Linear Regression with Many Controls of Limited Explanatory Power*

Chenchuan (Mark) Li and Ulrich K. Müller
Princeton University
Department of Economics
Princeton, NJ, 08544

First version: September 2015
This version: December 2016

Abstract

We consider inference about a scalar coefficient in a linear regression model. One previously considered approach to dealing with many controls imposes sparsity, that is it assumed known that nearly all control coefficients are zero. We instead impose a bound on a weighted sum of squared control coefficients, which is interpretable as a bound on the sample variation in the dependent variable induced by the controls. We develop a simple testing procedure that exploits this additional information in general heteroskedastic models. We also show that under asymptotics where the number of controls is a non-negligible fraction of the number of observations, and the bound is not too large, our suggested test comes close to being weighted average power maximizing in the Gaussian homoskedastic model. We compare our procedure to a sparsity-based approach in a Monte Carlo study and by revisiting the empirical relationship between crime and abortion.

Keywords: high dimensional linear regression, limit of experiments, L2 bound, invariance to linear reparameterizations

*We thank participants at various workshops for useful comments and advice. Müller gratefully acknowledges financial support from the National Science Foundation through grant SES-1627660.

1 Introduction

A classic issue that arises frequently in applied econometrics is how to deal with a potentially large number of control variables in a linear regression. As is well understood, excluding unnecessary controls leads to more precise estimation of the coefficient of interest, and thus to tighter confidence intervals and more powerful hypothesis tests. But excluding controls that have non-zero coefficients introduces omitted variable bias, which leads to oversized tests and confidence intervals with less than nominal coverage.

One might try to use a pre-test to identify which controls have non-zero coefficients, such as testing down procedures, or information criteria, and then proceed with standard inference using only the selected controls. As stressed by Leeb and Pötscher (2005) (also see Leeb and Pötscher (2008a, 2008b) and the references therein), however, this does not yield uniformly valid inference: in general, for any sample size, there exist sufficiently small values for the control coefficients so that they are not selected with probability one, yet they are still large enough to induce an omitted variable bias that implies oversized tests and confidence intervals.

Alternatively, there is a burgeoning statistical literature that assumes sparsity (Tibshirani (1996), Fan and Li (2001), etc.): most of the control coefficients are known to be zero, but it is not known which ones. A standard LASSO implementation does not lead to valid inference about the coefficient of interest. But by combining a sparsity assumption on the control coefficients with a sparsity assumption on the correlations between the regressor of interest and the control variables, recent work by Belloni, Chernozhukov, and Hansen (2014) shows how a novel LASSO based "double selection procedure" does yield uniformly valid inference.

While this work is important progress, a sparsity assumption might not always be a compelling starting point: in social science applications, it is usually not obvious why the large majority of control coefficients should be exactly zero. In addition, the sparsity restriction does not remain invariant to linear reparameterizations of the controls. For instance, in the context of technical controls that are functions of an underlying continuous variable, sparsity drives a distinction between specifying the controls as powers or Chebyshev polynomials.

This paper develops an alternative approach that exploits an *a priori* upper bound on the weighted sum of squared control coefficients, rather than on the number of non-zero control coefficients. To be more precise, consider testing the null hypothesis $H_0 : \beta = 0$ about the scalar parameter β from observing $\{y_i, x_i, z_{i,1}, \ldots, z_{i,k}\}_{i=1}^n$, where

$$y_i = x_i\beta + \sum_{j=1}^{k} \gamma_j z_{i,j} + \varepsilon_i, \tag{1}$$

the $k < n$ control coefficients γ_j, $j = 1, \ldots, k$ are scalar nuisance parameters, and ε_i is a conditionally mean zero error term. Let $\hat{\beta}^{\text{short}}$ and $\hat{\beta}^{\text{long}}$ be the coefficients on x_i from a linear regression of y_i on x_i, and from a linear regression of y_i on $(x_i, z_{i,1}, \ldots, z_{i,k})$, respectively. We combine the information in $(\hat{\beta}^{\text{short}}, \hat{\beta}^{\text{long}})$ and the bound

$$\kappa^2 = \sum_{i=1}^{n} \left(\sum_{j=1}^{k} \gamma_j z_{i,j} \right)^2 \leq \bar{\kappa}^2 \tag{2}$$

for some given value $\bar{\kappa}$ to develop inference that is more powerful than the t-test associated with $\hat{\beta}^{\text{long}}$. Since $(\hat{\beta}^{\text{short}}, \hat{\beta}^{\text{long}})$ and the bound (2) are invariant to linear transformations of the controls, so is our new test.

The parameter κ^2 in (2) is a measure of the sample variation of y_i that is induced by the controls $z_{i,j}$ in (1). The bound $\bar{\kappa}^2$ is thus interpretable as an *a priori* bound on the explanatory power of the controls for the observed variation in y_i. Our family of tests, indexed by $\bar{\kappa}$, has the feature that if the usual t-test based on $\hat{\beta}^{\text{short}}$ rejects, but the t-test based on $\hat{\beta}^{\text{long}}$ doesn't, then there exists a value $\bar{\kappa}^* > 0$ so that our test rejects for all $0 \leq \bar{\kappa} < \bar{\kappa}^*$, and it does not reject for $\kappa > \bar{\kappa}^*$. In applications, we expect that researchers don't necessarily have a specific $\bar{\kappa}$ in mind, but rather they report the threshold value $\bar{\kappa}^*$. Comparing this threshold with the observed variability in y_i leads to an interpretable condition about the magnitude of the control coefficients that is minimally sufficient for a significant result on β.

Our suggested test statistic is simply the Likelihood Ratio statistic based on the large sample normality of $(\hat{\beta}^{\text{short}}, \hat{\beta}^{\text{long}})$ and the bound on the omitted variable bias of $\hat{\beta}^{\text{short}}$ implied by (2), for which we tabulate appropriate critical values. From an econometric theory perspective, it is interesting to investigate whether this test comes close to efficiently exploiting the information contained in (2). To this end, we consider the Gaussian homoskedastic version of (1), and consider asymptotics where the number of controls k is of the same order of magnitude as the sample size n. Our main theoretical finding is that in this model, bivariate tests that depend on the data only through $(\hat{\beta}^{\text{short}}, \hat{\beta}^{\text{long}})$ are asymptotically efficient in a well defined sense as long as $\kappa^2 = o(n^{1/4})$.

L_2 penalties of the form (5) play a key role in ridge regression (Hoerl and Kennard (1970)), but our set-up uses (2) as a hard constraint on the nuisance parameters γ_j only.

Furthermore, our focus is on hypothesis testing and confidence intervals, and ridge regression estimators do not automatically lead to shorter confidence intervals (see, for instance, Obenchain (1977)). In recent work, Armstrong and Kolesár (2016) derive small sample minimax optimal confidence intervals in a class of Gaussian regression models with the regression function an element of a known convex set. As they point out in their Section 4.1.2., their generic results could be applied to (1) under the bound (2). Our approach of exploiting an *a priori* bound on the value of a nuisance parameter is also related in spirit to the analysis of Conley, Hansen, and Rossi (2012), who consider instrumental variable estimation with an imperfect instrument that has a direct effect on the outcome of bounded magnitude.

Section 2 contains the analysis of the Gaussian linear regression model (1). Section 2.1 introduces the model more formally and discusses a straightforward extension of (1) which includes additional baseline controls with *a priori* unconstrained coefficients. In this model, the likelihood ratio test is exact, and we analyze and compare its power properties in Section 2.2. Section 2.3 derives the asymptotic efficiency result for bivariate tests. Section 3 discusses the implementation of the likelihood ratio test for general non-normal, possibly heteroskedastic and clustered linear regressions. We compare our test with the method proposed by Belloni, Chernozhukov, and Hansen (2014) in a Monte Carlo exercise in Section 4, and in the empirical exercise from Donohue III and Levitt (2001) in Section 5. Section 6 concludes. All proofs are collected in an appendix.

2 Gaussian Linear Model

2.1 Set-up

Write model (1) in vector form as

$$\mathbf{y} = \mathbf{x}\beta + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon} \tag{3}$$

in obvious notation. Without loss of generality, consider tests of the null hypothesis $H_0 : \beta = 0$ (non-zero null values β_0 can be transformed into this problem by subtracting $\mathbf{x}\beta_0$ from $\mathbf{y}$, and confidence intervals about β can be obtained by inverting these tests). Initially, we focus on the canonical model where $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$ and the regressors $\mathbf{x}$ and $\mathbf{Z}$ are nonstochastic.

We assume throughout that $(\mathbf{x}, \mathbf{Z})$ is of full column rank. The $(k + 1)$ vector of OLS estimators

$$\begin{pmatrix} \hat{\beta}^{\text{long}} \\ \hat{\boldsymbol{\gamma}} \end{pmatrix} = \begin{pmatrix} \mathbf{x}'\mathbf{x} & \mathbf{x}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{x} & \mathbf{Z}'\mathbf{Z} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{x}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \beta \\ \boldsymbol{\gamma} \end{pmatrix}, \begin{pmatrix} \mathbf{x}'\mathbf{x} & \mathbf{x}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{x} & \mathbf{Z}'\mathbf{Z} \end{pmatrix}^{-1} \right) \tag{4}$$

forms a sufficient statistic. Inference about β thus becomes inference about one element of the mean of a $k+1$ dimensional multivariate normal with known covariance matrix. By Proposition 15.2 of van der Vaart (1998), for one-sided tests about β, the uniformly most powerful test statistic is hence simply given by $\hat{\beta}^{\text{long}}$, and the uniformly most powerful unbiased test rejects for large values of $|\hat{\beta}^{\text{long}}|$. Thus in this canonical model, no procedure whatsoever can do better than simply running the "long regression" that includes all controls, at least for one-sided inference, but also for two-sided unbiased inference. This shows that the corresponding observation for a single control variable in Elliott, Müller, and Watson (2015) generalizes to a vector of controls; Elliott, Müller, and Watson (2015) also show that only moderate power gains are possible for two-sided tests without the unbiasedness constraint.

Thus some constraint on $\boldsymbol{\gamma}$ is necessary to make further progress, motivating the *a priori* bound (2), which in vector notation becomes

$$\kappa^2 = \boldsymbol{\gamma}'\mathbf{Z}'\mathbf{Z}\boldsymbol{\gamma} \leq \bar{\kappa}^2. \tag{5}$$

Note that with $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$, κ^2 is directly linked to the population R^2 in a regression of y_i on $z_{i,j}$, $j = 1, \ldots, k$ under H_0, since $R^2_{yZ} = \kappa^2/(\kappa^2 + n)$.

It is straightforward to extend the model with additional baseline controls $q^e_{i,j}$, $j = 1, \ldots, m$ with unconstrained coefficients by suitably defining x_i and $z_{i,j}$ as projections off $q^e_{i,j}$. Specifically, assume we have $n+m$ observations of the model

$$y^e_i = x^e_i\beta + \sum_{j=1}^{k} \gamma_j z^e_{i,j} + \sum_{j=1}^{m} \delta_j q^e_{i,j} + \varepsilon^e_i, \tag{6}$$

such that in vector notation, $\mathbf{y}^e = \mathbf{x}^e\beta + \mathbf{Z}^e\boldsymbol{\gamma} + \mathbf{Q}^e\boldsymbol{\delta} + \boldsymbol{\varepsilon}^e$, where $\mathbf{y}^e$ and $\mathbf{x}^e$ are $(n+m)$ vectors, $\mathbf{Z}^e$ and $\mathbf{Q}^e$ are $(n+m)\times k$ and $(n+m)\times m$ matrices, respectively and $\boldsymbol{\varepsilon}^e \sim \mathcal{N}(0, \mathbf{I}_{n+m})$. Suppose $(\mathbf{Q}^e, \mathbf{Z}^e)$ is of full column rank. Let $\mathbf{P}^e$ be a $(n+m)\times n$ matrix such that $\mathbf{P}^{e\prime}\mathbf{P}^e = \mathbf{I}_n$ and $\mathbf{P}^e\mathbf{P}^{e\prime} = \mathbf{I}_{m+n} - \mathbf{Q}^e(\mathbf{Q}^{e\prime}\mathbf{Q}^e)^{-1}\mathbf{Q}^{e\prime}$. Then with $\mathbf{y} = \mathbf{P}^{e\prime}\mathbf{y}^e$, $\mathbf{x} = \mathbf{P}^{e\prime}\mathbf{x}^e$ and $\mathbf{Z} = \mathbf{P}^{e\prime}\mathbf{Z}^e$, (6) implies $\mathbf{y} = \mathbf{x}\beta + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}$ with $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \mathbf{I}_n)$, as in (3) above. What is more, $\mathbf{y}$ is a maximal invariant to the group of transformations $\mathbf{y}^e \to \mathbf{y}^e + \mathbf{Q}^{e\prime}\mathbf{d}$ for arbitrary $\mathbf{d} \in \mathbb{R}^m$. By Theorem 6.2.1 of Lehmann and Romano (2005), all invariant tests of $H_0 : \beta = 0$ in (6) can be written as functions of $\mathbf{y}$, and the discussion below (3) now applies to the properties of invariant tests. By the Frisch-Waugh theorem, $\hat{\beta}^{\text{long}}$ and $\hat{\boldsymbol{\gamma}}$ defined in (4) when applied to the transformed model (6) correspond to the usual OLS regression coefficients on x_i^e and $z_{i,j}^e$ in a regression of y_i^e on $\{x_i^e, z_{i,1}^e, \ldots, z_{i,k}^e, q_{i,1}^e, \ldots, q_{i,m}^e\}$. Correspondingly, the constraint (5) amounts to upper bound on the additional explanatory power of $z_{i,j}^e$, after controlling for the baseline controls $q_{i,j}^e$.

2.2 Bivariate Testing Problem

In order to exploit the bound (5) for inference, consider the coefficient estimator $\hat{\beta}^{\text{short}}$ from the regression of $\mathbf{y}$ on $\mathbf{x}$ that excludes the controls, $\hat{\beta}^{\text{short}} = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{y}$. Let $\rho^2 = \mathbf{x}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{x}/(\mathbf{x}'\mathbf{x})$, the R^2 of a regression of $\mathbf{x}$ on $\mathbf{Z}$. To avoid trivial complications in notation, assume $0 < \rho$ in the following. Straightforward algebra yields

$$\begin{pmatrix} \hat{\beta}^{\text{long}} \\ \hat{\beta}^{\text{short}} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \beta \\ \beta + \Delta \end{pmatrix}, (\mathbf{x}'\mathbf{x})^{-1} \begin{pmatrix} \frac{1}{1-\rho^2} & 1 \\ 1 & 1 \end{pmatrix} \right) \tag{7}$$

where $\Delta = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{Z}\boldsymbol{\gamma}$ is the omitted variable bias. Equation (7) is intuitive: the long regression provides an unbiased signal $\hat{\beta}^{\text{long}}$ about β, but with a variance that is larger than the (typically biased) signal $\hat{\beta}^{\text{short}}$ from the short regression. If $\rho \to 0$, then $\mathbf{Z}$ is orthogonal to $\mathbf{x}$, there is no bias from the short regression, and the two signals are identical, $\hat{\beta}^{\text{long}} = \hat{\beta}^{\text{short}}$.

Notice that $\kappa^2 = \boldsymbol{\gamma}'\mathbf{Z}'\mathbf{Z}\boldsymbol{\gamma}$ in (5) may be rewritten as

$$\begin{aligned} \kappa^2 &= \boldsymbol{\gamma}'\mathbf{Z}'\mathbf{x}(\mathbf{x}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\gamma}'\mathbf{Z}'\mathbf{M}_\rho\mathbf{Z}\boldsymbol{\gamma} \\ &= \rho^{-2}(\mathbf{x}'\mathbf{x})\Delta^2 + \boldsymbol{\gamma}'\mathbf{Z}'\mathbf{M}_\rho\mathbf{Z}\boldsymbol{\gamma} \end{aligned} \tag{8}$$

where $\mathbf{M}_\rho = \mathbf{I}_k - \mathbf{x}(\mathbf{x}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{x})^{-1}\mathbf{x}' = \mathbf{I}_k - \rho^{-2}\mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'$. The bound $\kappa^2 \leq \bar{\kappa}^2$ in (5) thus implies an upper bound[1] on the omitted variable bias,

$$|\Delta| \leq \rho\bar{\kappa}/\sqrt{\mathbf{x}'\mathbf{x}}. \tag{9}$$

This limit on the magnitude of the omitted variable bias in (7) makes $\hat{\beta}^{\text{short}}$ potentially valuable for inference about β, especially if ρ is close to one (so that $\hat{\beta}^{\text{short}}$ is much less variable than $\hat{\beta}^{\text{long}}$). The inference problem then becomes the problem of testing $H_0 : b = 0$ from observing the bivariate normal vector $\mathbf{W} = (W_1, W_2)' \sim (\sqrt{\mathbf{x}'\mathbf{x}}\hat{\beta}^{\text{long}}, \sqrt{\mathbf{x}'\mathbf{x}}\hat{\beta}^{\text{short}})'$,

$$\begin{pmatrix} W_1 \\ W_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} b \\ b + \rho d \end{pmatrix}, \boldsymbol{\Sigma}(\rho) \right), \ \boldsymbol{\Sigma}(\rho) = \begin{pmatrix} \frac{1}{1-\rho^2} & 1 \\ 1 & 1 \end{pmatrix}, \ |d| \leq \bar{\kappa} \tag{10}$$

[1] This bound is sharp: With $\boldsymbol{\gamma}$ proportional to $(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{x}$ and an equality in (5), $|\Delta| = \rho\bar{\kappa}/\sqrt{\mathbf{x}'\mathbf{x}}$.

with $b = \beta\sqrt{\mathbf{x}'\mathbf{x}}$, $d = \Delta\sqrt{\mathbf{x}'\mathbf{x}}$.

