Nonparametric Identification of a Binary Random Factor in Cross Section Data

and

Returns to Lying? Identifying the Effects of Misreporting When the Truth is Unobserved

Arthur Lewbel

Boston College

This packet contains two papers that will be discussed in the seminar:

"Nonparametric Identification of a Binary Random Factor in Cross Section Data," by Yingying Dong and Arthur Lewbel

and

"Returns to Lying? Identifying the Effects of Misreporting When the Truth is Unobserved," by Yingyao Hu and Arthur Lewbel

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 January 2010

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 illustrate the results with an empirical example looking at possible convergence over time in the world income distribution. 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. 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.

The symmetry of errors used for identification here can arise in a number of contexts. Normal errors are of course symmetric, and normality arises in many models such as those involving central limit theorems or Gibrat's law. Differences of independently, identically distributed errors (or more generally of exchangable errors) are symmetrically distributed. See, e.g., proposition 1 of Honore (1992). So, e.g., two period panel models with fixed effects will have errors that are symmetric after time differencing. Our results could therefore be applied in a two period panel where individuals can have an unobserved mean shift at any time (corresponding to the unobserved binary regressor), fixed effects (which are differenced away) and exchangable remaining errors (which yield symmetric errors after differencing). Below we give other more specific examples of models with symmetric errors.

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 question of whether less developed economies are catching up with the economies of richer countries 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 poor, $b_1 - b_0$ 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 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 parametric functional forms for both 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^*.

Another potential class of applications is where individuals are randomly assigned some treatment D^*, perhaps by a natural experiment, but in our data we do not ob-

serve who was treated and who wasn't. Assuming that treatment only induces a mean shift in outcomes, we would still be able to identify the probability of treatment and the difference in mean outcomes between treated and untreated in this context. If we have panel data (two periods of observations) and all treatments occur in one of the two periods, then as noted above the required symmetry of U errors would result automatically from time differencing the data, given the standard panel model assumption of individual specific fixed effects plus independently, identically distributed (or more generally exchangable) errors.

Dong (2008) 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 uses 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). Our model resembles a (restricted) finite mixture model, but differs crucially in that, for identification, finite mixture models require the distributions being mixed to be parametrically specified, while in our model U is nonparametric. While mixture models are more flexible than ours in allowing more than two groups, and for U to vary across groups, ours is more flexible in allowing U to be nonparametric, essentially allowing for

an infinite number of parameters versus finitely parameterized mixtures.

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 provide additional moments for estimation, and 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 $E\left(Y^9\right)$.

It can be readily shown using the method of proof in Theorem 1 that the parameters defining the distribution of V are generically locally identified using just a few low order moments of Y. Higher moments going up to the ninth moment are required only to distinguish amongst a small number of multiple roots and thereby provide global identification.

Let b_0 and b_1 denote 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 $V = b_1$. Note that Theorem 1 assumes asymmetry of V (since otherwise it would be indistinguishable from U) and hence requires $p \neq 1/2$. This suggests that the identification and associated estimation will be weak if the actual p is close to $1/2$. In practice, it would be easy to tell if this problem exists, because if it does then the observed Y will itself be close to symmetrically distributed. Applying a formal test of data symmetry such as Ahmed and Li (1997) to the Y data is equivalent in our model to testing if $p = 1/2$.

We next 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.

3 Estimation

Our estimator will take the form of the standard Generalized Method of Moments (GMM, as in Hansen 1982), since given data Y_1,...Y_n, we will below construct 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 the vector θ consists of the parameters interest h, b_0, p, as well as u_2, u_4, and u_6, where $u_d = E\left(U^d\right)$. The parameters u_2, u_4, and u_6 are nuisance parameters for estimating the V distribution, but in some applications they may be of interest as summary measures of the distribution of U.

Let $v_d = E\left(V^d\right)$. 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 we can write as $E[G(Y,\theta)] = 0$ in the six unknown parameters $\theta = (h,\ b_0,\ p,\ u_2,\ u_4,\ u_6)$, which we use for estimation via GMM. 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 global identification because of the nonlinearity of these

equations, and in particular the presence of multiple roots. Given an estimate of θ, the estimate of b_1 is then obtained by equation (1).

Based on Theorem 1, for this estimator we assume that Y_1, ...,Y_n are identically distributed (or more precisely, have identical first nine moments), however, the Y observations do not need to be independent, since GMM estimation theory permits some serial dependence in the data. To save space we do not write out detailed assumptions and associated limiting distribution theory for our estimator, because they are completely standard. Textbook GMM limiting distribution theory applied to our moments provides root n consistent, asymptotically normal estimates of θ and hence 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 $\widehat{b}_1 = \widehat{b}_0\widehat{p}/\left(\widehat{p}-1\right)$ from equation 1). In a free normalization, we assume $b_0 < b_1$ (if this is violated then the definitions of these two parameters can be switched to make the inequality hold). This along with $E\left(V\right) = 0$ implies that $\widehat{b}_0$ is negative and $\widehat{b}_1$ is positive, which may be imposed on estimation.

One might anticipate poor empirical results, and extreme sensitivity to outliers, given the use of such high order moments for estimation. However, we found that these problems did not arise in our empirical application, as long as Y was scaled appropriately (to avoid the effects of computer rounding errors based on inverting matrix entries of varying orders of magnitude). We believe the reason is that local identification and associated estimates primarily derives from the low order moments. The high order moments are only needed for global identification to distinguish between a few possible multiple solutions of the low order polynomials. So, e.g., if low order moments identify a parameter up to, say, two values, one positive and one negative, then substantial effects of outliers on estimated higher moments will not matter much if the high moments are only needed to distinguish between a positive mean and a negative mean.

In an extension section we describe how additional moments can be constructed for

estimation based on symmetry of U. These alternative moments could be employed in applications where the polynomial based moments are found to be problematic. Another possibility would be to remove outliers from the Y data prior to estimation (which can be interpreted as robustifying higher moment estimates), though we found this to be unnecessary in our empirical application.

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

$$\Psi(u) = \frac{\left[F_\varepsilon(-u+b_0)-1\right]p + F_\varepsilon(u+b_1)(1-p)}{1-2p}. \tag{11}$$

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

$$F_U(u) = \frac{\Psi(u) - \Psi(-u) + 1}{2}. \tag{12}$$

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. This can be used to construct an estimator for $F_U(u)$ as follows.

Let $I(\cdot)$ denote the indicator function that equals one if $\cdot$ is true and zero otherwise,

and let θ be a vector containing h, b_0, b_1, and p. Define the function $\omega(Y, u, \theta)$ by

$$\omega(Y, u, \theta) = \frac{[I(Y \leq h - u + b_0) - 1]\, p + I(Y \leq h + u + b_1)(1 - p)}{1 - 2p}. \tag{13}$$

Then using $Y = h + \varepsilon$ it follows immediately from equation (11) that

$$\Psi(u) = E(\omega(Y, u, \theta)). \tag{14}$$

An estimator for $F_U(u)$ can now be constructed by replacing the parameters in equation (14) with estimates, replacing the expectation with a sample average, and plugging the result into equation (12). The resulting estimator is

$$\widehat{F}_U(u) = \frac{1}{n} \sum_{i=1}^{n} \frac{\omega\left(Y_i, u, \widehat{\theta}\right) - \omega\left(Y_i, -u, \widehat{\theta}\right) + 1}{2}. \tag{15}$$

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 (12) and (14),

$$E\left[\eta_j - \frac{\omega(Y, u_j, \theta) - \omega(Y, u_j, \theta) + 1}{2}\right] = 0. \tag{16}$$

Adding equation (16) 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 (15) 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,

$$\begin{aligned}\Pr\left(V=b_0 \mid Y\leq y\right) &= \Pr\left(V=b_0, Y\leq y\right)/\Pr\left(Y\leq y\right)\\ &= F_U\left(y-h-b_0\right)/F_y\left(y\right)\end{aligned}$$

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 (11) and (12) with respect to u gives

$$\psi\left(u\right)=\frac{-f_\varepsilon\left(-u+b_0\right)p+f_\varepsilon\left(u+b_1\right)\left(1-p\right)}{1-2p}, \qquad f_U\left(u\right)=\frac{\psi\left(u\right)+\psi\left(-u\right)}{2}, \tag{17}$$

which suggests the estimators

$$\widehat{\psi}\left(u\right)=\frac{-\widehat{f}_\varepsilon\left(-u+\widehat{b}_0\right)\widehat{p}+\widehat{f}_\varepsilon\left(u+\widehat{b}_1\right)\left(1-\widehat{p}\right)}{1-2\widehat{p}}, \tag{18}$$

$$\widehat{f}_U\left(u\right)=\frac{\widehat{\psi}\left(u\right)+\widehat{\psi}\left(-u\right)}{2}, \tag{19}$$

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 $n^{2/5}$. 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 $\widehat{f}_U(u)$ with zero.

A potential numerical problem is that equation (18) 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 (19) 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 $\widehat{\varepsilon}_i$.

This construction also suggests a specification test for the model. Since symmetry of U implies that $\widehat{\psi}(u) = \widehat{\psi}(-u)$ one could base a test on whether $\int_0^L \left[\widehat{\psi}(u) - \widehat{\psi}(-u)\right]^2 w(u)\, du = 0$, where $w(u)$ is a weighting function that integrates to one, and L is in the range of values for which neither $\widehat{\psi}(-u)$ nor $\widehat{\psi}(u)$ requires evaluating $\widehat{f}_\varepsilon$ at a point outside the the range of observed values of $\widehat{\varepsilon}_i$. The limiting distribution theory for this type of test statistic (a degenerate U statistic under the null) based on functions of kernel densities is standard, and in this case would closely resemble Ahmed and Li (1997).

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

$$\sum_{i=1}^{n} \ln\left(\frac{p}{s\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{Y_i - h - b_0}{s}\right)^2\right] + \frac{1-p}{s\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{Y_i - h - \frac{b_0 p}{p-1}}{s}\right)^2\right]\right). \quad (20)$$

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 $\widehat{b}_1 = \widehat{b}_0\widehat{p}/\left(\widehat{p}-1\right)$. 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 $h_t = E_t\left(Y\right)$ 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 written 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.

Objections can be easily raised to this simplistic model, e.g., that other indicators in addition to income exist for grouping countries, that countries could be divided into more than two groups, and that there is not a strong economic argument for why the distribution of incomes around group means should be symmetric and the same for both groups. One could respond that it is common to dichotomize the world into groups of rich (well developed) and poor (less developed) countries, that Gibrat's law within groups could generate the required symmetry, and that the shape of the world income distribution suggests at least rough appropriateness of the model (including possible bimodality of Y with estimates of the U distribution close to normal). Still, given these valid concerns, we interpret our model as primarily descriptive rather than structural. Our main goal is to verify that the polynomial moments we use for identification and estimation can produce reasonable estimates with real data and small sample sizes.

Though simple, our 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.

Bianchi (1997) applies bimodality tests to the distribution of income across countries over time, to address questions regarding evidence for convergence. Bimodality versus unimodality of Y might be interpreted as evidence in favor of a '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.

For comparison we apply our model using the same data as Bianchi, 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.

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	GMM	.8575	-.1105	.6648	.7753	.3214	.0221	.0001$^{\$}$	.0024
		(.0352)	(.0244)	(.0664)	(.0590)	(.0284)	(.0042)	(.0002)	(.0009)
	MLE	.8098	-.1334	.5679	.7013	.3213	.0199		
		(.0362)	(.0260)	(.0487)	(.0477)	(.0280)	(.0031)		
1980	GMM	.8081	-.1722	.7252	.8974	.4223	.0294	.0016	.0017*
		(.0371)	(.0322)	(.0579)	(.0491)	(.0351)	(.0043)	(.0004)	(.0007)
	MLE	.8070	-.1692	.7077	.8769	.4222	.0350		
		(.0393)	(.0345)	(.0600)	(.0544)	(.0372)	(.0048)		
1989	GMM	.8125	-.2114	.9159	1.1273	.4804	.0384	.0051	.0028$^{\$}$
		(.0380)	(.0424)	(.1022)	(.1111)	(.0439)	(.0118)	(.0104)	(.0448)
	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 2: Estimates based on the scaled GDP per capita share data

		p	b0	b1	b1-b0	h	u2	u4	u6
	GMM	.8619	-.1392	.8682	1.0074	.4206	.0417	.0039$	.0057$
1970		(.0361)	(.0332)	(.1009)	(.0985)	(.0380)	(.0089)	(.0068)	(.0063)
	MLE	.8098	-.1744	.7425	.9169	.4202	.0340		
		(.0383)	(.0352)	(.0670)	(.0629)	(.0377)	(.0053)		
	GMM	.8080	-.1715	.7217	.8932	.4202	.0291	.0016	.0017
1980		(.0374)	(.0334)	(.0560)	(.0497)	(.0364)	(.0041)	(.0004)	(.0006)
	MLE	.8070	-.1684	.7043	.8727	.4202	.0347		
		(.0373)	(.0322)	(.0570)	(.0508)	(.0353)	(.0045)		
	GMM	.8117	-.1848	.7964	.9812	.4203	.0316	.0023	.0020*
1989		(.0360)	(.0344)	(.0609)	(.0518)	(.0388)	(.0049)	(.0007)	(.0009)
	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.

For each of the three years of data we provide two different estimates, labeled GMM and MLE in Tables 1 and 2. GMM is based on the identifying polynomial moments (3) to (10) (after substituting in equation (2)), while MLE is a maximum likelihood estimator that maximizes (20) assuming that U is normal.

Table 1 reports results based on per capita levels, $Y_{ti} = I_{ti}/10,000$, while Table 2 is based on scaled shares, $Y_{ti} = 50I_{ti}/\left(\sum_{i=1}^{n} I_{ti}\right)$. We scale by 10,000 in Table 1 and by 50 in Table 2 to put the Y_{ti} data in a range between zero and two in each case. These scalings are theoretically irrelevant, but in practice help ensure that the matrices involved in estimation (particularly the high order polynomial terms in the estimated second stage GMM weighting matrix) are numerically well conditioned despite computer round off error.

In both Tables 1 and 2, and in all three years, the GMM and maximum likelihood estimates are roughly comparable, for the most part lying within about 10% of each other. Looking across years, both Tables 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. In contrast, 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 scales out 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.

