Instrumental Variables Quantile Regression with Multivariate Endogenous Variable

Guillaume A. Pouliot

October 15, 2019

Abstract

We propose methodology for optimization and inference with the inverse quantile regression (IQR) estimator for the instrumental variable quantile regression (IVQR) problem. We suggest a mixed integer linear programming (MILP) formulation which computes the global optimum of the non-smooth, non-convex IQR estimation problem, is solved rapidly using modern solvers, and accommodates multivariate endogenous variables. This formulation accommodates subvector inference for the causal estimate via inversion of a regression rankscore test, thus adapting the standard method of inference for linear quantile regression to the instrumental variables case. In contrast to competing large sample approaches, this inference method does not require nonparametric density estimation in the homoskedastic case, and thus circumvents the need to select a bandwidth parameter. To accommodate subvector inference under weak identification, we suggest two complementary methods. We provide a mixed integer quadratically constrained programming (MIQCP) formulation to compute confidence sets for subvectors as a projection from regression rankscore test confidence sets for the entire weakly identified vector. We provide a new test producing rectangular confidence regions such that less power is foregone by projecting the regions on principal axes to obtain simultaneously valid confidence intervals, and approximately no loss of power is induced when all but one weakly identified variable are nuisance parameters which may be profiled out. We carry out a causal quantile regression analysis of the impact of different types of institutions on the wealth of nations using the data set of Acemoglu and Johnson (2005), allowing for three endogenous variables.

Keywords: causal inference, duality, instrumental variable quantile regression, inverse quantile regression, mixed integer linear programming, optimization, preprocessing, projection method, rectangular confidence regions, regression rankscore, subvector inference, weak identification.

1 Introduction

A main objective of quantile regression analysis is to approximate the conditional quantile function. While least-squares regression offers an approximation to the conditional expectation function, quantile regression affords the analyst a characterization of the entire conditional distribution under study. Starting with the work of Koenker and Bassett (1978) and further advocated in the work of Chamberlain (1994), Buchinsky (1994), and many others, quantile regression has gradually become an essential tool in regression analysis. As economists often carry out causal inference in the presence of endogenous variables (Imbens and Rubin, 2015), an essential extension of quantile regression has been to accommodate identification with instrumental variables. Many such extensions have been put forth, we refer the reader to Koenker et al. (2017) for detailed descriptions and comparisons.

We concern ourselves with the instrumental variables quantile regression (IVQR) model proposed by Chernozhukov and Hansen (2005). Within the general model they propose, identification obtains if the structural errors (a.k.a. rank variables) are identically distributed under all treatment assignments, conditional on the exogenous covariates, instruments, and remaining unobservables impacting the endogenous variable. Leveraging this implication, they produce moment conditions identifying the causal treatment effect (Koenker et al., 2017, Chernozhukov and Hansen, 2006).

Two broad approaches have been laid out for estimation in the IVQR model (Chen and Lee, 2018, Koenker et al., 2017, Chernozhukov and Hansen, 2006). The original approach is to directly use the orthogonality conditions implied by the conditional expectation conditions defining the instruments and to construct a generalized method of moments (GMM) estimator. A more recent approach is to solve for the regression coefficient on the endogenous variables which, when concentrated out, minimizes the norm of the regression coefficients on the instruments. This is traditionally called the inverse quantile regression (IQR) estimator, or simply the instrumental variables quantile regression estimator. A superficial intuition for the IQR procedure is that if the instrumental variables only affect the outcome variable through the endogenous variables, conditioning on the instruments should not affect the regression function as long as it is correctly specified in terms of the endogenous variable. We will work out a deeper intuition in Section 2 below.

In spite of the popularity of the GMM and IQR approaches to causal inference in the IVQR model, important computational hurdles pertaining to point estimation with multivariate treatment effects[1] as well as inference have made its use more difficult and

[1] Those arise naturally. One frequent instance is unordered categorical/factor variables with more than two levels, such as occupation choice. Another instance arises when the researcher interacts endogenous treatment variables with exogenous covariates to capture conditional treatment effects. Zhu (2018) offers

thus limited its popularity.

Many strategies have been proposed for computation of the GMM and IQR estimators of the linear quantile regression function, both of which involve finding the optimum of a nonconvex and nonsmooth surface.

For the GMM approach, point estimation is computationally challenging because the parameter space will typically be large and a grid search very computationally expensive –or imprecise. One may instead explore the space using Markov chain Monte Carlo (MCMC) methods (Chernozhukov and Hong, 2003, Hoff, 2009, Andrieu et al., 2003), which comes with no guarantee that a global optimum has been found. Inference may then either be carried out using standard asymptotic inference for extremum estimators (Newey and McFadden, 1994), which in this case would require nonparametric density estimation, or using MCMC output and building "credible intervals" from a quasi-posterior. Kaido and Wüthrich (2018) proposed a coordinate-descent approach to computing an approximate solution to the GMM problem for which they provide a valid bootstrap.

Inverse quantile regression may be phrased as a bi-level problem, and a grid search is only necessary over the parameter space for the regression coefficients of the endogenous variable. When the endogenous variable is a scalar, this is computationally cheap –and may be phrased as a single parametric linear programming exercise– but quickly becomes intractable for endogenous variables of higher dimension.

Common inference methods for the inverse quantile regression estimator likewise require nonparametric density estimation even in the homoskedastic case, and thus the choice of a bandwidth parameter with respect to which the variance estimate may be sensitive. Such sensitivity issues are discussed in Section 7.

Provably Optimal Solvers

In developing methodology for a general applied audience, we ought to avoid asking of the user to try multiple starting points when computing an econometric estimator, or to simply hope that the default procedure tried enough starting points for a locally convergent algorithm to reach the global optimum. In particular, when possible, off-the-shelf econometric methods should deliver solutions which have global optimality guarantees.

Consequently, when the computation of an econometric estimator involves a nonconvex optimization problem, we may wish to find a reformulation of the computational problem such that it can be solved efficiently, and such that a global optimality certificate obtains upon convergence. One such successful line of research has produced mixed integer programming reformulations for non-convex, combinatorially difficult econometric

this latter example as a key motivation for developing new computational tactics for IVQR.

and statistical estimation problems (Zubizarreta et al., 2013, Bertsimas et al., 2017). Advances in computing and branch-and-bound heuristics over the past 25 years have made mixed integer programming 200 billion times faster (Bertsimas et al., 2014), hence making this approach quite attractive anew.

Different provably optimal solvers have been put forth for estimation in the instrumental variables quantile regression model.

The closest method to that described herein is the exact GMM estimator of Chen and Lee (2018). They formulate the GMM estimation problem as a mixed integer quadratic program (MIQP) which they solve exactly, relying on modern solvers.[2]

Zhu (2018) remarks that by instead evaluating a GMM estimator with ℓ_∞-norm on the moments, one may formulate the estimation problem as a mixed integer linear program (MILP) which modern solvers often solve with greater efficiency than MIQP's.

Instead of proposing a MIQP formulation to solve with a modern solver –and essentially treating it as a blackbox– Xu and Burer (2017) develop a novel branch-and-bound algorithm using a relaxed problem –relieved of the complementary slackness condition– and other strategies to produce lower and upper bounds.

Inference Methodology for Instrumental Variables Quantile Regression

The default inference method for quantile regression consists in inverting a regression rankscore test (R package quantreg, Koenker, 2005). A main attribute of the method is that, under a simple albeit strong assumption of homoskedasticity,[3] it delivers large sample inference that is pivotal with respect to the density of regression errors, thus allowing the user to circumvent nonparametric density estimation and, importantly, the choice of a bandwidth parameter.

Of the aforementioned provably optimal methods, none circumvent nonparametric density estimation and the requirement to select a bandwidth parameter when carrying out inference. The inference methodology Chen and Lee (2018) propose requires estimating the density of the regression errors, even when assuming homoskedasticity, and the scale of the asymptotic variance estimate may be sensitive to the choice of bandwidth. Zhu (2018) proposes a method of inference by simulation which involves solving only linear programs but requires nonparametric density estimation and the choice of bandwidth even in the homoskedastic case. Xu and Burer (2017) do not develop on methodology for

[2] The authors use Gurobi.

[3] In the context of inference for quantile regression, the homoskedastic case refers to the assumption that the density of the regression errors evaluated at 0 does not depend on location, i.e., $f_\epsilon(0|X, D, Z) = f_\epsilon(0)$. The logical implications of this assumption have been the subject of much quantile regression folklore. They are investigated in detail in Section 7. Robustness to departures from the homoskedasticity assumption are likewise investigated in simulation in Section 7.

inference.

Different bootstrap approaches have been proposed for inference in the IVQR model. They are detailed in Chernozhukov et al. (2017), see also the more recent work of Kaido and Wüthrich (2018). Typical bootstrap approaches are rather unpalatable in the IVQR set-up with multivariate endogenous variable because the estimation of each individual estimator –for each resampling of the dataset– is computationally expensive. In addition, the branching algorithms underlying the MILP solvers we suggest to use are naturally amenable to parallelization, thus benefitting themselves from distributed computing (Gendron and Crainic, 1994). Bootstrapping is furthermore notoriously unreliable under weak identification (Moreira et al, 2009), whilst we lay out a robust implementation of regression rankscore test inversion.

Chernozhukov and Hansen (2006) propose a score resampling method which involves resampling a linear representation of the statistic of interest. The method cuts down on computational cost by avoiding the need to recompute the estimator, but does require nonparametric estimation of the regression error density in order to compute the linear representation.

The main takeaway is that, except for computationally intensive and non-robust resampling methods such as the bootstrap, available inference methodology requires nonparametric density estimation, even in the homoskedastic case.

Even though regression rankscore inference is the default method in standard quantile regression, its analog for instrumental variables quantile regression is not available. In fact, the question of the form of such a test, and of whether it would inherit similar pivotal properties, has remained open.

We find that such an analogous pivotal procedure does obtain. It is developed and displayed in Section 5.

Weak Identification Robust Inference

Complete inference methodology for instrumental variables estimator ought to accomodate the case of weak instruments (Staiger and Stock, 1997, Mikusheva, 2013, Dufour, 2003). The prevalence of this case in practice has been well documented (I. Andrews et al., 2018).

Although nontrivial, much progress has been made in the univariate endogenous variable case, and reliable inference methodology is available, both for linear regression (I. Andrews et al., 2018) and quantile regression (Chernozhukov and Hansen, 2004, Chernozhukov et al., 2009).

In the case of a multivariate endogenous variable, the weak identification robust inference problem is inherently more difficult because when producing a confidence interval for any of the weakly identified coefficient, the remaining weakly identified coefficients enter the problem as nuisance parameters with nonstandard asymptotic distributions.

The typical way around this problem is to obtain a simultaneous confidence region for the entire weakly identified vector, and project the set on individual axes to obtain simultaneously valid confidence intervals (Dufour and Taamouti, 2005, Dufour, 1997, Dufour and Jasiak, 2001).

Applied candidly, projection methods are computationally demanding, but important contributions have been made to attenuate the computational burden, particularly in the case of linear models (Dufour and Taamouti, 2005, Mikusheva, 2010). We provide computational strategies for projection with the regression rankscore test in Subsection 5.3.1.

The projection method typically induces conservative confidence intervals, in part because it implicitly accounts for worst cases in terms of nuisance parameters. As discussed below, the conservative aspect of this projection step is a direct consequence of the geometry of the multivariate confidence regions. In Subsection 5.3.2, we propose a joint test which inverts to produce rectangular joint confidence regions, thus reducing or eliminating the loss of power in the projection step.

User-Friendly Methods

The objective of this article is to provide “user-friendly” machinery for estimation and inference with the IQR estimator. We take a clear stance on what a “user-friendly” implementation of quantile regression methods ought to satisfy. For estimation, the analyst should only have to stipulate the same input information as when carrying out standard quantile regression analysis. Specifically, the method should run satisfyingly when provided only with the data set and model. Any tuning parameters required for stipulation of the program should have default values that are both principled and computationally efficient. Likewise for inference, the user should not be prompted for tuning parameter values. Importantly, this means that we ought to carry out, whenever possible, inference that does not require nonparametric estimation of the regression error density and thus selection of a bandwidth parameter. Estimation and inference should be computed fast enough to allow for exploratory data analysis on moderately large datasets, which we take to mean n in the realm of a few thousands.

Our endeavor to circumvent density estimation should not be contentious. It is, as Koenker (2005) puts it, an “unhappy fact of life” that nonparametric density estimation is intrinsic to asymptotic covariance estimation in quantile regression. In particular, the requirement for nonparametric estimation in the –typically more elementary– homoskedastic case once limited the user-friendliness of standard quantile regression. The important practical impact of regression rankscore inference is largely attributable to the fact that it does not require density estimation in the homoskedastic case.

Indeed, the default confidence intervals in the quantreg package are obtained by inversion of the univariate regression rankscores test (Koenker, 2005). The reason for this is explicitly the avoidance of the computation of the regression error density (Koenker, 1994), which scales[4] the covariance and is typically sensitive to the choice of bandwidth. Furthermore, the use of the regression rankscore test statistic comes at very little cost in power (Gutenbrunner and Jurecková, 1992, Gutenbrunner et al., 1993, Koenker, 2005, Sidak et al., 1999).

Contributions

We propose a MILP formulation of the IQR estimator, and provide evidence that it outperforms the other available provably optimal methods for estimation in the IVQR model. We provide regression rankscore inference methodology, akin to that which is the default for standard quantile regression, and which delivers pivotal inference under ahomoskedasticity assumption for both full vector inference and subvector inference. For subvector inference under weak identification, the regression rankscore method produces robust full vector inference and may thus be used to produce subvector inference using the projection method. We rephrase the projection problem as a mixed integer quadratically constrained problem, which may be solved directly using solvers like Gurobi and Cplex, hence circumventing the need for grid search. We produce a new test which inverts to produce rectangular regions, thus reducing the loss of power due to the projection step, and eliminating it altogether for a scalar subvector, when the confidence region of the remaining endogenous variables is allowed to be arbitrarily large.

Notation

Unindexed upper case letters, such as X, Y and D, stand for random variables and typically refer to the population problem. Indexed upper case letter, such as X_i, Y_i, and D_i, likewise stand for random variables but typically refer to sampled observations. Bold upper case letters refer to data vectors, e.g., $\mathbf{Y} = (Y_1, ..., Y_n)^T$, and $\mathbf{D} = (D_1^T, ..., D_n^T)^T$.

We use the left arrow "←" to refer to the output of a problem. For instance, we may designate the primal and dual solutions $\hat{\beta}$ and $\hat{a}$ of the quantile regression of Y on X as $(\hat{\beta}, \hat{u}, \hat{v}, \hat{a}) \leftarrow QR(\mathbf{Y}, \mathbf{X})$. When no confusion arises, we may use the same notation to define only part of the solution, e.g. introducing the primal quantile regression

[4] For instance, in the iid case with errors independent of the covariantes, the asymptotic variance is proportional to $1/f_\epsilon^2(0)$.

solution as $\hat{\beta} \leftarrow QR(\mathbf{Y}, \mathbf{X})$. We use the same notation for the IQR estimator, written $IQR(\mathbf{Y}, \mathbf{D}, \mathbf{Z}, \mathbf{X})$ where $\mathbf{Y}$ contains the outcome variables, $\mathbf{D}$ the endogenous variables, $\mathbf{Z}$ the instruments, and $\mathbf{X}$ the exogenous controls. We denote the length of vector A with p_A.

Outline

The remainder of the paper is divided as follows. Section 2 provides background material on quantile regression as well as the instrumental variables quantile regression model and estimators. Section 3 and 4 describe the MILP formulation for the inverse quantile regression estimator and preprocessing strategies, respectively. Section 5 describes inference methodology and theory for the inverse quantile regression estimator. Section 6 and 7 investigate the performance of the method in simulations and in an application, respectively. Section 8 concludes.

2 Background Material

An optimization-conscious approach to quantile regression offers a constructive narrative for the derivation of the IQR estimator. We first present quantile regression in its linear programming formulation, and then derive the inverse quantile regression estimator from orthogonality conditions, which can be recognized as dual feasibility conditions of a specific quantile regression problem.

2.1 Linear Programming Formulation of Linear Quantile Regression Problem

We will rely heavily on the linear programming structure of the quantile regression program and its dual. Recall that the quantile regression problem with outcome vector $\mathbf{Y} \in \mathbb{R}^n$ and design matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ may be expressed as a linear program with primal (P) and dual (D) formulations given below.

$$\begin{aligned} &\min_{\beta_X, u, v} \tau u^T \mathbf{1}_n + (1-\tau) v^T \mathbf{1}_n \\ &\text{s.t.} \\ &\quad \mathbf{X}\beta_X + u - v = \mathbf{Y} \\ &\quad u, v \geq 0 \end{aligned} , \quad \text{(P)} \qquad \begin{aligned} &\max_{a} a^T \mathbf{Y} - (1-\tau) \mathbf{1}_n^T \mathbf{Y} \\ &\text{s.t.} \\ &\quad \mathbf{X}^T a = (1-\tau) \mathbf{X}^T \mathbf{1}_n \\ &\quad a \in [0,1]^n \end{aligned} . \quad \text{(D)}$$

We speak of $\{\beta_X, u, v \ : \ \mathbf{X}\beta_X + u - v = \mathbf{Y}, \ u, v \geq 0\}$ as the primal feasible set and of its defining equalities and inequalities as primal feasibility conditions, and speak of $\left\{a \ : \ \mathbf{X}^T a = (1-\tau)\mathbf{X}^T \mathbf{1}_n, \ a \in [0,1]^n\right\}$ as the dual feasible set, and its defining equalities and inequalities as dual feasibility conditions.

An important element of duality theory relates the residuals from the primal problem solution to the dual solution. Complementary slackness stipulates that, at the optimal solution, $a_i = 1$ whenever $u_i > 0$, and $a_i = 0$ whenever $v_i > 0$. Furthermore, $n - p$ dual variables will be exactly 0 or 1. Note that as long as the data is in general position (Koenker, 2005), which is a probability one event if the outcome variable or any covariate is absolutely continuous with respect to Lebesgue measure, no more than $n - p$ dual variables will be exactly 0 or 1, with probability one. We may then call these inactive variables, while the p entries whose a_i's lie between 0 and 1 form the active basis.

2.2 Instrumental Variables Quantile Regression Model

The instrumental variables quantile regression (IVQR) model of Chernozhukov and Hansen (2005) gives sufficient conditions for the identification of causal effects.

