cemmap

centre for microdata methods and practice

Nonparametric instrumental variable estimation

Denis Chetverikov
Dongwoo Kim
Daniel Wilhelm

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP47/17

ESRC ECONOMIC & SOCIAL RESEARCH COUNCIL

An ESRC Research Centre

Nonparametric Instrumental Variable Estimation

Denis Chetverikov
Department of Economics
UCLA
Los Angeles, CA
chetverikov@econ.ucla.edu

Dongwoo Kim
Department of Economics
UCL
London, UK
dongwoo.kim.13@ucl.ac.uk

Daniel Wilhelm
Department of Economics
UCL
London, UK
d.wilhelm@ucl.ac.uk

Abstract. This paper introduces *Stata* commands [R] **npivreg** and [R] **npivregcv**, which implement nonparametric instrumental variable (NPIV) estimation methods without and with a cross-validated choice of tuning parameters, respectively. Both commands are able to impose monotonicity of the estimated function. The use of such a shape restriction may significantly improve the performance of the NPIV estimator (Chetverikov and Wilhelm 2017). This is because the ill-posedness of the NPIV estimation problem leads to unconstrained estimators that suffer from particularly poor statistical properties such as very high variance. The constrained estimator that imposes the monotonicity, on the other hand, significantly reduces variance by removing oscillations of the estimator that is nonmonotone. We provide a small Monte Carlo experiment to study the estimators' finite sample properties and an application to the estimation of gasoline demand functions.

Keywords: st0001, nonparametric instrumental variable estimation, shape restrictions, monotonicity, endogeneity, regression

1 Introduction

Instrumental variable methods are commonly used in economics to achieve identification and consistent estimation of models with endogeneity. Since economic theory does not provide any guidance for how to choose parameterizations of functions of interest, e.g. demand or production functions, it is desirable to avoid imposing such parameterizations when possible. Instead, the nonparametric instrumental variable (NPIV) model does not assume the function of interest is known up to a finite-dimensional parameter:

$$Y = g(X) + \varepsilon, \quad E[\varepsilon|W] = 0,$$

where Y is a scalar dependent variable, X a scalar endogenous explanatory variable, and W an instrumental variable (IV). In practice, however, a researcher often wants to include additional exogeneous covariates Z. To avoid the curse of dimensionality in nonparametric estimation, we do not allow the function g to arbitrarily vary with the covariates, but rather assume that they enter the model in an additively separable fashion. We will therefore study the following more general model:

$$Y = g(X) + \gamma' Z + \varepsilon, \quad E[\varepsilon|W, Z] = 0 \tag{1}$$

st0001

We are interested in the estimation of the function g based on a random sample $\{(Y_i, X_i, W_i, Z_i)\}_{i=1}^n$ without imposing any parametric functional form assumptions on g. Unfortunately, the nonparametric estimation of g is a very hard statistical problem that requires the solution of a so-called ill-posed inverse problem. These types of estimation problems are well-known to lead to estimators that are very poorly behaved in the sense that they can be extremely variable in finite samples and may converge to the true function g only at a very slow rate. The variance of the NPIV estimator is orders of magnitude larger than that of standard nonparametric regression estimators (based on exogeneous regressors) and, therefore, also much larger than the variance of parametric estimators. More details on the definition and properties of NPIV estimators can be found in Newey and Powell (2003), Hall and Horowitz (2005), Blundell et al. (2007), and Darolles et al. (2011), among others, and in the survey papers Horowitz (2011, 2014).

In many economic applications, economic theory implies that the function of interest, g, should be monotone, e.g. demand, production, or cost functions. Chetverikov and Wilhelm (2017) show that constraining the NPIV estimator to satisfy this monotonicity constraint may significantly improve the performance of the estimator in finite samples and, therefore, makes this constraint version of the NPIV estimator more attractive for applied work. To show that this improvement can be huge, we reproduce in Figure 1 a graph from that paper. It shows the square-root of the mean integrated square error ("MISE") of the unconstrained and the constrained NPIV estimators as functions of the sample size. The different panels show this function for different choices K of the number of terms in the series estimator. For $K = 4$ or $K = 5$ and a small sample size around $n = 100$, for example, the constrained estimator has a MISE that is only 20% of that of the unconstrained estimator and, in that sense, is therefore about 5 times more precise. The improvement in MISE from imposing the constraint is visible up to large sample sizes and then eventually disappears.

