cemmap

centre for microdata methods and practice

Instrumental variables estimation with flexible distributions

Christian Hansen
James B. McDonald
Whitney Newey

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP21/07

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

An ESRC Research Centre

Instrumental Variables Estimation with Flexible Distributions*

Christian Hansen
Graduate School of Business
University of Chicago

James B. McDonald
Brigham Young University
Department of Economics

Whitney K. Newey
Department of Economics
M.I.T.

August 20, 2007

Abstract

Instrumental variables are often associated with low estimator precision. This paper explores efficiency gains which might be achievable using moment conditions which are nonlinear in the disturbances and are based on flexible parametric families for error distributions. We show that these estimators can achieve the semiparametric efficiency bound when the true error distribution is a member of the parametric family. Monte Carlo simulations demonstrate low efficiency loss in the case of normal error distributions and potentially significant efficiency improvements in the case of thick-tailed and/or skewed error distributions.

* We appreciate outstanding research assistance provided by Brigham Frandsen, Samuel Dastrup, and Randall Lewis. We also thank two anonymous referees and an editor for constructive comments that have improved the paper. All remaining errors are, of course, ours. This work has been supported by NSF Grants SES-0136869 and SES-0617836 and funding from the William S. Fishman Faculty Research Fund and IBM Corporation Faculty Research Fund at the Graduate School of Business, the University of Chicago.

1. Introduction

Instrumental variables (IV) estimation is important in economics. A common finding is that the precision of IV estimators is low. This paper explores potential efficiency gains that might result from using moment conditions that are nonlinear in the disturbances. It is known that this approach can produce large efficiency gains in regression models. The hope is that such efficiency gains might also be present when models are estimated by IV. These gains could help in overcoming the low efficiency of IV estimators.

A simple approach to improving efficiency in IV estimation based on nonlinear functions of the residuals is to use flexible parametric families of disturbance distributions. This approach has proven useful in a variety of settings. For example, McDonald and Newey (1988) present a generalized t distribution which can be used to obtain partially adaptive estimators of regression parameters. McDonald and White (1993) use the generalized t and an exponential generalized beta distribution to show substantial efficiency gains can be obtained from partially adaptive estimators in applications characterized by skewed and/or thick tailed error distributions. Hansen, McDonald, and Theodossiou (2005) consider some additional partially adaptive regression estimators and find similar efficiency gains.

Here we follow an iterative approach to estimation with flexible distributions. We use residuals from a preliminary IV estimator to estimate the parameters of a density. We do this by quasi maximum likelihood on the residual distribution although other ways to estimate the parameters could be used. The product of the instrumental variables and the location score for the density, evaluated at the estimated distributional parameters, is then used to form moment conditions for nonlinear IV estimation. We give consistency and asymptotic normality results

for the estimator of the structural parameters. We also show that the asymptotic variance of the structural slope estimator does not depend on the estimator of the distributional parameters.

To help motivate the form of our estimator we derive the semiparametric efficiency bound for the structural slope estimators when the disturbance is independent of the instruments and the reduced form is unrestricted. This bound depends on the marginal distribution of the error and on the conditional expectation of the endogenous variable. When the reduced form for the endogenous right-hand side variables happens to be linear and additively separable in an independent disturbance, our nonlinear IV estimator achieves the semiparametric bound when the true distribution is included in the parametric family. Thus, the estimator has a "local" efficiency property, attaining the semiparametric bound in some cases.

To evaluate efficiency gains in practice we consider two empirical examples and carry out some Monte Carlo work. The empirical applications are taken from Card (1995) and Angrist and Krueger (1991). In the applications, we find that there may be moderate efficiency gains in estimation from using more flexible distributions. We also find evidence of potentially large efficiency gains in the Monte Carlo work.

Previous work on IV estimation with nonlinear functions of the residuals includes Newey (1990), Chernozhukov and Hansen (2005), and Honoré and Hu (2004). Newey (1990) considers efficiency in nonlinear simultaneous equations with disturbances independent of instruments, which specializes to the case considered here. Chernozhukov and Hansen (2005) consider IV estimation where the residual function corresponds to regression quantiles. Honoré and Hu (2004) also consider estimation based on residual ranks.

Section 2 of the paper introduces the model and estimators. The flexible distributions we consider are described in Section 3. Section 4 gives the asymptotic theory, including the

semiparametric variance bound. Section 5 reports results from the empirical applications with the results from the Monte Carlo simulations included in Section 6. Section 7 concludes.

2. The Model and Estimators

The model we consider is a regression model with a disturbance that is independent of instruments. This model takes the form

$$y_i = X_i'\beta_0 + \varepsilon_i \quad E[\varepsilon_i] = 0, \quad Z_i = Z(z_i),\ \varepsilon_i \text{ and } z_i \text{ independent.} \qquad (1)$$

where y_i is a left-hand side endogenous variable, X_i is a $p\times 1$ vector of right-hand side variables, β_0 is a $p\times 1$ vector of true parameter values, ε_i is a scalar disturbance, and Z_i is an $m\times 1$ vector of instrumental variables that is a function of variables z_i that are independent of the disturbance. We will assume throughout that the first element of X_i and of Z_i is 1, so that the mean zero restriction is just a normalization.

The nonlinear instrumental variables estimators (NLIV) we consider are based on a parametric family of pdf's. Let $f(\varepsilon,\gamma)$ denote a member of this family with parameter vector γ. In keeping with the normalization adopted above we will restrict the parameters so that the $f(\varepsilon,\gamma)$ has mean zero. Also, let

$$\rho(\varepsilon,\gamma) = \partial \ln f(\varepsilon,\gamma)/\partial\varepsilon.$$

If X_i were exogenous we could form an estimator of the parameters by maximizing $\sum_{i=1}^{n} \ln f(y_i - X_i'\beta,\tilde{\gamma})$, where $\tilde{\gamma}$ is a preliminary estimator of γ. This estimator has a first-order condition $0 = \sum_{i=1}^{n} X_i \rho(y_i - X_i'\beta,\tilde{\gamma})$. We generalize this estimator to the instrumental variables case by replacing X_i with Z_i outside ρ to form moment conditions. These moment conditions take the form

$$\hat{g}(\beta) = \sum_{i=1}^{n} Z_i \rho(y_i - X_i'\beta, \tilde{\gamma}).$$

The estimator is obtained by minimizing a quadratic form in $\hat{g}(\beta)$ where the weighting matrix is the usual one for NLIV. The estimator is given by

$$\hat{\beta} = \arg\min_{\beta \in B} \hat{g}(\beta)'\hat{Q}^{-1}\hat{g}(\beta), \quad \hat{Q} = \sum_{i=1}^{n} Z_i Z_i'.$$

The asymptotic variance of the slope parameters, the coefficients of the nonconstant elements of X_i, can be estimated in the usual way for NLIV. Let $\hat{\varepsilon}_i = y_i - X_i'\hat{\beta}$ and

$$\hat{\sigma}^2 = \sum_{i=1}^{n} \rho\left(\hat{\varepsilon}_i, \tilde{\gamma}\right)^2 / n, \quad \hat{G} = \sum_{i=1}^{n} Z_i X_i' \partial\rho\left(\hat{\varepsilon}_i, \tilde{\gamma}\right) / \partial\varepsilon.$$

Also, let S=[*0,I*] be the selection matrix that picks out the last *p-1* rows of β, where 0 is a (*p-1*)×1 vector of zeros and *I* is a *p-1* dimensional identity matrix. An estimator of the asymptotic variance of the slope parameter estimators $S\hat{\beta}$ is

$$\hat{V} = \hat{\sigma}^2 S(\hat{G}'\hat{Q}^{-1}\hat{G})^{-1} S'.$$

This variance estimator does not account for the presence of the preliminary estimator $\tilde{\gamma}$, but turns out to be consistent for the asymptotic variance of the slope parameter estimators under equation (1). In contrast, the asymptotic variance of the first component of $\hat{\beta}$ will depend on $\tilde{\gamma}$ in the usual way for two-step estimators. For simplicity and because interest often centers on slope coefficients we omit results on the asymptotic distribution of the first element of $\hat{\beta}$.

The NLIV estimator depends on a preliminary estimator $\tilde{\gamma}$ of γ. Two different approaches to estimation of γ are a quasi-maximum likelihood estimation (QMLE) and an approach that minimizes a scalar that affects the asymptotic distribution of the slope coefficients. Both are based on residuals $\tilde{\varepsilon}_i = y_i - X_i'\tilde{\beta}$ where $\tilde{\beta}$ is a preliminary estimator, such as limited

information maximum likelihood (LIML) or two-stage least squares (2SLS). The QMLE is given by

$$\tilde{\gamma} = \arg\max_{\gamma} \sum_{i=1}^{n} \ln f(\tilde{\varepsilon}_i, \gamma).$$

This estimator will be consistent for the true value of γ when the density of ε_i has the form $f(\varepsilon, \gamma)$ for some γ. The second approach is to minimize an estimator of a scalar that can affect the asymptotic variance. This estimator takes the form

$$\tilde{\gamma} = \arg\min_{\gamma} \left\{ \sum_{i=1}^{n} \rho(\tilde{\varepsilon}_i, \gamma)^2 \Big/ \left[\sum_{i=1}^{n} \partial\rho(\tilde{\varepsilon}_i, \gamma) / \partial\varepsilon \right]^2 \right\}.$$

When the reduced form for X_i is additive this estimator will minimize the asymptotic variance of $S\hat{\beta}$, as will be shown below. In general though this $\tilde{\gamma}$ need not minimize the asymptotic variance and so there will be no clear choice between the two estimators of γ in terms of asymptotic efficiency.