Assumption 1 (IVQR model) *Consider a common probability space (Ω, F, P) and the set of potential outcome variables $(Y_d, d \in \mathcal{D})$, endogenous variables D, exogenous covariates X, and instrumental variables Z. The following conditions hold jointly with probability 1*

A1 *(Potential outcomes). Conditional on X and for each d, $Y_d = Q(U_d, d, X)$, where $\tau \mapsto Q(\tau, d, X)$ is nondecreasing on $[0,1]$ and left-continuous and $U_d \sim U(0,1)$.*

A2 *(Independence). Conditional on X and for each d, U_d is independent of instrumental variables Z.*

A3 *(Selection). $D = \delta(Z, X, \nu)$ for some unknown function δ and random vector ν.*

A4 *(Rank similarity). Conditional on (Z, X, ν), $\{U_d\}$ are identically distributed.*

A5 *(Observables). The observed random vector consists of $Y = Y_d$, D, X, and Z.*

Theorem 1 (Chernozhukov and Hansen, 2005) *Suppose conditions A1-A5 hold.*

(i) *Then we have for $U = U_D$, with probability 1,*

$$Y = Q(U, D, X),\ U|X, Z \sim U(0, 1). \tag{1}$$

(ii) *If (1) holds and $\tau \mapsto Q(\tau, d, X)$ is strictly increasing for each d, then for each $\tau \in (0, 1)$, a.s.,*

$$P\left[Y \leq Q(\tau, D, X)|X, Z\right] = \tau. \tag{2}$$

(iii) *If (1) holds, then for any closed subset I of $(0, 1)$, a.s.,*

$$P(U \in I) \leq P\left[Y \in Q(I, D, X)|X, Z\right], \tag{3}$$

where $Q(I, d, x)$ is the image of I under the mapping $\tau \mapsto Q(\tau, d, x)$.

Chernozhukov and Hansen (2005, 2006, 2013) produce conditions for point identification in the IVQR model.

As detailed below, a key methodological implication of the result of Chernozhukov and Hansen (2005) is the justification of moment conditions identifying the causal effect coefficient.

We are interested in causal inference in the linear quantile regression model. Specifically, we keep, as do Chernozhukov and Hansen (2006), with the linear quantile regression function

$$Q(\tau, d, x) = d^T\beta_D(\tau) + x^T\beta_X(\tau), \tag{4}$$

for $\tau \in (0, 1)$.

A general method of moments (GMM) estimator may be constructed using moment conditions implied by Theorem 1. Let $\Phi := \Phi(D, X, Z)$ be a vector of functions of the transformed instruments. Chamberlain (1987) suggests a Φ and weighting matrix yielding pointwise efficiency. The popular $\Phi = (X^T, Z^T)^T$ remains a natural choice (Chernozhukov, Hansen and Wüthrich, 2017), and will be considered in the construction of our estimate, but largely dismissed when carrying out subvector inference.

Our suggested default transformation, in line with Chernozhukov and Hansen (2006), is to produce Φ by projecting D and the space spanned by Z and X.

This produces $p_\Phi + p_X$ moment conditions for estimation of the $p_D + p_X$ dimensional coefficient vector $(\beta_D^T, \beta_X^T)^T$. One can verify that these suffice for identification as long as $p_\Phi \geq p_D$ and standard regularity conditions are satisfied (Chernozhukov and Hansen, 2006).

The GMM estimator offers a very immediate way of imposing the implied moment conditions, which are

$$E\left[\left(\tau - \mathbf{1}\left\{Y - D^T\beta_D - X^T\beta_X \leq 0\right\}\right) X\right] = 0 \tag{5}$$

and

$$E\left[\left(\tau - \mathbf{1}\left\{Y - D^T\beta_D - X^T\beta_X \leq 0\right\}\right) \Phi\right] = 0, \tag{6}$$

but it is not the only approach to leverage these conditions so to construct an estimator. The IQR estimator is an alternative. The construction of the IQR estimator from the standard quantile regression estimator in its linear programming formulation captures its own connection with the identifying conditions (5) and (6).

2.3 Inverse Quantile Regression Estimator

The moment equations (5) and (6) have sample analogs

$$\left(\tau\mathbf{1}_n - (\mathbf{1}_n - a^*)\right)^T \mathbf{X} = 0, \tag{7}$$

and

$$\left(\tau\mathbf{1}_n - (\mathbf{1}_n - a^*)\right)^T \mathbf{\Phi} = 0, \tag{8}$$

where $a^* = (a_1^*, ..., a_n^*)^T$ is defined

$$a_i^* = 1 - 1\left\{\varepsilon_i \leq 0\right\},$$

$i = 1, ..., n$. Observe that, on the one hand, equations (7) and (8) are the dual feasibility conditions[5] of any quantile regression problem with independent variables X and Φ and that, moreover, the dual variables

$$a_i = 1 - \mathbf{1}\left\{\varepsilon_i \leq 0\right\},$$

for observations i corresponding to nonbasic dual variables of the quantile regression problem with dependent variable Y, independent variables X, D and Φ, and quantile regression function

$$Q_\tau(Y|X, D, \Phi) = X\beta_X(\tau) + D\beta_D(\tau).$$

[5] Note that they may be rewritten as

$$\mathbf{X}^T a^* = (1-\tau)\mathbf{X}^T\mathbf{1}_n$$

and

$$\mathbf{\Phi}^T a^* = (1-\tau)\mathbf{\Phi}^T\mathbf{1}_n,$$

respectively.

This naturally suggests that we define an estimate of β_D as the value for which, in the quantile regression of $\tilde{Y} := Y - D\beta_D$ on X and Φ, the coefficient β_Φ vanishes. Indeed, for such a solution, the sample orthogonality conditions (7) and (8) will approximate the moment orthogonality conditions (5) and (6) with a approximating[6] a^*.

We can thus define the IQR estimator $\hat{\beta}_D$ as the solution of

$$\min \|\beta_\Phi\| \tag{9}$$

subject to

$$\beta_\Phi \leftarrow QR(\mathbf{Y} - \mathbf{D}\beta_D, (\mathbf{X}, \mathbf{\Phi})), \tag{10}$$

for some choice of norm $\|\cdot\|$.[7] Chernozhukov and Hansen (2006) directly define the IQR estimator as (9)-(10).

As do Chernozhukov and Hansen (2006), we preconize the use of the instrument Φ_i formed by the least-squares projection of D_i on Z_i and X_i. The effect on point estimation appears to be small, it is however sizable and beneficial for subvector inference, as discussed in Section 5.

On the face of it, (9)-(10) is a bi-level problem. An outside loop optimizes over β_D, and for each value of β_D the quantile regression linear program is solved. This is indeed how the problem is typically solved in the univariate case, where the outer loop is, conveniently, a line search. However, for dimensions of the endogenous variable as small as three, a grid search in the outer loop may become both computationally expensive and imprecise. In particular, given that the surface we optimize over is nonsmooth and nonconvex, local optima are not guaranteed to be the global optimum and thus to be the optimal solution.

What is required for seamless implementation of IQR with multivariate endogenous variable is a one-level formulation of problem (9)-(10) which may be solved rapidly and provides a certificate of global optimality. Precisely to that end, we develop mixed integer linear programming formulations for the IQR program.

[6] Note that the approximation error vanishes for fixed p asymptotics.

[7] Chernozhukov and Hansen (2006) use a weighted ℓ_2-norm. We use the ℓ_1-norm. As discussed below, $\hat{\beta}_Z$ will asymptotically approach 0 (Chernozhukov and Hansen, 2006, subsection A4), making the choice of norm immaterial for point estimation. In the just identified case, under weak regularity conditions, it will be exactly 0 in sample, making the choice of norm immaterial for efficiency. Further remark that, as detailed below, the instrument both we and Chernozhukov and Hansen (2006) recommend always produces a just identified problem.

3 IVQR via Mixed Integer Linear Programming

The program (9)-(10) produces the IQR as a solution, but cannot be formulated as a convex program. We want to deliver a “one-click” software producing the IQR estimate along with a global optimality certificate. We achieve this by formulating the problem (9)-(10) as a mixed integer linear programming (MILP) problem, thus leveraging the excellent performance of modern MILP solvers.

Note that the difficulty of the problem at hand is inherent. Optimizing over quantile regression solutions is tantamount to imposing complementary slackness as a feasibility condition, which is known to produce a NP-hard program (Xu and Burer, 2017, Mangasarian, 1998). We have to deal with nonconvexity either in the form of continuous nonconvex constraints or of discrete variables. We preconize formulations with discrete variables rather than nonconvex quadratic constraints because global optimality guarantees obtain and they are more amenable to modern solvers.

Remarkably, it turns out to be possible to express the feasible set (10) as a set of linear equalities and inequalities in terms of continuous and binary variables, thus producing a MILP formulation for (9)-(10) if the ℓ_1-norm is used in (9). We derive this formulation.

Duality suggests two natural ways of expressing the set of quantile regression solutions. The first, and perhaps more obvious way, is to impose the primal feasibility conditions,

$$[\mathbf{X}, \mathbf{\Phi}, I, -I] \begin{pmatrix} \beta_X \\ \beta_\Phi \\ u \\ v \end{pmatrix} + \mathbf{D}\beta_D = \mathbf{Y}, \tag{11}$$

$$u, v \geq 0, \tag{12}$$

as well as the dual feasibility conditions,

$$\mathbf{X}^T a = (1-\tau)\mathbf{X}^T \mathbf{1}_n \tag{13}$$

$$\mathbf{\Phi}^T a = (1-\tau)\mathbf{\Phi}^T \mathbf{1}_n \tag{14}$$

$$a \in [0,1]^n, \tag{15}$$

and to require that the primal and dual objectives be equal. By weak duality, when we are for instance solving a minimization problem in the primal, the primal objective function evaluated at any primal feasible solution will be an upper bound to the dual objective function evaluated at any dual feasible solution. By strong duality, this bound is tight in that it holds with equality only at an optimal solution.

The second way is to impose the primal feasibility conditions (11)-(12), the dual feasibility conditions (13)-(15), and complementary slackness. These together suffice for optimality of the solution (Bertsimas and Tsitsiklis, 1997).

It turns out that the formulation obtained in the second approach is nested –and thus has fewer variables– in the formulation obtained in the first approach when the nonlinear condition of equality of the primal and dual objectives is linearized –which is necessary in order to produce a MILP formulation– using a McCormick envelope. We thus propose the formulation obtained by the second approach.

We will find that complementary slackness may be imposed by adding $4n$ constraints and $2n$ binary variables. Conveniently, this number can be brought down substantially with appropriate preprocessing, see Section 4.

The following theorem shows that discrete inequality conditions apparently emulating the complementary slackness conditions in fact do suffice to impose the said conditions.

Theorem 2 *Let $(\beta_D^*, \beta_X^*, \beta_\Phi^*, u^*, v^*, a^*, k^*, l^*)$ satisfy the primal feasibility conditions (11)-(12), the dual feasibility conditions (13)-(15), and*

$$v \leq l \cdot M, \; u \leq k \cdot M \tag{16}$$

$$a \geq k, \; a \leq \mathbf{1}_n - l \tag{17}$$

$$k \in \{0,1\}^n, \; l \in \{0,1\}^n. \tag{18}$$

Then, for M sufficiently large, $(\beta_X^*, \beta_\Phi^*, u^*, v^*)$ and a^* are, respectively, the solutions of the primal and dual quantile regression problems with data $(\tilde{\mathbf{Y}}, (\mathbf{X}, \mathbf{\Phi}))$, where $\tilde{\mathbf{Y}} = \mathbf{Y} - \mathbf{D}\beta_D^*$.

Proof

It suffices to show that the solutions $(\beta_X^*, \beta_\Phi^*, u^*, v^*)$ and a^* satisfy primal and dual feasibility, respectively, as well as complementary slackness. The primal solution is feasible because the primal feasibility constraints (11) and (12) are directly enforced. Likewise, the dual solution is feasible because the dual feasibility constraints (13)-(15) are directly enforced.

The complementary slackness conditions are

$$u_i (1 - a_i) = 0 \tag{19}$$

and

$$v_i a_i = 0, \tag{20}$$

$i = 1, ..., n$.

The complementary slackness conditions are likewise enforced by the set of conditions. For complementary slackness conditions (19), it suffices to show that equality holds even if $u_i > 0$ or $a_i \neq 1$. Note that

$$u_i > 0 \Rightarrow k_i = 1 \Rightarrow a_i = 1,$$

for all i. Consequently, if $u_i > 0$, it must be that $1 - a_i = 0$ and (19) holds. By contraposition, if $a_i \neq 1$, it cannot be that $u_i > 0$, and (19) must hold.

For complementary slackness conditions (20) to hold, it suffices to show that equality holds even if $v_i > 0$ or $a_i \neq 0$. Note that

$$v_i > 0 \Rightarrow l_i = 1 \Rightarrow a_i = 0,$$

for all i. Consequently, the complementary slackness condition must again hold by the same argument whether $v_i > 0$ or $a_i \neq 0$.

□

Theorem 2 tells us that the optimization problem of minimizing $\|\beta_\Phi\|_1$ subject to (11)-(18), i.e.,

$$\min \|\beta_\Phi\|_1$$

subject to

$$[\mathbf{X}, \mathbf{\Phi}, I, -I] \begin{pmatrix} \beta_X \\ \beta_\Phi \\ u \\ v \end{pmatrix} + \mathbf{D}\beta_D = \mathbf{Y}$$

$$\mathbf{X}^T a = (1-\tau)\mathbf{X}^T \mathbf{1}_n$$

$$\mathbf{\Phi}^T a = (1-\tau)\mathbf{\Phi}^T \mathbf{1}_n$$

$$u \leq k \cdot M, \ v \leq l \cdot M$$

$$a \geq k, \ a \leq \mathbf{1}_n - l$$

$$k, l \in \{0,1\}^n, \ u, v \geq 0, \ a \in [0,1]^n,$$

produces the exact IQR estimate, and may be submitted as such to a MILP solver.

See Figure 2 for the comparative performance of this approach with that of Chen and Lee (2017) and Zhu (2018). Software is available on the author's website.

3.1 Default Tuning Parameter Values

Although some tuning parameters are involved, researchers should be able to use the method without having to specify them. In fact, as they do not affect the statistical nature or value of the solution, but only numerical efficiency, researchers focusing on the application need not be concerned with them.

A tautological choice for the bound on the big-M constraint is

$$M_i = \max_{\beta \in \mathcal{B}} \left\{ Y_i - (X_i^T, D_i^T)\beta \right\},$$

where $\mathcal{B}$ is the compact set to which β is assumed to belong. A good rule of thumb is to take $M_i = 10\hat{\sigma}_{QR}$ for all i, where $\hat{\sigma}_{QR}^2$ is the sample variance of the errors of the standard quantile regression Y on X and D.

As detailed in Section 5, when inverting a null hypothesis, one will run the concentrated out regression of $Y - D\beta_D^0$ on X and Z, in which case M should be updated to $M_i + |D_i\beta_D^0|$ to avoid numerical errors.

4 Preprocessing

Preprocessing is typically an essential part of the formulation of any estimator one wishes to solve using modern MILP solvers. A good initial solution may be critical to the success of the method. For instance, Bertsimas et al. (2016) solve the OLS best subset selection problem as an MILP and, while they leverage the performance of solvers such as Gurobi, good preprocessing accounts for a speed-up of an order of magnitude (250 times faster in their largest instance), which is crucial to the amenability of the method.

For MILP's, the two perhaps most typical approaches to preprocessing are to provide an approximate solution believed to be close to the optimum, or to add to the problem formulation valid inequalities –a.k.a. cuts– which are non-redundant in the linear programming relaxation of the MILP. In the Supplementary Appendix, we describe these preprocessing strategies for the IQR MILP problem. Somewhat frustratingly, the problem, as solved using the heuristics of standard solvers such as Gurobi, is not very amenable to standard preprocessing strategies. In particular, as documented in the Supplementary Appendix 2.5, a feasible starting solution "near" the optimum is not helpful unless it is in a small neighborhood of the global optimum, even though solvers like Cplex and Gurobi attempt to exploit such information. In addition, as described in the Supplementary Appendix, adding valid inequalities does not appear to speed up the solution.

The preprocessing strategy we instead suggest builds on the preprocessing strategy of Koenker (2005) for large scale quantile regression, which finds outliers and fixes the sign of their residuals. This will be particularly lucrative for us, since fixing the sign of a residual eliminates two binary variable, the number of which largely drives the computational burden of the problem.

Remarkably, the structure of the problem at hand makes this rather aggressive preprocessing approach amenable. Recall that the instrumental variables regression problem is always just identified when we obtain the instruments by projection, consequently the optimal objective value is always (assuming rank conditions) $\hat{\beta}_Z = 0$. This avails us of a great opportunity for preprocessing because we may attempt very fast but highly constrained versions of the problem, and if the constraints turn out to be incompatible with the optimal solution (of the unconstrained problem) but the constrained problem is still feasible, the optimum objective will be some value $\hat{\beta}_Z \neq 0$, and we will detect that the problem was over-constrained. In that case we can relax the constraint in preprocessing.

Specifically, we run the quantile regression of Y on X and D using the full data, and pick out positive and negative outliers from its output. Again, we may gradually reduce the set of outliers if, at first, this produces an IQR regression which is either infeasible or has solution $\hat{\beta}_Z \neq 0$.

The preprocessing method is detailed in Algorithm 1, and requires three tuning parameters. One must specify α, which controls the quantity of residual signs that will be fixed in the first iterations, r, which controls the rate at which α is relaxed between iterations, and $T : \alpha \mapsto \mathbb{R}_+$, which is a menu of maximum running times passed which the problem is declared infeasible. Figure 1 gives an example of such a menu for small data sets. Of course, if α is such that no residual sign is prespecified, then $T(\alpha) = \infty$.

Algorithm 1 Preprocessing for IVQR MILP

```
1: Select $\alpha, r, T$
2: Run quantile regression of $Y \sim D + X$, extract residuals $\hat{\varepsilon}_1, ..., \hat{\varepsilon}_n$
3: Construct outlier sets
       $$\mathcal{O}_+ = \{i : \hat{\varepsilon}_i \geq u_i(\alpha)\}, \ \mathcal{O}_- = \{i : \hat{\varepsilon}_i \leq l_i(\alpha)\}$$
4: **IF** Running time reaches $T$
5: **OR** $\hat{\beta}_Z > 0$
6: **OR** the solver declares problem infeasible
7:   **THEN** Update $\alpha \leftarrow r \cdot \alpha$, $T \leftarrow T(\alpha)$ and return to line 3
8: **ELSE** Return solution $\hat{\beta}_D, \hat{\beta}_X, \hat{a}$ and **STOP**
```

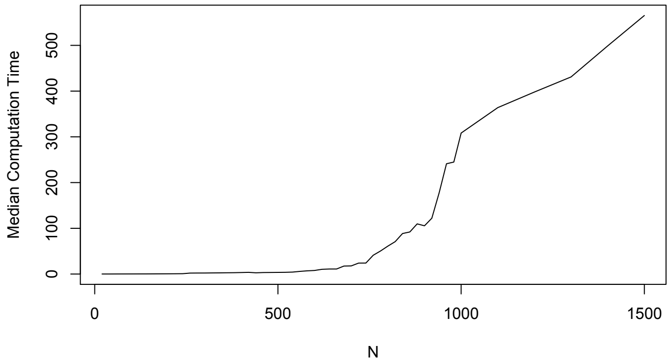


Figure 1 Preliminary. *Computation time versus the effective sample size.*

The preprocessing method is defined by the choice of u_i and l_i, $i = 1, ..., n$. Default values for the tuning parameter α, r, and T differ for different methods. The idea is to detect positive and negative outliers, and to fix the sign of their residuals in the programming problem.

