Yale University

EliScholar – A Digital Platform for Scholarly Publishing at Yale

Cowles Foundation Discussion Papers | Cowles Foundation

2-1-2015

Unbiased Instrumental Variables Estimation under Known First-Stage Sign

Isaiah Andrews

Timothy B. Armstrong

Follow this and additional works at: https://elischolar.library.yale.edu/cowles-discussion-paper-series

 Part of the Economics Commons

Recommended Citation

Andrews, Isaiah and Armstrong, Timothy B., "Unbiased Instrumental Variables Estimation under Known First-Stage Sign" (2015). *Cowles Foundation Discussion Papers*. 2406.
https://elischolar.library.yale.edu/cowles-discussion-paper-series/2406

This Discussion Paper is brought to you for free and open access by the Cowles Foundation at EliScholar – A Digital Platform for Scholarly Publishing at Yale. It has been accepted for inclusion in Cowles Foundation Discussion Papers by an authorized administrator of EliScholar – A Digital Platform for Scholarly Publishing at Yale. For more information, please contact elischolar@yale.edu.

UNBIASED INSTRUMENTAL VARIABLES ESTIMATION UNDER KNOWN FIRST-STAGE SIGN

By

Isaiah Andrews and Timothy B. Armstrong

March 2015
Revised November 2016

COWLES FOUNDATION DISCUSSION PAPER NO. 1984R5



COWLES FOUNDATION FOR RESEARCH IN ECONOMICS
YALE UNIVERSITY
Box 208281
New Haven, Connecticut 06520-8281

http://cowles.yale.edu/

Unbiased Instrumental Variables Estimation Under Known First-Stage Sign

Isaiah Andrews
MIT *

Timothy B. Armstrong
Yale University †

November 28, 2016

Abstract

We derive mean-unbiased estimators for the structural parameter in instrumental variables models with a single endogenous regressor where the sign of one or more first stage coefficients is known. In the case with a single instrument, there is a unique non-randomized unbiased estimator based on the reduced-form and first-stage regression estimates. For cases with multiple instruments we propose a class of unbiased estimators and show that an estimator within this class is efficient when the instruments are strong. We show numerically that unbiasedness does not come at a cost of increased dispersion in models with a single instrument: in this case the unbiased estimator is less dispersed than the 2SLS estimator. Our finite-sample results apply to normal models with known variance for the reduced-form errors, and imply analogous results under weak instrument asymptotics with an unknown error distribution.

*email: iandrews@mit.edu

†email: timothy.armstrong@yale.edu. We thank Gary Chamberlain, Jerry Hausman, Max Kasy, Frank Kleibergen, Anna Mikusheva, Daniel Pollmann, Jim Stock, and audiences at the 2015 Econometric Society World Congress, the 2015 CEME conference on Inference in Nonstandard Problems, Boston College, Boston University, Chicago, Columbia, Harvard, Michigan State, MIT, Princeton, Stanford, U Penn, and Yale for helpful comments, and Keisuke Hirano and Jack Porter for productive discussions. Andrews gratefully acknowledges support from the Silverman (1968) Family Career Development Chair at MIT. Armstrong gratefully acknowledges support from National Science Foundation Grant SES-1628939.

1 Introduction

Researchers often have strong prior beliefs about the sign of the first stage coefficient in instrumental variables models, to the point where the sign can reasonably be treated as known. This paper shows that knowledge of the sign of the first stage coefficient allows us to construct an estimator for the coefficient on the endogenous regressor which is unbiased in finite samples when the reduced form errors are normal with known variance. When the distribution of the reduced form errors is unknown, our results lead to estimators that are asymptotically unbiased under weak IV sequences as defined in Staiger & Stock (1997).

As is well known, the conventional two-stage least squares (2SLS) estimator may be severely biased in overidentified models with weak instruments. Indeed the most common pretest for weak instruments, the Staiger & Stock (1997) rule of thumb which declares the instruments weak when the first stage F statistic is less than 10, is shown in Stock & Yogo (2005) to correspond to a test for the worst-case bias in 2SLS relative to OLS. While the 2SLS estimator performs better in the just-identified case according to some measures of central tendency, in this case it has no first moment.[1] A number of papers have proposed alternative estimators to reduce particular measures of bias, e.g. Angrist & Krueger (1995), Imbens et al. (1999), Donald & Newey (2001), Ackerberg & Devereux (2009), and Harding et al. (2015), but none of the resulting feasible estimators is unbiased either in finite samples or under weak instrument asymptotics. Indeed, Hirano & Porter (2015) show that mean, median, and quantile unbiased estimation are all impossible in the linear IV model with an unrestricted parameter space for the first stage.

We show that by exploiting information about the sign of the first stage we can circumvent this impossibility result and construct an unbiased estimator. Moreover, the resulting estimators have a number of properties which make them appealing for

[1] If we instead consider median bias, 2SLS exhibits median bias when the instruments are weak, though this bias decreases rapidly with the strength of the instruments.

applications. In models with a single instrumental variable, which include many empirical applications, we show that there is a unique unbiased estimator based on the reduced-form and first-stage regression estimates. Moreover, we show that this estimator is substantially less dispersed that the usual 2SLS estimator in finite samples. Under standard ("strong instrument") asymptotics, the unbiased estimator has the same asymptotic distribution as 2SLS, and so is asymptotically efficient in the usual sense. In over-identified models many unbiased estimators exist, and we propose unbiased estimators which are asymptotically efficient when the instruments are strong. Further, we show that in over-identified models we can construct unbiased estimators which are robust to small violations of the first stage sign restriction. We also derive a lower bound on the risk of unbiased estimators in finite samples, and show that this bound is attained in some models.

In contrast to much of the recent weak instruments literature, the focus of this paper is on estimation rather than hypothesis testing or confidence set construction. Our approach is closely related to the classical theory of optimal point estimation (see e.g. Lehmann & Casella (1998)) in that we seek estimators which perform well according to conventional estimation criteria (e.g. risk with respect to a convex loss function) within the class of unbiased estimators. As we note in Section 2.4 below it is straightforward to use results from the weak instruments literature to construct identification-robust tests and confidence sets based on our estimators. As we also note in that section, however, optimal estimation and testing are distinct problems in models with weak instruments and it is not in general the case that optimal estimators correspond to optimal confidence sets or vice versa. Given the important role played by both estimation and confidence set construction in empirical practice, our results therefore complement the literature on identification-robust testing.

The rest of this section discusses the assumption of known first stage sign, introduces the setting and notation, and briefly reviews the related literature. Section 2 introduces the unbiased estimator for models with a single excluded instrument. Section 3 treats models with multiple instruments and introduces unbiased estimators which are robust

to small violations of the first stage sign restriction. Section 4 presents simulation results on the performance of our unbiased estimators. Section 5 discusses illustrative applications using data from Hornung (2014) and Angrist & Krueger (1991). Proofs and auxiliary results are given in a separate appendix.[2]

1.1 Knowledge of the First-Stage Sign

The results in this paper rely on knowledge of the first stage sign. This is reasonable in many economic contexts. In their study of schooling and earnings, for instance, Angrist & Krueger (1991) note that compulsory schooling laws in the United States allow those born earlier in the year to drop out after completing fewer years of school than those born later in the year. Arguing that quarter of birth can reasonably be excluded from a wage equation, they use this fact to motivate quarter of birth as an instrument for schooling. In this context, a sign restriction on the first stage amounts to an assumption that the mechanism claimed by Angrist & Krueger works in the expected direction: those born earlier in the year tend to drop out earlier. More generally, empirical researchers often have some mechanism in mind for why a model is identified at all (i.e. why the first stage coefficient is nonzero) that leads to a known sign for the direction of this mechanism (i.e. the sign of the first stage coefficient).

In settings with heterogeneous treatment effects, a first stage monotonicity assumption is often used to interpret instrumental variables estimates (see Imbens & Angrist 1994, Heckman et al. 2006). In the language of Imbens & Angrist (1994), the monotonicity assumption requires that either the entire population affected by the treatment be composed of "compliers," or that the entire population affected by the treatment be composed of "defiers." Once this assumption is made, our assumption that the sign of the first stage coefficient is known amounts to assuming the researcher knows which of these possibilities (compliers or defiers) holds. Indeed, in the examples where they argue that monotonicity is plausible (involving draft lottery numbers in one case and in-

[2]The appendix is available online at https://sites.google.com/site/isaiahandrews/working-papers

tention to treat in another), Imbens & Angrist (1994) argue that all individuals affected by the treatment are "compliers" for a certain definition of the instrument.

It is important to note, however, that knowledge of the first stage sign is not always a reasonable assumption, and thus that the results of this paper are not always applicable. In settings where the instrumental variables are indicators for groups without a natural ordering, for instance, one typically does not have prior information about signs of the first stage coefficients. To give one example, Aizer & Doyle Jr. (2015) use the fact that judges are randomly assigned to study the effects of prison sentences on recidivism. In this setting, knowledge of the first stage sign would require knowing a priori which judges are more strict.

