Nonparametric Identification of a Binary Random Factor in Cross Section Data

Yingying Dong and Arthur Lewbel*

California State University Fullerton and Boston College

Original January 2009, revised June 2009

Abstract

Suppose V and U are two independent mean zero random variables, where V has an asymmetric distribution with two mass points and U has a symmetric distribution. We show that the distributions of V and U are nonparametrically identified just from observing the sum $V+U$, and provide a rate root n estimator. We apply these results to the world income distribution to measure the extent of convergence over time, where the values V can take on correspond to country types, i.e., wealthy versus poor countries. We also extend our results to include covariates X, showing that we can nonparametrically identify and estimate cross section regression models of the form $Y = g(X, D^*)+U$, where D^* is an unobserved binary regressor.

JEL Codes: C35

Keywords: Random Effects, Binary, Unobserved Factor, Unobserved Regressor, Income distribution, Income Convergence, Nonparametric identification, Nonparametric Deconvolution

*Corresponding Author: Arthur Lewbel, Department of Economics, Boston College, 140 Commonwealth Avenue, Chestnut Hill, MA, 02467, USA. (617)-552-3678, lewbel@bc.edu,http://www2.bc.edu/~lewbel/

1 Introduction

We propose a method of nonparametrically identifying and estimating cross section regression models that contain an unobserved binary regressor, or equivalently an unobserved random effect that can take on two values. No instruments or proxies for the unobserved binary regressor are observed. Identification is obtained by assuming that the regression function errors are symmetrically distributed, while the distribution of the unobserved random effect is asymmetric. Moment conditions are derived based on these assumptions, and are used to construct either an ordinary generalized method of moments (GMM) estimator, or in the presence of covariates, a nonparametric local GMM estimator for the model.

Ignoring covariates for the moment, suppose $Y = h + V + U$, where V and U are independent mean zero random variables and h is a constant. The random V equals either b_0 or b_1 with unknown probabilities p and $1 - p$ respectively, where p does not equal a half, i.e., V is asymmetrically distributed. U is assumed to have a symmetric distribution. We observe a sample of observations of the random variable Y, and so can identify the marginal distribution of Y, but we do not observe h, V, or U.

We first show that the constant h and the distributions of V and U are nonparametrically identified just from observing Y. The only regularity assumption required is that some higher moments of Y exist.

We also provide estimators for the distributions of V and U. We show that the constant h, the probability mass function of V, moments of the distribution of U, and points of the distribution function of U can all be estimated using GMM. Unlike common deconvolution estimators that can converge at slow rates, we estimate the distributions of V and U, and the density of U (if it is continuous) at the same rates of convergence as if V and U were separately observed, instead of just observing their sum.

We do not assume that the supports of V or U are known, so estimation of the

distribution of V means identifying and estimating both of its support points b_0 and b_1, as well as the probabilities p and $1-p$, respectively, of V equaling b_0 or b_1.

To illustrate these results, we empirically apply them to the world economy convergence issue of whether less developed economies are catching up with richer economies over time. Cross country GDP data in different time periods are used in this application, where p in each time period is an estimate of the fraction of countries that are in the poor group, $b_1 - b_0$ provides a measure of the average difference in GDP between rich and poor countries, and the variance of U is a measure of the dispersion of countries within each group. Decreases in these numbers over time would indicate different forms of income convergence. A feature of these estimates is that they do not require an a priori definition of poor vs. rich, or any assignment of individual countries into the rich or poor groups.

The remainder of the paper then describes how these results can be extended to allow for covariates. If h depends on X while V and U are independent of X, then we obtain the random effects regression model $Y = h(X) + V + U$, which is popular for panel data, but which we identify and estimate just from cross section data.

More generally, we allow both h and the distributions of V and U to depend on X. This is equivalent to nonparametric identification and estimation of a regression model containing an unobserved binary regressor. The regression model is $Y = g(X, D^*) + U$, where g is an unknown function, D^* is an unobserved binary regressor that equals zero with unknown probability $p(X)$ and one with probability $1 - p(X)$, and U is a random error with an unknown symmetric mean zero conditional distribution $F_U(U \mid X)$. The unobserved random variables U and D^* are conditionally independent, conditioning upon X. By defining $h(x) = E(Y \mid X = x) = E[g(X, D^*) \mid X = x]$, $V = g(X, D^*) - h(X)$ and $U = Y - h(X) - V$, this regression model can then be rewritten as $Y = h(X) + V + U$, where $h(x)$ is a nonparametric regression function of Y on X, and the two support points

of V conditional on $X = x$ are then $b_d(x) = g(x, d) - h(x)$ for $d = 0, 1$.

The assumptions this regression model imposes on its error term U are standard, e.g., they hold if the error U is normal, and allow for the error U to be heteroskedastic with respect to X. Also, measurement errors are often assumed to be symmetric and U may be interpreted as measurement error in Y.

One possible application of these extensions is a wage equation, where Y is log wage and D^* indicates whether an individual is of low or high unobserved ability, which could be correlated with some covariates X such as education. This model may show how much wage variation could be explained by unobserved ability.

Another example is a stochastic frontier model, where Y is the log of a firm's output, X are factors of production, and D^* indicates whether the firm operates efficiently at the frontier, or inefficiently. Existing stochastic frontier models obtain identification either by assuming functional forms for the distributions of V and U, or by using panel data and assuming that each firm's individual efficiency level is a fixed effect that is constant over time. See, e.g., Kumbhakar et. al. (2007) and Simar and Wilson (2007). In contrast, with our model one could estimate a nonparametric stochastic frontier model using cross section data, given the restriction that unobserved efficiency is indexed by a binary D^*.

Dong (2008) identifies and estimates a model where $Y = h(X) + V + U$, and applies her results to data where Y is alcohol consumption, and the binary V is an unobserved indicator of health consciousness. Our results formally prove identification of Dong's model, and our estimator is more general in that it allows V and the distribution of U to depend in arbitrary ways on X. Hu and Lewbel (2007) also identify some features of a model containing an unobserved binary regressor. They employ two identification strategies, both of which differ from ours. One of their strategies employs a type of instrumental variable, while the other exploits an assumption of conditional independence

of low order moments, including homoskedasticity. They also use different estimators from ours, and the type of applications they focus on are also different.

Models that allocate individuals into various types, as D^* does, are common in the statistics and marketing literatures. Examples include cluster analysis, latent class analysis, and mixture models (see, e.g., Clogg 1995 and Hagenaars and McCutcheon 2002). Also related is the literature on mismeasured binary regressors, where identification generally requires instruments. An exception is Chen, Hu and Lewbel (2008). Like our Theorem 1 below, they exploit error symmetry for identification, but unlike this paper they assume that the binary regressor is observed, though with some measurement (classification) error, instead of being completely unobserved. A more closely related result is Heckman and Robb (1985), who like us use zero low order odd moments to identify a binary effect, though their's is a restricted effect that is strictly nested in our results. Error symmetry has also been used to obtain identification in a variety of other econometric contexts, e.g., Powell (1986).

There are a few common ways of identifying the distributions of random variables given just their sum. One method of identification assumes that the exact distribution of one of the two errors is known a priori, (e.g., from a validation sample as is common in the statistics literature on measurement error; see, e.g., Carroll, et. al. 2006) and using deconvolution to obtain the distribution of the other one. For example, if U were normal, one would need to know a priori its mean and variance to estimate the distribution of V. A second standard way to obtain identification is to parameterize both the distributions of V and U, as in most of the latent class literature or in the stochastic frontier literature (see, e.g., Kumbhakar and Lovell 2000) where a typical parameterization is to have V be log normal and U be normal. Panel data models often have errors of the form $V + U$ that are identified either by imposing specific error structures or assuming one of the errors is fixed over time (see, e.g., Baltagi 2008 for a survey of random effects and

fixed effects panel data models). Past nonparametric stochastic frontier models have similarly required panel data for identification, as described above. In contrast to all these identification methods, in our model both U and V have unknown distributions, and no panel data are required.

The next section contains our main identification result. We then provide moment conditions for estimating the model, including the distribution of V (its support points and the associated probability mass function), using ordinary GMM. Next we provide estimators for the distribution and density function of U. We empirically apply these results to estimating features of the distribution of per capita GDP across countries and use the results to examine the convergence hypothesis. This is followed by some extensions showing how our identification and estimation methods can be augmented to allow for covariates.

2 Identification

In this section, we first prove a general result about identification of the distribution of two variables given only their sum, and then apply it. Later we extend these results to including regressors X.

ASSUMPTION A1: Assume the distribution of V is mean zero, asymmetric, and has exactly two points of support. Assume $E\left(U^d \mid V\right) = E\left(U^d\right)$ exists for all positive integers $d \leq 9$, and $E\left(U^{2d-1}\right) = 0$ for all positive integers $d \leq 5$.

THEOREM 1: Let Assumption A1 hold, and assume the distribution of Y is identified, where $Y = h + V + U$. Then the constant h and the distributions of U and V are identified.

The proof of Theorem 1 is in the Appendix. Assumption A1 says that the first nine

moments of U conditional on V are the same as the moments that would arise if U were distributed symmetrically and independent of V. Given symmetry of U and an asymmetric, independent, two valued V, by Assumption A1 the only regularity condition required for Theorem 1 is existence of $E(Y^9)$.

Let b_0 and b_1 be the two support points of the distribution of V, where without loss of generality $b_0 < b_1$, and let p be the probability that $V = b_0$, so $1-p$ is the probability that $V = b_1$. We first consider estimation of h, b_0, b_1, and p, and then later show how the rest of the model, i.e., the distribution function of U, can be estimated.

We provide two different sets of moments that can be used for GMM estimation of h, b_0, b_1, and p. The first set of moments is based directly on Theorem 1, while the second provides additional moments that can be used assuming that U is symmetrically distributed.