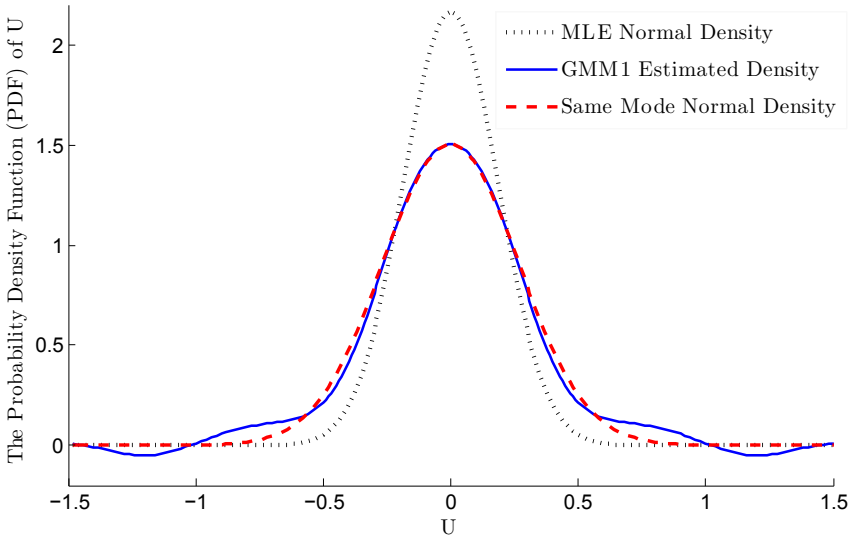

Figure 1: The estimated probability density function of U, using 1970 share data

Figure 1 shows $\widehat{f}_u$, the estimated density of U, given by equation (19) using the GMM estimates from Table 2 in 1970. Graphs of other years are very similar, so to save space we do not include them here. 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 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 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 log transformation changes the shape of the Y_{ti} distribution in a way that obscures 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. As noted earlier, one can readily tell a priori if p is close to .5, because this can happen only if the observed Y distribution is itself close to symmetric.

7 Extension 1: Alternative Moments For Estimation

Here we provide additional moments that may be used for estimating the parameters h, b_0, b_1 and p.

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 for any positive $\tau \leq T$ there exists a constant α_τ such that the following two equations hold:

$$E\left[\exp\left(\tau\left(Y-h\right)\right)-\left(p\exp\left(\tau b_0\right)+\left(1-p\right)\exp\left(\frac{\tau b_0 p}{p-1}\right)\right)\alpha_\tau\right]=0 \tag{21}$$

$$E\left[\exp\left(-\tau\left(Y-h\right)\right)-\left(p\exp\left(-\tau b_0\right)+\left(1-p\right)\exp\left(\frac{-\tau b_0 p}{p-1}\right)\right)\alpha_\tau\right]=0 \tag{22}$$

Given a set of L positive values for τ, i.e., constants τ_1,...,τ_L, each of which are less than T, equations (21) and (22) provide $2L$ moment conditions satisfied by the set of $L+3$ parameters τ_1,..., τ_L, h, p, and b_0. Although the order condition for identification is therefore satisfied with $L \geq 3$, we do not have a proof analogous to Theorem 1 showing that the parameters are actually globally identified based on any number these moments. Note also that Corollary 1 is based on means of exponents, and so requires Y to have a thinner tailed distribution than estimation based on the polynomial equations (3) to (10).

In theory, parameter estimates based on GMM just using the moments given by equations (3), (21) and (22) for various value of τ might not be globally identified, and hence if these moments are used they should in theory only be employed along with the polynomial based moments to improve efficiency. However, in some simulations (see also Dong 2008) we found that estimation based just on moments in Corollary 1, letting τ_1,...,τ_L be about a dozen equally spaced values between 1 and 2.5, yielded estimates that were both reasonable and similar to those based on the identified polynomial moments.

Corollary 1 actually provides a continuum of moments, so rather than just choose a finite number of values for τ, it would also be possible to efficiently combine all the moments given by an interval of values of τ using, e.g., Carrasco and Florens (2000).

8 Extension 2: h depends on covariates

We now extend our results by permitting 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.

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 equations (21) and (22) also hold replacing h with $h(X)$. This suggests a couple of 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 equations (21) and (22) 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. Root n limiting distribution theory may be immediately obtained by applying generic two step estimation theorems as in Newey and McFadden (1994).

An alternative estimator is to note that, given the assumptions of Corollary 2, equations (3) to (10) and/or equations (21) and (22) (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, or includes τ_1,...,τ_L if equations (21) and (22) are used. 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. This is also in the form of the local GMM of Lewbel (2008), which may be applied as described in the next section and the Appendix.

After replacing $\widehat{h}$ with $\widehat{h}(X_i)$, equation (15) can be used to estimate the distribution of U, or alternatively equation (16) for $j = 1, ..., J$, replacing h with $h(X)$, can be included in the set of functions defining G in the estimated described above. Since ε has the same properties here as before, the estimator (19) will still work for estimating the density of U if it is continuous, using as 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 (20) would be replaced by $Y_i - \widehat{h}(X_i)$ and the result maximized over b_0, p, and s to estimate those parameters.

9 Extension 3: 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 $\{0, 1\}$. Assume that the distribution of $g(X, D^*)$ conditional upon X is symmetric. 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, equations (21) and (22), will all hold after replacing the parameters h, b_0, b_1, p, u_j, and τ_ℓ, and with functions $h(X)$, $b_0(X)$, $b_1(X)$, $p(X)$, $u_j(X)$, and $\tau_\ell(X)$ and replacing the unconditional expectations in these equations with conditional expectations, conditioning on $X = x$. If desired, 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

$$E[G(q(x), Y) \mid X = x)] = 0 \tag{23}$$

for a vector of known functions G. Equation (23) 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 conditioned upon, that is, $X = x$.

We summarize here how this local GMM estimator could be implemented, while Appendix B provides details regarding the associated limiting distribution theory.

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. 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.

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 (23).

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

4. Repeat these steps using 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.

Note that this local GMM estimator can be used when X contains both continuous and discretely distributed elements. If all elements of X are discrete, then the estimator simplifies back to Hansen's (1982) original GMM, as described in Appendix B.

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 (20) with respect to the parameters, that is, using the likelihood score functions as moments.

10 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 $2K + (S+1)/2$ parameters b_1, b_2,..., b_K, p_1, p_2, ...,p_K, u_2, u_4, ..., u_{S-1}, h.[1] 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,

[1]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.

$E(Y^s)$ 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 attempt a proof of identification with more than two points of support. However, assuming that 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

$$E\left[\exp\left(\tau\left(Y-h\right)\right)\right]=E\left[\exp\left(\tau V\right)\right]\alpha_{\tau}$$

with $\alpha_{\tau}=\alpha_{-\tau}$ for any τ for which these expectations exist, and therefore by choosing constants τ_1,...,τ_L, GMM estimation could be based on the $2L$ moments

$$\begin{aligned} E\left[\sum_{k=0}^{K}\left[\left[\exp\left(\tau_{\ell}\left(Y-h\right)\right)\right]-\exp\left(\tau_{\ell}b_k\right)\alpha_{\tau_{\ell}}\right]p_k\right] &= 0 \\ E\left[\sum_{k=0}^{K}\left[\left[\exp\left(-\tau_{\ell}\left(Y-h\right)\right)\right]-\exp\left(-\tau_{\ell}b_k\right)\alpha_{\tau_{\ell}}\right]p_k\right] &= 0 \end{aligned}$$

for $\ell=1,...,L$. The number of parameters b_k, p_k and $\alpha_{\tau_{\ell}}$ to be estimated would be $2K+L$, so taking $L>2K$ provides more moments than unknowns.

11 Conclusions

We have proved global point identification and provided estimators for the models $Y=h+V+U$, $Y=h(X)+V+U$, 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 provide some summary measures for assessing whether income convergence has taken place over time, and show that our estimator works well with real data and small sample sizes, despite involving high order data moments.

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] Ahmed, I. A. and Q. Li (1997), "Testing Symmetry of an Unknown Density by Kernel Method," Nonparametric Statistics, 7, 279-293.

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

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

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

[6] 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.

[7] 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.

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

[9] 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.

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

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

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

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

[14] 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.

[15] Honore, B. (1992),"Trimmed Lad and Least Squares Estimation of Truncated and Censored Regression Models with Fixed Effects," Econometrica, 60, 533-565.

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

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

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

[19] 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.

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

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

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

[23] 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.

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

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

[26] Simar, L. and P. W. Wilson (2007) "Statistical Inference in Nonparametric Frontier Models: Recent Developments and Perspectives," in The Measurement of Produc-

tive Efficiency, 2nd edition, chapter 4, ed. by H. Fried, C.A.K. Lovell, and S.S. Schmidt, Oxford: Oxford University Press.

12 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

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

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

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

$$\begin{aligned} E\left(\varepsilon^5\right) &= E\left(V^5 + 5V^4U + 10V^3U^2 + 10V^2U^3 + 5VU^4 + U^5\right) \\ e_5 &= v_5 + 10v_3E\left(U^2\right) = v_5 + 10v_3\left(e_2 - v_2\right) \\ e_5 &= v_5 + 10e_3e_2 - 10e_3v_2 \end{aligned}$$

$$e_5 - 10e_3e_2 = v_5 - 10e_3v_2$$

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

$$\begin{aligned} E\left(\varepsilon^6\right) &= E\left(V^6 + 6V^5U + 15V^4U^2 + 20V^3U^3 + 15V^2U^4 + 6VU^5 + U^6\right) \\ e_6 &= v_6 + 15v_4E\left(U^2\right) + 15v_2E\left(U^4\right) + E\left(U^6\right) \\ E\left(U^6\right) &= e_6 - v_6 - 15v_4E\left(U^2\right) - 15v_2E\left(U^4\right) \\ &= e_6 - v_6 - 15v_4\left(e_2 - v_2\right) - 15v_2\left(e_4 - v_4 - 6v_2e_2 + 6v_2^2\right) \\ &= e_6 - v_6 - 15e_2v_4 - 15e_4v_2 + 30v_2v_4 - 90v_2^3 + 90e_2v_2^2 \end{aligned}$$

$$\begin{aligned} E\left(\varepsilon^7\right) &= E\left(V^7 + 7V^6U + 21V^5U^2 + 35V^4U^3 + 35V^3U^4 + 21V^2U^5 + 7VU^6 + U^7\right) \\ e_7 &= v_7 + 21v_5E\left(U^2\right) + 35v_3E\left(U^4\right) \\ e_7 &= v_7 + 21v_5\left(e_2 - v_2\right) + 35v_3\left(e_4 - v_4 - 6v_2e_2 + 6v_2^2\right) \end{aligned}$$

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.

$$\begin{aligned} E\left(\varepsilon^9\right) &= E\left(\begin{array}{c} V^9 + 9V^8U + 36V^7U^2 + 84V^6U^3 + 126V^5U^4 + \\ 126V^4U^5 + 84V^3U^6 + 36V^2U^7 + 9VU^8 + U^9 \end{array}\right) \\ e_9 &= v_9 + 36v_7E\left(U^2\right) + 126v_5E\left(U^4\right) + 84v_3E\left(U^6\right) \\ e_9 &= v_9 + 36v_7\left(e_2 - v_2\right) + 126v_5\left(e_4 - v_4 - 6v_2e_2 + 6v_2^2\right) \end{aligned}$$

$$+84v_3\left(e_6 - v_6 - 15e_2v_4 - 15e_4v_2 + 30v_2v_4 - 90v_2^3 + 90e_2v_2^2\right)$$

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, $E\left(V\right) = b_0p_0 + b_1p_1 = 0$ because $\varepsilon = V + U$ and both ε and U have mean zero, so $b_1 = -b_0p_0/\left(1 - p_0\right)$. Let $r = p_0/p_1 = p_0/\left(1 - p_0\right)$, so

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

and for any integer d

$$v_d = b_0^dp_0 + b_1^dp_1 = b_0^d\left(p_0 + \left(-r\right)^dp_1\right) = b_0^d\frac{r + \left(-r\right)^d}{1 + r}$$

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 $e_3 = b_0^3 r\,(1-r)$,

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

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

$$\begin{aligned}
w &= v_9 - 84 e_3 v_6 - 126 s v_4 - 36 q v_2 \\
&= \begin{pmatrix}
b_0^9 r\,(1-r)\left(r^2+1\right)\left(r^4+1\right) - 84\left(b_0^3 r\,(1-r)\right)\left(b_0^6 r\left(r^4-r^3+r^2-r+1\right)\right) \\
-126\left(b_0^5 r\,(1-r)\left(r^2-10r+1\right)\right)\left(b_0^4 r\left(r^2-r+1\right)\right) \\
-36\left(b_0^7 r\,(1-r)\left(r^4-56r^3+246r^2-56r+1\right)\right) b_0^2 r
\end{pmatrix} \\
w &= b_0^9 r\,(1-r)\left(r^6-246r^5+3487r^4-10452r^3+3487r^2-246r+1\right)
\end{aligned}$$

Summarizing the results so far we have

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

$$w = b_0^9 r\left(1-r\right)\left(r^6-246r^5+3487r^4-10452r^3+3487r^2-246r+1\right)$$

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 $r^2-\gamma r+1=0$, 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:

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

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

$$\begin{aligned}
0 &= \left(\left(1-R\right)\left(S-Q-3184\right)+12096-4320R\right)r^2 \\
&\quad +\left(\left(1-R\right)\left(9392+2Q-2S\right)+22032R-53136\right)r \\
&\quad +\left(\left(1-R\right)\left(S-Q-3184\right)+12096-4320R\right).
\end{aligned}$$

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: $Y = h + V + U$ and independence of U and V implies that

$$E\left[\exp\left(\tau\left(Y-h\right)\right)\right] = E\left[\exp\left(\tau V\right)\right]E\left[\exp\left(\tau U\right)\right]$$

Now $E\left[\exp\left(\tau V\right)\right] = p\exp\left(\tau b_0\right) + \left(1-p\right)\exp\left(\tau b_1\right)$. Define $\alpha_\tau = E\left(e^{\tau U}\right)$. By symmetry of U, $\alpha_\tau = E\left(e^{-\tau U}\right)$ also. It follows from these equation that

$$\begin{aligned} E\left[\exp\left(\tau\left(Y-h\right)\right)\right] &= \left(p\exp\left(\tau b_0\right) + \left(1-p\right)\exp\left(\tau b_1\right)\right)\alpha_\tau \\ E\left[\exp\left(-\tau\left(Y-h\right)\right)\right] &= \left(p\exp\left(-\tau b_0\right) + \left(1-p\right)\exp\left(-\tau b_1\right)\right)\alpha_\tau \end{aligned}$$

which with $b_1 = b_0 p/\left(p-1\right)$ gives equations (21) and (22).

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

$$F_\varepsilon\left(u+b_1\right) = \left(1-p\right)F_U\left(u\right) + pF_U\left(u+b_1-b_0\right) \tag{24}$$

and evaluating at $\varepsilon = -u+b_0$ gives $F_\varepsilon(-u+b_0) = (1-p)F_U(-u-b_1+b_0)+pF_U(-u)$. Apply symmetry of U which implies $F_U(u) = 1-F_U(-u)$ to this last equation to obtain

$$F_\varepsilon(-u+b_0) = (1-p)[1-F_U(U+b_1-b_0)] + p[1-F_U(u)] \tag{25}$$

Equations (24) and (25) are two equations in the two unknowns $F_U(U+b_1-b_0)$ and $F_U(U)$. Solving for $F_U(U)$ gives $F_U(U) = \Psi(U)$ with $\Psi(U)$ given by equation (11). It follows from symmetry of U that $F_U(U)$ must also equal $1-\Psi(-U)$, which gives equation (12).

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.

13 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 (23). To motivate this estimator, which is closely related to Gozalo 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 (23) 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

$$\widehat{\theta}_x = \arg\min_{\theta_x} \sum_{i=1}^{n} G(\theta_x, Y_i)' I(X_i = x) \Omega_n \sum_{i=1}^{n} G(\theta_x, Y_i)' I(X_i = x) \tag{26}$$

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 (26) 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

$$\widehat{q}(x) = \arg \inf_{q(x) \in \Theta(x)} S_n(q(x))^T \Omega_n S_n(q(x)) \tag{27}$$

The scaling of the kernal estimator $S_n(q(x))$ by b^d is convenient for deriving the properties of the estimator, but is numerically unnecessary because omitting it leaves the minimized value $\widehat{q}(x)$ unchanged.