1.2 Setting

For the remainder of the paper, we suppose that we observe a sample of T observations $(Y_t, X_t, Z_t'),\ t = 1, ..., T$ where Y_t is an outcome variable, X_t is a scalar endogenous regressor, and Z_t is a $k \times 1$ vector of instruments. Let Y and X be $T \times 1$ vectors with row t equal to Y_t and X_t respectively, and let Z be a $T \times k$ matrix with row t equal to Z_t'. The usual linear IV model, written in reduced-form, is

$$\begin{aligned} Y &= Z\pi\beta + U \\ X &= Z\pi + V \end{aligned} . \tag{1}$$

To derive finite-sample results, we treat the instruments Z as fixed and assume that the errors (U, V) are jointly normal with mean zero and known variance-covariance matrix $Var\left((U', V')'\right)$.[3] As is standard (see, for example, D. Andrews et al. (2006)), in contexts with additional exogenous regressors W (for example an intercept), we define

[3]Following the weak instruments literature we focus on models with homogeneous β, which rules out heterogeneous treatment effect models with multiple instruments. In models with treatment effect heterogeneity and a single instrument, however, our results immediately imply an unbiased estimator of the local average treatment effect. In models with multiple instruments, on the other hand, one can use our results to construct unbiased estimators for linear combinations of the local average treatment effects on different instruments. (Since the endogenous variable X is typically a binary treatment in

Y, X, Z as the residuals after projecting out these exogenous regressors. If we denote the reduced-form and first-stage regression coefficients by ξ_1 and ξ_2, respectively, we can see that

$$\begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} = \begin{pmatrix} (Z'Z)^{-1} Z'Y \\ (Z'Z)^{-1} Z'X \end{pmatrix} \sim N\left(\begin{pmatrix} \pi\beta \\ \pi \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right) \tag{2}$$

for

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} = \left(I_2 \otimes (Z'Z)^{-1} Z' \right) Var\left((U', V')' \right) \left(I_2 \otimes (Z'Z)^{-1} Z' \right)'. \tag{3}$$

We assume throughout that Σ is positive definite. Following the literature (e.g. Moreira & Moreira 2013), we consider estimation based solely on (ξ_1, ξ_2), which are sufficient for (π, β) in the special case where the errors (U_t, V_t) are iid over t. All uniqueness and efficiency statements therefore restrict attention to the class of procedures which depend on the data though only these statistics. The conventional generalized method of moments (GMM) estimators belong to this class, so this restriction still allows efficient estimation under strong instruments. We assume that the sign of each component π_i of π is known, and in particular assume that the parameter space for (π, β) is

$$\Theta = \left\{ (\pi, \beta) : \pi \in \Pi \subseteq (0, \infty)^k, \beta \in B \right\} \tag{4}$$

for some sets Π and B. Note that once we take the sign of π_i to be known, assuming $\pi_i > 0$ is without loss of generality since this can always be ensured by redefining Z.

In this paper we focus on models with fixed instruments, normal errors, and known error covariance, which allows us to obtain finite-sample results. As usual, these finite-sample results will imply asymptotic results under mild regularity conditions. Even in models with random instruments, non-normal errors, serial correlation, heteroskedasticity, clustering, or any combination of these, the reduced-form and first stage estimators will be jointly asymptotically normal with consistently estimable covariance matrix

such models, this discussion applies primarily to asymptotic unbiasedness as considered in Appendix B rather than the finite sample model where X and Y are jointly normal.)

Σ under mild regularity conditions. Consequently, the finite-sample results we develop here will imply asymptotic results under both weak and strong instrument asymptotics, where we simply define (ξ_1, ξ_2) as above and replace Σ by an estimator for the variance of ξ to obtain feasible statistics. Appendix B provides the details of these results.[4] In the main text, we focus on what we view as the most novel component of the paper: finite-sample mean-unbiased estimation of β in the normal problem (2).

1.3 Related Literature

Our unbiased IV estimators build on results for unbiased estimation of the inverse of a normal mean discussed in Voinov & Nikulin (1993). More broadly, the literature has considered unbiased estimators in numerous other contexts, and we refer the reader to Voinov & Nikulin for details and references. Recent work by Mueller & Wang (2015) develops a numerical approach for approximating optimal nearly unbiased estimators in variety of nonstandard settings, though they do not consider the linear IV model. To our knowledge the only other paper to treat finite sample mean-unbiased estimation in IV models is Hirano & Porter (2015), who find that unbiased estimators do not exist when the parameter space is unrestricted. In our setting, the sign restriction on the first-stage coefficient leads to a parameter space that violates the assumptions of Hirano & Porter (2015), so that the negative results in that paper do not apply.[5] The nonexistence of unbiased estimators has been noted in other nonstandard econometric contexts by Hirano & Porter (2012).

The broader literature on the finite sample properties of IV estimators is huge: see

[4]The feasible analogs of the finite-sample unbiased estimators discussed here are asymptotically unbiased in general models in the sense of converging in distribution to random variables with mean β. Note that this does not imply convergence of the mean of the feasible estimators to β, since convergence in distribution does not suffice for convergence of moments. Our estimator is thus asymptotically unbiased under weak and strong instruments in the same sense that LIML and just-identified 2SLS, which do not in general have finite-sample moments, are asymptotically unbiased under strong instruments.

[5]In particular, the sign restriction violates Assumption 2.4 of Hirano & Porter (2015), and so renders the negative result in Theorem 2.5 of that paper inapplicable. See Appendix C for details.

Phillips (1983) and Hillier (2006) for references. While this literature does not study unbiased estimation in finite samples, there has been substantial research on higher order asymptotic bias properties: see the references given in the first section of the introduction, as well as Hahn et al. (2004) and the references therein.

Our interest in finite sample results for a normal model with known reduced form variance is motivated by the weak IV literature, where this model arises asymptotically under weak IV sequences as in Staiger & Stock (1997) (see also Appendix B). In contrast to Staiger & Stock, however, our results allow for heteroskedastic, clustered, or serially correlated errors as in Kleibergen (2007). The primary focus of recent work on weak instruments has, however, been on inference rather than estimation. See Andrews (2014) for additional references.

Sign restrictions have been used in other settings in the econometrics literature, although the focus is often on inference or on using sign restrictions to improve population bounds, rather than estimation. Recent examples include Moon et al. (2013) and several papers cited therein, which use sign restrictions to partially identify vector autoregression models. Inference for sign restricted parameters has been treated by D. Andrews (2001) and Gouriéroux et al. (1982), among others.

2 Unbiased Estimation with a Single Instrument

To introduce our unbiased estimators, we first focus on the just-identified model with a single instrument, $k = 1$. We show that unbiased estimation of β in this context is linked to unbiased estimation of the inverse of a normal mean. Using this fact we construct an unbiased estimator for β, show that it is unique, and discuss some of its finite-sample properties. We note the key role played by the first stage sign restriction, and show that our estimator is equivalent to 2SLS (and thus efficient) when the instruments are strong.

In the just-identified context ξ_1 and ξ_2 are scalars and we write

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}.$$

The problem of estimating β therefore reduces to that of estimating

$$\beta = \frac{\pi\beta}{\pi} = \frac{E[\xi_1]}{E[\xi_2]}. \tag{5}$$

The conventional IV estimate $\hat{\beta}_{2SLS} = \frac{\xi_1}{\xi_2}$ is the natural sample-analog of (5). As is well-known, however, this estimator has no integer moments. This lack of unbiasedness reflects the fact that the expectation of the ratio of two random variables is not in general equal to the ratio of their expectations.

The form of (5) nonetheless suggests an approach to deriving an unbiased estimator. Suppose we can construct an estimator $\hat{\tau}$ which (a) is unbiased for $1/\pi$ and (b) depends on the data only through ξ_2. If we then define

$$\hat{\delta}(\xi, \Sigma) = \left(\xi_1 - \frac{\sigma_{12}}{\sigma_2^2}\xi_2\right), \tag{6}$$

we have that $E\left[\hat{\delta}\right] = \pi\beta - \frac{\sigma_{12}}{\sigma_2^2}\pi$, and $\hat{\delta}$ is independent of $\hat{\tau}$.[6] Thus, $E\left[\hat{\tau}\hat{\delta}\right] = E[\hat{\tau}]\,E\left[\hat{\delta}\right] = \beta - \frac{\sigma_{12}}{\sigma_2^2}$, and $\hat{\tau}\hat{\delta} + \frac{\sigma_{12}}{\sigma_2^2}$ will be an unbiased estimator of β. Thus, the problem of unbiased estimation of β reduces to that of unbiased estimation of the inverse of a normal mean.

2.1 Unbiased Estimation of the Inverse of a Normal Mean

A result from Voinov & Nikulin (1993) shows that unbiased estimation of $1/\pi$ is possible if we assume its sign is known. Let Φ and ϕ denote the standard normal cdf and pdf respectively.

Lemma 2.1. *Define*

$$\hat{\tau}\left(\xi_2, \sigma_2^2\right) = \frac{1}{\sigma_2}\frac{1 - \Phi(\xi_2/\sigma_2)}{\phi(\xi_2/\sigma_2)}.$$

[6]Note that the orthogonalization used to construct $\hat{\delta}$ is similar to that used by Kleibergen (2002), Moreira (2003), and the subsequent weak-IV literature to construct identification-robust tests.

For all $\pi > 0$, $E_\pi\left[\hat{\tau}\left(\xi_2, \sigma_2^2\right)\right] = \frac{1}{\pi}$.

The derivation of $\hat{\tau}\left(\xi_2, \sigma_2^2\right)$ in Voinov & Nikulin (1993) relies on the theory of bilateral Laplace transforms, and offers little by way of intuition. Verifying unbiasedness is a straightforward calculus exercise, however: for the interested reader, we work through the necessary derivations in the proof of Lemma 2.1.

From the formula for $\hat{\tau}$, we can see that this estimator has two properties which are arguably desirable for a restricted estimate of $1/\pi$. First, it is positive by definition, thereby incorporating the restriction that $\pi > 0$. Second, in the case where positivity of π is obvious from the data (ξ_2 is very large relative to its standard deviation), it is close to the natural plug-in estimator $1/\xi_2$. The second property is an immediate consequence of a well-known approximation to the tail of the normal cdf, which is used extensively in the literature on extreme value limit theorems for normal sequences and processes (see Equation 1.5.4 in Leadbetter et al. 1983, and the remainder of that book for applications). We discuss this further in Section 2.5.

2.2 Unbiased Estimation of β

Given an unbiased estimator of $1/\pi$ which depends only on ξ_2, we can construct an unbiased estimator of β as suggested above. Moreover, this estimator is unique.

Theorem 2.1. *Define*

$$\begin{aligned}\hat{\beta}_U(\xi, \Sigma) &= \hat{\tau}\left(\xi_2, \sigma_2^2\right)\hat{\delta}(\xi, \Sigma) + \tfrac{\sigma_{12}}{\sigma_2^2} \\ &= \tfrac{1}{\sigma_2}\tfrac{1-\Phi(\xi_2/\sigma_2)}{\phi(\xi_2/\sigma_2)}\left(\xi_1 - \tfrac{\sigma_{12}}{\sigma_2^2}\xi_2\right) + \tfrac{\sigma_{12}}{\sigma_2^2}.\end{aligned}$$

The estimator $\hat{\beta}_U(\xi, \Sigma)$ *is unbiased for* β *provided* $\pi > 0$.