For the first set of moments, define $u_d = E\left(U^d\right)$ and $v_d = E\left(V^d\right)$. Then $v_1 = E(V) = b_0 p + b_1(1-p) = 0$, so

$$b_1 = \frac{b_0 p}{p-1}, \tag{1}$$

and therefore,

$$v_d = E\left(V^d\right) = b_0^d p + \left(\frac{b_0 p}{p-1}\right)^d (1-p). \tag{2}$$

Now expand the expression $E\left[(Y-h)^d - (V+U)^d\right] = 0$ for integers d, noting by Assumption A1 that the first five odd moments of U are zero. The results are

$$E(Y-h) = 0 \tag{3}$$

$$E\left((Y-h)^2 - (v_2+u_2)\right) = 0 \tag{4}$$

$$E\left((Y-h)^3 - v_3\right) = 0 \tag{5}$$

$$E\left((Y-h)^4 - (v_4 + 6v_2u_2 + u_4)\right) = 0 \tag{6}$$

$$E\left((Y-h)^5-(v_5+10v_3u_2)\right)=0 \tag{7}$$

$$E\left((Y-h)^6-(v_6+15v_4u_2+15v_2u_4+u_6)\right)=0 \tag{8}$$

$$E\left((Y-h)^7-(v_7+21v_5u_2+35v_3u_4)\right)=0 \tag{9}$$

$$E\left((Y-h)^9-(v_9+36v_7u_2+126v_5u_4+84v_3u_6)\right)=0 \tag{10}$$

Substituting equation (2) into equations (3) to (10) gives eight moments in the six unknown parameters h, b_0, p, u_2, u_4, and u_6. The proof of Theorem 1 shows that these eight equations uniquely identify these parameters. As shown in the proof, more equations than unknowns are required for identification because of the nonlinearity of these equations, and in particular the presence of multiple roots. Given estimates of these parameters, an estimate of b_1 is obtained by equation (1).

Another set of conditional moments that can be used for estimation are given by the following Corollary.

COROLLARY 1: Let Assumption A1 hold. Assume U is symmetrically distributed and is independent of V. Assume $E\left[\exp\left(TU\right)\right]$ exists for some positive constant T. Then the following equation holds for all positive $\tau \leq T$,

$$E\left(\frac{\exp\left[\tau\left(Y-h\right)\right]}{p\exp\left(\tau b_0\right)+\left(1-p\right)\exp\left(\tau\frac{b_0p}{p-1}\right)}-\frac{\exp\left[-\tau\left(Y-h\right)\right]}{p\exp\left(-\tau b_0\right)+\left(1-p\right)\exp\left(-\tau\frac{b_0p}{p-1}\right)}\right)=0. \tag{11}$$

By choosing a large number of values of $\tau \leq T$, Corollary 1 provides a large number of additional moment conditions satisfied by the parameters h, b_0, and p. Estimation could be based on equations (3) to (10) (after substituting in equation (2)), or on moments given by equations (3) and (11) for some set of positive values of $\tau \leq T$, or on a combination of both sets of moments. In some simulations (see also Dong 2008) we found that equation (3) along with equation (11) letting τ be about a dozen equally spaced

values between 1 and 2.5 sufficed for identification and yielded reasonable estimates, though formally we have only proved that identification follows from moments (3) to (10), with moments based on equation (11) providing overidentifying information.

3 Estimation

Estimation takes the form of the standard Generalized Method of Moments (GMM, as in Hansen 1982), since given data Y_1,...Y_n, we have a set of moments of the form $E\left[G\left(Y,\theta\right)\right]=0$, where G is a set of known functions and θ is the vector of parameters h, b_0, p and also includes u_2, u_4, and u_6 if equations (4) to (10) (after substituting in equation (2)) are included in the set of moments G. Note that while all of the above assumes Y_1, ...,Y_n are identically distributed, they do not need to be independent, as GMM estimation theory permits some serial dependence in the data. To save space we do not write out the detailed assumptions and associated limiting distribution theory for these GMM estimators, which can be found in standard textbooks,

Estimation based on the first set of moments (3) to (10) entails estimation of the additional parameters u_2, u_4, and u_6, which in practice could be of direct interest. Note that it also depends on high order moments, which may be heavily influenced by outliers. Removing outliers from the Y data (which can be interpreted as robustifying higher moment estimation) and rescaling Y could be useful numerically.

The second set of moments given by equations (3) and (11) for some set of positive values of τ do not automatically follow from Theorem 1. However, they contain potentially many more moments for estimation. They can be of lower order and do not require estimation of the nuisance parameters u_2, u_4, and u_6. Both sets of moments can be combined into a single GMM estimator if desired.

Standard GMM limiting distribution theory provides root n consistent, asymptot-

ically normal estimates of h, and of the distribution of V, (i.e., the support points b_0 and b_1 and the probability p, where $\widehat{b}_1$ is obtained by $\widehat{b}_1 = \widehat{b}_0\widehat{p}/(\widehat{p}-1)$ from equation 1). We define b_0 as the smaller of the two support points of V. This along with $E(V) = 0$ requires that $\widehat{b}_0$ be negative, which may be imposed in estimation.

4 The Distribution of U

As noted in the proof of Theorem 1, once the distribution of V is recovered, then the distribution of U is identified by a deconvolution, in particular we have that the characteristic function of U is identified by $E\left(e^{i\tau U}\right) = E\left(e^{i\tau(Y-h)}\right)/E\left(e^{i\tau V}\right)$, where i denotes the square root of -1. However, under the assumption that U is symmetrically distributed, the following theorem provides a more convenient way to estimate the distribution function of U. For any random variable Z, let F_Z denote the marginal cumulative distribution function of Z. Also define $\varepsilon = V + U$ and define

$$\Psi(u) = \frac{[F_\varepsilon(-u+b_0)-1]\,p + F_\varepsilon(u+b_1)(1-p)}{1-2p}. \tag{12}$$

THEOREM 2: Let Assumption A1 hold. Assume U is symmetrically distributed. Then

$$F_U(u) = \frac{\Psi(u) + \Psi(-u)}{2}. \tag{13}$$

Theorem 2 provides a direct expression for the distribution of U in terms of b_0, b_1, p and the distribution of ε, all of which are previously identified. Let $I(\cdot)$ denote the indicator function that equals one if $\cdot$ is true and zero otherwise. Then using $Y = h + \varepsilon$ it follows immediately from equation (12) that

$$\Psi(u) = E\left(\frac{[I(Y \le h-u+b_0)-1]\,p + I(Y \le h+u+b_1)(1-p)}{1-2p}\right). \tag{14}$$

An estimator for $F_U(u)$ can then be constructed by replacing the parameters in equation (14) with estimates, replacing the expectation with a sample average, and plugging the result into equation (13), that is, define

$$\begin{aligned}\omega(Y,u,\theta) &= [I(Y \leq h-u+b_0)-1]\,p + I(Y \leq h+u+b_1)(1-p) \\ &\quad + [I(Y \leq h+u+b_0)-1]\,p + I(Y \leq h-u+b_1)(1-p),\end{aligned} \tag{15}$$

where θ contains h, b_0, b_1, and p. Then the estimator corresponding to equation (13) is

$$\widehat{F}_U(u) = \frac{1}{n}\sum_{i=1}^{n} \frac{\omega\left(Y_i,u,\widehat{\theta}\right)}{2-4\widehat{p}}. \tag{16}$$

Alternatively, $F_U(u)$ for a finite number of values of u, say $u_1,...,u_J$, can be estimated as follows. Recall that $E[G(Y,\theta)]=0$ was used to estimate the parameters h, b_0, b_1, p by GMM. For notational convenience, let $\eta_j = F_U(u_j)$ for each u_j. Then by equations (13) and (14),

$$E\left[(2-4p)\,\eta_j - \omega(Y,u_j,\theta)\right] = 0. \tag{17}$$

Adding equation (17) for $j=1,...,J$ to the set of functions defining G, including $\eta_1,...,\eta_J$ in the vector θ, and then applying GMM to this augmented set of moment conditions $E[G(Y,\theta)]=0$ simultaneously yields root n consistent, asymptotically normal estimates of h, b_0, b_1, p and $\eta_j = F_U(u_j)$ for $j=1,...,J$. An advantage of this approach versus equation (16) is that GMM limiting distribution theory then provides standard error estimates for each $\widehat{F}_U(u_j)$.

While p is the unconditional probability that $V=b_0$, given $\widehat{F}_U$ it is straightforward to estimate conditional probabilities as well. In particular,

$$\Pr(V=b_0 \mid Y \leq y) = \Pr(V=b_0, Y \leq y) / \Pr(Y \leq y)$$

$$= \quad F_U\left(y-h-b_0\right)/F_y\left(y\right)$$

which could be estimated as $\widehat{F}_U\left(y-\widehat{h}-\widehat{b}_0\right)/\widehat{F}_y\left(y\right)$ where $\widehat{F}_y$ is the empirical distribution of Y.

Let f_Z denote the probability density function of any continuously distributed random variable Z. So far no assumption has been made about whether U is continuous or discrete. However, if U is continuous, then ε and Y are also continuous, and then taking the derivative of equations (12) and (13) with respect to u gives

$$\psi\left(u\right)=\frac{-f_\varepsilon\left(-u+b_0\right)p+f_\varepsilon\left(u+b_1\right)\left(1-p\right)}{1-2p},\qquad f_U\left(u\right)=\frac{\psi\left(u\right)+\psi\left(-u\right)}{2}, \tag{18}$$

which suggests the estimators

$$\widehat{\psi}\left(u\right)=\frac{-\widehat{f}_\varepsilon\left(-u+\widehat{b}_0\right)\widehat{p}+\widehat{f}_\varepsilon\left(u+\widehat{b}_1\right)\left(1-\widehat{p}\right)}{1-2\widehat{p}}, \tag{19}$$

$$\widehat{f}_U\left(u\right)=\frac{\widehat{\psi}\left(u\right)+\widehat{\psi}\left(-u\right)}{2}, \tag{20}$$

where $\widehat{f}_\varepsilon\left(\varepsilon\right)$ is a kernel density or other estimator of $f_\varepsilon\left(\varepsilon\right)$, constructed using data $\widehat{\varepsilon}_i=Y_i-\widehat{h}$ for $i=1,...n$. Since densities converge at slower than rate root n, the limiting distribution of this estimator will be the same as if $\widehat{h}$, $\widehat{b}_0$, $\widehat{b}_1$, and $\widehat{p}$ were evaluated at their true values. The above $\widehat{f}_U\left(u\right)$ is just the weighted sum of kernel density estimators, each one dimensional, and so under standard regularity conditions will converge at the optimal one dimensional pointwise rate $n^{2/5}$. Note that it is possible for $\widehat{f}_U\left(u\right)$ to be negative in finite samples, so if desired one could replace negative values of $\widehat{f}_U\left(u\right)$ with zero.

