1 \le i \le n} \sum_{j=1}^{n} \mathbb{1}(M_{ij,n} \neq 0) = o_p(n^{1/3})$ is satisfied provided the distribution of the nuisance covariates $\mathbf{w}_{i,n}$ generates a projection matrix $\mathbf{M}_n$ that is approximately a band matrix (see below for concrete examples). Precise regularity conditions for this example are given in Section 4.1 of the supplemental appendix.

3.3.2 Fixed Effects Panel Data Regression Model

A second class of examples covered by our results are linear panel data models with multi-way fixed effects and related models such as those encountered in networks, spillovers or social interactions settings. A common feature in these examples is the presence of possibly many dummy variables in $\mathbf{w}_{i,n}$, capturing unobserved heterogeneity or other unobserved effects across units (e.g., network link or spillover effect). In many applications the number of distinct dummy-type variables is large because researchers often include multi-group indicators, interactions thereof, and similar regressors obtained from factor variables. In these complicated models the nuisance covariates need to be estimated explicitly, even in simple linear regression problems, because it is not possible to difference out the multi-way indicator variables for estimation and inference.

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

$$Y_{it} = \alpha_i + \boldsymbol{\beta}'\mathbf{X}_{it} + U_{it}, \qquad i = 1, \ldots, N, \qquad t = 1, \ldots, T, \tag{6}$$

where $\alpha_i \in \mathbb{R}$ is an individual-specific intercept, $\mathbf{X}_{it}$ is a regressor of dimension d, and U_{it} is an scalar error term, and the following assumptions are satisfied. To map this model into our framework, suppose that $\{(U_{i1}, \dots, U_{iT}, \mathbf{X}_{i1}' \dots, \mathbf{X}_{iT}') : 1 \leq i \leq n\}$ are independent over i, $\mathbb{E}[U_{it}|\mathbf{X}_{i1} \dots, \mathbf{X}_{iT}] = 0$, and $\mathbb{E}[U_{it}U_{is}|X_{i1} \dots, X_{iT}] = 0$ for $t \neq s$. Then, setting $n = NT$, $K_n = N$, $\boldsymbol{\gamma}_n = (\alpha_1, \dots, \alpha_N)'$, and $(y_{(i-1)T+t,n}, \mathbf{x}_{(i-1)T+t,n}', u_{(i-1)T+t,n}, \mathbf{w}_{(i-1)T+t,n}') = (Y_{it}, \mathbf{X}_{it}', U_{it}, \mathbf{e}_{i,N}')$, $1 \leq i \leq N$ and $1 \leq t \leq T$, where $\mathbf{e}_{i,N} \in \mathbb{R}^N$ is the i-th unit vector of dimension N, the model (6) is also of the form (1) and $\hat{\boldsymbol{\beta}}_n$ is the fixed effects estimator of $\boldsymbol{\beta}$. In general, this model does not satisfy an i.i.d. assumption, but Assumption 1 enables us to employ results for independent random variables when developing asymptotics. In particular, unlike Stock and Watson (2008), we do not require $(U_{i1}, \dots, U_{iT}, \mathbf{X}_{i1}' \dots, \mathbf{X}_{iT}')$ to be i.i.d. over i, nor we require any kind of stationarity on the part of $(U_{it}, \mathbf{X}_{it}')$. The amount of variance heterogeneity permitted is quite large, since we basically only require $\mathbb{V}[Y_{it}|\mathbf{X}_{i1}, \dots, \mathbf{X}_{iT}] = \mathbb{E}[U_{it}^2|\mathbf{X}_{i1}, \dots, \mathbf{X}_{iT}]$ to be bounded and bounded away from zero. (On the other hand, serial correlation is assumed away because our assumptions imply that $\mathbb{C}[Y_{it}, Y_{is}|\mathbf{X}_{i1}, \dots, \mathbf{X}_{iT}] = 0$ for $t \neq s$.) In other respects this model is in fact more tractable than the previous models due to the special nature of the covariates $\mathbf{w}_{i,n}$, that is, a dummy variable for each unit $i = 1, \dots, N$.

In this one-way fixed effects example, $K_n/n = 1/T$ and therefore a high-dimensional model corresponds to a short panel model: $\max_{1 \leq i \leq n} \sum_{j=1}^{n} \mathbb{1}(M_{ij,n} \neq 0) = T$ and hence the negligibility condition holds easily. If $T \geq 2$, our asymptotic Gaussian approximation for the distribution of the least-squares estimator $\hat{\boldsymbol{\beta}}_n$ is valid (see Theorem 1), despite the coefficients $\boldsymbol{\gamma}_n$ not being consistently estimated. On the other hand, consistency of our generic variance estimator requires $T \geq 3$ [implying $K_n/n < 1/2$]; see Theorems 3 and 4. Further details are given in Section 4.2 of the supplemental appendix, where we also discuss a case-specific consistent variance estimator when $T = 2$.