Moreover, if the parameter space (4) contains an open set then $\hat{\beta}_U(\xi, \Sigma)$ *is the unique non-randomized unbiased estimator for* β*, in the sense that any other estimator* $\hat{\beta}(\xi, \Sigma)$ *satisfying*

$$E_{\pi,\beta}\left[\hat{\beta}(\xi, \Sigma)\right] = \beta \ \forall \pi \in \Pi, \beta \in B$$

also satisfies

$$\hat{\beta}\left(\xi,\Sigma\right)=\hat{\beta}_U\left(\xi,\Sigma\right)\ a.s.\,\forall\pi\in\Pi,\beta\in B.$$

Note that the conventional IV estimator can be written as

$$\hat{\beta}_{2SLS}=\frac{\xi_1}{\xi_2}=\frac{1}{\xi_2}\left(\xi_1-\frac{\sigma_{12}}{\sigma_2^2}\xi_2\right)+\frac{\sigma_{12}}{\sigma_2^2}.$$

Thus, $\hat{\beta}_U$ differs from the conventional IV estimator only in that it replaces the plug-in estimate $1/\xi_2$ for $1/\pi$ by the unbiased estimate $\hat{\tau}$. From results in e.g. Baricz (2008), we have that $\hat{\tau}<1/\xi_2$ for $\xi_2>0$, so when ξ_2 is positive $\hat{\beta}_U$ shrinks the conventional IV estimator towards σ_{12}/σ_2^2.[7] By contrast, when $\xi_2<0$, $\hat{\beta}_U$ lies on the opposite side of σ_{12}/σ_2^2 from the conventional IV estimator. Interestingly, one can show that the unbiased estimator is uniformly more likely to correctly sign $\beta-\frac{\sigma_{12}}{\sigma_2^2}$ than is the conventional estimator, in the sense that for $\varphi(x)=1\{x\geq 0\}$,

$$Pr_{\pi,\beta}\left\{\varphi\left(\hat{\beta}_U-\frac{\sigma_{12}}{\sigma_2^2}\right)=\varphi\left(\beta-\frac{\sigma_{12}}{\sigma_2^2}\right)\right\}\geq Pr_{\pi,\beta}\left\{\varphi\left(\hat{\beta}_{2SLS}-\frac{\sigma_{12}}{\sigma_2^2}\right)=\varphi\left(\beta-\frac{\sigma_{12}}{\sigma_2^2}\right)\right\},$$

with strict inequality at some points.[8]

2.3 Risk and Moments of the Unbiased Estimator

The uniqueness of $\hat{\beta}_U$ among nonrandomized estimators implies that $\hat{\beta}_U$ minimizes the risk $E_{\pi,\beta}\ell\left(\tilde{\beta}(\xi,\Sigma)-\beta\right)$ uniformly over π,β and over the class of unbiased estimators $\tilde{\beta}$ for any loss function ℓ such that randomization cannot reduce risk. In particular, by Jensen's inequality $\hat{\beta}_U$ is uniformly minimum risk for any convex loss function ℓ. This includes absolute value loss as well as squared error loss or L^p loss for any $p\geq 1$. However, elementary calculations show that $|\hat{\beta}_U|$ has an infinite pth moment for $p>1$. Thus the fact that $\hat{\beta}_U$ has uniformly minimal risk implies that any unbiased estimator must have an infinite pth moment for any $p>1$. In particular, while $\hat{\beta}_U$ is the uniform

[7] Under weak instrument asymptotics as in Staiger & Stock (1997) and homoskedastic errors, σ_{12}/σ_2^2 is the probability limit of the OLS estimator, though this does not in general hold under weaker assumptions on the error structure.

[8] This property is far from unique to the unbiased estimator, however.

minimum mean absolute deviation unbiased estimator of β, it is minimum variance unbiased only in the sense that all unbiased estimators have infinite variance. We record this result in the following theorem.

Theorem 2.2. *For $\varepsilon > 0$, the expectation of $|\hat{\beta}_U(\xi, \Sigma)|^{1+\varepsilon}$ is infinite for all π, β. Moreover, if the parameter space (4) contains an open set then any unbiased estimator of β has an infinite $1 + \varepsilon$ moment.*

2.4 Relation to Tests and Confidence Sets

As we show in the next subsection, $\hat{\beta}_U$ is asymptotically equivalent to 2SLS when the instruments are strong and so can be used together with conventional standard errors in that case. Even when the instruments are weak the conditioning approach of Moreira (2003) yields valid conditional critical values for arbitrary test statistics and so can be used to construct conditional t-tests based on $\hat{\beta}_U$ which control size. We note, however, that optimal estimation and optimal testing are distinct questions in the context of weak IV (e.g. while $\hat{\beta}_U$ is uniformly minimum risk unbiased for convex loss, it follows from the results of Moreira (2009) that the Anderson-Rubin test, rather than a conditional t-test based on $\hat{\beta}_U$, is the uniformly most powerful unbiased two-sided test in the present just-identified context).[9] Since our focus in this paper is on estimation we do not further pursue the question of optimal testing in this paper. However, properties of tests based on unbiased estimators, particularly in contexts where the Anderson-Rubin test is not uniformly most powerful unbiased (such as one-sided testing and testing in the overidentified model of Section 3), is an interesting topic for future work.[10]

[9] Moreira (2009) establishes this result in the model without a sign restriction, and it is straightforward to show that the result continues to hold in the sign-restricted model.

[10] Absent such results, we suggest reporting the Anderson-Rubin confidence set to accompany the unbiased point estimate. As discussed in Section F.3, the 95% Anderson-Rubin confidence set contains $\hat{\beta}_U$ with probability exceeding 97%, and with probability near 100% except when π is extremely small.

2.5 Behavior of $\hat{\beta}_U$ When π is Large

While the finite-sample unbiasedness of $\hat{\beta}_U$ is appealing, it is also natural to consider performance when the instruments are highly informative. This situation, which we will model by taking π to be large, corresponds to the conventional strong-instrument asymptotics where one fixes the data generating process and takes the sample size to infinity.[11]

As we discussed above, the unbiased and conventional IV estimators differ only in that the former substitutes $\hat{\tau}\left(\xi_2, \sigma_2^2\right)$ for $1/\xi_2$. These two estimators for $1/\pi$ coincide to a high order of approximation for large values of ξ_2. Specifically, as noted in Small (2010) (Section 2.3.4), for $\xi_2 > 0$ we have

$$\sigma_2 \left| \hat{\tau}\left(\xi_2, \sigma_2^2\right) - \frac{1}{\xi_2} \right| \leq \left| \frac{\sigma_2^3}{\xi_2^3} \right|.$$

Thus, since $\xi_2 \xrightarrow{p} \infty$ as $\pi \to \infty$, the difference between $\hat{\tau}\left(\xi_2, \sigma_2^2\right)$ and $1/\xi_2$ converges rapidly to zero (in probability) as π grows. Consequently, the unbiased estimator $\hat{\beta}_U$ (appropriately normalized) has the same limiting distribution as the conventional IV estimator $\hat{\beta}_{2SLS}$ as we take $\pi \to \infty$.

Theorem 2.3. *As $\pi \to \infty$, holding β and Σ fixed,*

$$\pi\left(\hat{\beta}_U - \hat{\beta}_{2SLS}\right) \xrightarrow{p} 0.$$

Consequently, $\hat{\beta}_U \xrightarrow{p} \beta$ and

$$\pi\left(\hat{\beta}_U - \beta\right) \xrightarrow{d} N\left(0, \sigma_1^2 - 2\beta\sigma_{12} + \beta^2\sigma_2^2\right).$$

[11] Formally, in the finite-sample normal IV model (1), strong-instrument asymptotics will correspond to fixing π and taking $T \to \infty$, which under mild conditions on Z and $Var\left((U', V')'\right)$ will result in $\Sigma \to 0$ in (2). However, it is straightforward to show that the behavior of $\hat{\beta}_U$, $\hat{\beta}_{2SLS}$, and many other estimators in this case will be the same as the behavior obtained by holding Σ fixed and taking π to infinity. We focus on the latter case here to simplify the exposition. See Appendix B, which provides asymptotic results with an unknown error distribution, for asymptotic results under $T \to \infty$.

Thus, the unbiased estimator $\hat{\beta}_U$ behaves as the standard IV estimator for large values of π. Consequently, one can show that using this estimator along with conventional standard errors will yield asymptotically valid inference under strong-instrument asymptotics. See Appendix B for details.

3 Unbiased Estimation with Multiple Instruments

We now consider the case with multiple instruments, where the model is given by (1) and (2) with k (the dimension of Z_t, π, ξ_1 and ξ_2) greater than 1. As in Section 1.2, we assume that the sign of each element π_i of the first stage vector is known, and we normalize this sign to be positive, giving the parameter space (4).

Using the results in Section 2 one can construct an unbiased estimator for β in many different ways. For any index $i \in \{1, \ldots, k\}$, the unbiased estimator based on $(\xi_{1,i}, \xi_{2,i})$ will, of course, still be unbiased for β when $k > 1$. One can also take non-random weighted averages of the unbiased estimators based on different instruments. Using the unbiased estimator based on a fixed linear combination of instruments is another possibility, so long as the linear combination preserves the sign restriction. However, such approaches will not adapt to information from the data about the relative strength of instruments and so will typically be inefficient when the instruments are strong.

By contrast, the usual 2SLS estimator achieves asymptotic efficiency in the strongly identified case (modeled here by taking $\|\pi\| \to \infty$) when errors are homoskedastic. In fact, in this case 2SLS is asymptotically equivalent to an infeasible estimator that uses knowledge of π to choose the optimal combination of instruments. Thus, a reasonable goal is to construct an estimator that (1) is unbiased for fixed π and (2) is asymptotically equivalent to 2SLS as $\|\pi\| \to \infty$.[12] In the remainder of this section we first

[12]In the heteroskedastic case, the 2SLS estimator will no longer be asymptotically efficient, and a two-step GMM estimator can be used to achieve the efficiency bound. Because it leads to simpler exposition, and because the 2SLS estimator is common in practice, we consider asymptotic equivalence with 2SLS, rather than asymptotic efficiency in the heteroskedastic case, as our goal. As discussed in

introduce a class of unbiased estimators and then show that a (feasible) estimator in this class attains the desired strong IV efficiency property. Further, we show that in the over-identified case it is possible to construct unbiased estimators which are robust to small violations of the first stage sign restriction. Finally, we derive bounds on the attainable risk of any estimator for finite $\|\pi\|$ and show that, while the unbiased estimators described above achieve optimality in an asymptotic sense as $\|\pi\| \to \infty$ regardless of the direction of π, the optimal unbiased estimator for finite π will depend on the direction of π.