A numerical problem that can arise is that equation (19) may require evaluting $\widehat{f}_\varepsilon$

at a value that is outside the range of observed values of $\widehat{\varepsilon}_i$. Since both $\widehat{\psi}(u)$ and $\widehat{\psi}(-u)$ are consistent estimators of $\widehat{f}_U(u)$ (though generally less precise than equation (20) because they individually ignore the symmetry constraint), one could use either $\widehat{\psi}(u)$ or $\widehat{\psi}(-u)$ instead of their average to estimate $\widehat{f}_U(u)$ whenever $\widehat{\psi}(-u)$ or $\widehat{\psi}(u)$, respectively, requires evaluating $\widehat{f}_\varepsilon$ at a point outside the the range of observed values of $\widehat{\varepsilon}_i$.

5 A Parametric U Comparison

It might be useful to construct parametric estimates of the model, which could for example provide reasonable starting values for the GMM estimation. The parametric model we propose for comparison assumes that U is normal with mean zero and standard deviation s.

When U is normal the distribution of Y is finitely parameterized, and so can be estimated directly by maximum likelihood. The log likelihood function is given by

$$\sum_{i=1}^{n} \ln\left(\frac{p}{s\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{Y_i-h-b_0}{s}\right)^2\right]+\frac{1-p}{s\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{Y_i-h-\frac{b_0 p}{p-1}}{s}\right)^2\right]\right). \quad (21)$$

Maximizing this log likelihood function provides estimates of h, b_0, p, and s. As before, an estimate of b_1 would be given by $\widehat{b}_1 = \widehat{b}_0\widehat{p}/(\widehat{p}-1)$. Further, if U is normal then $u_2 = s^2$, $u_4 = 3s^2$, and $u_6 = 15s^2$. These estimates can be compared to the GMM estimates, which should be the same if the true distribution of U is indeed normal.

6 An Empirical Application: World Income Distribution

A large literature exists regarding the distribution of income across countries, much of which deals with the question of convergence, that is, whether poorer countries are catching up with richer countries as a result of increases in globalization of trade and diffusion of technology.

To measure the extent of convergence, if any, we propose a simple descriptive model of the income distribution across countries. Assume that there exist two types of countries, i.e., poor versus rich, or less developed versus more developed countries. Let I_{ti} denote the per capita income or GDP of country i in time t, and define Y_{ti} to be either income levels $Y_{ti} = I_{ti}$, or income shares $Y_{ti} = I_{ti} / \left(\sum_{i=1}^{n} I_{ti}\right)$. Assume that a poor country's income in year t is given by $Y_{ti} = g_{t0} + U_{ti}$, while that of a wealthy country is given by $Y_{ti} = g_{t1} + U_{ti}$, where g_{t0} and g_{t1} are the mean income levels or mean shares for poor and rich countries, respectively, and U_{ti} is an individual country's deviation from its group mean. Here U_{ti} embodies both the relative ranking of country i within its (poor or rich) group, and may also include possible measurement errors in Y_{ti}. We assume that the distribution of U_{ti} is symmetric and mean zero with a probability density function f_{tu}.

Let $h_t = E_t(Y)$ be the mean income or income share for the whole population of countries in year t. Then the income measure for country i in year t can be rewritten as $Y_{ti} = h_t + V_{ti} + U_{ti}$, where V_{ti} is the deviation of rich or poor countries' group mean from the grand mean h_t. Then V_{ti} equals $b_{t0} = g_{t0} - h_t$ with probability p_t and V_{ti} equals $b_{t1} = g_{t1} - h_t$ with probability $1 - p_t$, so p_t is the fraction of countries that are in the poor group in year t, and $b_{t1} - b_{t0}$ is the difference in mean income or income shares between poor and wealthy countries.

This simple model provides measures of a few different possible types of convergence.

Having p_t decrease over time would indicate that on average countries are leaving the poor group and joining the set of wealthy nations. A finding that $b_{t1} - b_{t0}$ decreases over time would mean that the differences between rich and poor nations are diminishing, and a finding that the spread (e.g. the variance) of the density f_{tu} decreases over time would mean that there is convergence within but not necessarily across the poor and rich groups.

A feature of this model is that it does not require arbitrarily choosing a threshold level of Y to demarcate the line between rich and poor countries, and so avoids this potential source of misspecification. This model also allows for the possibility that a poor country has higher income than some wealthy country in a given time period due to random factors (e.g., natural disaster in a wealthy country i, implying a low draw of U_{ti} in time t). More generally, the model does not require specifying or estimating the group to which each country belongs.

Bimodality versus unimodality of Y may be interpreted as evidence in favor of this 'two group' model, though note that even if U is unimodal, e.g., normal, then Y can be either unimodal or bimodal (with possibly large differences in the heights of the two modes), depending on p and on the magnitudes of b_0 and b_1. The density for Y can also be quite skewed, even though U is symmetric.

Bianchi (1997) applies bimodality tests to the distribution of income across countries over time, to address questions regarding evidence for convergence. For comparison we apply our model using the same data as Bianchi's, which consists of I_{it} defined as annual per capita GDP in constant U.S. dollars for 119 countries, measured in 1970, 1980 and 1989.

For each of the three years of data we provide four different estimates: GMM1, which is GMM based on moments (3) to (10) (after substituting in equation (2)); GMM2, which is GMM based on moments given by equations (3) and (11) where τ takes 11 equally

spaced values between 0.19 and 2.09;[1] GMM3, which uses both these sets of moments, and MLE, which is the maximum likelihood estimator that maximizes (21), assuming that U is normal.

Table 1: Estimates based on the GDP per capita level data (in 10,000 1985 dollars)

		p	b0	b1	b1-b0	h	u2	u4	u6
1970	GMM1	.8575	-.1105	.6648	.7753	.3214	.0221	.0001$^{\$}$	.0024
		(.0352)	(.0244)	(.0664)	(.0590)	(.0284)	(.0042)	(.0002)	(.0009)
	GMM2	.8605	-.1089	.6719	.7808	.3213			
		(.0300)	(.0199)	(.0656)	(.0601)	(.0246)			
	GMM3	.8573	-.1110	.6667	.7777	.3214	.0210	.0012	.012*
		(.0329)	(.0230)	(.0644)	(.0584)	(.0287)	(.0040)	(.0004)	(.0005)
	MLE	.8098	-.1334	.5679	.7013	.3213	.0199		
		(.0362)	(.0260)	(.0487)	(.0477)	(.0280)	(.0031)		
1980	GMM1	.8081	-.1722	.7252	.8974	.4223	.0294	.0016	.0017*
		(.0371)	(.0322)	(.0579)	(.0491)	(.0351)	(.0043)	(.0004)	(.0007)
	GMM2	.8129	-.1684	.7316	.900	.4222			
		(.0315)	(.0267	(.0560)	(.0493)	(.0283)			
	GMM3	.8068	-.1742	.7275	.9017	.4221	.0277	.0010$^{\$}$	.0025*
		(.0396)	(.0354)	(.0676)	(.0634)	(.0371)	(.0050)	(.0148)	(.0011)
	MLE	.8070	-.1692	.7077	.8769	.4222	.0350		
		(.0393)	(.0345)	(.0600)	(.0544)	(.0372)	(.0048)		
1989	GMM1	.8125	-.2114	.9159	1.1273	.4804	.0384	.0051	.0028$^{\$}$
		(.0380)	(.0424)	(.1022)	(.1111)	(.0439)	(.0118)	(.0104)	(.0448)
	GMM2	.8183	-.2049	.9230	1.1279	.4806			
		(.0283)	(.0293)	(.0683)	(.0612)	(.0322)			
	GMM3	.8123	-.2120	.9176	1.1296	.4804	.0373	.0051	.0027$^{\$}$
		(.0374)	(.0410)	(.0758)	(.0697)	(.444)	(.0068)	(.0016)	(.0452)
	MLE	.7948	-.2192	.8491	1.0683	.4805	.0489		
		(.0393)	(.0413)	(.0754)	(.0679)	(.0441)	(.0076)		

Note: $^{\$}$ not significant; * significant at the 5% level; all the others are significant at the 1% level. Standard errors are in parentheses.

Table 1 reports results based on per capita levels, $Y_{ti} = I_{ti}/10,000$, while Table 2 is based on scaled shares, $Y_{ti} = 50I_{ti}/\left(\sum_{i=1}^{n} I_{ti}\right)$.[2] For each data set, estimates based on all three GMM estimators are quite similar, with estimates of p, b_0, and b_1 across the

[1] We found our results were relatively insensitive to the exact range and number of values of τ used.

[2] We scale by 10,000 or by 50 to put the Y_{ti} data in a range between zero and two. Such scalings helped ensure that the matrices involved in estimation (e.g., the estimated weighting matrix used for efficiency in the the second stage of GMM) were numerically well conditioned.

Table 2: Estimates based on the scaled GDP per capita share data

		p	b0	b1	b1-b0	h	u2	u4	u6
1970	GMM1	.8619	-.1392	.8682	1.0074	.4206	.0417	.0039$	.0057$
		(.0361)	(.0332)	(.1009)	(.0985)	(.0380)	(.0089)	(.0068)	(.0063)
	GMM2	.8640	-.1392	.8844	1.0236	.4202			
		(.0291)	(.0241)	(.0909)	(.0816)	(.0292)			
	GMM3	.8579	-.1448	.8737	1.0185	.4203	.0347	.0044$	.0053$
		(.0348)	(.0319)	(.0998)	(.0976)	(.0361)	(.0085)	(.0281)	(.0246)
	MLE	.8098	-.1744	.7425	9169	4202	.0340		
		(.0383)	(.0352)	(.0670)	(.0629)	(.0377)	(.0053)		
1980	GMM1	.8080	-.1715	.7217	.8932	.4202	.0291	.0016	.0017
		(.0374)	(.0334)	(.0560)	(.0497)	(.0364)	(.0041)	(.0004)	(.0006)
	GMM2	.8128	-.1676	.7280	8956	4202			
		(.0323)	(.0274)	(.0565)	(.0496)	(.0304)			
	GMM3	.8067	-.1734	.7240	.8974	.4200	.0274	.0009	.0025*
		(.0364)	(.0322)	(.0551)	(.0483)	(.0359)	(.0041)	(.0006)	(.0011)
	MLE	.8070	-.1684	.7043	8727	4202	.0347		
		(.0373)	(.0322)	(.0570)	(.0508)	(.0353)	(.0045)		
1989	GMM1	.8117	-.1848	.7964	.9812	.4203	.0316	.0023	.0020*
		(.0360)	(.0344)	(.0609)	(.0518)	(.0388)	(.0049)	(.0007)	(.0009)
	GMM2	.8167	-.1808	.8056	.9864	.4202			
		(.0311)	(.0276)	(.0659)	(.0570)	(.0301)			
	GMM3	.8106	-.1870	.8002	.9872	.4201	.0288	.0024$	.0021$
		(.0365)	(.0355)	(.0707)	(.0680)	(.0382)	(.0055)	(.0242)	(.0015)
	MLE	.7948	-.1916	.7424	934	4202	.0374		
		(.0387)	(.0355)	(.0655)	(.0589)	(.0395)	(.0058)		

Note: $ not significant; *significant at the 5% level; all the others are significant at the 1% level. Standard errors are in parentheses.