In this paper, we describe how to use the new *Stata* command [R] **npivreg** which implements both the constrained and unconstrained NPIV estimators for a user-chosen number of series terms in the estimator. A second command [R] **npivregcv**, provides the same estimators, but uses a cross-validation criterion to choose the number of series terms in a data-driven fashion.

1.1 The NPIV Estimator Without Exogenous Covariates

We first introduce the NPIV estimator in the absence of additional exogeneous covariates Z. It is a series estimator that takes the form of the standard two-stage least squares estimator for linear models, except that the nonparametric version considered here does not regress Y on X using W as instrument, but instead regresses Y on a set of transformations of X using a set of transformations of W as instruments.

Let (Y_i, X_i, W_i), $i = 1, \dots, n$, be an i.i.d. sample from the distribution of (Y, X, W). To define our estimator, we first introduce some notation. Let $\{p_k(x), k \geq 1\}$ and $\{q_k(w), k \geq 1\}$ be two orthonormal bases in $L^2[0,1]$. For $K = K_n \geq 1$ and $J = J_n \geq K_n$, let $p(x) := (p_1(x), \dots, p_K(x))'$ and $q(w) := (q_1(w), \dots, q_J(w))'$ be vectors

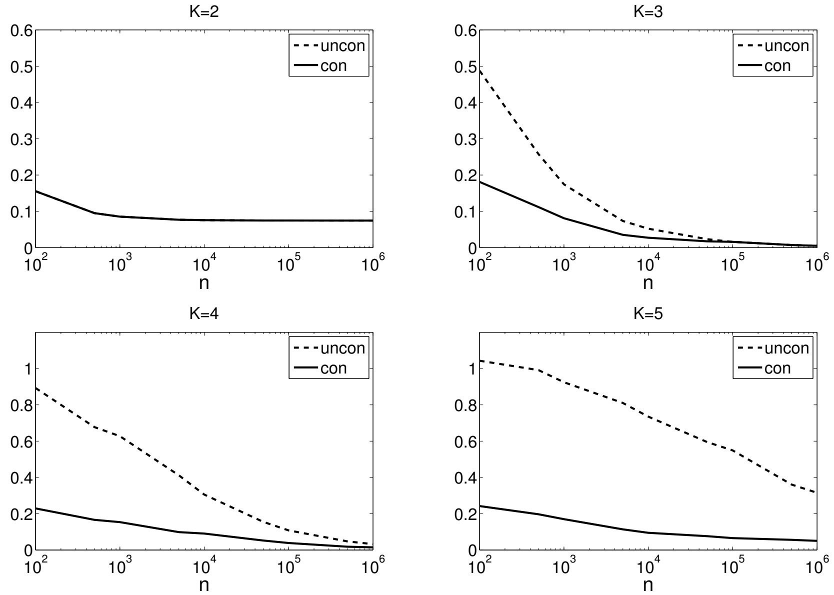

Figure 1: an example demonstrating performance gains from imposing the monotonicity constraint. In this example, $g(x) = x^2 + 0.2x$, $W = \Phi(\zeta)$, $X = \Phi(\rho\zeta + \sqrt{1-\rho^2}\epsilon)$, $\varepsilon = \sigma(\eta\epsilon + \sqrt{1-\eta^2}\nu)$, where (ζ, ϵ, ν) is a triple of independent $N(0,1)$ random variables, $\rho = 0.3$, $\eta = 0.3$, $\sigma = 0.5$, and $\Phi(\cdot)$ is the cdf of the $N(0,1)$ distribution. The four panels of the figure show the square root of the MISE of the constrained (con) and the unconstrained (uncon) series estimators defined in Section 1.1 as a function of the sample size n depending on the number of series functions used, K. We use the series estimators based on quadratic regression splines. The figure shows that the constrained estimator substantially outperforms the unconstrained one as long as $K \geq 3$ even in large samples.

of basis functions. Define $\mathbf{P} := (p(X_1), \dots, p(X_n))'$, $\mathbf{Q} := (q(W_1), \dots, q(W_n))'$, and $\mathbf{Y} := (Y_1, \dots, Y_n)'$.