3.1 A Class of Unbiased Estimators

Let

$$\xi(i) = \begin{pmatrix} \xi_{1,i} \\ \xi_{2,i} \end{pmatrix} \text{ and } \Sigma(i) = \begin{pmatrix} \Sigma_{11,ii} & \Sigma_{12,ii} \\ \Sigma_{21,ii} & \Sigma_{22,ii} \end{pmatrix}$$

be the reduced form and first stage coefficients on the ith instrument and their variance matrix, respectively, so that $\hat{\beta}_U(\xi(i), \Sigma(i))$ is the unbiased estimator based on the ith instrument. Given a weight vector $w \in \mathbb{R}^k$ with $\sum_{i=1}^k w_i = 1$, let

$$\hat{\beta}_w(\xi, \Sigma; w) = \sum_{i=1}^k w_i \hat{\beta}_U(\xi(i), \Sigma(i)).$$

Clearly, $\hat{\beta}_w$ is unbiased so long as w is nonrandom. Allowing w to depend on the data ξ, however, may introduce bias through the dependence between the weights and the estimators $\hat{\beta}_U(\xi(i), \Sigma(i))$.

To avoid this bias we first consider a randomized unbiased estimator and then take its conditional expectation given the sufficient statistic ξ to eliminate the randomization. Let $\zeta \sim N(0, \Sigma)$ be independent of ξ, and let $\xi^{(a)} = \xi + \zeta$ and $\xi^{(b)} = \xi - \zeta$. Then $\xi^{(a)}$ and $\xi^{(b)}$ are (unconditionally) independent draws with the same marginal distribution as ξ, save that Σ is replaced by 2Σ. If T is even, $Z'Z$ is the same across the first and

Appendix A.2, however, our approach generalizes directly to efficient estimators in non-homoskedastic settings.

second halves of the sample, and the errors are iid, then $\xi^{(a)}$ and $\xi^{(b)}$ have the same joint distribution as the reduced form estimators based on the first and second half of the sample. Thus, we can think of these as split-sample reduced-form estimates.

Let $\hat{w} = \hat{w}(\xi^{(b)})$ be a vector of data dependent weights with $\sum_{i=1}^{k} \hat{w}_i = 1$. By the independence of $\xi^{(a)}$ and $\xi^{(b)}$,

$$E\left[\hat{\beta}_w(\xi^{(a)}, 2\Sigma; \hat{w}(\xi^{(b)}))\right] = \sum_{i=1}^{k} E\left[\hat{w}_i(\xi^{(b)})\right] \cdot E\left[\hat{\beta}_U(\xi^{(a)}(i), 2\Sigma(i))\right] = \beta. \tag{7}$$

To eliminate the noise introduced by ζ, define the "Rao-Blackwellized" estimator

$$\hat{\beta}_{RB} = \hat{\beta}_{RB}(\xi, \Sigma; \hat{w}) = E\left[\hat{\beta}_w(\xi^{(a)}, 2\Sigma; \hat{w}(\xi^{(b)}))\middle|\xi\right].$$

This gives a class of unbiased estimators, where the estimator depends on the choice of the weight $\hat{w}$. Unbiasedness of $\hat{\beta}_{RB}$ follows immediately from (7) and the law of iterated expectations. While $\hat{\beta}_{RB}$ does not, to our knowledge, have a simple closed form, it can be computed by integrating over the distribution of ζ. This can easily be done by simulation, taking the sample average of $\hat{\beta}_w$ over simulated draws of $\xi^{(a)}$ and $\xi^{(b)}$ while holding ξ at its observed value.

3.2 Equivalence with 2SLS under Strong IV Asymptotics

We now propose a set of weights $\hat{w}$ which yield an unbiased estimator asymptotically equivalent to 2SLS. To motivate these weights, note that for $W = Z'Z$ and e_i the ith standard basis vector, the 2SLS estimator can be written as

$$\hat{\beta}_{2SLS} = \frac{\xi_2' W \xi_1}{\xi_2' W \xi_2} = \sum_{i=1}^{k} \frac{\xi_2' W e_i e_i' \xi_2}{\xi_2' W \xi_2} \frac{\xi_{1,i}}{\xi_{2,i}},$$

which is the GMM estimator with weight matrix $W = Z'Z$. Thus, the 2SLS estimator is a weighted average of the 2SLS estimates based on single instruments, where the weight for estimate $\xi_{1,i}/\xi_{2,i}$ based on instrument i is equal to $\frac{\xi_2' W e_i e_i' \xi_2}{\xi_2' W \xi_2}$. This suggests the unbiased Rao-Blackwellized estimator with weights $\hat{w}_i^*(\xi^{(b)}) = \frac{\xi_2^{(b)\prime} W e_i e_i' \xi_2^{(b)}}{\xi_2^{(b)\prime} W \xi_2^{(b)}}$:

$$\hat{\beta}_{RB}^* = \hat{\beta}_{RB}(\xi, \Sigma; \hat{w}) = E\left[\hat{\beta}_w(\xi^{(a)}, 2\Sigma; \hat{w}^*(\xi^{(b)}))\middle|\xi\right]. \tag{8}$$

The following theorem shows that $\hat{\beta}^*_{RB}$ is asymptotically equivalent to $\hat{\beta}_{2SLS}$ in the strongly identified case, and is therefore asymptotically efficient if the errors are iid.

Theorem 3.1. *Let* $\|\pi\| \to \infty$ *with* $\|\pi\|/\min_i \pi_i = \mathcal{O}(1)$. *Then* $\|\pi\|(\hat{\beta}^*_{RB} - \hat{\beta}_{2SLS}) \xrightarrow{p} 0$.

The condition that $\|\pi\|/\min_i \pi_i = \mathcal{O}(1)$ amounts to an assumption that the "strength" of all instruments is of the same order. As discussed below in Section 3.3, this assumption can be relaxed by redefining the instruments.

To understand why Theorem 3.1 holds, consider the "oracle" weights $w^*_i = \frac{\pi' W e_i e_i' \pi}{\pi' W \pi}$. It is easy to see that $\hat{w}^*_i - w^*_i \xrightarrow{p} 0$ as $\|\pi\| \to \infty$. Consider the oracle unbiased estimator $\hat{\beta}^o_{RB} = \hat{\beta}_{RB}(\xi, \Sigma; w^*)$, and the oracle combination of individual 2SLS estimators $\hat{\beta}^o_{2SLS} = \sum_{i=1}^k w^*_i \frac{\xi_{1,i}}{\xi_{2,i}}$. By arguments similar to those used to show that statistical noise in the first stage estimates does not affect the 2SLS asymptotic distribution under strong instrument asymptotics, it can be seen that $\|\pi\|(\hat{\beta}^o_{2SLS} - \hat{\beta}_{2SLS}) \xrightarrow{p} 0$ as $\|\pi\| \to \infty$. Further, one can show that $\hat{\beta}^o_{RB} = \hat{\beta}_w(\xi, \Sigma; w^*) = \sum_{i=1}^k w^*_i \hat{\beta}_U(\xi(i), \Sigma(i))$. Since this is just $\hat{\beta}^o_{2SLS}$ with $\hat{\beta}_U(\xi(i), \Sigma(i))$ replacing $\xi_{i,1}/\xi_{i,2}$, it follows by Theorem 2.3 that $\|\pi\|(\hat{\beta}^o_{RB} - \hat{\beta}^o_{2SLS}) \xrightarrow{p} 0$. Theorem 3.1 then follows by showing that $\|\pi\|(\hat{\beta}_{RB} - \hat{\beta}^o_{RB}) \xrightarrow{p} 0$, which follows for essentially the same reasons that first stage noise does not affect the asymptotic distribution of the 2SLS estimator but requires some additional argument. We refer the interested reader to the proof of Theorem 3.1 in Appendix A for details.

3.3 Robust Unbiased Estimation

So far, all the unbiased estimators we have discussed required $\pi_i > 0$ for all i. Even when the first stage sign is dictated by theory, however, we may be concerned that this restriction may fail to hold exactly in a given empirical context. To address such concerns, in this section we show that in over-identified models we can construct estimators which are robust to small violations of the sign restriction. Our approach has the further benefit of ensuring asymptotic efficiency when, while $\|\pi\| \to \infty$, the elements π_i may increase at different rates.

Let M be a $k \times k$ invertible matrix such that all elements are strictly positive, and

$$\tilde{\xi} = (I_2 \otimes M)\xi, \qquad \tilde{\Sigma} = (I_2 \otimes M)\Sigma(I_2 \otimes M)', \qquad \tilde{W} = M^{-1\prime} W M^{-1}.$$

The GMM estimator based on $\tilde{\xi}$ and $\tilde{W}$ is numerically equivalent to the GMM estimator based on ξ and W. In particular, for many choices of W, including all those discussed above, estimation based on $(\tilde{\xi}, \tilde{W}, \tilde{\Sigma})$ is equivalent to estimation based on instruments ZM^{-1} rather than Z.