A final point that needs to be considered is how to select which parametric family to use for obtaining the NLIV estimates. There are a variety of approaches one might consider. For example, one could choose a particular parametric family, estimate the distributional parameters using the first-stage LIML or 2SLS residuals, and then test that the fitted distribution is consistent with the data using a modification of a conventional testing procedure, such as the information matrix test of White (1982) or Kolmogorov-Smirnov tests. The tests would need to be modified to account for the two-step nature of the procedure but would otherwise be quite standard. While this testing approach is intuitive and appealing on a number of dimensions, it suffers from the usual drawback that considering multiple candidate distributions raises concerns

about pretesting and related size and power considerations. It also fails to get directly at the question of interest which is the efficiency of the estimator of β.

A different approach which we pursue in this paper is to choose the parametric family based on a model selection procedure. Again, there are a variety of procedures that one may wish to consider, but for simplicity, we focus on one intuitive and rather simple approach. Specifically, we select the model that produces the smallest value of

$$H = \left\{ \sum_{i=1}^{n} \rho(\hat{\varepsilon}_i, \tilde{\gamma})^2 / \left[\sum_{i=1}^{n} \partial\rho(\hat{\varepsilon}_i, \tilde{\gamma}) / \partial\varepsilon \right]^2 \right\} + (k-2)\frac{\ln n}{n^2}$$

where $\tilde{\gamma}$ is the preliminary estimate of the distributional parameters which have dimension k obtained by QMLE, minimizing the first term of H, or some other method and $\hat{\varepsilon}_i$ is a residual, $\hat{\varepsilon}_i = y_i - X_i'\hat{\beta}$, for $\hat{\beta}$ a consistent estimator of β_0 such as the LIML, 2SLS, or NLIV estimator. The first term is a scalar quantity that is related to the asymptotic variance of the NLIV estimator, and the second term is a BIC-type penalty for the number of parameters used to fit the residual distribution. As noted above, the first component of H relates to the asymptotic variance of $S\hat{\beta}$ which will be minimized when the first component of H is minimized when the reduced form for X_i is additive. H is simple to compute and is directly related to the variance of the object of interest in a leading case and so seems like a natural object upon which to base model selection. Under the reduced form conditions and regularity conditions given in Section 5 of this paper, one could establish the properties of this procedure as in Andrews (1999). We end by noting that while this procedure is simple, it may not select the estimator that produces the smallest asymptotic variance when the reduced form conditions given above are not satisfied. We believe that it is still likely to select a model that captures much of the efficiency gain

available from non-Gaussian disturbances in more general settings though pursuing other approaches to estimation and model selection may be an interesting avenue for other research.

3. Distributions

Many distributions could be considered in the generalized IV estimation procedure outlined in the previous section. The use of such distributions as the normal or Laplace would not model distributions that are both thick-tailed and asymmetric, both of which are often observed with economic and financial data. The skewed generalized *t*, the exponential generalized beta of the second kind, and inverse hyperbolic sine distributions involve a small number of distributional parameters and permit modeling a wider range of data characteristics than the normal, Laplace, *t*, and many other common distributions. These distributions will be defined with basic properties and special cases summarized. We note that there are many other flexible families of distributions that could be considered. Examples include the stable distributions, the generalized hyperbolic distribution, and mixture distributions to name a few. We have chosen to focus on our particular set of distributions because they involve few distributional parameters and are relatively simple to implement while containing as special cases many of the common distributions employed in practice. Of course, there are a variety of reasons for which one may prefer to use a different parametric family, and the main results of the paper will continue to hold regardless of the family considered.

3.1 Skewed Generalized t distribution (SGT)

The skewed generalized t distribution (SGT) was obtained by Theodossiou (1998) and can be defined by

$$SGT\left(y;m,\lambda,\phi,p,q\right)=\frac{p}{\left[2\phi q^{1/p}B\left(1/p,q\right)\left(1+\frac{\left|y-m\right|^{p}}{\left(\left(1+\lambda sign\left(y-m\right)\right)^{p}q\phi^{p}\right)}\right)^{q+1/p}\right]}.$$

where *B(.,.)* is the beta function, m is the mode of y and the parameters *p* and *q* control the height and tails of the density. The parameter ϕ is a scale parameter and λ determines the degree of skewness with the area to the left of the mode equal to $\left(1-\lambda\right)/2$; thus, positive (negative) values for λ correspond to positive (negative) skewness. Setting λ = 0 in the SGT yields the generalized *t* (GT) of McDonald and Newey (1988). Similarly, setting p=2 yields the skewed t (ST) of Hansen (1994) which includes the student t distribution if $\lambda = 0$. The skewed t also includes the skewed Cauchy if pq =1. Standardized values for skewness and kurtosis in the ranges (-∞,∞) and (1.8,∞), respectively, can be modeled with the SGT. The SGT has all moments of order less than the degrees of freedom (*pq*).

Another important class of flexible density functions corresponds to a limiting case of the SGT. When the parameter q grows indefinitely large, we obtain the skewed generalized error distribution (SGED) defined by

$$SGED\left(y;m,\lambda,\phi,p\right)=\frac{pe^{-\left(\frac{\left|y-m\right|^{p}}{\left(\left(1+\lambda sign\left(y-m\right)\right)^{p}\phi^{p}\right)}\right)}}{\left[2\phi\,\Gamma\left(1/p\right)\right]}.$$

The parameter p in the SGED controls the height and tails of the density and λ controls the skewness. The SGED is symmetric for λ = 0 and positively (negatively) skewed for positive (negative) values of λ. The symmetric SGED is also known as the generalized power (Subbotin (1923)) distribution. The SGED can easily be seen to include the skewed ($\lambda \neq 0$) or symmetric ($\lambda = 0$) Laplace (SLaplace or Laplace respectively) when p = 1 and the skewed ($\lambda \neq 0$) or

symmetric ($\lambda = 0$) normal (SNormal or Normal respectively) when p = 2. The interrelationships between the SGT and many of its special cases can be visualized as in figure 1.

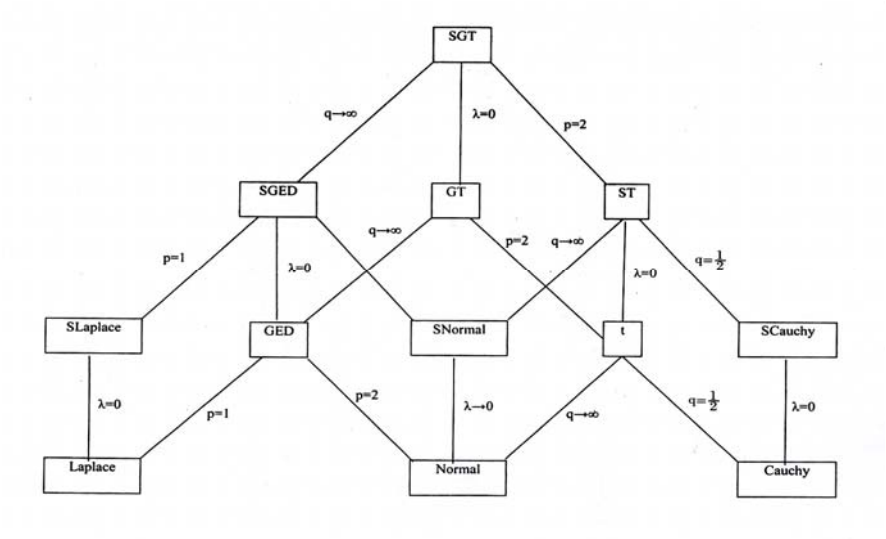

Figure 1. SGT distribution tree

3.2 Exponential generalized beta of the second kind (EGB2)

The four parameter EGB2 distribution is defined by the probability density function

$$EGB2\left(y;m,\phi,p,q\right)=\frac{e^{p(y-m)/\phi}}{\left[\phi\ \mathrm{B}\left(p,q\right)\left(1+e^{(y-m)/\phi}\right)^{p+q}\right]}$$

where the parameters ϕ, p, and q are assumed to be positive, cf. McDonald and Xu (1995). m and ϕ are respectively location and scale parameters. The parameters p and q are shape parameters. The EGB2 pdf is symmetric if and only if p and q are equal. The normal distribution is a limiting case of the EGB2 where the parameters p and q are equal and grow indefinitely large. Other special or limiting cases of the EGB2 include the Gumbel, Burr 2, generalized Gompertz, extreme value, and logistic distributions. Standardized values for kurtosis are limited to the range (3.0, 9.0), and the standardized skewness coefficient can assume values

in the range (-2.0, 2.0).

3.3 Inverse hyperbolic sine (IHS)

The hyperbolic sine pdf was proposed by Johnson (1949) and allows for modeling a wide range of skewness and kurtosis. The parameterization used here is slightly different than used by Johnson and is based on the transformation $y = a + b\sinh\left(\lambda + z/k\right) = a + bw$ where *sinh* is the hyperbolic sign, z is a standard normal, and *a, b, λ,* and *k* are scaling constants related respectively to the mean (μ), variance (σ^2), skewness, and kurtosis of the random variable *y*. The pdf of *y* can be written as

$$IHS\left(y;\mu,\sigma,k,\lambda\right) = \frac{ke^{-\left(\frac{k^2}{2}\left(\ln\left(u/\sigma+\sqrt{\theta^2+u^2/\sigma^2}\right)-\left(\lambda+\ln\theta\right)\right)^2\right)}}{\sqrt{2\pi\left(\theta^2+u^2/\sigma^2\right)\sigma^2}}$$

where $u = y - \mu + \delta\sigma$, $\theta = 1/\sigma_w$, $\delta = \mu_w/\sigma_w$, $\mu_w = .5\left(e^{\lambda} - e^{-\lambda}\right)e^{.5k^{-2}}$, and $\sigma_w = .5\left(e^{2\lambda+k^{-2}} + e^{-2\lambda+k^{-2}} + 2\right)^{.5}\left(e^{k^{-2}} - 1\right)^{.5}$; see Hansen, McDonald, Theodossiou (2005). Positive (negative) values of λ generate positive (negative) skewness, and zero corresponds to symmetry. Smaller values of k result in more leptokurtic distributions with the normal corresponding to the limiting case of $k \to \infty$ with $\lambda = 0$. The IHS allows skewness and kurtosis in the range (3,∞) and (-∞,∞), respectively.