Under larger and larger bounds $\bar{\kappa} \to \infty$ with d close to the bound, $|\bar{\kappa} - d| \to a$ for some fixed a, the "two-sided" problem (10) turns into one of two "one-sided" problems indexed by $i \in \{-1, 1\}$ via the recentering $\mathbf{W}^o = (W_1^o, W_2^o)' = (W_1, W_2 - i\bar{\kappa})'$

$$\begin{pmatrix} W_1^o \\ W_2^o \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} b \\ b + a \end{pmatrix}, \mathbf{\Sigma}(\rho)\right) \tag{11}$$

where under $H_0 : b = 0$ and $a \in A_i$, with $A_1 = (-\infty, 0]$ and $A_{-1} = [0, \infty)$. Finally, if $\bar{\kappa} \to \infty$ and $\bar{\kappa} - |d| \to \infty$, then with or without recentering by $i\bar{\kappa}$, one obtains the problem (11) with unrestricted $a \in A_0 = \mathbb{R}$ under the null hypothesis.

The inference problem (10) is a fairly transparent small sample problem indexed by two known parameters $(\rho^2, \bar{\kappa}) \in [0, 1] \times [0, \infty)$, and involves a one-dimensional unknown nuisance parameter $d \in \mathbb{R}$. The second observation W_2 augments the usual Gaussian shift experiment, and there are a variety of potential approaches to exploiting this additional information. We found that a simple but effective test of $H_0 : b = 0$ is generated by the generalized likelihood ratio statistic

$$\begin{aligned} \mathrm{LR}(\bar{\kappa}) &= \min_{|d| \le \bar{\kappa}} \begin{pmatrix} W_1 \\ W_2 - \rho d \end{pmatrix}' \mathbf{\Sigma}(\rho)^{-1} \begin{pmatrix} W_1 \\ W_2 - \rho d \end{pmatrix} \\ &\quad - \min_{b, |d| \le \bar{\kappa}} \begin{pmatrix} W_1 - b \\ W_2 - b - \rho d \end{pmatrix}' \mathbf{\Sigma}(\rho)^{-1} \begin{pmatrix} W_1 - b \\ W_2 - b - \rho d \end{pmatrix}. \end{aligned} \tag{12}$$

For the one-sided problems (11), define the analogous statistics

$$\begin{aligned} \mathrm{LR}_i^o &= \min_{a \in A_i} \begin{pmatrix} W_1^o \\ W_2^o - a \end{pmatrix}' \mathbf{\Sigma}(\rho)^{-1} \begin{pmatrix} W_1^o \\ W_2^o - a \end{pmatrix} \\ &\quad - \min_{b, a \in A_i} \begin{pmatrix} W_1^o - b \\ W_2^o - a - b \end{pmatrix}' \mathbf{\Sigma}(\rho)^{-1} \begin{pmatrix} W_1^o - b \\ W_2^o - a - b \end{pmatrix}. \end{aligned}$$

Figure 1 plots the 5% level critical value $\mathrm{cv}_\rho(\bar{\kappa})$ of $\mathrm{LR}(\bar{\kappa})$ that is just large enough to ensure size control for all values of $|d| \le \bar{\kappa}$ for $\rho \in \{0.5, 0.95, 0.99\}$, and Figure 2 plots the rejection region of the resulting 5% level test for $\rho = 0.95$ and $\bar{\kappa} \in \{0, 1, 3, 10\}$. For $\bar{\kappa} = 0$, the LR test reduces to rejecting for large values of $W_2^2 > \mathrm{cv}_\rho(0) = 1.96^2$, that is it reduces to the usual t-test based on the short regression. More generally, whenever $|W_2| \gg \bar{\kappa}$, that is the short regression coefficient value is much larger than $\bar{\kappa}$ in absolute value, then the

Figure 1: Five percent critical value of LR($\bar{\kappa}$) as a function of $\bar{\kappa}$

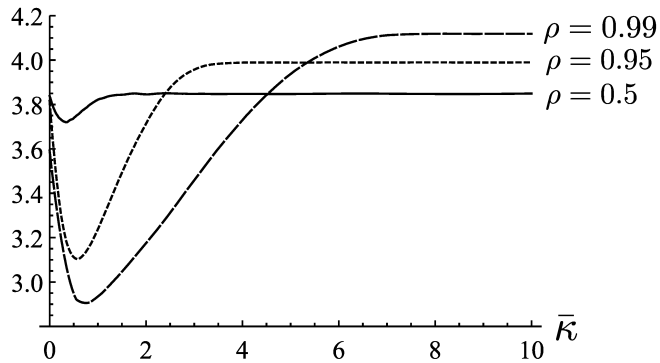
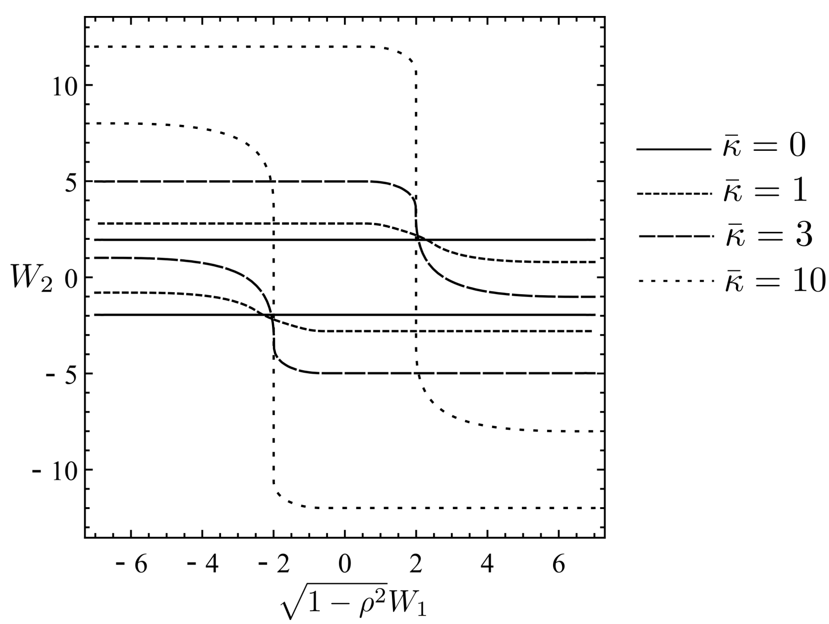


Figure 2: Acceptance region of LR($\bar{\kappa}$) for $\rho = 0.95$



Notes: The lines are the boundaries of the acceptance region. For all values of $\bar{\kappa}$, $(0,0)$ is in the acceptance region.

LR test rejects. On the other hand, for $|W_2| \ll \bar{\kappa}$ and $\bar{\kappa}$ large, the LR test rejects when $(1-\rho^2)W_1^2 > \mathrm{cv}_\rho(\bar{\kappa})$, that is whenever the long regression coefficient is too large in absolute value, with a critical value that is slightly larger than what one would employ in a standard chi-squared test with one degree of freedom. Once $\bar{\kappa}$ is moderately large (say, larger than 8), the critical value $\mathrm{cv}_\rho(\bar{\kappa})$ stabilizes at $\mathrm{cv}_\rho(\infty)$, and further increases of $\bar{\kappa}$ simply amount to an additional outward shift of the acceptance region. Indeed, after a recentering by $i\bar{\kappa}$, this acceptance region is equal to the acceptance region of a test based on LR_i^o, $i \in \{-1, 1\}$, whose smallest valid critical value also equals $\mathrm{cv}_\rho(\infty)$. The LR_i^o test is hence simply the smooth extension of the fixed $\bar{\kappa}$ test based on (10) to the one-sided $\bar{\kappa} \to \infty$ problems (11). Formally, for any $(w_1, w_2^o) \in \mathbb{R}^2$,

$$\lim_{\bar{\kappa}\to\infty} \varphi_{\mathrm{LR}}(\bar{\kappa}, (w_1, w_2^o + i\bar{\kappa})') = \varphi_{\mathrm{LR}_i^o}((w_1, w_2^o)'). \tag{13}$$

In the unconstrained test of observing (11) with $a \in A_0 = \mathbb{R}$ under H_0, the LR_0^o based test rejects for large values of $(1-\rho^2)(W_1^o)^2$, and relying on $\mathrm{cv}_\rho(\infty)$ instead of 1.96^2 induces a slight power loss.

We consider this LR approach attractive for a number of reasons. First, it is easy to implement (we discuss implementation issues in more detail in Section 3 below). Second, it has close to maximal weighted average power under a weighting function where $b \sim \mathcal{N}(0, 10)$ and d is uniform between $[0, \bar{\kappa}]$. This is shown in panel A of Table 1, which reports an upper bound on this weighted average power for all 5% level tests of $H_0 : b = 0$ and $\rho|d| \leq \bar{\kappa}$ for selected values of $(\rho, \bar{\kappa})$, along with the weighted average power of the LR test.[2] Third, its average power against symmetric alternatives $b = \pm b_1$ is nearly uniformly higher over all $|\rho d| \leq \bar{\kappa}$ and $b_1 \in \mathbb{R}$ than the standard long regression test that rejects for large $|W_1|$; Table 1 panel B provides corresponding numerical evidence. And finally, the LR approach has the potentially attractive feature that if $|W_2| > 1.96$ and $\sqrt{1-\rho^2}|W_1| < 1.96$ (that is, the short regression rejects, but the long regression doesn't), then there is a unique threshold value $\bar{\kappa}^* > 0$ such that the LR test rejects only when $\bar{\kappa} < \bar{\kappa}^*$, so in this sense, imposing a smaller value of $\bar{\kappa}$ always leads to more informative inference.

More abstractly, a (potentially randomized) test φ of $H_0 : \beta = 0$ under the bound $\bar{\kappa}$ maps the data $\mathbf{Y} = (\mathbf{y}, \mathbf{x}, \mathbf{Z})$ into the unit interval $[0, 1]$, with the value $\varphi(\bar{\kappa}, \mathbf{Y})$ interpreted as the conditional rejection probability given the observation $\mathbf{Y}$. The inversion of these tests yields a confidence set for $\bar{\kappa}$: the set of values for the bound (5) that is compatible with the null hypothesis $\beta = 0$ (or, equivalently, the slice of the confidence region for $(\beta, \bar{\kappa})$ along

[2] These were computed using the algorithm developed by Elliott, Müller, and Watson (2015).

Table 1: Power Properties of Bivariate LR Test

Panel A: Weighted Average Power

	Bivariate LR test				Upper Bound			
$\rho\backslash\bar{\kappa}$	0	1	4	10	0	1	4	10
0.50	0.60	0.56	0.55	0.55	0.60	0.56	0.56	0.56
0.95	0.79	0.72	0.60	0.57	0.79	0.72	0.62	0.58
0.99	0.93	0.91	0.81	0.68	0.93	0.91	0.81	0.70

Panel B: Smallest and Largest Difference in Symmetrized Power between Bivariate LR test and Long Regression t-Test

	Smallest				Largest			
$\rho\backslash\bar{\kappa}$	0	1	4	10	0	1	4	10
0.50	0.00	0.00	0.00	-0.01	0.12	0.06	0.06	0.05
0.95	0.00	-0.03	-0.02	-0.02	0.62	0.48	0.28	0.28
0.99	0.00	-0.04	-0.05	-0.03	0.88	0.88	0.76	0.48

Panel C: Weighted Expected Length of Confidence Intervals for $\bar{\kappa}$

	Bivariate LR test				Lower Bound			
$\rho\backslash\bar{\kappa}_{\max}$	0	1	4	10	0	1	4	10
0.50	0.40	0.42	0.37	0.31	0.40	0.38	0.28	0.25
0.95	0.21	0.24	0.31	0.31	0.21	0.24	0.30	0.28
0.99	0.07	0.08	0.13	0.20	0.07	0.08	0.13	0.19

Notes: All tests are of level 5%. Weighted average power in Panel A has $b \sim \mathcal{N}(0,10)$ and $d \sim U[0,\bar{\kappa}]$. Panel B reports largest and smallest difference of average power against alternatives $(+b_1,d_1)$, $(-b_1,d_1)$ over the grid with $d_1 \in \{0.05\bar{\kappa}, 0.1\bar{\kappa}, \ldots, \bar{\kappa}\}$ and $\sqrt{1-\rho^2}b_1 \in \{0.5, 1.0, \ldots, 4.0\}$. Panel C reports expected length of confidence intervals for $\bar{\kappa}$ as in equation (14), with a weighting function F_0 with $b \sim \mathcal{N}(0,10)$ and $d \sim U[0, K_{\max}]$, and Π_0 has $\bar{\kappa}$ uniform between 0 and $\bar{\kappa}_{\max}$. Based on 250,000 importance sampling draws.

$\beta = 0$). As discussed in the introduction, we expect that researchers will not necessarily approach a problem with a given value of $\bar{\kappa}$ in mind. Rather, they might be interested in the weakest assumption (that is, the largest value of $\bar{\kappa}$) that still induces rejection of the null hypothesis. From that perspective the objective is to construct $\varphi(\bar{\kappa}, \mathbf{Y})$ in a way such that the length $\int(1-\varphi(\bar{\kappa}, \mathbf{Y}))d\bar{\kappa}$ of the confidence set tends to be as small as possible. Since $\bar{\kappa}$ is potentially unbounded, length can be infinite. A more generally applicable objective is the weighted average length $\int(1-\varphi(\bar{\kappa}, \mathbf{Y}))d\Pi(\bar{\kappa})$ for some probability measure Π with support in $\mathbb{R}^+$. (If on is willing to bound $\bar{\kappa}$ from above, then one can simply set Π to be uniform between zero and this upper bound, and $\int(1-\varphi(\bar{\kappa}, \mathbf{Y}))d\Pi(\bar{\kappa})$ reduces to a scalar multiple of $\int(1-\varphi(\bar{\kappa}, \mathbf{Y}))d\bar{\kappa}$.)

As noted by Pratt (1961), there is a tight connection between the power of tests and the length of the resulting confidence set. In the following discussion, we focus on *bivariate tests* that depend on $\mathbf{Y}$ only through $\psi(\mathbf{Y}) = (\sqrt{\mathbf{x}'\mathbf{x}}\hat{\beta}^{\text{long}}, \sqrt{\mathbf{x}'\mathbf{x}}\hat{\beta}^{\text{short}})$. The distribution of $\psi(\mathbf{Y})$ depends on $b = \beta\sqrt{\mathbf{x}'\mathbf{x}}$ and $d = \Delta\sqrt{\mathbf{x}'\mathbf{x}}$, so the rejection probability of bivariate tests may be written as $E_{b,d}[\varphi_W(\bar{\kappa}, \psi(\mathbf{Y}))]$, where the subscript of the expectation operator indicates the parameters that govern the distribution of $\psi(\mathbf{Y})$. Let F_0 be a bivariate probability measure, so that $\int E_{b,d}[\varphi_W(\bar{\kappa}, \psi(\mathbf{Y}))]dF_0(b,d)$ is the F_0-weighted average power of the test φ_W. Then, for any probability measure Π on $\mathbb{R}^+$, Tonelli's Theorem yields

$$\int E_{b,d}[\int(1-\varphi_W(\bar{\kappa}, \psi(\mathbf{Y})))d\Pi(\bar{\kappa})]dF_0(b,d) = 1 - \int\int E_{b,d}[\varphi_W(\bar{\kappa}, \psi(\mathbf{Y}))]dF_0(b,d)d\Pi(\bar{\kappa}) \tag{14}$$

so that there is a one-to-one correspondence between F_0-weighted expected average power and F_0-weighted expected length. Panel C of Table 1 provides evidence that the LR test has fairly attractive average expected length properties among all bivariate tests under weighting scheme where $b \sim \mathcal{N}(0, 10)$, $d \sim U[0, \bar{\kappa}]$ and $\bar{\kappa} \sim U[0, \bar{\kappa}_{\max}]$ for a range of values of $\bar{\kappa}_{\max}$.