Our generic results go beyond one-way fixed effect linear regression models, as they can be used to obtain valid inference in other contexts where multi-way fixed effects or similar discrete regressors are included. For a second concrete example, consider the recent work of Verdier (2017, and references therein) in the context of linear models with two-way unobserved heterogeneity and sparsely matched data. This model is isomorphic to a network model, where students and teacher (or workers and firms, for another example) are "matched" or "connected" over time, but potential unobserved heterogeneity at both levels is a concern. In this setting, under random sampling, Verdier (2017) offers primitive conditions for our high-level assumptions when two-way fixed effect models are used for estimation and inference. In particular, using a clever Markov chain argument (see his Lemma 1), he is able to provide different restriction on T and the number of matches in the network to ensure consistent variance estimation using the methods developed in this paper. To give one concrete example, he finds that if $T \geq 5$ and for any pair of teachers (firms), the number of students (workers) assigned to both teachers (firms) in the pair is either zero or greater than three, then our key high-level condition in Theorem 4 below is verified.

3.3.3 Semiparametric Partially Linear Model

Another model covered by our results is the partially linear model

$$y_i = \boldsymbol{\beta}'\mathbf{x}_i + g(\mathbf{z}_i) + \varepsilon_i, \qquad i = 1, \ldots, n, \tag{7}$$

where $\mathbf{x}_i$ and $\mathbf{z}_i$ are explanatory variables, ε_i is an error term satisfying $\mathbb{E}[\varepsilon_i|\mathbf{x}_i, \mathbf{z}_i] = 0$, the function $g(\mathbf{z})$ is unknown, and sampling is i.i.d. across i is assumed. Suppose $\{\mathbf{p}^k(\mathbf{z}) : k = 1, 2, \cdots\}$ are functions having the property that linear combinations can approximate square-integrable functions of $\mathbf{z}$ well, in which case $g(\mathbf{z}_i) \approx \boldsymbol{\gamma}_n'\mathbf{p}_n(\mathbf{z}_i)$ for some $\boldsymbol{\gamma}_n$, where $\mathbf{p}_n(\mathbf{z}) = (\mathbf{p}^1(\mathbf{z}), \ldots, \mathbf{p}^{K_n}(\mathbf{z}))'$. Defining $y_{i,n} = y_i$, $\mathbf{x}_{i,n} = \mathbf{x}_i$, $\mathbf{w}_{i,n} = \mathbf{p}_n(\mathbf{z}_i)$, and $u_{i,n} = \varepsilon_i + g(\mathbf{z}_i) - \boldsymbol{\gamma}_n'\mathbf{w}_{i,n}$, the model (7) is of the form (1), and $\hat{\boldsymbol{\beta}}_n$ is the series estimator of $\boldsymbol{\beta}$; see, e.g., Donald and Newey (1994) and Cattaneo, Jansson, and Newey (2017) and references therein.

Constructing the basis $\mathbf{p}_n(\mathbf{z}_i)$ in applications may require using a large K_n, either because the underlying functions are not smooth enough or because $\dim(\mathbf{z}_i)$ is large. For example, if a $\mathfrak{p} = 3$ cubic polynomial expansion is used, also known as a power series of order 3, then $\dim(\mathbf{w}_i) = (\mathfrak{p} + \dim(\mathbf{z}_i))!/(\mathfrak{p}!\dim(\mathbf{z}_i)!) = 286$ when $\dim(\mathbf{z}_i) = 10$, and therefore flexible estimation and inference using the semi-linear model (7) with a sample size of $n = 1,000$ gives $K_n/n = 0.286$. For further technical details on series-based methods see, e.g., Newey (1997), Chen (2007), Cattaneo and Farrell (2013), and Belloni, Chernozhukov, Chetverikov, and Kato (2015), and references therein. For another example, when the basis functions $\mathbf{p}_n(\mathbf{z})$ are constructed using partitioning estimators, the OLS estimator of $\boldsymbol{\beta}$ becomes a subclassification estimator, a method that has been proposed in the literature on program evaluation and treatment effects; see, e.g., Cochran (1968), Rosenbaum and Rubin (1983), Cattaneo and Farrell (2011), and references therein. When a Partitioning estimator of order 0 is used, the semi-linear model becomes a one-way fixed effects linear regression model, where each dummy variable corresponds to one (disjoint) partition on the support of $\mathbf{z}_i$; in this case, K_n is to the number of partitions or fixed effects included in the estimation.

Our primitive regularity conditions for this example include

$$\varrho_n = \min_{\boldsymbol{\gamma}\in\mathbb{R}^{K_n}} \mathbb{E}[|g(\mathbf{z}_i) - \boldsymbol{\gamma}'\mathbf{p}_n(\mathbf{z}_i)|^2] = o(1), \qquad \chi_n = \min_{\boldsymbol{\delta}\in\mathbb{R}^{K_n\times d}} \mathbb{E}[\|\mathbb{E}[\mathbf{x}_i|\mathbf{z}_i] - \boldsymbol{\delta}'\mathbf{p}_n(\mathbf{z}_i)\|^2] = O(1),$$

$n\varrho_n\chi_n = o(1)$, and the negligibility condition $\max_{1\leq i\leq n}\|\hat{\mathbf{v}}_{i,n}\|/\sqrt{n} = o_p(1)$. A key finding implied by these regularity conditions is that we only require minimal smoothness conditions on $g(\mathbf{z}_i)$ and $\mathbb{E}[\mathbf{x}_i|\mathbf{z}_i]$. The negligibility condition is automatically satisfied if $\chi_n = o(1)$, as discussed above, but in fact our results do not require any approximation of $\mathbb{E}[\mathbf{x}_i|\mathbf{z}_i]$, as usually assumed in the literature, provided a "locally supported" basis is used; i.e., any basis $\mathbf{p}_n(\mathbf{z})$ that generates an approximately band projection matrix $\mathbf{M}_n$; examples of such basis include partitioning and spline estimators. See Section 4.3 in the supplemental appendix for further discussion and technical details.