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 (26) and (27) 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).

Returns to Lying? Identifying the Effects of Misreporting When the Truth is Unobserved*

Yingyao Hu
Johns Hopkins University

Arthur Lewbel
Boston College

original April, 2007, revised June 2009

Abstract

Consider an observed binary regressor D and an unobserved binary variable D^*, both of which affect some other variable Y. This paper considers nonparametric identification and estimation of the effect of D on Y, conditioning on $D^* = 0$. For example, suppose Y is a person's wage, the unobserved D^* indicates if the person has been to college, and the observed D indicates whether the individual claims to have been to college. This paper then identifies and estimates the difference in average wages between those who falsely claim college experience versus those who tell the truth about not having college. We estimate this average effect of lying to be about 6% to 20%. Nonparametric identification without observing D^* is obtained either by observing a variable V that is roughly analogous to an instrument for ordinary measurement error, or by imposing restrictions on model error moments.

JEL Codes: C14, C13, C20, I2. Keywords: Binary regressor, misclassification, measurement error, unobserved factor, discrete factor, program evaluation, treatment effects, returns to schooling, wage model.

*We would like to thank Xiaohong Chen for her help on this paper. We also thank anonymous referees and participants of UCL, IFS, CEMMAP, CREST, BC, LAMES, UCSD, Montreal, York, and Brown seminars for helpful comments, and Douglas Staiger for providing data. Any errors are our own.

Department of Economics, Johns Hopkins University, 440 Mergenthaler Hall, 3400 N. Charles Street, Baltimore, MD 21218, USA Tel: 410-516-7610. Email: yhu@jhu.edu, http://www.econ.jhu.edu/people/hu/

Department of Economics, Boston College, 140 Commonwealth Avenue, Chestnut Hill, MA 02467 USA. Tel: 617-552-3678. email: lewbel@bc.edu http://www2.bc.edu/~lewbel

1 Introduction

Consider an observed binary regressor D and an unobserved binary variable D^*, both of which affect some other variable Y. This paper considers nonparametric identification and estimation of the effect of D on Y, conditioning on a value of the unobserved D^* (and possibly on a set of other observed covariates X). Formally, what is identified is the function $R(D, X)$ defined by

$$R(D, X) = E(Y \mid D^* = 0, D, X).$$

This can then be used to evaluate

$$r(X) = R(1, X) - R(0, X)$$

and $r = E[r(X)]$, which are respectively, the conditional and unconditional effects of D on Y, holding D^* fixed. When D^* is observed, identification and estimation of R is trivial. Here we obtain identification and provide estimators when D^* is unobserved.

Assuming $E(Y \mid D^*, D, X)$ exists, define a model H and an error η by

$$Y = E(Y \mid D^*, D, X) + \eta = H(D^*, D, X) + \eta \tag{1}$$

where the function H is unknown and the error η is mean zero and uncorrelated with D, D^*, and X. Then, since D and D^* are binary, we may without loss of generality rewrite this model in terms of the unknown R, r, and an unknown function s as

$$Y = R(D, X) + s(D, X)D^* + \eta \tag{2}$$

or equivalently

$$Y = R(0, X) + r(X)D + s(D, X)D^* + \eta. \tag{3}$$

This paper provides conditions that are sufficient to point identify the unknown functions R and r, even though D^* is unobserved. We also show set (interval) identification under weaker assumptions.

For a specific example, suppose for a sample of individuals the observed D is one if an individual claims or is reported to have some college education (and zero otherwise), and the unobserved D^* is one if the individual actually has some college experience. Let Y be the individual's wage rate. Then r is the difference in average wages Y between those who claim to have a degree when they actually do not, versus those who honestly report not having a college degree. This paper provides nonparametric identification

and associated estimators of the function r. We empirically apply these methods to estimate this average difference in outcomes between truth tellers and liars, when the truth D^* is not observed. Notice that we are not focusing on the effects of misreporting on estimates of returns to schooling, as in, e.g., Ashenfelter and Krueger (1994), but rather on the direct effects of misreporting on wages.

Only responses and not intent can be observed, so we cannot distinguish between intentional lying and false beliefs about D^*. For example, suppose D^* as an actual treatment and D is a perceived treatment (i.e., D is the treatment an individual thinks he received, and so is a false belief rather than an intentional lie). Then r is the average placebo effect, that is, the average difference in outcomes between those who were untreated but believe they received treatment versus those who correctly perceive that they were untreated. This paper then provides identification and an estimator for this placebo effect when the econometrician does not observe who actually received treatment.

Given a Rubin (1974) type unconfoundedness assumption, r will equal the average placebo effect, or the average returns to lying (which could be positive or negative). Unconfoundedness may be a reasonable assumption in the placebo example, but is less likely to hold when lying is intentional. Without unconfoundedness, the difference r in outcomes Y that this paper identifies could be due in part to unobserved differences between truth tellers and liars. For example, r could be positive even if lying itself has no direct effect on wages, if those who misreport their education level are on average more aggressive in pursuing their goals than others, or if some of them have spent enough time and effort studying (more on average than other nongraduates) to rationalize claiming that they have college experience. Alternatively r could be negative even if the returns to lying itself is zero, if the liars are more likely to arouse suspicion, or if there exist other negative character flaws that correlate with misreporting. Even with unconfoundedness, r might not equal the true returns to lying if Y is self reported data and the propensity to misreport D^* is correlated with misreporting Y, e.g., individuals who lie about their education level may also lie about their income.

Given that unconfoundedness may often be implausible in this context, we will call r the "effects of lying," and use the phrase "returns to lying" only when unconfoundedness is assumed.

The interpretation of r as a placebo effect or as effects or returns to lying also assumes that D^* and D are respectively the true and reported values of the same variable. This paper's identification and associated estimator does not require D and D^* to be related in this way (they can be completely different binary variables), however, for the purposes of interpreting the required assumptions and associated results, we will throughout this paper refer to D as the reported value of a true D^*.

Discreteness of D and D^* is also not essential for this paper's identification method, but it does simplify

the associated estimators and limiting distribution theory. In particular, if we more generally have a reported Z and an unobserved Z^*, we could apply this paper's identification method for any particular values z and z^* of interest by letting $D^* = I(Z^* \neq z)$ and $D = I(Z \neq z)$, where I is the indicator function. Then $D = 1$ when $D^* = 0$ means lying by claiming a value z when the truth is not z. Although our identification theory still holds in that case, having D or D^* be zero could then be zero probability events, resulting in estimation problems analogous to weak instruments which we do not address here.

When D is a possibly mismeasured or misclassified observation of D^*, then $D - D^*$ is the measurement or misclassification error. Virtually all of the literature on mismeasured binary regressors (which goes back at least as far as Aigner 1973) that attempts to estimate or bound the effect of D^* on Y (a treatment effect) assumes $r(X) = 0$, or equivalently, that any misclassification or measurement errors have no effect on the outcome Y after conditioning on the true D^*. Examples include Ashenfelter and Krueger (1994), Kane and Rouse (1995), Card (1996), Bollinger (1996), Hotz, Mullin, and Sanders (1997), Klepper, (1988), Manski (1990), Hu (2006), Mahajan (2006), Lewbel (2007a), Chen, Hu, and Lewbel (2008a, 2008b), and Molinari (2008). The same is true for general endogenous binary regressor estimators when they are interpreted as arising from mismeasurement. See, e.g., Das (2004), Blundell and Powell (2004), Newey and Powell (2003), and Florens and Malavolti (2003). The assumption that $r(X) = 0$ may be reasonable if the reporting errors $D - D^*$ are due to data collection errors such as accidently checking the wrong box on a survey form. Having $r(X) = 0$ would also hold if the outcome Y could not be affected by the individual's beliefs or reports regarding D, e.g., if D^* were an indicator of whether the individual owns stock and Y is the return on his investment, then that return will only depend on the assets he actually owns and not on his beliefs or self reports about what he owns. Still, there are many applications where it is not reasonable to assume a priori that $r(X)$ is zero, so even when $r(X)$ is not of direct interest, it may be useful to apply this paper's methods to test if it is zero, which would then permit the application of the existing mismeasured or misclassified binary regressor estimators that require that $r(X) = 0$.

We propose two different methods of obtaining nonparametric identification without observing D^*. One is by observing a variable V that has some special properties, analogous to an instrument. The second way we obtain identification is through restrictions on the first three moments of the model error η. Identification using an instrument V requires V to have some of the properties of a repeated measurement. In particular, Kane and Rouse (1995) and Kane, Rouse, and Staiger (1999) obtain data on both self reports of educational attainment D, and on transcript reports. They provide evidence that this transcript data (like the self reports D) may contain considerable reporting errors on questions like, "Do you have some years of college?"

These transcript reports therefore cannot be taken to equal D^*, but we show these transcripts may satisfy the conditions we require for use as an instrument V.

The alternative method we propose for identification does not require an instrument V, but is instead based primarily on assuming that the first three moments of the model error η are independent of the covariates. For example, if η is normal, as might hold by Gibrat's (1931) law for Y being log wages, and homoskedastic, then η will satisfy this assumption. This second method of identification is similar to Chen, Hu, and Lewbel (2008a, 2008b), though (as we will show later) those papers could not be used to identify the effects of lying in our context without additional information.

The next two sections describe identification with and without an instrument. We then propose estimators based on each of these methods of identification, and provide an empirical application estimating the effects on wages of lying about educational attainment.

2 Identification Using an Instrument

ASSUMPTION A1: *The variable Y, the binary variable D, and a (possibly empty) vector of other covariates X are all observable. The binary variable D^* is unobserved. $E(Y \mid D^*, D, X)$ exists. The functions H, R, r, s and the variable η are defined by equations (1), (2) and (3).*

ASSUMPTION A2: *A variable V is observed with*

$$E(\eta V \mid D, X) = 0, \tag{4}$$

$$E\left(V \mid D, D^* = 1, X\right) = E\left(V \mid D^* = 1, X\right), \tag{5}$$

$$E(V \mid D = 1, X) \neq E(V \mid X). \tag{6}$$

Equation (4) says that the instrument V is uncorrelated with the model error η for any value of the observable regressors D and X. A sufficient condition for equation (4) to hold is if $E(Y \mid D^*, D, X, V) = E(Y \mid D^*, D, X)$. This is a standard property for an instrument. The following very simple Lemmas are useful for interpreting and applying the other equations that comprise Assumption A2:

LEMMA 1: *Assume $E(D \mid D^* = 1, X) \neq 0$. Equation (5) holds if and only if*

$$Cov\left(D, V \mid D^* = 1, X\right) = 0 \tag{7}$$

LEMMA 2: *Assume* $E(D \mid X) \neq 0$. *Equation (6) holds if and only if*

$$Cov(D, V \mid X) \neq 0. \tag{8}$$

Proofs of Lemmas and Theorems are in the Appendix. As shown by Lemmas 1 and 2, equations (5) and (6) say that D and V are correlated, but at least for $D^* = 1$, this relationship only occurs through D^*. Equation (5) means that when $D^* = 1$, the variable D has no additional power to explain V given X. If V is a second mismeasurement of D^*, then (5) or its equivalent (7) is implied by a standard assumption of repeated measurements, namely, that the error in the measurement D be unrelated to the error in the measurement V, while equation (6) can be expected to hold because both measurements are correlated with the true D^*. Equation (6) is close to a standard instrument assumption, if we are thinking of V as an instrument for D (since we are trying to identify the effect of D on Y). Note that equation (6) or Lemma 2 can be easily tested, since they only depend on observables.

To facilitate interpretation of the identifying assumptions, we discuss them in the context of the example in which Y is a wage, D^* is the true indicator of whether an individual has some college experience, D is the individual's self report of college experience, and V is transcript reports of educational attainment, which are an alternative mismeasure of D^*. Let X denote a vector of other observable covariates we may be interested in that can affect either wages, schooling, and/or lying, so X could include observed attributes of the individual and of her job.

In the college and wages example, equation (4) will hold if wages depend on both actual and self reported education, i.e., D^* and D, but not on the transcript reports V. This should hold if employers rely on resumes and worker's actual knowledge and abilities, but don't see college transcripts. Equation (5) or equivalently (7) makes sense, in that errors in college transcripts depend on the actual D^*, but not on what individuals later self report. However, this assumption could be violated if individuals see their own transcripts and base their decision to lie in part on what the transcripts say. Finally, (6) is likely to hold assuming transcripts and self reports are accurate enough on average to both be positively correlated with the truth.

Define the function $g(X)$ by

$$g(X) = E\left(V \mid D^* = 1, X\right).$$

THEOREM 1: *If Assumptions A1 and A2 hold then* $R(D, X)$ *satisfies*

$$R(D, X) = \frac{E(YV \mid D, X) - E(Y \mid D, X)\, g(X)}{E(V \mid D, X) - g(X)}. \tag{9}$$

and $r(X) = R(1, X) - R(0, X)$ satisfies

$$r(X) = E(Y \mid D = 1, X) - E(Y \mid D = 0, X) + \frac{cov(Y, V \mid D = 0, X)}{g(X) - E(V \mid D = 0, X)} - \frac{cov(Y, V \mid D = 1, X)}{g(X) - E(V \mid D = 1, X)}. \tag{10}$$

We now consider set identification of $r(X)$ based on equation (10), and then follow that with additional assumptions that suffice for point identification of $R(D, X)$, and hence of $r(X)$, based on equation (9).

2.1 Set Identification Bounds Using an Instrument

ASSUMPTION A3: *Assume that* $0 \leq E(V \mid D = 0, X) < E(V \mid D = 1, X) \leq g(X)$

Assumption A3 is a very mild set of inequalities. Having the support of V be nonnegative suffices to make the expectations in Assumption A3 nonnegative. $E(V \mid D = 0, X) < E(V \mid D = 1, X)$ essentially means that self reports are positively correlated with the instrument, which should hold since both would typically be positively correlated with the truth. In the college example, this inequality is equivalent to $\Pr(V = 1 \mid D = 0, X) < \Pr(V = 1 \mid D = 1, X)$, meaning that people reporting going to college are more likely to have a transcript that says they went to college than people who report not going to college. Given equation (7), violation of this inequality would require a relatively large fraction of people to reverse lie, that is, claim to not have college when they have in fact gone to college.

Define $\delta^*(X)$ by

$$\delta^*(X) = g(X) - E(V \mid D = 1, X)$$

So the last inequality in Assumption A3 is $\delta^*(X) \geq 0$. When V is a mismeasure of D^*, having $\delta^*(X) \geq 0$ is equivalent to $\Pr(V = 1 \mid D = 1, X) \leq \Pr(V = 1 \mid D^* = 1, X)$, which basically says that the instrument is closer to the truth than to the self report. This holds if a transcript is more likely to say you went to

college when you are in the set of people that actually did go to college than when you are in the set of people that claimed to have been to college. It can also be readily shown that this last equality holds if $\Pr(V = 1 \mid D = 1, D^* = 1, X) \geq \Pr(V = 1 \mid D = 1, D^* = 0, X)$, which means that among people who claim college, those who actually went to college have a higher chance of their transcript saying they went to college than those that who's claims to college are misreports. As with some earlier assumptions, this assumption in any of its forms will hold if people's decision to lie, or accidental misreporting, is unrelated to transcript errors.