GMM estimators all within 2% of each other. The maximum likelihood estimates for these parameters are also roughly comparable.

Looking across years, both Tables 1 and 2 tell similar stories in terms of percentages of poor countries. Using either levels or shares, by GMM p is close to .86 in 1970, and close to .81 in 1980 and 1989, showing a decline in the number of poor countries in the 1970's, but no further decline in the 1980's (MLE shows p close to .81 in all years). The average difference between rich and poor, $b_1 - b_0$, increases steadily over time in the levels data, but this may be due in part to the growth of average income over time,

given by h. Share data takes into account this income growth over time. Estimates based on shares in Table 2 show that $b_1 - b_0$ decreased by a small amount in the 1970's, but then increased again in the 1980's, so by this measure there is no clear evidence of convergence or divergence.

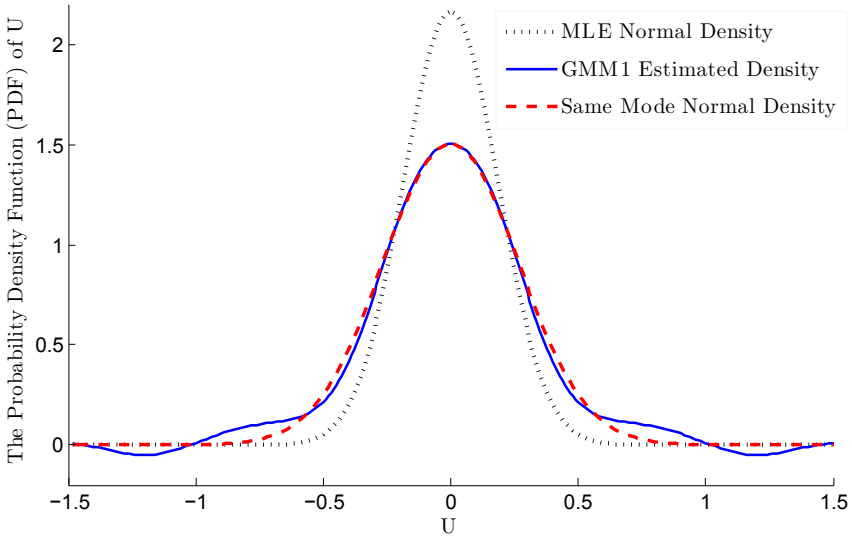

Figure 1: The estimated probability density function of U, using 1970 share data

Figure 1 shows $\widehat{f}_u$, the estimated density of U, given by equation (20) using the GMM1 estimates from Table 2 in 1970.[3] This estimated density is compared to a normal density with the same mode, $\widehat{f}_u(0)$. It follows that this normal density has standard deviation $(2\pi)^{-1/2}\left[\widehat{f}_u(0)\right]^{-1}$. With the same central tendency given by construction, these two densities can be compared for differences in dispersion and tail behaviors. As Figure 1 shows, the semiparametric $\widehat{f}_u$ matches the normal density rather closely except near the tails of its distribution where data are sparse. Also shown in Figure 1 is the maximum likelihood estimate of f_u, which assumes U is normal. Although close to normal in shape, the semiparametric $\widehat{f}_u$ appears to have a larger variance than the maximum likelihood estimate. The graphs of $\widehat{f}_u$ in other years are very similar, and they

[3]Graphs of other years and other GMM estimates are very similar, so to save space we do not include them here.

along with the variance estimates in Table 2 show no systematic trends in the dispersion of U over time, and hence no evidence of income convergence within groups of rich or poor countries.

In this analysis of U, note that Y is by construction nonnegative so U cannot literally be normal; however, the value of U where $Y = h + V + U$ crosses zero is far out in the left tail of the U distribution (beyond the values graphed in Figure 1), so imposing the constraint on U that Y be nonnegative (e.g., making the parametric comparison U a truncated normal) would have no discernable impact on the resulting estimates.

In addition to analyzing levels I_{ti} and shares $I_{ti}/\left(\sum_{i=1}^{n} I_{ti}\right)$, Bianchi (1997) also considers logged data, but finds that the logarithmic transformation changes the shape of the Y_{ti} distribution in a way that obscures any bimodality. We found similar results, in that with logged data our model yields estimates of p close to .5, which is basically ruled out by our model, as $p = .5$ would make V be symmetric and hence unidentifiable relative to U.

7 Extension 1: h depends on covariates

We now consider some extensions of our main results. The first extension allows h to depend on covariates X. Estimators associated with this extension will take the form of standard two step estimators with a uniformly consistent first step, so after showing identification we will omit technical details regarding estimator assumptions to save space.

COROLLARY 2: Assume the conditional distribution of Y given X is identified and its mean exists. Let $Y = h(X) + V + U$. Let Assumption A1 hold. Assume V and U are independent of X. Then the function $h(X)$ and distributions of U and V are identified.

Corollary 2 extends Theorem 1 by allowing the conditional mean of Y to nonparametrically depend on X. Given the assumptions of Corollary 2, it follows immediately that equations (3) to (10) hold replacing h with $h(X)$, and if U is symmetrically distributed and independent of V and X then equation (11) also holds replacing h with $h(X)$. This suggests two ways of extending the GMM estimators of the previous section. One method is to first estimate $h(X)$ by a uniformly consistent nonparametric mean regression of Y on X (e.g., a kernel regression), then replace $Y - h$ in equations (3) to (10) and/or equation (11) with $\varepsilon = Y - h(X)$, and apply ordinary GMM to the resulting moment conditions (using as data $\widehat{\varepsilon}_i = Y_i - \widehat{h}(X_i)$ for $i = 1, ..., n$) to estimate the parameters b_0, b_1, p, u_2, u_4, and u_6. Consistency of this estimator follows immediately from the uniform consistency of $\widehat{h}$ and ordinary consistency of GMM. This estimator is easy to implement because it only depends on ordinary nonparametric regression and ordinary GMM, but note that the usual standard error formulas from the second step GMM will not be correct because they do not account for the first stage estimation error in h.

An alternative estimator is to note that, given the assumptions of Corollary 2, equations (3) to (10) and/or equation (11) (the latter assuming symmetry and independence of U) hold by replacing h with $h(X)$ and replacing the unconditional expectations in these equations with conditional expectations, conditioning on $X = x$. The resulting set of equations can be written as $E[G(Y, \theta, h(X)) \mid X = x] = 0$ where G is a set of known functions and θ is the vector of parameters b_0, b_1, p, and also includes u_2, u_4, and u_6 if equations (4) to (10) (after substituting in equation (2)) are included in the set of moments G. This is now in the form of conditional GMM given by Ai and Chen (2003), who provide a Sieve estimator and associated limiting distribution theory.

After replacing $\widehat{h}$ with $\widehat{h}(X_i)$, equation (16) can be used to estimate the distribution of U, or alternatively equation (17) for $j = 1, ..., J$, replacing h with $h(X)$, can be

included in the set of functions defining G in the conditional GMM estimator. The estimator (20) will still work for estimating the density of U if it is continuous, using as before data $\widehat{\varepsilon}_i = Y_i - \widehat{h}(X_i)$ for $i = 1, ..., n$ to estimate the density function f_ε.

If desired, this model can be easily compared to a semiparametric specification where U is normal while $h(X)$ is unknown. In this case the first step would still be to construct an estimate $\widehat{h}(X)$ by a nonparametric regression of Y on X, and then $Y_i - h$ in the likelihood function (21) would be replaced by $Y_i - \widehat{h}(X_i)$ and the result maximized over b_0, p, and s to estimate those parameters.

8 Extension 2: Nonparametric regression with an Unobserved Binary Regressor

This section extends previous results to a more general nonparametric regression model of the form $Y = g(X, D^*) + U$. Specifically, we have the following corollary.

COROLLARY 3: Assume the joint distribution of Y, X is identified, and that $g(X, D^*) = E(Y \mid X, D^*)$ exists, where D^* is an unobserved variable with support $\{0, 1\}$. Define $p(X) = E(1 - D^* \mid X)$ and define $U = Y - g(X, D^*)$. Assume $E\left(U^d \mid X, D^*\right) = E\left(U^d \mid X\right)$ exists for all integers $d \leq 9$ and $E\left(U^{2d-1} \mid X\right) = 0$ for all positive integers $d \leq 5$. Then the functions $g(X, D^*)$, $p(X)$, and the distribution of U are identified.

Corollary 3 permits all of the parameters of the model to vary nonparametrically with X. It provides identification of the regression model $Y = g(X, D^*) + U$, allowing the unobserved model error U to be heteroskedastic (and have nonconstant higher moments as well), though the variance and other low order even moments of U can only depend on X and not on the unobserved regressor D^*. As noted in the introduction and in the proof of this Corollary, $Y = g(X, D^*) + U$ is equivalent to $Y = h(X) + V + U$ but, unlike

Corollary 2, now V and U have distributions that can depend on X. As with Theorem 1, symmetry of U (now conditional on X) suffices to make the required low order odd moments of U be zero.