4 Results

This section presents our main theoretical results for inference in linear regression models with many covariates and heteroskedasticity. Mathematical proofs, and other technical results that may be of independent interest, are given in the supplemental appendix.

4.1 Asymptotic Normality

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

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

$$\boldsymbol{\Omega}_n^{-1/2}\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \to_d \mathcal{N}(\mathbf{0}, \mathbf{I}), \qquad \boldsymbol{\Omega}_n = \hat{\boldsymbol{\Gamma}}_n^{-1}\boldsymbol{\Sigma}_n\hat{\boldsymbol{\Gamma}}_n^{-1}, \tag{8}$$

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

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 furthermore exploits the fact that the parameter of interest is given by the first d coordinates of the vector $(\boldsymbol{\beta}', \boldsymbol{\gamma}_n')'$ (i.e., in Mammen (1993) notation, it considers the case $\mathbf{c} = (\boldsymbol{\iota}', \mathbf{0}')'$ with ι denoting a d-dimensional vector of ones and $\mathbf{0}$ denoting a K_n-dimensional vector of zeros).

In isolation, the fact that Theorem 1 removes the requirement $K_n/n \to 0$ may seem like little more than a subtle technical improvement over results currently available. It should be recognized, however, that conducting inference turn out to be considerably harder when $K_n/n \not\to 0$. The latter is an important insight about large-dimensional models that cannot be deduced from results obtained under the assumption $K_n/n \to 0$, but can be obtained with the help of Theorem 1. In addition, it is worth mentioning that Theorem 1 is a substantial improvement over Cattaneo, Jansson, and Newey (2017, Theorem 1) because here it is not required that $K_n \to \infty$ nor $\chi_n = o(1)$ —a different method of proof is also used. This improvement applies not only to the partially linear model example, but more generally to linear models with many covariates, because Theorem 1 applies to quite general form of nuisance covariate $\mathbf{w}_{i,n}$ beyond specific approximating basis functions. In the specific case of the partially linear model, this implies that we are able to weaken smoothness assumptions (or the curse of dimensionality), otherwise required to satisfy the condition $\chi_n = o(1)$.

Remark 2. Theorem 1 concerns only distributional properties of $\hat{\boldsymbol{\beta}}_n$. First, this theorem implies $\sqrt{n}$-consistency of $\hat{\boldsymbol{\beta}}_n$ because $\boldsymbol{\Omega}_n^{-1} = O_p(1)$ (see Lemmas SA-1 and SA-2 of the supplemental appendix). Second, this theorem does require nor imply consistency of the (implicit) least squares estimate of $\boldsymbol{\gamma}_n$, as in fact such a result will not be true in most applications with many nuisance covariates $\mathbf{w}_{n,i}$. For example, in a partially linear model (7) the approximating coefficients $\boldsymbol{\gamma}_n$ will not be consistently estimated unless $K_n/n \to 0$, or in a one-way fixed effect panel data model (6) the unit-specific coefficients in $\boldsymbol{\gamma}_n$ will not be consistently estimated unless $K_n/n = 1/T \to 0$. Nevertheless, Theorem 1 shows that $\hat{\boldsymbol{\beta}}_n$ can still be root-n asymptotically normal under fairly general conditions; this result is due to the intrinsic linearity and additive separability of the model (1).

4.2 Variance Estimation

Achieving (2), the counterpart of (8) in which the unknown matrix $\boldsymbol{\Sigma}_n$ is replaced by the estimator $\hat{\boldsymbol{\Sigma}}_n$, 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 $\hat{\boldsymbol{\Sigma}}_n = \hat{\boldsymbol{\Sigma}}_n^{\texttt{HO}}$.*

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 $\hat{\boldsymbol{\Sigma}}_n^{\texttt{HO}}$ 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 $\hat{\mathbf{\Sigma}}_n^{\mathtt{EW}}$ under the assumption $\mathcal{M}_n \to_p 0$, and also derives an asymptotic representation for estimators of the form $\hat{\mathbf{\Sigma}}_n(\boldsymbol{\kappa}_n)$ without imposing this assumption, which is useful to study the asymptotic properties of other members of the HCk class of standard error estimators.

Theorem 3 *Suppose the assumptions of Theorem 1 hold.*

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

(b) If $\|\boldsymbol{\kappa}_n\|_\infty = \max_{1\le i\le n} \sum_{j=1}^n |\kappa_{ij,n}| = O_p(1)$, then

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

The conclusion of part (a) typically fails when the condition $K_n/n \to 0$ is dropped. For example, when specialized to $\boldsymbol{\kappa}_n = \mathbf{I}_n$ part (b) implies that in the homoskedastic case (i.e., when the assumptions of Theorem 2 are satisfied) the standard estimator $\hat{\mathbf{\Sigma}}_n^{\mathtt{EW}}$ is asymptotically downward biased in general (unless $K_n/n \to 0$). In the following section we make this result precise and discuss similar results for other popular variants of the HCk standard error estimators mentioned above.