One specific choice of u_i and l_i amounts to picking the band to be a "tube" around the fitted quantile regression; we the fix the sign of the residuals outside of a margin of size 2α around the fitted plane. In that case, $u_i = \alpha$, $l_i = -\alpha$, in an array of simulations, we found that picking α so that 30% of the errors are fixed in the first pass of the algorithm fix $r = 1.25$ was a good default. The speed-ups from preprocessing are documented with simulation evidence in the Appendix.

The reason for adding a time limit is that, in practice, the time to establish infeasibility of a problem for which some errors were attributed the wrong sign may be substantially longer than the typical time needed for solving a problem of that effective size. One then wants to stop the solver, relax the preprocessing, and start over for a larger value of α. As a default, we suggest picking for a value of $T(\alpha)$ twice the median time for solving a typical problem of the same effective size without preprocessing; such typical computation times for different (n, p) pairs are plotted in Figure 1, and further detailed in the Appendix. We remark that there is a large set of values of T which speed-up the preprocessing overall. Indeed, there is a large set of menus T such that total solving time is decreased. In fact, even by picking a fixed value of T to for all but full data, we found sizable speed-ups.

We tried using IQR on subset(s), or IQR in separate univariate instrumental variable quantile regressions, instead of standard quantile regression in order to obtain the approximate fit but we did not find that it accelerated the procedure.

Nevertheless, in special cases where the endogeneity is suspected to be severe, it may be preferable to use the fitted values from the inverse quantile regression on a subset of the data to detect residuals. One way to assess the reliability of the quantile regression as a guess for the inverse quantile regression solution is to extract the fitted $\hat{\beta}_D$ from the standard quantile regression of Y on D and X and look at the magnitude of $\hat{\beta}_Z$ in the quantile regression of $Y - D\hat{\beta}_D$ on X and Z.

It is important to keep in mind that, although the gains in computation speed are affected by the choice of tuning parameters, the estimated solution is not; there are no statistical consequences from picking good or bad tuning parameters.

5 Inference

Inference in standard linear quantile regression analysis can be done by inverting a test based on the regression rankscore statistic, which is defined below. This is, for instance, the default inference method in the R package quantreg. The main motivation for its use is that it circumvents, under the homoskedasticity assumption, the requirement for nonparametric density estimation, which makes the asymptotic variance estimate sensitive to the choice of bandwitdh parameters.

To the best or our knowledge, no large sample approximation based on the inversion of regression rankscore tests or otherwise circumventing the need to estimate the density of regression errors and select a bandwidth parameter have been put forth in the context of IVQR.

We suggest a new testing procedure generalizing the regression rankscore inference methodology to the IVQR framework. Regression rankscore inference in the standard case consists in producing the scores of a simple linear rank statistic from the short dual problem in which the tested coefficient is concentrated out and using the tested covariate as the coefficient of the rank statistic, thus producing a rank test statistic (Koenker, 2005). We find that concentrating out the tested endogenous covariate in the short IQR problem to produce the scores, and using the set of instruments as the –possibly multivariate– coefficients of the –possibly quadratic– rank statistic naturally produces a well-performing regression rankscore inference procedure for IQR.

From a computational perspective, the test can be implemented as a sequence of MILP's, made faster by systematically using the previous previous solution as a starting solution for the next program in the sequence, thus emulating the parametric programming exercise of Koenker and D'Orey (1987).[8]

[8] The quantreg package has a particularly fast implementation of the the simplex algorithm for quantile regression because it uses a modification of the tailor-made Barrodale and Roberts (1973) algorithm. For

For clarity of exposition, we first present results in the special case of full vector inference.

5.1 Full Vector Inference

We present the theory and the methodology to construct tests and confidence intervals for β_D, the regression coefficient of the potentially endogenous covariate D, which may be multivariate. We consider inverting tests of the form

$$H_0 \ : \ \beta_D = \beta_D^0$$

against the alternative

$$H_1 \ : \ \beta_D \neq \beta_D^0.$$

For univariate instruments, i.e. $p_Z = p_D = 1$, we suggest the test statistic

$$L_n = \frac{S_n}{\sqrt{\tau(1-\tau)\tilde{\mathbf{\Phi}}^T\tilde{\mathbf{\Phi}}}}, \tag{21}$$

where

$$S_n = n^{-1/2}\tilde{\mathbf{\Phi}}^T\hat{b}, \tag{22}$$

$\tilde{\mathbf{\Phi}} = Q\mathbf{\Phi}$, $Q = I - \mathbf{X}(\mathbf{X}^T\mathbf{\Psi}\mathbf{X})^{-1}\mathbf{\Psi}\mathbf{X}^T$, $\mathbf{\Psi} = \mathrm{diag}\left(f_\varepsilon\left(0\,|X_i, D_i, Z_i\,\right)\right)$ and $\hat{b} = \hat{a} - (1-\tau)\mathbf{1}_n$, with

$$\hat{a} = \arg\max\left\{\left(\mathbf{Y} - \mathbf{D}\beta_D^0\right)^T a \ : \ \mathbf{X}^T a = (1-\tau)\mathbf{X}^T\mathbf{1}_n, \ a \in [0,1]^n\right\}. \tag{23}$$

Note that, in the univariate instrument case, S_n is a scalar.

For multivariate instruments, we suggest the test statistic

$$Q_n = \frac{S_n^T M_n^{-1} S_n}{\tau(1-\tau)}, \tag{24}$$

where $M_n = \frac{1}{n}\tilde{\mathbf{\Phi}}^T\tilde{\mathbf{\Phi}}$.

Such tests are extensions of classical rank tests (Sidak et al., 1999) and have been developed and detailed in Gutenbrunner and Jurecková (1992) and further adapted to quantile regression by Gutenbrunner et al. (1993).

As detailed below, the test statistics L_n and Q_n converge, respectively, to standard Gaussian and Chi-Square random variables. Their "primary virtue" (Koenker, 1994) is that under a –artificial, albeit transparent– homoskedasticity assumption, the statistics

an analysis of this simplex algorithm, consult Pouliot (2017) or Bai, Pouliot and Shaikh (2019).

are pivotal with respect to the density of regression errors, thus circumventing the choice of a bandwidth parameters scaling the asymptotic variance.

Indeed, a remarkable fact about the asymptotic theory for regression rankscore statistics is that in the homoskedastic case the density of the regression errors, f_ϵ, does not enter the asymptotic covariance and thus need not be computed. Specifically, if $f_\varepsilon(0\,|X, D, Z) = f_\varepsilon(0)$ for all $X, D, Z \in \text{supp}(X, D, Z)$, which we refer to as the homoskedasticity assumption, then

$$Q = I - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T.$$

This property is established and explored in the study of the asymptotic distribution below, and its empirical performance is assessed in Sections 6 and 7.

The intuition for the statistical power of the statistic and testing procedure is best gleaned from the form of S_n. If the value of β_D^0 used in the short problem (23) is correct, then D should have no predictive power over the residual $\tilde{Y} = Y - D\beta_D^0$. In particular, because the instrument ought to affect $\tilde{Y} - X\hat{\beta}_X$ only through the endogenous variable D, we would expect the inner product between Z and $\tilde{Y} - X\hat{\beta}_X$ or a transformation thereof to be small. On the contrary, if β_D^0 is far from the true value of the regression coefficient of D, we would expect such a term to be large. The statistic S_n is precisely such an inner product, with $\hat{b}_i$ taking on value τ if $\tilde{Y}_i - X_i\hat{\beta}_X > 0$, $\tau - 1$ if $\tilde{Y}_i - X_i\hat{\beta}_X < 0$, and a value between $\tau - 1$ and τ if $\tilde{Y}_i - X_i\hat{\beta}_X = 0$.

Another intuition for the statistic is that it evaluates the omitted dual feasibility condition –equivalently, the IVQR orthogonality conditions– with respect to the instrument, $\mathbf{\Phi}^T\hat{a} - (1-\tau)\mathbf{\Phi}^T\mathbf{1}_n = \mathbf{\Phi}^T\hat{b}$, which should be small for the true β_D, since conditioning on Z in the primal should produce an estimate of β_Z of zero, and thus adding $\mathbf{\Phi}^T a - (1-\tau)\mathbf{\Phi}^T\mathbf{1}_n = 0$ to the dual should not affect the optimal solution if the posited value β_D^0 is correct.

The regression rankscore approach is close in spirit to the "dual inference" approach of Chernozhukov and Hansen (2004) which use as a test statistic a quadratic form in the estimate $\hat{\beta}_Z$ from the short regression.

Subsubsection 5.1.1 gives the asymptotic approximation to the distribution of (21) and (24).

5.1.1 Asymptotic Approximations

We obtain asymptotic distributions for the test statistics of interest, thus allowing for test inversion using asymptotic approximations to the critical values.

Because the short problem (23) is the standard dual quantile regression problem, the central limit theorem for the full vector case is an immediate adaptation of its analogue for standard quantile regression.

Theorem 3 (Lindeberg-Feller CLT, *Bai, Pouliot, and Shaikh, 2019*) *Suppose that*

(a) There exists $\varepsilon > 0$ such that $E\left[\|X\|^2\right] < \infty$.

(b) $E[f_\varepsilon(0|X_i, Z_i)X_iX_i^T]$ is positive definite.

(c) For all x, z, d, $f_U(\cdot|x,z,d)$ exists and is bounded. In addition, there exists $\delta > 0$ and $C > 0$ such that

$$\sup_{|u|<\delta} \sup_{(x,z,d)\in\mathbb{R}^{p_X+2}} \left| \frac{f_U(u|x,z,d) - f_U(0|x,z,d)}{u} \right| \leq C.$$

(d) $B = g(X,Z,D)$ for some map $g : \mathbb{R}^{p_X+2} \to \mathbb{R}^{p'}$, $p' \in \mathbb{N}$, is measurable and $0 < E[BB^T] < \infty$.

Then,

$$n^{-1/2}\mathbf{B}^T\hat{b} \xrightarrow{d} N(0,\sigma^2),$$

where

$$\sigma^2 = \tau(1-\tau)E\left[\tilde{B}_i\tilde{B}_i^T\right],$$

with $\tilde{B}_i = B_i - E[f_U(0|X_i,Z_i)B_iX_i']E[f_U(0|X_i,Z_i)X_iX_i']^{-1}X_i$.

Theorem 3 follows directly from an application of Theorem 3.1 of Bai, Pouliot and Shaikh (2019). It is also a special case of Proposition 1, below.

The pivotal nature of the regression rankscore statistic comes from the cancelation of the density term evaluated at zero when conditioning is immaterial.

Corollary 1 (Lindeberg-Feller CLT, Computable Formula) *Suppose that assumptions a-d of Theorem 3 hold, and further suppose that $f_U(0) = f_U(0|X,D,Z)$, $\forall$ X, D, Z. Then*

$$\tilde{B}_i = B_i - E[B_iX_i']E[X_iX_i']^{-1}X_i.$$

If $\mathbf{B} = \mathbf{\Phi}_\perp := \mathbf{\Phi} - \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{\Phi}$, $\forall$ i, then the plug-in estimate

$$\hat{\sigma}^2 = \frac{\tau(1-\tau)}{n}{\mathbf{\Phi}_\perp}^T\mathbf{\Phi}_\perp$$

is consistent, i.e.,

$$\hat{\sigma}^2 \xrightarrow{P} \sigma^2.$$

Proof

Notice that $f_U(0) = f_U(0|X, D, Z), \forall X, D, Z$ implies that

$$\tilde{B}_i = B_i - E[B_i X_i']E[X_i X_i']^{-1} X_i,$$

and $B_i = \tilde{\Phi}_i$ implies that

$$\tilde{B}_i = \tilde{\Phi}_i,$$

because

$$E[\tilde{\Phi}_i X_i']E[X_i X_i']^{-1} X_i = E[\left(\tilde{\Phi}_i - E[\tilde{\Phi}_i X_i']E[X_i X_i']^{-1} X_i\right) X_i']E[X_i X_i']^{-1} X_i$$

$$= E[\tilde{\Phi}_i X_i']E[X_i X_i']^{-1} X_i - E[E[\tilde{\Phi}_i X_i']E[X_i X_i']^{-1} X_i X_i']E[X_i X_i']^{-1} X_i = 0.$$

The result then follows because $\frac{1}{n}\mathbf{\Phi}_\perp^T \mathbf{\Phi}_\perp$ converges to $\frac{1}{n}\tilde{\mathbf{\Phi}}^T \tilde{\mathbf{\Phi}}$, which itself converges to $\frac{\sigma^2}{\tau(1-\tau)}$.

□

The equivalent result for a multivariate instrument is then immediate.

Corollary 2 (Lindeberg-Feller CLT Multivariate Instrument) *Suppose that assumptions a-d of Theorem 5 hold. Then*

$$\frac{S_n^T M_n^{-1} S_n}{\tau(1-\tau)} \xrightarrow{d} \chi^2_{p_B}.$$

If $f_\varepsilon(0) = f_\varepsilon(0|X, D, Z), \forall X, D, Z$, then $\mathbf{\Phi}_\perp$ may be computed as in Corollary 1.

Because the test statistic, and thus its implied p-value, are discontinuous functions of β_D^0, slightly more accurate confidence intervals are obtained by using interpolation, as described in Hall and Beran (1993). This is also part of the default procedure in standard quantile regression packages such as quantreg.

The results are readily extended to uniform central limit theorems, and a confidence band for inference on the process $\beta_D(\tau)$, $\tau \in [\epsilon, 1-\epsilon] \subset (0, 1)$ may be obtained in the usual fashion. The statistic $\mathcal{B} = \sup_{\tau \in [\epsilon, 1-\epsilon]} Q_n(\tau)$ is a Bessel process of order p_D. A distribution-free asymptotic distribution obtains and no Durbin problem arises as we pose a hypothesis directly on the value of the regression coefficients. Different null hypotheses may produce tests with nuisance parameters, however these may be handled by Khmaladzation (Koenker and Xiao, 2003) or subsampling methods (Chernozhukov and Fernandez-Val, 2005).

5.2 Subvector Inference

Our interest lies in developing methodology for point estimation and inference with a multivariate endogenous variable. In that case, full vector inference is scarcely of interest, as researchers typically want to produce confidence intervals for individual coefficients. We therefore extend and adapt the regression rankscore statistic and test inversion apparatus to accommodate subvector inference. In this subsection, we consider the strongly identified case, the weakly identified case is considered in the Subsection 5.3.

The test we want to carry out and invert is

$$H_0 \; : \; \beta_{D,J} = \beta_{D,J}^0$$

against the alternative

$$H_{1,J} \; : \; \beta_{D,J} \neq \beta_{D,J}^0,$$

where $J \subset \{1, ..., p_D\}$.

The structure of the problem allows for a convenient formulation of the testing problem and procedure.

We propose to build

$$L_n = \frac{S_n}{\sqrt{\tau(1-\tau)\tilde{\mathbf{\Phi}}_{\cdot,J}^T \tilde{\mathbf{\Phi}}_{\cdot,J}}}$$

in the univariate case, and

$$Q_n = \frac{S_n^T M_n^{-1} S_n}{\tau(1-\tau)}$$

where $M_n = \frac{1}{n}\tilde{\mathbf{\Phi}}_{\cdot,J}^T \tilde{\mathbf{\Phi}}_{\cdot,J}$, in the multivariate case $|J| > 1$, from

$$S_n = n^{-1/2}\mathbf{\Phi}_{\cdot,J}^T \hat{b},$$

with

$$\hat{a} \leftarrow IQR(\tilde{\mathbf{Y}}_{\cdot,J}, \mathbf{D}_{\cdot,-J}, \mathbf{X}, \mathbf{\Phi}_{\cdot,-J}),$$

where $\tilde{\mathbf{Y}}_{\cdot,J} = \mathbf{Y} - \mathbf{D}_{\cdot,J}\beta_{D,J}^0$,$\tilde{\mathbf{\Phi}} = Q\mathbf{\Phi}$, $Q = I - \mathbf{X}(\mathbf{X}^T\mathbf{\Psi}\mathbf{X})^{-1}\mathbf{\Psi}\mathbf{X}^T$, $\mathbf{\Psi} = \text{diag}\left(f_\varepsilon\left(0\,|X_i, D_i, Z_i\right)\right)$ and $\hat{b} = \hat{a} - (1-\tau)\mathbf{1}_n$.

We obtain critical values from asymptotic theory. The proof technique of Bai, Pouliot, and Shaikh (2019) generalizes to this case, and a central limit theorem obtains. One ingredient of the proof is an asymptotic representation for $\sqrt{n}(\hat{\beta} - \beta)$, which obtains under Assumption 2, given below.

Assumption 2 (Sufficient Conditions for CLT, Chernozhukov and Hansen, 2006) *Let $\mathcal{B}_D$, $\mathcal{B}_Z$ and $\mathcal{B}_\Phi$ be the supports, respectively, of D, Z, and Φ. Let $\mathcal{T} = (0, 1)$.*

A1 *Compactness and convexity: for all τ, $(\beta_D(\tau), \beta_X(\tau)) \in \operatorname{int} \mathcal{B}_D \times \mathcal{B}_Z$, $\mathcal{B}_D \times \mathcal{B}_Z$ is compact and convex.*

A2 *Full rank and continuity: Y has a bounded conditional density, i.e., $\sup_{y \in \mathbb{R}} f_{Y|(X,D,Z)}(y) < A$ a.s. for some A, and for $\pi := (\beta_D, \beta_X, \beta_\Phi)$, $\theta := (\beta_D^T, \beta_X^T)$, and*

$$\Pi(\pi, \tau) := E\left[\left(\tau - \mathbf{1}\left\{Y < D^T\beta_D + X^T\beta_X + \Phi(\tau)^T\beta_\Phi\right\}\right)\Psi(\tau)\right],$$

$$\Pi(\theta, \tau) := E\left[\left(\tau - \mathbf{1}\left(Y < D^T\beta_D + X\beta_X\right)\right)\Psi(\tau)\right], \;\; \Psi_i(\tau) = \left[\Phi_i(\tau)^T, X_i^T\right]^T,$$

Jacobian matrices $\frac{\partial}{\partial(\beta_D^T, \beta_X^T)}\Pi(\theta, \tau)$ and $\frac{\partial}{\partial(\beta_X^T, \beta_Z^T,)}\Pi(\pi, \tau)$ are continuous and have full rank, uniformly over $\mathcal{B}_D \times \mathcal{B}_Z \times \mathcal{B}_\Phi \times \mathcal{T}$, and the image of $\mathcal{B}_D \times \mathcal{B}_Z$ under the mapping $(\beta_D, \beta_X) \mapsto \Pi(\theta, \tau)$ is simply connected.