We define two estimators of g. The first one is the *unconstrained estimator* $\widehat{g}^u(x) := p(x)'\widehat{\beta}^u$ which is a linear combination of basis functions with the coefficients estimated by the two-stage least squares optimization problem:

$$\widehat{\beta}^u := \operatorname{argmin}_{b \in \mathbb{R}^K} (\mathbf{Y} - \mathbf{P}b)'\mathbf{Q}(\mathbf{Q}'\mathbf{Q})^{-1}\mathbf{Q}'(\mathbf{Y} - \mathbf{P}b). \tag{2}$$

This estimator is similar to the one defined in Horowitz (2012) and is a special case of the estimator considered in Blundell et al. (2007). The second one is the *constrained estimator* $\widehat{g}^c(x) := p(x)'\widehat{\beta}^c$ which is a linear combination of basis functions with the coefficients estimated by the two-stage least squares optimization problem *subject to* *the constraint that the resulting estimator is nondecreasing*:

$$\widehat{\beta}^c := \text{argmin}_{b \in \mathbb{R}^K:\, p(\cdot)'b \text{ is nondecreasing}} (\mathbf{Y} - \mathbf{P}b)'\mathbf{Q}(\mathbf{Q}'\mathbf{Q})^{-1}\mathbf{Q}'(\mathbf{Y} - \mathbf{P}b). \tag{3}$$

Figure 1 shows that the constrained estimator $\widehat{g}^c(x)$ may substantially outperform the unconstrained estimator $\widehat{g}^u(x)$.

1.2 The NPIV Estimator With Exogenous Covariates

We now introduce two procedures for accommodating the presence of L additional, exogenous covariates Z in the NPIV estimation of g. The first procedure gives *one-step estimators* that are identical to those in (2) and (3) except that the bases $\mathbf{P}$ and $\mathbf{Q}$ are replaced by the enlarged bases

$$\widetilde{\mathbf{P}} := [\mathbf{P}, \mathbf{Z}], \quad \widetilde{\mathbf{Q}} := [\mathbf{Q}, \mathbf{Q} \times \mathbf{Z}]$$

where $\mathbf{Z} := (Z_1, \cdots, Z_n)'$ and $\mathbf{Q} \times \mathbf{Z}$ is the tensor product of the columns of matrices $\mathbf{Q}$ and $\mathbf{Z}$, i.e. the matrix $\mathbf{Q} \times \mathbf{Z}$ is such that its i-th row consists of all products of the form $q_j(W_i)Z_{i,l}$, $j = 1, \ldots, J$ and $l = 1, \ldots, L$. The second procedure gives *two-step estimators*, which are defined as follows. First, compute the constrained or unconstrained one-step estimator as described above. Let $\widehat{\gamma}$ denote the estimator of γ from this first step. Then, remove the covariates from the outcome by defining $\widetilde{Y} := Y - \widehat{\gamma}'Z$ and estimate g by NPIV estimation of $\widetilde{Y}$ on X only, i.e. using (2) or (3). We find that two-step estimators may outperform one-step estimators if the dimension of Z is large.

2 The npivreg and npivregcv Commands

The commands [R] **npivreg** and [R] **npivregcv** compute the fitted values of the NPIV estimator of g. The former requires the user to specify the number of series terms used in each of the two bases $p(x)$ and $q(w)$. The latter constrains the two tuning parameters to be equal and then uses cross-validation to choose them in a data-driven fashion.

The commands require two *Stata* packages, '`bspline`' and '`polyspline`'. These can be installed by simply typing '`ssc install bspline`' and '`ssc install polyspline`.'

When the user does not specify any options, [R] **npivreg** computes the estimator in (2) or the corresponding one-step estimator if there are additional, exogenous covariates using B-spline bases of degrees 2 and 3, with 2 and 3 knots, for $p(x)$ and $q(w)$, respectively. These default choices and further parameters can be modified through the options described below.

[R] **npivregcv** is built upon [R] **npivreg**. Its default choice of bases are also B-splines of degrees 2 and 3 for $p(x)$ and $q(w)$, respectively. It constrains the number of knots to be the same for both bases and then chooses the number of knots by minimizing the cross-validation criterion.

In the absence of monotonicity constraints on g, the commands avoid solving the optimization problem in (2) by computing the closed-form two-stage least squares solution of the problem. When the one of the options `decreasing` or `increasing` is specified, the commands compute the solution to the constrained optimization problem in (3) to ensure that the resulting estimator is decreasing or increasing, respectively. This optimization problem is implemented using a constrained optimization routine in *Mata* which typically requires significantly more computation time than the closed-form estimator in the absence of the monotonicity restrictions. As discussed above, however, the gains in precision of the estimator when imposing the constraints may be substantial.