COROLLARY 1.: *Let Assumptions A1, A2, and A3 hold. Then $r(X)$ lies in an identified interval that is bounded from below if $cov(Y, V \mid D = 0, X) > 0$ and bounded from above if $cov(Y, V \mid D = 0, X) < 0$. If there exists an identified positive $\delta(X)$ such that $\delta(X) \leq \delta^*(X)$ then $r(X)$ lies in an identified bounded interval.*

Corollary 1 provides bounds on $r(X)$ whether an identified $\delta(X)$ exists or not, but the bounds are improved given a $\delta(X)$. For an example of a $\delta(X)$, suppose that $E(V \mid D^* = 1, X) = E(V \mid D^* = 1)$, that is, the probability that a school produces the transcript error $V = 0$ when $D^* = 1$ is unrelated to an individual's observed attributes X, e.g., this would hold if all college graduates are equally likely to have the school lose their file or otherwise mistakenly report that they are not graduates. Then $g(X)$ is independent of X, and $\delta(X) = \sup_x E(V \mid D = 1, X = x) - E(V \mid D = 1, X)$ which may be strictly positive for many values of X.

Corollary 1 follows immediately from inspection of equation (10), as does the construction of bounds for $r(X)$. All of the terms on the right of equation (10) are moments of observable data, and hence are identified, except for $g(X)$. By Assumption A3, a lower bound on $g(X)$ is $E(V \mid D = 1, X)$. An upper bound of $g(X)$ is $\sup\left[supp(V)\right]$, since $g(X)$ is an expectation of V and so cannot exceed the largest value V can take on. Note that when V is a mismeasure of D^* as in the college example, this upper bound of $g(X)$ is one. From Assumptions A1 and A2, all of the expectations and covariances on the right of equation (10) exist. The function $g(x)$ appears only in the denominators of the last two terms in equation (10). By Assumption A3, the third term in equation (10) lies in the interval bounded by the two points

$$\frac{cov(Y, V \mid D = 0, X)}{E(V \mid D = 1, X) - E(V \mid D = 0, X)} \quad \text{and} \quad \frac{cov(Y, V \mid D = 0, X)}{\sup\left[supp(V)\right] - E(V \mid D = 0, X)}$$

Both of which are finite. Similarly, the last term in equation (10) lies in the interval bounded by the two

points

$$\frac{cov\left(Y, V \mid D=1, X\right)}{\delta^{*}\left(X\right)} \quad \text{and} \quad \frac{cov\left(Y, V \mid D=1, X\right)}{\sup\left[supp\left(V\right)\right]-E\left(V \mid D=1, X\right)}$$

The second of these points is finite. Given only assumptions A1, A2, and A3, $\delta^{*}\left(X\right) \leq 0$ so the first of the above points can be infinite. Whether it is plus or minus infinity, and hence whether we only have a lower or upper bound for $r\left(X\right)$, depends on the sign of $cov\left(Y, V \mid D=1, X\right)$. If we have a $\delta\left(X\right)$ with $0<\delta\left(X\right) \leq \delta^{*}\left(X\right)$, then we instead obtain the finite bound $cov\left(Y, V \mid D=1, X\right)/\delta\left(X\right)$.

To construct the identified interval that contains $r\left(X\right)$, we must consider four cases corresponding to the four possible pairs of signs that $cov\left(Y, V \mid D=0, X\right)$ and $cov\left(Y, V \mid D=1, X\right)$ can take on. Note that the denominators of the last two terms in equation (10) are positive. If $cov\left(Y, V \mid D=0, X\right)$ and $cov\left(Y, V \mid D=1, X\right)$ have opposite signs, then $r\left(X\right)$ is strictly increasing or decreasing in $g\left(X\right)$, so the interval that $r\left(X\right)$ can lie in is bounded by equation (10) evaluated at the lowest and highest values $g\left(X\right)$ can take on, the highest being $\sup\left[supp\left(V\right)\right]$ and lowest either $E\left(V \mid D=1, X\right)$ or $E\left(V \mid D=1, X\right)+\delta\left(X\right)$ if a $\delta\left(X\right)$ is known. If $cov\left(Y, V \mid D=0, X\right)$ and $cov\left(Y, V \mid D=1, X\right)$ have the same signs, then these could still be bounds on $r\left(X\right)$, but it is also possible in that case that $r\left(X\right)$ either first increases and then decreases in $g\left(X\right)$ or vice versa, in which case the point where the derivative of $r\left(X\right)$ with respect to $g\left(X\right)$ equals zero may also be a bound.

Although Assumption A3 is already rather weak, one could similarly obtain a looser bound by replacing it with the weaker assumption that $0 \leq E\left(V \mid D^{*}=0, X\right) \leq E\left(V \mid D^{*}=1, X\right)$. This is little more than the assumption that transcripts be right more often than they are wrong, that is, people with college education will have a higher probability of transcripts reporting college education than those without college education.

2.2 Point Identification Using an Instrument

We now consider additional assumptions that permit point identification of $r\left(X\right)$.

COROLLARY 2: *Let Assumptions A1 and A2 hold. Assume the function* $g\left(X\right)$ *is known* and $E\left(V \mid D, X\right) \neq g\left(X\right)$. *Then* $R(D, X)$ *is identified by*

$$R(D, X)=\frac{E\left(Y\left(V-g\left(X\right)\right) \mid D, X\right)}{E\left(\left(V-g\left(X\right)\right) \mid D, X\right)} \tag{11}$$

Identification of $r(X)$ is then given by $r(X) = R(1, X) - R(0, X)$. Corollary 2 follows immediately from Theorem 1 by substituting $g(X)$ for $E(V \mid D^* = 1, X)$ in equation (9), and observing that all the other terms in equation (9) are expectations of observables, conditioned on other observables, and hence are themselves identified. One way Corollary 2 might hold is if a form of validation data exists. For example if D and D^* refer to graduating from college, then $g(X)$ could be obtained from a survey of transcripts just of people known to have graduated college. A special case of this assumption holding is if V is a mismeasure of D^*, as when V is the transcript report, and $g(X) = 1$, that is, if transcript errors of the form $V = 0$ when $D^* = 1$ are ruled out.

Another example or variant of Corollary 2 is the following.

ASSUMPTION A3: *There exists an x_1 such that*

$$E\left(V \mid D^* = 1, X\right) = E\left(V \mid X = x_1\right) \tag{12}$$

and

$$E(V \mid D, X) \neq E(V \mid X = x_1) \tag{13}$$

Equation (12) assumes that V has the same mean for people who have $X = x_1$ as for people that have $D^* = 1$ and any value of X. One set of sufficient (but stronger than necessary) conditions for equation (12) to hold is if $E(V \mid D^* = 1, X = x_1) = E(V \mid D^* = 1)$, so for people having college ($D^* = 1$), the probability of a transcript error is unrelated to one's personal attribute information X, and if

$$\Pr\left(D^* = 1 \mid X = x_1\right) = 1, \tag{14}$$

so people who have $X = x_1$ are an observable subpopulation that definitely have some college. In our application, we use Corollary 3 below for identification and we take this subpopulation x_1 to be individuals with very high test scores and self reported advanced degrees. Note that if equation (14) holds then equation (12) would only be violated if colleges systematically made more or fewer errors when producing transcripts for individuals with attributes $X = x_1$ than for students with other attribute values.

Equation (13) is a technicality that, analogous to the assumption that $E(V \mid D, X) \neq g(X)$ in Corollary 2, will avoid division by zero in Corollary 3 below. It is difficult to see why it should not hold in general, and it is empirically testable since it depends only on observables. However, if both equations (12) and (14)

hold then equation (13) will not hold for $X = x_1$. This means that $R(D, x_1)$ cannot be identified in this case, though we still identify $R(D, X)$ for $X \neq x_1$. This is logical because if all individuals having $X = x_1$ have $D^* = 1$ by equation (14), then none of them can be lying when reporting $D = 1$.

COROLLARY 3: *If Assumptions A1, A2, and A3 hold then* $R(D, X)$ *is identified by*

$$R(D, X) = \frac{E(YV \mid D, X) - E(Y \mid D, X) E(V \mid X = x_1)}{E(V \mid D, X) - E(V \mid X = x_1)}. \tag{15}$$

Corollary 3 follows Theorem 1, by substituting equation (12) into equation (9) to obtain equation (15), and equation (13) makes the denominator in equation (15) be nonzero.

Given identification of $R(D, X)$ by Corollary 2 or 3, the effects of lying $r(X)$ is also identified by $r(X) = R(1, X) - R(0, X)$.

Although rather more difficult to interpret and satisfy than the assumptions in Corollaries 2 and 3, yet another alternative set of identifying assumptions is equations (4), (6) and $Cov(D^*, V \mid D, X) = 0$, which by equation (3) implies $Cov(Y, V \mid X) = r(X)Cov(D, V \mid D, X)$ which can then be solved for, and hence identifies, $r(X)$.

3 Identification Without an Instrument

We now consider identification based on restrictions on moments of η rather than on the presence of an instrument. In particular, we will assume that the second and third moments of η do not depend on D^*, D, and X. The method of identification here is similar to that of Chen, Hu, and Lewbel (2008b), though that paper imposes the usual measurement error assumption that the outcome Y is conditionally independent of the mismeasure D, conditioning on the true D^*, or equivalently, it assumes that $r(X) = 0$. One could modify Chen, Hu, and Lewbel (2008b) to identify our effects of lying model in part by including D in the list of regressors and treating our V from the previous section as the observed mismeasure of D^*. However, in that case one would need both an instrument V with certain properties and restrictions on higher moments of η, while in the present paper these are alternative methods of identification.

ASSUMPTION B2:

$$E\left(\eta \mid D^*, D, X\right) = 0, \tag{16}$$

$$E\left(\eta^k \mid D^*, D, X\right) = E\left(\eta^k\right) \quad \text{for } k = 2, 3, \tag{17}$$

there exists an x_0 *such that*

$$\Pr\left(D = 0 \mid D^* = 1, X = x_0\right) = 0 \quad \text{and} \quad \Pr\left(D = 0 \mid X = x_0\right) > 0, \tag{18}$$

and

$$E\left(Y \mid D^* = 1, D, X\right) \geq E\left(Y \mid D^* = 0, D, X\right) \tag{19}$$

Equation (16) can be assumed to hold without loss of generality by definition of the model error η. Equation (17) says that the second and third moments of the model error η do not depend on D^*, D, X, and so would hold under the common modeling assumption that the error η in a wage equation is independent of the regressors.

Equation (18) implies that people, or at least those in some subpopulation $\{X = x_0\}$, will not underreport and claim to not have been to college if they in fact have been to college. At least in terms of wages, this is plausible in that it is hard to see why someone would lie to an employer by claiming to have less education or training than he or she really possesses.

Finally, equation (19) implies that the impact of D^* on Y conditional on D and X is known to be positive. This makes sense when Y is wages and D^* is the true education level, since ceteris paribus, higher education on average should result in higher wages on average.

Define

$$\begin{aligned}
\sigma^2_{Y|D,X}(D, X) &= E\left(Y^2|D, X\right) - \left[E\left(Y|D, X\right)\right]^2, \\
v^3_{Y|D,X}(D, X) &= E\left(\left[Y - E\left(Y|D, X\right)\right]^3 |D, X\right),
\end{aligned}$$

$$\begin{aligned}
\alpha(D, X) &= \sigma^2_{Y|D,X}(D, X) - \sigma^2_{Y|D,X}(0, x_0), \\
\beta(D, X) &= v^3_{Y|D,X}(D, X) - v^3_{Y|D,X}(0, x_0) + 2E\left(Y|D, X\right)\alpha(D, X), \\
\gamma\left(D, X\right) &= \alpha(D, X)^2 + \left[E\left(Y|D, X\right)\right]^2 \alpha(D, X) - E\left(Y|D, X\right)\beta(D, X).
\end{aligned}$$

THEOREM 2: *Suppose that Assumptions A1 and B2 hold and that* $\alpha(D, X) \neq 0$ *for* $(D, X) \neq (0, x_0)$. *Then,* $R(D, X)$ *and* $s(D, X)$ *are identified as follows:*

i) if $(D, X) = (0, x_0)$*, then* $R(D, X) = E(Y|D, X)$*;*

ii) if $(D, X) \neq (0, x_0)$*, then*

$$R(D, X) = \frac{\beta(D, X) - \sqrt{\beta(D, X)^2 + 4\alpha(D, X)\gamma(D, X)}}{2\alpha(D, X)},$$

and

$$s(D, X) = \frac{\alpha(D, X)}{E(Y|D, X) - R(D, X)} + E(Y|D, X) - R(D, X).$$

As before given $R(D, X)$ we may identify the effects of lying $r(X)$ using $r(x) = R(1, X) - R(0, X)$. Identification of $s(D, X)$ in Theorem 2 means that the entire conditional mean function H in equation 1 is identified.

Some intuition for this identification comes from observing that, conditional on X, the number of equality constraints imposed by the assumptions equal the number of unknowns. One of these equations is a quadratic, and the inequality (19) is only needed to identify which of the two roots is correct. Based on this intuition, identification based on alternative equality restrictions should be possible, e.g., in place of equation (18) one could consider the constraint that the third moment $E\left(\eta^3\right)$ equal zero. Also, dropping inequality assumptions like (19) will result in set rather than point identification, where the sets are finite and consist of only two or three possible values.

4 Unconfoundedness

By construction the function $r(X)$ is the difference in the conditional mean of Y (conditioning on D, X, and on $D^* = 0$) when D changes from zero to one. Assuming D is the reported response and D^* is the truth, here we formally provide the unconfoundedness condition required to have this $r(X)$ equal the returns to lying. Consider the weak version of the Rubin (1974) or Rosenbaum and Rubin (1984) unconfoundedness assumption given by equation (20), interpreting D as a treatment. Letting $Y(d)$ denote what Y equals given the response $D = d$, if

$$E[Y(d) \mid D, D^* = 0, X] = E[Y(d) \mid D^* = 0, X] \tag{20}$$

then it follows immediately from applying, e.g., Heckman, Ichimura, and Todd (1998), that $E[Y(1) - Y(0) \mid D^* = 0, X] = r(X)$ is the conditional average effect of D, and so is the conditional on X average returns to lying.

5 Estimation Using an Instrument

We now provide estimators of $R(D, X)$ and hence of $r(X)$ following from Corollary 2 or 3 of Theorem 1. We first describe nonparametric estimators that are based on ordinary sample averages, which can be used if X is discrete. We then discuss kernel based nonparametric estimation, and finally we provide a simple least squares based semiparametric estimator that does not require any kernels, bandwidths, or other smoothers regardless of whether X contains continuous or discrete elements.