4. Large Sample Properties

In this Section we give an account of the asymptotic theory of the estimator. To keep things relatively simple we restrict $\rho(\varepsilon,\gamma)$ to be smooth in γ, although the non-smooth case could be considered as in McDonald and Newey (1988). The first condition imposes the model of equation (1) and identification.

Assumption 1: a) Equation (1) is satisfied, $W_i = (y_i, X_i, Z_i), (i = 1,...,n)$ are i.i.d, $X_{i1} \equiv 1$, and $Z_{i1} \equiv 1$; b) there is γ^* such that $\widetilde{\gamma} = \gamma^* + O_p(1/\sqrt{n})$; c) there a unique solution α^* to $E[\rho(\varepsilon_i - \alpha, \gamma^*)] = 0$; d) there is at most one solution to $E[Z_i \rho(y_i - X_i'\beta, \gamma^*)] = 0$.

The next condition imposes smoothness and dominance conditions.

Assumption 2: B is compact, $\rho(\varepsilon, \gamma)$ is continuously differentiable in ε and γ, and there is a function *d(w)* such $E[d(W_i)]$ exists and for $\beta \in B$ and all γ in a neighborhood of γ*,

$$|\rho(y - x'\beta, \gamma)|^2 \le d(w), \quad \|Z\|^2 \le d(w),$$
$$\|ZX'\partial\rho(y - X'\beta, \gamma)/\partial\varepsilon\| \le d(w), \quad \|Z\partial\rho(y - X'\beta, \gamma)/\partial\gamma\| \le d(w)$$

The final condition imposes rank conditions for asymptotic normality.

Assumption 3: $Q = E[Z_i Z_i']$ is nonsingular, and $G = E[Z_i X_i' \partial\rho(\varepsilon_i - \alpha^*, \gamma^*)/\partial\varepsilon]$ has rank *p*.

It is worth noting that the conditions imposed in Assumptions 1-3 place our theoretical results in the conventional asymptotic framework. In particular, the full rank condition in Assumption 3 rules out weak identification, and we have assumed a fixed number of instruments and thus are not considering many instrument asymptotic sequences. While considering inference issues for the NLIV estimator under these conditions is an interesting question, we focus in this paper on the potential efficiency gains that may be achieved by considering NLIV. We note that due to the GMM formulation of the problem, the approaches to weak-identification robust inference of, for example, Stock and Wright (2000) and Kleibergen (2005) could readily be adopted. In many instrument settings, one could also consider the GMM approach of Newey and Windmeijer (2007).

It is also important to note that the model assumes independence between the structural errors and the instruments, ruling out heteroskedasticity. In principle, it would be simple to accommodate parametric forms of heteroskedasticity by suitably modifying the family of

distributions to allow its parameters to depend on the instruments and then suitably modifying the moment conditions. However, the NLIV estimator will likely be inconsistent as formulated in the presence of heteroskedasticity. As such, researchers may wish to test for the presence of heteroskedasticity in the LIML or 2SLS residuals obtained in the first stage estimation which could be done using any standard test for heteroskedasticity. A particular simple test that is available is a Hausman test of the difference between the NLIV and 2SLS or LIML estimates of the structural parameters. Under the assumption that the conditions given below such that NLIV attains the efficiency bound hold, this test can be performed simply by taking a quadratic form in the difference in estimated coefficients with the difference in estimated variances as the weighting matrix. We consider this in the empirical examples in Section 5.

To describe the asymptotic variance of the slope coefficients let $\sigma^2 = E[\rho(\varepsilon_i - \alpha^*, \gamma^*)^2]$. The asymptotic variance of $S\hat{\beta}$ will be

$$V = \sigma^2 S(G'Q^{-1}G)^{-1}S'.$$

The following result shows the consistency and asymptotic normality of the slope coefficient estimator $S\hat{\beta}$.

Theorem 1: If Assumptions 1 - 3 are satisfied then

$$\sqrt{n}S\left(\hat{\beta} - \beta_0\right) \xrightarrow{d} N(0,V), \quad \hat{V} \xrightarrow{p} V.$$

We now turn to the efficiency of the slope estimators. We motivated the estimator by analogy with the exogenous *X* case, but it is not clear *a priori* what the efficiency properties of such an estimator might be. In particular, the form of the estimator seems to use only information about the marginal distribution of ε, and one might wonder whether more information is available. We analyze efficiency in the semiparametric model where the only substantive assumption imposed is independence of z and ε. This is a "limited information" semiparametric

model, where no restrictions are placed on the conditional distribution of the endogenous regressors given the instruments z and the disturbance ε. This model also does not restrict the form of the distribution of ε or the other random variables.

We derive the efficiency bound without a full statement of regularity conditions to avoid much additional notation and clutter. This corresponds to a "formal" derivation, as is common in the semiparametric efficiency literature, e.g. see Newey (1990). To state the efficiency result let *x* denote the nonconstant elements of *X*, so that $X = (1, x')'$. Let $\rho_0(\varepsilon)$ denote the location score for ε, that is $\rho_0(\varepsilon) = \partial \ln f_0(\varepsilon)/\partial\varepsilon$ where $f_0(\varepsilon)$ is the marginal pdf of ε, let $\Pi(\varepsilon, z) = E[x|\varepsilon, z]$, and $\Pi_\varepsilon(\varepsilon, z) = \partial\Pi(\varepsilon, z)/\partial\varepsilon$. The following result is based on equation (23) of Newey (1990) and further calculations.

Theorem 2: *In the semiparametric model of equation (1) the semiparametric variance bound for Sβ is* $V^* = (E[s^* s^{*\prime}])^{-1}$ *where*

$$s^* = -\rho_0(\varepsilon)\{\Pi(\varepsilon, z) - E[\Pi(\varepsilon, z)|\varepsilon]\} - \Pi_\varepsilon(\varepsilon, z) + E[\Pi_\varepsilon(\varepsilon, z)|\varepsilon].$$

If $x = \pi(z) + \eta$, *and* (ε, η) *is independent of z, then*

$$s^* = -\rho_0(\varepsilon)\{\pi(z) - E[\pi(z)]\},\ V^* = \{E[\rho_0(\varepsilon)^2]\}^{-1}\ \mathrm{var}(\pi(z))^{-1},$$

Furthermore, for $\pi = \pi(z)$ *and* $\bar{\pi} = E[\pi]$ *we have,*

$$V = \sigma^2\{E[\partial \ln f(\varepsilon - \alpha^*, \gamma^*)/\partial\alpha]\}^{-2}\{E[(\pi - \bar{\pi})Z']Q^{-1}E[Z(\pi - \bar{\pi})']\}^2,$$

Finally, if π(z) is linear in Z and the pdf $f_0(\varepsilon)$ *of* ε_i *also satisfies* $f_0(\varepsilon) = f(\varepsilon - \alpha^* | \gamma^*)$ *then* $S\hat{\beta}$ *is an efficient estimator.*

The semiparametric bound is the inverse of the variance of the efficient score *s**. It is interesting to note that *s** depends only the score $\rho_0(\varepsilon)$ for ε and the conditional mean $E[x|\varepsilon, Z]$. When there is an additively-separable, reduced-form π(z)+η for *x* the efficient score

takes a more familiar form. In that case s^* is analogous to the efficient score in a linear model with exogenous regressors, where the regressors are replaced by the reduced-form variables $\pi(z)$. In particular, when the disturbance is Gaussian, the bound corresponds to the variance of an efficient instrumental variables estimator. More generally, it corresponds to a GMM estimator where the location score for the disturbance appears in place of the disturbance itself.

We also find that when the reduced form is additive, the asymptotic variance of the NLIV estimator depends only on the scalar function

$$E[\rho(\varepsilon_i - \alpha^*, \gamma^*)^2] / E[\partial\rho(\varepsilon_i - \alpha^*, \gamma^*) / \partial\varepsilon]^2 .$$

and could be minimized by choosing α^* and γ^* to minimize that function. Also, the NLIV estimator will attain the semiparametric variance bound when the reduced form is linear in Z, additive in an independent error, and the parametric family $f(\varepsilon - \alpha|\gamma)$ includes the truth at α^* and γ^*. That is, among all estimators that are consistent, asymptotically normal, and satisfy appropriate regularity conditions under the semiparametric model of equation (1), the estimator we consider will be efficient under the aforementioned conditions. This kind of efficiency property is sometimes termed "local efficiency," referring to the efficiency of the estimator over a subset of the whole semiparametric model.

When $\Pi(\varepsilon, z)$ is not additive in z and ε, attaining efficiency would require an approach different than NLIV based on flexible families of distributions. We focus here on NLIV because it is relatively simple and parsimonious and seems likely to capture much of the efficiency gain available from non-Gaussian disturbances.

5. Applications

In this section we apply the NLIV estimators described in section 2 to two models previously discussed in the literature. The first application is to the problem considered by Card (1995)

which uses 1976 wage and schooling data from the 1966 cohort from the NLS to estimate returns to schooling. The second example uses the model outlined by Angrist and Krueger (1991) with quarter of birth as instrumental variables to estimate returns to schooling based on the 1980 Census for men born between 1930 and 1939. Table 1 summarizes each of these models and related data sets.