When $\bar{\kappa}$ diverges, length can be defined by focussing on local deviations from some baseline value $\bar{\kappa}_0 \to \infty$. For instance, from (13), the likelihood ratio test with $\bar{\kappa} = \bar{\kappa}_0 + K$ satisfies $\lim_{\bar{\kappa}_0\to\infty} \varphi_{\text{LR}}(\bar{\kappa}_0 + K, (w_1, w_2^o + i\bar{\kappa})) = \varphi_{\text{LR}_i^o}((w_1, w_2^o - iK))$ for all $(K, w_1, w_2^o) \in \mathbb{R}^3$. More generally, a test $\tilde{\varphi}_i^o(\mathbf{W}^o)$ in the one-sided problem (11) induces the family of tests $\varphi_i^o(K, \mathbf{W}^o) = \tilde{\varphi}_i^o((W_1^o, W_2^o - iK)')$ indexed by K. With length measured by $\int(1 - \varphi_i^o(K, \mathbf{W}^o))d\Pi(K)$, the analogue to (14) is

$$\int E_{b,d}[\int(1-\varphi_i^o(K, \psi^o(\mathbf{Y})))d\Pi(K)]dF_0(b,d) = 1 - \int\int E_{b,d}[\varphi_i^o(K, \psi^o(\mathbf{Y}))]dF_0(b,d)d\Pi(K) \tag{15}$$

for $\psi^o(\mathbf{Y}) = \psi(\mathbf{Y}) + (0, -i\bar{\kappa}_0)'$.

2.3 Asymptotic Efficiency of Bivariate Tests

Regardless how exactly they are constructed, good tests of $H_0 : b = 0$ based on $\mathbf{W} = (W_1, W_2)'$ in (10) will in general be more powerful than those based on W_1 alone under many alternatives. But this does not mean that they necessarily fully exploit the information in the bound (5). After all, the reduction to the bivariate problem (7) was not based on any sufficiency argument. So the question arises whether one can do systematically better than constructing good tests for the bivariate problem (10). This section takes up this issue under asymptotics where $k/n \to c \in (0,1)$, that is when the number of controls is a non-negligible fraction of the number of observations.

In general, the rejection probability of a test $\varphi(\bar{\kappa}, \mathbf{Y})$ based on the entire set of observations $\mathbf{Y}$ is not only a function of β, the bias $\Delta = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{Z}\boldsymbol{\gamma}$ of the short regression and the slackness in the inequality (9) $\tau^2 = \kappa^2 - \rho^{-2}(\mathbf{x}'\mathbf{x})\Delta^2 = \boldsymbol{\gamma}'\mathbf{Z}'\mathbf{M}_\rho\mathbf{Z}\boldsymbol{\gamma}$, but also of the direction of $\boldsymbol{\gamma}$ that leads to identical values of Δ and τ. Consider the linear reparameterization $\mathbf{Z}\boldsymbol{\gamma} = \mathbf{Z}^r\boldsymbol{\gamma}^r$ with $\boldsymbol{\gamma}^r = (\mathbf{Z}'\mathbf{Z})^{1/2}\boldsymbol{\gamma}$ and $\mathbf{Z}^r = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1/2}$, and corresponding estimator $\hat{\boldsymbol{\gamma}}^r$=$(\mathbf{Z}'\mathbf{Z})^{1/2}\hat{\boldsymbol{\gamma}}$. Further, let $\mathbf{P}_{xZ^r}$ be a $k \times (k-1)$ matrix such that $\mathbf{P}'_{xZ^r}\mathbf{Z}^{r\prime}\mathbf{x}$ =0 and $\mathbf{P}'_{xZ^r}\mathbf{P}_{xZ^r} = \mathbf{I}_{k-1}$. Then with $\hat{\boldsymbol{\phi}} = \mathbf{P}'_{xZ^r}\hat{\boldsymbol{\gamma}}^r$, it follows from (4) that

$$\begin{pmatrix} \hat{\beta}^{\text{long}} \\ \hat{\beta}^{\text{short}} \\ \hat{\boldsymbol{\phi}} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \beta \\ \beta + \Delta \\ \tau\boldsymbol{\omega} \end{pmatrix}, (\mathbf{x}'\mathbf{x})^{-1} \begin{pmatrix} \frac{1}{1-\rho^2} & 1 & \mathbf{0} \\ 1 & 1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & (\mathbf{x}'\mathbf{x})\mathbf{I}_{k-1} \end{pmatrix} \right) \tag{16}$$

where $\boldsymbol{\omega} = \mathbf{P}'_{xZ^r}\boldsymbol{\gamma}^r/||\mathbf{P}'_{xZ^r}\hat{\boldsymbol{\gamma}}^r|| = \mathbf{P}'_{xZ^r}\boldsymbol{\gamma}^r/\tau$. The parameter $\boldsymbol{\omega}$ is an element of the surface of the $k-1$ dimensional unit hypersphere and indicates the direction of $\boldsymbol{\gamma}^r$ in the $k-1$ dimensional subspace orthogonal to $\mathbf{Z}^{r\prime}\mathbf{x}$.

In the parameterization $(\beta, \Delta, \tau, \boldsymbol{\omega})$, the rejection probability of φ is $E_{\beta,\Delta,\tau,\boldsymbol{\omega}}[\varphi(\bar{\kappa}, \mathbf{Y})]$, where the subscript indicates the parameters that govern the distribution of $\mathbf{Y}$. In absence of any additional information about the controls, it seems natural to consider tests whose rejection probability does not vary with $\boldsymbol{\omega}$. The following lemma shows that for any such test, there exists another test with the same power function that is a function of the three dimensional statistic $\mathbf{T} = (\hat{\beta}^{\text{long}}, \hat{\beta}^{\text{short}}, \hat{\tau})'$, where

$$\hat{\tau}^2 = \hat{\boldsymbol{\gamma}}'\mathbf{Z}'\mathbf{M}_\rho\mathbf{Z}\hat{\boldsymbol{\gamma}} = \hat{\boldsymbol{\phi}}'\hat{\boldsymbol{\phi}}.$$

Note that the distribution of $\mathbf{T}$ only depends on (β, Δ, τ). Thus, to the extent that one is willing to restrict attention to tests whose power function is symmetric in this sense, one might focus on tests that are functions of $\mathbf{T}$, with an effective parameter space equal to $\theta = (\beta, \Delta, \tau) \in \mathbb{R}^2 \times [0, \infty)$.

Lemma 1 *For any $n > 2$, if φ is a test such that $E_{\beta,\Delta,\tau,\boldsymbol{\omega}}[\varphi(\mathbf{Y})]$ does not vary in $\boldsymbol{\omega}$ for all (β, Δ, τ), then there exists a test $\tilde{\varphi}(\mathbf{T})$ such that $E_{\beta,\Delta,\tau,\boldsymbol{\omega}}[\tilde{\varphi}(\mathbf{T})] = E_{\beta,\Delta,\tau,\boldsymbol{\omega}}[\varphi(\mathbf{Y})]$ for all $(\beta, \Delta, \tau, \boldsymbol{\omega})$.*

Given the decomposition of κ^2 in (8), it would clearly be beneficial to know the value of $\tau^2 = \boldsymbol{\gamma}'\mathbf{Z}'\mathbf{M}_\rho\mathbf{Z}\boldsymbol{\gamma}$, as it would allow the strengthening of the bound on Δ under (5) to $|\Delta| \leq \rho\sqrt{\bar{\kappa}^2 - \tau^2}/\sqrt{\mathbf{x}'\mathbf{x}}$. This suggests that $\hat{\tau}^2$ could potentially be used for the same purpose. But to be actually useful, the estimation error in $\hat{\tau}^2$ must not be too large relative to $\bar{\kappa}^2$. The following Lemma shows that in large samples, $\hat{\tau}$ does not contain useful information about τ as long as τ is small. From now on, we use subscripts to denote the value of quantities and functions that depend on the sample size n.

Lemma 2 *Let $L_n(\tau)$ be the likelihood of τ based on the observation $\hat{\tau}_n$ in the regression model (1) with n observations and $\varepsilon_i \sim iid\mathcal{N}(0, 1)$. If $k_n/n \to c \in (0, 1)$, $\tau_n = o(n^{1/4})$ and $t_n = o(n^{1/4})$, then $L_n(t_n)/L_n(0) \xrightarrow{p} 1$.*

The lemma shows that even the likelihood ratio statistic for the observation $\hat{\tau}_n$ does not drive an asymptotic wedge between the values $\tau_n = 0$ and $\tau_n = o(n^{1/4})$, suggesting that $\mathbf{T}_n$ does not help to determine the value of τ_n of order $o(n^{1/4})$. Combining the observations in Lemmas 1 and 2 with limit of experiments arguments leads to the following result.

Theorem 1 *Consider a sequence of observations from the linear regression model with $\varepsilon_i \sim iid\mathcal{N}(0, 1)$ where $k_n/n \to c \in (0, 1)$ and $\rho_n^2 \to \rho^2 \in [0, 1)$, and let $\varphi_n(\bar{\kappa}_n, \mathbf{Y}_n)$ be a sequence of tests that, for all sufficiently large n, satisfy the assumption of Lemma 1.*

(a) Suppose $\bar{\kappa}_n = \bar{\kappa}$ for all n. If for all $(b, d) \in \mathbb{R}^2$, $E_{\theta_n}[\varphi_n(\bar{\kappa}, \mathbf{Y}_n)]$ converges along a sequence θ_n with $\sqrt{\mathbf{x}_n'\mathbf{x}_n}(\beta_n, \Delta_n) = (b, d)$ and $\tau_n = o(n^{1/4})$ (where τ_n may depend on (b, d)), then $\lim_{n\to\infty} E_{\theta_n}[\varphi_n(\bar{\kappa}_n, \mathbf{Y}_n)] = E_{b,d}[\varphi(\mathbf{W})]$ for some test $\varphi : \mathbb{R}^2 \mapsto [0, 1]$ with $\mathbf{W}$ distributed as in (10).

(b) Suppose $\bar{\kappa}_n \to \infty$, and $i \in \{-1, 1\}$. If for all $(b, a) \in \mathbb{R}^2$, $E_{\theta_n}[\varphi_n(\bar{\kappa}_n, \mathbf{Y}_n)]$ converges along a sequence θ_n with $(\sqrt{\mathbf{x}_n'\mathbf{x}_n}\beta_n, \sqrt{\mathbf{x}_n'\mathbf{x}_n}\Delta_n - i\bar{\kappa}_n) = (b, a)$ and $\tau_n = o(n^{1/4})$ (where τ_n may depend on (b, a)), then $\lim_{n\to\infty} E_{\theta_n}[\varphi_n(\bar{\kappa}_n, \mathbf{Y}_n)] = E_{b,a}[\varphi(\mathbf{W}^o)]$ for some test $\varphi : \mathbb{R}^2 \mapsto [0, 1]$ with $\mathbf{W}^o$ distributed as in (11), where $a \in A_i$ under the null hypothesis of $b = 0$.

(c) Suppose $\bar{\kappa}_n \to \infty$. If for some sequence s_n and all $(b, a) \in \mathbb{R}$, $E_{\theta_n}[\varphi_n(\bar{\kappa}_n, \mathbf{Y}_n)]$ converges along a sequence θ_n with $\sqrt{\mathbf{x}_n'\mathbf{x}_n}\beta_n = b$, $\sqrt{\mathbf{x}_n'\mathbf{x}_n}\Delta_n - s_n = a$, $\bar{\kappa}_n - |s_n| \to \infty$ and $\tau_n = o(n^{1/4})$ (where τ_n may depend on (b, a)), then $\lim_{n\to\infty} E_{\theta_n}[\varphi_n(\bar{\kappa}_n, \mathbf{Y}_n)] = E_{b,a}[\varphi(\mathbf{W}^o)]$ for some test $\varphi : \mathbb{R}^2 \mapsto [0, 1]$ with $\mathbf{W}^o$ distributed as in (11), where $a \in \mathbb{R}$ under the null hypothesis of $b = 0$.

The theorem demonstrate that under $\tau_n = o(n^{1/4})$ the asymptotic power of any test satisfying the condition of Lemma 1 can always be matched by the power of a bivariate test that depends on the data only through the short and long regression coefficient estimators. The limiting experiments described by the different parts correspond to the two-sided ($\bar{\kappa}$ fixed) and one-sided ($\bar{\kappa} \to \infty$) experiments (10) and (11) introduced in the last subsection. The determination of tests with attractive asymptotic power properties is hence reduced to the problem of identifying good bivariate tests.

Note that the implementation of asymptotically valid bivariate tests is straightforward in the Gaussian homoskedastic model, since the distribution of $\psi(\mathbf{Y}) = (\sqrt{\mathbf{x}_n'\mathbf{x}_n}\hat{\beta}_n^{\text{long}}, \sqrt{\mathbf{x}_n'\mathbf{x}_n}\hat{\beta}_n^{\text{short}})'$ then exactly matches the distribution of $\mathbf{W}$ in (10). In particular, the test

$$\varphi_{\text{LR},n}(\bar{\kappa}_n, \mathbf{Y}_n) = \mathbf{1}[\text{LR}_n(\bar{\kappa}_n) > \text{cv}_{\rho_n}(\bar{\kappa}_n)]$$

with $\text{LR}_n(\bar{\kappa})$ equal to (12) with this choice of $\mathbf{W}$ and $\text{cv}_{\rho_n}(\bar{\kappa}_n)$ as defined in Section 2.2 has asymptotic rejection probabilities under the sequences described in parts (a), (b) and (c) of Theorem 1 that are equal to the small sample rejection probability of the two- and one-sided LR test in (10) and (11). The near small sample optimal weighted average power properties of the LR test discussed in Section 2.2 among bivariate tests thus translate via Theorem 1 into near asymptotic weighted average power optimality in the Gaussian homoskedastic model among the larger class of tests that are only required to satisfy Lemma 1.

Since $\tau \leq \bar{\kappa}$ under the null hypothesis, a bound of the order $\bar{\kappa} = o(n^{1/4})$ implies the assumption $\tau_n = o(n^{1/4})$ under the null hypothesis. Recalling the relationship $R_{yZ}^2 = \kappa^2/(\kappa^2 + n)$, fixed values of $\bar{\kappa}$ correspond to $R_{yZ}^2 = O(n^{-1})$, and $\bar{\kappa} = o(n^{1/4})$ corresponds to $R_{yZ}^2 = o(n^{-1/2})$. This suggests that the efficiency implication of Theorem 1 is particularly pertinent in small samples for values of $\bar{\kappa}$ that imply that the explanatory power of the controls is small compared to the variation in y_i.

Note, however, that the theorem does not require τ_n to be smaller than $\bar{\kappa}_n$; for instance, in part (a), $\bar{\kappa}_n$ is bounded, but τ_n is allowed to diverge. This corresponds to a situation where the bound $\bar{\kappa}$ is much smaller than the actual value κ. This is most easily interpreted along the lines discussed at the end of Section 2.2 above: The limited information in the data can lead to confidence sets for $\bar{\kappa}$ that include values which are smaller than the true value κ. Hence, the usefulness of a confidence set relies not only on having correct coverage probability, but also on its expected length, where shorter sets are preferable to longer ones. The following corollary translates the asymptotic efficiency results of tests in terms of their power of Theorem 1 into a corresponding claim about the asymptotic length of the implied confidence set, in analogy to equations (14) and (15) of Section 2.2.

Corollary 1 *Suppose the assumptions of Theorem 1 hold. Let s_n be an arbitary positive sequence of order $o(n^{1/4})$, and let G_n be sequence of probability distributions with support equal to a subset of $\mathbb{R}^2 \times [0, \infty)$.*

(a) Suppose G_n is such that under $(\beta, \Delta, \tau) \sim G_n$, $(\sqrt{\mathbf{x}_n'\mathbf{x}_n}\beta, \sqrt{\mathbf{x}_n'\mathbf{x}_n}\Delta) \sim F_0$ for some bivariate probability distribution F_0, and $G_n(\tau < s_n) \to 1$. Further, let Π_0 be some probability distribution with support equal to a subset of $[0, \infty)$. Then for any sequence φ_n such that $E_{\theta_n}[\varphi_n(\bar{\kappa}, \mathbf{Y}_n)]$ converges under a sequence of Theorem 1 (a) for all $\bar{\kappa}$ in the support of Π_0, there exists $\varphi_W : \mathbb{R} \times \mathbb{R}^2 \mapsto [0, 1]$ such that

$$\int E_\theta[\int (1 - \varphi_n(\bar{\kappa}, \mathbf{Y}_n))d\Pi_0(\bar{\kappa})]dG_n(\theta) \to \int E_{a,b}[\int (1 - \varphi_W(\bar{\kappa}, \mathbf{W}))d\Pi_0(\bar{\kappa})]dF_0(a, b).$$

(b) For some $\bar{\kappa}_{0,n} \to \infty$ and $i \in \{-1, 1\}$, let G_n be such that with $(\beta, \Delta, \tau) \sim G_n$, $(\sqrt{\mathbf{x}_n'\mathbf{x}_n}\beta, \sqrt{\mathbf{x}_n'\mathbf{x}_n}\Delta_n - i\bar{\kappa}_{0,n}) \sim F_0$ and $G_n(\tau < s_n) \to 1$. Further, let Π_n be such that with $\bar{K} \sim \Pi_n$, $\bar{K} - \bar{\kappa}_{0,n} \sim \Pi_0$, where Π_0 is a probability distribution with support equal to a subset of $\mathbb{R}$. Then for any sequence φ_n such that $E_{\theta_n}[\varphi_n(\bar{\kappa}_{0,n} + K, \mathbf{Y}_n)]$ converges under the sequences of Theorem 1 (b) for all K in the support of Π_0, there exists $\varphi_i^o : \mathbb{R} \times \mathbb{R}^2 \mapsto [0, 1]$ such that

$$\int E_\theta[\int (1 - \varphi_n(\bar{\kappa}, \mathbf{Y}_n))d\Pi_n(\bar{\kappa})]dG_n(\theta) \to \int E_{a,b}[\int (1 - \varphi_i^o(K, \mathbf{W}^o))d\Pi_0(K)]dF_0(a, b).$$

In summary, under $k/n \to c \in (0, 1)$ asymptotics, as long as $\tau_n = o(n^{1/4})$, the quality of asymptotic inference in the Gaussian homoskedastic model is limited from above by the performance of bivariate tests. The attractive small sample features of the LR approach discussed in Section 2.2 thus translate into attractive large sample inference.