Given the assumptions of Corollary 3, equations (3) to (10), and given symmetry of U, equation (11), will all hold after replacing the parameters h, b_0, b_1, p, u_2, u_4, and u_6 with functions $h(X)$, $b_0(X)$, $b_1(X)$, $p(X)$, $u_2(X)$, $u_4(X)$, and $u_6(X)$ and replacing the unconditional expectations in these equations with conditional expectations, conditioning on $X = x$. We can further replace $b_0(X)$ and $b_1(X)$ with $g(x,0) - h(x)$ and $g(x,1) - h(x)$, respectively, to directly obtain estimates of the function $g(X, D^*)$ instead of $b_0(X)$ and $b_1(X)$.

Let $q(x)$ be the vector of all of the above listed unknown functions. Then these conditional expectations can be written as

$$E[G(q(x), Y) \mid X = x)] = 0 \tag{22}$$

for a vector of known functions G. Equation (22) is again in the form of conditional GMM which could be estimated using Ai and Chen (2003), replacing all of the unknown functions $q(x)$ with sieves (related estimators are Carrasco and Florens, 2000 and Newey and Powell, 2003). However, given independent, identically distributed draws of X, Y, the local GMM estimator of Lewbel (2008) may be easier to use because it exploits the special structure we have here where all the functions $q(x)$ to be estimated depend on the same variables that the moments are conditioning on, $X = x$.

We summerize here how this estimator could be implemented, while Appendix B provides details regarding the associated limiting distribution theory. Note that this estimator can be used when X contains both continuous and discretely distributed elements. If all elements of X are discrete, then the estimator can again be simplified to

Hansen's (1982) original GMM, as described in Appendix B.

1. For any value of x, construct data $Z_i = K(x - X_i)/b$ for $i = 1, ..., n$, where K is an ordinary kernel function (e.g., the standard normal density function) and b is a bandwidth parameter.[4]

2. Obtain $\widehat{\theta}$ by applying standard two step GMM based on the moment conditions $E(G(\theta, Y)Z) = 0$ for G from equation (22).

3. For the given value of x, let $\widehat{q}(x) = \widehat{\theta}$.

4. Repeat these steps for every value of x for which one wishes to estimate the vector of functions $q(x)$. For example, one may repeat these steps for a fine grid of x points on the support of X, or repeat these steps for x equal to each data point X_i to just estimate the functions $q(x)$ at the observed data points.

For comparison, one could also estimate a semiparametric specification where U is normal but all parameters of the model still vary with x. Analogous to the local GMM estimator, this comparison model could be estimated by applying the local GMM estimator described in Appendix B to moment conditions defined as the derivatives of the expected value of log likelihood function (21) with respect to the parameters, that is, using the likelihood score functions as moments.

9 Discrete V With More Than Two Support Points

A simple counting argument suggests that it may be possible to extend this paper's identification and associated estimators to applications where V is discrete with more than two points of support, as follows. Suppose V takes on the values b_0, b_1, ..., b_K with probabilities p_0, p_1,..., p_K. Let $u_j = E(U^j)$ for integers j as before. Then for any positive odd integer S, the moments $E(Y^s)$ for $s = 1$, ..., S equal known functions of the

[4]As is common practice when using kernel functions, it is a good idea to first standardize the data by scaling each continuous element of X by its sample standard deviation.

$2K + (S+1)/2$ parameters b_1, b_2,..., b_K, p_1, p_2, ...,p_K, u_2, u_4, ..., u_{S-1}, h.[5] Therefore, with any odd $S \geq 4K+1$, $E(Y^s)$ for $s = 1, ..., S$ provides at least as many moment equations as unknowns, which could be used to estimate these parameters by GMM. These moments include polynomials with up to $S-1$ roots, so having S much larger than $4K+1$ may be necessary for identification, just as the proof of Theorem 1 requires $S = 9$ even though in that theorem $K = 1$. Still, as long as U has sufficiently thin tails, $E(Y^s)$ can exist for arbitrarily high integers s, thereby providing far more identifying equations than unknowns.

The above analysis is only suggestive. Given how long the proof is for our model where V takes on only two values, we do not provide a proof of identification with more than two points of support. However, assuming a model where V takes on more than two values is identified, the moment conditions for estimation analogous to those we provided earlier are readily available. For example, as in the proof of Corollary 1 it follows from symmetry of U that

$$E\left(\frac{\exp(\tau(Y-h))}{\exp(\tau V)} - \frac{\exp(-\tau(Y-h))}{\exp(-\tau V)}\right) = 0$$

for any τ for which these expectations exist, and therefore GMM estimation could be based on the moments

$$E\left[\sum_{k=0}^{K}\left(\frac{\exp(\tau(Y-h))}{\exp(\tau b_k)} - \frac{\exp(-\tau(Y-h))}{\exp(-\tau b_k)}\right)p_k\right] = 0$$

for a large number of different values of τ.

[5]Here p_0 and b_0 can be expressed as functions of the other parameters by probabilities summing to one and V having mean zero. Also u_s for odd values of s are zero by symmetry of U.

10 Conclusions

We have provided identification and estimators for $Y = h + V + U$, $Y = h(X) + V + U$, and more generally for $Y = g(X, D^*) + U$. In these models, D^* or V are unobserved regressors with two points of support, and the unobserved U is drawn from an unknown symmetric distribution. No instruments, measures, or proxies for D^* or V are observed. To illustrate the results, we apply our basic model to the distribution of income across countries, where the two values V can take on correspond to country types such as more developed versus less developed countries. The estimates from this model provide some summary measures for assessing whether income convergence has taken place over time.

Interesting work for the future could include derivation of semiparametric efficiency bounds for the model, and conditions for identification when V can take on more than two values.

References

[1] Ai, C. and X. Chen (2003), "Efficient Estimation of Models With Conditional Moment Restrictions Containing Unknown Functions," Econometrica, 71, 1795-1844.

[2] Baltagi, B. H. (2008), Econometric Analysis of Panel Data, 4th ed., Wiley.

[3] Bianchi, M. (1997), "Testing for Convergence: Evidence from Nonparametric Multimodality Tests," Journal of Applied Econometrics, 12, 393-409.

[4] Carrasco, M. and J. P. Florens (2000), "Generalization of GMM to a Continuum of Moment Conditions," Econometric Theory, 16, 797-834.

[5] Carroll, R. J., D. Ruppert, L. A. Stefanski, and C. M. Crainiceanu, (2006), Measurement Error in Nonlinear Models: A Modern Perspective, 2nd edition, Chapman & Hall/CRC.

[6] Chen, X., Y. Hu, and A. Lewbel, (2008) "Nonparametric Identification of Regression Models Containing a Misclassified Dichotomous Regressor Without Instruments," Economics Letters, 2008, 100, 381-384.

[7] Chen, X., O. Linton, and I. Van Keilegom, (2003) "Estimation of Semiparametric Models when the Criterion Function Is Not Smooth," Econometrica, 71, 1591-1608,

[8] Clogg, C. C. (1995), Latent class models, in G. Arminger, C. C. Clogg, & M. E. Sobel (Eds.), Handbook of statistical modeling for the social and behavioral sciences (Ch. 6; pp. 311-359). New York: Plenum.

[9] Dong, Y.,(2008), "Nonparametric Binary Random Effects Models: Estimating Two Types of Drinking Behavior," Unpublished manuscript.

[10] Gozalo, P, and Linton, O. (2000). Local Nonlinear Least Squares: Using Parametric Information in Non-parametric Regression. Journal of econometrics, 99, 63-106.

[11] Hagenaars, J. A. and McCutcheon A. L. (2002), Applied Latent Class Analysis Models, Cambridge: Cambridge University Press.

[12] Hansen, L., (1982), "Large Sample Properties of Generalized Method of Moments Estimators," Econometrica, 50, 1029-1054.

[13] Heckman, J. J. and R. Robb. (1985) "Alternative Methods for Evaluating the Impact of Interventions, " in Longitudinal Analysis of Labor Market Data. James J. Heckman and B. Singer, eds. New York: Cambridge University Press, 156-245.

[14] Hu, Y. and A. Lewbel, (2008) "Identifying the Returns to Lying When the Truth is Unobserved," Boston College Working paper.

[15] Kasahara, H. and Shimotsu, K. (2007), "Nonparametric Identification and Estimation of Multivariate Mixtures," Working Papers 1153, Queen's University, Department of Economics

[16] Kumbhakar, S. C. and C. A. K. Lovell , (2000), Stochastic Frontier Analysis, Cambridge University Press.

[17] Kumbhakar, S.C., B.U. Park, L Simar, and E.G. Tsionas, (2007) "Nonparametric stochastic frontiers: A local maximum likelihood approach," Journal of Econometrics, 137, 1-27.

[18] Lewbel, A. (2007) "A Local Generalized Method of Moments Estimator," Economics Letters, 94, 124-128.

[19] Lewbel, A. and O. Linton, (2007) "Nonparametric Matching and Efficient Estimators of Homothetically Separable Functions," Econometrica, 75, 1209-1227.

[20] Li, Q. and J. Racine (2003), "Nonparametric estimation of distributions with categorical and continuous data," Journal of Multivariate Analysis, 86, 266-292

[21] Newey, W. K. and D. McFadden (1994), "Large Sample Estimation and Hypothesis Testing," in Handbook of Econometrics, vol. iv, ed. by R. F. Engle and D. L. McFadden, pp. 2111-2245, Amsterdam: Elsevier.

[22] Newey, W. K. and J. L. Powell, (2003), "Instrumental Variable Estimation of Nonparametric Models," Econometrica, 71 1565-1578.

[23] Powell, J. L. (1986), "Symmetrically Trimmed Least Squares Estimation of Tobit Models," Econometrica, 54, 1435-1460.

[24] Simar, L. and P. W. Wilson (2007) "Statistical Inference in Nonparametric Frontier Models: Recent Developments and Perspectives," in The Measurement of Productive Efficiency, 2nd edition, chapter 4, ed. by H. Fried, C.A.K. Lovell, and S.S. Schmidt, Oxford: Oxford University Press.

11 Appendix A: Proofs

PROOF of Theorem 1: First identify h by $h = E(Y)$, since V and U are mean zero. Then the distribution of ε defined by $\varepsilon = Y - h$ is identified, and $\varepsilon = U + V$. Define $e_d = E\left(\varepsilon^d\right)$ and $v_d = E\left(V^d\right)$.