In each of the applications, we estimate models of the form

$$y_{1i} = \beta_0 + \beta_1 y_{2i} + \gamma_1' x_i + \varepsilon_{1i}$$
$$y_{2i} = \pi_1' x_i + \pi_2' z_i + \eta_{2i}$$

where y_{1i}, $y_{2i,}$ x_i, and z_i respectively denote observations on the dependent variable, the explanatory variable of interest, a $K_1 x\ 1$ vector of control variables, and a $K_2 x\ 1$ vector of instrumental variables with γ_1, π_1, and π_2 being conformable column vectors of structural and reduced form parameters and ε_{1i} and η_{2i} denoting structural and reduced form random disturbances.

We start by estimating the parameters of the structural equation using ordinary least squares (OLS), limited information maximum likelihood (LIML), and two stage least squares (2SLS) for the two examples and report the estimated schooling coefficients in Table 2. Figures 2 and 3 depict the estimated distributions of the first-step LIML residuals for the two examples along with the normal distribution and SGT distribution implied by the ML estimates of their respective parameters. The Card residual distribution is much more similar to a normal than is the residual distribution for the Angrist-Krueger data, though the SGT provides an improved fit in both cases.

We also report the 2-step NLIV estimates of β_1 based on the t, GT, EGB2, IHS, ST, and SGT pdf's with first step estimated by LIML in Table 2. In both examples, we report results

based on NLIV imposing homoskedasticity on the error terms. We also report results based on NLIV where we allow for heteroskedasticity, specifically by allowing all distributional parameters to be different depending on an individual's quarter of birth, for the Angrist-Krueger example; these estimates are reported in the column labeled "Angrist-Krueger heteroskedasticity".

Looking at the results, we see that the NLIV estimates agree fairly closely with the LIML estimates in the Card example but are quite different in the Angrist-Krueger example when homoskedasticity is imposed. This difference is essentially eliminated when we allow for heteroskedasticity. In both cases, the estimated standard errors associated with the NLIV show evidence of improved efficiency. These improvements range from around five percent for the Card example with rather normal structural error distributions to approximately thirty percent improvement for the Angrist-Krueger model with a much more non-normal error distribution. In the Card example, we find the model selection procedure chooses the ST which gives an estimate of the returns to schooling of .128 with an estimated standard error of .0502; in this example, the LIML estimate is .132 with standard error of .0550. In the Angrist-Krueger example, the model selection procedure chooses the IHS which produces an estimate of .071 with standard error of .0139 in the homoskedastic case and .111 with standard error of .0151 in the heteroskedastic case while the LIML estimate is .109 with standard error .0198. Also of interest is the value of the concentration parameter $\mu^2 = (\pi_2' Z' \left(I - X \left(X' X \right)^{-1} X' \right) Z \pi_2 / Var\left(\eta_{2i} \right))$ which provides a measure of the strength of the instruments. It takes on a rather small value, 13, in the Card data and is large, 108, in the Angrist and Krueger data. Finally, we can compare the estimated schooling coefficient from LIML (or 2SLS) to the NLIV estimate to test for heteroskedasticity (and potentially other types of misspecification). Under the assumption that the conditions

required for the NLIV estimator to attain the efficiency bound are satisfied, the standard error of this difference coefficients is simply given as the square root of the difference in the estimated variances, and the difference between the coefficients divided by this standard error will be asymptotically standard normal. For the Card example, we obtain a value of this test statistic of .178 using LIML and the ST results and would thus fail to reject the hypothesis of homoskedasticity at conventional levels. On the other hand, the value of this statistic is 2.62 using LIML and the IHS results in the Angrist-Krueger example under homoskedasticity, and we would reject the hypothesis of independence at usual levels. However, in the heteroskedastic specification, we obtain a test statistic of -.156 and would fail to reject the hypothesis at conventional levels.

6. Simulation Results

We investigate the properties of some NLIV estimators using Monte Carlo simulations which are similar to the data generating process considered in Newey and Windmeijer (2007). Let the structural relation of interest be

$$y_{1i} = y_{2i}\beta + \varepsilon_i$$

with the corresponding reduced form representation of y_{2i} being

$$y_{2i} = z_i'\pi + \eta_{2i}$$

where the structural disturbance is written in terms of reduced form disturbances as follows

$$\varepsilon_i = \rho\eta_{2i} + \sqrt{1-\rho^2}\eta_{1i}\,.$$

To complete the data generating process for the Monte Carlo study observations of the exogenous variables (instruments) will be generated as

$$z_i \sim N(0, I_K)$$

for i=1,2,…,n. The reduced form coefficients (π) will be specified to be of the form

$$\pi = \sqrt{\frac{\mu^2}{Kn}}\, i_K$$

where i_K denotes a KX1 column vector of ones. μ^2 denotes the concentration parameter $\left[E(\pi' Z' Z\pi / Var(\eta_{2i})) = (n)\pi'\pi / Var(\eta_{2i})\right]$. The distribution of the estimators of β depends upon the values of the concentration parameter (μ^2), the number of instruments (K), the correlation (ρ) between the structural and reduced form (for y_{2i}) disturbances, the distribution of the disturbances, and the sample size. In the sample design we generate samples of size 200 with β = .1, μ^2 = 15, 30, 60, K =3 or 10, and ρ =.3 or .5. We consider three different distributions for η_{1i} and η_{2i} : (1) standard normal; (2) mixture of normal variables or a variance-contaminated normal, U*N[0,1/9]+(1-U)*N[0,9] where U is an independent Bernoulli(.9) random variable; and (3) lognormal. In order for each error distribution to have a zero mean and unitary variance the third reduced form error distribution is generated as $\left(\frac{e^{N[0,1]} - e^{.5}}{\sqrt{e(e-1)}}\right)$.

Monte Carlo simulation results of the alternative estimators of the structural slope coefficient (β =.1) based on 20,000 simulation replications are summarized in Tables 3, 4, and 5. We report median bias (Bias), interquartile range (IQR), median absolute deviation (MAD), and 95% confidence interval coverage probability (CP) for estimates of β obtained by OLS, 2SLS, and LIML, as well as the 2-step NLIV estimators of β corresponding to the t, GT, EGB2, IHS ST, and SGT pdf's using LIML first step estimates and QMLE based on the LIML residuals to obtain estimates of the distributional parameters.

Results for the normal error distribution are summarized in Table 3. In this case, the median sample bias is minimized by LIML among the estimators we consider. Among the NLIV estimators, the bias and spread as measured by the interquartile range are largely insensitive to the pdf. In this case, the NLIV estimators all perform worse than LIML in terms of median bias but appear to have considerably less dispersion as measured by IQR. NLIV estimators also dominate LIML in terms of estimator risk as measured by MAD. We note that 2SLS, which is numerically identical to NLIV using a normal distribution, does slightly better than the other NLIV estimators. As would be expected, we see that coverage probabilities for NLIV estimators may be distorted when instruments are weak ($\mu^2 = 15$) or many (K = 10) and that this distortion decreases as μ^2 /K increases.

The results corresponding to the case of a mixed normal or variance contaminated error distribution which is symmetric with thick tails (standardized kurtosis is approximately 20) are reported in Table 4. The median bias of all estimators decreases as μ^2 /K increases. The NLIV estimators produce substantially smaller values of IQR than LIML or 2SLS with IQR's less than 50% those of LIML or 2SLS in some cases. The gains are also apparent in MAD terms where improvements are similar to those in IQR. It is interesting that in this case, unlike the normal design, the majority of NLIV estimators have median biases which are comparable to LIML. The NLIV estimators, especially those based on the SGT, do suffer somewhat relative to LIML in terms of coverage probability. We also see that using the simple model selection procedure produces an estimator with quite favorable properties.

The impact of a skewed and leptokurtic error distribution on estimator performance can be seen in Table 5. As in the thick-tailed case above, we see that the NLIV estimators show large improvements in efficiency that are not necessarily accompanied by large increases in bias

relative to LIML or 2SLS. Not surprisingly, the NLIV estimators based on possibly skewed pdf's show the greatest improvement with the exception of the SGT which may need a larger sample size to accurately model the underlying error distribution. As before, we see that the NLIV estimators suffer somewhat in terms of coverage relative to LIML. This is especially true for the SGT and GT which perform very poorly. We also see that model selection procedure produces an estimator that performs well overall.

Overall, the simulation results are encouraging for the NLIV estimators. In the case of nonnormal disturbances, the NLIV estimators show substantial gains relative to LIML or 2SLS in terms of dispersion and MAD. These gains are accompanied by only minor losses in the case of normal errors. As expected, the coverage probabilities of interval estimates based on the NLIV estimators are somewhat distorted in cases of weak or many instruments, though this could likely be remedied by adopting existing results from the weak and many instruments literatures in these cases. There also appear to be some distortions in coverage probabilities even when the instruments are stronger, though they are generally minor. This could be the result of the small sample or may suggest that pursuing other approaches to estimating standard errors, such as the bootstrap, may be desirable in this context.

7. Summary and conclusions

In this paper, we consider efficiency gains that might be available using moment conditions which are nonlinear in the disturbances. The nonlinear functions we consider are based on the use of flexible parametric families of disturbance distributions. We illustrate the approach in two empirical examples. In both examples, the NLIV estimators are associated with smaller standard errors than conventional IV estimators. Monte Carlo simulations demonstrate

that while NLIV estimators may be associated with modest efficiency loss in the case of normal error distributions, they offer the possibility of significant efficiency improvements in the presence of thick-tailed and/or skewed error distributions.