Note that for $\tilde{\pi} = M\pi$, $\tilde{\xi}$ is normally distributed with mean $(\tilde{\pi}'\beta, \tilde{\pi}')'$ and variance $\tilde{\Sigma}$. Thus, if we construct the estimator $\hat{\beta}^*_{RB}$ from $(\tilde{\xi}, \tilde{W}, \tilde{\Sigma})$ instead of (ξ, W, Σ), we obtain an unbiased estimator provided $\tilde{\pi}_i > 0$ for all i. Since all elements of M are strictly positive this is a strictly weaker condition than $\pi_i > 0$ for all i. By Theorem 3.1, $\hat{\beta}^*_{RB}$ constructed from from $\tilde{\xi}$ and $\tilde{W}$ will be asymptotically efficient as $\|\tilde{\pi}\| \to \infty$ so long as $\tilde{\pi} = M\pi$ is nonnegative and satisfies $\|\tilde{\pi}\|/\min_i \tilde{\pi}_i = \mathcal{O}(1)$. Note, however, that

$$\min_i \tilde{\pi}_i \geq (\min_{i,j} M_{ij})\|\pi\| = (\min_{i,j} M_{ij})\frac{\|\pi\|}{\|M\pi\|}\|\tilde{\pi}\| \geq (\min_{i,j} M_{ij})\left(\inf_{\|u\|=1}\frac{\|u\|}{\|Mu\|}\right)\|\tilde{\pi}\|$$

so $\|\tilde{\pi}\|/\min_i \tilde{\pi}_i = \mathcal{O}(1)$ now follows automatically from $\|\pi\| \to \infty$.

Conducting estimation based on $\tilde{\xi}$ and $\tilde{W}$ offers a number of advantages for many different choices of M. One natural class of transformations M is

$$M = \begin{bmatrix} 1 & c & c & \cdots & c \\ c & 1 & c & \cdots & c \\ c & c & 1 & \cdots & c \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c & c & c & \cdots & 1 \end{bmatrix} Diag(\Sigma_{22})^{-\frac{1}{2}} \tag{9}$$

for $c \in [0, 1)$ and $Diag(\Sigma_{22})$ the matrix with the same diagonal as Σ_{22} and zeros elsewhere. For a given c, denote the estimator $\hat{\beta}^*_{RB}$ based on the corresponding $(\tilde{\xi}, \tilde{W}, \tilde{\Sigma})$ by $\hat{\beta}^*_{RB,c}$. One can show that $\hat{\beta}^*_{RB,0} = \hat{\beta}^*_{RB}$ based on (ξ, W, Σ), and going forward we let $\hat{\beta}^*_{RB}$ denote $\hat{\beta}^*_{RB,0}$.

We can interpret c as specifying a level of robustness to violations on the sign restriction for π_i. In particular, for a given choice of c, $\tilde{\pi}$ will satisfy the sign restriction provided that for each i,

$$-\pi_i/\sqrt{\Sigma_{22,ii}} < c \cdot \sum_{j \neq i} \pi_j/\sqrt{\Sigma_{22,jj}},$$

that is, provided the expected z-statistic for testing that each wrong-signed π_i is equal to zero is less than c times the sum of the expected z-statistics for $j \neq i$. Larger values of c provide a greater degree of robustness to violations of the sign restriction, while all choices of $c \in (0, 1)$ yield asymptotically equivalent estimators as $\|\pi\| \to \infty$. For finite values of π however, different choices of c yield different estimators, so we explore the effects of different choices below using the Angrist & Krueger (1991) dataset. Determining the optimal choice of c for finite values of π is an interesting topic for future research.

3.4 Bounds on the Attainable Risk

While the class of estimators given above has the desirable property of asymptotic efficiency as $\|\pi\| \to \infty$, it is useful to have a benchmark for the performance for finite π. In Appendix D, we derive a lower bound for the risk of any unbiased estimator at a given π^*, β^*. The bound is based on the risk in a submodel with a single instrument and, as in the single instrument case, shows that any unbiased estimator must have an infinite $1 + \varepsilon$ absolute moment for $\varepsilon > 0$. In certain cases, which include large parts of the parameter space under homoskedastic errors (U_t, V_t), the bound can be attained. The estimator that attains the bound turns out to depend on the value π^*, which shows that no uniform minimum risk unbiased estimator exists. See Appendix D for details.

4 Simulations

In this section we present simulation results on the performance of our unbiased estimators. We study the model with normal errors and known reduced-form variance. We

first consider models with a single instrument and then turn to over-identified models. Since the parameter space in the single-instrument model is small, we are able to obtain comprehensive simulation results in this case, studying performance over a wide range of parameter values. In the over-identified case, by contrast, the parameter space is too large to comprehensively explore by simulation so we instead calibrate our simulations to the Staiger & Stock (1997) specifications for the Angrist & Krueger (1991) dataset.

4.1 Performance with a Single Instrument

The estimator $\hat{\beta}_U$ based on a single instrument plays a central role in all of our results, so in this section we examine the performance of this estimator in simulation. For purposes of comparison we also discuss results for the two-stage least squares estimator $\hat{\beta}_{2SLS}$. The lack of moments for $\hat{\beta}_{2SLS}$ in the just-identified context renders some comparisons with $\hat{\beta}_U$ infeasible, however, so we also consider the performance of the Fuller (1977) estimator with constant one,

$$\hat{\beta}_{FULL} = \frac{\xi_2\xi_1 + \sigma_{12}}{\xi_2^2 + \sigma_2^2}$$

which we define as in Mills et al. (2014).[13] Note that in the just-identified case considered here $\hat{\beta}_{FULL}$ also coincides with the bias-corrected 2SLS estimator (again, see Mills et al.).

While the model (2) has five parameters in the single-instrument case, $(\beta, \pi, \sigma_1^2, \sigma_{12}, \sigma_2^2)$, an equivariance argument implies that for our purposes it suffices to fix $\beta = 0$, $\sigma_1 = \sigma_2 = 1$ and consider the parameter space $(\pi, \sigma_{12}) \in (0, \infty) \times [0, 1)$. See Appendix E for details. Since this parameter space is just two-dimensional, we can fully explore it via simulation.

[13] In the case where U_t and V_t are correlated or heteroskedastic across t, the definition of $\hat{\beta}_{FULL}$ above is the natural extension of the definition considered in Mills et al. (2014).

4.1.1 Estimator Location

We first compare the bias of $\hat{\beta}_U$ and $\hat{\beta}_{FULL}$ (we omit $\hat{\beta}_{2SLS}$ from this comparison, as it does not have a mean in the just-identified case). We consider $\sigma_{12} \in \{0.1, 0.5, 0.95\}$ and examine a wide range of values for $\pi > 0$.[14] These results are plotted in the first panel of Figure 1.

If rather than mean bias we instead consider median bias, we find that $\hat{\beta}_U$ and $\hat{\beta}_{2SLS}$ generally exhibit smaller median bias than $\hat{\beta}_{FULL}$. There is no ordering between $\hat{\beta}_U$ and $\hat{\beta}_{2SLS}$ in terms of median bias, however, as the median bias of $\hat{\beta}_U$ is smaller than that of $\hat{\beta}_{2SLS}$ for very small values of π, while the median bias of $\hat{\beta}_{2SLS}$ is smaller for larger values π. A plot of median bias is given in Appendix F.1.

4.1.2 Estimator Absolute Deviation

We examine the distribution of the absolute deviation of each estimator from the true parameter value. The last three panels of Figure 1 plot the 10th, 50th, and 90th percentiles of absolute deviation of the estimators considered from the true value β for three values of σ_{12}. We plot the log quantiles of absolute deviation (or equivalently the quantiles of log absolute deviation) for the sake of visibility. Here, and in additional unreported simulation results, we find that $\hat{\beta}_U$ has smaller median absolute deviation than $\hat{\beta}_{IV}$ uniformly over the parameter space. The 10th and 90th percentiles of the absolute deviation are also lower for $\hat{\beta}_U$ than $\hat{\beta}_{IV}$ for much of the parameter space, though we find that there is not a uniform ranking for all percentiles. The Fuller estimator has low median absolute deviation over much of the parameter space, but performs worse than both $\hat{\beta}_U$ and $\hat{\beta}_{IV}$ in certain cases, such as when $\sigma_{12} = 0.95$ and the first stage coefficient is small. Turning to mean absolute deviation, we find that

[14] We restrict attention to $\pi \geq 0.16$ in the bias plots. Since the first stage F-statistic is $F = \xi_2^2$ in the present context, this corresponds to $E[F] \geq 1.026$. The expectation of $\hat{\beta}_U$ ceases to exist at $\pi = 0$, and for π close to zero the heavy tails of $\hat{\beta}_U$ make computing the expectation very difficult. Indeed, we use numerical integration rather than monte-carlo integration here because it allows us to consider smaller values π. We thank an anonymous referee for this suggestion.

the mean absolute deviation of $\hat{\beta}_U$ from β exceeds that of $\hat{\beta}_{FULL}$ except in cases with very high ρ and small π, while as already noted the mean absolute deviation of $\hat{\beta}_{IV}$ is infinite.

Thus, over much of the parameter space the unbiased estimator is more concentrated around the true parameter value than the 2SLS estimator, according to a variety of different measures of concentration. It would be interesting to decompose the deviations from the true parameter value into bias and variance components. Unfortunately, however, the lack of second moments of both the 2SLS and unbiased estimators means that the variance is infinite in both cases and therefore does not yield a useful comparison. To get around this, we consider the distribution of the absolute deviation of each estimator from the median of the estimator as a location free measure of dispersion. In Appendix F.2, we examine this numerically and find a stochastic dominance relation in which the unbiased estimator is less dispersed than the 2SLS estimator and more dispersed than the Fuller estimator uniformly over the parameter space.