3 Implementation in Non-Gaussian and Potentially Heteroskedastic Models

In the Gaussian linear regression model, the bivariate tests introduced in Section 2.2 have exact small sample properties. But for applied use, it is important to have a valid implementation in non-Gaussian and potentially heteroskedastic models. With the regressors nonstochastic (or after conditioning on the regressors with a conditionally mean zero error term), the general model is still of the form (1), where now $\varepsilon_i \sim (0, \sigma_i^2)$ independent across i. Under weak technical conditions on the tails of the distribution of ε_i, on the sequence $\{\sigma_i^2\}_{i=1}^n$ and on the regressors $\{x_i, z_{i,1}, \ldots, z_{i,k}\}_{i=1}^n$, a central limit theorem yields

$$\mathbf{\Omega}_n^{-1/2} \begin{pmatrix} \hat{\beta}_n^{\text{long}} - \beta_n \\ \hat{\beta}_n^{\text{short}} - \beta_n - \Delta_n \end{pmatrix} \Rightarrow \mathcal{N}(\mathbf{0}, \mathbf{I}_2) \tag{17}$$

for some suitably defined $\mathbf{\Omega}_n$, since $(\hat{\beta}^{\text{long}} - \beta_n, \hat{\beta}^{\text{short}} - \beta_n - \Delta_n)$ are linear combinations of the heterogeneous but mean zero and independent random variables $\{\varepsilon_i\}_{i=1}^n$. We provide a corresponding result in Appendix A.6 that allows for dependence among the ε_i due to clustering, and accommodates the presence of baseline controls.

Suppose $\hat{\mathbf{\Omega}}_n$ is a consistent estimator of $\mathbf{\Omega}_n$ in the sense that $\mathbf{\Omega}_n^{-1}\hat{\mathbf{\Omega}}_n \xrightarrow{p} \mathbf{I}_2$. Recall from Section 2.2 that the L_2 bound (5) implies the inequality(9), that is $|\Delta_n| \leq \bar{K}_\Delta = \rho_n \bar{\kappa}_n / \sqrt{\mathbf{x}_n' \mathbf{x}_n}$. The natural LR statistic of $H_0 : \beta_n = 0$ then becomes

$$\begin{aligned} \widehat{\text{LR}}_n(\bar{\kappa}_n) &= \min_{|d| \leq \bar{K}_\Delta} \begin{pmatrix} \hat{\beta}_n^{\text{long}} \\ \hat{\beta}_n^{\text{short}} - d \end{pmatrix}' \hat{\mathbf{\Omega}}_n^{-1} \begin{pmatrix} \hat{\beta}_n^{\text{long}} \\ \hat{\beta}_n^{\text{short}} - d \end{pmatrix} \\ &\quad - \min_{b, |d| \leq \bar{K}_\Delta} \begin{pmatrix} \hat{\beta}_n^{\text{long}} - b \\ \hat{\beta}_n^{\text{short}} - b - d \end{pmatrix}' \hat{\mathbf{\Omega}}_n^{-1} \begin{pmatrix} \hat{\beta}_n^{\text{long}} - b \\ \hat{\beta}_n^{\text{short}} - b - d \end{pmatrix}. \end{aligned}$$

Exploiting the invariance of the LR statistic to reparameterizations, the distribution of $\widehat{\text{LR}}_n(\bar{\kappa}_n)$ under the approximations (17), $\hat{\mathbf{\Omega}}_n = \mathbf{\Omega}_n$ and $|\Delta_n| \leq \bar{K}_\Delta$ is effectively indexed by two scalar parameters

$$\psi_1 = \frac{\Omega_{12} - \Omega_{11}}{\sqrt{\Omega_{11}\Omega_{22} - \Omega_{12}^2}} \tag{18}$$

$$\psi_2 = \frac{\sqrt{\Omega_{11}}}{\sqrt{\Omega_{11}\Omega_{22} - \Omega_{12}^2}} \bar{K}_\Delta \tag{19}$$

where Ω_{ij} is the i,jth element of $\mathbf{\Omega}_n$, and $\widehat{\mathrm{LR}}_n(\bar{\kappa}_n)$ under the null hypothesis of $\beta_n = 0$ has the same distribution as

$$\min_{|g|\leq\psi_2}\begin{pmatrix} Z_1 \\ Z_2+g_0-g \end{pmatrix}'\begin{pmatrix} Z_1 \\ Z_2+g_0-g \end{pmatrix} - \min_{h,|g|\leq\psi_2}\begin{pmatrix} Z_1-h \\ Z_2+g_0-\psi_1 h-g \end{pmatrix}'\begin{pmatrix} Z_1-h \\ Z_2+g_0-\psi_1 h-g \end{pmatrix} \tag{20}$$

where $(Z_1, Z_2)' \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_2)$ and $g_0 = \sqrt{\Omega_{11}}\Delta_n/\sqrt{\Omega_{11}\Omega_{22}-\Omega_{12}^2}$. Let $\mathrm{cv}_\Omega(\boldsymbol{\psi})$, $\boldsymbol{\psi} = (\psi_1, \psi_2)$ be the corresponding critical value which ensures size control of the statistic in (20) for all values of $|g_0| \leq \psi_2$. In the replication files, we provide an look-up table for all possible values of $\boldsymbol{\psi}$. A subsequence argument then yields asymptotic validity of this feasible LR test $\hat{\varphi}_{\mathrm{LR},n}(\bar{\kappa}_n, \mathbf{Y}_n) = \mathbf{1}[\widehat{\mathrm{LR}}_n(\bar{\kappa}_n) > \mathrm{cv}_\Omega(\hat{\boldsymbol{\psi}}_n)]$, where $\hat{\boldsymbol{\psi}}_n = (\hat{\psi}_{n,1}, \hat{\psi}_{n,2})$ are as in (18) and (19), with the elements of $\mathbf{\Omega}_n$ replaced by those of $\hat{\mathbf{\Omega}}_n$.

Lemma 3 *(a) If* $\mathbf{\Omega}_n^{-1}\hat{\mathbf{\Omega}}_n \xrightarrow{p} \mathbf{I}_2$ *and (17) holds, then* $\limsup_{n\to\infty} E_{\theta_n}[\hat{\varphi}_{\mathrm{LR},n}(\bar{\kappa}_n, \mathbf{Y}_n)] \leq \alpha$ *for all sequences* θ_n *with* $\beta_n = 0$ *and* $|\Delta_n| \leq \rho_n\bar{\kappa}_n/\sqrt{\mathbf{x}_n'\mathbf{x}_n}$.

(b) Under the assumptions of Theorem 1 part (a), (b) or (c), $E_{\theta_n}[\hat{\varphi}_{\mathrm{LR},n}(\bar{\kappa}_n, \mathbf{Y}_n)] - E_{\theta_n}[\varphi_{\mathrm{LR},n}(\bar{\kappa}_n, \mathbf{Y}_n)] \to 0$.

Note that the asymptotic validity in part (a) holds without any assumptions about the sequences k_n or $\bar{\kappa}_n$. In particular, it is not required that $k_n/n \to c \in (0,1)$ or $\bar{\kappa}_n = o(n^{1/4})$. In the Gaussian homoskedastic model, $\mathbf{\Omega}_n$ is equal to $(\mathbf{x}_n'\mathbf{x}_n)^{-1}\mathbf{\Sigma}(\rho_n)$, and in large samples, $\hat{\varphi}_{\mathrm{LR},n}$ reduces to the bivariate LR test introduced in Section 2.2. Formally, part (b) of the Lemma shows that the large sample power properties of $\hat{\varphi}_{\mathrm{LR},n}$ in the Gaussian homoskedastic model are equal to the small sample power properties of the LR test as introduced in Section 2.2. Thus, $\hat{\varphi}_{\mathrm{LR},n}(\bar{\kappa}_n, \mathbf{Y}_n)$ has the same asymptotic efficiency properties as $\varphi_{\mathrm{LR},n}(\bar{\kappa}_n, \mathbf{Y}_n)$ discussed below Theorem 1, even among tests that depend on the data beyond the short and long regression coefficient.

Given Lemma 3, the only obstacle to a straightforward implementation of the LR test in a more general model is the estimation of the asymptotic variance $\mathbf{\Omega}_n$. If the number of controls k is fixed, or only slowly increasing with n, the usual heteroskedasticity robust White (1980) estimator for $\mathbf{\Omega}_n$ is consistent under reasonably weak assumptions. However, under asymptotics where $k_n/n \to c \in (0,1)$, as employed for the asymptotic efficiency argument in Theorem 1, Cattaneo, Jansson, and Newey (forthcoming) show that the White (1980) estimator is no longer consistent, and Cattaneo, Jansson, and Newey (2015) provide an alternative estimator that remains consistent. Alternatively, if one considers a not-too-large bound $\bar{\kappa}_n = o(n)$, one may also consistently estimate $\hat{\mathbf{\Omega}}_n$ from the usual White formula based on the residuals from the short regression that only includes the baseline controls. This has the advantage of being readily implementable also with clustering. We provide a corresponding result in Appendix A.6.

4 Small Sample Simulations

In this section, we use Monte Carlo simulations to evaluate the finite-sample size and power properties of $\hat{\varphi}_{\mathrm{LR}}$, and we compare it to the performance of the LASSO-based post-double-selection technique of Belloni, Chernozhukov, and Hansen (2014) (abbreviated BCH in the following two sections).

We generate data from the extended linear model (6) with $m = 1$ and $q_{1,i} = 1$, that is

$$y_i^e = \delta_1 + x_i^e \beta + \sum_{j=1}^{k} \gamma_j z_{i,j}^e + \varepsilon_i^e, \quad , i = 1, \ldots, n+1$$

so that the single ($m = 1$) unconstrained regressor is a constant δ_1. We let $\varepsilon_i^e \sim iid\mathcal{N}(0, 1)$ independent of the regressors throughout. As in BCH's Monte Carlo, we set the total number of observations to $n + 1 = 500$, let $k = 200$, and generate x_i^e by a linear model from the controls $z_{i,j}^e$

$$x_i^e = \sum_{j=1}^{k} \mu_j z_{i,j}^e + \varepsilon_i^x. \tag{21}$$

Our designs vary according to the value of four parameters: the previously introduced $\rho^2 \in \{0.7, 0.95\}$ and $\kappa \in \{5, 10\}$; the scalar $\eta \in \{0.1, 0.3\}$ determines the degree of sparsity of γ_j and μ_j; and $\nu \in \{0, 0.5, 1\}$ determines the overlap between the non-zero indices of γ_j and μ_j. Specifically, the data is generated as follows:

1. $(z_{i,1}^e, \ldots, z_{i,k}^e, \varepsilon_i^x)' \sim iid\mathcal{N}(\mathbf{0}, \mathbf{I}_{k+1})$ independent of $\{\varepsilon_i^e\}$, $i = 1, \ldots, 500$.
2. $\gamma_j = c_\gamma \mathbf{1}[j \leq \lfloor \eta k \rfloor]$, where the scalar c_γ is chosen such that the implied value of $\kappa^2 = \sum_{i=1}^{n+1} \left(\sum_{j=1}^{k} \gamma_j \tilde{z}_{i,j} \right)^2$ with $\tilde{z}_{i,j} = z_{i,j}^e - (n+1)^{-1} \sum_{l=1}^{n+1} z_{l,j}^e$ is equal to the specified value.
3. $\mu_j = c_\mu \mathbf{1}[\lfloor \eta(1-\nu)k \rfloor + 1 \leq j \leq \lfloor \eta \nu k \rfloor]$, $j = 1, \ldots, k$, where $c_\mu \in \mathbb{R}$ is chosen such that the sample R^2 of a regression of $\tilde{x}_i = x_i^e - (n+1)^{-1} \sum_{l=1}^{n+1} x_l^e$ on $\{z_{i,j}^e\}_{j=1}^{k}$ is equal to ρ^2.

4. $\beta = 0$ under H_0 and $\beta = 0.06$ under H_1.

With $n+1 = 500$, the values $\kappa \in \{5, 10\}$ correspond to a population R^2 of a regression of y_i^e on $\{z_{i,j}^e\}_{j=1}^k$ of approximately $\{0.05, 0.16\}$, conditional on $\{z_{i,j}^e\}_{j=1}^k$. The parameter η plays a crucial role for the BCH method, since the method requires that the number of non-zero values of γ_j and μ_j is not too large. In contrast, the test $\hat{\varphi}_{\mathrm{LR}}$ remains numerically invariant to any linear reparameterizations of the regressors.[3] Finally, the parameter ν determines the omitted variable bias in the short regression coefficient $\hat{\beta}^{\mathrm{short}}$ (which is the coefficient on x_i^e in the regression of y_i^e on $(1, x_i^e)$). Under $\nu = 0$, there is no overlap, and the variables $z_{i,j}^e$ with non-zero coefficient γ_j are uncorrelated with the regressor of interest x_i^e, so there is no omitted variable bias, at least over repeated samples with random regressors. In the other extreme, with $\nu = 1$, every variable $z_{i,j}^e$ with non-zero coefficient γ_j is correlated with x_i^e, leading to a large omitted variable bias.

We consider four types of tests of the null hypothesis $H_0 : \beta = 0$. First, inference based on the t-test associated with the short regression coefficient $\hat{\beta}^{\mathrm{short}}$. Second, inference based on the t-test associated with the long regression coefficient $\hat{\beta}^{\mathrm{long}}$. Third, the feasible test $\hat{\varphi}_{\mathrm{LR}}$ introduced in the last subsection, where we set $\bar{\kappa}$ equal to the two values $\{5, 10\}$. Fourth, the LASSO-based post-double-selection method "LPDS" from BCH, as specified in their Monte Carlo Section 4.2. For the first three types of methods, we estimate standard errors of $(\hat{\beta}^{\mathrm{long}}, \hat{\beta}^{\mathrm{short}})'$ with the heteroskedasticity-robust estimator of Cattaneo, Jansson, and Newey (2015).

Table 2 contains the results. Inference based on $\hat{\beta}^{\mathrm{short}}$, $\hat{\beta}^{\mathrm{long}}$ and $\hat{\varphi}_{\mathrm{LR}}$ is asymptotically valid conditional on the realization of the regressors. In contrast, the BCH method relies on the randomness of the regressors for its asymptotic validity, so correspondingly, we report unconditional rejection probabilities. In unreported results, we computed conditional rejection probabilities (computed from 100 independent draws of the regressors) for the first three methods, but found limited variation, especially under the null hypothesis. For simplicity, we therefore simply compute unconditional rejection probabilities for all considered methods.

The t-test based on $\hat{\beta}^{\mathrm{short}}$ is substantially oversized whenever the overlap parameter ν is positive. This shows that the considered values of κ are large enough to severely distort inference that simply sets the control coefficients to zero. In contrast, the t-test associated

[3] Given $\{z_{i,j}\}$ and κ, $\hat{\varphi}_{\mathrm{LR}}$ is numerically invariant to the sparsity of $\{\gamma_j\}$. But under (21), the sparsity of $\{\mu_j\}$ determines the realization of $\{x_j^e\}$ for given $\{\varepsilon_j^x\}$, which affects the distribution of $\hat{\varphi}_{\mathrm{LR}}$ through the realization of heteroskedasticity-robust estimators $\hat{\mathbf{\Omega}}_n$.