Appendix: Proofs of Theorems

Proof of Theorem 1: Let e_1 denote the first unit vector and $\beta^*=\beta_0+\alpha^* e_1$. By Assumption 1,

$$\begin{aligned} E[Z_i\rho(y_i - X_i'\beta^*,\gamma^*)] &= E[Z_i\rho(\varepsilon_i - \alpha^*,\gamma^*)] \\ &= E[Z_i]E[\rho(\varepsilon_i - \alpha^*,\gamma^*)] = 0. \end{aligned}$$

Also, this β^* is unique by Assumption 1. Let $Z = (Z_1,\ldots,Z_n)'$ and $W=Q^{-1}$. Note that $n\hat{Q}^{-1} = (Z'Z/n)^{-1} \xrightarrow{p} W$ by the law of large numbers (LLN) and the continuous mapping theorem (CMT). Let $g(\beta) = E[Z_i\rho(y_i - X_i'\beta,\gamma^*)]$ and $Q(\beta) = g(\beta)'Wg(\beta)$. By Assumption 2 and a standard uniform convergence argument,

$$\sup_{\beta\in B}\left\|\hat{g}(\beta)/n - g(\beta)\right\| \xrightarrow{p} 0.$$

It follows as in the proof of Theorem 2.6 of Newey and McFadden (1994) that

$$\sup_{\beta\in B}\left\|\hat{g}(\beta)'\hat{Q}^{-1}\hat{g}(\beta)/n - Q(\beta)\right\| \xrightarrow{p} 0.$$

Also, the objective function $Q(\beta)$ has a unique minimum at β^*, so it follows as in the proof of Theorem 2.6 of Newey and McFadden (1994) that $\hat{\beta} \xrightarrow{p} \beta^*$.

Let $\hat{G}(\beta) = \sum_{i=1}^{n} Z_i X_i'\partial\rho(y_i - X_i'\beta,\tilde{\gamma})/\partial\varepsilon/n$. It follows by standard arguments that for any $\hat{\beta} \xrightarrow{p} \beta^*$,

$$\hat{G}(\bar{\beta}) \xrightarrow{p} G$$

where $\bar{\beta}$ is an intermediate value between $\hat{\beta}$ and $\bar{\beta}$.

Next, for $u_i = \varepsilon_i - \alpha *$ and any $\bar{\gamma} \xrightarrow{p} \gamma *$, by standard arguments,

$$\frac{1}{n}\sum_{i=1}^{n} Z_i \frac{\partial \rho}{\partial \gamma}(u_i, \bar{\gamma})' \xrightarrow{p} E\left[Z_i \frac{\partial \rho}{\partial \gamma}(u_i, \gamma *)' \right] = E[Z_i] E\left[\frac{\partial \rho}{\partial \gamma}(u_i, \gamma *)' \right] = Ge_1 r,$$

where $r = E[\partial \rho(u_i, \gamma *)/\partial \gamma']/E[\partial \rho(u_i, \gamma *)/\partial u]$, and e_1 is the first unit vector. It then follows by an expansion that

$$\begin{aligned} \sqrt{n}\hat{g}(\beta *) &= \frac{1}{\sqrt{n}}\sum_i Z_i \rho(u_i, \tilde{\gamma}) \\ &= \frac{1}{\sqrt{n}}\sum_i Z_i \rho(u_i, \gamma *) + \frac{1}{n}\sum_i Z_i \frac{\partial \rho}{\partial \gamma}(u_i, \tilde{\gamma})' \sqrt{n}(\tilde{\gamma} - \gamma *) \\ &= \frac{1}{\sqrt{n}}\sum_i Z_i \rho(u_i, \gamma *) + Ge_1 r \sqrt{n}(\tilde{\gamma} - \gamma *) + o_p(1). \end{aligned}$$

It also follows by standard GMM arguments that

$$\begin{aligned} \sqrt{n}(\hat{\beta} - \beta *) &= -(G'WG)^{-1} G'W \sqrt{n}\hat{g}(\beta *) + o_p(1) \\ &= -(G'WG)'^{-1} G^1 W \frac{1}{\sqrt{n}}\sum_i Z_i \rho(u_i, \gamma *) \\ &\quad - e_1 r \sqrt{n}(\tilde{\gamma} - \gamma *) + o_p(1). \end{aligned}$$

By the Lindberg-Levy central limit theorem, $\sum_i Z_i \rho(u_i, \gamma *)/\sqrt{n} \xrightarrow{d} N(0, \sigma^2 Q^{-1})$.

Premultiplying by S we obtain (by $Se_1 = 0$),

$$\begin{aligned} \sqrt{n}S(\hat{\beta} - \beta *) &= S(G'WG)^{-1} G'W \frac{1}{\sqrt{n}}\sum_i Z_i \rho(u_i, \gamma *) + o_p(1) \\ &\xrightarrow{d} S(G'WG)^{-1} G'W N(0, \sigma^2 Q^{-1}) \\ &= N(0, \sigma^2 S(G'WG)^{-1} S). \end{aligned}$$

giving the first conclusion. The second conclusion also follows by a standard argument. *Q.E.D.*

Proof of Theorem 2: Here let δ denote the slope coefficients $S\beta$. By the assumption that z and ε are independent the joint pdf of ε, z, and X takes the form

$$f(\varepsilon)g(z)h(x|\varepsilon, z),$$

where *f* and *g* denote the marginal densities of ε and z respectively, and *h* is the conditional pdf of *X* given z and ε. Substituting *y*-*x*′δ for ε and differentiating we find the score for δ to be

$$s_\delta = -x\{\rho_0(\varepsilon) + \partial \ln h(x|\varepsilon, Z)\partial\varepsilon\}, \quad \rho_0(\varepsilon) = \partial \ln f_0(\varepsilon)/\partial\varepsilon.$$

Applying eq. (23) of Newey (1990a), the efficient score is

$$s^* = E[s_\beta|z,\varepsilon] - E[s_\beta|\varepsilon].$$

Note also that by interchanging the order of differentiation and integration we have

$$E[x\partial \ln h(x|\varepsilon,z)\partial\varepsilon|\varepsilon,z] = \int x[\partial h(x|\varepsilon,z)/\partial\varepsilon]dx = \Pi_\varepsilon(\varepsilon,z).$$

Then applying iterated expectations gives the first conclusion.

Next, if *x*=π(z)+η for η independent of *Z* we have

$$\begin{aligned} \Pi(\varepsilon,z) &= \pi(z) + \lambda(\varepsilon), \quad \lambda(\varepsilon) = E[\eta|\varepsilon], \\ \Pi_\varepsilon(\varepsilon,z) &= \lambda_\varepsilon(\varepsilon) = \partial\lambda(\varepsilon)/\partial\varepsilon. \end{aligned}$$

It then follows that

$$\begin{aligned} \Pi(\varepsilon,z) - E[\Pi(\varepsilon,z)|\varepsilon] &= \pi(z) + \lambda(\varepsilon) - E[\pi(z)] - \lambda(\varepsilon) = \pi(z) - E[\pi(z)], \\ \Pi_\varepsilon(\varepsilon,z) - E[\Pi_\varepsilon(\varepsilon,z)|\varepsilon] &= \lambda_\varepsilon(\varepsilon) - \lambda_\varepsilon(\varepsilon) = 0. \end{aligned}$$

Substituting these expressions in the formula for *s** gives the second conclusion.

Next, suppose that π(*Z*) is linear in *Z*, i.e. that π(*Z*)=Π*Z* for a constant matrix Π. Let $\pi_i = \Pi Z_i$, $\bar{\pi} = E[\pi_i]$, and $\rho_{\varepsilon i} = \partial\rho(\varepsilon_i - \alpha^*, \gamma^*)/\partial\varepsilon$. Note that $X_i' = (1, \pi_i') + (0, \eta_i')$ so that

$$\begin{aligned} G &= E[Z_i X_i' \rho_{\varepsilon i}] = E[Z_i(1,\pi_i')]E[\rho_{\varepsilon i}] + E[Z_i]E[\rho_{\varepsilon i}(0,\eta_i')] \\ &= E[Z_i(0,\pi_i' - \bar{\pi}')]E[\rho_{\varepsilon i}] + E[Z_i]\{E[\rho_{\varepsilon i}](1,\bar{\pi}') + E[\rho_{\varepsilon i}(0,\eta_i')]\} \\ &= E[\rho_{\varepsilon i}]\{E[Z_i(0,\pi_i' - \bar{\pi}')] + E[Z_i](1,a')\}, \; a = \bar{\pi} + E[\rho_{\varepsilon i}\eta_i]/E[\rho_{\varepsilon i}]. \end{aligned}$$

Note that $Q^{-1}E[Z_i] = Q^{-1}Qe_1 = e_1$ and that $e_1'E[Z_i(0,\pi_i' - \bar{\pi}')] = E[1\cdot(0,\pi_i' - \bar{\pi}')] = 0$. Also, for

$$E[(0,\pi_i' - \bar{\pi}')'Z_i']Q^{-1}E[Z_i(0,\pi_i' - \bar{\pi}')] = \text{diag}[0,\tilde{Q}^{-1}], \tilde{Q}^{-1} = E[(\pi_i' - \bar{\pi}')'Z_i']Q^{-1}E[Z_i(\pi_i' - \bar{\pi}')].$$

Then we have

$$G'Q^{-1}G = (E[\rho_{\varepsilon i}])^2\{\text{diag}[0,\widetilde{Q}\ ] + (1,a')'(1,a')\}$$
$$= (E[\rho_{\varepsilon i}])^2\begin{bmatrix}1 & a' \\ a & \widetilde{Q}\ + aa'\end{bmatrix}.$$

By the partitioned inverse formula it follows that

$$V = \sigma^2 S'(G'Q^{-1}G)^{-1}S = \{\sigma^2/(E[\rho_{\varepsilon i}])^2\}(\widetilde{Q}\ + aa' - aa'/1)^{-1}$$
$$= \sigma^2/(E[\rho_{\varepsilon i}])^2\widetilde{Q}^{-1},$$

giving the third conclusion.