Now evaluate e_d for integers $d \leq 9$. These e_d exist by assumption, and are identified because the distribution of ε is identified. The first goal will be to obtain expressions for v_d in terms of e_d for various values of d. Using independence of V and U, the fact that both are mean zero, and U being symmetric we have

$$\begin{aligned} E\left(\varepsilon^2\right) &= E\left(V^2 + 2VU + U^2\right) \\ e_2 &= v_2 + E\left(U^2\right) \\ E\left(U^2\right) &= e_2 - v_2 \end{aligned}$$

$$\begin{aligned} E\left(\varepsilon^3\right) &= E\left(V^3 + 3V^2U + 3VU^2 + U^3\right) \\ e_3 &= v_3 \end{aligned}$$

$$\begin{aligned} E\left(\varepsilon^4\right) &= E\left(V^4 + 4V^3U + 6V^2U^2 + 4VU^3 + U^4\right) \\ e_4 &= v_4 + 6v_2E\left(U^2\right) + E\left(U^4\right) \\ E\left(U^4\right) &= e_4 - v_4 - 6v_2E\left(U^2\right) \\ &= e_4 - v_4 - 6v_2\left(e_2 - v_2\right) \\ E\left(U^4\right) &= e_4 - v_4 - 6v_2e_2 + 6v_2^2 \end{aligned}$$

$$E\left(\varepsilon^5\right) = E\left(V^5 + 5V^4U + 10V^3U^2 + 10V^2U^3 + 5VU^4 + U^5\right)$$

$$\begin{aligned} e_5 &= v_5 + 10v_3 E\left(U^2\right) = v_5 + 10v_3\left(e_2 - v_2\right) \\ e_5 &= v_5 + 10e_3e_2 - 10e_3v_2 \\ e_5 - 10e_3e_2 &= v_5 - 10e_3v_2 \end{aligned}$$

Define $s = e_5 - 10e_3e_2$, and note that s depends only on identified objects and so is identified. Then $s = v_5 - 10e_3v_2$,

$$\begin{aligned} E\left(\varepsilon^6\right) &= E\left(V^6 + 6V^5U + 15V^4U^2 + 20V^3U^3 + 15V^2U^4 + 6VU^5 + U^6\right) \\ e_6 &= v_6 + 15v_4 E\left(U^2\right) + 15v_2 E\left(U^4\right) + E\left(U^6\right) \\ E\left(U^6\right) &= e_6 - v_6 - 15v_4 E\left(U^2\right) - 15v_2 E\left(U^4\right) \\ &= e_6 - v_6 - 15v_4\left(e_2 - v_2\right) - 15v_2\left(e_4 - v_4 - 6v_2e_2 + 6v_2^2\right) \\ &= e_6 - v_6 - 15e_2v_4 - 15e_4v_2 + 30v_2v_4 - 90v_2^3 + 90e_2v_2^2 \end{aligned}$$

$$\begin{aligned} E\left(\varepsilon^7\right) &= E\left(V^7 + 7V^6U + 21V^5U^2 + 35V^4U^3 + 35V^3U^4 + 21V^2U^5 + 7VU^6 + U^7\right) \\ e_7 &= v_7 + 21v_5 E\left(U^2\right) + 35v_3 E\left(U^4\right) \\ e_7 &= v_7 + 21v_5\left(e_2 - v_2\right) + 35v_3\left(e_4 - v_4 - 6v_2e_2 + 6v_2^2\right) \end{aligned}$$

plug in $v_5 = s + 10e_3v_2$ and $v_3 = e_3$ and expand:

$$\begin{aligned} e_7 &= v_7 + 21\left(s + 10e_3v_2\right)\left(e_2 - v_2\right) + 35e_3\left(e_4 - v_4 - 6v_2e_2 + 6v_2^2\right) \\ &= v_7 + 21se_2 - 21sv_2 + 35e_3e_4 - 35e_3v_4 \end{aligned}$$

Bring terms involving identified objects e_d and s left:

$$e_7 - 21se_2 - 35e_3e_4 = v_7 - 35e_3v_4 - 21sv_2.$$

Define $q = e_7 - 21se_2 - 35e_3e_4$ and note that q depends only on identified objects and so is identified. Then

$$q = v_7 - 35e_3v_4 - 21sv_2.$$

Next consider e_9.

$$E\left(\varepsilon^9\right) = E\left(\begin{array}{c} V^9 + 9V^8U + 36V^7U^2 + 84V^6U^3 + 126V^5U^4 + \\ 126V^4U^5 + 84V^3U^6 + 36V^2U^7 + 9VU^8 + U^9 \end{array}\right)$$

$$\begin{aligned} e_9 &= v_9 + 36v_7 E\left(U^2\right) + 126v_5 E\left(U^4\right) + 84v_3 E\left(U^6\right) \\ e_9 &= v_9 + 36v_7\left(e_2 - v_2\right) + 126v_5\left(e_4 - v_4 - 6v_2e_2 + 6v_2^2\right) \\ &\quad +84v_3\left(e_6 - v_6 - 15e_2v_4 - 15e_4v_2 + 30v_2v_4 - 90v_2^3 + 90e_2v_2^2\right) \end{aligned}$$

Use q and s to substitute out $v_7 = q + 35e_3v_4 + 21sv_2$ and $v_5 = s + 10e_3v_2$, and use $v_3 = e_3$ to get

$$\begin{aligned} e_9 &= v_9 + 36\left(q + 35e_3v_4 + 21sv_2\right)\left(e_2 - v_2\right) + 126\left(s + 10e_3v_2\right)\left(e_4 - v_4 - 6v_2e_2 + 6v_2^2\right) \\ &\quad +84e_3\left(e_6 - v_6 - 15e_2v_4 - 15e_4v_2 + 30v_2v_4 - 90v_2^3 + 90e_2v_2^2\right) \end{aligned}$$

Expand and bring terms involving identified objects e_d, s, and q to the left:

$$e_9 - 36qe_2 - 126se_4 - 84e_3e_6 = v_9 - 36qv_2 - 126sv_4 - 84e_3v_6$$

Define $w = e_9 - 36qe_2 - 126se_4 - 84e_3e_6$ and note that w depends only on identified objects and so is identified. Then

$$w = v_9 - 36qv_2 - 126sv_4 - 84e_3v_6$$

Summarizing, we have w, s, q, e_3 are all identified and

$$\begin{aligned} e_3 &= v_3 \\ s &= v_5 - 10e_3v_2 \\ q &= v_7 - 35e_3v_4 - 21sv_2 \\ w &= v_9 - 84e_3v_6 - 126sv_4 - 36qv_2. \end{aligned}$$

Now V only takes on two values, so let V equal b_0 with probability p_0 and b_1 with probability p_1. Probabilities sum to one, so $p_1 = 1 - p_0$. Also, $E(V) = b_0p_0 + b_1p_1 = 0$ because $\varepsilon = V + U$ and both ε and U have mean zero, so $b_1 = -b_0p_0/(1 - p_0)$. Let $r = p_0/p_1 = p_0/(1 - p_0)$, so

$$p_0 = r/(1 + r), \qquad p_1 = 1/(1 + r), \qquad b_1 = -b_0r,$$

and for any integer d

$$v_d = b_0^d p_0 + b_1^d p_1 = b_0^d \left(p_0 + (-r)^d p_1\right) = b_0^d \frac{r + (-r)^d}{1+r}$$

so in particular

$$\begin{aligned}
v_2 &= b_0^2 r \\
v_3 &= b_0^3 r\,(1-r) \\
v_4 &= b_0^4 r\left(r^2 - r + 1\right) \\
v_5 &= b_0^5 r\,(1-r)\left(r^2+1\right) \\
v_6 &= b_0^6 \frac{r + (-r)^6}{1+r} = b_0^6 r\left(r^4 - r^3 + r^2 - r + 1\right) \\
v_7 &= b_0^7 r\,(1-r)\left(r^4 + r^2 + 1\right) \\
v_9 &= b_0^9 \frac{r + (-r)^9}{1+r} = b_0^9 r\,(1-r)\left(r^2+1\right)\left(r^4+1\right)
\end{aligned}$$

Substituting these v_d expressions into the expression for e_3, s, q, and w gives $e_3 = b_0^3 r\,(1-r)$,

$$\begin{aligned}
s &= b_0^5 r\,(1-r)\left(r^2+1\right) - 10 b_0^3 r\,(1-r)\, b_0^2 r \\
&= b_0^5\left(r\,(1-r)\left(r^2+1\right) - 10 r\,(1-r)\, r\right) \\
s &= b_0^5 r\,(1-r)\left(r^2 - 10r + 1\right)
\end{aligned}$$

$$\begin{aligned}
q &= v_7 - 35 e_3 v_4 - 21 s v_2 \\
&= b_0^7 r\,(1-r)\left(r^4 + r^2 + 1\right) - 35 b_0^3 r\,(1-r)\, b_0^4 r\left(r^2 - r + 1\right) - 21 b_0^5 r\,(1-r)\left(r^2 - 10r + 1\right) b_0^2 r \\
&= b_0^7\left(r\,(1-r)\left(r^4 + r^2 + 1\right) - 35 r\,(1-r)\, r\left(r^2 - r + 1\right) - 21 r\,(1-r)\left(r^2 - 10r + 1\right) r\right) \\
q &= b_0^7 r\,(1-r)\left(r^4 - 56r^3 + 246r^2 - 56r + 1\right)
\end{aligned}$$

$$\begin{aligned}
w &= v_9 - 84 e_3 v_6 - 126 s v_4 - 36 q v_2 \\
&= \begin{pmatrix}
b_0^9 r\,(1-r)\left(r^2+1\right)\left(r^4+1\right) - 84\left(b_0^3 r\,(1-r)\right)\left(b_0^6 r\left(r^4 - r^3 + r^2 - r + 1\right)\right) \\
-126\left(b_0^5 r\,(1-r)\left(r^2 - 10r + 1\right)\right)\left(b_0^4 r\left(r^2 - r + 1\right)\right) \\
-36\left(b_0^7 r\,(1-r)\left(r^4 - 56r^3 + 246r^2 - 56r + 1\right)\right) b_0^2 r
\end{pmatrix}
\end{aligned}$$

$$
\begin{aligned}
&= b_0^9 r\left(1-r\right)\left(\begin{array}{c}\left(r^2+1\right)\left(r^4+1\right)-84\left(r\left(r^4-r^3+r^2-r+1\right)\right)\\ -126\left(\left(r^2-10r+1\right)\right)\left(r\left(r^2-r+1\right)\right)-36\left(\left(r^4-56r^3+246r^2-56r+1\right)\right)r\end{array}\right)\\
w &= b_0^9 r\left(1-r\right)\left(r^6-246r^5+3487r^4-10452r^3+3487r^2-246r+1\right)
\end{aligned}
$$

Summarizing the results so far we have

$$
\begin{aligned}
e_3 &= b_0^3 r\left(1-r\right)\\
s &= b_0^5 r\left(1-r\right)\left(r^2-10r+1\right)\\
q &= b_0^7 r\left(1-r\right)\left(r^4-56r^3+246r^2-56r+1\right)\\
w &= b_0^9 r\left(1-r\right)\left(r^6-246r^5+3487r^4-10452r^3+3487r^2-246r+1\right)
\end{aligned}
$$

These are four equations in the two unknowns b_0 and r. We require all four equations for identification, and not just two or three of them, because these are polynomials in r and so have multiple roots. We will now show that these four equations imply that $r^2-\gamma+1=0$, where γ is finite and identified.