5.1 Nonparametric, Discrete X Estimation

Note that while identification only requires Assumption A3 to hold for a single value of X, that is, x_1, it may be the case that this assumption is known to hold for a range of values of x_1. We may then replace $E(V \mid X = x_1)$ with the expected value of V conditional on X equalling any value in this range. This may then improve the accuracy with which we can estimate this conditional expectation. In particular if X has any continuous components then $E(V \mid X = x_1)$ for a single value of x_1 is conditioning on a zero probability event, the estimate of which will converge at a slower rate than conditioning on a range of values X that has nonzero probability. Therefore, define U_i to be a dummy variable such that

$$U_i = I\left(X_i \in \{x_1\text{: Assumption A3 is known to hold}\}\right), \tag{21}$$

where $I(.)$ is the indicator function. In other words, let U_i equal one if equations (12) and (13) are assumed to hold when replacing x_1 in those equations with X_i , otherwise let U_i equal zero. It then follows immediately from Corollary 3 that equation (15) holds replacing $E(V \mid X = x_1)$ with $E(V \mid U = 1)$, so

$$R(D, X) = \frac{E(YV \mid D, X) - E(Y \mid D, X)\, E(V \mid U = 1)}{E(V \mid D, X) - E(V \mid U = 1)}. \tag{22}$$

We first consider estimation in the simple case where X is discrete. Replacing the expectations in equation (22) with sample averages in this case gives the estimators

$$\widehat{R}(d, x) = \frac{\widehat{\mu}_{Y,V,X,d} - \widehat{\mu}_{Y,X,d}\widehat{\mu}}{\widehat{\mu}_{V,X,d} - \widehat{\mu}_{X,d}\widehat{\mu}}, \quad \widehat{r}(x) = \widehat{R}(1, x) - \widehat{R}(0, x). \tag{23}$$

with

$$\begin{aligned}
\widehat{\mu}_{Y,V,X,d} &= \frac{1}{n}\sum_{i=1}^{n} Y_i V_i I(X_i = x, D_i = d), \quad \widehat{\mu}_{Y,X,d} = \frac{1}{n}\sum_{i=1}^{n} Y_i I(X_i = x, D_i = d), \\
\widehat{\mu}_{V,X,d} &= \frac{1}{n}\sum_{i=1}^{n} V_i I(X_i = x, D_i = d), \quad \widehat{\mu}_{X,d} = \frac{1}{n}\sum_{i=1}^{n} I(X_i = x, D_i = d), \\
\widehat{\mu}_{V,U} &= \frac{1}{n}\sum_{i=1}^{n} V_i U_i, \quad \widehat{\mu}_{U} = \frac{1}{n}\sum_{i=1}^{n} U_i, \quad \widehat{\mu} = \widehat{\mu}_{V,U}/\widehat{\mu}_{U}
\end{aligned}$$

Estimation based on equation (11) is the same replacing $\widehat{\mu}$ with $g(X)$ in equation (23)

One may also consider the unconditional mean wages $R_d = E[R(d, X)]$ and unconditional average effects of lying $r = E[r(X)]$, which may be estimated by

$$\widehat{R}_d = \frac{1}{n}\sum_{i=1}^{n} \widehat{R}(d, X_i), \quad \widehat{r} = \frac{1}{n}\sum_{i=1}^{n} \widehat{r}(X_i). \tag{24}$$

Assuming independent, identically distributed draws of $\{Y_i, V_i, X_i, D_i, U_i\}$, and existence of relevant variances, it follows immediately from the Lindeberg-Levy central limit theorem and the delta method that $\widehat{R}(d, x)$, $\widehat{r}(x)$, $\widehat{R}_d$, and $\widehat{r}$ are root n consistent and asymptotically normal, with variance formulas as provided in the appendix, or that can be obtained by an ordinary bootstrap. Analogous limiting distribution results will hold with heteroskedastic or dependent data generating processes, as long as a central limit theorem still applies.

5.2 General Nonparametric Estimation

Letting $\mu = E(V \mid U = 1)$, equation (22) can be rewritten as

$$R(D, X) = \frac{E[Y(V - \mu) \mid D, X]}{E[(V - \mu) \mid D, X]}. \tag{25}$$

Equation (11) can also be written in the form of equation (25) by replacing μ with $g(X)$.

Assume n independent, identically distributed draws of $\{Y_i, V_i, X_i, D_i, U_i\}$. Let $X_i = (Z_i, C_i)$ where Z and C are, respectively, the vectors of discretely and continuously distributed elements of X. Similarly let $x = (z, c)$. Let $\widehat{\mu} = \widehat{\mu}_{V,U}/\widehat{\mu}_{U}$ if estimation is based on equation (22), otherwise replace $\widehat{\mu}$ with $g(x)$. Using equation (25), a kernel based estimator for $R(D, X)$ is

$$\widehat{R}(d, x) = \frac{\Sigma_{i=1}^{n} Y_i (V_i - \widehat{\mu}) K[(C_i = c)/b] I(Z_i = z) I(D_i = d)}{\Sigma_{i=1}^{n} (V_i - \widehat{\mu}) K[(C_i = c)/b] I(Z_i = z) I(D_i = d)} \tag{26}$$

where K is a kernel function and b is a bandwidth that goes to zero as n goes to infinity. Equation (26) is numerically identical to the ratio of two ordinary nonparametric Nadaraya-Watson kernel regressions of $Y(V-\widehat{\mu})$ and $V-\widehat{\mu}$ on X, D, which under standard conditions are consistent and asymptotically normal. These will have the same slower than root n rate of convergence as regressions that use a known μ in place of the estimator $\widehat{\mu}$, because an estimated $\widehat{\mu}$ converges at the rate root n by the law of large numbers. Alternatively, equation (25) can be rewritten as the conditional moment

$$E\left[\left(Y-R(D,X)\right)(V-\mu)\mid D,X\right]=0 \tag{27}$$

which may be estimated using, e.g., the functional GMM estimator of Ai and Chen (2003), or by Lewbel's (2007b) local GMM estimator , with limiting distributions as provided by those references.

Given $\widehat{R}(d,x)$ from equation (26) we may as before construct $\widehat{r}(x)=\widehat{R}(1,x)-\widehat{R}(0,x)$, and unconditional estimates $\widehat{R}_d$ and $\widehat{r}$ by equation (24). We also construct trimmed unconditional effects $\widehat{r}_t=\frac{1}{n}\sum_{i=1}^{n}\widehat{r}(X_i)I_{ti}$ and similarly for $\widehat{R}_{dt}$, where I_{ti} is a trimming parameter that equals one for most observations i, but equals zero for tail observations. Assuming regularity conditions such as Newey (1994) these trimmed unconditional effects are root n consistent and asymptotically normal estimates of the trimmed means r_t and R_{dt}.

5.3 Simple Semiparametric Estimation

Assume we have a parameterization $R(D,X,\theta)$ for the function $R(D,X)$ with a vector of parameters θ. The function $s(D,X)$ and the distribution of the model error η are not parameterized. Then based on the definition of μ and equation (27), θ and μ could be jointly estimated based on Corollary 3 by applying GMM to the moments

$$E\left[(V-\mu)U\right]=0 \tag{28}$$

$$E\left[\psi(D,X)\left(Y-R(D,X,\theta)\right)(V-\mu)\right]=0 \tag{29}$$

for a chosen vector of functions $\psi(D,X)$. For estimation based on Corollary 2, the estimator would just use the moments given by equation (29) replacing μ with $g(X)$.

Let $W=\left(1,D,X'\right)'$. If R has the linear specification $R(D,X,\theta)=W'\theta$ then let $\psi(D,X)=W$ to yield moments $E\left[W\left(Y-W'\theta\right)(V-\mu)\right]=0$, so $\theta=E\left[(V-\mu)WW'\right]^{-1}E\left[(V-\mu)WY\right]$. This then

yields a weighted linear least squares regression estimator

$$\widehat{\theta} = \left[\sum_{i=1}^{n} (V_i - \widehat{\mu})\, W_i W_i'\right]^{-1} \left[\sum_{i=1}^{n} (V_i - \widehat{\mu})\, W_i Y_i\right] \tag{30}$$

based on Corollary 3, or the same expression replacing $\widehat{\mu}$ with $g(X_i)$ based on Corollary 2. Given $\widehat{\theta}$ we then have $\widehat{R}(D, X) = W'\widehat{\theta}$. In this semiparametric specification $r(x)$ is a constant with $\widehat{r}(x) = \widehat{r} = \widehat{\theta}_1$, the first element of $\widehat{\theta}$. Note that both GMM based on equation (29) and the special case of weighted linear regression based on equation (30) do not require any kernels, bandwidths, or other smoothers for their implementation.

6 Estimation Without an Instrument

We now consider estimation based on Theorem 2. As in the previous section, let K be a kernel function, b be a bandwidth, and $X_i = (Z_i, C_i)$ where Z and C are, respectively, the vectors of discretely and continuously distributed elements of X. Also let $x = (z, c)$. For $k = 1, 2, 3$, define

$$\widehat{E}\left(Y^k | D = d, X = x\right) = \frac{\Sigma_{i=1}^{n} Y_i^k K[(C_i = c)/b] I(Z_i = z) I(D_i = d)}{\Sigma_{i=1}^{n} K[(C_i = c)/b] I(Z_i = z) I(D_i = d)} \tag{31}$$

This is a standard Nadayara-Watson Kernel regression combining discrete and continuous data, which provides a uniformly consistent estimator of $E\left(Y^k | D = d, X = x\right)$ under standard conditions. Define

$$\begin{aligned}
\widehat{\sigma}^2_{Y|D,X}(d, x) &= \widehat{E}\left(Y^2 | D = d, X = x\right) - \left[\widehat{E}\left(Y | D = d, X = x\right)\right]^2, \\
\widehat{v}^3_{Y|D,X}(d, x) &= \widehat{E}\left(\left[Y - \widehat{E}\left(Y | D = d, X = x\right)\right]^3 | D = d, X = x\right),
\end{aligned}$$

$$\begin{aligned}
\widehat{\alpha}(d, x) &= \widehat{\sigma}^2_{Y|D,X}(d, x) - \widehat{\sigma}^2_{Y|D,X}(0, x_0), \\
\widehat{\beta}(d, x) &= \widehat{v}^3_{Y|D,X}(d, x) - v^3_{Y|D,X}(0, x_0) + 2\widehat{E}\left(Y | D = d, X = x\right)\widehat{\alpha}(d, x), \\
\widehat{\gamma}(d, x) &= \widehat{\alpha}(d, x)^2 + \left[\widehat{E}\left(Y | D = d, X = x\right)\right]^2 \widehat{\alpha}(d, x) - \widehat{E}\left(Y | D = d, X = x\right)\widehat{\beta}(d, x).
\end{aligned}$$

Based on Theorem 2 and uniform consistency of the kernel regressions, a consistent estimator of $R(d, x)$ is then

$$\widehat{R}(0, x_0) = \widehat{E}\left(Y | D = 0, X = x_0\right),$$

$$\widehat{R}(d,x) = \frac{\widehat{\beta}(d,x) - \sqrt{\widehat{\beta}(d,x)^2 + 4\widehat{\alpha}(d,x)\widehat{\gamma}(d,x)}}{2\widehat{\alpha}(d,x)} \text{ for } (d,x) \neq (0,x_0).$$

As before, every conditional expectation above that conditions on $X = x_0$ can be replaced by an expectation conditional on X equalling any value x having the property that the assumptions of Theorem 2 hold replacing x_0 with that value x.

If X does not contain any continuously distributed elements, then these estimators are smooth functions of cell means, and so are root n consistent and asymptotically normal by the Lindeberg Levy central limit theorem and the delta method. Given $\widehat{R}(d,x)$ from equation (26) we may as before construct $\widehat{r}(x) = \widehat{R}(1,x) - \widehat{R}(0,x)$, and unconditional effects $\widehat{R}_d$ and $\widehat{r}$ by equation (24). Also as before, root n consistent, asymptotically normal convergence of trimmed means of $\widehat{R}_d$ and $\widehat{r}$ is possible using regularity conditions as in Newey (1994) for two step plug in estimators.

7 Effects of Misreporting College Attainment

Here we report results of empirically implementing our estimators of $r(x)$ where D is self reports of schooling and Y is log wages. In this context, our effects of lying estimates should be interpreted only as the difference in means between accurate reporters and misreporters of college for a limited sample, rather than as actual returns to lying about schooling, for many reasons. First, our conditional mean estimates cannot control for the selection effects that are at the heart of the modern literature on wages and schooling going back at least to Heckman (1979). Similarly, unconfoundedness with respect to lying based on equation (20) may not hold. Also, people who misreport college may similarly misreport their wages. Our results may also differ from actual returns to lying by the fact that both the risks and the returns to misreporting on a survey are lower than for lying on a job application, though presumably the cost of potentially being caught in a lie in any context provides some incentive to report the same education level on a survey as was reported to one's employer. Finally, our sample may not be representative of the general population.

7.1 Preliminary Data Analysis

Kane, Rouse, and Staiger (1999) estimate a model of wages as a function of having either some college, an associate degree or higher, or a bachelors degree or higher. Their model also includes other covariates, and they use data on both self reports and transcript reports of education level. Their data is from the National

Longitudinal Study of High School Class of 1972 (NLS-72) and a Post-secondary Education Transcript Survey (PETS). We use their data set of $n = 5912$ observations to estimate the effects of lying, defining Y to be log wage in 1986, D to be one if an individual self reports having "some college" and zero otherwise, while V is one for a transcript report of having "some college" and zero otherwise (both before 1979). We also provide estimates where D and V are self and transcript reports of having an associate degree or more, and reports of having a bachelor's degree or more. We take X to be the same set of other regressors Kane, Rouse, and Staiger (1999) used, which are a 1972 standardized test score and zero-one dummy variables for female, black nonhispanic, hispanic, and other nonhispanic.

The means of D and V (which equal the fractions of our sample that report having that level of college or higher) are 0.6739 and 0.6539 for "some college," 0.4322 and 0.3884 respectively for "Associate degree," and 0.3557 and 0.3383 for "Bachelors degree." The average log wage Y is 2.228.

Table 1: Effects of Lying and Schooling Treating Transcripts as True

	Some college	Associate degree	Bachelor's degree
r if V=D*	0.1266 (0.03129)	0.2322 (0.02748)	0.1948 (0.04451)
r if V=D*, linear	0.07868 (0.02864)	0.1681 (0.02777)	0.1269 (0.04082)
s if V=D*	0.2831 (0.01366)	0.2958 (0.01288)	0.3181 (0.01280)
E(DV)	0.6204	0.3794	0.3325
E[D(1-V)]	0.05345	0.05277	0.02317
E[(1-D)V]	0.03349	0.008965	0.005751
E[(1-D)(1-V)]	0.2926	0.5589	0.6385

Standard Errors are in Parentheses

If D^* were observed along with Y and D, then the functions $r(x)$ and $s(d, x)$ could be immediately estimated from equation (3). Table 1 provides preliminary estimates of r and s based on this equation, under the assumption that transcripts have no errors. The row "r if V=D*" in Table 1 is the sample estimates of $E(Y|V = 0, D = 1) - E(Y|V = 0, D = 0)$, which would equal an estimate of $r = E[r(X)]$ if $V = D^*$, that is, if the transcripts V were always correct. The row, "r if V=D*, linear" is the coefficient of D in a linear regression of Y on D, V, and X, and so is another estimate of r that would be valid if if $V = D^*$ and given a linear model for log wages.