We can now give the main distributional result.

Proposition 1 *Consider identically and independently distributed draws $(Y_i, D_i, X_i, Z_i) \sim \mathcal{F}$, $i = 1, ..., n$, for some distribution $\mathcal{F}$, taking value on a compact set and suppose that the quantile regression function has linear form (4). Suppose Assumptions 1 and 2 hold. Further suppose that*

a) (Y_i, D_i, X_i) is almost surely in general position,

b) $\Phi(X, Z, D)$ is measurable, and $0 < E\left[\Phi_{i,J}\Phi_{i,J}^T\right] < \infty$,

c) For all x, d and z, $f_\varepsilon(\cdot|x, d, z)$ exists and is bounded. In addition, there exists $\delta > 0$ and $C > 0$ such that

$$\sup_{|u|<\delta} \sup_{(x,d,z)\in\mathbb{R}^p} \left| \frac{f_\varepsilon(u|x, d, z) - f_\varepsilon(0|x, d, z)}{u} \right| \leq C,$$

d) $E[(D_{i,J}^T, X_i^T, Z_i^T)^T(D_{i,J}^T, X_i^T, Z_i^T)]$ is bounded.

Then

$$n^{-1/2}\sum_{i=1}^{n} \Phi_{i,J}\hat{b}_i \xrightarrow{d} N\left(0, \tau(1-\tau)E\left[\tilde{\Phi}_{i,J}\tilde{\Phi}_{i,J}^T\right]\right)$$

where

$$\tilde{\Phi}_{i,J} = \Phi_{i,J} - E\left[f_{\varepsilon_i}(0)\Phi_{i,J}(D_{i,-J}^T, X_i^T)\right] E\left[f_{\varepsilon_i}(0)\left(\Phi_{-J}^T, X_i^T\right)^T\left(D_{-J}^T, X_i^T\right)\right]^{-1}\left[\Phi_{i,-J}^T, X_i^T\right]^T,$$

with $f_{\varepsilon_i}(0) = f_\varepsilon(0|X_i, D_{i,-J}, \Phi_i)$.

Remark 4 The result may be extended, as in Chernozhukov and Hanson (2006), to allow for consistently estimated instruments and weights on the observations. The result may likewise be extended to inference over the quantile regression process.

Remark 5 The form of $\tilde{\Phi}_{i,j}$ calls for a comparison with its analog in the standard quantile regression case, where the tested covariate is projected on the span of the remaining covariates, with inner products weighted by the density $f_U(0|D,X,Z)$. We can recuperate an analogous interpretation here. Since we can express the instruments and exogenous controls as

$$E\left[f_U(0|D_i,X_i)\Phi_{i,J}(D_{i,-J}^T,X_i^T)\right]E\left[f_U(0|D_i,X_i,Z_i)\begin{pmatrix}\Phi_{i,-J}\\ X_i\end{pmatrix}\left(D_{i,-J}^T,X_i^T\right)\right]^{-1}\begin{pmatrix}\Phi_{i,-J}\\ X_i\end{pmatrix},$$

under the assumption $f_U(0)=f_U(0|D_i,X_i,Z_i)$, the sample estimate is

$$(\mathbf{\Phi}_{\cdot,-J},\mathbf{X})\left((\mathbf{D}_{\cdot,-J},\mathbf{X})^T(\mathbf{\Phi}_{\cdot,-J},\mathbf{X})\right)^{-1}(\mathbf{D}_{\cdot,-J},\mathbf{X})^T\mathbf{\Phi}_{\cdot,J}.$$

Note that if $\mathbf{\Phi}$ is obtained by projecting $\mathbf{D}$ on the exogenous variables, then there exists a positive definite matrix Ω such that $(\mathbf{D}_{\cdot,-J},\mathbf{X})=\Omega(\mathbf{\Phi}_{\cdot,-J},\mathbf{X})$ and thus

$$(\mathbf{\Phi}_{\cdot,-J},\mathbf{X})\left((\mathbf{\Phi}_{\cdot,-J},\mathbf{X})^T\Omega(\mathbf{\Phi}_{\cdot,-J},\mathbf{X})\right)^{-1}(\mathbf{\Phi}_{\cdot,-J},\mathbf{X})^T\Omega\mathbf{\Phi}_{\cdot,j},$$

which is an oblique projection.

This does mean that we have to solve an MILP for every null we want to test. We should thus be economical in our estimations when inverting the test. One can design a line search using the standard quantile confidence regions as an order of magnitude and reduce the computational burden by reoptimization.[9]

5.3 Subvector Inference Under Weak Identification

A common occurrence in causal inference using instrumental variables is that of weak instruments, i.e., they capture only a small part of the variation of the endogeneous variable. This is known to make subvector inference particularly difficult. Typically, when accounting for weak identification in the asymptotics, one obtains a nonstandard distribution for the regression coefficient which will furthermore depend on nuisance parameters quantifying the strength of identification. When the nuisance parameters, although not consistently estimable, are identifiable, a Bonferoni procedure may be developed –but this of course requires modeling the strength of identification (Staiger and Stock, 1997).

Perhaps more commonly, confidence intervals are obtained by first inverting a null on the full weakly identified vector and then projecting the resulting multivariate confidence

[9] Alternatively, one could do the search as a single parametric mixed integer program (Jenkins, 1982).

region on the axis corresponding to the coefficient –or linear combination of coefficients– of interest, thus producing robust confidence intervals.

In order to make multivariate IQR inference methodology amenable to regression analysis under weak identification, we provide two strategies. First, we provide a mixed integer quadratically constrained program (MIQCP) formulation of the standard projection problem. Second, we provide a multivariate test with rectangular confidence regions which may then be used to produce confidence intervals by the projection method, thus reducing the loss of power due to the projection step, and eliminating it altogether when the confidence interval of only one coefficient is of interest and the confidence region for the remaining weakly identified coefficients is allowed to be arbitrarily large.

5.3.1 Projection from the Regression Rankscore Confidence Region

Applying the projection method to regression rankscore confidence regions may seem problematic since these are notoriously difficult to compute (Bai et al., 2019). The candid grid search approach to test inversion and subvector inference can make for a substantial computational burden. In the linear regression case, Dufour and Taamouti (2005) and Mikusheva (2010) have put forth results alleviating the said burden. In the case of IQR, one way in which the specific structure of the problem at hand facilitates subvector inference under weak identification is that the projection problem may itself be formulated as a mixed integer quadratically constrained (MIQCP) problem. Specifically, for a given $j \in \{1, ..., p\}$, we want to minimize or maximize $\beta_{D,j}$ subject to the p-value associated with $T_n(\beta_{D,1}, ...\beta_{D,p})$ being less than a pre-specified critical level. This may be written, for the case of, say, the maximum, as

$$\max_{\beta_D, \beta_X, a, k, l} \beta_{D,j}$$

$$[\mathbf{X}, I, -I] \begin{pmatrix} \beta_X \\ u \\ v \end{pmatrix} + \mathbf{D}\beta_D = \mathbf{Y} \tag{25}$$

$$\mathbf{X}^T a = (1-\tau)\mathbf{X}^T \mathbf{1}_n \tag{26}$$

$$n\left(a - (1-\tau)\right)^T \mathbf{\Phi} Q \mathbf{\Phi}^T \left(a - (1-\tau)\right) \leq \tau(1-\tau) \cdot c_{1-\alpha} \tag{27}$$

$$v \leq l \cdot M, \;\; u \leq k \cdot M \tag{28}$$

$$a \geq k, \;\; a \leq \mathbf{1}_n - l \tag{29}$$

$$k \in \{0,1\}^n, \;\; l \in \{0,1\}^n, \;\; a \in [0,1]^n \tag{30}$$

$$u, v \geq 0, \;\; \beta_D, \beta_X \text{ free.} \tag{31}$$

By Theorem 2, conditions (30), (31) and (33)-(36) ensure that the feasible set is that of quantile regression primal and dual solutions.

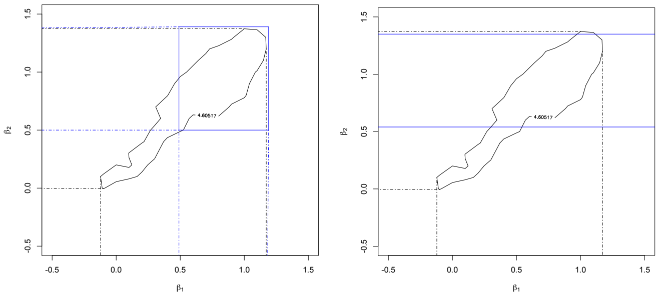
5.3.2 Regression Rankscore Rectangular Regions

Although projection methods control test size and are robust to weak identification, they tend to be problematically conservative. Even statistically powerful joint tests, producing joint confidence regions of small volume, may produce quite conservative confidence intervals due to the projection step. The geometry of the problem is immediate; oblique confidence region (as in Figure 2) will have a projection on any Euclidian basis vector that is much wider than any intersection of a translation of the said vector with the confidence region. See Pouliot (2019) for a thorough discussion and for the linear instrumental variables case.

A natural response is to try and use test statistics which produce rectangular regions, so to incur a lesser loss in power when projecting. Typically, the main challenge in doing so is that one needs to design a statistically powerful test statistic $t : \mathbb{R}^p \to \mathbb{R}$ whose level sets have a rectangular preimage. But of course, it is inherently difficult to accommodate such design constraints for a test statistic mapping a p-dimensional space, $p > 1$, to a one-dimensional space.

The suggestion of Pouliot (2019) is to circumvent this issue by considering test statistics with multivariate domain, $T : \mathbb{R}^p \to \mathbb{R}^p$. Intuitively, if we directly use a p-dimensional statistic whose inner product makes for a powerful test statistic, we should be producing a powerful test.

By getting the test size as the integral under the null of the image of the confidence region with respect to the test statistic, the construction attributes exact size to every subset of the support of the regression coefficient vector. In particular, every rectangular region is a valid joint confidence region.



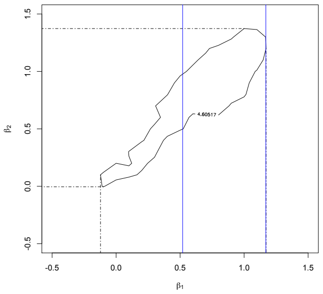


Figure 2 *Three 90% rectangular regression rankscore confidence regions (blue) are overlaid atop the 90% regression rankscore region (black). Dashed lines indicate the projections.*

In general, the projection method is still conservative even though the joint confidence region is rectangular.[10] However, in the case in which one endogenous variable is of interest and the other weakly identified control variables are nuisance parameters, one may pick a rectangular region that is arbitrarily wide in the dimension of nuisance variables, thus making the subvector test of exact nominal size.

Figure 2 gives examples in the $p = 2$ case, where both visualization and numerical integration (see below) are easy. We can see from the top-left plot that using the rectangular region, the projection on both axes produces smaller jointly valid confidence intervals than projection from the powerful but non-rectangular standard regression rankscore region. By profiling out one of the two regression coefficients, which is to say by allowing it to have arbitrarily large projected confidence intervals, we can make one confidence interval even shorter. In this given simulation instance, the gains are noticeable for β_1 but not for β_2.

Explicitly, for a weakly identified subvector $\beta_{D,J}$, we suggest testing the null $H_0 : \beta_{D,J} = \beta^0_{D,J}$ against the alternative $H_1 : \beta_{D,J} \neq \beta^0_{D,J}$ using the test statistic

$$\hat{T}_n(\beta^0_{D,j}) := \frac{1}{\sqrt{\tau(1-\tau)}} M_n^{-1/2} S_n,$$

where $S_n = n^{-1/2}\tilde{\mathbf{\Phi}}_J^T \hat{b}$, $\hat{a} \leftarrow IQR(\tilde{\mathbf{Y}}_{\cdot,J}, \mathbf{D}_{\cdot,-J}, \mathbf{X}, \mathbf{\Phi}_{-J})$, and $\tilde{\mathbf{Y}}_{\cdot,J} = \mathbf{Y} - \mathbf{D}_{\cdot,J}\beta^0_{D,J}$. Observe that $\hat{T}_n(\beta^0_{D,J}) \in \mathbb{R}^{p_{D_J}}$.

Further observe that $\hat{T}_n(\beta^0_{D,J}) \xrightarrow{d} N_p(0, I)$ under the null hypothesis $H_0 : \beta_{D,J} = \beta^0_{D,j}$, for any $\beta^0_{D,J} \in \mathbb{R}^{p_{D,J}}$. Testing the null $H_0 : \beta_{D,J} = \beta^0_{D,J}$ corresponds to asking if

[10] This simply refers to the fact that, for say the rectangular region $I_1 \times I_2 \subset \mathbb{R}^2$, where I_1 and I_2 are finite intervals, $P((\beta_1, \beta_2) \in I_1 \times I_2) \leq P(\beta_1 \in I_1)$. Of course, if $I_2 = \mathbb{R}$, then $P((\beta_1, \beta_2) \in I_1 \times I_2) = P(\beta_1 \in I_1)$.

$\hat{T}_n(\beta^0_{D,J}) \in \mathcal{C}_{\hat{T}_n}$ for some predefined "acceptance region" $\mathcal{C}_{\hat{T}_n}$ satisfying

$$P_{H_0}\left(\hat{T}_n(\beta^0_{D,j}) \in \mathcal{C}_{\hat{T}_n}\right) \equiv \int_{\mathcal{C}_{\hat{T}_n}} \phi(t)dt = 1-\alpha, \tag{32}$$

where ϕ is the standard Gaussian density.

A valid confidence region obtains by selecting a region $\mathcal{R}_\beta \subset \mathbb{R}^p$ such that its image $\mathcal{C}_{\hat{T}_n} = \hat{T}_n(\mathcal{R}_\beta)$ has Gaussian volume of size $1-\alpha$, as specified in (32).

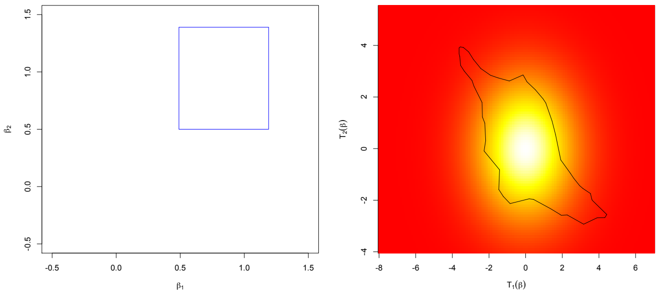


Figure 3 *The rectangular confidence region $\mathcal{R}_\beta$ is plotted on the left-hand side, and its image per the test statistic $\hat{T}_n(\mathcal{R}_\beta)$, overlaying a heat map of the density of the null distribution ϕ, plotted on the right-hand side. The coverage probability (32) is the integral of ϕ over the region $\hat{T}_n(\mathcal{R}_\beta)$.*

Searching for an optimal $\mathcal{R}$ requires an optimality criterion. Finding the narrowest interval for a coefficient of interest, when profiling out nuisance control parameters, is straightforward. At any given value for the (wlog) left-hand side bound of the confidence interval for the coefficient of interest, say β_1, all other intervals are taken to be arbitrarily large, and the requirement (32) uniquely determines the right-hand side bound of the interval. The search for the narrowest interval may then be formulated as a line search over the left-hand side bound. For routines optimizing over other optimality criteria such as minimum volume, or the minimum of the maximum interval, as well as for numerical integration of (32), see Pouliot (2019).

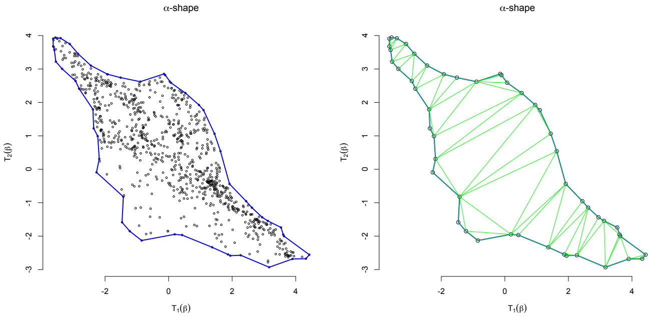


Figure 4 *The left-hand plot displays the set $\mathcal{S} = \left\{\hat{T}_n(b_1), ..., \hat{T}_n(b_N)\right\}$ in black and its α-shape in blue. The right-hand side plot displays the Delaunay triangulation of the α-extreme points of $\mathcal{S}$.*

The description up to this point was generic and may have been given for a general test statistic T. What makes the implementation different for different choice of test statistic is the computation of the numerical integral (32). In case at hand, the rectangular regression rankscore statistic makes for a particularly challenging numerical integration problem. Indeed, consider the following: we cannot get a closed form for, or directly compute $\hat{T}_n(\mathcal{R}_\beta)$ or its Gaussian volume; establishing membership of a point to $\hat{T}_n(\mathcal{R}_\beta)$ is computationally expensive; computing the volume is inherently challenging because $\hat{T}_n(\mathcal{R}_\beta)$ may not be convex. We cannot directly sample uniformly from $\hat{T}_n(\mathcal{R}_\beta)$ and importance sampling techniques with change-of-variables are intractable because T is not smooth. Our main “point of access” to $\hat{T}_n(\mathcal{R}_\beta)$ is that we can sample $b_1, ..., b_N$, say uniformly, from $\mathcal{R}_\beta$, and produce draws $\hat{T}_n(b_1), ..., \hat{T}_n(b_N)$, all in $\hat{T}_n(\mathcal{R}_\beta)$, and distributed according to some unknown g.