On the other hand, because $\sum_{1\le k\le n} \kappa_{ik,n}^{\mathtt{HC}} M_{kj,n}^2 = \mathbb{1}(i=j)$ by construction, part (b) implies that $\hat{\mathbf{\Sigma}}_n^{\mathtt{HC}}$ is consistent provided $\|\boldsymbol{\kappa}_n^{\mathtt{HC}}\|_\infty = O_p(1)$. A simple condition for this to occur can be stated in terms of $\mathcal{M}_n$. Indeed, if $\mathcal{M}_n < 1/2$, then $\boldsymbol{\kappa}_n^{\mathtt{HC}}$ is diagonally dominant and it follows from Theorem 1 of Varah (1975) that

$$\|\boldsymbol{\kappa}_n^{\mathtt{HC}}\|_\infty \le \frac{1}{1/2 - \mathcal{M}_n}.$$

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* $\hat{\boldsymbol{\Sigma}}_n = \hat{\boldsymbol{\Sigma}}_n^{\mathtt{HC}}$.

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 (6)), 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 $\hat{\Sigma}_n^{\mathtt{HC}}$ to perform well.

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. Following the recommendation of a reviewer, we explored the numerical performance of the standard nonparametric bootstrap in our simulation study, where we found that indeed bootstrap validity seems to fail in the high-dimensional settings we considered.

4.3 HCk Standard Errors with Many Covariates

The HCk variance estimators are very popular in empirical work, and in our context are of the form $\hat{\boldsymbol{\Sigma}}_n(\boldsymbol{\kappa}_n)$ with $\kappa_{ij,n} = \mathbb{1}(i=j)\Upsilon_{i,n} M_{ii,n}^{-\xi_{i,n}}$ for some choice of $(\Upsilon_{i,n}, \xi_{i,n})$. See Long and Ervin (2000) and MacKinnon (2012) for reviews. Theorem 3(b) can be used to formulate conditions, including $K_n/n \to 0$, under which these estimators are consistent in the sense that

$$\hat{\mathbf{\Sigma}}_n(\boldsymbol{\kappa}_n) = \mathbf{\Sigma}_n + o_p(1), \qquad \mathbf{\Sigma}_n = \frac{1}{n}\sum_{i=1}^{n} \hat{\mathbf{v}}_{i,n}\hat{\mathbf{v}}_{i,n}' \mathbb{E}[U_{i,n}^2|\mathcal{X}_n, \mathcal{W}_n].$$

More generally, Theorem 3(b) shows that, if $\kappa_{ij,n} = \mathbb{1}(i=j)\Upsilon_{i,n}M_{ii,n}^{-\xi_{i,n}}$, then

$$\hat{\mathbf{\Sigma}}_n(\boldsymbol{\kappa}_n) = \bar{\mathbf{\Sigma}}_n(\boldsymbol{\kappa}_n) + o_p(1), \qquad \bar{\mathbf{\Sigma}}_n(\boldsymbol{\kappa}_n) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} \Upsilon_{i,n}M_{ii,n}^{-\xi_{i,n}}M_{ij,n}^2 \hat{\mathbf{v}}_{i,n}\hat{\mathbf{v}}_{i,n}' \mathbb{E}[U_{j,n}^2|\mathcal{X}_n, \mathcal{W}_n].$$

We therefore obtain the following (mostly negative) results about the properties of HCk estimators when $K_n/n \nrightarrow 0$, that is, when potentially many covariates are included.

HC0: $(\Upsilon_{i,n}, \xi_{i,n}) = (1, 0)$. If $\mathbb{E}[U_{j,n}^2|\mathcal{X}_n, \mathcal{W}_n] = \sigma_n^2$, then

$$\bar{\mathbf{\Sigma}}_n(\boldsymbol{\kappa}_n) = \mathbf{\Sigma}_n - \frac{\sigma_n^2}{n}\sum_{i=1}^{n}(1 - M_{ii,n})\hat{\mathbf{v}}_{i,n}\hat{\mathbf{v}}_{i,n}' \leq \mathbf{\Sigma}_n,$$

with $n^{-1}\sum_{i=1}^{n}(1-M_{ii,n})\hat{\mathbf{v}}_{i,n}\hat{\mathbf{v}}_{i,n}' \neq o_p(1)$ in general (unless $K_n/n \to 0$). Thus, $\hat{\mathbf{\Sigma}}_n(\boldsymbol{\kappa}_n) = \hat{\mathbf{\Sigma}}_n^{\mathtt{EW}}$ is inconsistent in general. In particular, inference based on $\hat{\mathbf{\Sigma}}_n^{\mathtt{EW}}$ is asymptotically liberal (even) under homoskedasticity.

HC1: $(\Upsilon_{i,n}, \xi_{i,n}) = (n/(n-K_n), 0)$. If $\mathbb{E}[U_{j,n}^2|\mathcal{X}_n, \mathcal{W}_n] = \sigma_n^2$ and if $M_{11,n} = \ldots = M_{nn,n}$, then $\bar{\mathbf{\Sigma}}_n(\boldsymbol{\kappa}_n) = \mathbf{\Sigma}_n$, but in general this estimator is inconsistent when $K_n/n \nrightarrow 0$ (and so is any other scalar multiple of $\hat{\mathbf{\Sigma}}_n^{\mathtt{EW}}$).