2.1 Syntax

The syntax of the commands [R] **npivreg** and [R] **npivregcv** is similar:

`npivreg` *depvar expvar inst exovar* [*if*] [*in*] [, `power_exp(#)` `power_inst(#)` `num_exp(#)` `num_inst(#)` `pctile(#)` `polynomial` `increasing` `decreasing`]

`npivregcv` *depvar expvar inst exovar* [*if*] [*in*] [, `power_exp(#)` `power_inst(#)` `maxknot(#)` `pctile(#)` `polynomial` `increasing` `decreasing`]

The only difference between the two is that [R] **npivreg** possesses two additional options for the specification of the number of knots. The four required arguments of the commands are *depvar* (the outcome variable Y), *expvar* (the endogenous regressor X), *inst* (the instrumental variable Z), and *exovar* (a list of exogenous covariates W).

2.2 Options

We now describe the options of the two commands. If options are left unspecified, the commands run on the default settings.

`power_exp(integer)` is a positive integer for the degree of the spline basis for the endogenous regressor. The default is 2.

`power_inst(integer)` is a positive integer for the degree of the spline basis for the instrument. This number needs to be equal to or larger than `power_exp`. The default is 3.

`num_exp(integer)` is a positive integer greater than 1 for the number of knots of the spline basis for the endogenous regressor. The default is 2. The user need not specify this if `polynomial` is used.

`num_inst(integer)` is a positive integer greater than 1 for the number of knots of the spline basis for the instrument. The default is 3. The user need not specify this if `polynomial` is used.

`maxknot(integer)` is a positive integer for maximum number of knots to be considered in the cross-validation procedure. With a sample size N and the option `maxknot(k)` the cross-validation procedure evaluates the performance of the NPIV estimator with numbers of knots from 3 to $\max(N^{1/5}, k)$ and executes the NPIV regression with the optimal number of knots from that range. If the option `polynomial` is used, then `maxknot` specifies the maximum power of the polynomial to be considered. 5 is the default.

`pctile(integer)` specifies the domain of the endogenous regressor over which the NPIV estimator of g is to be computed. This needs to be a positive integer smaller than 50. For a given value k, the NPIV estimator is computed at fine grid points within the k-th and the $(100 - k)$-th percentiles of the empirical distribution of X. The default is 5.

`polynomial` specifies the type of both bases $p(x)$ and $q(w)$ to be power polynomials. Choices of numbers of knots are ignored under this option. Shape restrictions cannot be imposed for this basis and an error message is generated if this option is used together with `decreasing` or `increasing`.

`increasing` imposes that the NPIV estimator is an increasing function of the endogenous regressor. If this option is specified, the basis $p(x)$ is forced to be quadratic B-spline and the option `power_exp(integer)` is not used. An error occurs when this option is used together with one of the options `decreasing` or `polynomial`. The basis $q(w)$ for the instrument is also restricted to be a B-spline, but the power and number of knots can be freely chosen through `power_inst(integer)` and `num_inst(integer)`.

`decreasing` imposes that the NPIV estimator is a decreasing function of the endogenous regressor. The same restrictions as for `increasing` apply to this option.

2.3 Saved results

The commands `npivreg` and `npivregcv` each generate an output variable in the *Stata* data space, called *npest*. *npest* contains the fitted values of the NPIV estimator of g over the range indicated by `pctile(integer)`. Both commands are of the e-class. The following results are stored in `e()`:

Scalars

`e(N)`	number of observations	`e(powerexp)`	power of basis for X
`e(powerinst)`	power of basis for W	`e(numexp)`	# of knots of basis for X
`e(numinst)`	# of knots of basis for W	`e(pct)`	percentile
`e(xmin)`	min of X in domain	`e(xmax)`	max of X in domain
`e(zmin)`	min of Z in domain	`e(zmax)`	max of Z in domain
`e(gmin)`	min of grid on which g evaluated	`e(gmax)`	max of grid on which g evaluated
`e(optknot)`	optimal number of knots	`e(maxknot)`	maximum knots to be evaluated

Macros

`e(cmd)`	name of command		
`e(depvar)`	name of Y	`e(expvar)`	name of X
`e(inst)`	name of Z	`e(exovar)`	list of exogenous regressors
`e(basis)`	type of spline basis	`e(title)`	nonparametric IV regression
`e(shape)`	type of shape restriction		

Matrices

`e(b)`	coefficient vector of basis

In the presence of additional, exogenous covariates Z of dimension k, the estimates of γ correspond to the last k elements of the coefficient vector in `e(b)`.