We suggest to approach the exercise as a support detection problem. The computational strategy is to use the draws $\hat{T}_n(b_1), ..., \hat{T}_n(b_N)$ in order to find a stable estimate of the contour of the set $\hat{T}_n(\mathcal{R}_\beta)$, along with a subdivision into easily integrable regions such as convex polytopes, the integral evaluations of which may then be summed to produce an estimate of (32). Separating the integration problem into a sum of integrals over convex polytopes is convenient for numerical approximation by quadrature methods –in lower dimensions– and Monte Carlo integration methods –for higher dimensional integration– where the partitioning furthermore helps reduce the estimation variance.[11]

Furthermore, we remark that the support detection problem and triangulation problem are intertwined. Consider the specifically the $p = 2$ case illustrated in Figure 3. The

[11] When the integrand may be represented as a polynomial, partitioning the region of integration into simplicies allows for exact polynomial integration methods (Lasserre and Avrachenkov, 2001).

case for $p > 2$ is analogous, with triangles replaced by polytopes. We map a large number of points b_i, $i = 1, ..., n$, into $\hat{T}_n(\mathcal{R}_\beta)$, and then estimate the α-shape of $\hat{T}_n(\mathcal{R}_\beta)$ using $\mathcal{S} = \left\{\hat{T}_n(b_1), ..., \hat{T}_n(b_N)\right\}$.[12] The α-shape corresponds to a Voronoi diagram and thus a Delaunay triangulation for $\mathcal{S}$, but it is very dense with small Delaunay triangles, and thus poorly suited for integration in the triangles. We instead compute the Delaunay triangulation of the α-extreme points $\mathcal{S}_{\text{extreme}} \subset \mathcal{S}$, which is much sparser.[13] The estimates tend to be very stable.[14] The main tuning parameter is the value of α. However, using a large number N of draws makes the procedure robust to choice of α (Pouliot, 2019).

5.4 Overidentification Restriction Test

A simple overidentification restriction test obtains. As discussed above, one of the interpretations of the regression rankscore test is that it evaluates the feasibility condition omitted in the short dual problem. This naturally invites an overidentification restriction test. Suppose the instrument may be split in two subvectors, $Z = (Z_1, Z_2)$, with Z_1 suspect and $p_{Z_2} \geq p_D$, then

$$n^{-1/2}\frac{\mathbf{Z}_1^T\hat{b}}{\sqrt{\tau(1-\tau)\mathbf{Z}_1^T Q \mathbf{Z}_1}}$$

where

$$\hat{a} \leftarrow IQR(\mathbf{Y}, \mathbf{D}, \mathbf{\Phi}, \mathbf{X}), \text{ and } \Phi_i = b_i(\mathbf{D}, \mathbf{X}, \mathbf{Z}_2) \; \forall \; i,$$

will be distributed as a standard Gaussian random variable (the multivariate case with $\chi^2_{p_{Z_1}}$ asymptotic distribution is analogous) under the null of correct specification.

[12] The literature developing the α-shape to characterize the shape of non-convex bodies begins with Edelsbrunner et al. (1983). Edelsbrunner (2010) and Graham and Yao (1990) give good surveys. Rodríguez Casal and Pateiro-López (2010) produce efficient software for computing the α-shape and α-hull, expanding on software from Renka (1996) and Renka et al. (2009), and computing the Delaunay triangulation (Delaunay, 1934) and Voronoi diagram (Voronoi, 1908) of a set of points. Rodríguez Casal and Pateiro-López (2010) also give a good review of the literature.

[13] Note that one must be careful to keep only Delaunay triangles inside the α-shape. Furthermore, the triangulation may be made even sparser by combining Delaunay triangles whose union is itself a convex polytope. See Pouliot (2019) for more details and algorithms.

[14] A triangulation made of many very long and narrow triangles would make for an unstable numerical integral. This can happen when computing the size of, say, very tall and narrow confidence regions, as these tend to map to elongated acceptance regions. However, it is typically unnecessary to use vary tall rectangles to profile out a variable, because statistical error –from having a tall as opposed to infinitely tall rectangle– vanished for moderately tall whose triangulation will produce easily integrable triangles.

6 Application

We revisit the application of Acemoglu and Johnson (2005). The authors investigate the average impact of different types of institutions on the wealth and growth of nations. The dataset aims at capturing two types of institutions. The first are "contracting institutions", defined as the rules and regulations governing contracting between ordinary citizens. The second are "property rights institutions", defined as the rules and regulations protecting citizens against the power of government and elites. Both types of institutions are proxied with observables, which are in turn instrumented for.

Property rights institutions are captured by three proxies. The first is a measure of "constraint on executive", which reports the degree of constraints on politicians and politically powerful elites. The second is a measure of "protection against expropriation" by the government. The third is a private property index (La Porta et al., 1999, Beck et al., 2003).

The proxies for property rights institutions are interpretable and interesting on their own. However, as they are meant to capture similar variation, careful multivariate analysis is called for if we are interested in which variables matter most for a given outcome.

Proxy for	Property Rights Institutions	Contracting Institutions
Endogenous Variables	protection against expropriation private property index	legal formalism
Instruments	log settler mortality initial indigenous population	legal origins

Table 1 *Description of the data used in regression analysis with multivariate endogenous variables by Acemoglu and Johnson (2005). Both property rights institutions and contracting institutions are proxied by their own set of (endogenous) variables, which in turn have their designated set of instruments.*

Contracting institutions are proxied for by three indexes of legal formalism (Djankov et al., 2003). One quantifies the formal procedures associated with collecting on a bounced check, one quantifies the overall procedural complexity of resolving a court case involving nonpayment of commercial debt, and one counts the number of procedures involved in that process.

The authors give three instruments, each of which is assigned either to the "property rights institution" or "contracting institutions" set of variables. The authors argue that the (log) settler mortality in countries that were colonized by European nations, as a well as the initial indigenous population density, are valid instruments for property rights institutions. The argument is that colonizers were more likely to develop extracting institutions when there was a larger population to put to work, and where their own survival was more likely.

The authors further argue that the legal origin of the country is a valid instrument for contracting institutions. The argument is that the British imposed common law systems on the countries they colonized, whilst other European powers had civil-law systems, and the system which a country was imposed impacts its degree of legal formalism (Djankov et al., 2003).

	OLS	2SLS	Median	IV-Median
β_{LF}	0.19 $(0.05, 0.32)$	0.44 $(0.11, 0.77)$	0.18 $(0.12, 0.39)$	0.57 $(-0.22, 0.70)$
β_{Exp}	0.52 $(0.39, 0.65)$	1.00 $(0.59, 1.41)$	0.62 $(0.46, 0.64)$	0.81 $(0.51, 2.04)$
β_{PP}	0.37 $(0.19, 0.55)$	0.06 $(-0.68, 0.80)$	0.2198 $(0.09, 0.55)$	-0.03 $(-1.86, 0.67)$

Table 2 *Regression output. The outcome variable is log GDP per capita in 1995. The rows are β_{LF} for legal formalism, β_{Exp} for protection against expropriation, β_{PP} for private property. The confidence intervals have size $\alpha = 0.1$.*

Table 2 speaks to the importance of instrumentation. As pointed out in Acemoglu and Johnson (2005), attenuating endogeneity using instrumental variables has an important impact on the linear regression estimates. Likewise, we find that instrumenting has an important impact on quantile regression estimate.

Table 3 illustrates the robustness of the pivotal test, in spite of the failure of the homoskedasticity assumption suggested by Table 3 and Figure 5 (see discussion in Section 7 concerning the failure of the homoskedasticity assumption). We can see that the confidence interval obtained using the pivotal statistic is well within the range of confidence intervals obtained using the robust test statistic, for reasonable choices of bandwidth parameters.

	homoskedastic	$0.8h_{n,HS}$	$h_{n,HS}$	$1.2h_{n,HS}$
β_{LF}	$(-0.21, 0.67)$	$(-0.22, 0.69)$	$(-0.22, 0.70)$	$(-0.28, 0.70)$
β_{Exp}	$(0.51, 2.05)$	$(0.48, 2.23)$	$(0.51, 2.04)$	$(0.51, 2.23)$
β_{PP}	$(-1.86, 0.67)$	$(-1.91, 0.68)$	$(-1.86, 0.67)$	$(-1.90, 0.68)$

Table 3 *Confidence intervals for the median regression, $\tau = 0.5$, with statistical size $1 - \alpha = 0.9$. "Homoskedastic" corresponds to the regression rankscore test under the homoskedasticity assumption, the other columns correspond to the robust regression rankscore test and are designated by the bandwidth choice, where $h_{n,HS}$ is the optimal bandwidth (Koenker, 2005).*

Any value around the optimal bandwidth may be considered as reasonable since the exact choice of constant ultimately depends on a parametric assumption. One could argue that any value between $0.8h_{n,HS}$ and $1.2h_{n,HS}$ is reasonable, and that we pick arbitrarily amongst them. Table 3 illustrates the point that the homoskedastic statistic, while devoid of the tuning parameters, tends to produce output closely resembling that of the robust statistic for some reasonable tuning parameter. Bai, Pouliot, Shaikh (2019) present simulation results displaying the resilience of the homoskedastic statistic's good coverage to adversarial (and heteroskedastic) data generating processes, further encouraging its use, at the very least for data exploration purposes.

The results of Table 3 should be complemented and compared with analogous coverage results in Table 7 of Section 7.

	model a	model b	model c	model d	model e	model f	model g
β_{LF}	-0.03 $(-0.41,0.62)$			0.23 (0.11,0.59)		0.44 (0.23,0.79)	0.57 $(-0.21,0.67)$
β_{Exp}		0.68 (0.63,1.34)		0.90 (0.69,1.00)	0.67 (0.65,3.30)		0.81 (0.51,2.05)
β_{PP}			1.07 (0.01,2.50)		1.03 $(-2.18,0.19)$	1.51 (1.12,1.88)	0.48 $(-1.86,0.67)$

Table 4 *Outcome is log GDP per capita in 1995. Model a is instrumented with legal origins. Models b, c, and e are instrumented using log settler mortality and initial indigenous population. Models d, f, and g are instrumented using all three instruments.*

The results of Table 4 emphasize the possible issues of "omitted variable" like errors in estimation, encouraging the use of all available and pertinent endogenous variables.

The quantile regression output, as displayed in Figure 3, suggests that greater legal formalism reduced credit –perhaps it made credit harder to obtain– for small creditors, but increased it for large creditors. Protection against expropriation has a positive impact on contemporaneous credit across the distribution, but a larger effect in the left tail of the distribution. Curiously, private property institutions as captured by the private property index appear to have a negative effect on contemporaneous credit, especially so for small creditors.

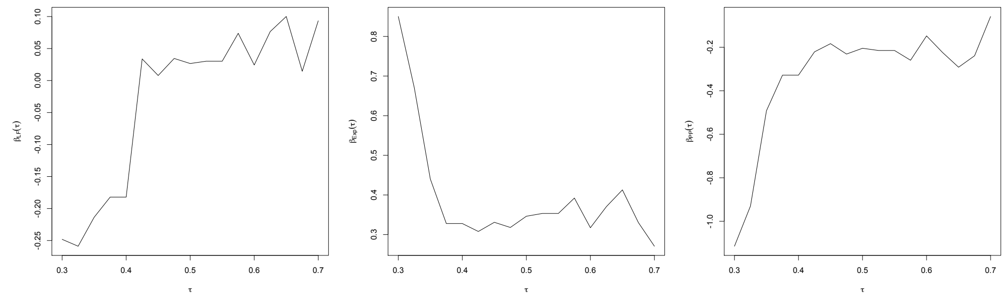


Figure 5 $\beta(\tau)$ *versus* τ. *The outcome is credit to private sector in 1998. From Left to write, the plotted regression coefficients are legal formalism, protection against expropriation, and private property index.*

The legal origins of a country may strike one as rather weakly identifying its legal formalism. The formality of procedures for collecting on a bounced check or for resolving a nonpayment court case may certainly find much more explanatory variation in conjunctural elements such as cultural norms, size of government, measures of corruption –which presumably find themselves very little explanatory variation in the legal origin of the country– than in the legal origins of the country.

Producing robust confidence intervals for β_{LF} requires subvector inference under weak identification. As discussed in Section 5, two typical approaches are the projection method using standard statistic and inverting a null for the entire weakly identified vector of parameters, and the Bonferroni method, which requires modeling the weak identification.

The validity of the instrument inspires more skepticism, as the colonizing country affects many aspects of a colony other than its legal formalism –culture, early technology, etc– which may in turn affect GDP. The inference result should thus be interpreted with commensurate care.

	IVQR	IVQR-Proj	IVQR-Square-Proj
$\tau = 0.25$	0.38 $(-1.21, 1.60)$	0.38 $(-2.91, 3.34)$	0.38 $(-1.97, 2.17)$
$\tau = 0.5$	0.57 $(-0.22, 0.70)$	0.57 $(-3.26, 4.34)$	0.57 $(-0.95, 0.98)$
$\tau = 0.75$	0.48 $(-0.08, 0.56)$	0.48 $(-3.06, 3.28)$	0.48 $(-1.95, 2.08)$

Table 5 *Regression summary for the legal formalism variable, with weak identification robust confidence intervals. The protection against expropriation and private property index variables are controlled and instrumented for.*

Although the confidence intervals do not allow us to conclude that there is a significant effect of legal formalism at either of the $\tau = 0.25$, 0.5, or 0.75 quantile, the output is interesting from a methodological standpoint. The confidence intervals obtained by projection via a rectangular regression rankscore statistic are much narrower than those obtained by the standard projection method.

7. Simulations and Discussions

Chen and Lee (2018) introduce a simulation design and, using the Skorohod representation, produce oracle values with which to compare estimates.

We first compare our mixed integer linear programming formulation to the GMM formulations of Chen and Lee (2017) and Zhu (2018),

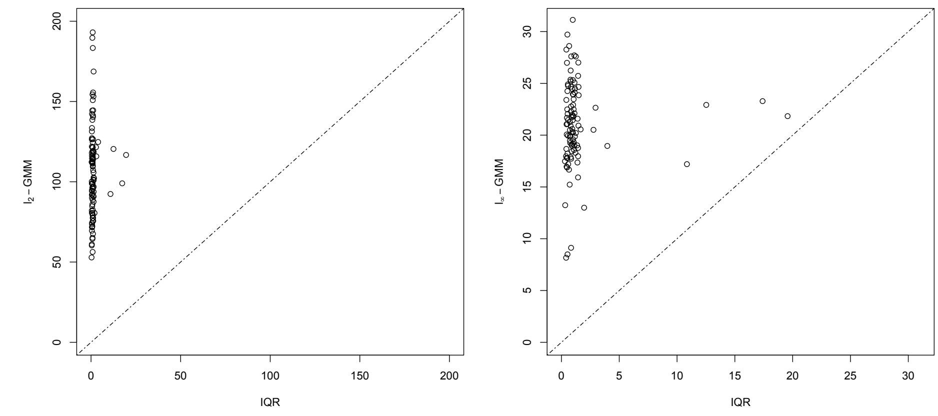


Figure 5 *The left-hand side plots the IQR MILP versus the ℓ_2-GMM of Chen and Lee (2017). The right-hand side plots the ℓ_∞-GMM of Zhu (2018). The implementation does not use preprocessing or additional cuts.*

Since both estimation methods are exact,[15] the simulation design provides interesting material for comparing the method of moments and IQR estimator for causal inference with the quantile regression function. For the same size dataset, both methods are expected to perform very similarly, as suggested by the alternative derivation of IQR in Subsection 2.3. When comparing both methods with $n = 100$ observations, the GMM approach appears to have slightly lower bias, while IQR has typically slightly lower variance, producing largely comparable root mean squared errors (RMSE). The GMM approach however takes

[15] They are exact in the computational sense that they solve for the global optimum.

94 seconds to run on average, while the IQR takes 29 seconds. In typical applications, one may have a larger dataset and, in particular for data exploration purposes, may be more constrained on time than on data. The gains from IQR are striking if we consider solving the problem with a fixed computational budget. If we cap the time/computational budget to what is required by GMM with the100 observations, we find that we can run IQR with 500 observations, be well "under budget" with an average running time still of about half that of GMM, and RMSE almost systematically half of that of GMM.

	$n = 100$						$n = 500$		
	GMM			IQR			IQR		
	Bias	RMSE	Time	Bias	RMSE	Time	Bias	RMSE	Time
$\theta_0(0.25)$	0.0109	0.2436	(94,197)	-0.1092	0.1568	(29,30)	-0.2215	0.2474	(53, 59)
$\theta_1(0.25)$	-0.0327	0.3554	*	-0.3345	0.4479	*	-0.3534	0.3840	*
$\theta_2(0.25)$	0.0003	0.1642	*	-0.2400	0.3055	*	-0.0509	0.0905	*
$\theta_3(0.25)$	0.0064	0.2332	*	0.0166	0.1907	*	-0.0277	0.0908	*
$\theta_0(0.5)$	0.0161	0.2498	(348, 989)	-0.2020	0.2420	(30,35)	-0.0063	0.0843	(44, 59)
$\theta_1(0.5)$	-0.0412	0.3241	*	-0.2195	0.2456	*	-0.0230	0.1494	*
$\theta_2(0.5)$	-0.0012	0.1561	*	0.0626	0.0813	*	0.0099	0.0580	*
$\theta_3(0.5)$	0.0012	0.2047	*	0.1347	0.1347	*	0.0217	0.0654	*
$\theta_0(0.75)$	0.0187	0.3046	(86,186)	0.1222	0.2187	(29, 30)	0.2049	0.2447	(41,54)
$\theta_1(0.75)$	-0.0358	0.3425	*	0.1445	0.4541	*	0.3454	0.3773	*
$\theta_2(0.75)$	-0.0022	0.1820	*	0.1110	0.2711	*	0.0639	0.0788	*
$\theta_3(0.75)$	0.0035	0.2393	*	0.2020	0.3046	*	0.0434	0.1296	*

Table 6 *Bias, RMSE, and computation times (mean, max) for GMM and IQR for first quantile regression coefficients.*

Table 7 describes coverage of different inference methods with different tuning. We can see that the regression rankscore test offers coverage very close to the nominal level of 95%. Importantly, and in spite of the DGP not satisfying the homoskedasticity assumption, we find that the pivotal test statistic whose derivation relies on the homoskedasticity assumption, performs remarkably well. Even to the extend that it may have incorrect coverage, its error is less than one would obtain with the robust test statistic using some reasonable value of the tuning parameters.