HC2: $(\Upsilon_{i,n}, \xi_{i,n}) = (1, 1)$. If $\mathbb{E}[U_{j,n}^2|\mathcal{X}_n, \mathcal{W}_n] = \sigma_n^2$, then $\bar{\mathbf{\Sigma}}_n(\boldsymbol{\kappa}_n) = \mathbf{\Sigma}_n$, but in general this estimator is inconsistent under heteroskedasticity when $K_n/n \nrightarrow 0$. For instance, if $d = 1$ and if $\mathbb{E}[U_{j,n}^2|\mathcal{X}_n, \mathcal{W}_n] = \hat{v}_{j,n}^2$, then

$$\bar{\mathbf{\Sigma}}_n(\boldsymbol{\kappa}_n) - \mathbf{\Sigma}_n = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}[\frac{M_{ij,n}^2}{2}(M_{ii,n}^{-1} + M_{jj,n}^{-1}) - \mathbb{1}(i=j)]\hat{v}_{i,n}^2\hat{v}_{j,n}^2 \neq o_p(1)$$

in general (unless $K_n/n \to 0$).

HC3: $(\Upsilon_{i,n}, \xi_{i,n}) = (1, 2)$. Inference based on this estimator is asymptotically conservative because

$$\bar{\boldsymbol{\Sigma}}_n(\boldsymbol{\kappa}_n) - \boldsymbol{\Sigma}_n = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1, j\neq i}^{n} M_{ii,n}^{-2} M_{ij,n}^2 \hat{\mathbf{v}}_{i,n}\hat{\mathbf{v}}_{i,n}' \mathbb{E}[U_{j,n}^2 | \mathcal{X}_n, \mathcal{W}_n] \geq 0,$$

where $n^{-1}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n} M_{ii,n}^{-2} M_{ij,n}^2 \hat{\mathbf{v}}_{i,n}\hat{\mathbf{v}}_{i,n}' \mathbb{E}[U_{j,n}^2 | \mathcal{X}_n, \mathcal{W}_n] \neq o_p(1)$ in general (unless $K_n/n \to 0$).

HC4: $(\Upsilon_{i,n}, \xi_{i,n}) = (1, \min(4, nM_{ii,n}/K_n))$. If $M_{11,n} = \ldots = M_{nn,n} = 2/3$ (as occurs when $T = 3$ in the fixed effects panel data model), then HC4 reduces to HC3, so this estimator is also inconsistent in general.

Among other things these results show that (asymptotically) conservative inference in linear models with many covariates (i.e., even when $K/n \not\to 0$) can be conducted using standard linear methods (and software), provided the HC3 standard errors are used.

In the numerical work reported in the following sections and the supplemental appendix, we present evidence comparing all these standard error estimators. In particular, we find that indeed standard OLS-based confidence intervals employing HC3 standard errors are always quite conservative. Furthermore, we also find that our proposed variance estimator $\hat{\boldsymbol{\Sigma}}_n^{\mathtt{HC}}$ delivers confidence intervals with close-to-correct empirical coverage.

5 Simulations

We conducted a simulation study to assess the finite sample properties of our proposed inference methods as well as those of other standard inference methods available in the literature. Based on the generic linear regression model (1), we consider 15 distinct data generating processes (DGPs) motivated by the three examples discussed above. To conserve space, here we only discuss results from Model 1, a representative case, but the supplemental appendix contains the full set of results and further details (see Table 1 in the supplement for a synopsis of the DGPs used).

We discuss results for a linear model (1) with i.i.d. data, $n = 700$, $d = 1$ and $x_{i,n} \sim \mathsf{Normal}(0, 1)$, $\mathbf{w}_{i,n} = \mathbb{1}(\mathbf{v}_{i,n} \geq 2.5)$ with $\mathbf{v}_{i,n} \sim \mathsf{Normal}(\mathbf{0}, \mathbf{I}_{K_n})$, and $u_{i,n} \sim \mathsf{Normal}(0, 1)$, all independent of each other. Thus, this design considers (possibly overlapping) sparse dummy variables entering $\mathbf{w}_{i,n}$; each column assigns a value of 1 to approximately five units out of $n = 700$. We set $\beta = 1$ and $\boldsymbol{\gamma}_n = \mathbf{0}$, and considered five different model dimensions: $\dim(\mathbf{w}_{i,n}) = K_n \in \{1, 71, 141, 211, 281\}$. In the supplemental appendix we also present results for more sparse dummy variables in the context of one-way and two-way linear panel data regression models, and for non-binary covariates $\mathbf{w}_{i,n}$ in both increasing dimension linear regression settings and semiparametric partially linear regression settings (where $\boldsymbol{\gamma}_n \neq \mathbf{0}$ and $\mathbf{w}_{i,n}$ is constructed using power series expansions). Furthermore, we also consider an asymmetric and a bimodal distribution for the unobservable error terms. In all cases the numerical results are qualitatively similar to those discussed herein. For each DGP, we investigate both homoskedastic as well as (conditional on $x_{i,n}$ and/or $\mathbf{w}_{i,n}$) heteroskedastic models, following closely the specifications in Stock and Watson (2008) and MacKinnon (2012). In particular, our heteroskedastic model takes the form: $\mathbb{V}[u_{i,n}|x_{i,n}, \mathbf{w}_{i,n}] = \varkappa_u(1 + (t(x_{i,n}) + \boldsymbol{\iota}'\mathbf{w}_{i,n})^2)$ and $\mathbb{V}[x_{i,n}|\mathbf{w}_{i,n}] = \varkappa_v(1 + (\boldsymbol{\iota}'\mathbf{w}_{i,n})^2)$, where the constants $\varkappa_u$ and $\varkappa_v$ are chosen so that $\mathbb{V}[u_{i,n}] = \mathbb{V}[x_{i,n}] = 1$, and $t(a) = a\mathbb{1}(-2 \leq a \leq 2) + 2\,\mathrm{sgn}(a)(1 - \mathbb{1}(-2 \leq a \leq 2))$.