First we have $e_3 = v_3 \neq 0$ and $r \neq 1$ by asymmetry of V. Also $r \neq 0$ because then V would only have one point of support instead of two, and these together imply by $e_3 = b_0^3 r\left(1-r\right)$ that $b_0 \neq 0$. Applying these results to the s equation shows that if s (which is identified) is zero then $r^2-10r+1=0$, and so in that case γ is identified. So now consider the case where $s \neq 0$.

Define $R = qe_3/s^2$, which is identified because its components are identified. Then

$$
\begin{aligned}
R &= \frac{b_0^7 r\left(1-r\right)\left(r^4-56r^3+246r^2-56r+1\right)b_0^3 r\left(1-r\right)}{b_0^5 r\left(1-r\right)\left(r^2-10r+1\right)b_0^5 r\left(1-r\right)\left(r^2-10r+1\right)}\\
&= \frac{r^4-56r^3+246r^2-56r+1}{\left(r^2-10r+1\right)^2}
\end{aligned}
$$

So

$$
\begin{aligned}
0 &= \left(r^4-56r^3+246r^2-56r+1\right)-\left(r^2-10r+1\right)^2 R\\
0 &= \left(1-R\right)r^4+\left(-56+20R\right)r^3+\left(246-102R\right)r^2+\left(-56+20R\right)r+\left(1-R\right)
\end{aligned}
$$

Which yields a fourth degree polynomial in r. If $R=1$, then (using $r \neq 0$) this polynomial reduces to the quadratic $0 = r^2-4r+1$, so in this case $\gamma=-4$ is identified. Now consider the case where $R \neq 1$.

Define $Q = s^3/e_3^5$ which is identified because its components are identified. Then

$$\begin{aligned}
Q &= \frac{\left(b_0^5 r\left(1-r\right)\left(r^2-10r+1\right)\right)^3}{\left(b_0^3 r\left(1-r\right)\right)^5} = \frac{\left(r^2-10r+1\right)^3}{\left(r\left(1-r\right)\right)^2} \\
0 &= \left(r^2-10r+1\right)^3 - \left(r\left(1-r\right)\right)^2 Q \\
0 &= r^6 - 30r^5 + \left(303-Q\right) r^4 + \left(2Q-1060\right) r^3 + \left(303-Q\right) r^2 - 30r + 1
\end{aligned}$$

which is a sixth degree polynomial in r. Also define $S = w/e_3^2$ which is identified because its components are identified. Then

$$\begin{aligned}
\frac{w}{e_3^3} &= S = \frac{b_0^9 r\left(1-r\right)\left(r^6-246r^5+3487r^4-10452r^3+3487r^2-246r+1\right)}{\left(b_0^3 r\left(1-r\right)\right)^3} \\
S &= \frac{\left(r^6-246r^5+3487r^4-10452r^3+3487r^2-246r+1\right)}{\left(r\left(1-r\right)\right)^2} \\
0 &= \left(r^6-246r^5+3487r^4-10452r^3+3487r^2-246r+1\right) - \left(r\left(1-r\right)\right)^2 S \\
0 &= r^6 - 246r^5 + \left(3487-S\right) r^4 + \left(2S-10452\right) r^3 + \left(3487-S\right) r^2 - 246r + 1
\end{aligned}$$

which is another sixth degree polynomial in r. Subtracting the second of these sixth degree polynomials from the other and dividing the result by r gives the fourth order polynomial:

$$0 = 216r^4 + \left(S-Q-3184\right) r^3 + \left(9392+2Q-2S\right) r^2 + \left(S-Q-3184\right) r + 216.$$

Multiply this fourth order polynomial by $(1-R)$, multiply the previous fourth order polynomial by 216, subtract one from the other. and divide by r to obtain a quadratic in r:

$$\begin{aligned}
0 &= 216\left(1-R\right) r^4 + \left(1-R\right)\left(S-Q-3184\right) r^3 + \left(1-R\right)\left(9392+2Q-2S\right) r^2 \\
&\quad + \left(1-R\right)\left(S-Q-3184\right) r + 216\left(1-R\right) - 216\left(1-R\right) r^4 - 216\left(-56+20R\right) r^3 \\
&\quad -216\left(246-102R\right) r^2 - 216\left(-56+20R\right) r - 216\left(1-R\right) \\
0 &= \left(\left(1-R\right)\left(S-Q-3184\right) - 216\left(-56+20R\right)\right) r^3 \\
&\quad + \left(\left(1-R\right)\left(9392+2Q-2S\right) - 216\left(246-102R\right)\right) r^2 \\
&\quad + \left(\left(1-R\right)\left(S-Q-3184\right) - 216\left(-56+20R\right)\right) r \\
0 &= \left(\left(1-R\right)\left(S-Q-3184\right) + 12096 - 4320R\right) r^2
\end{aligned}$$

$$+\left(\left(1-R\right)\left(9392+2Q-2S\right)+22032R-53136\right)r$$
$$+\left(\left(1-R\right)\left(S-Q-3184\right)+12096-4320R\right).$$

which simplifies to

$$0=Nr^2-\left(2\left(1-R\right)\left(6320+S-Q\right)+31104\right)r+N$$

where $N=\left(1-R\right)\left(1136+S-Q\right)+7776$. The components of N can be written as

$$\begin{aligned}1-R &= 1-\frac{r^4-56r^3+246r^2-56r+1}{\left(r^2-10r+1\right)^2}=\frac{\left(r^2-10r+1\right)^2-\left(r^4-56r^3+246r^2-56r+1\right)}{\left(r^2-10r+1\right)^2}\\ &= \frac{36r^3-144r^2+36r}{\left(r^2-10r+1\right)^2}\end{aligned}$$

$$\begin{aligned}&1136+S-Q\\ =\ &\left(1136+\left(\frac{\left(r^6-246r^5+3487r^4-10452r^3+3487r^2-246r+1\right)}{\left(r\left(1-r\right)\right)^2}\right)-\frac{\left(r^2-10r+1\right)^3}{\left(r\left(1-r\right)\right)^2}\right)\\ =\ &\frac{1136\left(r\left(1-r\right)\right)^2+\left(r^6-246r^5+3487r^4-10452r^3+3487r^2-246r+1\right)-\left(r^2-10r+1\right)^3}{\left(r\left(1-r\right)\right)^2}\\ =\ &\frac{-216r^5+4320r^4-11664r^3+4320r^2-216r}{\left(r\left(1-r\right)\right)^2}\end{aligned}$$

so

$$\begin{aligned}N &= \left(\left(\frac{36r^3-144r^2+36r}{\left(r^2-10r+1\right)^2}\right)\left(\frac{-216r^5+4320r^4-11664r^3+4320r^2-216r}{\left(r\left(1-r\right)\right)^2}\right)+7776\right)\\ &= \frac{\left(36r^3-144r^2+36r\right)\left(-216r^5+4320r^4-11664r^3+4320r^2-216r\right)}{\left(r^2-10r+1\right)^2\left(r\left(1-r\right)\right)^2}\\ &\quad+\frac{7776\left(r^2-10r+1\right)^2\left(r\left(1-r\right)\right)^2}{\left(r^2-10r+1\right)^2\left(r\left(1-r\right)\right)^2}\\ &= \frac{15552r^3+62208r^4+93312r^5+62208r^6+15552r^7}{\left(r^2-10r+1\right)^2\left(r\left(1-r\right)\right)^2}=\frac{15552r^3\left(r+1\right)^4}{\left(r^2-10r+1\right)^2\left(r\left(1-r\right)\right)^2}\\ N &= \frac{15552r\left(r+1\right)^4}{\left(r^2-10r+1\right)^2\left(1-r\right)^2}\end{aligned}$$

The denominator of this expression for N is not equal to zero, because that would imply

$s = 0$, and we have already considered that case, and ruled it out in the derivation of the quadratic involving N. Now N could only be zero if $15552r\left(r+1\right)^4 = 0$, and this cannot hold because $r \neq 0$, and $r > 0$ (being a ratio of probabilities) so $r \neq -1$ is ruled out. We therefore have $N \neq 0$, so the quadratic involving N can be written as $0 = r^2 - \gamma r + 1$ where $\gamma = \left(2\left(1-R\right)\left(6320+S-Q\right)+31104\right)/N$, which is identified because all of its components are identified.