Table 2: Small Sample Rejection Probabilities

			$\hat{\beta}^{\text{short}}$		$\hat{\beta}^{\text{long}}$		$\hat{\varphi}_{\text{LR}}(5, \mathbf{Y})$		$\hat{\varphi}_{\text{LR}}(10, \mathbf{Y})$		LPDS	
		β	0.00	0.06	0.00	0.06	0.00	0.06	0.00	0.06	0.00	0.06
η	ν	ρ					$\kappa = 5$					
0.10	0.00	0.70	0.06	0.49	0.06	0.19	0.05	0.19	0.05	0.19	0.05	0.28
0.10	0.50	0.70	0.44	0.96	0.06	0.19	0.05	0.19	0.06	0.19	0.08	0.43
0.10	1.00	0.70	0.95	1.00	0.05	0.19	0.04	0.36	0.05	0.19	0.13	0.58
0.10	0.00	0.95	0.06	1.00	0.06	0.20	0.03	0.22	0.06	0.19	0.05	0.26
0.10	0.50	0.95	0.68	1.00	0.06	0.20	0.03	0.69	0.06	0.19	0.05	0.26
0.10	1.00	0.95	1.00	1.00	0.06	0.20	0.06	1.00	0.04	0.22	0.05	0.27
0.30	0.00	0.70	0.06	0.50	0.05	0.19	0.05	0.19	0.05	0.19	0.05	0.42
0.30	0.50	0.70	0.45	0.96	0.05	0.19	0.05	0.19	0.05	0.19	0.32	0.90
0.30	1.00	0.70	0.95	1.00	0.06	0.19	0.04	0.36	0.06	0.19	0.84	1.00
0.30	0.00	0.95	0.06	1.00	0.06	0.19	0.03	0.21	0.06	0.19	0.05	0.53
0.30	0.50	0.95	0.69	1.00	0.06	0.19	0.03	0.70	0.06	0.19	0.15	0.71
0.30	1.00	0.95	1.00	1.00	0.05	0.19	0.06	1.00	0.04	0.21	0.38	0.80
η	ν	ρ					$\kappa = 10$					
0.10	0.00	0.70	0.08	0.49	0.05	0.20	0.05	0.20	0.05	0.20	0.05	0.26
0.10	0.50	0.70	0.94	1.00	0.05	0.19	0.04	0.37	0.05	0.19	0.12	0.51
0.10	1.00	0.70	1.00	1.00	0.05	0.20	0.83	1.00	0.04	0.25	0.32	0.76
0.10	0.00	0.95	0.08	0.99	0.05	0.19	0.03	0.23	0.05	0.19	0.05	0.23
0.10	0.50	0.95	1.00	1.00	0.05	0.19	0.06	0.99	0.04	0.21	0.05	0.25
0.10	1.00	0.95	1.00	1.00	0.05	0.19	1.00	1.00	0.05	1.00	0.05	0.26
0.30	0.00	0.70	0.08	0.50	0.06	0.19	0.06	0.19	0.06	0.19	0.05	0.37
0.30	0.50	0.70	0.94	1.00	0.05	0.19	0.04	0.37	0.05	0.19	0.77	0.99
0.30	1.00	0.70	1.00	1.00	0.05	0.19	0.83	1.00	0.04	0.25	1.00	1.00
0.30	0.00	0.95	0.08	0.99	0.05	0.19	0.03	0.22	0.05	0.19	0.05	0.48
0.30	0.50	0.95	1.00	1.00	0.05	0.19	0.06	0.99	0.04	0.21	0.35	0.78
0.30	1.00	0.95	1.00	1.00	0.06	0.19	1.00	1.00	0.05	1.00	0.67	0.87

Notes: Entries are rejection probabilities of nominal 5% level tests of $H_0 : \beta = 0$. Rows correspond to different DGPs, with η measuring the sparsity of the design, ν the overlap between the non-zero indices on $z^e_{i,j}$ in the regressions of y^e_i on $z^e_{i,j}$ and of $x^e_{i,j}$ on $z_{i,j}$, and ρ^2 is R^2 of a regression of the demeaned values of x^e_i on $z^e_{i,j}$. The columns are different tests, with $\hat{\beta}^{\text{short}}$ and $\hat{\beta}^{\text{long}}$ the t-statistics associated with the short and long regressions, $\hat{\varphi}_{\text{LR}}(\bar{\kappa}, \mathbf{Y})$ is the test developed in this paper that imposes the bound $\kappa \leq \bar{\kappa}$, and LPDS is BCH's LASSO-based post-double-selection procedure. Based on 20,000 Monte Carlo simulations.

with $\hat{\beta}^{\text{long}}$ has size very close to the nominal level throughout, but at the cost of fairly low power. Our tests $\hat{\varphi}_{\text{LR}}(\bar{\kappa}, \mathbf{Y})$ with $\bar{\kappa} \in \{5, 10\}$ control size well whenever $\kappa \leq \bar{\kappa}$, and have higher power than the t-test based on $\hat{\beta}^{\text{long}}$ when $\nu = 0$, with larger gains for larger values of ρ, as expected. The test that imposes the smaller bound $\bar{\kappa} = 5$, $\hat{\varphi}_{\text{LR}}(5, \mathbf{Y})$, has close to nominal null rejection probability even when $\kappa = 10$ as long as $\nu \leq 0.5$, and it has higher power than the t-test based on $\hat{\beta}^{\text{long}}$ when $\nu = 0.5$. This is an example where the true value of κ is larger than the imposed bound; rejections under the alternative are then still desirable, as they lead to shorter sets of values of $\bar{\kappa}$ that are deemed compatible with the null hypothesis of $H_0 : \beta = 0$. The LPDS method is sometimes oversized even in the relatively sparse design with $\eta = 0.1$, especially if $\kappa = 5$. Apparently, the relatively small values of γ_j make it difficult for the method to correctly pick up the $\eta \cdot k = 20$ non-zero coefficients, leading to a remaining omitted variable bias that is large enough to induce non-negligible overrejections. When $\kappa = 10$ and $\rho = 0.95$, so that both γ_j and μ_j are relatively larger, the LPDS method reliably controls size in the $\eta = 0.1$ sparse design, and leads to moderate power gains over the t-test based on $\hat{\beta}^{\text{long}}$.

Looking over the table, it is tempting to conclude that one should use $\hat{\beta}^{\text{short}}$ whenever there is no overlap, $\nu = 0$, as this leads to the highest power by far, and only slight size distortions (or, related, to use $\hat{\varphi}_{\text{LR}}(5, \mathbf{Y})$ even when $\kappa = 10$ when $\nu = 0.5$). However, it is not possible to consistently determine the value of ν from the observations. This is the result of the asymptotic efficiency derivations in Section 2.3: With $\bar{\kappa}$ fixed or slowly increasing, it is impossible to do better than to construct tests based on the bivariate statistics $(\hat{\beta}^{\text{long}}, \hat{\beta}^{\text{short}})'$ at least in large samples, and our test $\hat{\varphi}_{\text{LR}}(\bar{\kappa}, \mathbf{Y})$ comes close to exploiting the information contained in this pair of statistics under the constraint $\kappa \leq \bar{\kappa}$.

5 Empirical Application

The empirical exercise of this paper is based on a panel study of 48 states from 1985 to 1997 by Donohue III and Levitt (2001), which is also considered in BCH as an empirical example of their methodology. In regressions of crime rates on lagged abortion rates, Donohue and Levitt find evidence for negative causal effects of abortion on crime, but these results were disputed in follow-up studies (see, for instance, Foote and Goetz (2008), Joyce (2004, 2009) and BCH). Here, we ask which *a priori* assumption on an expansive set of potential controls lead to a significant effect.

We apply the same specification as in BCH: In three separate regressions corresponding

Table 3: Empirical effect of abortion on crime

	Violent Crime	Property Crime	Murder
$\hat{\beta}^{\text{short}}$	-0.159^*	-0.121^*	-0.201^*
	(0.033)	(0.023)	(0.069)
$\hat{\beta}^{\text{long}}$	0.013	-0.195	2.34
	(0.674)	(0.212)	(2.00)
LPDS	-0.082	-0.031	-0.068
	(0.106)	(0.057)	(0.200)
$\bar{\kappa}^*$	0.134	0.143	0.074
$\text{SSR}^{\text{short}}$	2.78	1.18	36.9
$(\bar{\kappa}^*)^2/\text{SSR}^{\text{short}}$	0.65%	1.7%	0.015%

Notes: Standard errors in parentheses are clustered at the state level, and for $\hat{\beta}^{\text{short}}$, $\hat{\beta}^{\text{long}}$ and in the computation of $\bar{\kappa}^*$, estimated using short regression residuals.

to crime types {violent, property, murder}, let all variables, implicitly indexed by crime type, follow the relationship

$$y_{i,t}^e = x_{i,t}^e \beta + \sum_{j=1}^{k} \gamma_j z_{i,t,j}^e + \sum_{j=1}^{m} \delta_j q_{i,t,j}^e + \varepsilon_{i,t}^e, \tag{22}$$

where $i = 1, \ldots, 48$ denotes states, $t = 1, \ldots, 12$ denotes year, and all variables are expressed in first differences to account for state fixed-effects. Among the observables, $y_{i,t}^e$ is a measure of crime, $x_{i,t}^e$ is a measure of lagged abortion rates, $q_{i,t,j}^e$ is a set of 20 controls (including 12 time dummies) present in Donohue and Levitt's original specification, and $z_{i,t,j}^e$ is a set of $k = 276$ potential controls proposed by BCH. The coefficient of interest is β. We view $q_{i,t,j}^e$ as a set of baseline controls that are always included. The short regression with estimated coefficient $\hat{\beta}^{\text{short}}$ thus uses only Donohue and Levitt's choice of controls, while the long regression also includes the additional controls from BCH. Using standard errors clustered at the state level, estimated using short regression residuals (see Appendix A.6), the t-tests based on $\hat{\beta}^{\text{short}}$ rejects for all three crime types, but the t-test using the estimator $\hat{\beta}^{\text{long}}$ from the long regression does not. The LASSO-based post-double-selection of BCH, constrained to include the baseline controls, fails to reject for all three crime types (see Table 3).

Let $\tilde{z}_{i,t,j}$ be the residuals from a regression of $z_{i,t,j}^e$ on $q_{i,t,j}^e$. The new test $\hat{\varphi}_{\text{LR}}$ developed here improves inference about β under the assumption that $\kappa^2 =$

$\sum_{i=1}^{48} \sum_{t=1}^{12} \left(\sum_{j=1}^{k} \gamma_j \tilde{z}_{i,t,j} \right)^2 \leq \bar{\kappa}^2$, for some given bound $\bar{\kappa}$. In Table 3, we report the threshold value $\bar{\kappa}^*$ such that our 5% level test rejects for all $\bar{\kappa} < \bar{\kappa}^*$, and it does not for all $\bar{\kappa} > \bar{\kappa}^*$. To put these values into perspective, we also report the sum of squared residuals from the short regression, $\mathrm{SSR}^{\mathrm{short}}$, and the ratio $(\bar{\kappa}^*)^2/\mathrm{SSR}^{\mathrm{short}}$. The resulting ratios are always smaller than 1.7%, often substantially so. This quantifies the sensitivity of Donohue and Levitt's results with respect to the additional controls $z^e_{i,t,j}$: The coefficient of interest β loses significance unless the explanatory power of $z^e_{i,t,j}$ is assumed to be very small. This empirical finding of small threshold values $\bar{\kappa}^*$ also underlines the relevance of the asymptotic efficiency results in Section 2.3 derived under the assumption that $\bar{\kappa}$ is not large.

6 Conclusion

Improving inference over tests based on the "long regression" estimator requires some *a priori* knowledge about the control coefficients. In this paper, we explore inference under a bound on the weighted sum of squared control coefficients, which corresponds to a limit of the explanatory power of the controls. This is a substantively different assumption from a limit of the number of non-zero coefficients. Potentially attractive features of our method are the straightforward interpretation of the central condition, and the invariance to linear reparameterizations. At the same time, the sparsity-based technique of Belloni, Chernozhukov, and Hansen (2014) allows for cases where the number of potential controls is larger than the number of observations, which our approach does not accommodate.

It might be interesting to investigate generalizations to instrumental variable regressions, where the validity of the instruments requires a large number of control variables. Another generalization could consider inference about vector valued parameters of interest. We leave those to future research.

A Appendix

A.1 Proof of Lemma 1

By sufficiency, we may without loss of generality assume that $\varphi(\mathbf{Y})$ is a function of $(\hat{\beta}^{\text{long}}, \hat{\beta}^{\text{short}}, \hat{\phi}')$. Thus, with $\mathbf{O}$ a $(k-1)\times(k-1)$ rotation matrix

$$\begin{aligned}
E_{\beta,\Delta,\tau,\boldsymbol{\omega}}[\varphi(\mathbf{Y})] &= E_{\beta,\Delta,\tau,\boldsymbol{\omega}}[\varphi((\hat{\beta}^{\text{long}}, \hat{\beta}^{\text{short}}, \hat{\boldsymbol{\phi}}'))] \\
&= E_{\beta,\Delta,\tau}[\varphi((\hat{\beta}^{\text{long}}, \hat{\beta}^{\text{short}}, (\tau\boldsymbol{\omega}+\mathbf{e})'))] \\
&= E_{\beta,\Delta,\tau}[\varphi((\hat{\beta}^{\text{long}}, \hat{\beta}^{\text{short}}, (\tau\boldsymbol{\omega}+\mathbf{Oe})'))] \\
&= E_{\beta,\Delta,\tau}[\varphi((\hat{\beta}^{\text{long}}, \hat{\beta}^{\text{short}}, (\tau\mathbf{O}\boldsymbol{\omega}+\mathbf{Oe})'))]
\end{aligned}$$

where $\mathbf{e}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_{k-1})$ is independent of $(\hat{\beta}^{\text{long}}, \hat{\beta}^{\text{short}})$, the before last equality follows from the spherical symmetry of the distribution of $\mathbf{e}$, and the last equality from the assumption about φ. Since $\mathbf{O}$ was arbitrary, we also have

$$E_{\beta,\Delta,\tau,\boldsymbol{\omega}}[\varphi(\mathbf{Y})] = \int E_{\beta,\Delta,\tau}[\varphi((\hat{\beta}^{\text{long}}, \hat{\beta}^{\text{short}}, (\mathbf{O}\tau\boldsymbol{\omega}+\mathbf{Oe})'))]dH_{k-1}(\mathbf{O})$$

where H_{k-1} is the Haar measure on the $k-1$ rotation matrices. Now set $\tilde{\varphi}(\mathbf{T}) = \int\varphi((\hat{\beta}^{\text{long}}, \hat{\beta}^{\text{short}}, \hat{\tau}\iota_{k-1}'\mathbf{O}))dH_{k-1}(\mathbf{O}) = \int\varphi((\hat{\beta}^{\text{long}}, \hat{\beta}^{\text{short}}, \hat{\boldsymbol{\phi}}'\mathbf{O}))dH_{k-1}(\mathbf{O})$, with ι an arbitrary fixed vector of unit length in $\mathbb{R}^{k-1}$. Then

$$\begin{aligned}
E_{\beta,\Delta,\tau}[\tilde{\varphi}(\mathbf{T})] &= E_{\beta,\Delta,\tau}[\int\varphi((\hat{\beta}^{\text{long}}, \hat{\beta}^{\text{short}}, \hat{\boldsymbol{\phi}}'\mathbf{O}'))dH_{k-1}(\mathbf{O})] \\
&= \int E_{\beta,\Delta,\tau,\boldsymbol{\omega}}[\varphi((\hat{\beta}^{\text{long}}, \hat{\beta}^{\text{short}}, \hat{\boldsymbol{\phi}}'\mathbf{O}'))]dH_{k-1}(\mathbf{O}) \\
&= \int E_{\beta,\Delta,\tau}[\varphi((\hat{\beta}^{\text{long}}, \hat{\beta}^{\text{short}}, (\tau\mathbf{O}\boldsymbol{\omega}+\mathbf{Oe})'))]dH_{k-1}(\mathbf{O})
\end{aligned}$$

and the result follows.

A.2 Proof of Lemma 2

We will make use of the following Lemma.

Lemma 4 *Let $\mathcal{I}_m$ denote the modified Bessel function of the first kind of degree $m>0$. Then for any positive sequence $s_m = o(m^{1/2})$,*

$$\lim_{m\to\infty}\mathcal{I}_m(s_m)\frac{\Gamma(m+1)}{(\frac{1}{2}s_m)^m} = 1$$

where Γ is the Gamma function.