The third row of Table 1 is the sample analog of $E(Y|V=1) - E(Y|V=0)$, which if $V = D^*$ would be an estimate of the effects of schooling $s = E[s(D, X)]$ (that is, the difference in conditional means of log wages between those with $D^* = 1$, versus those with $D^* = 0$, which equals returns to schooling if the effects of schooling satisfy an unconfoundedness condition). In this and all other tables, standard errors are obtained by 400 bootstrap replications, and are given in parentheses.

Table 1 also shows the fraction of truth tellers and liars, if the transcripts V were always correct. The rows labeled E(DV) and E[(1-D)(1-V)] give the fraction of observations where self and transcript reports agree that the individual respectively either has or does not have the given level of college. The row labeled E[D(1-V)] gives the fraction of relevant liars if the transcripts are correct, that is, it is the fraction who claim to have the given level of college, $D = 1$, while their transcripts say they do not, $V = 0$. This fraction is a little over 5% of the sample for some college or Associate degree, but only about half that amount appear to misreport having a Bachelor's degree.

If V has no errors, then Table 1 indicates a small amount of lying in the opposite direction, given by the row labeled E[(1-D)V]. These are people who self report having less education than is indicated by their transcripts, ranging from a little over half a percent of the sample regarding college degrees to almost 3% for "some college." It is difficult to see a motive for lying in this direction, which suggests ordinary reporting errors in self reports, transcript reports, or both.

Prior to estimating $r(x)$, we examined equation (6) of Assumption A2, which is testable. A sufficient condition for equation (6) to hold is that $E(V|D=1) - E(V) \neq 0$. In our data the t-statistic for the null hypothesis $E(V|D=1) = E(V)$ is over 40 for each of the three levels of schooling considered, which strongly supports this assumption.

7.2 Instrumental Variable Based Estimates

We now report instrumental variable based estimates, specifically, Table 2 summarizes estimates of $r(x)$ based on Corollary 3. We define U in equation (21) to equal one for individual's that both self report having a masters degree or a PhD and are in the top decile of the standardized test scores. We are therefore assuming that Assumption A3 holds for x_0 equal to any X that includes these attributes of a self reported advanced degree and a high test score.

In our data the mean of U is 0.03468, so about 3.5% of our sample have both very high test scores and self report an advance degree. We could have based U on transcript reports of a graduate degree instead, but then by construction we would have $\widehat{\mu}_{V|U} = 1$. In our data, $\widehat{\mu}_{V|U}$ is .971 for a Bachelor's degree, .981 for

an Associate degree, and 1.000 for some college. Nonparametric estimates of $\widehat{r}(x) = \widehat{R}(1,x) - \widehat{R}(0,x)$ are obtained with $\widehat{R}(d,x)$ given by equation (26) with these estimates of $\widehat{\mu}_{V|U}$, and where the variable C in X is the test score, while Z is the vector of other elements of X. The first row of Table 2 contains r, the sample average of $\widehat{r}(X)$, while the second row has the estimated trimmed mean r_t, which is the sample average of $\widehat{r}(X)$ after removing the highest 5% and lowest 5% of $\widehat{r}(X)$ in the sample. Next are the lower quartile, middle quartile (median) and upper quartile r_{q1}, r_{med}, and r_{q3}, of $\widehat{r}(X)$ in the sample. The final row, "r semi, linear" is a semiparametric estimate of r using equation (30). Standard errors, reported in parentheses, are based on 400 bootstrap replications. One set of sufficient regularity conditions for bootstrapping here is Theorem B in Chen, Linton, and Van Keilegom (2003).

Table 2: Effects of Lying, Nonparametric and Semiparametric Corollary 2 IV Estimates

	Some college	Associate degree	Bachelor's degree
r nonparametric	0.07052 (0.03420)	0.1696 (0.3335)	0.1250 (1.918)
r_t nonparametric	0.07355 (0.03166)	0.1796 (0.04158)	0.07109 (0.1217)
r_{q1} nonparametric	-0.05768 (0.04930)	0.09099 (0.06185)	-0.1654 (0.1841)
r_{med} nonparametric	0.06447 (0.03663)	0.1287 (0.04903)	0.06696 (0.1003)
r_{q3} nonparametric	0.1421 (0.03903)	0.3214 (0.05156)	0.3002 (0.1596)
r semi, linear	0.08008 (0.02940)	0.1610 (0.03362)	0.05613 (1.138)

For the nonparametric estimates, the kernel function K is a standard normal density function, with bandwidth $b = 0.1836$ given by Silverman's rule. Doubling or halving this bandwidth changed most estimates by less than 10%, indicating that the results were generally not sensitive to bandwidth choice. An exception is that mean and trimmed mean estimates for the Bachelor's degree, which are small in Table 2, become larger (closer to the median r estimate) when the bandwidth is doubled. The results for the bachelor's degree are also much less precisely estimated than for some college or associate degree, with generally more than twice as large standard errors. Based on Table 1, we might expect that far fewer individuals misreport having a bachelor's degree, so the resulting imprecision in the Bachelor's degree estimates could be due to a much smaller fraction of data points that are informative about misreporting.

The nonparametric mean and median estimates of r are significant in Table 2, except for the Bachelor's degree. Overall, these results indicate that those who misreport by claiming to have have some college have about 6% to 8% higher wages than those who tell the truth about not having any college on average,

and those who misreport by claiming to have an associate degree have about 13% to 18% higher wages. The point estimates for lying about having a Bachelor's degree are lower, but they also have much larger standard errors. The variability in these estimated effects is quite large, ranging from a zero or negative effect at the first quartile to effects of 14% for some college to 32% for a degree at the third quartile. The semiparametric estimates of r are similar to the mean of the nonparametric estimates, though the variation in the quantiles of the nonparametric estimates suggests that the semiparametric specification, which assumes r is constant, is not likely to hold.

If transcripts V are very accurate, then V should be close to D^*, and the estimates of r in Table 1 should be close to those in Table 2. The linear model estimates in Table 1 are close to the semiparametric linear model estimates in Table 2 (for some college and associate degrees), however, the nonparametric estimates of r in Table 1 are much larger than the mean and median nonparametric estimates in Table 2. In linear models measurement error generally causes attenuation bias, but in contrast here the potentially mismeasured data estimates appear too large rather than too small. This could be due to nonlinearity, or because the potentially mismeasured variable V is highly correlated with another regressor, D.

We should expect that the effects of lying would be smaller than the returns to actually having some college or a degree. These effects of actual schooling are not identified from the assumptions in Corollary 2 or 3. Table 1 gives estimates of the effects of schooling s ranging from 28% for some college to 32% for a bachelor's degree, though these estimates are only reliable if transcripts V are accurate. These are indeed higher than the effects of lying, as one would expect. Also, while we would expect the effects of schooling to increase monotonically with the level of schooling, we do not necessarily expect the effects of lying to increase in the same way, because those effects depend on other factors like the plausibility of the misreport.

Kane, Rouse, and Staiger (1999) report some substantial error rates in transcripts, however, those findings are based on model estimates that could be faulty, rather than any type of direct verification. It is possible that transcripts are generally accurate, and in that case the ability of our estimator to produce reasonable estimates of r would not be impressive, since one could then just as easily generate good estimates of r using regressions or cell means as in Table 1. Therefore, to check the robustness of our methodology, we reestimated the model after randomly changing 20% of the observations of V to $1 - V$, thereby artificially making V a much weaker instrument. The resulting estimates of the mean and trimmed mean of r were generally higher than those reported in Tables 1 and 2 (consistent with our earlier result that, in our application, measurement error in V seems to raise rather than lower estimates of the effects of lying). As

with the other estimates, the numbers for bachelor's degrees are unstable with very large standard errors. However, the estimates of the median of r with this noisy V data are very close to the median estimates in table 2 (though of course with larger standard errors) for some college and associate degree. Specifically, the r_{med} estimates with substantial measurement error added to V were 0.070, 0.133, and 0.190, compared to the r_{med} estimates in Table 2 of 0.064, 0.129, and 0.067.

Table 3: Nonparametric Corollary 3 IV Effects of Lying Linearized Coefficient Estimates

X	Some college	Associate degree	Bachelor's degree
blacknh	-0.09208 (0.1246)	-0.2429 (2.674)	-0.3735 (2.288)
hispanic	0.01220 (0.1289)	-0.1529 (1.627)	-0.1541 (1.588)
othernh	0.2176 (0.1304)	0.1444 (1.265)	0.5398 (4.763)
female	0.09291 (0.06570)	0.2306 (0.5377)	0.2876 (3.370)
mscore	-0.009755 (0.03807)	0.03345 (0.3496)	-0.09489 (2.471)
constant	0.02449 (0.04635)	0.07127 (0.2840)	0.01803 (2.900)

To summarize how $\widehat{r}(x)$ varies with regressors x, Table 3 reports the estimated coefficients from linearly regressing the nonparametric estimates $\widehat{r}(x)$ on x and on a constant. The results show a few interesting patterns, including that women appear to have a larger effect of (possibly indicating higher returns to) lying than men, and that individuals with above average high school test scores also have above average effects of misreporting a higher degree of education. These results are consistent with the notion that the effects of lying should be highest for those who can lie most plausibly (e.g., those with high ability) or for those who may be perceived as less likely to lie (such as women). However, these results should not be over interpreted, since they are mostly not statistically significant.

8 Alternative Estimates Without IV

To check the robustness of our results to alternative identifying assumptions, in Table 4 we report the effects of lying using the estimator based on Theorem 2, which does not use data on the instrument V. These estimates are based only on self reports, and so do not use the transcript data in any way. For these estimates we assume equation (18) holds for x_0 equal to any value of X, which implies the assumption that

that no one understates their education level by reporting $D = 0$ when $D^* = 1$ (and hence that transcripts are wrong for the few observations in the data that have $D = 0$ and $V = 1$).

Table 4: Effects of Lying, Nonparametric and Semiparametric Theorem 2 Estimates Without IV

	Some college	Associate degree	Bachelor's degree
r nonparametric	-0.4127 (28.66)	0.1917 (2.915)	0.1247 (18.27)
r_t nonparametric	0.05064 (0.1402)	0.1684 (0.1738)	0.09186 (0.2489)
r_{q1} nonparametric	-0.05096 (0.1446)	-0.1065 (0.2406)	-0.5425 (0.3659)
r_{med} nonparametric	0.1179 (0.06115)	0.1495 (0.06191)	0.1958 (0.05549)
r_{q3} nonparametric	0.2570 (0.1019)	0.2813 (0.1428)	0.3308 (0.2038)

As should be expected, the estimates in Table 4 are mostly less precise than those in Table 2, in part because they do not exploit any transcript information, and they assume no heteroskedasticity in the model error η, which may not hold in this application. They are also more variable in part because they depend on higher moments of the data, and so will be more sensitive to outliers in the first stage nonparametric estimates. Still, the estimates in Table 4 are generally consistent with those in Table 2, and in particular almost all of the differences between Tables 2 and 4 are not statistically significant. Given the substantial differences in estimators and identifying assumptions between Corollary 3 and Theorem 2, it is reassuring that the resulting estimates are robust across the two methodologies.

In the Appendix we report the estimates of $E\left[R(d, X)\right]$ corresponding to Tables 2 and 4. As one would expect, these are generally more stable than the estimates of $E\left[r(X)\right]$ reported in Tables 2 and 4, since $r(X)$ is a difference $R(1, X) - R(0, X)$ rather than a level $R(d, X)$.

9 Conclusions

We provide identification and associated estimators for the conditional mean of an outcome Y, conditioned upon an observed discrete variable D and an unobserved discrete variable D^*. In particular, we identify the effects of lying, that is, the average difference in the mean level of Y between individuals having the unobserved $D^* = 0$ and those having $D^* = 1$ when the observed $D = 0$. Given an unconfoundedness assumption this difference in conditional means equals either the returns to lying (if misreports of D are intentional) or a placebo effect.

In our empirical application, Y is log wages, while D and D^* are self reports and actual levels of educational attainment. We find that wages are on average about 6% to 12% higher for those who lie about having some college, and from 8% to 20% higher on average for those who lie about having a college degree, relative to those who tell the truth about not having college or a diploma. Median and trimmed mean estimates appear to be more reliable and robust than estimates of raw mean returns and returns at other quantiles. Our results are about the same based on either semiparametric or nonparametric estimation, and are roughly comparable whether identification and associated estimation is based on using transcript reports as an instrument, or is based on higher moment error independence assumptions without exploiting transcript data. Our results are also robust to artificially adding a great deal of noise to the instrument.

The plausibility of our particular identifying assumptions may be debated, but we believe much of the value of this paper is in demonstrating that these effects of misreporting can be identified at all, and we expect future research will yield alternative assumptions that may be better suited to this and other applications. It would be particularly useful in the future to investigate how these results may be extended to handle confounding correlations with the unobserved treatment D^*, to obtain returns to lying without unconfoundedness assumptions.

In this application D and D^* refer to the same binary event (educational attainment), with D a self report of D^*. However, our theorems do require having D and D^* refer to the same binary event. More generally, one could estimate the average effect of any binary treatment or choice D (e.g., exposure to a law, a tax, or an advertisement) on any outcome Y (e.g., compliance with a law, income, expenditures on a product) where the effect is averaged only over some subpopulation of interest indexed by D^* (e.g., potential criminals, the poor, or a target audience of potential buyers), and where we do not observe exactly who is in the subpopulation of interest. Our identification strategy may thereby be relevant to a wide variety of applications, not just effects of lying.

10 Appendix

Proof of Lemmas 1 and 2: Consider Lemma 2 first:

$$\begin{aligned} Cov\,(D, V \mid X) &= E\,(DV \mid X) - E\,(D \mid X)\,E\,(V \mid X) \\ &= E\,[DE\,(V \mid D, X) \mid X] - E\,(D \mid X)\,E\,(V \mid X) \\ &= \Pr\,(D = 1 \mid X)\,E\,(V \mid D = 1, X) - E\,(D \mid X)\,E\,(V \mid X) \\ &= E\,(D \mid X)\,[E\,(V \mid D = 1, X) - E\,(V \mid X)] \end{aligned}$$

so $Cov\,(D, V \mid X) \neq 0$ if and only if the right side of the above expression is nonzero. The proof of Lemma 1 works exactly the same way.

Proof of Theorem 1:
First observe that

$$\begin{aligned} E\left(D^* V \mid D, X\right) &= \sum_{d^*=0}^{1} \Pr\left(D^* = d^* \mid D, X\right) E\left(D^* V \mid D^* = d^*, D, X\right) \\ &= \Pr\left(D^* = 1 \mid D, X\right) E\left(V \mid D^* = 1, D, X\right) \\ &= E\left(D^* \mid D, X\right) E\left(V \mid D^* = 1, X\right) \end{aligned}$$

and using this result we have

$$\begin{aligned} E\,(YV \mid D, X) &= R(D, X)E\,(V \mid D, X) + s(D, X)E\left[D^* V \mid D, X\right] + E\,(\eta V \mid D, X) \\ &= R(D, X)E\,(V \mid D, X) + s(D, X)E\left(D^* \mid D, X\right) E\left(V \mid D^* = 1, X\right). \end{aligned}$$

Also

$$E\,(Y \mid D, X) = R(D, X) + s(D, X)E\left[D^* | D, X\right]$$

Use the latter equation to substitute $s(D, X)E\left[D^* | D, X\right]$ out of the former equation, and solve what remains for $R(D, X)$ to obtain equation (9). Equation (10) then follows immediately from equation (9) using $r(X) = R(1, X) - R(0, X)$ and the properties of a covariance.