We have now shown that $0 = r^2 - \gamma r + 1$ where γ is identified. This quadratic has solutions

$$r = \frac{1}{2}\gamma + \frac{1}{2}\sqrt{\gamma^2 - 4} \quad \text{and} \quad r = \frac{1}{\frac{1}{2}\gamma + \frac{1}{2}\sqrt{\gamma^2-4}}$$

so one of these must be the true value of r. Given r, we can then solve for b_0 by $b_0 = e_3^{1/3}\left(r\left(1-r\right)\right)^{1/3}$. Recall that $r = p_0/p_1$. By symmetry of the set up of the problem, if we exchanged b_0 with b_1 and exchanged p_0 with p_1 everywhere, all of the above equations would still hold. It follows that one of the above two values of r must equal p_0/p_1, and the other equals p_1/p_0. The former when substituted into $e_3\left(r\left(1-r\right)\right)$ will yield b_0^3 and the latter must by symmetry yield b_1^3. Without loss of generality imposing the constraint that $b_0 < 0 < b_1$, shows that the correct solution for r will be the one that satisfies $e_3\left(r\left(1-r\right)\right) < 0$, and so r and b_0 is identified. The remainder of the distribution of V is then given by $p_0 = r/\left(1+r\right)$, $p_1 = 1/\left(1+r\right)$, and $b_1 = -b_0 r$. Finally, given that the distributions of ε and of V are identified, the distribution of U is identified by a deconvolution, in particular we have that the characteristic function of U is identified by $E\left(e^{i\tau U}\right) = E\left(e^{i\tau\varepsilon}\right)/E\left(e^{i\tau V}\right)$.

PROOF of Corollary 1: By $\varepsilon = Y - h = V + U$ and by symmetry of U, equation (11) equals

$$E\left(\frac{\exp\left(\tau\varepsilon\right)}{\exp\left(\tau V\right)} - \frac{\exp\left(-\tau\varepsilon\right)}{\exp\left(-\tau V\right)}\right) = E\left(\exp\left(\tau U\right) - \exp\left(-\tau U\right)\right) = 0$$

and $\tau \leq T$ ensures that these expectations exist.

PROOF of Theorem 2: By the probability mass function of the V distribution, $F_\varepsilon\left(\varepsilon\right) = \left(1-p\right)F_U\left(\varepsilon - b_1\right) + pF_U\left(\varepsilon - b_0\right)$. Evaluating this expression at $\varepsilon = u + b_1$ gives

$$F_\varepsilon\left(u+b_1\right) = \left(1-p\right)F_U\left(u\right) + pF_U\left(u+b_1-b_0\right) \tag{23}$$

and evaluating at $\varepsilon = -u + b_0$ gives $F_\varepsilon\left(-u+b_0\right) = \left(1-p\right)F_U\left(-u-b_1+b_0\right) + pF_U\left(-u\right)$.

Apply symmetry of U which implies $F_U(u) = 1 - F_U(-u)$ to this last equation to obtain

$$F_\varepsilon(-u + b_0) = (1-p)[1 - F_U(U + b_1 - b_0)] + p[1 - F_U(u)] \tag{24}$$

Equations (23) and (24) are two equations in the two unknowns $F_U(U + b_1 - b_0)$ and $F_U(U)$. Solving for $F_U(U)$ gives equation (13).

PROOF of Corollary 2: First identify $h(x)$ by $h(x) = E(Y \mid X = x)$, since $E(Y - h(X) \mid X = x) = E(V + U \mid X = x) = E(V + U) = 0$. Next define $\varepsilon = Y - h(X)$ and then the rest of the proof is identical to the proof of Theorem 1.

PROOF of Corollary 3: Define $h(x) = E(Y \mid X)$ and $\varepsilon = Y - h(X)$. Then $h(x)$ and the distribution of ε conditional upon X is identified and $E(\varepsilon \mid X) = 0$. Define $V = g(X, D^*) - h(X)$ and let $b_d(X) = g(X, d) - h(X)$ for $d = 0, 1$. Then $\varepsilon = V + U$, where V (given X) has the distribution with support equal to the two values $b_0(X)$ and $b_1(X)$ with probabilities $p(X)$ and $1 - p(X)$, respectively. Also U and ε have mean zero given X so $E(V \mid X) = 0$. Applying Theorem 1 separately for each value x that X can take on shows that $b_0(x)$, $b_1(x)$ and $p(x)$ are identified for each x in the support of X, and it follows that the function $g(x, d)$ is identified by $g(x, d) = b_d(x) + h(x)$. Applying Theorem 1 separately for each value X can take on also directly provides identification of $p(X)$ and the conditional distribution of U given X.

12 Appendix B: Asymptotic Theory

Most of the estimators in the paper are either standard GMM or well known variants of GMM. However, we here briefly summarize the application of the local GMM estimator of Lewbel (2008) to estimation based on Corollary 3, which as described in the text reduces to estimation based on equation (22). To motivate this estimator, which is closely related to Gonzalo and Linton (2000), first consider the case where all the elements of X are discrete, or more specifically, the case where X has one or more mass points and we only wish to estimate $q(x)$ at those points. Let $q_0(x)$ denote the true value of $q(x)$, and let $\theta_{x0} = q_0(x)$. If the distribution of X has a mass point with positive probability at x, then

$$E[G(\theta_x, Y) \mid X = x] = \frac{E[G(\theta_x, Y) I(X = x)]}{E[I(X = x)]}$$

so equation (22) holds if and only if $E[G(\theta_{x0}, Y)I(X = x)] = 0$. It therefore follows that under standard regularity conditions we may estimate $\theta_{x0} = q_0(x)$ using the ordinary GMM estimator

$$\widehat{\theta}_x = \arg\min_{\theta_x} \sum_{i=1}^{n} G(\theta_x, Y_i)' I(X_i = x) \Omega_n \sum_{i=1}^{n} G(\theta_x, Y_i)' I(X_i = x) \tag{25}$$

for some sequence of positive definite Ω_n. If Ω_n is a consistent estimator of $\Omega_{x0} = E[G(\theta_{x0}, Y)G(\theta_{x0}, Y)'I(X = x)]^{-1}$, then standard efficient GMM gives

$$\sqrt{n}(\widehat{\theta}_x - \theta_{x0}) \to^d N\left(0, \left[E\left(\frac{\partial G(\theta_{x0}, Y)I(X = x)}{\partial \theta_x{}'}\right)\Omega_{x0} E\left(\frac{\partial G(\theta_{x0}, Y)I(X = x)}{\partial \theta_x{}'}\right)'\right]^{-1}\right)$$

Now assume that X is continuously distributed. Then the local GMM estimator consists of applying equation (25) by replacing the average over just observations $X_i = x$ with local averaging over observations X_i in the neighborhood of x.

Assumption B1. Let X_i, Y_i, $i = 1, ..., n$, be an independently, identically distributed random sample of observations of the random vectors X, Y. The d vector X is continuously distributed with density function $f(X)$. For given point x in the interior of $supp(X)$ having $f(x) > 0$ and a given vector valued function $G(q, y)$ where $G(q(x), y)$ is twice differentiable in the vector $q(x)$ for all $q(x)$ in some compact set $\Theta(x)$, there exists a unique $q_0(x) \in \Theta(x)$ such that $E[G(q_0(x), Y) \mid X = x] = 0$. Let Ω_n be a finite positive definite matrix for all n, as is $\Omega = plim_{n\to\infty}\Omega_n$.

Assumption B1 lists the required moment condition structure and identification for the estimator. Corollary 1 in the paper provides the conditions required for Assumption B1, in particular uniqueness of $q_0(x)$. Assumption B2 below provides conditions required for local averaging. Define $e[q(x), Y]$, $\Sigma(x)$, and $\Psi(x)$ by

$$\begin{aligned} e[q(x), Y] &= G(q(x), Y)f(x) - E[G(q(x), Y)f(X) \mid X = x] \\ \Sigma(x) &= E\left[e(q_0(x), Y)e(q_0(x), Y)^T \mid X = x\right] \\ \Psi(x) &= E\left(\frac{\partial G[q_0(x), Y]}{\partial q_0(x)^T} f(X) \mid X = x\right) \end{aligned}$$

Assumption B2. Let η be some constant greater than 2. Let K be a nonnegative symmetric kernel function satisfying $\int K(u)du = 1$ and $\int ||K(u)||^\eta du$ is finite. For all $q(x) \in$

$\Theta(x)$, $E[||G(q(x),Y)f(X)||^{\eta} \mid X=x]$, $\Sigma(x)$, $\Psi(x)$, and $Var[[\partial G(q(x),Y)/\partial q(x)]f(X) \mid X=x]$ are finite and continuous at x and $E[G(q(x),Y)f(X) \mid X=x]$ is finite and twice continuously differentiable at x.

Define

$$S_n(q(x)) = \frac{1}{nb^d}\sum_{i=1}^{n} G[q(x),Y_i]K\left(\frac{x-X_i}{b}\right)$$

where $b=b(n)$ is a bandwidth parameter. The proposed local GMM estimator is

$$\widehat{q}(x) = \arg \inf_{q(x)\in\Theta(x)} S_n(q(x))^T \Omega_n S_n(q(x)) \tag{26}$$

THEOREM 3 (Lewbel 2008): Given Assumptions B1 and B2, if the bandwidth b satisfies $nb^{d+4} \to 0$ and $nb^d \to \infty$, then $\widehat{q}(x)$ is a consistent estimator of $q_0(x)$ with limiting distribution

$$(nb)^{1/2}[\widehat{q}(x) - q_0(x)] \to^d N\left[0, (\Psi(x)^T\Omega\Psi(x))^{-1}\Psi(x)^T\Omega\Sigma(x)\Omega\Psi(x)(\Psi(x)^T\Omega\Psi(x))^{-1}\int K(u)^2 du\right]$$

Applying the standard two step GMM procedure, we may first estimate $\widetilde{q}(x) = \arg\inf_{q(x)\in\Theta(x)} S_n(q(x))^T S_n(q(x))$, then let Ω_n be the inverse of the sample variance of $S_n(\widetilde{q}(x))$ to get $\Omega = \Sigma(x)^{-1}$, making

$$(nb)^{1/2}[\widehat{q}(x) - q_0(x)] \to^d N\left[0, (\Psi(x)^T\Omega\Psi(x))^{-1}\int K(u)^2 du\right]$$

where $\Psi(x)$ can be estimated using

$$\Psi_n(x) = \frac{1}{nb^d}\sum_{i=1}^{n} \frac{\partial G[\widehat{q}(x),Y_i]}{\partial \widehat{q}(x)^T} K\left(\frac{x-X_i}{b}\right)$$

At the expense of some additional notation, the two estimators (25) and (26) can be combined to handle X containing both discrete and continuous elements, by replacing the kernel function in S_n with the product of a kernel over the continuous elements and an indicator function for the discrete elements, as in Li and Racine (2003).