4.2 Performance with Multiple Instruments

In models with multiple instruments, if we assume that errors are homoskedastic an equivariance argument closely related to that in just-identified case again allows us to reduce the dimension of the parameter space. Unlike in the just-identified case, however, the matrix $Z'Z$ and the direction of the first stage, $\pi/\|\pi\|$, continue to matter (see Appendix E for details). As a result, the parameter space is too large to fully explore by simulation, so we instead calibrate our simulations to the Staiger & Stock (1997) specifications for the 1930-1939 cohort in the Angrist & Krueger (1991) data. While there is statistically significant heteroskedasticity in this data, this significance appears to be the result of the large sample size rather than substantively important deviations from homoskedasticity. In particular, procedures which assume homoskedasticity produce very similar answers to heteroskedasticity-robust procedures when applied to this data. Thus, given that homoskedasticity leads to a reduction of the parameter space

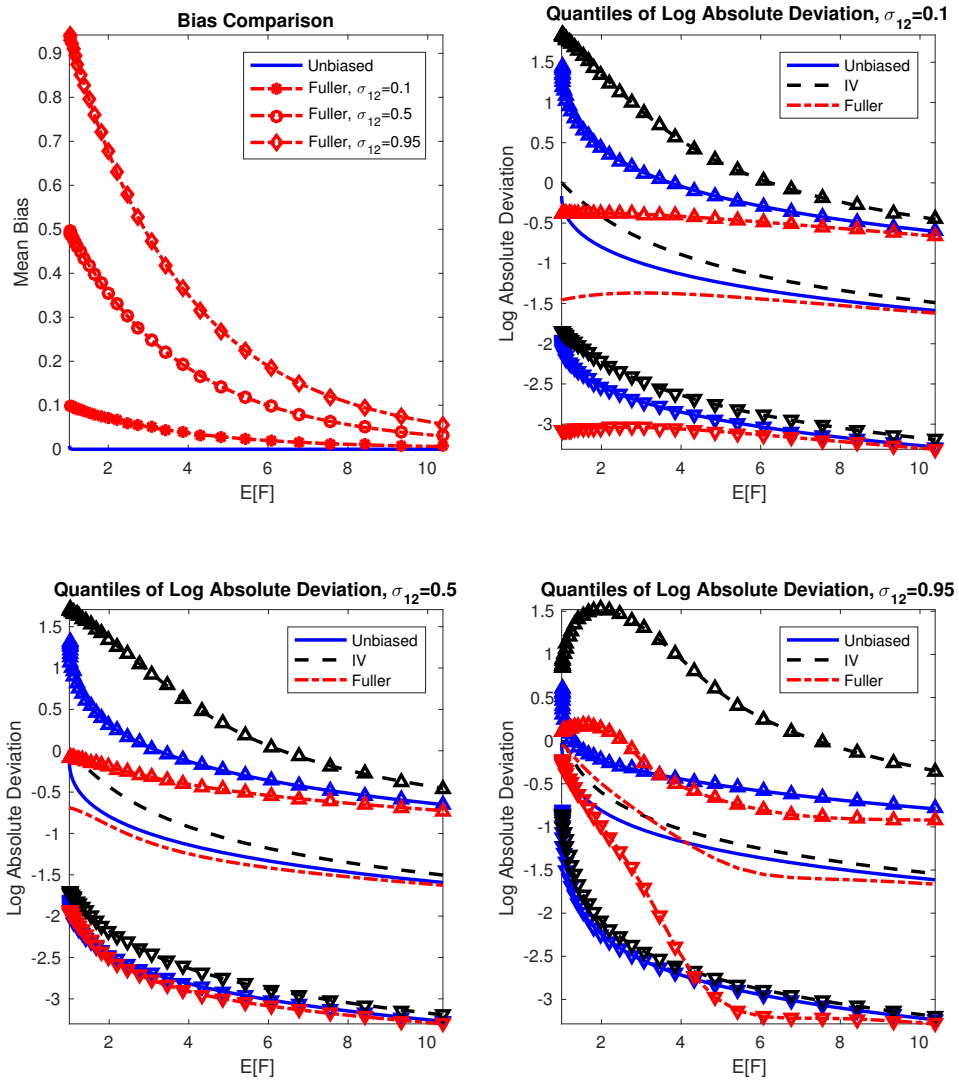


Figure 1: The first panel plots the bias of single-instrument estimators, calculated by numerical integration, against the mean $E[F]$ of first-stage F-statistic. The remaining panels plot log quantiles of absolute deviation from the true value of β for unbiased estimator, 2SLS, and Fuller, for three values of σ_{12}. The lines corresponding to the median are plotted without markers, while the lines corresponding to the 90th and 10th percentiles are plotted with upward and downward pointing triangles, respectively. The absolute deviation results are based on 10 million simulation draws.

as discussed above, we impose homoskedasticity in our simulations.

In each of the four Staiger & Stock (1997) specifications we estimate $\pi/\|\pi\|$ and $Z'Z$ from the data (ensuring, as discussed in Appendix G, that $\pi/\|\pi\|$ satisfies the sign restriction). After reducing the parameter space by equivariance and calibrating $Z'Z$ and $\pi/\|\pi\|$ to the data, the model has two remaining free parameters: the norm of the first stage, $\|\pi\|$, and the correlation σ_{UV} between the reduced-form and first-stage errors. We examine behavior for a range of values for $\|\pi\|$ and for $\sigma_{UV} \in \{0.1, 0.5, 0.95\}$. Further details on the simulation design are given in Appendix G.

For each parameter value we simulate the performance of $\hat{\beta}_{2SLS}$, $\hat{\beta}_{FULL}$ (which is again the Fuller estimator with constant equal to one), and $\hat{\beta}^*_{RB}$ as defined in Section 3.2. We also consider the robust estimators $\hat{\beta}^*_{RB,c}$ discussed in Section 3.3 for $c \in \{0.1, 0.5, 0.9\}$, but find that all three choices produce very similar results and so focus on $c = 0.5$ to simplify the graphs.[15] Even with a million simulation replications, simulation estimates of the bias for the unbiased estimators (which we know to be zero from the results of Section 3) remain noisy relative to e.g. the bias in 2SLS in some calibrations, so we do not plot the bias estimates and instead focus on the mean absolute deviation (MAD) $E_{\pi,\beta}\left[\left|\hat{\beta} - \beta\right|\right]$ since, unlike in the just-identified case, the MAD for 2SLS is now finite. We also plot the lower bound on the mean absolute deviation of unbiased estimators discussed in Section 3.4. The results are plotting in Figure 2.

Several features become clear from these results. As expected, the performance of 2SLS is typically worse for models with more instruments or with a higher degree of correlation between the reduced-form and first-stage errors (i.e. higher σ_{UV}). The robust unbiased estimator $\hat{\beta}_{RB,0.5}$ generally outperforms $\hat{\beta}^*_{RB} = \hat{\beta}^*_{RB,0}$. Since the estimators with $c = 0.1$ and $c = 0.9$ perform very similarly to that with $c = 0.5$, they outperform $\hat{\beta}^*_{RB}$ as well. The gap in performance between the RB estimators and the lower bound on MAD over the class of all unbiased estimators is typically larger in specifications with more instruments. Interestingly, we see that the Fuller estimator often performs quite well, and has MAD close to or below the lower bound for the class of unbiased

[15] All results for the RB estimators are based on $1,000$ draws of ζ.

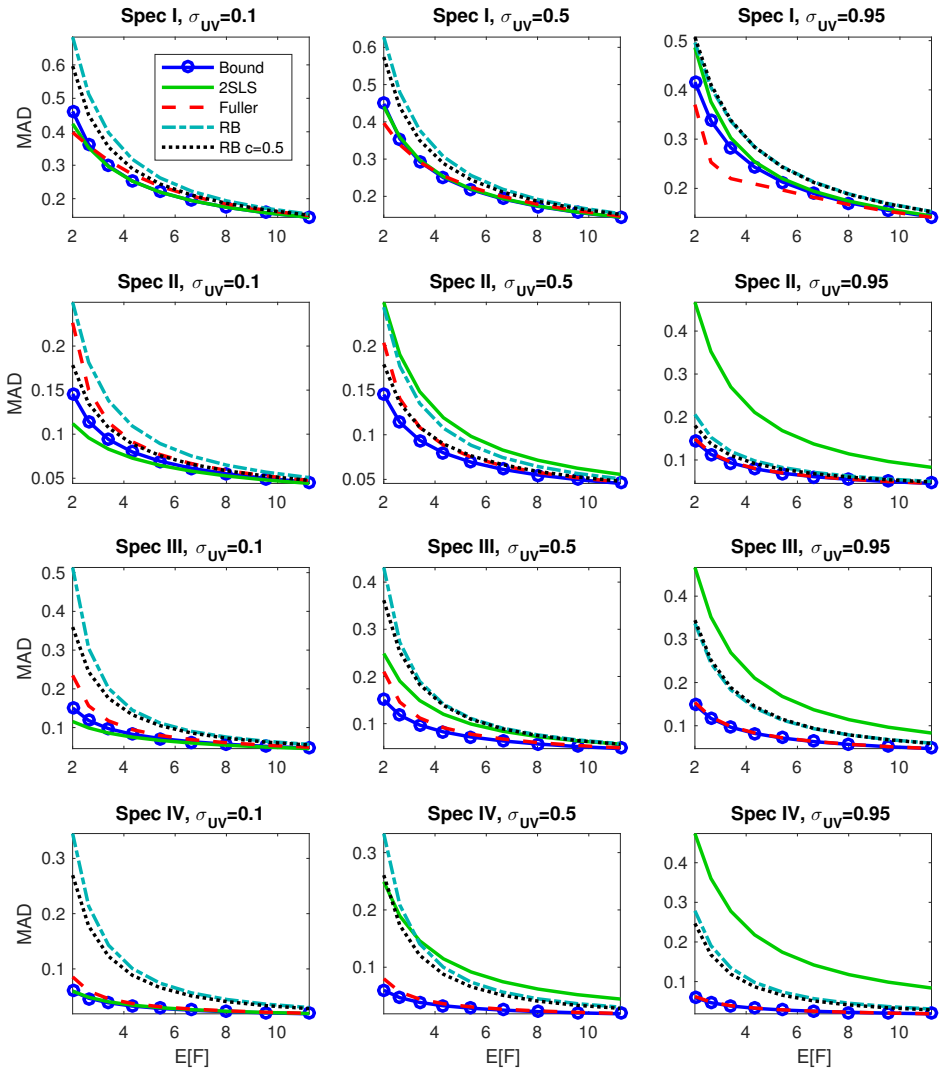


Figure 2: Mean absolute deviation of estimators in simulations calibrated to specification I-IV of Staiger & Stock (1997). These specifications have $k = 3, 30, 28$, and 178 instruments, respectively. Results for specifications I-III are based on 1 million simulation draws, while results for specification IV are based on 100,000 simulation draws.

estimators in most designs. While this estimator is biased, its bias decreases quickly in $\|\pi\|$ in the designs considered. Thus, at least in the homoskedastic case, this estimator seems a potentially appealing choice if we are willing to accept bias for small values of π.