Proof. From the definition of $\mathcal{I}_m$,

$$
\begin{aligned}
\mathcal{I}_m(s) &= (\tfrac{1}{2}s)^m \sum_{j=0}^{\infty} \frac{(\frac{1}{4}s^2)^j}{j!\Gamma(m+j+1)} \\
&= \frac{(\frac{1}{2}s)^m}{\Gamma(m+1)} \left(1 + \sum_{j=1}^{\infty} \frac{(\frac{1}{4}s^2)^j}{j!} \frac{\Gamma(m+1)}{\Gamma(m+j+1)}\right).
\end{aligned}
$$

Now

$$
\begin{aligned}
\sum_{j=1}^{\infty} \frac{(\frac{1}{4}s^2)^j}{j!} \frac{\Gamma(m+1)}{\Gamma(m+j+1)} &\leq \sum_{j=1}^{\infty} \frac{s^{2j}}{j!} \frac{\Gamma(m+1)}{\Gamma(m+j+1)} \\
&\leq \sum_{j=1}^{\infty} \frac{(s^2/m)^j}{j!} = \exp[s^2/m] - 1
\end{aligned}
$$

where the second inequality uses the elementary inequality $\Gamma(m+1)m^j/\Gamma(m+j+1) \leq 1$ obtained from reapeatedly applying $\Gamma(m+i+1) = (m+i)\Gamma(m+i) \leq m\Gamma(m+i)$ for all $i \geq 0$ and $m > 0$. The result now follows from $s_m^2/m \to 0$ under $s_m = o(m^{1/2})$. ■

For ease of notation, we omit the dependence on n (and $k = k_n$), except for t_n. From (16), it follows that $\hat{\tau}^2 = \hat{\boldsymbol{\phi}}'\hat{\boldsymbol{\phi}}$ with $\hat{\boldsymbol{\phi}} \sim \mathcal{N}(\tau\boldsymbol{\omega}, \mathbf{I}_{k-1})$ is distributed non-central χ^2 with $k-1$ degrees of freedom and non-centrality parameter τ^2. Without loss of generality, assume $\boldsymbol{\omega} = \boldsymbol{\iota} = (1, 0, \ldots, 0)'$. Then, with $\hat{\boldsymbol{\omega}} = \hat{\boldsymbol{\phi}}/\|\hat{\boldsymbol{\phi}}\|$, from the density of $\hat{\boldsymbol{\phi}}$,

$$
L_n(t) = C \int \exp[-\tfrac{1}{2}||\hat{\tau}\mathbf{O}\hat{\boldsymbol{\omega}} - \boldsymbol{\iota}t||^2] dH_{k-1}(\mathbf{O})
$$

for some constant C that does not depend on t (and note that $L_n(t)$ does not depend on the realization of $\hat{\boldsymbol{\omega}}$). Thus

$$
L_n(t)/L_n(0) = \int \exp[\hat{\tau}\hat{\boldsymbol{\omega}}'\mathbf{O}\boldsymbol{\iota}t - \tfrac{1}{2}t^2] dH_{k-1}(\mathbf{O}).
$$

We initially show the convergence under $\tau = 0$. It then suffices to show that $E[(L_n(t_n)/L_n(0) - 1)^2] \to 1$ under $\hat{\boldsymbol{\phi}} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{k-1})$ and an arbitrary sequence $t_n = o(n^{1/4})$. Observe that

$$
\begin{aligned}
& E\left[(L_n(t_n)/L_n(0) - 1)^2\right] \\
= \;& E\left[\left(\int \exp[t_n\hat{\boldsymbol{\phi}}'\mathbf{O}\boldsymbol{\iota} - \tfrac{1}{2}t_n^2]\, dH_{k-1}(\mathbf{O}) - 1\right)^2\right] \\
= \;& E\left[\left(\int \exp[t_n\hat{\boldsymbol{\phi}}'\mathbf{O}\boldsymbol{\iota} - \tfrac{1}{2}t_n^2]\, dH_{k-1}(\mathbf{O}) - 1\right)\left(\int \exp[t_n\hat{\boldsymbol{\phi}}'\tilde{\mathbf{O}}\boldsymbol{\iota} - \tfrac{1}{2}t_n^2]\, dH_{k-1}(\tilde{\mathbf{O}}) - 1\right)\right]
\end{aligned}
$$

$$
\begin{aligned}
&= E\left[\left(\int \exp[t_n\hat{\boldsymbol{\phi}}'\mathbf{O}\boldsymbol{\iota} - \tfrac{1}{2}t_n^2]\, dH_{k-1}(\mathbf{O})\right)\left(\int \exp[t_n\hat{\boldsymbol{\phi}}'\tilde{\mathbf{O}}\boldsymbol{\iota} - \tfrac{1}{2}t_n^2]\, dH_{k-1}(\tilde{\mathbf{O}})\right)\right] \\
&\quad -2\cdot E\left[\int \exp[t_n\hat{\boldsymbol{\phi}}'\mathbf{O}\boldsymbol{\iota} - \tfrac{1}{2}t_n^2]\, dH_{k-1}(\mathbf{O})\right] + 1 \\
&= h_n(t_n) - 2\tilde{h}_n(t_n) + 1
\end{aligned}
$$

In what follows, we show that $h_n(t_n) \to 1$. The convergence $\tilde{h}_n(t_n) \to 1$ follows from the same arguments and is omitted for brevity.

Tonelli's Theorem and $\hat{\boldsymbol{\phi}} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_{k-1})$ imply

$$
\begin{aligned}
h_n(t_n) &= E\left[\int\int \exp[t_n\hat{\boldsymbol{\phi}}'\mathbf{O}\boldsymbol{\iota} + t_n\hat{\boldsymbol{\phi}}'\tilde{\mathbf{O}}\boldsymbol{\iota} - t_n^2]\, dH_{k-1}(\mathbf{O})\, dH_{k-1}(\tilde{\mathbf{O}})\right] \\
&= \int\int E\left[\exp[t_n\hat{\boldsymbol{\phi}}'(\mathbf{O}\boldsymbol{\iota} + \tilde{\mathbf{O}}\boldsymbol{\iota}) - t_n^2]\right] dH_{k-1}(\mathbf{O})\, dH_{k-1}(\tilde{\mathbf{O}}) \\
&= \int\int \exp[\tfrac{1}{2}t_n^2\|\mathbf{O}\boldsymbol{\iota} + \tilde{\mathbf{O}}\boldsymbol{\iota}\|^2 - t_n^2]\, dH_{k-1}(\mathbf{O})\, dH_{k-1}(\tilde{\mathbf{O}}) \\
&= \int\int \exp[t_n^2(\mathbf{O}\boldsymbol{\iota})'\tilde{\mathbf{O}}\boldsymbol{\iota}]\, dH_{k-1}(\mathbf{O})\, dH_{k-1}(\tilde{\mathbf{O}}) \\
&= \int \exp[t_n^2\boldsymbol{\iota}'\mathbf{O}\boldsymbol{\iota}]\, dH_{k-1}(\mathbf{O}).
\end{aligned}
$$

Using the notation of Lemma 4, the formula for the normalizing constant of the von Mises-Fisher distribution (see, for instance, equation (9.3.4) of Mardia and Jupp (2000)) implies $h_n(t_n) = 2^{\tilde{k}/2-1}\cdot\mathcal{I}_{\tilde{k}/2-1}(t_n^2)\cdot\Gamma(\tilde{k}/2)/t_n^{2(\tilde{k}/2-1)}$, where $\tilde{k} = k-1$. Application of Lemma 4 with $s_n = t_n^2$ now yields $h_n(t_n) \to 1$, since under $t_n = o(n^{1/4})$ and $k/n \to c \in (0,1)$, $s_n^2 = t_n^4 = o(\tilde{k}/2-1)$.

This concludes the proof under $\tau_n = 0$. Now apply this very result to another sequence t_n, $t_n = t_n'$. Then $L_n(t_n')/L_n(0) \xrightarrow{p} 1$ implies via LeCam's first lemma (see, for instance, Lemma 6.4 in van der Vaart (1998)) that in the experiment of observing $\hat{\tau}_n^2$, the sequence $\tau_n = t_n'$ is contiguous to $\tau_n = 0$. Thus, $L_n(t_n)/L_n(0) \xrightarrow{p} 1$ also holds under $\tau_n = t_n' = o(n^{1/4})$ by definition of contiguity, which was to be shown.

A.3 Proof of Theorem 1

We prove a slightly more general Lemma that implies Theorem 1.

Theorem 2 *Consider a sequence of observations from the linear regression model with $\varepsilon_i \sim iid\mathcal{N}(0,1)$ where $k_n/n \to c \in (0,1)$ and $\rho_n^2 \to \rho^2 \in [0,1)$, and let $\varphi_n(\bar{\kappa}_n, \mathbf{Y}_n)$ be a sequence of tests that, for all sufficiently large n, satisfy the assumption of Lemma 1. If for some sequence $\tilde{s}_n$ and all $(b,\tilde{a}) \in \mathbb{R}^2$, $E_{\theta_n}[\varphi_n(\bar{\kappa}_n, \mathbf{Y}_n)]$ converges along a sequence θ_n with $(\sqrt{\mathbf{x}_n'\mathbf{x}_n}\beta_n, \sqrt{\mathbf{x}_n'\mathbf{x}_n}\Delta_n - \tilde{s}_n) = (b,\tilde{a})$ and $\tau_n = o(n^{1/4})$ (where τ_n may depend on $(b,\tilde{a})$), then*

$\lim_{n\to\infty} E_{\theta_n}[\varphi_n(\bar{\kappa}_n, \mathbf{Y}_n)] = E_{b,\tilde{a}}[\varphi(\tilde{\mathbf{W}})]$ *for some test* $\varphi : \mathbb{R}^2 \mapsto [0,1]$ *with*

$$\tilde{\mathbf{W}} = \begin{pmatrix} \tilde{W}_1 \\ \tilde{W}_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} b \\ b+\rho\tilde{a} \end{pmatrix}, \mathbf{\Sigma}(\rho)\right),\ \mathbf{\Sigma}(\rho) = \begin{pmatrix} \frac{1}{1-\rho^2} & 1 \\ 1 & 1 \end{pmatrix},$$

and $\lim_{n\to\infty} E_{\theta_n}[\varphi_n(\bar{\kappa}_n, \mathbf{Y}_n)] = E_{b,\tilde{a}}[\varphi(\tilde{\mathbf{W}})]$ *holds under* all *sequences* θ_n *with* $(\sqrt{\mathbf{x}_n'\mathbf{x}_n}\beta_n, \sqrt{\mathbf{x}_n'\mathbf{x}_n}\Delta_n - \tilde{s}_n) = (b, \tilde{a})$ *and* $\tau_n = o(n^{1/4})$.

Proof. Let ℓ_n be the likelihood ratio statistic based on $\mathbf{Y}_n$ of testing $H_0 : (b, \tilde{a}, \tau_n) = (b_0, \tilde{a}_0, \tau_{n,0})$ against $H_1 : (b, \tilde{a}, \tau_n) = (b_1, \tilde{a}_1, \tau_{n,1})$. Let $h_{j,n} = (b_j, b_j + \rho_n\tilde{a}_j)'$ and $h_j = (b_j, b_j + \rho\tilde{a}_j)$, $j = 0, 1$. From (16),

$$\begin{aligned}\log \ell_n &= \sqrt{\mathbf{x}_n'\mathbf{x}_n}\begin{pmatrix} \hat{\beta}_n^{\text{long}} \\ \hat{\beta}_n^{\text{short}} - \tilde{s}_n \end{pmatrix}' \mathbf{\Sigma}(\rho_n)^{-1}(h_{1,n} - h_{0,n}) - \tfrac{1}{2}h_{1,n}'\mathbf{\Sigma}(\rho_n)^{-1}h_{1,n} \\ &\quad + \tfrac{1}{2}h_{0,n}'\mathbf{\Sigma}(\rho_n)^{-1}h_{0,n} + \log\left(\frac{L_n(\tau_{n,1})}{L_n(\tau_{n,0})}\right)\end{aligned}$$

and with $\ell(b, \tilde{a})$ the likelihood ratio statistic based on $\tilde{\mathbf{W}}$ of of testing $H_0 : (b, \tilde{a}) = (0, 0)$ against $H_1 : (b, \tilde{a}) = (b_1, \tilde{a}_1)$,

$$\log \ell = \begin{pmatrix} \tilde{W}_1 \\ \tilde{W}_2 \end{pmatrix}' \mathbf{\Sigma}(\rho)^{-1}(h_1 - h_0) - \tfrac{1}{2}h_1'\mathbf{\Sigma}(\rho)^{-1}h_1 + \tfrac{1}{2}h_0'\mathbf{\Sigma}(\rho)^{-1}h_0.$$

By Lemma 2,

$$\frac{L_n(\tau_{n,1})}{L_n(\tau_{n,0})} = \frac{L_n(\tau_{n,1})}{L_n(0)}\frac{L_n(0)}{L_n(\tau_{n,0})} \xrightarrow{p} 1$$

and from $\rho_n \to \rho$, $(\hat{\beta}_n^{\text{long}}, \hat{\beta}_n^{\text{short}})' \Rightarrow \tilde{\mathbf{W}}$ and $h_{j,n} \to h_j$ for $j = 0, 1$. Thus, under H_0, $\ell_n \Rightarrow \ell$. This straightforwardly extends more generally to $\{\ell_n\}_{(b,\tilde{a})\in H} \Rightarrow \{\ell\}_{(b,\tilde{a})\in H}$ for any finite $H \subset \mathbb{R}^2$. Thus, by Definition 9.1 in van der Vaart (1991), under the assumptions of the Lemma, the sequence of experiments of observing $\mathbf{Y}_n$ with local parameter space $(b, \tilde{a}) \in \mathbb{R}^2$ converges to the experiment of observing $\tilde{\mathbf{W}}$. The first claim now follows from Theorem 15.1 in van der Vaart (1991).

For the second claim, for given $(b, \tilde{a})$, suppose $E_{\theta_n}[\varphi_n(\bar{\kappa}_n, \mathbf{Y}_n)] \to E_{b,\tilde{a}}[\varphi(\tilde{\mathbf{W}})]$ along $\theta_n = \theta_{n,1}$ with $\tau_n = \tau_{n,1} = o(n^{1/4})$. Let $\tau_{n,2} = o(n^{1/4})$ be another sequence, and denote $\theta_{n,2}$ the corresponding sequence of θ. Suppose $E_{\theta_{n,2}}[\varphi_n(\bar{\kappa}_n, \mathbf{Y}_n)]$ does not converge to $E_{b,\tilde{a}}[\varphi(\tilde{\mathbf{W}})]$. By Prohorov's Theorem (see, for instance, Theorem 2.4 in van der Vaart (1998)) and $0 \le \varphi_n(\bar{\kappa}_n, \mathbf{Y}_n) \le 1$, there exists a subsequence of n such that $\varphi_n(\bar{\kappa}_n, \mathbf{Y}_n)$ converges in distribution along that subsequence. Furthermore, by Lemma 2, the likelihood ratio statistic between the corresponding sequences $\theta_{n,1}$ and $\theta_{n,2}$ with identical values of $(b, \tilde{a})$ converges in probability to one, and this convergence automatically holds jointly with $\varphi_n(\bar{\kappa}_n, \mathbf{Y}_n)$ along the subsequence. Thus, a trivial application of LeCam's Third Lemma (see, for instance, Theorem 6.6 in van der Vaart (1998)) yields that under

$\theta_{n,2}$, $\varphi_n(\bar{\kappa}_n, \mathbf{Y}_n)$ converges to the same weak limit as under $\theta_{n,1}$ under the subsequence. But convergence in distribution implies convergence of expectations given that $0 \leq \varphi_n \leq 1$, and the desired contradiction follows. ■

A.4 Proof of Corollary 1

We prove part (b). The proof of part (a) is similar, but slightly easier, so we omit it for brevity.

Using the assumption of the Corollary, Theorem 1 (b) (or, equivalently, Theorem 2) shows that there exists $\varphi_i^o : \mathbb{R} \times \mathbb{R}^2 \mapsto [0,1]$ such that for Π_0-almost all K, and all (a,b)

$$E_{\theta_n}[\varphi_n(\bar{\kappa}_{0,n} + K, \mathbf{Y}_n)] \to E_{a,b}[\varphi_i^o(K, \mathbf{W}^o)]$$

for some sequence θ_n with $(\sqrt{\mathbf{x}_n'\mathbf{x}_n}\beta_n, \sqrt{\mathbf{x}_n'\mathbf{x}_n}\Delta_n - i\bar{\kappa}_{0,n}) = (b,a)$ and $\tau_n = o(n^{1/4})$. Write $\theta_n = T_n(a,b,t)$ for θ_n with $(\sqrt{\mathbf{x}_n'\mathbf{x}_n}\beta_n, \sqrt{\mathbf{x}_n'\mathbf{x}_n}\Delta_n - i\bar{\kappa}_{0,n}) = (b,a)$ and $\tau_n = t$. We claim that for Π_0-almost all K, and all (a,b)

$$\sup_{0 \leq t \leq s_n} |E_{T_n(a,b,t)}[\varphi_n(\bar{\kappa}_{0,n} + K, \mathbf{Y}_n)] - E_{a,b}[\varphi_i^o(K, \mathbf{W}^o)]| \to 0. \tag{23}$$

Suppose otherwise. Then there exists a sequence $t_n^* = o(n^{1/4})$ such that under $\theta_n = T_n(a,b,t_n^*)$, $E_{T_n(a,b,t_n^*)}[\varphi_n(\bar{\kappa}_{0,n} + K, \mathbf{Y}_n)]$ does not converge to $E_{a,b}[\varphi_i^o(K, \mathbf{W}^o)]$. But this contradicts the second claim of Theorem 2.