For the fourth conclusion, note that when π_i' is a linear combination of Z_i

then $(0,\pi_i' - \overline{\pi}')$ is too, so that

$$E[(0,\pi_i' - \overline{\pi}')'Z_i']Q^{-1}E[Z_i(0,\pi_i' - \overline{\pi}')]$$
$$= E[(0,\pi_i' - \overline{\pi}')'(0,\pi_i' - \overline{\pi}')] = \text{diag}[0, \text{var}(\pi_i)].$$

Furthermore, if $f_0(\varepsilon) = f(\varepsilon - \alpha^*, \gamma^*)$ then $\rho_0(\varepsilon) = \rho(\varepsilon)$ and the information matrix equality for a location parameter gives $E[\rho\,\varepsilon_i] = -\sigma^2 = E[\rho_0(\varepsilon)^2]$, so that

$$V = \{E[\rho_0(\varepsilon)^2]\}\,\text{var}(\pi_i)^{-1} = (E[s^* s^{*\prime}])^{-1}.\ Q.E.D.$$

References

Andrews, D. W. K. (1999). "Consistent moment selection procedures for generalized method of moments," *Econometrica*, 67, 543-564.

Angrist, J. and A. K. Krueger (1991). "Does Compulsory School Attendance Affect Schooling and Earnings?" *Quarterly Journal of Economics*, 106, 979-1014.

Card, D. (1995). "Using Geographic Variation in College Proximity to Estimate the Return to Schooling,"in *Aspects of Labour Market Behavior: Essays in Honour of John Vanderkamp*. Ed. L.N. Christophides, E. K. Grant, and R. Swidinsky, 201-222. Toronto: University of Toronto Press.

Chernozhukov, V. and C. Hansen (2005). "An IV model of quantile treatment effects," *Econometrica*, 73, 245-261.

Hansen, B. E. (1994). "Autoregressive conditional density estimation," *International Economic Review*, 35, 705-730.

Hansen, C., J. B. McDonald, and P. Theodossiou (2005). "Some flexible parametric models for partially adaptive estimators of econometric models," working paper.

Honore, B.E. and L. Hu (2004). "On the performance of some robust instrumental variables estimators," *Journal of Business & Economic Statistics* 22, 30-39.

Johnson, N. L. (1949). "Systems of frequency curves generated by methods of translation," *Biometrica* 36, 149-176.

Kleibergen, F. (2005). "Testing parameters in GMM without assuming that they are identified," *Econometrica* 73, 1103-1124.

McDonald, J. B. and W. K. Newey (1988). "Partially adaptive estimation of regression models via the generalized t distribution," *Econometric Theory* 4, 428-457.

McDonald, J. B. and S. B. White (1993). "Comparison of robust, adaptive, and partially adaptive estimators of regression models," *Econometric Reviews* 37, 273-278.

McDonald, J. B. and Y. J. Xu (1995). "A generalization of the beta distribution with applications," *Journal of Econometrics* 66, 133-152. Errata 69 (1995), 427-428.

Newey, W. K. (1988). "Adaptive estimation of regression models via moment restrictions," *Journal of Econometrics* 38, 301-339.

Newey, W.K. (1990). "Semiparametric Efficiency Bounds," *Journal of Applied Econometrics* 5, 99-135.

Newey, W. K. and D. McFadden (1994). "Large sample estimation and hypothesis testing," Handbook of Econometrics, eds. R. Engle and D. McFadden, vol. 4, chapter 36. North Holland.

Newey, W. K. and F. Windmeijer (2007). "GMM with many weak moment conditions," working paper, MIT.

Subbotin, M. T. (1923). "On the law of frequency of error," *Mathematicheskii Sbornik* 31, 296-301.

Stock, J. H. and J. H. Wright (2000). "GMM with weak identification," *Econometrica*, 68, 1055-1096.

Theodossiou, P., (1998). "Financial data and the skewed generalized t distribution," *Management Science* 44, 1650-1661.

White, H. (1982). "Maximum likelihood estimation of misspecified models," *Econometrica* 50(1), 1-25.

Table 1. Model Summary

	Card (1995)	**Angrist and Krueger (1991)**
Data	NLS Young Men (1966 Cohort) 1976 sample. N = 3010	1980 U.S. Census Sample of men born between 1930 and 1939. N=329,509
Dependent Variable (y_{1i})	Log (wage)	Log (wage)
Explanatory variable of interest (y_{2i})	Years of education	Years of education
Control variables (x_i)	Race, Experience, SMSA, Region	Year of birth (9 variables) State of birth (50 variables)
Instruments (z_i)	Binary: grew up near a 4 year college	Binary variables: quarter of birth

Table 2. Estimation results of Schooling Coefficient

	Card		Angrist-Krueger homoskedastic		Angrist-Krueger heteroskedastic	
Estimator	$\hat{\beta}_1$	$s_{\hat{\beta}_1}$	$\hat{\beta}_1$	$s_{\hat{\beta}_1}$	$\hat{\beta}_1$	$s_{\hat{\beta}_1}$
	.075	.0035	.067	.0004	.067	.0004
LIML	.132	.0550	.109	.0198	.109	.0198
Normal (2SLS)	.132	.0550	.108	.0195	.108	.0196
T	.131	.0508	.078	.0143	.082	.0144
GT	.130	.0504	.081	.0137	.084	.0138
EGB2	.136	.0554	.094	.0170	.109	.0155
IHS	.132	.0522	.071*	.0139*	.111*	.0151*
ST	.128*	.0502*	.074	.0144	.112	.0147
SGT	.124	.0575	.075	.0139	.112	.0140
μ^2	13.3			108.1		

Note: Estimated schooling coefficient and standard error using various estimators for Card (1995) and Angrist and Krueger (1991) examples summarized in Table 1. The first two rows correspond to OLS and LIML, and the remaining rows give results using NLIV estimators corresponding to the specified distribution. * denotes the model chosen by the simple model selection procedure.

Table 3. Simulation Results from Normal Design

	K = 3								K = 10							
	ρ = 0.3				ρ = 0.5				ρ = 0.3				ρ = 0.5			
Estimator	**Bias**	**IQR**	**MAD**	**CP**	**Bias**	**IQR**	**MAD**	**CP**	**BIAS**	**IQR**	**MAD**	**CP**	**Bias**	**IQR**	**MAD**	**CP**
	A. $\mu^2 = 15$								A. $\mu^2 = 15$							
OLS	0.279	0.088	0.279	0.012	0.465	0.080	0.465	0.000	0.279	0.088	0.279	0.011	0.466	0.082	0.466	0.000
LIML	0.004	0.384	0.191	0.960	-0.003	0.376	0.185	0.950	0.001	0.458	0.228	0.936	0.007	0.439	0.216	0.921
2SLS-Normal	0.041	0.329	0.168	0.956	0.059	0.316	0.169	0.933	0.116	0.267	0.162	0.908	0.192	0.251	0.206	0.793
EGB2	0.037	0.336	0.172	0.949	0.054	0.322	0.171	0.927	0.107	0.283	0.164	0.894	0.179	0.263	0.199	0.790
IHS	0.039	0.333	0.171	0.950	0.056	0.320	0.170	0.927	0.110	0.278	0.163	0.897	0.183	0.258	0.201	0.790
T	0.038	0.335	0.170	0.951	0.054	0.319	0.170	0.930	0.110	0.277	0.163	0.901	0.182	0.259	0.201	0.794
GT	0.043	0.335	0.170	0.940	0.064	0.323	0.173	0.918	0.118	0.275	0.168	0.879	0.197	0.259	0.215	0.759
ST	0.038	0.337	0.172	0.948	0.054	0.324	0.171	0.924	0.109	0.282	0.165	0.890	0.180	0.264	0.200	0.784
SGT	0.043	0.336	0.172	0.935	0.063	0.326	0.174	0.912	0.115	0.281	0.168	0.865	0.194	0.265	0.213	0.748
BIC	0.041	0.332	0.170	0.953	0.060	0.319	0.171	0.930	0.115	0.271	0.163	0.903	0.191	0.255	0.207	0.790
	B. $\mu^2 = 30$								B. $\mu^2 = 30$							
OLS	0.260	0.086	0.260	0.017	0.435	0.079	0.435	0.000	0.260	0.085	0.260	0.019	0.435	0.080	0.435	0.000
LIML	0.001	0.259	0.129	0.956	0.001	0.257	0.128	0.954	-0.001	0.288	0.143	0.939	-0.001	0.282	0.139	0.939
2SLS-Normal	0.019	0.240	0.122	0.955	0.033	0.238	0.123	0.943	0.071	0.212	0.119	0.921	0.117	0.202	0.139	0.850
EGB2	0.018	0.245	0.124	0.948	0.029	0.241	0.124	0.936	0.064	0.222	0.121	0.911	0.107	0.208	0.136	0.851
IHS	0.018	0.243	0.124	0.949	0.031	0.240	0.124	0.937	0.066	0.219	0.120	0.915	0.111	0.205	0.137	0.849
T	0.018	0.243	0.123	0.952	0.030	0.241	0.124	0.940	0.065	0.218	0.120	0.916	0.109	0.206	0.137	0.852
GT	0.019	0.245	0.125	0.942	0.035	0.242	0.125	0.929	0.071	0.217	0.123	0.897	0.122	0.208	0.147	0.822
ST	0.018	0.245	0.124	0.947	0.030	0.242	0.125	0.935	0.064	0.224	0.122	0.907	0.108	0.209	0.137	0.847
SGT	0.019	0.247	0.125	0.937	0.035	0.244	0.126	0.921	0.070	0.222	0.123	0.889	0.120	0.212	0.146	0.817
BIC	0.020	0.242	0.123	0.952	0.033	0.240	0.123	0.939	0.069	0.215	0.120	0.918	0.115	0.204	0.139	0.848
	C. $\mu^2 = 60$								C. $\mu^2 = 60$							
OLS	0.231	0.081	0.231	0.032	0.385	0.075	0.385	0.000	0.231	0.081	0.231	0.033	0.384	0.076	0.384	0.000
LIML	0.001	0.178	0.089	0.953	0.001	0.177	0.088	0.950	-0.001	0.187	0.094	0.947	0.001	0.186	0.093	0.949
2SLS-Normal	0.011	0.171	0.087	0.951	0.017	0.170	0.086	0.943	0.039	0.161	0.086	0.934	0.067	0.157	0.095	0.899
EGB2	0.010	0.174	0.088	0.943	0.016	0.171	0.087	0.936	0.035	0.165	0.087	0.924	0.061	0.162	0.095	0.897
IHS	0.011	0.174	0.088	0.945	0.017	0.171	0.087	0.938	0.037	0.164	0.086	0.925	0.063	0.160	0.095	0.896
T	0.010	0.173	0.087	0.947	0.016	0.170	0.087	0.940	0.036	0.163	0.086	0.929	0.062	0.160	0.094	0.900
GT	0.012	0.175	0.088	0.937	0.018	0.171	0.088	0.931	0.041	0.164	0.088	0.913	0.070	0.162	0.099	0.876
ST	0.010	0.175	0.088	0.942	0.016	0.171	0.087	0.935	0.036	0.165	0.087	0.921	0.061	0.162	0.095	0.894
SGT	0.011	0.177	0.089	0.932	0.018	0.173	0.088	0.925	0.041	0.167	0.089	0.905	0.069	0.163	0.100	0.868
BIC	0.011	0.173	0.087	0.948	0.017	0.171	0.087	0.940	0.038	0.162	0.086	0.929	0.065	0.159	0.096	0.897