	GMM			IQR			IQR*
	$0.8h_n^*$	h_n^*	$1.2h_n^*$	$0.8h_n^*$	h_n^*	$1.2h_n^*$	
$\theta_1(0.25)$	0.906	0.914	0.918	0.950	0.952	0.944	0.948
$\theta_2(0.25)$	0.912	0.924	0.934	0.961	0.949	0.962	0.950
$\theta_3(0.25)$	0.916	0.926	0.938	0.962	9.957	0.964	0.962
$\theta_1(0.5)$	0.938	0.944	0.952	0.962	0.948	0.938	0.954
$\theta_2(0.5)$	0.950	0.958	0.966	0.960	0.964	0.950	0.952
$\theta_3(0.5)$	0.896	0.916	0.928	0.952	0.930	0.968	0.954
$\theta_1(0.75)$	0.922	0.938	0.944	0.950	0.958	0.958	0.924
$\theta_2(0.75)$	0.892	0.896	0.908	0.964	0.952	0.948	0.950
$\theta_3(0.75)$	0.918	0.928	0.942	0.952	0.972	0.960	0.970

Table 7 *IQR implements inference by inversion of the robust regression rankscore test, which requires the specification of a bandwidth parameter. IQR* implements inference by inversion of the "homoskedastic" regression rankscore test, which does not require the specification of a bandwidth parameter under the homoskedasticity assumption.*

Table 7 extends the simulation results from Chen and Lee (2018). It details how the choice of bandwidth amongst reasonable values may have nontrivial impacts on inference in the GMM approach. We also remark that the same is true, to a lesser extent, for regression rankscore inference.

A key observation is that the pivotal procedure, which relies on a homoskedasticity assumption (see below for further discussion) typically not satisfied in practice, and specifically not satisfied in the data generating process of this simulation, shows itself to be quite robust. Its coverage is quite accurate and, perhaps more to the point, is at least in this simulation typically within the coverages obtained over a range of "reasonable" choices of bandwidth parameters. Similar characteristics were documented in simulation exercises in Bai et al. (2019), where it was also observed that it was particularly challenging to design a adversarial DGP which would seriously invalidate coverage with the homoskedastic regression rankscore test.

As discussed in Koenker (1994) and Bai, Pouliot, and Shaikh (2019), the principal virtue of the homoskedasticity assumption –i.e., $f_\epsilon(0|X) = f_\epsilon(0)$ for all X, where $\epsilon = Y - X\beta(\tau)$ – is to produce an asymptotic variance formula for regression rankscore tests which does not require estimation of the density of regression errors, and thus the choice of a bandwidth parameter. As we saw in Section 5, this virtue is inherited by regression rankscore inference for instrumental variables quantile regression, even in the case of subvector inference. The main drawback of the homoskedasticity assumption is that it is understood to impose strong restrictions on the regression coefficient. It has been conventional wisdom in the quantile regression folklore that some variant of the homoskedasticity assumption essentially imposes that hyperplanes of conditional quantiles for different quantile values τ had to be parallel.[16] We give a formal result characterizing the implications of the homoskedasticity assumption.

Proposition 2 *Suppose that you have $(X, Y) \sim \mathcal{F}$ and a quantile regression coefficient function $\beta : (0,1) \to \mathbb{R}^p$ such that, for*

$$\epsilon(\tau)|X := Y - X\beta(\tau) \tag{33}$$

with density $f_{\epsilon(\tau)|X}$, we have that

$$\int_{-\infty}^{0} f_{\epsilon(\tau)|X}(r)dr = \tau. \tag{34}$$

Suppose that $f_{\epsilon(\tau)|X}$ is monotone decreasing away from zero. Assume that for some $\tau \in (0,1)$,

$$f_{\epsilon(\tau)|X}(0) = f_{\epsilon(\tau)|X'}(0),$$

for all $X, X' \in \mathbb{R}^p$. Then,

$$\beta(\tau) = \beta(\tau'),\ \forall\ \tau' \in (0,1).$$

Note that assumption (33) and (34) below refer to the well-specified case, which allows for a straightforward and intuitive proof.[17]

Although it was part of the folklore that the homoskedasticity assumption implied parallel hyperplanes, it was perhaps not clear whether or not it was innocuous to assume homoskedasticity at a single quantile. In fact, Proposition 2 states that assuming homoskedasticity at a single quantile implies parallel hyperplanes at all quantiles.

Remark that the assumption of such parallel conditional quantile hyperplanes is not

[16] I thank Manuel Arellano, Stéphane Bonhomme, and Josh Angrist for informative conversations to that effect.

[17] The argument for the proof may be laid out as follows.

Pick any $\tau' \neq \tau$. Without loss of generality, assume $\tau > \tau'$. Pick some $X \in \mathbb{R}^p$. Then,

$$\begin{aligned}\tau - \tau' &= \int_0^{X^T(\beta(\tau)-\beta(\tau'))} f_{\epsilon(\tau')|X}(r)dr \\ &\geq X^T(\beta(\tau) - \beta(\tau')) \cdot f_{\epsilon(\tau')|X}(X^T(\beta(\tau) - \beta(\tau'))) \\ &= X^T(\beta(\tau) - \beta(\tau')) \cdot f_{\epsilon(\tau)|X}(0) \\ &= X^T(\beta(\tau) - \beta(\tau')) \cdot f_{\epsilon(\tau)}(0),\end{aligned}$$

where the last equality holds by the homoskedasticity assumption. For X_j large enough, for any $j = 1, ..., p$, we find that $\beta_j(\tau) = \beta_j(\tau')$.

□

only typically implausible, it challenges a very premise of quantile regression analysis, which is that more information is to be gotten by inspecting the conditional quantile regression across quantiles.

The homoskedasticity assumption is thus a salient instance of the celebrated aphorism according to which all models all wrong, yet some remain useful. The relevant empirical point is indeed that although the homoskedasticity assumption and the closed-form asymptotic variance formula it entails are wrong in theory, they are useful in practice. In fact, evidence such as that displayed in Table 7 suggests that they are "often correct in practice", comparatively to the robust covariance statistic which is correct in theory under weaker assumptions. Specifically, we document in Table 7 how, in the simulation design described above, the homoskedastic test statistic typically behaves like the robust test statistic for some reasonable choice of bandwidth parameter.

We find that the rank test, even though it is not in this case asymptotically valid, performs quite well with coverage close to the nominal size of the test.

8. Conclusion

As exemplified in this article, modern optimization technology may be employed to solve combinatorially difficult econometrics problems, and to provide global optimality guarantees, even when the surface over which we optimize is not convex or even unimodal. There are large benefits to doing so, as grid search methods tend to be slow and imprecise, and do not provide global optimality certificates.

Serious consideration of the optimization problem underlying econometric estimators instructs methodological development beyond computations. The duality structure of the quantile regression linear program within the inverse quantile regression mixed integer linear program revealed the intimate relation between the GMM and the IQR estimators, and offered a natural framework within which to construct and interpret the regression rankscore test for IQR, thus extending the default inference methodology of quantile regression to the instrumental variables case.

We could reduce the loss in power incurred when carrying out the projection method by producing rectangular joint confidence regions. Weak identification robust inference strategies often involved sophisticated constructions and conditioning approaches. In a sense, we shifted the burden of difficulty from the statistical methodology to the computational methodology, thus delivering more powerful robust subvector inference methodology that remains as easy to interpret as the classical projection method.

Taking a wider perspective, we find that an optimization-conscious approach to econometrics allows the researcher to not only develop more computationally efficient estimators but, by conceptualizing the statistical object as the output of an algorithm or optimization problem instead of the minimand of an objective function, provides a fresh angle from which to develop methods and theory.

Acknowledgements

I am indebted, for their comments and suggestions, to Gary Chamberlain, Isaiah Andrews, Christian Genest, Mikkel Plagborg-Muller, Manuel Arellano, Stephane Bonhomme, Chris Hansen and Rahul Mazumder. Hung Ho has provided outstanding and insightful research assistance.

I thank the seminar participants at Columbia, the University of Southern California, and the Booth School of Business for valuable comments.

Appendix

A.1 Technical Material

We provide the the proofs deferred in the main text.

Proof of Proposition 1

Define the object of interest

$$\hat{W}_n = n^{-1/2} \sum_{i=1}^{n} \Phi_{i,J}(\hat{a}_i - (1-\tau)),$$

and its population counterpart

$$W_n = n^{-1/2} \sum_{i=1}^{n} \Phi_{i,J}(a_i^* - (1-\tau)),$$

where $a_i^* = \mathbf{1}\left\{\varepsilon_i > 0\right\}$ and $\varepsilon_i := Y_i - D_i\beta_D - X_i\beta_X$.

Let $\tilde{Y}_{i,J} = Y - D_{i,J}^T\beta_{D,J}$, for $i = 1, ..., n$. Consider

$$\hat{W}_n = W_n + n^{-1/2} \sum_{i=1}^{n} \Phi_{i,J}(\hat{a}_i - a_i^*)$$

$$= W_n - n^{-1/2} \sum_{i=1}^{n} \Phi_{i,J}\mathbf{1}\left\{\tilde{Y}_{i,J} - D_{i,-J}\hat{\beta}_{D,-J} \leq X_i^T\hat{\beta}_X\right\}$$

$$-n^{-1/2} \sum_{i=1}^{n} \Phi_{i,J}\mathbf{1}\left\{\tilde{Y}_{i,J} - D_{i,-J}\beta_{D,-J} \leq X_i^T\beta_X\right\} + n^{-1/2} \sum_{i=1}^{n} \Phi_{i,J}\mathbf{1}\left\{\tilde{Y}_{i,J} - D_{i,-J}\hat{\beta}_{D,-J} = X_i^T\hat{\beta}_X\right\}$$

$$= W_n - n^{-1/2} \sum_{i=1}^{n} \Phi_{i,J} \mathbf{1} \left\{ \tilde{Y}_{i,J} - D_{i,-J} \hat{\beta}_{D,-J} \leq X_i^T \hat{\beta}_X \right\}$$

$$+ n^{-1/2} \sum_{i=1}^{n} \Phi_{i,j} \mathbf{1} \left\{ \tilde{Y}_{i,J} - D_{i,-J} \beta_{D,-J} \leq X_i^T \beta_X \right\} + o_p(1).$$

The last equality obtains by lemma ?? of Bai, Pouliot and Shaikh (2019).

Adding and subtracting $\sqrt{n}\ E\left[\Phi_{i,J} \mathbf{1}\left\{\varepsilon_i \leq n^{-1/2}(D_{i,J}^T, X_i^T, Z_i^T)t\right\}\right]\big|_{t=n^{1/2}(\hat{\beta}-\beta)}$, $\sqrt{n} E\left[\Phi_{i,J} \mathbf{1}\left\{\varepsilon_i \leq 0\right\}\right]$, and $E\left[f_{\varepsilon_i}(0)\Phi_{i,J}(D_{i,J}^T, X_i^T, Z_i^T)t\right]\big|_{t=n^{1/2}(\hat{\beta}-\beta)}$, we obtain

$$\hat{W}_n = W_n - E\left[f_{\varepsilon_i}(0)\Phi_{i,J}(D_{i,J}^T, X_i^T, Z_i^T)\right] n^{1/2}(\hat{\beta}-\beta) - R_{n,3} - R_{n,4} + o_p(1),$$

where

$$R_{n,3} = \sqrt{n} \Bigg(E\left[\Phi_{i,J} \mathbf{1}\left\{\varepsilon_i \leq n^{-1/2}(D_{i,J}^T, X_i^T, Z_i^T)t\right\}\right]\Big|_{t=n^{1/2}(\hat{\beta}-\beta)}$$

$$- E\left[\Phi_{i,J} \mathbf{1}\left\{\varepsilon_i \leq 0\right\}\right] - n^{-1/2}\ E\left[f_{\varepsilon_i}(0)\Phi_{i,J}(D_{i,J}^T, X_i^T, Z_i^T)t\right]\big|_{t=n^{1/2}(\hat{\beta}-\beta)}\Bigg)$$

$$= \sqrt{n} \Bigg(E\left[\Phi_{i,J} E\left[\mathbf{1}\left\{\varepsilon_i \leq n^{-1/2}(D_{i,J}^T, X_i^T, Z_i^T)t\right\}\Big| D_{i,J}, X_i, Z_i\right]\right]\Big|_{t=n^{1/2}(\hat{\beta}-\beta)}$$

$$- E\left[\Phi_{i,J} E\left[\mathbf{1}\left\{\varepsilon_i \leq 0\right\}\right| D_{i,J}, X_i, Z_i\right]\right] - n^{-1/2}\ E\left[f_{\varepsilon_i}(0)\Phi_{i,J}(D_{i,J}^T, X_i^T, Z_i^T)t\right]\big|_{t=n^{1/2}(\hat{\beta}-\beta)}\Bigg)$$

$$= \sqrt{n} \Bigg(E\left[\Phi_{i,J} F_{\varepsilon_i}(n^{-1/2}(D_{i,J}^T, X_i^T, Z_i^T)t)\right]\Big|_{t=n^{1/2}(\hat{\beta}-\beta)}$$

$$- E\left[\Phi_{i,J} F_{\varepsilon_i}(0)\right] - n^{-1/2}\ E\left[f_{\varepsilon_i}(0)\Phi_{i,J}(D_{i,J}^T, X_i^T, Z_i^T)t\right]\big|_{t=n^{1/2}(\hat{\beta}-\beta)}\Bigg)$$

and

$$R_{n,4} = -n^{-1/2} \sum_{i=1}^{n} \Bigg(\Phi_{i,J} \mathbf{1}\left\{\tilde{Y}_{i,J} - D_{i,-J}\hat{\beta}_{D,-J} \leq X_i^T \hat{\beta}_X + Z_i^T \hat{\beta}_Z\right\}$$

$$- E\left[\Phi_{i,J} \mathbf{1}\left\{\varepsilon_i \leq n^{-1/2}(D_{i,J}^T, X_i^T, Z_i^T)t\right\}\right]\Big|_{t=n^{1/2}(\hat{\beta}-\beta)}\Bigg)$$

$$+ n^{-1/2} \sum_{i=1}^{n} \left(\Phi_{i,j} \mathbf{1}\left\{\tilde{Y}_{i,J} - D_{i,-J}\beta_{D,-J} \leq X_i^T \beta_X\right\} - E\left[\Phi_{i,J} \mathbf{1}\left\{\varepsilon_i \leq 0\right\}\right]\right) + o_p(1)$$

By lemmas ?? and ?? of Bai, Pouliot and Shaikh (2019), $R_{n,3}$ and $R_{n,4}$ are $o_p(1)$.

Given assumptions e, f , and Assumptions 1 and 2, Theorem 3 of Chernozhukov and

Hansen (2006) states that

$$\sqrt{n}\left((\hat\beta_{D,-J}^T(\cdot),\hat\beta_X^T(\cdot))^T-(\beta_{D,-J}^T(\cdot),\beta_X^T(\cdot))^T\right)$$

$$=J(\cdot)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n(\tau-\mathbf{1}\left\{\varepsilon_i(\tau)<0\right\})\left[\Phi_i(\tau)^T,X_i^T\right]^T+o_p(1)\xrightarrow{d}b(\cdot),$$

where $b(\cdot)$ is a mean zero Gaussian process with covariance function $Eb(\tau)b(\tau')=J(\tau)^{-1}S(\tau,\tau')J(\tau')^{-T}$,

$$J(\tau)=E\left[f_{\varepsilon(\tau)}(0|X,D,Z)\Psi(\tau)\left(D^T,X^T\right)\right],\ S(\tau,\tau')=(\min(\tau,\tau')-\tau\tau')E\Psi(\tau)\Psi^T(\tau').$$

Which is to say, omitting τ for brevity, e.g., $\hat\beta_X=\hat\beta_X(\tau)$,

$$\sqrt{n}\left((\hat\beta_{D,-J}^T,\hat\beta_X^T)^T-(\beta_{D,-J}^T,\beta_X^T)^T\right)=J^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n(\tau-\mathbf{1}\left\{\varepsilon_i(\tau)<0\right\})\left[\Phi_i(\tau)^T,X_i^T\right]^T+o_p(1),$$

$$E\left[f_{\varepsilon_i}(0)\Phi_{i,J}(D_{i,J}^T,X_i^T)\right]n^{1/2}\left((\hat\beta_{D,-J}^T,\hat\beta_X^T)^T-(\beta_{D,-J}^T,\beta_X^T)^T\right)$$

$$=E\left[f_{\varepsilon_i}(0)\Phi_{i,J}(D_{i,J}^T,X_i^T)\right]E\left[f_{\varepsilon_i}(0)\left(\Phi_i(\tau)^T,X_i^T\right)^T\left(D^T,X^T\right)\right]^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n(\tau-\mathbf{1}\left\{\varepsilon_i(\tau)<0\right\})\left[\Phi_i(\tau)^T,X_i^T\right]^T$$

$+o_p(1)$

$$=E\left[f_{\varepsilon_i}(0)\Phi_{i,J}(D_{i,J}^T,X_i^T)\right]E\left[f_{\varepsilon_i}(0)\left(\Phi_i(\tau)^T,X_i^T\right)^T\left(D^T,X^T\right)\right]^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n(a_i^*-(1-\tau))\left[\Phi_i(\tau)^T,X_i^T\right]^T$$

$+o_p(1)$,

using assumption 2.

Then,

$$\hat W_n=n^{-1/2}\sum_{i=1}^n(a_i^*-(1-\tau))\tilde\Phi_{i,j}+o_p(1),$$

where

$$\tilde\Phi_{i,j}=\Phi_{i,J}-E\left[f_\varepsilon(0|X,D_{-J},\Phi)\Phi_{i,J}(D_{-J}^T,X_i^T)\right]E\left[f_\varepsilon(0|X,D_{-J},\Phi)\left(\Phi_{-J}^T,X_i^T\right)^T\left(D_{-J}^T,X^T\right)\right]^{-1}\left[\Phi_{-J}^T,X^T\right]^T.$$

□

A.2 The Reduced Program Using Parameter Specifications From Advanced Warm-Starts

The program, when fixing outliers as described in Section 3, may be submitted to the solver in the following formulation.

$$\min \mathbf{1}_{p_Z}^T \beta_{Z,-} + \mathbf{1}_{p_Z}^T \beta_{Z,+}$$

subject to

$$[\mathbf{X}, \mathbf{Z}, I, -I] \begin{pmatrix} \beta_X \\ \beta_Z \\ u \\ v \end{pmatrix} + \mathbf{D}\beta_D = \mathbf{Y}$$

$$\mathbf{X}_{\mathcal{I}}^T a_{\mathcal{I}} = -\mathbf{X}_{O}^T a_{\mathcal{O}} + (1-\tau)\mathbf{X}^T \mathbf{1}_n$$

$$\mathbf{Z}_{\mathcal{I}}^T a_{\mathcal{I}} = -\mathbf{Z}_{O}^T a_{\mathcal{O}} + (1-\tau)\mathbf{Z}^T \mathbf{1}_n$$

$$v_{\mathcal{I}} \leq l_{\mathcal{I}} \cdot M,\ u_{\mathcal{I}} \leq k_{\mathcal{I}} \cdot M,\ v_{\mathcal{O}} \leq l_{\mathcal{O}} \cdot M,\ u_{\mathcal{O}} \leq k_{\mathcal{O}} \cdot M$$

$$a_{\mathcal{I}} \geq k_{\mathcal{I}},\ a_{\mathcal{I}} \leq \mathbf{1}_{|\mathcal{I}|} - l_{\mathcal{I}}$$

$$k_{\mathcal{I}} \in \{0,1\}^{|\mathcal{I}|},\ l_{\mathcal{I}} \in \{0,1\}^{|\mathcal{I}|}$$

$$a_{\mathcal{I}} \in [0,1]^{|\mathcal{I}|}$$

$$\beta_{Z,-}, \beta_{Z,+}, u, v \geq 0$$

$$\beta_D, \beta_X \text{ free.}$$

The upper bounds bound on $u_{\mathcal{O},i}$ should be set to 0 when $a_{\mathcal{O},i} = 0$, and the upper bound for $v_{\mathcal{O},i}$ should be set to 0 when $a_{\mathcal{O},i} = 1$. Alternatively/equivalently the program may be rewritten with these provided as constants –not variables– in the program, which is what we do in our implementation.