3 Example: Estimating Gasoline Demand Functions

In this section, we illustrate the use of the NPIV estimation commands by applying them to the estimation of gasoline demand functions on the household level as in Blundell et al. (2012, 2016); Chetverikov and Wilhelm (2017). We consider the partially linear specification in (1) where Y is annual log-gasoline consumption (`log_q`), X denotes log-price of gasoline (`log_p`), and the instrument W is distance to major oil platform (`distance_oil1000`). We include additional controls Z such as household characteristics, urbanity, population density, and availability of public transit, among others. The controls are listed in the local macro '`exolist`'. The data set is from the 2001 National Household Travel Survey and the sample size is 4,812. See Blundell et al. (2012) for more details on the data set, sample selection and definition of the variables. Price and income elasticities are constant in a fully linear model, but it is more likely that those elasticities vary with price and income levels. By using the nonparametric function g, we allow for non-constant price elasticity. We focus on three subsamples in which household income levels are within ± 0.5 (in log) of \$72,500 (high), \$57,500 (middle) and \$42,500 (low).

Figures 2 and 3 show the unconstrained and constrained (with a decreasing shape) NPIV estimates of g. In both estimation procedures, a quadratic B-spline basis with 3 knots is used for $p(x)$ and a cubic B-spline with 10 knots for $q(w)$. The NPIV demand function is estimated on price levels from 5th to 95th percentiles of the given data.

We now describe how to produce these results using the command `npivreg` for the middle income group. First, consider estimation of the one-step unconstrained NPIV estimator. The following code defines the parameters of the B-spline basis, drops observations outside the middle income group, defines the list of additional exogenous covariates, then executes the npivreg command, and finally stores the estimated demand function in the variable `one_step`:

```
.  local powerx = 2
.  local powerz = 3
.  local numx = 3
.  local numz = 10
.  local p    = 5
.  local income = 57500
.  local inclevel "Middle"

.  drop in 4813/5254
.  drop if log_y < log(`income´) - 0.5
.  drop if log_y > log(`income´) + 0.5

.  local exolist log_hhsize log_driver log_hhr_age total_wrkr /*
>  */ publictransit_d cl5_smtown_d cl5_suburban_d cl5_secondcity_d cl5_urban_d /*
>  */ popdensity_d2 popdensity_d3 popdensity_d4 popdensity_d5 popdensity_d6 /*
>  */ state_fips region popdensity_d7 popdensity_d8

.  npivreg log_q log_p distance_oil1000 `exolist´, /*
>  */ power_exp(`powerx´) power_inst(`powerz´) num_exp(`numx´) num_inst(`numz´) pctile(`p´)
Domain over which the estimator is computed: .05-quantile of X to .95-quantile of X
(189 observations deleted)
Basis for X: B-spline of order 2
Basis for Z: B-spline of order 3
Number of equally spaced knots for X: 10
Number of equally spaced knots for Z: 5
no shape restriction

.  generate one_step = npest
```

For the estimation of the two-step unconstrained NPIV estimator, we need to subtract the index of covariates, $\widehat{\gamma}'W$, from the outcome. The following code estimates the index coefficients $\widehat{\gamma}$ and stores the difference between the outcome and the index of covariates in `Y_tilde`.

```
.  local exovar = e(exovar)

.  mata
----------------------- mata (type end to exit) -----------------------
:  bw = st_matrix("e(b)")´

:  W  = st_data(., "`exovar´", 0)

:  nw = cols(W)

:  nb = rows(bw)

:  Ey = W*bw[(nb-nw+1)..nb]

:  end
-----------------------------------------------------------------------

.  getmata Ey, force

.  quietly generate Y_tilde = log_q - Ey
```

The two-step unconstrained NPIV estimator is then computed as the unconstrained NPIV estimator without additional covariates, using `Y_tilde` as the new outcome:

```
.  npivreg Y_tilde log_p distance_oil1000, /*
>  */ power_exp(`powerx´) power_inst(`powerz´) num_exp(`numx´) num_inst(`numz´) pctile(`p´)
```

```
. mata
----------------------- mata (type end to exit) -----------------------
: npest    = st_data(., "npest", 0)

: one_step = st_data(., "one_step", 0)

: onestep = one_step :+ bw[nb-nw+1]*log(`income´)

: twostep = npest :+ bw[nb-nw+1]*log(`income´)

: end
------------------------------------------------------------------------
```

Finally, we recover the one-step and two-step estimates from mata and plot them together in the same graph:

```
. getmata onestep, force
. getmata twostep, force

. quietly line twostep grid || line onestep grid, title("`inclevel´ income group")
```

The resulting graph is the second one in Figure 2. The other two graphs for the lower and upper income group are generated in a similar fashion. Similarly, the constrained estimators are constructed similarly, making use of the option `decreasing` of the `npivreg` command.