Note: Results for normal simulation model described in text. The design uses 200 observations, and all results are based on 20,000 simulation replications. We report median bias (Bias), interquartile range (IQR), median absolute deviation (MAD), and 95% confidence interval coverage (CP). Rows labeled EGB2, IHS, t, GT, ST, and SGT correspond to NLIV estimates based on the distribution given in the row label. At each iteration, BIC uses the estimator selected by the model selection procedure outlined in the text. We also note that 2SLS corresponds to NLIV when the assumed error distribution is normal.

Table 4. Simulation Results from Normal Mixture Design

	K = 3								K = 10							
	ρ = 0.3				ρ = 0.5				ρ = 0.3				ρ = 0.5			
Estimator	**Bias**	**IQR**	**MAD**	**CP**	**Bias**	**IQR**	**MAD**	**CP**	**BIAS**	**IQR**	**MAD**	**CP**	**Bias**	**IQR**	**MAD**	**CP**
	A. $\mu^2 = 15$								A. $\mu^2 = 15$							
OLS	0.278	0.244	0.279	0.192	0.464	0.227	0.464	0.025	0.280	0.248	0.281	0.186	0.464	0.225	0.464	0.026
LIML	0.000	0.362	0.180	0.958	0.002	0.357	0.177	0.953	0.004	0.435	0.217	0.933	0.005	0.410	0.201	0.925
2SLS-Normal	0.036	0.315	0.162	0.952	0.059	0.307	0.164	0.932	0.112	0.279	0.163	0.881	0.186	0.260	0.203	0.780
EGB2	0.002	0.146	0.073	0.924	0.002	0.145	0.072	0.929	0.006	0.160	0.080	0.861	0.013	0.153	0.077	0.874
IHS	0.002	0.163	0.081	0.904	0.001	0.162	0.081	0.911	0.007	0.191	0.096	0.794	0.011	0.185	0.094	0.799
T	0.001	0.145	0.073	0.918	-0.001	0.146	0.073	0.923	0.003	0.166	0.083	0.831	0.006	0.161	0.081	0.844
GT	0.001	0.164	0.082	0.830	0.002	0.164	0.082	0.831	0.010	0.194	0.098	0.655	0.016	0.184	0.094	0.667
ST	0.001	0.147	0.073	0.912	0.000	0.148	0.074	0.917	0.003	0.169	0.085	0.821	0.007	0.163	0.082	0.835
SGT	0.000	0.168	0.084	0.816	0.002	0.166	0.083	0.819	0.009	0.199	0.100	0.639	0.014	0.190	0.097	0.647
BIC	0.000	0.150	0.075	0.918	0.000	0.151	0.075	0.923	0.005	0.172	0.086	0.839	0.009	0.167	0.084	0.845
	B. $\mu^2 = 30$								B. $\mu^2 = 30$							
OLS	0.259	0.225	0.260	0.195	0.432	0.212	0.432	0.028	0.258	0.234	0.259	0.208	0.431	0.213	0.431	0.028
LIML	-0.001	0.246	0.123	0.960	0.002	0.244	0.121	0.954	-0.001	0.278	0.138	0.942	0.001	0.271	0.134	0.939
2SLS-Normal	0.018	0.230	0.116	0.956	0.032	0.227	0.117	0.939	0.067	0.215	0.119	0.910	0.114	0.210	0.138	0.842
EGB2	0.001	0.103	0.052	0.933	0.001	0.101	0.050	0.934	0.003	0.108	0.054	0.894	0.005	0.106	0.053	0.903
IHS	0.001	0.113	0.057	0.915	0.000	0.112	0.056	0.920	0.002	0.124	0.062	0.853	0.004	0.123	0.062	0.852
T	0.000	0.103	0.051	0.925	-0.001	0.100	0.050	0.929	0.000	0.110	0.055	0.881	0.002	0.108	0.054	0.886
GT	0.000	0.111	0.056	0.852	0.000	0.111	0.055	0.855	0.002	0.125	0.063	0.723	0.004	0.125	0.062	0.730
ST	0.000	0.103	0.051	0.921	-0.001	0.101	0.051	0.926	0.000	0.111	0.055	0.870	0.002	0.108	0.054	0.876
SGT	0.000	0.115	0.057	0.838	0.000	0.114	0.057	0.845	0.002	0.128	0.064	0.709	0.005	0.127	0.064	0.714
BIC	0.000	0.105	0.053	0.928	-0.001	0.103	0.052	0.931	0.002	0.113	0.056	0.879	0.002	0.112	0.056	0.885
	C. $\mu^2 = 60$								C. $\mu^2 = 60$							
OLS	0.227	0.207	0.228	0.220	0.380	0.193	0.380	0.033	0.228	0.206	0.229	0.217	0.378	0.191	0.378	0.033
LIML	0.001	0.172	0.086	0.955	-0.001	0.175	0.088	0.949	0.000	0.178	0.089	0.945	-0.001	0.179	0.089	0.947
2SLS-Normal	0.009	0.166	0.084	0.952	0.015	0.169	0.085	0.941	0.039	0.157	0.084	0.925	0.062	0.156	0.092	0.888
EGB2	0.001	0.071	0.036	0.934	0.001	0.071	0.036	0.936	0.002	0.074	0.037	0.914	0.002	0.073	0.037	0.911
IHS	0.000	0.078	0.039	0.922	0.001	0.078	0.039	0.924	0.002	0.084	0.042	0.877	0.002	0.085	0.043	0.873
T	0.000	0.070	0.035	0.931	0.000	0.071	0.035	0.934	0.001	0.075	0.037	0.902	0.001	0.075	0.037	0.901
GT	0.000	0.076	0.038	0.869	0.000	0.077	0.038	0.871	0.002	0.084	0.042	0.772	0.002	0.084	0.042	0.769
ST	0.000	0.071	0.036	0.926	0.000	0.071	0.036	0.929	0.001	0.075	0.037	0.892	0.000	0.075	0.038	0.892
SGT	0.000	0.077	0.039	0.860	0.001	0.078	0.039	0.862	0.002	0.085	0.043	0.753	0.002	0.086	0.043	0.754
BIC	0.000	0.072	0.036	0.931	0.001	0.072	0.036	0.934	0.002	0.077	0.038	0.902	0.000	0.076	0.038	0.898

Note: Results for normal mixture simulation model described in text. The design uses 200 observations, and all results are based on 20,000 simulation replications. We report median bias (Bias), interquartile range (IQR), median absolute deviation (MAD), and 95% confidence interval coverage (CP). Rows labeled EGB2, IHS, t, GT, ST, and SGT correspond to NLIV estimates based on the distribution given in the row label. At each iteration, BIC uses the estimator selected by the model selection procedure outlined in the text. We also note that 2SLS corresponds to NLIV when the assumed error distribution is normal.