Now by Tonelli's Theorem,

$$\int E_\theta[\int (1 - \varphi_n(\bar{\kappa}, \mathbf{Y}_n)) d\Pi_n(\bar{\kappa})] dG_n(\theta) = 1 - \int \int E_\theta[\varphi_n(\bar{\kappa}, \mathbf{Y}_n)] dG_n(\theta) d\Pi_n(\bar{\kappa})$$

and with $G_{\tau,n}^{a,b}$ the conditional distribution of τ given (a,b) under G_n, by definition of G_n and Π_n,

$$\begin{aligned}
& \int \int E_\theta[\varphi_n(\bar{\kappa}, \mathbf{Y}_n)] dG_n(\theta) d\Pi_n(\bar{\kappa}) \\
= & \int \int \int E_{T_n(a,b,t)}[\varphi_n(\bar{\kappa}_{0,n} + K, \mathbf{Y}_n)] dG_{\tau,n}^{a,b}(t) dF_0(a,b) \Pi_0(K) \\
= & \int \int \int \mathbf{1}[t < s_n] E_{T_n(a,b,t)}[\varphi_n(\bar{\kappa}_{0,n} + K, \mathbf{Y}_n)] dG_{\tau,n}^{a,b}(t) dF_0(a,b) \Pi_0(K) + o(1)
\end{aligned}$$

where the last line follows from the assumption about G_n and $0 \leq \int \int E_\theta[\varphi_n(\bar{\kappa}_{0,n} + K, \mathbf{Y}_n)] dF_n(\theta) d\Pi_0(K) \leq \int \int dF_n(\theta) d\Pi_0(K) = 1$. Now by (23), for all a, b and Π_0 almost all values of K,

$$|\int \mathbf{1}[t < s_n] E_{T_n(a,b,t)}[\varphi_n(\bar{\kappa}_{0,n} + K, \mathbf{Y}_n)] dG_{\tau,n}^{a,b}(t) - E_{a,b}[\varphi_i^o(K, \mathbf{W}^o)]| \to 0.$$

Thus, by dominated convergence,

$$\int \int E_\theta[\varphi_n(\bar{\kappa}, \mathbf{Y}_n)] dF_n(\theta) d\Pi_n(\bar{\kappa}) \to \int \int E_{a,b}[\varphi_i^o(K, \mathbf{W}^o)] dF_0(a,b) \Pi_0(K)$$

and the claim thus follows from

$$1 - \int\int E_{a,b}[\varphi_i^o(K, \mathbf{W}^o)]dF_0(a,b)\Pi_0(K) = \int E_{a,b}[\int(1-\varphi_i^o(K,\mathbf{W}^o))d\Pi_0(K)]dF_0(a,b)$$

using Tonelli's Theorem.

A.5 Proof of Lemma 3

We will make use of the following Lemma.

Lemma 5 *Let $\mathbf{H}_n$ and $\hat{\mathbf{H}}_n$ be the Choleski decompositions of $\mathbf{\Omega}_n = \mathbf{H}_n\mathbf{H}_n'$ and $\hat{\mathbf{\Omega}}_n = \hat{\mathbf{H}}_n\hat{\mathbf{H}}_n'$, respectively, and let $\mathbf{Q}_n = (\hat{\beta}_n^{long} - \beta_n, \hat{\beta}_n^{short} - \beta_n - \Delta_n)'$. Then under (17) and $\mathbf{\Omega}_n^{-1}\hat{\mathbf{\Omega}}_n \xrightarrow{p} \mathbf{I}_2$*

(a) $\hat{\mathbf{H}}_n^{-1}\mathbf{H}_n \xrightarrow{p} \mathbf{I}_2$

(b) $\hat{\mathbf{H}}_n^{-1}\mathbf{Q}_n \Rightarrow \mathcal{N}(\mathbf{0}, \mathbf{I}_2)$.

Proof. (a) Note that $\hat{\mathbf{H}}_n^{-1}\mathbf{H}_n\mathbf{H}_n'\hat{\mathbf{H}}_n'^{-1}$, by similarity, has the same eigenvalues as $\hat{\mathbf{H}}_n'^{-1}\hat{\mathbf{H}}_n^{-1}\mathbf{H}_n\mathbf{H}_n' = \hat{\mathbf{\Omega}}_n^{-1}\mathbf{\Omega}_n \xrightarrow{p} \mathbf{I}_2$, so they both converge to one in probability. But $\hat{\mathbf{H}}_n^{-1}\mathbf{H}_n\mathbf{H}_n'\hat{\mathbf{H}}_n'^{-1}$ is symmetric, and all symmetric matrices with eigenvalues converging to one converge to the identity matrix. Thus $\hat{\mathbf{H}}_n^{-1}\mathbf{H}_n\mathbf{H}_n'\hat{\mathbf{H}}_n'^{-1} \xrightarrow{p} \mathbf{I}_2$, and since $\hat{\mathbf{H}}_n^{-1}\mathbf{H}_n$ is lower triangular, this further implies $\hat{\mathbf{H}}_n^{-1}\mathbf{H}_n \xrightarrow{p} \mathbf{I}_2$.

(b) Note that $\mathbf{H}_n$ is related to $\mathbf{\Omega}_n^{1/2}$ via $\mathbf{H}_n = \mathbf{\Omega}_n^{1/2}\mathbf{O}_n$ for some rotation matrix $\mathbf{O}_n$. Thus, also $\mathbf{H}_n^{-1}\mathbf{Q}_n = \mathbf{O}_n'\mathbf{\Omega}^{-1/2}\mathbf{Q}_n \Rightarrow \mathcal{N}(\mathbf{0}, \mathbf{I}_2)$. (Suppose otherwise. Then, by the Cramer-Wold device, for some 2×1 vector $\boldsymbol{v}$ and $c \in \mathbb{R}$, $\liminf_{n\to\infty} |P(\boldsymbol{v}'\mathbf{O}_n'\mathbf{\Omega}^{-1/2}\mathbf{Q}_n > c) - P(\mathcal{N}(\mathbf{0}, \boldsymbol{v}'\boldsymbol{v}) > c)| > 0$. Pick a subsequence along which the liminf is attained, and $\mathbf{O}_n$ converges. Then we have a contradiction, because the continuous mapping theorem implies the convergence $P(\boldsymbol{v}'\mathbf{O}_n'\mathbf{\Omega}^{-1/2}\mathbf{Q}_n > c) - P(\mathcal{N}(\mathbf{0}, \boldsymbol{v}'\boldsymbol{v}) > c) \to 0$ along that subsequence.) Invoking Lemma 5 (a), also $\hat{\mathbf{H}}_n^{-1}\mathbf{Q}_n = (\hat{\mathbf{H}}_n^{-1}\mathbf{H}_n)\mathbf{H}_n^{-1}\mathbf{Q}_n \Rightarrow \mathcal{N}(\mathbf{0}, \mathbf{I}_2)$ by the continuous mapping theorem. ■

(a) Write $L(Z_1, Z_2 + g_0, \psi_1, \psi_2)$ for the expression in equation (20). Reparametrize $\boldsymbol{\psi}$ and g_0 in (20) in terms of $(r, \phi, u) \in [0, \infty) \times [0, \pi/2) \times [0, 1]$ via $\psi_1 = r\cos(\phi)$, $\psi_2 = r\sin(\phi)$ and $u = g_0/\psi_2$ (with $u = 0$ if $\psi_2 = 0$). By a direct calculation, the limit of $L(Z_1, Z_2 + ur\cos(\phi), r\cos(\phi), r\sin(\phi))$ as $r \to \infty$ exists for almost all Z_1, Z_2 and all $(u, \phi) \in \mathbb{R} \times [0, \pi/2)$ and is equal to

$$L^\infty(Z_1, Z_2, u, \phi) = \begin{cases} (Z_1 - (1+u)\tan\phi)^2 \text{ if } Z_1 > (1+u)\tan\phi \\ (Z_1 + (1-u)\tan\phi)^2 \text{ if } Z_1 < -(1-u)\tan\phi \\ 0 \text{ otherwise.} \end{cases}$$

Correspondingly, $\lim_{r\to\infty} \mathrm{cv}_\Omega((r\cos(\phi), r\sin(\phi)) = \mathrm{cv}_\Omega^\infty(\phi)$ exists, too, and satisfies $\sup_{0\le u\le 1} P(L^\infty(Z_1, Z_2, u, \phi) \ge \mathrm{cv}_\Omega^\infty(\phi)) \le \alpha$. (In general, this inequality is not sharp, since the definition of $\text{cv}_\Omega^\infty(\phi)$ also requires $P(L(Z_1, ur\cos(\phi) + Z_2, r\cos(\phi), r\sin(\phi)) \geq \text{cv}_\Omega^\infty(\phi)) \leq \alpha$ for all finite r). If $r \to \infty$ and $\phi \to \pi/2$, then the limit still exists and is equal to $L^\infty(Z_1, Z_2, u, \pi/2) = 0$.

Suppose the assertion of the Lemma is false. Then there exists a subsequence of n such that along that subsequence, $\lim_{n\to\infty} E_{\theta_n}[\hat{\varphi}_{\text{LR},n}(\bar{\kappa}_n, \mathbf{Y}_n)] = \limsup_{n\to\infty} E_{\theta_n}[\hat{\varphi}_{\text{LR},n}(\bar{\kappa}_n, \mathbf{Y}_n)] > \alpha$. Pick a sub-subsequence, such that with (r_n, ϕ_n, u_n) the parameters computed from $\mathbf{\Omega} = \mathbf{\Omega}_n$ and $g_{0,n} = \sqrt{\Omega_{n,11}}\Delta_n/\sqrt{\Omega_{n,11}\Omega_{n,22} - \Omega_{n,12}^2}$, (r_n, ϕ_n, u_n) converge along that sub-subsequence to some value (r_0, ϕ_0, u_0) in $(\mathbb{R} \cup \{\infty\}) \times [0, \pi/2] \times [0, 1]$. Correspondingly, let cv_0 be the limit of $\text{cv}_\Omega((r_n\cos(\phi_n), r_n\sin(\phi_n))$ along that sub-subsequence (which exists by the above observations also when $r_n \to \infty$, even when $\phi_0 = \pi/2$).

By Lemma 5 (a), $(\hat{Z}_{n,1}, \hat{Z}_{n,2})' = \hat{\mathbf{H}}_n^{-1}\mathbf{Q}_n \Rightarrow (Z_1, Z_2)' \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_2)$ by the continuous mapping theorem. Since

$$\hat{\mathbf{H}}_n^{-1} = \begin{pmatrix} 1/\sqrt{\hat{\Omega}_{n,11}} & 0 \\ -\frac{\hat{\Omega}_{n,12}}{\sqrt{\hat{\Omega}_{n,11}}\sqrt{\hat{\Omega}_{n,11}\hat{\Omega}_{n,22} - \hat{\Omega}_{n,12}^2}} & \frac{\sqrt{\hat{\Omega}_{n,11}}}{\sqrt{\hat{\Omega}_{n,11}\hat{\Omega}_{n,22} - \hat{\Omega}_{n,12}^2}} \end{pmatrix}$$

the definitions of $\hat{\psi}_{1,n} = \hat{\psi}_1$ and $\hat{\psi}_{2,n} = \hat{\psi}_2$ yield

$$\widehat{\text{LR}}_n(\bar{\kappa}_n) = \min_{|g|\leq\hat{\psi}_{2,n}} \left\| \begin{array}{c} \hat{Z}_{n,1} \\ \hat{Z}_{n,2} + \hat{g}_{0,n} - g \end{array} \right\|^2 - \min_{c,|g|\leq\hat{\psi}_{2,n}} \left\| \begin{array}{c} \hat{Z}_{n,1} - c \\ \hat{Z}_{n,2} + \hat{g}_{0,n} - g - \hat{\psi}_{1,n}c \end{array} \right\|^2$$

where $\hat{g}_{0,n} = \sqrt{\hat{\Omega}_{n,11}}\Delta_n/\sqrt{\hat{\Omega}_{n,11}\hat{\Omega}_{n,22} - \hat{\Omega}_{n,12}^2}$. Let $(\hat{r}_n, \hat{\phi}_n, \hat{u}_n) \in [0, \infty) \times [0, \pi/2) \times [0, 1]$ be such that $\hat{\psi}_{1,n} = \hat{r}_n\cos(\hat{\phi}_n)$, $\hat{\psi}_{2,n} = \hat{r}_n\sin(\hat{\phi}_n)$ and $\hat{u}_n = \hat{g}_{0,n}/\hat{\psi}_{2,n}$ (with $\hat{u}_n = 0$ if $\hat{\psi}_{2,n} = 0$). Write $\hat{\mathbf{H}}_n^{-1}\mathbf{H}_n \xrightarrow{p} \mathbf{I}_2$ from Lemma 5 (b) element-by-element to conclude that $\hat{\Omega}_{11,n}/\Omega_{11,n} \xrightarrow{p} 1$, $(\hat{\Omega}_{11,n}\hat{\Omega}_{22,n} - \hat{\Omega}_{12,n}^2)/(\Omega_{11,n}\Omega_{22,n} - \Omega_{12,n}^2) \xrightarrow{p} 1$ and $(\hat{\Omega}_{12,n} - \Omega_{12,n})/(\Omega_{11,n}\Omega_{22,n} - \Omega_{12,n}^2) \xrightarrow{p} 0$. Therefore, also $(\hat{r}_n - r_n)/\max(r_n, 1) \xrightarrow{p} 0$, $\hat{u}_n - u_n \xrightarrow{p} 0$ and $\hat{\phi}_n - \phi_n \xrightarrow{p} 0$. Thus, along the sub-subsequence defined above, by the continuous mapping theorem

$$\widehat{\text{LR}}_n(\bar{\kappa}_n) \Rightarrow L_0 = \begin{cases} L(Z_1, u_0r_0\cos(\phi_0) + Z_2, r_0\cos(\phi), r_0\sin(\phi_0)) \text{ if } r_0 < \infty \\ L^\infty(Z_1, Z_2, u_0, \phi_0) \text{ otherwise} \end{cases}$$

and $E_{\theta_n}[\hat{\varphi}_{\text{LR},n}(\bar{\kappa}_n, \mathbf{Y}_n)] \to P(L_0 > \text{cv}_0)$. But by the definition of cv_0, $P(L_0 > \text{cv}_0) \leq \alpha$, yielding the desired contradiction.

A.6 Additional theoretical results in general model

In the following results, we consider the linear regression model (6) with additional baseline controls $q_{i,j}^e$,

$$y_i^e = x_i^e\beta + \sum_{j=1}^{k}\gamma_j z_{i,j}^e + \sum_{j=1}^{m}\delta_j q_{i,j}^e + \varepsilon_i^e,\ i = 1, \ldots, m+n \tag{24}$$

where the regressors $\{x_i^e, z_{i,1}^e, \ldots, z_{i,k}^e, q_{i,1}^e, \ldots, q_{i,m}^e\}_{i=1}^{m+n}$ are non-stochastic. Write (24) in vector form as

$$\begin{aligned}\mathbf{y}^e &= \mathbf{x}^e\beta + \mathbf{Z}^e\boldsymbol{\gamma} + \mathbf{Q}^e\boldsymbol{\delta} + \boldsymbol{\varepsilon}^e \\ &= \mathbf{R}\boldsymbol{\alpha} + \tilde{\mathbf{Z}}\boldsymbol{\gamma} + \mathbf{e}\end{aligned}$$

where $\mathbf{R} = (\mathbf{x}^e, \mathbf{Q}^e)$, $\tilde{\mathbf{Z}} = (\mathbf{I}_{\tilde{n}} - \mathbf{Q}^e(\mathbf{Q}^{e\prime}\mathbf{Q}^e)^{-1}\mathbf{Q}^{e\prime})\mathbf{Z}^e$, $\boldsymbol{\alpha} = (\beta, (\boldsymbol{\delta} + (\mathbf{Q}^{e\prime}\mathbf{Q}^e)^{-1}\mathbf{Q}^{e\prime}\mathbf{Z}^e\boldsymbol{\gamma})')'$ and, to ease notation, $\mathbf{e} = \boldsymbol{\varepsilon}^e$.

We consider a set-up with M clusters of not necessarily equal size. Write

$$\mathbf{y}_j^e = \mathbf{R}_j\boldsymbol{\alpha} + \tilde{\mathbf{Z}}_j\boldsymbol{\gamma} + \mathbf{e}_j$$

for the observations in the jth cluster (so that the sum of the lengths of the $\mathbf{y}_j^e$ vectors over $j = 1, \ldots, M$ equals $m + n$). We allow the sequence of regressors $\mathbf{R}$ and $\tilde{\mathbf{Z}}$, the coefficients $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$, the number of observations per cluster, and the distribution of $\mathbf{e}_j$ to depend on the sample size in a double array fashion, but do not make this explicit in the notation.

Define the $(m + n) \times 2$ matrix $\mathbf{v} = (\mathbf{v}_1', \ldots, \mathbf{v}_M')'$ such that $M^{-1}\mathbf{v}'\mathbf{y}^e = M^{-1}\sum_{j=1}^M \mathbf{v}_j'\mathbf{y}_j^e = (\hat{\beta}^{\text{long}}, \hat{\beta}^{\text{short}})'$. Let $||\cdot||$ be the spectral norm. We make the following assumptions.

Condition 1 *(a)* $\mathbf{e}_j$, $j = 1, \ldots, M$ *are independent with* $E[\mathbf{e}_j] = 0$ *and* $E[\mathbf{e}_j\mathbf{e}_j'] = \boldsymbol{\Sigma}_j$.

(b) $||(M^{-1}\sum_{j=1}^M \mathbf{v}_j'\boldsymbol{\Sigma}_j\mathbf{v}_j)^{-1}|| = O(1)$, $\max_j ||\mathbf{v}_j||^4 \cdot \sum_{j=1}^M E[||\mathbf{e}_j||^4] = o(M^2)$.