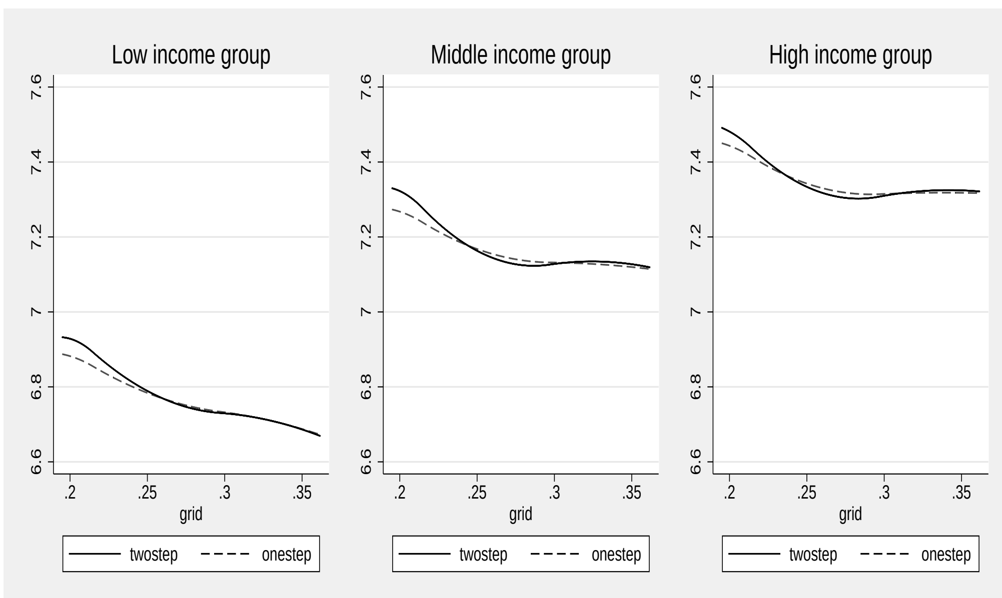

Figure 2: Unconstrained NPIV estimates of g in the three income groups using the one-step and two-step estimators to accommodate additional covariates.

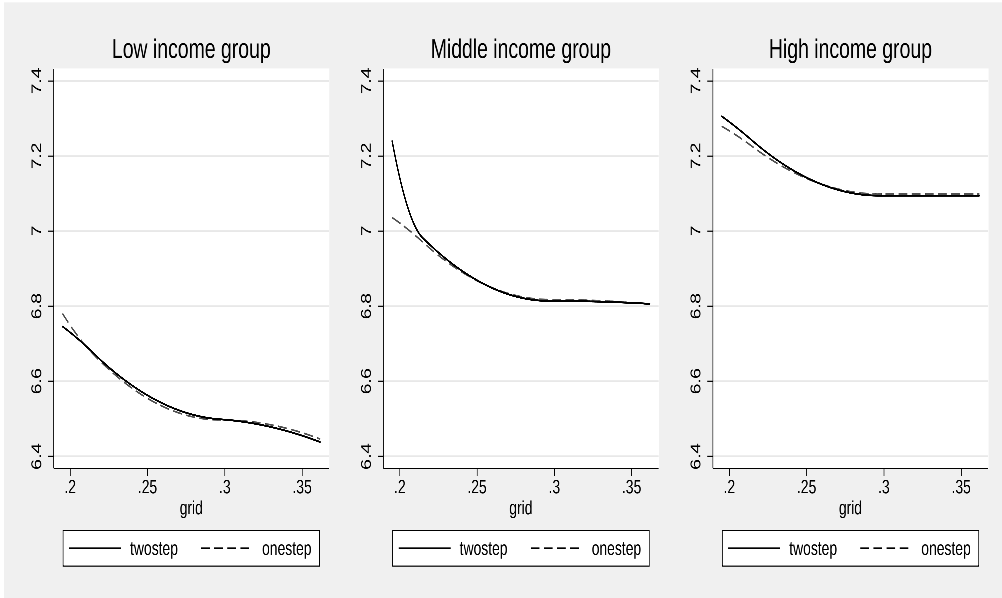

Figure 3: Constrained NPIV estimates of g in the three income groups using the one-step and two-step estimators to accommodate additional covariates.