5 Empirical Applications

We calculate our proposed estimators in two empirical applications. First, we consider the data and specifications used in Hornung (2014) to examine the effect of seventeenth century migrations on productivity. For our second application, we study the Staiger & Stock (1997) specifications for the Angrist & Krueger (1991) dataset on the relationship between education and labor market earnings. Before continuing, we present a step-by-step description of the implementation of our estimators.

5.1 Implementation

To describe the implementation in a general setup, we introduce additional notation to explicitly allow for additional exogenous variables (such as a constant). We have observations $t = 1, \ldots, T$ with $\tilde{Y}_t$ a scalar outcome variable, $\tilde{X}_t$ a scalar endogenous variable, $\tilde{Z}_t$ a $k \times 1$ vector of instruments and W_t a vector of additional control variables. Let $\tilde{Y} = (\tilde{Y}_1, \ldots, \tilde{Y}_T)'$, $\tilde{X} = (\tilde{X}_1, \ldots, \tilde{X}_T)'$, $\tilde{Z} = (\tilde{Z}_1, \ldots, \tilde{Z}_T)'$ and $W = (W_1, \ldots, W_T)'$. Let $Y = (I - W(W'W)^{-1}W')\tilde{Y}$, $X = (I - W(W'W)^{-1}W')\tilde{X}$ and $Z = (I - W(W'W)^{-1}W')\tilde{Z}$ denote the residuals from regressing $\tilde{Y}$, $\tilde{X}$ and $\tilde{Z}$ on W, as described in the introduction.

Our estimates are obtained using the following steps.

1.) Let ξ_1 and ξ_2 denote the estimates of the coefficient on $\tilde{Z}_t$ in the regressions of $\tilde{Y}_t$ and $\tilde{X}_t$ respectively on $\tilde{Z}_t$ and W_t, and let $\hat{U}_t$ and $\hat{V}_t$ denote residuals from these regressions. Let $\hat{\Sigma}$ denote an estimate of the variance-covariance matrix of $(\xi_1', \xi_2')'$. If the observations are independent (but possibly heteroskedastic), we

can use the heteroskedasticity robust estimate

$$(I_2 \otimes (Z'Z)^{-1}) \left[\sum_{t=1}^{T} \begin{pmatrix} \hat{U}_t^2 Z_t Z_t' & \hat{U}_t \hat{V}_t Z_t Z_t' \\ \hat{U}_t \hat{V}_t Z_t Z_t' & \hat{V}_t^2 Z_t Z_t' \end{pmatrix} \right] (I_2 \otimes (Z'Z)^{-1}).$$

We use this estimate in our application based on Angrist & Krueger (1991), while for our application based on Hornung (2014) we follow Hornung and use a clustering-robust variance estimator. Likewise, in time-series contexts one could use a serial-correlation robust variance estimator, e.g. Newey & West (1987) here.

2.) In the case of a single instrument (so Z_t is scalar), the estimate is given by $\hat{\beta}_U(\xi, \hat{\Sigma})$ where $\hat{\beta}_U(\cdot, \cdot)$ is defined in Theorem 2.1.

3.) In the case with $k > 1$ instruments, let $\hat{\Sigma}_{22}$ denote the lower-right $k \times k$ submatrix of $\hat{\Sigma}$, and let M be the matrix given in (9) with Σ_{22} replaced by $\hat{\Sigma}_{22}$ for some choice of c between 0 and 1 (we find that $c = .5$ works well in our Monte Carlos). Let $\tilde{\xi} = (I_2 \otimes M)\xi$ and $\tilde{\Sigma} = (I_2 \otimes M)\hat{\Sigma}(I_2 \otimes M)'$. Let $\tilde{\Sigma}(i)$ denote the 2×2 symmetric matrix with diagonal elements given by the i, i and $(k+i), (k+i)$ elements of $\tilde{\Sigma}$ respectively and off-diagonal element given by the $i, (k+i)$ element of $\tilde{\Sigma}$. Generate S independent $N(0, \tilde{\Sigma})$ vectors $\zeta_1, \ldots, \zeta_S$. Let $\tilde{\xi}_1$ and $\zeta_{s,1}$ denote the $k \times 1$ vectors with elements 1 through k of $\tilde{\xi}$ and ζ_s respectively and let $\tilde{\xi}_2$ and $\zeta_{s,2}$ denote the $k \times 1$ vectors with elements $k+1$ through $2k$ of $\tilde{\xi}$ and ζ_s respectively. Let $\tilde{\xi}(i) = (\tilde{\xi}_{1,i}, \tilde{\xi}_{2,i})'$ and let $\tilde{\zeta}_s(i) = (\tilde{\zeta}_{s,1,i}, \tilde{\zeta}_{s,2,i})'$. Let

$$\hat{\beta}_s = \sum_{i=1}^{k} w_{i,s} \hat{\beta}_U(\tilde{\xi}(i) + \zeta_s(i), 2\tilde{\Sigma}(i))$$

where $\hat{\beta}_U(\cdot, \cdot)$ is defined in Theorem 2.1 and

$$w_{i,s} = \frac{(\tilde{\xi}_2 - \zeta_{s,2})' M^{-1\prime}(Z'Z) M^{-1} e_i e_i' (\tilde{\xi}_2 - \zeta_{s,2})}{(\tilde{\xi}_2 - \zeta_{s,2})' M^{-1\prime}(Z'Z) M^{-1} (\tilde{\xi}_2 - \zeta_{s,2})}.$$

The estimator is given by the average over S simulation draws:

$$\hat{\beta} = \frac{1}{S} \sum_{i=1}^{n} \hat{\beta}_s.$$

In our application, we use $S = 100,000$ simulation draws.

5.2 Hornung (2014)

Hornung (2014) studies the long term impact of the flight of skilled Huguenot refugees from France to Prussia in the seventeenth century. He finds that regions of Prussia which received more Huguenot refugees during the late seventeenth century had a higher level of productivity in textile manufacturing at the start of the nineteenth century. To address concerns over endogeneity in Huguenot settlement patterns and obtain an estimate for the causal effect of skilled immigration on productivity, Hornung (2014) considers specifications which instrument Huguenot immigration to a given region using population losses due to plague at the end of the Thirty Years' War. For more information on the data and motivation of the instrument, see Hornung (2014).

Hornung's argument for the validity of his instrument clearly implies that the first-stage effect should be positive, but the relationship between the instrument and the endogenous regressors appears to be fairly weak. In particular, the four IV specifications reported in Tables 4 and 5 of Hornung (2014) have first-stage F-statistics of 3.67, 4.79, 5.74, and 15.35, respectively. Thus, it seems that the conventional normal approximation to the distribution of IV estimates may be unreliable in this context. In each of the four main IV specifications considered by Hornung, we compare 2SLS and Fuller (again with constant equal to one) to our estimator. Since there is only a single instrument in this context, the model is just-identified and the unbiased estimator is unique. In each specification we also compute and report an identification-robust Anderson-Rubin confidence set for the coefficient on the endogenous regressor. The results are reported in Table 1.

As we can see from Table 1, our unbiased estimates in specifications I-III are smaller than the 2SLS estimates computed in Hornung (2014) (the unbiased estimate is smaller in specification IV as well, though the difference only appears in the fourth decimal place). Fuller estimates are, in turn, smaller than our unbiased estimates. Nonetheless, difference between the 2SLS and unbiased estimates is less than half of the 2SLS standard error in every specification. In specifications I-III, where the instruments are

Specification	Estimator	I	II	III	IV
X : Percent Huguenots in 1700	2SLS	3.48	3.38	1.67	
	Fuller	3.17	3.08	1.59	
	Unbiased	3.24	3.14	1.61	
X : log Huguenots in 1700	2SLS				0.07
	Fuller				0.07
	Unbiased				0.07
95% AR Confidence Set		$(-\infty,59.23]\cup[1.55,\infty)$	[1.64,19.12]	[-0.45,5.93]	[-0.01,0.16]
Other controls		Yes	Yes	Yes	Yes
Observations		150	150	186	186
Number of Towns		57	57	71	71
First Stage F-Statistic		3.67	4.79	5.74	15.35

Table 1: Results in Hornung (2014) data. Specifications in columns I and II correspond to Table 4 columns (3) and (5) in Hornung (2014), respectively, while columns III and IV correspond to Table 5 columns (3) and (6) in Hornung (2014). Y =log output, X as indicated, and Z =unadjusted population losses in I, interpolated population losses in II, and population losses averaged over several data sources in III and IV. See Hornung (2014). The 2SLS and Fuller rows report two stage least squares and Fuller estimates, respectively, while Unbiased reports $\hat{\beta}_U$. Other controls include a constant, a dummy for whether a town had relevant textile production in 1685, measurable inputs to the production process, and others as in Hornung (2014). As in Hornung (2014), all covariance estimates are clustered at the town level. Note that the unbiased and Fuller estimates, as well as the AR confidence sets, have been updated to correct an error in the March 22, 2015 version of the present paper.

relatively weak, the 95% AR confidence sets are substantially wider than 95% confidence sets calculated using 2SLS standard errors, while in specification IV the AR confidence set is fairly similar to the conventional 2SLS confidence set.

5.3 Angrist & Krueger (1991)

Angrist & Krueger (1991) are interested in the relationship between education and labor market earnings. They argue that students born later in the calendar year face a longer period of compulsory schooling than those born earlier in the calendar year, and that quarter of birth is a valid instrument for years of schooling. As we note above their argument implies that the sign of the first-stage effect is known. A substantial literature, beginning with Bound et al. (1995), notes that the relationship between the instruments and the endogenous regressor appears to be quite weak in some specifications considered in Angrist & Krueger (1991). Here we consider four specifications from Staiger & Stock (1997), based on the 1930-1939 cohort. See Angrist & Krueger (1991) and Staiger & Stock (1997) for more on the data and specification.