Table 5. Simulation Results from Lognormal Design

	K = 3								K = 10							
	ρ = 0.3				ρ = 0.5				ρ = 0.3				ρ = 0.5			
Estimator	**Bias**	**IQR**	**MAD**	**CP**	**Bias**	**IQR**	**MAD**	**CP**	**BIAS**	**IQR**	**MAD**	**CP**	**Bias**	**IQR**	**MAD**	**CP**
	A. $\mu^2 = 15$								A. $\mu^2 = 15$							
OLS	0.271	0.079	0.271	0.055	0.455	0.075	0.455	0.002	0.271	0.081	0.271	0.054	0.456	0.075	0.456	0.002
LIML	0.000	0.352	0.175	0.957	0.003	0.353	0.174	0.944	0.006	0.417	0.208	0.929	0.006	0.409	0.202	0.918
2SLS-Normal	0.035	0.308	0.159	0.952	0.063	0.303	0.162	0.922	0.114	0.260	0.160	0.885	0.185	0.253	0.203	0.772
EGB2	0.009	0.099	0.050	0.899	0.004	0.109	0.054	0.911	0.031	0.106	0.058	0.804	0.027	0.115	0.060	0.835
IHS	0.008	0.108	0.055	0.885	0.005	0.116	0.058	0.908	0.033	0.126	0.066	0.777	0.030	0.125	0.066	0.827
T	0.001	0.195	0.097	0.899	0.001	0.208	0.103	0.917	0.023	0.215	0.111	0.773	0.020	0.224	0.114	0.833
GT	0.003	0.229	0.114	0.684	0.005	0.235	0.117	0.713	0.020	0.273	0.138	0.427	0.023	0.280	0.141	0.460
ST	0.014	0.146	0.073	0.747	0.006	0.144	0.073	0.791	0.041	0.199	0.099	0.546	0.036	0.182	0.096	0.603
SGT	0.017	0.177	0.088	0.702	0.008	0.165	0.084	0.755	0.047	0.244	0.121	0.494	0.040	0.214	0.110	0.546
BIC	0.007	0.140	0.070	0.894	0.007	0.152	0.077	0.906	0.031	0.164	0.085	0.794	0.030	0.167	0.088	0.829
	B. $\mu^2 = 30$								B. $\mu^2 = 30$							
OLS	0.252	0.079	0.252	0.067	0.423	0.083	0.423	0.003	0.252	0.081	0.252	0.071	0.424	0.082	0.424	0.003
LIML	-0.001	0.242	0.121	0.953	0.003	0.245	0.122	0.949	0.002	0.265	0.132	0.940	-0.001	0.258	0.128	0.939
2SLS-Normal	0.018	0.227	0.115	0.950	0.033	0.226	0.118	0.933	0.070	0.206	0.116	0.906	0.112	0.199	0.137	0.841
EGB2	0.004	0.067	0.034	0.913	0.003	0.076	0.038	0.920	0.014	0.070	0.037	0.854	0.014	0.079	0.041	0.861
IHS	0.003	0.074	0.037	0.901	0.002	0.080	0.040	0.921	0.014	0.083	0.042	0.821	0.013	0.083	0.043	0.862
T	-0.001	0.132	0.066	0.923	-0.001	0.144	0.072	0.928	0.008	0.142	0.072	0.849	0.008	0.150	0.075	0.887
GT	-0.001	0.148	0.073	0.745	0.001	0.154	0.077	0.769	0.008	0.171	0.086	0.507	0.009	0.176	0.089	0.549
ST	0.003	0.093	0.047	0.776	0.002	0.100	0.050	0.801	0.018	0.123	0.063	0.590	0.015	0.119	0.061	0.636
SGT	0.005	0.110	0.055	0.738	0.003	0.112	0.056	0.771	0.021	0.145	0.073	0.544	0.019	0.136	0.070	0.582
BIC	0.004	0.097	0.049	0.906	0.004	0.108	0.054	0.915	0.015	0.109	0.056	0.839	0.013	0.111	0.057	0.866
	C. $\mu^2 = 60$								C. $\mu^2 = 60$							
OLS	0.223	0.081	0.223	0.094	0.371	0.089	0.371	0.006	0.222	0.081	0.222	0.098	0.370	0.090	0.370	0.007
LIML	0.002	0.169	0.084	0.952	0.002	0.169	0.084	0.951	-0.001	0.177	0.088	0.947	-0.002	0.176	0.087	0.946
2SLS-Normal	0.012	0.162	0.082	0.948	0.018	0.162	0.082	0.942	0.038	0.154	0.082	0.924	0.062	0.153	0.092	0.884
EGB2	0.002	0.048	0.024	0.922	0.001	0.053	0.027	0.917	0.007	0.048	0.025	0.877	0.007	0.055	0.028	0.873
IHS	0.000	0.052	0.026	0.905	0.001	0.056	0.028	0.919	0.005	0.056	0.028	0.836	0.006	0.058	0.030	0.873
T	-0.001	0.093	0.046	0.928	0.001	0.099	0.050	0.937	0.003	0.097	0.049	0.893	0.004	0.104	0.052	0.908
GT	-0.001	0.098	0.049	0.774	0.001	0.103	0.052	0.794	0.003	0.114	0.057	0.569	0.004	0.119	0.059	0.597
ST	0.001	0.063	0.031	0.791	0.000	0.068	0.034	0.799	0.008	0.082	0.041	0.603	0.007	0.083	0.042	0.646
SGT	0.002	0.073	0.037	0.757	0.001	0.076	0.038	0.772	0.011	0.094	0.048	0.556	0.009	0.093	0.047	0.595
BIC	0.002	0.069	0.035	0.910	0.002	0.075	0.038	0.917	0.006	0.073	0.037	0.859	0.006	0.076	0.039	0.883

Note: Results for lognormal simulation model described in text. The design uses 200 observations, and all results are based on 20,000 simulation replications. We report median bias (Bias), interquartile range (IQR), median absolute deviation (MAD), and 95% confidence interval coverage (CP). Rows labeled EGB2, IHS, t, GT, ST, and SGT correspond to NLIV estimates based on the distribution given in the row label. At each iteration, BIC uses the estimator selected by the model selection procedure outlined in the text. We also note that 2SLS corresponds to NLIV when the assumed error distribution is normal.

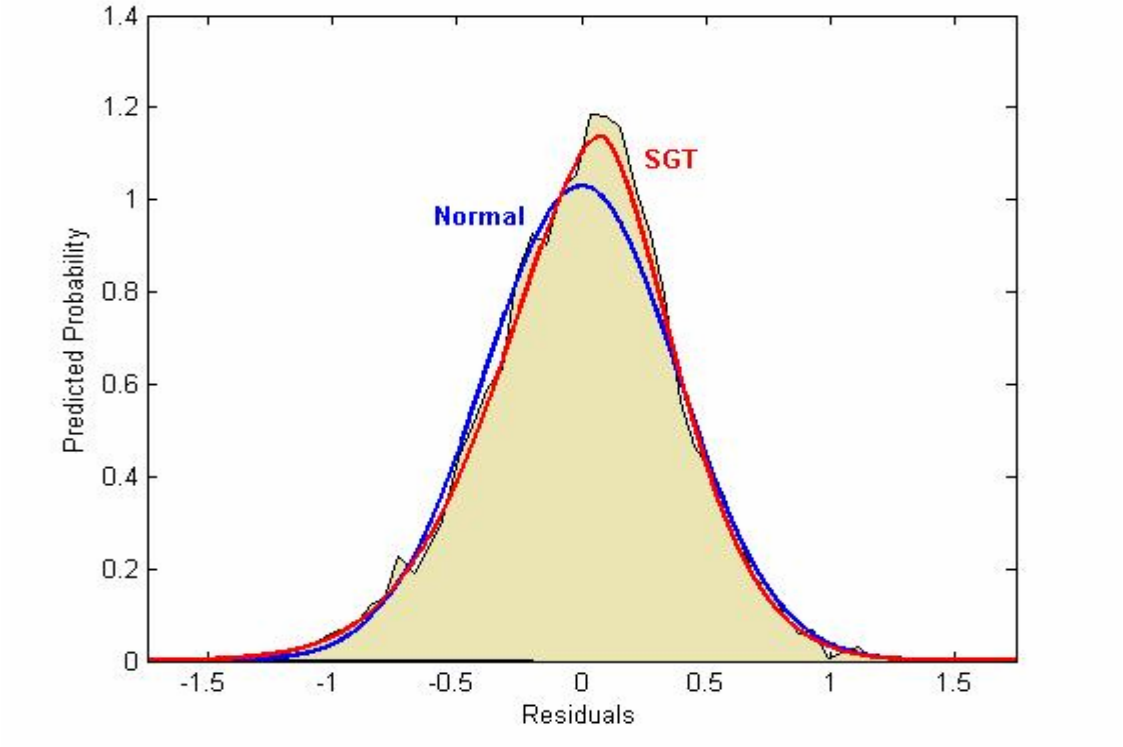

Figure 2. LIML residual distribution from Card data.

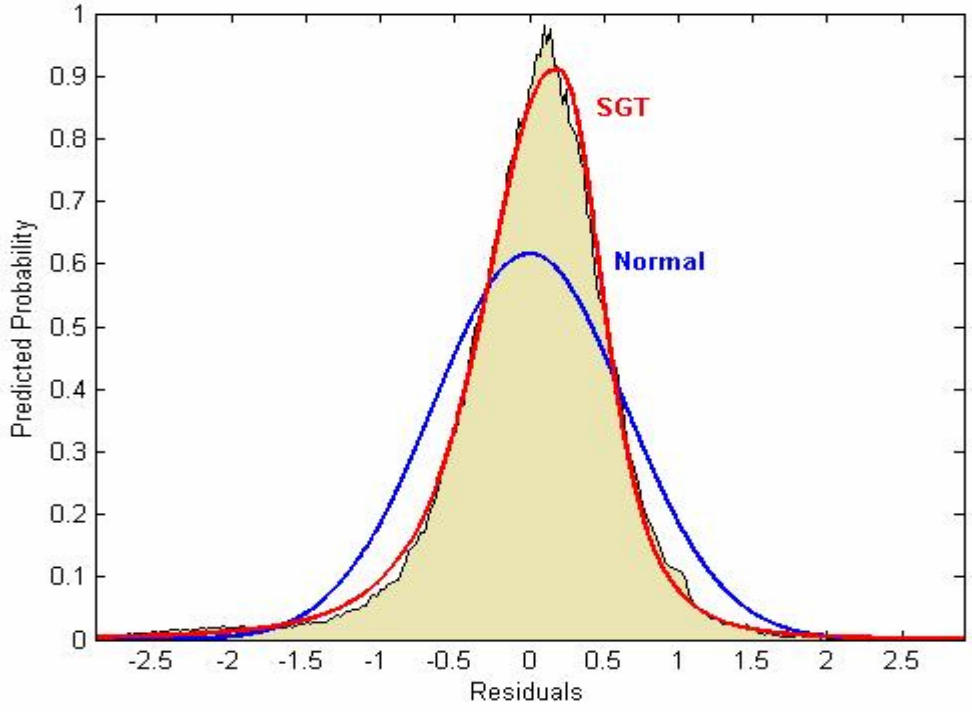

Figure 3. LIML residual distribution from Angrist and Krueger data.