4 Monte Carlo Simulation

In this section, we conduct a small simulation experiment to study the finite sample properties of the NPIV estimators discussed in this paper and to show that imposing monotonicity constraints can substantially improve the performance of the estimators.

Consider the following data generating process:

$$\begin{bmatrix}\varepsilon\\ \eta\end{bmatrix} \sim N\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1 & 0.5\\ 0.5 & 1\end{bmatrix}\right),$$

$$W \sim N(0,1), \quad X = 2W + \eta,$$

$$Y = g(X) + \varepsilon \quad \text{where} \quad g(X) = \frac{\exp(X/2)}{1+\exp(X/2)}$$

Y, X and W denote the outcome, the endogenous regressor, and the instrument respectively. η is the endogenous component of X. W is independent of ε and η by construction. We generate $1,000$ Monte Carlo samples of size 100 and of size 800.

First, we compute the unconstrained NPIV estimator:

```
. npivreg Y X W, power_exp(2) power_inst(3) num_exp(3) num_inst(3) pctile(1)
```

Second, we compute the constrained NPIV estimate that imposes that the estimator is increasing:

```
. npivreg Y X W, power_exp(2) power_inst(3) num_exp(3) num_inst(3) pctile(1) increasing
```

Figures 4 and 5 show the true g function and the NPIV estimate averaged across the $1,000$ samples. Red lines show bands that indicate the variability of the estimator. The lower and upper band were generated by subtracting and adding 2-times the empirical standard deviations from the average estimate at each grid point.

For both sample sizes, the two estimators incur only a small bias, but the constrained NPIV estimator is much less variable than the unconstrained one. The difference in variability decreases as the sample size increases. Eventually, as the sample size becomes large enough the constrained and unconstrained estimators will be equal to each other, although this may happen only for extremely large samples. These findings are consistent with the Monte Carlo simulation results in Chetverikov et al. (2017).