Proof of Theorem 2: Begin with equation (2), $Y = R(D, X) + s(D, X)D^* + \eta$ with $R(D, X) =$

$R(X) + r(X)D$. Assumption B2 implies that

$$\begin{aligned} \mu_{Y|D,X} &\equiv E\left(Y|D,X\right) \\ &= E\left(\left(R(D,X)+s(D,X)D^*\right)|D,X\right) \\ &= R(D,X)+s(D,X)E\left(D^*|D,X\right), \end{aligned} \tag{32}$$

$$\begin{aligned} \mu_{Y^2|D,X} &\equiv E\left(Y^2|D,X\right) \\ &= E\left(\left(R(D,X)+s(D,X)D^*+\eta\right)^2|D,X\right) \\ &= E\left(\left(R(D,X)+s(D,X)D^*\right)^2|D,X\right)+E\eta^2 \\ &= R(D,X)^2+2R(D,X)s(D,X)E\left(D^*|D,X\right)+s(D,X)^2E\left(D^*|D,X\right)+E\eta^2 \\ &= R(D,X)^2+2R(D,X)\left(\mu_{Y|D,X}-R(D,X)\right)+s(D,X)\left(\mu_{Y|D,X}-R(D,X)\right)+E\eta^2 \\ &= \mu_{Y|D,X}R(D,X)+\left(R(D,X)+s(D,X)\right)\left(\mu_{Y|D,X}-R(D,X)\right)+E\eta^2, \end{aligned} \tag{33}$$

and

$$\begin{aligned} \mu_{Y^3|D,X} &\equiv E\left(Y^3|D,X\right) \\ &= E\left(\left(R(D,X)+s(D,X)D^*+\eta\right)^3|D,X\right) \\ &= E\left[\left(R(D,X)+s(D,X)D^*\right)^3|D,X\right]+3E\left[\left(R(D,X)+s(D,X)D^*\right)|D,X\right]E\eta^2+E\eta^3 \\ &= R(D,X)^3+3R(D,X)^2s(D,X)E\left(D^*|D,X\right) \\ &\quad +3R(D,X)s(D,X)^2E\left(D^*|D,X\right)+s(D,X)^3E\left(D^*|D,X\right) \\ &\quad +3\mu_{Y|D,X}E\eta^2+E\eta^3. \end{aligned} \tag{34}$$

We now show that assumption B2 implies the identification of $E\left(\eta^k\right)$ for $k=2,3$. This assumption implies that

$$\begin{aligned} & E\left(D^*|D=0,X=x_0\right) \\ =\ & \Pr\left(D^*=1|D=0,X=x_0\right) \\ =\ & \Pr\left(D=0|D^*=1,X=x_0\right)\frac{\Pr\left(D^*=1|X=x_0\right)}{\Pr\left(D=0|X=x_0\right)} \\ =\ & 0, \end{aligned}$$

and therefore,

$$\begin{aligned}\mu_{Y|0,x_0} &\equiv E\left(Y|D=0,X=x_0\right)\\ &= R(0,x_0)+s(0,x_0)E\left(D^*|D=0,X=x_0\right)\\ &= R(0,x_0),\end{aligned}$$

$$\begin{aligned}\mu_{Y^2|0,x_0} &\equiv E\left(Y^2|D=0,X=x_0\right)\\ &= R(0,x_0)^2+2R(D,X)s(D,X)E\left(D^*|D=0,X=x_0\right)\\ &\quad +s(D,X)^2E\left(D^*|D=0,X=x_0\right)+E\eta^2\\ &= R(0,x_0)^2+E\eta^2\\ &= \mu_{Y|0,x_0}^2+E\eta^2,\end{aligned}$$

and

$$\begin{aligned}\mu_{Y^3|0,x_0} &= E\left(Y^3|D=0,X=x_0\right)\\ &= R(0,x_0)^3+3\mu_{Y|0,x_0}E\eta^2+E\eta^3\\ &= \mu_{Y|0,x_0}^3+3\mu_{Y|0,x_0}\left(\mu_{Y^2|0,x_0}-\mu_{Y|0,x_0}^2\right)+E\eta^3.\end{aligned}$$

Therefore, we have

$$\begin{aligned}E\eta^2 &= \mu_{Y^2|0,x_0}-\mu_{Y|0,x_0}^2\\ &\equiv \sigma_{Y|0,x_0}^2,\end{aligned}$$

and

$$\begin{aligned}E\eta^3 &= \mu_{Y^3|0,x_0}+2\mu_{Y|0,x_0}^3-3\mu_{Y|0,x_0}\mu_{Y^2|0,x_0}\\ &= E\left(\left(Y-\mu_{Y|0,x_0}\right)^3|D=0,X=x_0\right)\\ &\equiv \upsilon_{Y|0,x_0}^3.\end{aligned}$$

In the next step, we eliminate $s(D,X)$ and $E\left(D^*|D,X\right)$ in equations 32-34 to obtain a restriction only containing $R(D,X)$ and known variables. We will use the following two equations repeatedly.

$$\left(R(D,X)+s(D,X)\right)\left(\mu_{Y|D,X}-R(D,X)\right)=\mu_{Y^2|D,X}-E\eta^2-\mu_{Y|D,X}R(D,X) \tag{35}$$

$$s(D,X)E\left(D^*|D,X\right) = \mu_{Y|D,X} - R(D,X) \tag{36}$$

Notice that

$$s(D,X) = \frac{\mu_{Y^2|D,X} - \mu^2_{Y|D,X} - \sigma^2_{Y|0,x_0}}{\mu_{Y|D,X} - R(D,X)} + \mu_{Y|D,X} - R(D,X)$$

which also implies that we can't identify $s(0,x_0)$ because $\mu_{Y|D=0,x_0} = R(0,x_0)$.

From here on we will for clarity drop the term (D,X) when it is obvious from context. Consider

$$\begin{aligned}
\mu_{Y^3|D,X} &\equiv E\left(Y^3|D,X\right) \\
&= E\left(\left(R(D,X)+s(D,X)D^*+\eta\right)^3|D,X\right) \\
&= E\left(\left(R+sD^*\right)^3|D,X\right) + 3E\left(\left(R+sD^*\right)|D,X\right)E\eta^2 + E\left(\eta^3\right) \\
&= R(D,X)^3 + 3R(D,X)^2 s(D,X)E\left(D^*|D,X\right) \\
&\quad +3R(D,X)s(D,X)^2E\left(D^*|D,X\right) + s(D,X)^3E\left(D^*|D,X\right) \\
&\quad +3\left[R(D,X)+s(D,X)E\left(D^*|D,X\right)\right]E\eta^2 + E\eta^3 \\
&= R^3 + 3R^2\left(\mu_{Y|D,X}-R\right) + 3Rs\left(\mu_{Y|D,X}-R\right) + s^2\left(\mu_{Y|D,X}-R\right) + 3\mu_{Y|D,X}E\eta^2 + E\eta^3 \\
&= R^3 + 3R^2\left(\mu_{Y|D,X}-R\right) + 2Rs\left(\mu_{Y|D,X}-R\right) + s\left(R+s\right)\left(\mu_{Y|D,X}-R\right) + 3\mu_{Y|D,X}E\eta^2 + E\eta^3 \\
&= R^3 + 3R^2\left(\mu_{Y|D,X}-R\right) + 2Rs\left(\mu_{Y|D,X}-R\right) + s\left(\mu_{Y^2|D,X} - E\eta^2 - \mu_{Y|D,X}R\right) \\
&\quad +3\mu_{Y|D,X}E\eta^2 + E\eta^3
\end{aligned}$$

Which, with a little algebra can be written as

$$\begin{aligned}
\mu_{Y^3|D,X} &= R\left(\mu_{Y^2|D,X} - E\eta^2\right) + (R+s)\left(\mu_{Y^2|D,X} - E\eta^2 - \mu_{Y|D,X}R\right) + 3\mu_{Y|D,X}E\eta^2 + E\eta^3 \\
&= R\left(\mu_{Y^2|D,X} - E\eta^2\right) + \frac{\mu_{Y^2|D,X} - E\eta^2 - \mu_{Y|D,X}R}{\left(\mu_{Y|D,X}-R\right)}\left(\mu_{Y^2|D,X} - E\eta^2 - \mu_{Y|D,X}R\right) \\
&\quad +3\mu_{Y|D,X}E\eta^2 + E\eta^3.
\end{aligned}$$

That is

$$\begin{aligned}
0 &= \left(\mu_{Y^2|D,X} - E\eta^2 - \mu_{Y|D,X}R\right)^2 + \left(\mu_{Y^2|D,X} - E\eta^2\right)\left(\mu_{Y|D,X}-R\right)R \\
&\quad -\left(\mu_{Y^3|D,X} - \left(3\mu_{Y|D,X}E\eta^2 + E\eta^3\right)\right)\left(\mu_{Y|D,X}-R\right).
\end{aligned}$$

The restrictions on R simplify to the quadratic equation

$$-\alpha R^2 + \beta R + \gamma = 0,$$

where

$$\begin{aligned}
\alpha &= -\left(\mu_{Y|D,X}^2 - \left(\mu_{Y^2|D,X} - E\eta^2\right)\right), \\
\beta &= \left(-\left(\mu_{Y^2|D,X} - E\eta^2\right)\mu_{Y|D,X} + \mu_{Y^3|D,X} - \left(3\mu_{Y|D,X}E\eta^2 + E\eta^3\right)\right), \\
\gamma &= \left(\mu_{Y^2|D,X} - E\eta^2\right)^2 - \left(\mu_{Y^3|D,X} - \left(3\mu_{Y|D,X}E\eta^2 + E\eta^3\right)\right)\mu_{Y|D,X}.
\end{aligned}$$

Notice that

$$\sigma_{Y|D,X}^2 = \mu_{Y^2|D,X} - \mu_{Y|D,X}^2,$$

$$\begin{aligned}
v_{Y|D,X}^3 &\equiv E\left(\left(Y - \mu_{Y|D,X}\right)^3 |D, X\right) \\
&= \mu_{Y^3|D,X} + 2\mu_{Y|D,X}^3 - 3\mu_{Y|D,X}\mu_{Y^2|D,X}.
\end{aligned}$$

We then simplify the expressions of α, β, and γ as follows:

$$\begin{aligned}
\alpha &= -\left(\mu_{Y|D,X}^2 - \left(\mu_{Y^2|D,X} - E\eta^2\right)\right) \\
&= \left(\sigma_{Y|D,X}^2 - \sigma_{Y|0,x_0}^2\right),
\end{aligned}$$

$$\begin{aligned}
\beta &= \left(-\left(\mu_{Y^2|D,X} - E\eta^2\right)\mu_{Y|D,X} + \mu_{Y^3|D,X} - \left(3\mu_{Y|D,X}E\eta^2 + E\eta^3\right)\right) \\
&= \left(\mu_{Y^3|D,X} - 2\mu_{Y|D,X}E\eta^2 - E\eta^3 - \mu_{Y|D,X}\mu_{Y^2|D,X}\right) \\
&= v_{Y|D,X}^3 - 2\mu_{Y|D,X}^3 + 3\mu_{Y|D,X}\mu_{Y^2|D,X} - 2\mu_{Y|D,X}E\eta^2 - E\eta^3 - \mu_{Y|D,X}\mu_{Y^2|D,X} \\
&= v_{Y|D,X}^3 - E\eta^3 - 2\mu_{Y|D,X}^3 - 2\mu_{Y|D,X}E\eta^2 + 2\mu_{Y|D,X}\mu_{Y^2|D,X} \\
&= v_{Y|D,X}^3 - E\eta^3 - 2\mu_{Y|D,X}^3 - 2\mu_{Y|D,X}E\eta^2 + 2\mu_{Y|D,X}\left(\sigma_{Y|D,X}^2 + \mu_{Y|D,X}^2\right) \\
&= v_{Y|D,X}^3 - E\eta^3 + 2\mu_{Y|D,X}\left(\sigma_{Y|D,X}^2 - E\eta^2\right) \\
&= v_{Y|D,X}^3 - v_{Y|0,x_0}^3 + 2\mu_{Y|D,X}\left(\sigma_{Y|D,X}^2 - \sigma_{Y|0,x_0}^2\right) \\
&= v_{Y|D,X}^3 - v_{Y|0,x_0}^3 + 2\mu_{Y|D,X}\alpha,
\end{aligned}$$

$$
\begin{aligned}
\gamma &= \left(\mu_{Y^2|D,X} - E\eta^2\right)^2 - \left(\mu_{Y^3|D,X} - \left(3\mu_{Y|D,X}E\eta^2 + E\eta^3\right)\right)\mu_{Y|D,X} \\
&= \left(\sigma^2_{Y|D,X} + \mu^2_{Y|D,X} - E\eta^2\right)^2 - \left(\mu_{Y^3|D,X} - \left(3\mu_{Y|D,X}E\eta^2 + E\eta^3\right)\right)\mu_{Y|D,X} \\
&= \mu^4_{Y|D,X} + 2\mu^2_{Y|D,X}\left(\sigma^2_{Y|D,X} - E\eta^2\right) + \left(\sigma^2_{Y|D,X} - E\eta^2\right)^2 \\
&\quad -\mu_{Y^3|D,X}\mu_{Y|D,X} + 3\mu^2_{Y|D,X}E\eta^2 + \mu_{Y|D,X}E\eta^3 \\
&= \mu^4_{Y|D,X} + 2\mu^2_{Y|D,X}\sigma^2_{Y|D,X} + \left(\sigma^2_{Y|D,X} - E\eta^2\right)^2 - \mu_{Y^3|D,X}\mu_{Y|D,X} + \mu^2_{Y|D,X}E\eta^2 + \mu_{Y|D,X}E\eta^3 \\
&= \mu^4_{Y|D,X} + 2\mu^2_{Y|D,X}\sigma^2_{Y|D,X} + \left(\sigma^2_{Y|D,X} - E\eta^2\right)^2 \\
&\quad - \left(v^3_{Y|D,X} - 2\mu^3_{Y|D,X} + 3\mu_{Y|D,X}\mu_{Y^2|D,X}\right)\mu_{Y|D,X} + \mu^2_{Y|D,X}E\eta^2 + \mu_{Y|D,X}E\eta^3 \\
&= \mu^4_{Y|D,X} + 2\mu^2_{Y|D,X}\sigma^2_{Y|D,X} + \left(\sigma^2_{Y|D,X} - E\eta^2\right)^2 \\
&\quad +2\mu^4_{Y|D,X} - 3\mu^2_{Y|D,X}\mu_{Y^2|D,X} + \mu^2_{Y|D,X}E\eta^2 + \mu_{Y|D,X}\left(E\eta^3 - v^3_{Y|D,X}\right) \\
&= \mu^4_{Y|D,X} + 2\mu^2_{Y|D,X}\sigma^2_{Y|D,X} + \left(\sigma^2_{Y|D,X} - E\eta^2\right)^2 \\
&\quad +2\mu^4_{Y|D,X} - 3\mu^2_{Y|D,X}\left(\sigma^2_{Y|D,X} + \mu^2_{Y|D,X}\right) + \mu^2_{Y|D,X}E\eta^2 + \mu_{Y|D,X}\left(E\eta^3 - v^3_{Y|D,X}\right) \\
&= \left(\sigma^2_{Y|D,X} - E\eta^2\right)^2 - \mu^2_{Y|D,X}\left(\sigma^2_{Y|D,X} - E\eta^2\right) - \mu_{Y|D,X}\left(v^3_{Y|D,X} - E\eta^3\right) \\
&= \left(\sigma^2_{Y|D,X} - \sigma^2_{Y|0,x_0}\right)^2 - \mu^2_{Y|D,X}\left(\sigma^2_{Y|D,X} - \sigma^2_{Y|0,x_0}\right) - \mu_{Y|D,X}\left(v^3_{Y|D,X} - v^3_{Y|0,x_0}\right) \\
&= \alpha^2 - \mu^2_{Y|D,X}\alpha - \mu_{Y|D,X}\left(\beta - 2\mu_{Y|D,X}\alpha\right) \\
&= \alpha^2 + \mu^2_{Y|D,X}\alpha - \mu_{Y|D,X}\beta.
\end{aligned}
$$