We conducted $S = 5,000$ simulations to study the finite sample performance of 16 confidence intervals: eight based on a Gaussian approximation and eight based on a bootstrap approximation. Our paper offers theory for Gaussian-based inference methods, but we also included bootstrap-based inference methods for completeness (as discussed in Remark 3, the bootstrap is invalid when $K_n \propto n$ in linear regression models). For each inference method, we report both average coverage frequency and interval length of 95% nominal confidence intervals; the latter provides a summary of efficiency/power for each inference method. To be more specific, for $\alpha = 0.05$, the confidence intervals take the form:

$$\mathsf{I}_\ell = \left[\,\hat{\beta}_n - q^{-1}_{\ell,1-\alpha/2}\cdot\sqrt{\frac{\hat{\Omega}_{n,\ell}}{n}}\;,\;\hat{\beta}_n - q^{-1}_{\ell,\alpha/2}\cdot\sqrt{\frac{\hat{\Omega}_{n,\ell}}{n}}\,\right], \qquad \hat{\Omega}_{n,\ell} = \hat{\Gamma}_n^{-1}\hat{\Sigma}_{n,\ell}\hat{\Gamma}_n^{-1},$$

where $q_{\ell,a}^{-1} = q_\ell^{-1}(a)$ and $q_\ell(a)$ denotes a cumulative distribution function, and $\hat{\Sigma}_{n,\ell}$ with $\ell \in \{\text{HO0, HO1, HC0, HC1, HC2, HC3, HC4, HC}K\}$ corresponds the variance estimators discussed in Sections 2 and 4.3. Gaussian-based methods set $q(a)$ equal to the standard normal distribution for all ℓ, while bootstrap-based methods are based on the nonparametric bootstrap distributional approximation to the distribution of the t-test $\mathsf{T}_\ell = (\hat{\beta}_n - \beta)/\sqrt{\hat{\Omega}_{n,\ell}/n}$. The empirical coverage of these 16 confidence intervals are reported in Panel (a) of Table 1. In addition, Panel (b) of Table 1 reports the average interval length of each confidence intervals, which is computed as $\mathsf{L}_\ell = (q_{\ell,1-\alpha/2}^{-1} - q_{\ell,\alpha/2}^{-1}) \cdot \sqrt{\hat{\Omega}_{n,\ell}/n}$, which offers a summary of finite sample power/efficiency of each inference method.

The main findings from the simulation study are in line with our theoretical results. To be precise, we find that the confidence interval estimators constructed using our proposed standard errors formula $\hat{\Sigma}_n^{\mathtt{HC}}$, denoted HC$K$, offer close-to-correct empirical coverage. 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 find that inference based on HC3 standard errors is conservative, a general asymptotic result that is formally established in this paper. Bootstrap-based methods seem to perform better than their Gaussian-based counterparts, but they never outperform our proposed Gaussian-based inference procedure nor do they provide close-to-correct empirical coverage across all cases. Finally, our proposed confidence intervals also exhibit very good average interval length.

6 Empirical Illustration

We illustrate the different linear regression inference methods discussed in this paper using a real data set to study the effect of ability on earnings. In particular, we employ the dataset constructed by Carneiro, Heckman, and Vytlacil (2011, CHV, hereafter). [The dataset is available at `https://www.aeaweb.org/articles?id=10.1257/aer.101.6.2754`.]. The data comes from the 1979 National Longitudinal Survey of Youth (NLSY79), which surveys individuals born in 1957–1964 and includes basic demographic, economic and educational information for each individual. It also includes a well-known proxy for ability (beyond schooling and work experience): the Armed Forces Qualification Test (AFQT), which gives a measure usually understood as a proxy for their intrinsic ability for the respondent. This data has been used repeatedly to either control for or estimate the effects of ability in empirical studies in economics and other disciplines. See CHV for further details and references.

The sample is composed of white males of ages between 28 and 34 years of old in 1991, at most 5 siblings, and with at least incomplete secondary education. We split the sample into individuals with high school dropouts and high school graduates, and individuals with some college, college graduates, and postgraduates. For each subsample, we consider the linear regression model (1) with $y_{i,n} = \log(\texttt{wages}_i)$, where $\texttt{wages}_i$ is the log wage in 1991 of unit i, $x_{i,n} = \texttt{afqt}_i$ denotes the (adjusted) standardized AFQT score for unit i, and $\mathbf{w}_{i,n}$ collects several survey, geographic and dummy variables for unit i. In particular, $\mathbf{w}_{i,n}$ includes the 14 covariates described in CHV (Table 2, p. 2763), a dummy variable for wether the education level was completed, eight cohort fixed effects, county fixed effects, and cohort-county fixed effects. For our illustration, we further restrict the sample to units in counties with at least 3 survey respondents, giving a total of $K_n = 122$ and $n = 436$ ($K_n/n = 0.280$, $\mathcal{M}_n = 0.422$) for high school educated units and $K_n = 123$ and $n = 452$ ($K_n/n = 0.272$, $\mathcal{M}_n = 0.411$) college educated units.