We calculate unbiased estimators $\hat{\beta}^*_{RB}$, $\hat{\beta}^*_{RB,0.1}$, $\hat{\beta}^*_{RB,0.5}$, and $\hat{\beta}^*_{RB,0.9}$. In all cases we take $W = Z'Z$. To calculate confidence sets we use the quasi-CLR (or GMM-M) test of Kleibergen (2005), which simplifies to the CLR test of Moreira (2003) under homoskedasticity and so delivers nearly-optimal confidence sets in that case (see Mikusheva 2010). Thus, since as discussed above the data in this application appears reasonably close to homoskedasticity, we may reasonably expect the quasi-CLR confidence set to perform well. All results are reported in Table 2.

A few points are notable from these results. First, we see that in specifications I and II, which have the largest first stage F-statistics, the unbiased estimates are quite close to the other point estimates. Moreover, in these specifications the choice of c makes little difference. By contrast, in specification III, where the instruments appear to be quite weak, the unbiased estimates differ substantially, with $\hat{\beta}^*_{RB}$ yielding a negative point estimate and $\hat{\beta}^*_{RB,c}$ for $c \in \{0.1, 0.5, 0.9\}$ yielding positive estimates substantially larger

Specification	I	II	III	IV
	$\hat{\beta}$	$\hat{\beta}$	$\hat{\beta}$	$\hat{\beta}$
2SLS	0.099	0.081	0.060	0.081
Fuller	0.100	0.084	0.058	0.098
LIML	0.100	0.084	0.057	0.098
$\hat{\beta}^*_{RB}$,	0.097	0.085	-0.041	0.056
$\hat{\beta}_{RB}$, $c = 0.1$	0.098	0.083	0.135	0.066
$\hat{\beta}_{RB}$, $c = 0.5$	0.098	0.083	0.135	0.066
$\hat{\beta}_{RB}$, $c = 0.9$	0.098	0.083	0.135	0.066
First Stage F	30.582	4.625	1.579	1.823
QCLR CS	[0.059,0.144]	[0.046,0.127]	[-0.588,0.668]	[0.056,0.150]
Controls				
Base Controls	Yes	Yes	Yes	Yes
Age, Age^2	No	No	Yes	Yes
SOB	No	No	No	Yes
Instruments				
QOB	Yes	Yes	Yes	Yes
QOB*YOB	No	Yes	Yes	Yes
QOB*SOB	No	No	No	Yes
# instruments	3	30	28	178
Observations	329,509	329,509	329,509	329,509

Table 2: Results for Angrist & Krueger (1991) data. Specifications as in Staiger & Stock (1997): Y =log weekly wages, X=years of schooling, instruments Z and exogenous controls as indicated. QCLR is the is the quasi-CLR (or GMM-M) confidence set of Kleibergen (2005). Unbiased estimators calculated by averaging over 100,000 draws of ζ.

than the other estimators considered.[16] A similar, though less pronounced, version of this phenomenon arises in specification IV, where unbiased estimates are smaller than those based on conventional methods and $\hat{\beta}^*_{RB}$ is almost 20% smaller than estimates based on other choices of c.

As in the simulations there is very little difference between the estimates for $c \in \{0.1, 0.5, 0.9\}$. In particular, while not exactly the same, the estimates coincide once rounded to three decimal places in all specifications. Given that these estimators are more robust to violations of the sign restriction than that with $c = 0$, we think it makes more sense to focus on these estimates.

6 Conclusion

In this paper, we show that a sign restriction on the first stage suffices to allow finite-sample unbiased estimation in linear IV models with normal errors and known reduced-form error covariance. Our results suggest several avenues for further research. First, while the focus of this paper is on estimation, recent work by Mills et al. (2014) finds good power for particular identification-robust conditional t-tests, suggesting that it may be interesting to consider tests based on our unbiased estimators, particularly in over-identifed contexts where the Anderson-Rubin test is no longer uniformly most powerful unbiased. More broadly, it may be interesting to study other ways to use the knowledge of the first stage sign, both for testing and estimation purposes.

[16] All unbiased estimates are calculated by averaging over 100,000 draws of ζ. For all estimates except $\hat{\beta}^*_{RB}$ in specification III, the residual randomness is small. For $\hat{\beta}^*_{RB}$ in specification III, however, redrawing ζ yields substantially different point estimates. This issue persists even if we increase the number of ζ draws to 1,000,000.

References

Ackerberg, D. & Devereux, P. (2009), 'Improved JIVE estimators for overidentified linear models with and without heteroskedasticity', *Review of Economics and Statistics* **91**, 351–362.

Aizer, A. & Doyle Jr., J. J. (2015), 'Juvenile incarceration, human capital and future crime: Evidence from randomly-assigned judges', *Quarterly Journal of Economics* **130**(2), 759–803.

Andrews, D., Moreira, M. & Stock, J. (2006), 'Optimal two-sided invariant similar tests of instrumental variables regression', *Econometrica* **74**, 715–752.

Andrews, D. W. K. (2001), 'Testing when a parameter is on the boundary of the maintained hypothesis', *Econometrica* **69**(3), 683–734.

Andrews, I. (2014), Conditional linear combination tests for weakly identified models. Unpublished Manuscript.

Angrist, J. D. & Krueger, A. B. (1991), 'Does compulsory school attendance affect schooling and earnings?', *The Quarterly Journal of Economics* **106**(4), 979–1014.

Angrist, J. & Krueger, A. (1995), 'Split-sample instrumental variables estimates of the return to schooling', *Journal of Business and Economic Statistics* **13**, 225–235.

Baricz, Á. (2008), 'Mills' ratio: Monotonicity patterns and functional inequalities', *Journal of Mathematical Analysis and Applications* **340**(2), 1362–1370.

Bound, J., Jaeger, D. & Baker, R. (1995), 'Problems with instrumental variables estimation when the correlation between the instruments and the endogeneous explanatory variable is weak', *Journal of the American Statistical Association* **90**, 443–450.

Donald, S. & Newey, W. (2001), 'Choosing the number of instruments', *Econometrica* **69**, 1161–1191.

Fuller, W. (1977), 'Some properties of a modification of the limited information maximum likelihood estimator', *Econometrica* **45**, 939–953.

Gouriéroux, C., Holly, A. & Monfort, A. (1982), 'Likelihood ratio test, wald test, and kuhn-tucker test in linear models with inequality constraints on the regression parameters', *Econometrica* **50**(1), 63–80.

Hahn, J., Hausman, J. & Kuersteiner, G. (2004), 'Estimation with weak instruments: Accuracy of higher order bias and mse approximations', *Econometrics Journal* **7**, 272–306.

Harding, M., Hausman, J. & Palmer, C. (2015), Finite sample bias corrected IV estimation for weak and many instruments. Unpublished Manuscript.

Heckman, J. J., Urzua, S. & Vytlacil, E. (2006), 'Understanding instrumental variables in models with essential heterogeneity', *Review of Economics and Statistics* **88**(3), 389–432.

Hillier, G. (2006), 'Yet more on the exact properties of IV estimators', *Econometric Theory* **22**, 913–931.

Hirano, K. & Porter, J. (2015), 'Location properties of point estimators in linear instrumental variables and related models', *Econometric Reviews* . Forthcoming.

Hirano, K. & Porter, J. R. (2012), 'Impossibility results for nondifferentiable functionals', *Econometrica* **80**(4), 1769–1790.

Hornung, E. (2014), 'Immigration snd the diffusion of technology: The Huguenot diaspora in Prussia', *American Economic Review* **104**, 84–122.

Imbens, G., Angrist, J. & Krueger, A. (1999), 'Jacknife instrumental variables estimation', *Journal of Applied Econometrics* **14**, 57–67.

Imbens, G. W. & Angrist, J. D. (1994), 'Identification and estimation of local average treatment effects', *Econometrica* **62**(2), 467–475.

Kleibergen, F. (2002), 'Pivotal statistics for testing structural parameters in instrumental variables regression', *Econometrica* **70**, 1781–1803.

Kleibergen, F. (2005), 'Testing parameters in gmm without assuming they are identified', *Econometrica* **73**, 1103–1123.

Kleibergen, F. (2007), 'Generalizing weak IV robust statistics towards multiple parameters, unrestricted covariance matrices and identification statistics', *Journal of Econometrics* **133**, 97–126.

Leadbetter, M. R., Lindgren, G. & Rootzen, H. (1983), *Extremes and Related Properties of Random Sequences and Processes*, 1 edn, Springer, New York.

Lehmann, E. & Casella, G. (1998), *Theory of Point Estimation*, Springer.

Lehmann, E. L. & Romano, J. P. (2005), *Testing statistical hypotheses*, Springer.

Mikusheva, A. (2010), 'Robust confidence sets in the presence of weak instruments', *Journal of Econometrics* **157**, 236–247.

Mills, B., Moreira, M. & Vilela, L. (2014), 'Estimator tests based on t-statistics for IV regression with weak instruments', *Journal of Econometrics* **182**, 351–363.

Moon, H. R., Schorfheide, F. & Granziera, E. (2013), 'Inference for VARs identified with sign restrictions', *Unpublished Manuscript* .

Moreira, H. & Moreira, M. (2013), Contributions to the theory of optimal tests. Unpublished Manuscript.

Moreira, M. (2003), 'A conditional likelihood ratio test for structural models', *Econometrica* **71**, 1027–1048.

Moreira, M. (2009), 'Tests with correct size when instruments can be arbitrarily weak', *Journal of Econometrics* **152**, 131–140.

Mueller, U. & Wang, Y. (2015), Nearly weighted risk minimal unbiased estimation. Unpublished Manuscript.

Newey, W. & West, K. (1987), 'A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix', *Econometrica* **55**(3), 703–708.

Phillips, P. (1983), *Handbook of Econometrics*, Amsterdam: North Holland, chapter Exact Small Sample Theory in the Simultaneous Equations Model, pp. 449–516.

Small, C. (2010), *Expansions and Asymptotics for Statistics*, Chapman and Hall.

Staiger, D. & Stock, J. H. (1997), 'Instrumental variables regression with weak instruments', *Econometrica* **65**(3), 557–586.

Stock, J. & Yogo, M. (2005), Testing for weak instruments in linear IV regression, *in* D. Andrews & J. Stock, eds, 'Identifcation and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg', Cambridge University Press, pp. 80–108.

Voinov, V. & Nikulin, M. (1993), *Unbiased Estimators and Their Applications*, Vol. 1, Springer Science+Business Media.