In summary, we have

$$-\alpha R^2 + \beta R + \gamma = 0$$

$$
\begin{aligned}
\alpha &= \sigma^2_{Y|D,X} - \sigma^2_{Y|0,x_0} \\
\beta &= v^3_{Y|D,X} - v^3_{Y|0,x_0} + 2\mu_{Y|D,X}\alpha \\
\gamma &= \alpha^2 + \mu^2_{Y|D,X}\alpha - \mu_{Y|D,X}\beta
\end{aligned}
$$

That means

$$R = \frac{\beta + \sqrt{\beta^2 + 4\alpha\gamma}}{2\alpha} \text{ or } \frac{\beta - \sqrt{\beta^2 + 4\alpha\gamma}}{2\alpha}.$$

In fact, we may show that equations 36 and 35 implies

$$\alpha \geq 0$$

Consider

$$\begin{aligned} s &= \frac{\mu_{Y^2|D,X} - \mu^2_{Y|D,X} - E\eta^2}{\mu_{Y|D,X} - R} + \mu_{Y|D,X} - R \\ &= \frac{\alpha}{\mu_{Y|D,X} - R} + \mu_{Y|D,X} - R \end{aligned}$$

and

$$\begin{aligned} E\left(D^*|D,X\right) &= \frac{\mu_{Y|D,X} - R}{s} \\ &= \frac{\left(\mu_{Y|D,X} - R\right)^2}{\left(\mu_{Y|D,X} - R\right)^2 + \alpha}. \end{aligned}$$

Therefore, $0 \leq E\left(D^*|D,X\right) \leq 1$ implies that $\alpha \geq 0$.

The last step is to eliminate one of the two roots to achieve point identification. Notice that

$$E\left(Y|D^*,D,X\right) = R(D,X) + s(D,X)D^*.$$

Assumption B2 implies that

$$s(D,X) \geq 0.$$

Consider

$$\begin{aligned} \mu_{Y|D,X} &= R + sE\left(D^*|D,X\right) \\ &= R\left[1 - E\left(D^*|D,X\right)\right] + (R+s)\,E\left(D^*|D,X\right). \end{aligned}$$

Therefore, $0 \leq E\left(D^*|D,X\right) \leq 1$ and $s(D,X) \geq 0$ imply

$$R \leq \mu_{Y|D,X} \leq s + R,$$

Thus, we may identify R as the smaller root if $\mu_{Y|D,X}$ is between the two roots. , i.e.,

$$-\alpha\mu^2_{Y|D,X} + \beta\mu_{Y|D,X} + \gamma \geq 0,$$

which holds because

$$\begin{aligned} & -\alpha\mu_{Y|D,X}^2 + \beta\mu_{Y|D,X} + \gamma \\ = \; & -\alpha\mu_{Y|D,X}^2 + \beta\mu_{Y|D,X} + \alpha^2 + \mu_{Y|D,X}^2\alpha - \mu_{Y|D,X}\beta \\ = \; & \alpha^2 \geq 0. \end{aligned}$$

Therefore, we have

$$R(D, X) = \frac{\beta - \sqrt{\beta^2 + 4\alpha\gamma}}{2\alpha}.$$

Notice that R equals the larger root if $s(D, X) \leq 0$. The function $s(D, X)$ then follows.

Discrete Limiting Distributions for equation (15). Let

$$\begin{aligned} \widehat{\alpha}(x) &= \left(\widehat{\mu}_{Y,V,X,1}, \widehat{\mu}_{Y,V,X,0}, \widehat{\mu}_{Y,X,1}, \widehat{\mu}_{Y,X,0}, \widehat{\mu}_{V,X,1}, \widehat{\mu}_{V,X,0}, \widehat{\mu}_{X,1}, \widehat{\mu}_{X,0}, \widehat{\mu}_{VU}, \widehat{\mu}_{U}\right)^T, \\ \alpha_0 &= E\left[\widehat{\alpha}(x)\right], \\ \widehat{R}(d, \widehat{\alpha}(x)) &\equiv \frac{\left(\widehat{\mu}_{Y,V,X,1}^d \widehat{\mu}_{Y,V,X,0}^{1-d}\right)\widehat{\mu}_U - \left(\widehat{\mu}_{Y,X,1}^d \widehat{\mu}_{Y,X,0}^{1-d}\right)\widehat{\mu}_{VU}}{\left(\widehat{\mu}_{V,X,1}^d \widehat{\mu}_{V,X,0}^{1-d}\right)\widehat{\mu}_U - \left(\widehat{\mu}_{X,1}^d \widehat{\mu}_{X,0}^{1-d}\right)\widehat{\mu}_{VU}}, \\ \widehat{r}(x) &= \widehat{R}(1, \widehat{\alpha}(x)) - \widehat{R}(0, \widehat{\alpha}(x)), \end{aligned}$$

$$\begin{aligned} \gamma &= \frac{\partial}{\partial t} R\left(d, \alpha_0 + t\left(\widehat{\alpha} - \alpha_0\right)\right)\Big|_{t=0} \\ &\equiv G\left(d, \alpha_0\right)^T \left(\widehat{\alpha} - \alpha_0\right), \end{aligned}$$

$$V\left(\widehat{\alpha}(x)\right) = n \times E\left[\left(\widehat{\alpha} - \alpha_0\right)\left(\widehat{\alpha} - \alpha_0\right)^T\right].$$

Assuming independent, identically distributed draws and existence of $V\left(\widehat{\alpha}(x)\right)$, by the Lindeberg-Levy central limit theorem and the delta method

$$\begin{aligned} \sqrt{n}\left[\widehat{R}(d, x) - R(d, x)\right] &\rightarrow {}^d N\left(0, \Omega_R\right) \\ \Omega_R &= G\left(d, \alpha_0(x)\right)^T V\left(\widehat{\alpha}(x)\right) G\left(d, \alpha_0(x)\right) \end{aligned}$$

and

$$\begin{aligned} \sqrt{n}\left[\widehat{r}(x) - r(x)\right] &\rightarrow {}^d N\left(0, \Omega_r\right) \\ \Omega_r &= \left[G\left(1, \alpha_0(x)\right) - G\left(0, \alpha_0(x)\right)\right]^T V\left(\widehat{\alpha}(x)\right)\left[G\left(1, \alpha_0(x)\right) - G\left(0, \alpha_0(x)\right)\right]. \end{aligned}$$

Table 5: R(0,X), Nonparametric and Semiparametric Corollary 3 IV Estimates

	Some college	Associate degree	Bachelor's degree
R0 nonparametric	2.072 (0.01514)	2.125 (0.009665)	2.143 (0.007997)
$R0_t$ nonparametric	2.065 (0.01536)	2.125 (0.01016)	2.144 (0.008579)
$R0_{q1}$ nonparametric	1.863 (0.02520)	1.939 (0.01940)	1.975 (0.01834)
$R0_{med}$ nonparametric	2.003 (0.03859)	2.089 (0.03225)	2.143 (0.02788)
$R0_{q3}$ nonparametric	2.309 (0.02681)	2.326 (0.01763)	2.319 (0.01665)
R0 semi, linear	2.025 (0.01174)	2.094 (0.008754)	2.114 (0.007451)

Table 6: R(1,X), Nonparametric and Semiparametric Corollary 3 IV Estimates

	Some college	Associate degree	Bachelor's degree
R1 nonparametric	2.142 (0.03011)	2.295 (0.3326)	2.268 (1.918)
$R1_t$ nonparametric	2.152 (0.02986)	2.319 (0.04103)	2.223 (0.1219)
$R1_{q1}$ nonparametric	1.997 (0.04430)	2.181 (0.06094)	2.092 (0.1694)
$R1_{med}$ nonparametric	2.173 (0.04633)	2.380 (0.04084)	2.189 (0.1026)
$R1_{q3}$ nonparametric	2.340 (0.04635)	2.449 (0.04508)	2.397 (0.1731)
R1 semi, linear	2.188 (0.02898)	2.341 (0.03397)	2.267 (1.149)

Table 7: R(0,X), Nonparametric and Semiparametric Theorem 2 Estimates Without IV

	Some college	Associate degree	Bachelor's degree
R0 nonparametric	2.078 (0.01383)	2.126 (0.009571)	2.144 (0.007897)
$R0_t$ nonparametric	2.074 (0.01447)	2.123 (0.01025)	2.148 (0.008459)
$R0_{q1}$ nonparametric	1.891 (0.02270)	1.942 (0.01916)	1.974 (0.01820)
$R0_{med}$ nonparametric	2.022 (0.03761)	2.095 (0.03227)	2.146 (0.02767)
$R0_{q3}$ nonparametric	2.288 (0.02247)	2.321 (0.01708)	2.324 (0.01620)

Table 8: R(1,X), Nonparametric and Semiparametric Theorem 2 Estimates Without IV

	Some college	Associate degree	Bachelor's degree
R1 nonparametric	1.666 (28.66)	2.318 (2.915)	2.269 (18.27)
$R1_t$ nonparametric	2.141 (0.1418)	2.310 (0.1719)	2.227 (0.2483)
$R1_{q1}$ nonparametric	1.832 (0.1459)	2.069 (0.2132)	1.525 (0.3052)
$R1_{med}$ nonparametric	2.223 (0.07273)	2.247 (0.08727)	2.222 (0.07330)
$R1_{q3}$ nonparametric	2.419 (0.09065)	2.501 (0.1239)	2.552 (0.2033)

References

[1] AI, C. AND X. CHEN (2003), "Efficient Estimation of Models With Conditional Moment Restrictions Containing Unknown Functions," *Econometrica*, 71, 1795-1844.

[2] AIGNER, D. J. (1973), "Regression With a Binary Independent Variable Subject to Errors of Obervation," *Journal of Econometrics*, 1, 249-60.

[3] ASHENFELTER, O. AND A. KRUEGER (1994), "Estimates of the Economic Return to Schooling from a New Sample of Twins," *The American Economic Review*, 84, 1157-1173

[4] BLUNDELL, R. AND J. L. POWELL, (2004), "Endogeneity in Semiparametric Binary Response Models" *Review of Economic Studies*, 71, 655-679.

[5] BOLLINGER, C. R. (1996), "Bounding Mean Regressions When a Binary Regressor is Mismeasured," *Journal of Econometrics*, 73, 387-399.

[6] CARD, D. (1996), "The Effect of Unions on the Structure of Wages: A Longitudinal Analysis," *Econometrica*, 64, 957-979.

[7] CHEN, X., Y. HU AND A. LEWBEL, (2008a), "Nonparametric Identification of Regression Models Containing a Misclassified Dichotomous Regressor Without Instruments," *Economics Letters*, 100, 381-384.

[8] CHEN, X., Y. HU AND A. LEWBEL, (2008b), "A Note on the Closed-form Identification of Regression Models with a Mismeasured Binary Regressor," Statistics and Probability Letters, 78, 1473-1479.

[9] CHEN, X., O. LINTON, AND I. VAN KEILEGOM, (2003) "Estimation of Semiparametric Models when the Criterion Function Is Not Smooth," *Econometrica*, 71, 1591-1608,

[10] DAS, M., (2004), "Instrumental Variables Estimators of Nonparametric Models With Discrete Endogenous Regressors," *Journal of Econometrics*, 124, 335-361.

[11] FLORENS, J.-P. AND L. MALAVOLTI, (2003), "Instrumental Regression with Discrete Variables," unpublished manuscript.

[12] GIBRAT, R. (1931), Les Inegalites Economiques, Librairie du Recueil Sirey, Paris

[13] HECKMAN, J. J. (1979), "Sample selection bias as a specification error," *Econometrica*, 47, 153–161.

[14] HECKMAN, J. J., H. ICHIMURA AND P. TODD, (1998), "Matching as an Econometric Evaluations Estimator, Review of Economic Studies, 65, 261-294.

[15] HOTZ, V. J., C. MULLIN, AND S. SANDERS, (1997), "Bounding Causal Effects Using Data from a Contaminated Natural Experiment: Analyzing the Effects of Teenage Childbearing," *Review of Economic Studies*, 64, 575-603.

[16] HU, Y. (2006), "Identification and estimation of nonlinear models with misclassification error using instrumental variables," U. Texas at Austin unpublished manuscript.

[17] KANE, T. J., AND C. E. ROUSE, (1995), "Labor market returns to two- and four- year college," *American Economic Review*, 85, 600-614

[18] KANE, T. J., C. E. ROUSE, AND D. STAIGER, (1999), "Estimating Returns to Schooling When Schooling is Misreported," NBER working paper #7235.

[19] KLEPPER, S., (1988), "Bounding the Effects of Measurement Error in Regressions Involving Dichotomous Variables," *Journal of Econometrics*, 37, 343-359.

[20] LEWBEL, A., (2007a), "Estimation of Average Treatment Effects With Misclassification," *Econometrica*, 75, 537-551forthcoming.

[21] LEWBEL, A., (2007b), "A Local Generalized Method of Moments Estimator," *Economics Letters*, 94, 124-128.

[22] MAHAJAN, A. (2006) "Identification and Estimation of Regression Models with Misclassification," *Econometrica*, 74, 631-665.

[23] MANSKI, C. F. (1990) "Nonparametric Bounds on Treatment Effects," *American Economic Review Papers and Proceedings*, 80, 319-323.

[24] NEWEY, W. K., (1994), "Kernel Estimation of Partial Means and a General Variance Estimator," *Econometric Theory*, 10, 233-253.

[25] NEWEY, W. K. AND J. L. POWELL, (2003), "Instrumental Variable Estimation of Nonparametric Models," *Econometrica*, 71, 1565-1578.

[26] MOLINARI, F. (2008), "Partial Identification of Probability Distributions with Misclassified Data," *Journal of Econometrics,* 144, 81-117.

[27] ROSENBAUM, P. AND D. RUBIN, (1984), "Reducing Bias in Observational Studies Using Subclassification on the Propensity Score," *Journal of the American Statistical Association*, 79, 516-524.

[28] RUBIN, D. B. (1974), "Estimating Causal Effects of Treatments in Randomized and Non-Randomized Studies, " *Journal of Educational Psychology*, 76, 688-701.