The equality of primal and dual objective, and the McCormick inequalities, take the following form

$$\tau u^T \mathbf{1}_n + (1-\tau)\, v^T \mathbf{1}_n = a_{\mathcal{I}}^T \mathbf{Y}_{\mathcal{I}} + a_{\mathcal{O}}^T \mathbf{Y}_{\mathcal{O}} - w_{\mathcal{I}}^T \mathbf{D}_{\mathcal{I}} - a_{\mathcal{O}}^T \mathbf{D}_{\mathcal{O}} \beta_D + (\tau - 1)\mathbf{1}_n^T \left(\mathbf{Y} - \mathbf{D}\beta_D\right)$$

$$w_{\mathcal{I}} \geq a_{\mathcal{I}} \underline{\beta_D}$$

$$w_{\mathcal{I}} \geq \beta_D \mathbf{1}_n + a_{\mathcal{I}} \overline{\beta_D} - \overline{\beta_D} \mathbf{1}_n$$

$$w_{\mathcal{I}} \leq \beta_D \mathbf{1}_n + a_{\mathcal{I}} \underline{\beta_D} - \underline{\beta_D} \mathbf{1}_n$$

$$w_{\mathcal{I}} \leq a_{\mathcal{I}} \overline{\beta_D}.$$

A.3 Extension of the Simulation Design of Chen and Lee (2018)

Chen and Lee (2018) produce a simulation design for a problem of dimension $(n, p) = (100, 3)$. Specifically, they model

$$Y = 1 + D_1 + D_2 + D_3 + (0.5 + D_1 + 0.25D_2 + 0.15D_3)\epsilon,$$

where the endogenous variables are themselves modeled as

$$D_1 = \Phi(Z_1 + \nu_1), D_2 = 2\Phi(Z_2 + \nu_2), D_3 = 1.5\Phi(Z_3 + \nu_3),$$

and the sources of variation are captured by

$$(\epsilon, \nu_1, \nu_2, \nu_3) \sim \mathcal{N}\left(\mathbf{0}_4, 0.25V\right),$$

where

$$V = \begin{pmatrix} 1 & 0.4 & 0.6 & -0.2 \\ 0.4 & 1 & 0 & 0 \\ 0.6 & 0 & 1 & 0 \\ -0.2 & 0 & 0 & 1 \end{pmatrix}.$$

In the article, we use an extension of this data generating process to general dimensions (n, p). Specifically, we use the model

$$Y = 1 + D\beta_D + (0.5 + D\beta_\epsilon)\epsilon,$$

where the endogenous variables are themselves modeled as

$$D_i = s_i\Phi(Z_i + \nu_i), i = 1, 2, \ldots, p,$$

and the sources of variation are captured by

$$(\epsilon, \nu_1, \ldots, \nu_p) \sim \mathcal{N}(\mathbf{0}_{p+1}, 0.25V),$$

where

$$V = \begin{pmatrix} 1 & a_1 & \dots & a_p \\ a_1 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ a_p & 0 & \dots & 1 \end{pmatrix}.$$

For example, the original Chen and Lee (2018) uses

$$n = 100, p = 3, \beta_D = \mathbf{1}_3, \beta_\epsilon = (1, 0.25, 0.15)^T,$$

$$(s_1, s_2, s_3) = (1, 2, 1.5), (a_1, a_2, a_3) = (0.4, 0.6, -0.2).$$

Our default design when p is greater than 3 is

$$\beta_D = \mathbf{1}_p, \beta_\epsilon = \left(1, 0.25, 0.15, (0.1, 0.1 + \frac{0.9}{p-3}, 0.1 + 2 \cdot \frac{0.9}{p-3}, ..., 1)\right)^T,$$

$$(s_1, \dots, s_p) = \left(1, 2, 1.5, (1, 1 + \frac{1}{p-3}, 1 + 2 \cdot \frac{1}{p-3}, ..., 2)\right).$$

Put otherwise, we extend β_ϵ using the sequence command $\text{seq}(0.1, 1, \mathit{length} = p - 3)$, and we extend s using the sequence command $\text{seq}(1, 2, \mathit{length} = p - 3)$.

We need to make sure V is positive-definite. The eigenvalues of V are $\left\{\mathbf{1}_{p-1}, 1 \pm \sqrt{a_1^2 + \dots + a_p^2}\right\}$. So to ensure positive-definiteness, we need to choose $\mathbf{a} = (a_1, \dots, a_p)$ such that $\|\mathbf{a}\| < 1$. We are using

$$\mathbf{a} = \|(0.4, 0.6, -0.2)\| \cdot \frac{(0.4, 0.6, -0.2, \text{seq}(-1, 1, \mathit{length} = p - 3))}{\|(0.4, 0.6, -0.2, \text{seq}(-1, 1, \mathit{length} = p - 3))\|}.$$

This way, when $p = 3$, we get exactly the original data generating process of Chen and Lee.

References

Andrews, Isaiah, James Stock, and Liyang Sun. *Weak instruments in iv regression: Theory and practice.* manuscript (2018).

Andrieu, Christophe, Nando De Freitas, Arnaud Doucet, and Michael I. Jordan. "An introduction to MCMC for machine learning." *Machine learning* 50, no. 1-2 (2003): 5-43.

Barrodale, Ian, and Frank DK Roberts. "An improved algorithm for discrete l_1 linear approximation." SIAM *Journal on Numerical Analysis* 10, no. 5 (1973): 839-848.

Beran, Rudolf, and Peter Hall. "Interpolated nonparametric prediction intervals and confidence intervals." *Journal of the Royal Statistical Society. Series B (Methodological)* (1993): 643-652.

Bertsimas, Dimitris, Angela King, and Rahul Mazumder. "Best subset selection via a modern optimization lens." *The annals of statistics* 44, no. 2 (2016): 813-852.

Bertsimas, Dimitris, and Robert Weismantel. *Optimization Over Integers.* Vol. 13. Belmont: Dynamic Ideas, 2005.

Bertsimas, Dimitris, Martin S. Copenhaver, and Rahul Mazumder. "Certifiably optimal low rank factor analysis." *The Journal of Machine Learning* Research 18, no. 1 (2017): 907-959.

Buchinsky, Moshe. "Estimating the asymptotic covariance matrix for quantile regression models a Monte Carlo study." *Journal of Econometrics* 68, no. 2 (1995): 303-338.

Buchinsky, Moshe. "Changes in the US wage structure 1963-1987: Application of quantile regression." *Econometrica* (1994): 405-458.

Chen, Le-Yu, and Sokbae Lee. "Exact computation of GMM estimators for instrumental variable quantile regression models." *Journal of Applied Econometrics* (2018).

Chamberlain, Gary. "Quantile regression, censoring, and the structure of wages." *In Advances in econometrics: sixth world congress*, vol. 2, pp. 171-209. 1994.

Chernozhukov, Victor, and Iván Fernández-Val. "Subsampling inference on quantile regression processes." *Sankhyā*: The Indian Journal of Statistics (2005): 253-276.

Chernozhukov, Victor, and Christian Hansen. "An IV model of quantile treatment effects." *Econometrica* 73, no. 1 (2005): 245-261.

Chernozhukov, Victor, and Christian Hansen. "Instrumental quantile regression inference for structural and treatment effect models." *Journal of Econometrics* 132, no. 2 (2006): 491-525.

Chernozhukov, Victor, and Christian Hansen. "Instrumental variable quantile regression." (2004).

Chernozhukov, Victor, and Han Hong. "An MCMC approach to classical estimation." *Journal of Econometrics* 115, no. 2 (2003): 293-346.

Chernozhukov, Victor, and Christian Hansen. "Instrumental variable quantile regression." (2004).

Chernozhukov, Victor, Christian Hansen, and Michael Jansson. "Finite sample inference for quantile regression models." *Journal of Econometrics* 152, no. 2 (2009): 93-103.

Delaunay, Boris. "Sur la sphere vide." *Izv. Akad. Nauk SSSR, Otdelenie Matematicheskii i Estestvennyka Nauk 7*, no. 793-800 (1934): 1-2.

Dufour, Jean-Marie. "Some impossibility theorems in econometrics with applications to structural and dynamic models." *Econometrica: Journal of the Econometric Society*

(1997): 1365-1387.

Dufour, Jean-Marie, and Joann Jasiak. "Finite sample limited information inference methods for structural equations and models with generated regressors." *International Economic Review* 42, no. 3 (2001): 815-844.

Dufour, Jean-Marie, and Mohamed Taamouti. "Projection-based statistical inference in linear structural models with possibly weak instruments." *Econometrica 73*, no. 4 (2005): 1351-1365.

Dufour, Jean-Marie. "Identification, weak instruments, and statistical inference in econometrics." *Canadian Journal of Economics/Revue canadienne d'économique* 36, no. 4 (2003): 767-808.

Edelsbrunner, Herbert, David Kirkpatrick, and Raimund Seidel. "On the shape of a set of points in the plane." *IEEE Transactions on information theory 29*, no. 4 (1983): 551-559.

Edelsbrunner, Herbert. "Alpha shapes—a survey." *Tessellations in the Sciences 27* (2010): 1-25.

Gendron, Bernard, and Teodor Gabriel Crainic. "Parallel branch-and-branch algorithms: Survey and synthesis." *Operations research* 42, no. 6 (1994): 1042-1066.

Graham, Ron, and Frances Yao. "*A whirlwind tour of computational geometry.*" The American Mathematical Monthly 97, no. 8 (1990): 687-701.

Gutenbrunner, C., and J. Jurecková. "Regression rank scores and regression quantiles." *The Annals of Statistics* (1992): 305-330.

Gutenbrunner, C., J. K. R. S. Jurečková, R. Koenker, and S. Portnoy. "Tests of linear hypotheses based on regression rank scores." *Journal of Nonparametric Statistics* 2, no. 4 (1993): 307-331.

Hoff, Peter D. *A first course in Bayesian statistical methods.* Springer Science & Business Media, 2009.

Imbens, Guido W., and Donald B. Rubin. *Causal inference in statistics, social, and biomedical sciences.* Cambridge University Press, 2015.

Kaido, Hiroaki, and Kaspar Wuthrich. "Decentralization estimators for instrumental variable quantile regression models." *arXiv preprint arXiv*:1812.10925 (2018).

Kennedy, Fetter E. "Randomization tests in econometrics." *Journal of Business & Economic Statistics* 13, no. 1 (1995): 85-94.

Koenker, Roger, and Zhijie Xiao. "Inference on the quantile regression process." *Econometrica* 70, no. 4 (2002): 1583-1612.

Koenker. *Quantile Regression.* Econometric Society monographs; no. 38. Cambridge University Press, 2005.

Koenker, Roger, and Gilbert Bassett Jr. "Regression quantiles." *Econometrica: journal of the Econometric Society* (1978): 33-50.

Koenker, Roger, Victor Chernozhukov, Xuming He, and Limin Peng, eds. *Handbook of Quantile Regression.* CRC Press, 2017.

Koenker, Roger. "Confidence intervals for regression quantiles." *Asymptotic Statistics*, pp. 349-359. Physica, Heidelberg, 1994.

Lasserre, Jean B., and Konstantin E. Avrachenkov. "The Multi-Dimensional Version of$\int$ ba xp dx." *The American Mathematical Monthly* 108, no. 2 (2001): 151-154.

Lehmann, Erich L., and Joseph P. Romano. *Testing statistical hypotheses.* Springer Science & Business Media, 2006.

Mangasarian, Olvi L. "Regularized linear programs with equilibrium constraints." In *Reformulation: Nonsmooth, Piecewise Smooth, Semismooth and Smoothing Methods*, pp. 259-268. Springer, Boston, MA, 1998.

Mikusheva, Anna. "Robust confidence sets in the presence of weak instruments." *Journal of Econometrics* 157, no. 2 (2010): 236-247.

Mikusheva, Anna. 'Survey on statistical inferences in weakly-identified instrumental variable models." *Прикладная эконометрика* 1 (29) (2013).

Moreira, Marcelo J., Jack R. Porter, and Gustavo A. Suarez. "Bootstrap validity for the score test when instruments may be weak." *Journal of Econometrics* 149, no. 1 (2009): 52-64.

Newey, Whitney K., and Daniel McFadden. "Large sample estimation and hypothesis testing." *Handbook of econometrics* 4 (1994): 2111-2245.

Pouliot, Guillaume, A. *Optimization-Conscious Econometrics.* PhD Course Manuscript. available online.

Pouliot, Guillaume, A. *Rectangular Confidence Regions with Applications to Weakly Identified Subvector Inference.* Manuscript.

Renka, Robert J. "*Algorithm 751: TRIPACK: a constrained two-dimensional Delaunay triangulation package.*" ACM Transactions on Mathematical Software (TOMS) 22, no. 1 (1996): 1-8.

Renka, R. J., A. Gebhardt, S. Eglen, S. Zuyev, and D. White. "*tripack: Triangulation of irregularly spaced data.*" R package version (2009): 1-3.

Rodríguez Casal, Alberto, and Beatriz Pateiro-López. "*Generalizing the convex hull of a sample: the R package alphahull.*" (2010).

Sidak, Zbynek, Pranab K. Sen, and Jaroslav Hajek. *Theory of rank tests.* Elsevier, 1999.

Staiger, Douglas, and James H. Stock. Instrumental Variables Regression with Weak Instruments. *Econometrica: Journal of the Econometric Society* (1997): 557-586.

Van der Vaart, Aad W. *Asymptotic statistics.* Vol. 3. Cambridge university press, 2000.

Voronoi, Georges. "Nouvelles applications des paramètres continus à la théorie des formes quadratiques. Deuxième mémoire. Recherches sur les parallélloèdres primitifs." *Journal für die reine und angewandte Mathematik 134* (1908): 198-287.

Xu, Guanglin, and Samuel Burer. "A branch-and-bound algorithm for instrumental variable quantile regression." *Mathematical Programming Computation* 9, no. 4 (2017): 471-497.

Zhu, Yinchu. “Mixed Integer Linear Programming: a new approach for instrumental variable quantile regressions and related problems”. manuscript. 2018.

Zubizarreta, José R., Dylan S. Small, Neera K. Goyal, Scott Lorch, and Paul R. Rosenbaum. "Stronger instruments via integer programming in an observational study of late preterm birth outcomes." *The Annals of Applied Statistics* 7, no. 1 (2013): 25-50.

Supplementary Appendix

A.1 Strong Formulations

Stronger formulations are here defined as mathematically equivalent formulations of a mixed integer linear programming problem that are solved faster using standard or available solution software and heuristics. Let $\mathbf{x} = (\mathbf{x}_I, \mathbf{x}_C)$, where $\mathbf{x}_I$ are integer variables and $\mathbf{x}_C$ continuous variables. Let $\mathcal{L}$ and $\mathcal{L}$' be two sets of linear equalities and inequalities such that

$$\{\mathbf{x} \ : \mathbf{x} \in \mathcal{L}, \ \mathbf{x}_I \in \{0,1\}\} = \{\mathbf{x} \ : \mathbf{x} \in \mathcal{L},' \ \mathbf{x}_I \in \{0,1\}\}.$$

Clearly, for any given $\mathbf{c} \in \mathbb{R}^n$, the problems $\min\{\mathbf{x}^T\mathbf{c} \ \ : \mathbf{x} \in \mathcal{L}, \ \mathbf{x}_I \in \{0,1\}\}$ and $\min\{\mathbf{x}^T\mathbf{c} \ : \mathbf{x} \in \mathcal{L}', \ \mathbf{x}_I \in \{0,1\}\}$ are equivalent. Suppose however that

$$\{\mathbf{x} \ : \mathbf{x} \in \mathcal{L}', \ \mathbf{x}_I \in [0,1]\} \supset \{\mathbf{x} \ : \mathbf{x} \in \mathcal{L}, \ \mathbf{x}_I \in [0,1]\}.$$

Then

$$\min_{\mathbf{x} \in \mathcal{L}', \mathbf{x}_I \in [0,1]} \mathbf{x}^T\mathbf{c} \leq \min_{\mathbf{x} \in \mathcal{L}, \mathbf{x}_I \in [0,1]} \mathbf{x}^T\mathbf{c} \leq \min_{\mathbf{x} \in \mathcal{L}, \mathbf{x}_I \in \{0,1\}} \mathbf{x}^T\mathbf{c}.$$

The $\mathcal{L}$ formulation produces a tighter bound in the continuous relaxation on the problem.

Typical, MILP solvers incrementally ameliorate the candidate solution by updating and solving such continuous relaxation of the mixed integer problem. Consequently, tighter relaxations accelerate the procedure. The impact of stronger formulation on solving time can be very important, whence the need for strong formulations.

A.1.1 Size of the Active Basis

The minimum number of a_i's that are strictly 0 or 1 depends on the number of independent variables. Hence, one strengthening constraint we include is the size of the inactive basis, which may be represented as the linear equality

$$\mathbf{1}_n^T k + \mathbf{1}_n^T l = n - p_X - p_Z. \tag{35}$$

Not that this is valid even if the data is not in standard position, which may occur in bootstrap samples.[18]

[18] By bootstrap samples, we mean samples obtained by resampling with replacement from the original data.

A.1.2 Integer Dual Feasibility

Consider for example a data of size $n = 9$ and $\tau = 0.5$. Then the dual feasibility condition with respect to the intercept implies that $\sum_i a_i = 4.5$, which implies $\sum_i k_i \leq 4.5$ and $\sum_i l_i \leq 4.5$. By integrality, these imply $\sum_i k_i \leq 4$ and $\sum_i l_i \leq 4$. In this case, the integral bounds obtained trivially. However, in the general case, i.e., for dual feasibility conditions with respect to an independent variable other than a constant, they are not trivial.