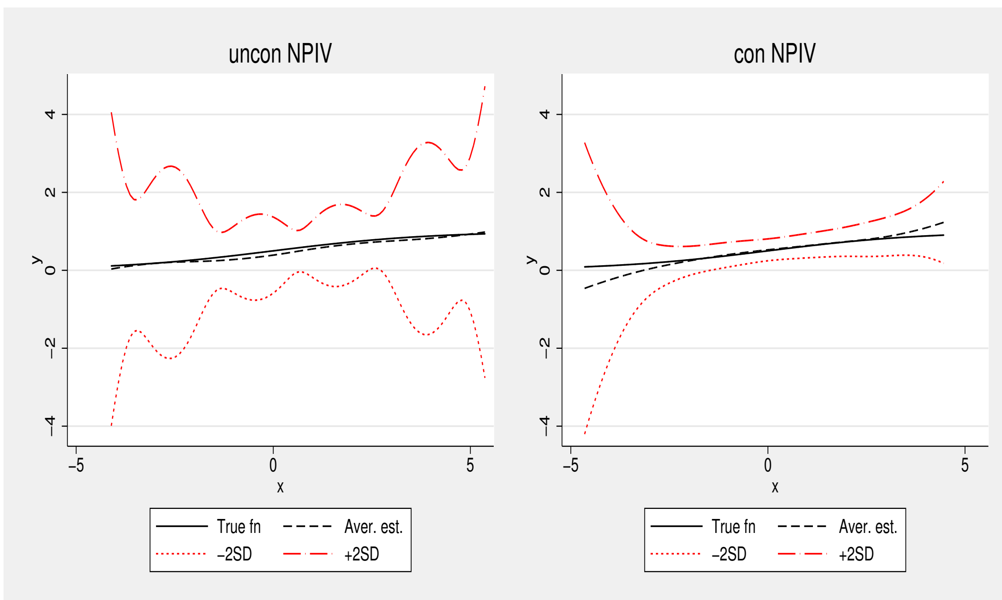

Figure 4: Monte Carlo simulation results on small samples

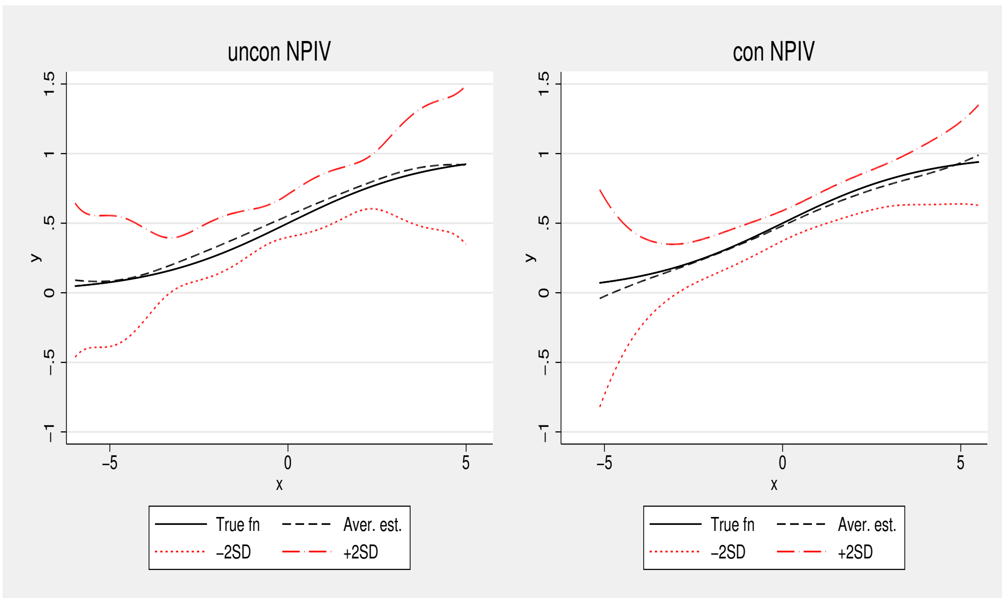

Figure 5: Monte Carlo simulation results on large samples

5 References

Blundell, R., X. Chen, and D. Kristensen. 2007. Semi-Nonparametric IV Estimation of Shape-Invariant Engel Curves. *Econometrica* 75(6): 1613–1669.

Blundell, R., J. Horowitz, and M. Parey. 2016. Nonparametric Estimation of a Non-separable Demand Function under the Slutsky Inequality Restriction. *Review of Economics and Statistics* forthcoming.

Blundell, R., J. L. Horowitz, and M. Parey. 2012. Measuring the price responsiveness of gasoline demand: Economic shape restrictions and nonparametric demand estimation. *Quantitative Economics* 3(1): 29–51.

Chetverikov, D., D. Kim, and D. Wilhelm. 2017. An Adaptive Test of Stochastic Monotonicity. Technical report.

Chetverikov, D., and D. Wilhelm. 2017. Nonparametric Instrumental Variable Estimation Under Monotonicity. *Econometrica* 85(4): 1303–1320.

Darolles, S., Y. Fan, J. P. Florens, and E. Renault. 2011. Nonparametric Instrumental Regression. *Econometrica* 79(5): 1541–1565.

Hall, P., and J. L. Horowitz. 2005. Nonparametric Methods for Inference in the Presence of Instrumental Variables. *The Annals of Statistics* 33(6): 2904–2929.

Horowitz, J. L. 2011. Applied Nonparametric Instrumental Variables Estimation. *Econometrica* 79(2): 347–394.

———. 2012. Specification Testing in Nonparametric Instrumental Variable Estimation. *Journal of Econometrics* 167(2): 383–396.

———. 2014. Ill-Posed Inverse Problems in Economics. *Annual Review of Economics* 6: 21–51.

Newey, W. K., and J. L. Powell. 2003. Instrumental Variable Estimation of Nonparametric Models. *Econometrica* 71(5): 1565–1578.

About the authors

Denis Chetverikov is an Assistant Professor in the Department of Economics at University of California at Los Angeles.

Dongwoo Kim is a PhD candidate in the Department of Economics at University College London.

Daniel Wihelm is a Lecturer in Economics at University College London, a staff member of the ESRC Centre for Microdata Methods and Practice (CeMMAP), a Research Fellow at the Institute for Fiscal Studies, and a Research Fellow at the Centre for Research and Analysis of Migration (CReAM).