The empirical findings are reported in Table 2. For high school educated individuals, we find an estimated returns to ability of $\hat{\beta} = 0.060$. The statistical significance of this effect, however, depends on the inference method employed. If homoskedastic consistent standard errors are used, then the effect is statistical significant (p-values are 0.010 and 0.029 for unadjusted and degrees-of-freedom adjusted standard errors, respectively). If heteroskedasticity consistent standard errors are used, the default method in most empirical studies, then the statistical significance depends on the which inference method is used; see Section 4.3. In particular, HC0 also gives a statistically significant result (p-value is 0.020), while HC1 and HC2 deliver marginal significance (both p-values are 0.048). On the other hand, HC3 and

HC4 give p-values of 0.092 and 0.122, respectively, and hence suggest that the point estimate is not statistically distinguishable from zero. Finally, our proposed standard error, HCK, gives a p-value of 0.058, also making $\hat{\beta} = 0.060$ statistically insignificant at the conventional 5-pecent level. In contrast, for college educated individuals, we find an effect of $\hat{\beta} = 0.091$, and all inference methods indicate that this estimated returns to ability is statistically significant at conventional levels. In particular, HC3 and our proposed standard errors HCK give p-values of 0.037 and 0.017, respectively.

This illustrative empirical application showcases the role of our proposed inference method for empirical work employing linear regression with possibly many covariates; in this application, K_n large relative to n ($K_n/n \approx 0.3$) is quite natural due to the presence of many county and cohort fixed effects. Specifically, when studying the effect of ability on earnings for high school educated individuals, the statistical significance of the results crucially depend on the inference methods used: as predicted by our theoretical findings, inference methods that are not robust to the inclusion of many covariates tend to deliver statistically significant results, while methods that are robust (HC3 is asymptotically conservative and HCK is asymptotically correct) do not deliver statistically significant results, giving an example where the empirical conclusion may change depending on whether the presence of many covariates is taken into account when conducting inference. In contrast, the empirical findings for college educated individuals appear to be statistically significant and robust across all inference methods.

7 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 (2014), Farrell (2015), Belloni, Chernozhukov, Hansen, and Fernandez-Val (2017), 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). An implication of the results obtained in this paper is that the latter assumption cannot be dropped if post covariate selection inference is based on conventional standard errors. It would therefore 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

Abadie, A., G. W. Imbens, and F. Zheng (2014): "Inference for Misspecified Models With Fixed Regressors," *Journal of the American Statistical Association*, 109(508), 1601–1614.

Angrist, J., and J. Hahn (2004): "When to Control for Covariates? Panel Asymptotics for Estimates of Treatment Effects," *Review of Economics and Statistics*, 86(1), 58–72.

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 (2014): "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 (2017): "Program Evaluation and Causal Inference with High-Dimensional Data," *Econometrica*, forthcoming.

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.

Carneiro, P., J. J. Heckman, and E. J. Vytlacil (2011): "Estimating Marginal Returns to Education," *American Economic Review*, 101(6), 2754–2781.

Cattaneo, M. D., and M. H. Farrell (2011): "Efficient Estimation of the Dose-Response Function under Ignorability using Subclassification on the Covariates," in *Missing-Data Methods: Cross-sectional Methods and Applications (Advances in Econometrics, vol. 27)*, ed. by D. Drukker, pp. 93–127. Emerald Group Publishing.

——— (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 (2017): "Alternative Asymptotics and the Partially Linear Model with Many Regressors," *Econometric Theory*, forthcoming.

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, pp. 5549–5632. 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.

Cochran, W. G. (1968): "The Effectiveness of Adjustment by Subclassification in Removing Bias in Observational Studies," *Biometrics*, 24(2), 295–313.

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*, 189(1), 1–23.

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.

Kauermann, G., and R. J. Carroll (2001): "A Note on the Efficiency of Sandwich Covariance Matrix Estimation," *Journal of the American Statistical Association*, 96(456), 1387–1396.

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.

Li, C., and U. K. Müller (2017): "Linear Regression with Many Controls of Limited Explanatory Power," working paper, Princeton University.

Long, J. S., and L. H. Ervin (2000): "Using Heteroscedasticity Consistent Standard Errors in the Linear Regression Model," *The American Statistician*, 54(3), 217–224.

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.

Rosenbaum, P. R., and D. B. Rubin (1983): "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70(1), 41–55.

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.

Verdier, V. (2017): "Estimation and Inference for Linear Models with Two-Way Fixed Effects and Sparsely Matched Data," Working paper, UNC.

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.

Zheng, S., D. Jiang, Z. Bai, and X. He (2014): "Inference on multiple correlation coefficients with moderately high dimensional data," *Biometrika*, 101(3), 748–754.

Table 1: Simulation Results (Model 1 in Supplemental Appendix).