Strengthening inequalities –a.k.a., cuts or generated columns– may be generated for the dual feasibility constraint corresponding to nonconstant covariates, as applied to k and l. Indeed, it must be true that $X_{\cdot,j}^T k$ is less than the solution of

$$\max X_{\cdot,j}^T k$$

subject to

$$\mathbf{X}^T a = (1-\tau)\mathbf{X}^T \mathbf{1}_n \tag{36}$$

$$\mathbf{Z}^T a = (1-\tau)\mathbf{Z}^T \mathbf{1}_n \tag{37}$$

$$\mathbf{1}_n^T k + \mathbf{1}_n^T l = n - p_X - p_Z \tag{38}$$

$$a \geq k, \;\; a \leq \mathbf{1}_n - l \tag{39}$$

$$k \in \{0,1\}^n, \;\; l \in \{0,1\}^n, \;\; a \in [0,1]^n. \tag{40}$$

Note that generating this bound requires solving a MILP, and that the solution to the relaxed problem produces a redundant bound. Remark that (27) is a redundant constraint, there to strengthen the formulation.

A.1.3 Equality of Objective Functions

As indicated above, another sufficient characterization of the the quantile regression solution is that, in addition to primal and dual feasibility, the primal and dual objective functions be equal, i.e.,

$$\tau u^T \mathbf{1}_n + (1-\tau) v^T \mathbf{1}_n = a^T \mathbf{Y} - a^T \mathbf{D}\beta_D - (1-\tau)\mathbf{1}_n^T(\mathbf{Y} - \mathbf{D}\beta_D). \tag{41}$$

Adding condition (27) strengthens the formulation in that it shrinks the size of the feasible set of the continuously relaxed problem. However it does not produce, as such, a MILP formulation because the quadratic term $a^T \mathbf{D}\beta_D$ makes the constraint (27) quadratic and nonconvex. In order to add the constraint to the IQR problem formulation such that it remains a MILP, we can linearize the bilinear quadratic term by implementing the change of variable[19] $w_i = \beta_D a_i$, $i = 1, ..., n$, and introducing of the defining inequalities

$$w_i \geq a_i \underline{\beta_D},\ w_i \leq a_i \overline{\beta_D}, \tag{42}$$

$$w_i \geq \beta_D + a_i \overline{\beta_D} - \overline{\beta_D},\ w_i \leq \beta_D + a_i \underline{\beta_D} - \underline{\beta_D}. \tag{43}$$

These make up the McCormick envelope $\mathcal{M} = \mathcal{M}\left(\underline{\beta_D}, \overline{\beta_D}\right) \subset \mathbb{R}^{np_D+n+p_D}$, defined as

$$\mathcal{M} := \left\{w, a, \beta_D \left| a_i \underline{\beta_D} \leq w_i \leq a_i \overline{\beta_D}, \beta_D + a_i \overline{\beta_D} - \overline{\beta_D} \leq w_i \leq \beta_D + a_i \underline{\beta_D} - \underline{\beta_D} \right.\right\}.$$

In other words, the quadratic equality constraint (27) maybe approximated by

$$\tau u^T \mathbf{1}_n + (1-\tau)\, v^T \mathbf{1}_n = a^T \mathbf{Y} - w^T \mathbf{D} + (\tau - 1) \mathbf{1}_n^T \left(\mathbf{Y} - \mathbf{D}\beta_D\right) \tag{44}$$

together with the McCormick inequalities (28)-(29).

A key observation is that the McCormick inequalities are tight if one of the variable entering the bilinear terms is binary. Fortunately, at the quantile regression solution, $n - p$ of the dual variables a_i, $i = 1, ..., n$, are either 0 or 1 (Pouliot, 2017). Intuitively, if we enforce this characterization of the solution, the relaxation will be much tighter. Naturally, this is precisely what is enforced by the "complementary slackness" conditions (16)-(18). Without those, the program offers a linear programming relaxation of the IQR program, whose usefulness will be explored in Section 4.

A.2 Alternative Approaches for Preprocessing

A.2.1 Linear Programming Relaxation

As detailed in Subsection 3.1, the linear programing problem minimizing $\|\beta_Z\|_1$ subject to (11)-(15) and (28)-(30) is an approximation to the IQR estimator.

Because it is an approximation, the solution may not be feasible in the IQR program. It may thus be helpful to project the solution in order to produce a starting solution which is feasible for the IQR program.[20] The projection may be computed as follows. Obtain $\hat{\epsilon}_i = \hat{u}_i - \hat{v}_i$, $i = 1, ..., n$, from the relaxed problem. Rearrange the residuals in increasing order, as

$$\hat{\epsilon}_{(1)} \leq \hat{\epsilon}_{(2)} \leq \cdots \leq \hat{\epsilon}_{(n)},$$

[19] For β_D of dimension p_D we would have $w_{ij} = \beta_{D,j} a_i$, $i = 1, ..., n$, $j = 1, ..., p_D$.

[20] The solver may accept an infeasible solution. It would be nice to assess the gains from an informed projection.

and set

$$a_{(i)} = 0,\ i = 1, ..., \lfloor n \cdot (1-\tau) \rfloor - \lceil p/2 \rceil\,,$$

$$a_{(i)} = 1,\ i = \lfloor n \cdot (1-\tau) \rfloor - \lceil p/2 \rceil + p + 1, ..., n,$$

$$\mathcal{B} = \{ i\ :\ (i) \in \{ \lfloor n \cdot (1-\tau) \rfloor - \lceil p/2 \rceil + 1, ..., \lfloor n \cdot (1-\tau) \rfloor - \lceil p/2 \rceil + p \} \}\,,$$

$$\left(\tilde{\beta}_X^T, \tilde{\beta}_Z^T \right)^T = [\mathbf{X}, \mathbf{Z}]_{\cdot,\mathcal{B}}^{-1} \tilde{\mathbf{Y}}_{\mathcal{B}},$$

$$u = \max\left\{ 0, \tilde{\mathbf{Y}} - [\mathbf{X}, \mathbf{Z}] \left(\tilde{\beta}_X^T, \tilde{\beta}_Z^T \right)^T \right\},\ v = \max\left\{ 0, -\left(\tilde{\mathbf{Y}} - [\mathbf{X}, \mathbf{Z}] \left(\tilde{\beta}_X^T, \tilde{\beta}_Z^T \right)^T \right) \right\},$$

and

$$a_{\mathcal{B}} = [\mathbf{X}, \mathbf{Z}]_{\mathcal{B},\cdot}^{-T} \left((1-\tau) [\mathbf{X}, \mathbf{Z}]^T \mathbf{1} - [\mathbf{X}, \mathbf{Z}]^T a_{(1:n) \backslash \mathcal{B}} \right).$$

This produces a feasible solution.

A.2.2 Integer Programming Relaxation

Another preprocessing problem is to minimize $\|\beta_Z\|_1$ subject to (11)-(15) and (28)-(30), but to enforce $a \in \{0, 1\}$. Note that this problem may not be feasible, but that will typically be declared quickly by the solver. Further note that some instances of this problem may themselves be computationally intensive, hence it is important to set a low time limit on the allowed solving time. If the problem stops due to the time constraint, one may use the early-stopping solution. The same projection as in Subsection 4.1 may be applied if one wishes to produce a feasible solution.

A.2.3 Grid Search

The preprocessing consists in quickly solving for a good starting solution, but may also be used to specify part of the solution. A fast and approximative method such as a –coarse– grid search or a variant of gradient descent can be used to find an approximate solution which will then be used as an initial solution. Furthermore, we may use the preprocessing log to partially specify the optimal solution. If some a_j's are constant in the neighborhood of the preprocessing solution, we may call them "stable" and fix them in the optimization program. Note that this partially fixes the solution rather than suggesting an initial value for an iterative procedure, and as such requires either an ex post exact guarantee.

The index of the thus fixed dual variables are collected in the set $\mathcal{O}$ of outliers. We also define the complement set $\mathcal{I} = \{1, ..., n\} \backslash \mathcal{O}$.

We are now interested in the reduced program. Note that $v_{\mathcal{I}}$, $l_{\mathcal{I}}$, $u_{\mathcal{I}}$, $k_{\mathcal{I}}$, are still variables, as well as $v_{\mathcal{O}}$ and $u_{\mathcal{O}}$, but $l_{\mathcal{O}}$ and $k_{\mathcal{O}}$ are fixed. The reduced program, given explicitely in Appendix Section A.3, is obtained by simply fixing a_i, k_i and l_i , $i \in \mathcal{O}$, to their stable values, i.e., their value in the preprocessing solution.

One may get an ex post exact guarantee that the partial specfication of the solution was correct. Under weak regularity conditions, the optimal $\hat{\beta}_Z$ will be identically zero, which may be used as a confirmation that the partial specification was innocuous.

(n,p)	default	LPR	IPR	SS	SS_{fix}	QR	QR_{fix}	GS	GS_{fix}
$(100,3)$									
$(100,5)$									
$(500,5)$									
$(500,10)$									

Table 1 *Comparison of strategies for preprocessing. LPR and IPR use the warm start for the linear programming and integer programming relaxations, respectively.* SS, QR*, GS implement, respectively, subsampling, quantile regression, and grid search without fixing part of the solution.* S_{fix}, QR_{fix}, GS_{fix} *fix part of the solution. The data is generated according to extensions of the simulation design of Chen and Lee (2017), as detailed in the Appendix.*

A.2.4 Subsampling

The $\sqrt{n}$-convergence of the estimated regression coefficients, combined with the exponential rate of computational complexity, suggests it may be beneficial to extract statistical information from estimates evaluated on subsamples of the data. Indeed, running few IQR regressions on small random subsamples of the data allows for the detection of outliers, which may ten be fixed in the solution of the full-scale problem, thus reducing its size and computational complexity. Of course, these separate regressions on separate subsets are "embarrassingly parallel".

The subsampling preprocessing may be expressed in pseudocode as follows. Pick a small number of subsamples, say $S = \log(n)^2$ and a small subsample size, say $n_{\text{sub}} = \lfloor n/S \rfloor$. Tune the rejection with α, and the update rate with r_α. Reasonable values are $\alpha = 1$ and $r_\alpha = 1.1$.

1. Run S IQR regressions on n_{sub} observations sampled without replacement from the original data
2. Collect the residual vectors $\hat{\varepsilon}^{(s)} = \hat{u}^{(s)} - \hat{v}^{(s)}$ and $\hat{\beta}^{(s)} = \left(\hat{\beta}_X^{(s)}, \hat{\beta}_D^{(s)}\right)$ for each subsample $s = 1, ..., S$
3. Compute the sample covariance matrix $\hat{\Sigma}_{\hat{\beta}} = \frac{1}{S}\sum_{s=1}^{S}\left(\hat{\beta}^{(s)} - \overline{\hat{\beta}}\right)\left(\hat{\beta}^{(s)} - \overline{\hat{\beta}}\right)^T$ and sample residual variance $\hat{\sigma}^2_{\hat{\varepsilon}}$

4. For $i = 1, ..., n$, if $\hat{\varepsilon}_i^{(s)} < -\alpha\sqrt{(X_i^T, D_i^T)\hat{\Sigma}_{\hat{\beta}}(X_i^T, D_i^T)^T + \hat{\sigma}_{\hat{\varepsilon}}^2}$ for all s, then set i as a *negative outlier*. Likewise, if $\hat{\varepsilon}_i^{(s)} > \alpha\sqrt{(X_i^T, D_i^T)\hat{\Sigma}_{\hat{\beta}}(X_i^T, D_i^T)^T + \hat{\sigma}_{\hat{\varepsilon}}^2}$ for all s, then set i as a *positive outlier*.

5. Run the IQR regression with the outliers fixed.

   (a) If $\hat{\beta}_Z = 0$, stop.

   (b) If $\hat{\beta}_Z \neq 0$, update $\alpha \leftarrow r_\alpha \cdot \alpha$ and return to step 4.

In step 5, you may produce the starting solution for the remaining variables using the $\hat{\beta}_D^{(s)}$ value which, in the quantile regression of $Y - D\hat{\beta}_D^{(s)}$ on X and Z, produces the smallest $\left\|\hat{\beta}_Z\right\|_1$, and using the projection from subsection 4.1. In 5.b, one may alternatively leave α intact, and run an additional subsample, i.e., update $S \leftarrow S + 1$.

A.2.5 Reoptimization

Reoptimization, or the use of a starting solution which is "close" to optimal solution, is a standard strategy for accelerating an optimization algorithm. Modern samplers receive starting solution as arguments, even for mixed integer linear programming problems. Although this is an attractive feature as, for instance, it suggests we can use scalable yet non-exact solvers (Chernhozukhov and Hansen, 2006, Kaido and Wüthrich, 2018), our experience has been that starting solutions need to be extremely close to the optimal solution in order to improve the computational performance.

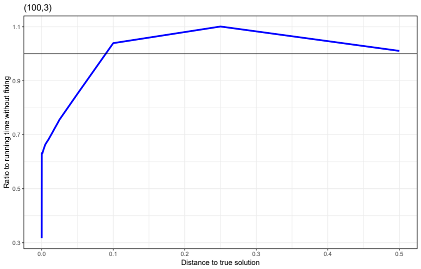


Table *Average time improvement for a starting solution given the ℓ-1 distance to the optimal solution. The data generating process is that of the simulation design of Chen and Lee (2017).*

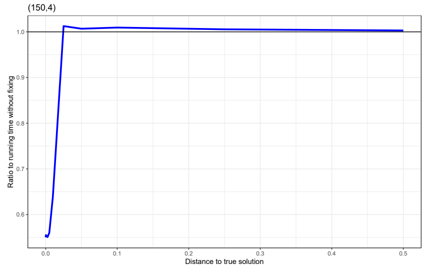


Table *Average time improvement for a starting solution given the ℓ-1 distance to the optimal solution. The data generating process is an extension of the simulation design of Chen and Lee (2017).*

A.3 Alternative Test

It is not possible to implement the tests described above as a parametric programming exercise directly using native R functions. This motivates the development of a different test whose implementation is made trivial by the use of native functions from standard quantile regression packages.

We propose an alternative statistic which requires stronger assumptions to produce a valid exact test, but from which confidence intervals may be computed seamlessly using existing quantile regression packages.

The test statistic is

$$T_n = n^{-1/2} \frac{\mathbf{D}^T \hat{b}}{\sqrt{\tau(1-\tau)\mathbf{D}^T Q \mathbf{D}}}$$

where , $Q = I - \mathbf{X}(\mathbf{X}^T \mathbf{\Psi} \mathbf{X})^{-1} \mathbf{\Psi} \mathbf{X}^T$, $\mathbf{\Psi} = \operatorname{diag}\left(f_\varepsilon\left(0\,|X_i, D_i, Z_i\right)\right)$ and $\hat{b} = \hat{a} - (1-\tau)\mathbf{1}_n$ with

$$\hat{a} = \arg\max \left\{ \left(\mathbf{Y} - \mathbf{D}\beta_D^0\right)^T a \ : \ \mathbf{X}^T a = (1-\tau)\mathbf{X}^T \mathbf{1}_n, \ \mathbf{Z}^T a = (1-\tau)\mathbf{Z}^T \mathbf{1}_n, \ a \in [0,1]^n \right\}.$$

Asymptotic distribution theory akin to that of Subsection 5.3 obtains for these statistics.

Corollary 3 (Lindeberg-Feller CLT, *Bai, Pouliot, and Shaikh, 2018*) *Suppose that*

(a) There exists $\varepsilon > 0$ such that $E\left[\|X\|^2\right] < \infty$.

(b) $E[f_U(0|X_i, Z_i)\left(X_i^T, Z_i^T\right)^T \left(X_i^T, Z_i^T\right)]$ is positive definite.

(c) For all x, z, d, $f_U(\cdot|x, z, d)$ exists and is bounded. In addition, there exists $\delta > 0$

and $C > 0$ *such that*

$$\sup_{|u|<\delta} \sup_{(x,z,d)\in\mathbb{R}^{p_X+2}} \left| \frac{f_U(u|x,z,d) - f_U(0|x,z,d)}{u} \right| \le C.$$

(d) $0 < E[DD^T] < \infty$.
Then

$$T_n \xrightarrow{d} N(0,\sigma^2),$$

where

$$\sigma^2 = \tau(1-\tau)E\left[\tilde{D}_i\tilde{D}_i^T\right],$$

where $\tilde{D}_i = D_i - E\left[f_\varepsilon\left(0\left|X_i, D_i, Z_i\right.\right) D_i X_i'\right] E\left[f_\varepsilon\left(0\left|X_i, D_i, Z_i\right.\right) X_i X_i'\right]^{-1} X_i$.

Under the homoskedasticity assumption $f_\varepsilon\left(0\left|X_i, D_i, Z_i\right.\right) = f_\varepsilon(0)$, we again obtain that $\tilde{D}_i = D_i - E\left[D_i X_i'\right] E\left[X_i X_i'\right]^{-1} X_i$ and one can produce a plug-in estimate for the asymptotic variance without having to cary out nonparametric density estimation and pick a bandwidth parameter.

The computations involved in regression rankscore inference using T_n can be presented as procedurally equivalent to carried out in the standard linear quantile regression case. Consequently, they can be computed using the quantreg package "as if" computing standard confidence regions. To obtain a confidence interval for the j^{th} endogenous variable, one inverts the null against alternative $H_{1,j}$ as follows:

1. Generate the variable $\tilde{Y}$: Compute $\tilde{Y}_i := Y_i - D_i^T\hat{\beta}_{D,IQR}$, $i = 1, ..., n$,
2. Regress $\tilde{Y}$ on X, Φ, D_j,
3. Extract the confidence interval for D_j from the standard regression output,
4. The regression output will center the interval at 0, shit if by $\hat{\beta}_{D,IQR,j}$ to obtain a confidence interval for $\beta_{D,IQR,j}$, $j = 1, .., p_D$.

The permutation test using statistic T_n or $T_{\perp,n}$ is exact under strong assumptions. [as detailed above, only robust for $T_{\perp,n}$...reorganize order of result]

Corollary 4 *Suppose Assumption 1 holds, and data is generated according to the linear model*

$$Y = Q(U,D,X),\ q(\tau,d,x) = d^T\beta_D(\tau) + x^T\beta_X(\tau). \tag{45}$$

Suppose Φ *is independent of* D, X *and* U. *Let* $\phi_n = \phi(L_n)$ *if* $p_Z = 1$ *and* $\phi_n = \phi(Q_n)$ *if* $p_Z > 1$, *where* $\hat{a}$ *is obtained as defined for the alternative test. Then,*

$$E_P[\phi_n] = \alpha,$$

for $P \in \mathbf{P}_0$*, the subset of* $\mathbf{P}$ *satisfying* H_0*.*

Corollary 5 states that a sufficient condition for exact validity is that D be independent of the covariates and exogenous. Although large sample validity obtains under more plausible assumptions in the strong identification case and the test is easy to implement, it has poor permutation test interpretation.