(c) $||M^{-1}\sum_{j=1}^M \mathbf{R}_j\mathbf{R}_j'|| = O(1)$, $||(M^{-1}\sum_{j=1}^M \mathbf{R}_j\mathbf{R}_j')^{-1}|| = O(1)$, $\max_j ||\boldsymbol{\Sigma}_j|| = o(M)$, $\max_j ||\boldsymbol{\Sigma}_j|| \cdot \max_j ||\mathbf{v}_j||^4 = O(M)$ *and* $\max_j ||\mathbf{v}_j||^2 = O(M)$.

(d) $\max_j ||\mathbf{v}_j||^2 \cdot \kappa^2 = o(M)$ *and* $\max_j ||\boldsymbol{\Sigma}_j|| \cdot \max_j ||\mathbf{v}_j||^4 \cdot \kappa^2 = o(M^2)$*, where* $\kappa^2 = \boldsymbol{\gamma}'\tilde{\mathbf{Z}}'\tilde{\mathbf{Z}}\boldsymbol{\gamma} = \boldsymbol{\gamma}'\mathbf{Z}'\mathbf{Z}\boldsymbol{\gamma}$ *with* $\mathbf{Z}$ *defined as below (6).*

Theorem 3 *(a) Under Condition 1 (a) and (b),*

$$\boldsymbol{\Omega}_n^{-1/2}\begin{pmatrix} \hat{\beta}_n^{long} - \beta_n \\ \hat{\beta}_n^{short} - \beta_n - \Delta_n \end{pmatrix} \Rightarrow \mathcal{N}(\mathbf{0}, \mathbf{I}_2)$$

where $\boldsymbol{\Omega}_n = M^{-2}\sum_{j=1}^M \mathbf{v}_j'\boldsymbol{\Sigma}_j\mathbf{v}_j$*;*

(b) Under Condition 1 (a)-(e), $\boldsymbol{\Omega}_n^{-1}\hat{\boldsymbol{\Omega}}_n \xrightarrow{p} \mathbf{I}_2$*, where*

$$\hat{\boldsymbol{\Omega}}_n = M^{-2}\sum_{j=1}^M \mathbf{v}_j'\hat{\mathbf{e}}_j\hat{\mathbf{e}}_j'\mathbf{v}_j \text{ and } \hat{\mathbf{e}} = \mathbf{y}^e - \mathbf{R}(\mathbf{R}'\mathbf{R})^{-1}\mathbf{R}'\mathbf{y}^e.$$

Remark 4 *Since* $(\hat{\beta}^{long}, \hat{\beta}^{short})$ *are* $\mathbf{v}_j$*-weighted averages of* $\mathbf{e}_j$*, some bound on the relative magnitude of* $||\mathbf{v}_j||$ *is necessary to obtain asymptotic normality. The bounds in Condition 1 are relatively weak, allowing for* $\max_j ||\mathbf{v}_j|| = o(M^{1/4})$ *(if* $\sum_{j=1}^M E[||\mathbf{e}_j||^4] = O(M)$*,* $\max_j ||\boldsymbol{\Sigma}_j|| = O(1)$ *and*

$\kappa^2 = o(M^{1/2})$). At the same time, one could also imagine that $\max_j ||\mathbf{v}_j|| = O(1)$, which would then allow for $E[||\mathbf{e}_j||^4] = o(M)$, either because of increasingly fat tails, or because the number of observations per cluster is growing.

The definition of $\hat{\mathbf{\Omega}}_n$ is the standard clustered variance estimator, except that the regression residuals are computed from the short regression. Under $\max_j ||\mathbf{v}_j|| = O(1)$ and $\max_j ||\mathbf{\Sigma}_j|| = O(1)$, $\kappa^2 = o(M)$ is enough to obtain consistency of $\hat{\mathbf{\Omega}}_n$. The important special case of independent but heteroskedastic disturbances ε_i^e (so that $\hat{\mathbf{\Omega}}_n$ reduces to the White (1980) standard errors based on short regression residuals), is obtained by setting $M = m + n$.

Proof. (a) By the Cramer-Wold device, it suffices to show that $M^{-1}\boldsymbol{\upsilon}'\mathbf{v}'\mathbf{e}/\sqrt{\boldsymbol{\upsilon}'\mathbf{\Omega}_n\boldsymbol{\upsilon}} \Rightarrow \mathcal{N}(0,1)$ for all 2×1 vectors $\boldsymbol{\upsilon}$ with $\boldsymbol{\upsilon}'\boldsymbol{\upsilon} = 1$. This follows from the (triangular array version of the) Lyapunov central limit theorem applied to the M independent variables $\boldsymbol{\upsilon}'\mathbf{v}_j'\mathbf{e}_j \sim (0, \boldsymbol{\upsilon}'\mathbf{v}_j'\mathbf{\Sigma}_j\mathbf{v}_j\boldsymbol{\upsilon})$ and Condition 1 (b), since

$$\frac{\sum_{j=1}^M E[(\boldsymbol{\upsilon}'\mathbf{v}_j'\mathbf{e}_j)^4]}{(\sum_{j=1}^M \boldsymbol{\upsilon}'\mathbf{v}_j'\mathbf{\Sigma}_j\mathbf{v}_j\boldsymbol{\upsilon})^2} \leq \max_j ||\mathbf{v}_j||^4 \cdot M^{-2}\sum_{j=1}^M E[||\mathbf{e}_j||^4] \cdot ||(M^{-1}\sum_{j=1}^M \mathbf{v}_j'\mathbf{\Sigma}_j\mathbf{v}_j)^{-1}||^2 \to 0$$

and $\mathrm{Var}[M^{-1}\boldsymbol{\upsilon}'\mathbf{v}'\mathbf{e}/\sqrt{\boldsymbol{\upsilon}'\mathbf{\Omega}_n\boldsymbol{\upsilon}}] = 1$.

(b) We show convergence of $\boldsymbol{\upsilon}'\hat{\mathbf{\Omega}}_n\boldsymbol{\upsilon}/(\boldsymbol{\upsilon}'\mathbf{\Omega}_n\boldsymbol{\upsilon}) \xrightarrow{p} 1$ for all 2×1 vectors $\boldsymbol{\upsilon}$ with $\boldsymbol{\upsilon}'\boldsymbol{\upsilon} = 1$. Note that $\boldsymbol{\upsilon}'\hat{\mathbf{\Omega}}_n\boldsymbol{\upsilon} = M^{-2}\sum_{j=1}^M \hat{\mathbf{e}}_j'\mathbf{V}_j\hat{\mathbf{e}}_j = M^{-2}\hat{\mathbf{e}}'\mathbf{V}\hat{\mathbf{e}}$ with $\mathbf{V}_j = \mathbf{v}_j\boldsymbol{\upsilon}\boldsymbol{\upsilon}'\mathbf{v}_j'$ and $\mathbf{V} = \mathrm{diag}(\mathbf{V}_1, \ldots, \mathbf{V}_M)$, and

$$\begin{aligned} \hat{\mathbf{e}} &= \mathbf{M}_R\mathbf{e} + \mathbf{M}_R\tilde{\mathbf{Z}}\boldsymbol{\gamma} \\ &= \mathbf{e} - \mathbf{R}(\mathbf{R}'\mathbf{R})^{-1}\mathbf{R}'\mathbf{e} + \mathbf{M}_R\tilde{\mathbf{Z}}\boldsymbol{\gamma} \end{aligned} \tag{25}$$

with $\mathbf{M}_R = \mathbf{I}_{n+m} - \mathbf{R}(\mathbf{R}'\mathbf{R})^{-1}\mathbf{R}'$, so that

$$\begin{aligned} \hat{\mathbf{e}}'\mathbf{V}\hat{\mathbf{e}} = \mathbf{e}'\mathbf{V}\mathbf{e} + \boldsymbol{\gamma}'\tilde{\mathbf{Z}}'\mathbf{M}_R\mathbf{V}\mathbf{M}_R\tilde{\mathbf{Z}}\boldsymbol{\gamma} + 2\boldsymbol{\gamma}'\tilde{\mathbf{Z}}'\mathbf{M}_R\mathbf{V}\mathbf{M}_R\mathbf{e} - 2\mathbf{e}'\mathbf{V}\mathbf{R}(\mathbf{R}'\mathbf{R})^{-1}\mathbf{R}'\mathbf{e} \\ + \mathbf{e}'\mathbf{R}(\mathbf{R}'\mathbf{R})^{-1}\mathbf{R}'\mathbf{V}\mathbf{R}(\mathbf{R}'\mathbf{R})^{-1}\mathbf{R}'\mathbf{e}. \end{aligned}$$

Now

$$\begin{aligned} \boldsymbol{\gamma}'\tilde{\mathbf{Z}}'\mathbf{M}_R\mathbf{V}\mathbf{M}_R\tilde{\mathbf{Z}}\boldsymbol{\gamma} &\leq ||\mathbf{V}|| \cdot \boldsymbol{\gamma}'\tilde{\mathbf{Z}}'\mathbf{M}_R\tilde{\mathbf{Z}}\boldsymbol{\gamma} \\ &\leq \max_j ||\mathbf{v}_j||^2 \cdot \boldsymbol{\gamma}'\tilde{\mathbf{Z}}'\tilde{\mathbf{Z}}\boldsymbol{\gamma} = \max_j ||\mathbf{v}_j||^2 \cdot \kappa^2 \end{aligned}$$

and, with $\mathbf{\Sigma} = \mathrm{diag}(\mathbf{\Sigma}_1, \ldots, \mathbf{\Sigma}_M)$,

$$\begin{aligned} \mathrm{Var}[\boldsymbol{\gamma}'\tilde{\mathbf{Z}}'\mathbf{M}_R\mathbf{V}\mathbf{M}_R\mathbf{e}] &= \boldsymbol{\gamma}'\tilde{\mathbf{Z}}'\mathbf{M}_R\mathbf{V}\mathbf{M}_R\mathbf{\Sigma}\mathbf{M}_R\mathbf{V}\mathbf{M}_R\tilde{\mathbf{Z}}\boldsymbol{\gamma} \\ &\leq \max_j ||\mathbf{\Sigma}_j|| \cdot \boldsymbol{\gamma}'\tilde{\mathbf{Z}}'\mathbf{M}_R\mathbf{V}\mathbf{M}_R\mathbf{V}\mathbf{M}_R\tilde{\mathbf{Z}}\boldsymbol{\gamma} \end{aligned}$$

$$
\begin{aligned}
&\leq \max_j ||\mathbf{\Sigma}_j|| \cdot \boldsymbol{\gamma}'\tilde{\mathbf{Z}}'\mathbf{M}_R\mathbf{V}^2\mathbf{M}_R\tilde{\mathbf{Z}}\boldsymbol{\gamma} \\
&\leq \max_j ||\mathbf{\Sigma}_j|| \cdot \max_j ||\mathbf{v}_j||^4 \cdot \kappa^2
\end{aligned}
$$

and

$$
\begin{aligned}
||\operatorname{Var}[\mathbf{R}'\mathbf{e}]|| &= ||\mathbf{R}'\mathbf{\Sigma}\mathbf{R}|| \leq \max_j ||\mathbf{\Sigma}_j|| \cdot \sum_{j=1}^{M} ||\mathbf{R}_j||^2 \\
||\operatorname{Var}[\mathbf{R}'\mathbf{V}\mathbf{e}]|| &= ||\mathbf{R}'\mathbf{V}\mathbf{\Sigma}\mathbf{V}\mathbf{R}|| \leq \max_j ||\mathbf{\Sigma}_j|| \cdot \max_j ||\mathbf{v}_j||^4 \cdot \sum_{j=1}^{M} ||\mathbf{R}_j||^2 \\
||\mathbf{R}'\mathbf{V}\mathbf{R}|| &\leq \max_j ||\mathbf{v}_j||^2 \cdot \sum_{j=1}^{M} ||\mathbf{R}_j||^2.
\end{aligned}
$$

Furthermore, $(\boldsymbol{\upsilon}'\mathbf{\Omega}_n\boldsymbol{\upsilon})^{-1} = (M^{-2}\sum_{j=1}^{M}\boldsymbol{\upsilon}'\mathbf{v}_j'\mathbf{\Sigma}_j\mathbf{v}_j\boldsymbol{\upsilon})^{-1} \leq ||(M^{-2}\sum_{j=1}^{M}\mathbf{v}_j'\mathbf{\Sigma}_j\mathbf{v}_j)^{-1}|| = O(M^{-1})$, so that under Condition 1 (b)-(d), $M^{-2}(\hat{\mathbf{e}}'\mathbf{V}\hat{\mathbf{e}} - \mathbf{e}'\mathbf{V}\mathbf{e})/(\boldsymbol{\upsilon}'\mathbf{\Omega}_n\boldsymbol{\upsilon}) \xrightarrow{p} 0$.

Finally, rewrite $\mathbf{e}'\mathbf{V}\mathbf{e} = \sum_{j=1}^{M}\boldsymbol{\upsilon}'\mathbf{v}_j'\mathbf{e}_j\mathbf{e}_j'\mathbf{v}_j\boldsymbol{\upsilon}$. Then $E[M^{-1}\mathbf{e}'\mathbf{V}\mathbf{e} - M\boldsymbol{\upsilon}'\mathbf{\Omega}_n\boldsymbol{\upsilon}] = 0$, and

$$
\begin{aligned}
\operatorname{Var}[M^{-1}\mathbf{e}'\mathbf{V}\mathbf{e} - M\boldsymbol{\upsilon}'\mathbf{\Omega}_n\boldsymbol{\upsilon}] &= M^{-2}\sum_{j=1}^{M}\operatorname{Var}[\boldsymbol{\upsilon}'\mathbf{v}_j'(\mathbf{e}_j\mathbf{e}_j' - \mathbf{\Sigma}_j)\mathbf{v}_j\boldsymbol{\upsilon}] \\
&\leq M^{-2}\max_j ||\mathbf{v}_j||^4 \cdot \sum_{j=1}^{M} E[||\mathbf{e}_j||^4]
\end{aligned}
$$

and the result follows from $(\boldsymbol{\upsilon}'\mathbf{\Omega}_n\boldsymbol{\upsilon})^{-1} = O(M^{-1})$ and Condition 1 (a). ∎

References

Armstrong, T. B., and M. Kolesár (2016): "Optimal inference in a class of regression models," *Working Paper, Princeton University.*

Belloni, A., V. Chernozhukov, and C. Hansen (2014): "Inference on treatment effects after selection among high-dimensional controls," *The Review of Economic Studies*, 81(2), 608–650.

Cattaneo, M. D., M. Jansson, and W. K. Newey (2015): "Heteroskedastic Consistent Standard Errors with Many Covariates," *Working Paper.*

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

Conley, T. G., C. B. Hansen, and P. E. Rossi (2012): "Plausibly exogenous," *Review of Economics and Statistics*, 94(1), 260–272.

Donohue III, J. J., and S. D. Levitt (2001): "The Impact of Legalized Abortion on Crime," *Quarterly Journal of Economics*, CXVI, 379–420.

Elliott, G., U. K. Müller, and M. W. Watson (2015): "Nearly Optimal Tests When a Nuisance Parameter is Present Under the Null Hypothesis," *Econometrica*, 83, 771–811.

Fan, J., and R. Li (2001): "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American Statistical Association*, 96(456), 1348–1360.

Foote, C. L., and C. F. Goetz (2008): "The impact of legalized abortion on crime: Comment," *The Quarterly Journal of Economics*, 123, 407–423.

Hoerl, A. E., and R. W. Kennard (1970): "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, 12(1), 55–67.

Joyce, T. (2004): "Did legalized abortion lower crime?," *Journal of Human Resources*, 39(1), 1–28.

——— (2009): "A simple test of abortion and crime," *The Review of Economics and Statistics*, 91(1), 112–123.

Leeb, H., and B. M. Pötscher (2005): "Model Selection and Inference: Facts and Fiction," *Econometric Theory*, 21, 21–59.

Leeb, H., and B. M. Pötscher (2008a): "Can one estimate the unconditional distribution of post-model-selection estimators?," *Econometric Theory*, 24(2), 338–376.

——— (2008b): "Recent Developments in Model Selection and Related Areas," *Econometric Theory*, 24(2), 319–322.

Lehmann, E. L., and J. P. Romano (2005): *Testing Statistical Hypotheses*. Springer, New York.

Mardia, K. V., and P. E. Jupp (2000): *Directional statistics*, Wiley Series in Probability and Statistics. John Wiley & Sons, Chichester.

Obenchain, R. L. (1977): "Classical F-Tests and Confidence Regions for Ridge Regression," *Technometrics*, 19, 429–439.

Pratt, J. W. (1961): "Length of Confidence Intervals," *Journal of the American Statistical Association*, 56, 549–567.

Tibshirani, R. (1996): "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, 58, 267–288.

van der Vaart, A. W. (1991): "An asymptotic representation theorem," *International Statistical Review*, 259, 97–121.

——— (1998): *Asymptotic Statistics*. Cambridge University Press, Cambridge, UK.

White, H. (1980): "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," *Econometrica*, 48, 817–830.