(a) Empirical Coverage

	Gaussian Distributional Approximation								Bootstrap Distributional Approximation							
	HO0	HO1	HC0	HC1	HC2	HC3	HC4	HCK	HO0	HO1	HC0	HC1	HC2	HC3	HC4	HCK
Homoskedastic Model																
$K/n = 0.001$	0.949	0.950	0.948	0.948	0.948	0.948	0.948	0.948	0.946	0.946	0.943	0.943	0.943	0.943	0.943	0.943
$K/n = 0.101$	0.939	0.956	0.939	0.952	0.952	0.962	0.980	0.951	0.951	0.951	0.947	0.947	0.948	0.949	0.947	0.942
$K/n = 0.201$	0.916	0.947	0.919	0.947	0.946	0.968	0.989	0.945	0.965	0.965	0.950	0.950	0.949	0.946	0.944	0.939
$K/n = 0.301$	0.900	0.950	0.904	0.954	0.951	0.977	0.983	0.949	0.980	0.980	0.961	0.961	0.949	0.931	0.948	0.933
$K/n = 0.401$	0.881	0.954	0.884	0.955	0.952	0.989	0.972	0.949	0.989	0.989	0.976	0.976	0.956	0.928	0.967	0.944
Heteroskedastic Model																
$K/n = 0.001$	0.880	0.880	0.945	0.945	0.945	0.945	0.946	0.945	0.939	0.939	0.937	0.937	0.937	0.937	0.937	0.937
$K/n = 0.101$	0.725	0.750	0.885	0.904	0.926	0.957	0.989	0.948	0.897	0.897	0.916	0.906	0.907	0.909	0.902	0.919
$K/n = 0.201$	0.762	0.804	0.853	0.901	0.924	0.973	0.995	0.945	0.919	0.919	0.919	0.909	0.908	0.907	0.908	0.920
$K/n = 0.301$	0.784	0.856	0.837	0.903	0.926	0.981	0.977	0.947	0.944	0.944	0.936	0.926	0.919	0.903	0.920	0.920
$K/n = 0.401$	0.758	0.875	0.792	0.908	0.929	0.990	0.950	0.948	0.975	0.975	0.962	0.962	0.936	0.900	0.953	0.926

(b) Interval Length

	Gaussian Distributional Approximation								Bootstrap Distributional Approximation							
	HO0	HO1	HC0	HC1	HC2	HC3	HC4	HCK	HO0	HO1	HC0	HC1	HC2	HC3	HC4	HCK
Homoskedastic Model																
$K/n = 0.001$	0.148	0.148	0.148	0.148	0.148	0.148	0.148	0.148	0.148	0.148	0.149	0.149	0.149	0.149	0.149	0.149
$K/n = 0.101$	0.148	0.156	0.148	0.157	0.156	0.165	0.186	0.156	0.161	0.161	0.158	0.158	0.158	0.158	0.158	0.157
$K/n = 0.201$	0.148	0.166	0.149	0.167	0.165	0.185	0.225	0.165	0.180	0.180	0.170	0.170	0.169	0.167	0.166	0.164
$K/n = 0.301$	0.148	0.177	0.150	0.179	0.177	0.212	0.219	0.177	0.210	0.210	0.189	0.189	0.182	0.172	0.180	0.174
$K/n = 0.401$	0.148	0.192	0.150	0.194	0.191	0.247	0.213	0.190	0.260	0.260	0.223	0.223	0.200	0.174	0.212	0.189
Heteroskedastic Model																
$K/n = 0.001$	0.148	0.148	0.186	0.186	0.186	0.186	0.187	0.186	0.186	0.186	0.188	0.188	0.188	0.188	0.188	0.188
$K/n = 0.101$	0.148	0.156	0.213	0.225	0.241	0.273	0.357	0.254	0.243	0.243	0.264	0.264	0.266	0.268	0.273	0.269
$K/n = 0.201$	0.148	0.166	0.187	0.209	0.226	0.276	0.353	0.244	0.243	0.243	0.252	0.252	0.251	0.248	0.251	0.249
$K/n = 0.301$	0.148	0.177	0.170	0.203	0.219	0.287	0.278	0.240	0.259	0.259	0.254	0.254	0.244	0.232	0.247	0.239
$K/n = 0.401$	0.148	0.191	0.159	0.206	0.220	0.310	0.239	0.241	0.300	0.300	0.276	0.276	0.248	0.218	0.269	0.243

Notes: (i) DGP is Model 1 from the supplemental appendix, sample size is $n = 700$, number of bootstrap replications is $B = 500$, and number of simulation replications is $S = 5,000$; (ii) Columns HO0 and HO1 correspond to confidence intervals using homoskedasticity consistent standard errors without and with degrees of freedom correction, respectively, columns HC0–HC4 correspond to confidence intervals using the heteroskedasticity consistent standard errors discussed in Sections 2 and 4.3, and columns HCK correspond to confidence intervals using our proposed standard errors estimator.

Table 2: Empirical Application (Returns to Ability, AFQT Score).

(a) Secondary Education

	Outcome:	log(wages)
$\hat{\beta}$	0.060	
	Std.Err.	**p-value**
HO0	0.023	0.010
HO1	0.028	0.029
HC0	0.026	0.020
HC1	0.030	0.048
HC2	0.030	0.048
HC3	0.036	0.092
HC4	0.039	0.122
HCK	0.032	0.058
K_n	122	
n	436	
K_n/n	0.280	
$\mathcal{M}_n$	0.422	

(b) College Education

	Outcome:	log(wages)
$\hat{\beta}$	0.091	
	Std.Err.	**p-value**
HO0	0.032	0.005
HO1	0.038	0.016
HC0	0.033	0.006
HC1	0.039	0.018
HC2	0.038	0.016
HC3	0.044	0.037
HC4	0.048	0.058
HCK	0.038	0.017
K_n	123	
n	452	
K_n/n	0.272	
$\mathcal{M}_n$	0.411	