Estimation of Average Treatment Effects With Misclassification

Arthur Lewbel*
Boston College

October, 2003; revised June 2006

Abstract

This paper considers identification and estimation of the effect of a mismeasured binary regressor in a nonparametric or semiparametric regression, or the conditional average effect of a binary treatment or policy on some outcome where treatment may be misclassified. Failure to account for misclassification is shown to result in attenuation bias in the estimated treatment effect. An identifying assumption that overcomes this bias is existence of an instrument for the binary regressor that is conditionally independent of the treatment effect. A discrete instrument suffices for nonparametric identification.

JEL Codes: C14, C13. Keywords: Binary regressor, Program Evaluation, Treatment Effects, Misclassification, Contamination Bias, Measurement error, Binary Choice, Binomial Response.

The author wishes to thank Todd Prono for excellent research assistance, Alberto Abadie, Jinyong Hahn, Francesca Molinari, James Heckman, Whitney Newey, and anonymous referees for many helpful comments. Any errors are my own.

1 Introduction

This paper provides conditions for identification and estimation of the average effect, conditioned on covariates, of a binary treatment, program, or policy on a scalar outcome when treatment may be misclassified. More generally, what is provided is an estimator of the effect of a binary regressor in a conditional mean

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

regression (which may also include other regressors), when the binary regressor is observed with error. This equals the conditional average treatment effect given a weak unconfoundedness assumption.

Misclassification occurs when a binary variable (the treatment indicator) is measured with error, that is, some units are reported to have received treatment when they actually have not, and vice versa. For example, in a returns to schooling analysis the outcome could be wages, the binary variable could be attaining a college degree, and misclassification could arise from school transcript reporting errors. See, e.g., Kane and Rouse (1995) and Kane, Rouse, and Staiger (1999). Bollinger (1996) considers misclassification of union status in wage equations. If treatment is a job training program, misclassification may arise if individuals assigned to the program fail to attend, or if those not assigned obtain training elsewhere. Similarly, for medical treatment individuals assigned a drug might fail to take it, and those assigned a placebo might obtain treatment from another source. More generally, misclassification describes any binary variable that is sometimes mismeasured.

This paper first shows that misclassification causes attenuation bias in estimated treatment effects, analogous to the attenuation bias of classically mismeasured variables in linear regression models. Additional assumptions on existence of an instrument are then provided that identify misclassification probabilities and the true probability of treatment (conditional on covariates), in addition to identifying the conditional average treatment effect. Corresponding estimators are variants of Hansen's (1982) Generalized Method of Moments (GMM).

Let Y be an outcome, let X be a vector of covariates, and let T^* be an unobserved true treatment indicator, or more generally T^* is any unobserved binary variable that affects Y and can only equal zero or one. Define

$$h^*(X, T^*) = E(Y \mid X, T^*) \tag{1}$$

So $h^*(X, T^*)$ is the conditional mean outcome, given X and treatment T^*. Since T^* is binary we may without loss of generality rewrite the conditional mean outcome as

$$h^*(X, T^*) = h_0^*(X) + \tau^*(x)T^* \tag{2}$$

where $h_0^*(X) = h^*(X, 0)$ and

$$\tau^*(x) = h^*(x, 1) - h^*(x, 0) \tag{3}$$

Assume we observe draws of Y, X, and T but not of T^*. The binary variable T is a proxy for T^* that is subject to measurement error. Misclassification occurs when $T \neq T^*$. If T^* were observed instead of T, then $h^*(X, T^*)$ and hence $\tau^*(x)$ would be identified, and could be estimated by nonparametric regression.

Das (2004) identifies and estimates $\widetilde{h}(X, T) = \widetilde{h}_0(X) + \widetilde{\tau}(x)T$ in the model $Y = \widetilde{h}(X, T) + e$ where $E(e \mid X) = 0$. Estimation is then based on $E(Y \mid X) = \widetilde{h}_0(X) + \widetilde{\tau}(x)E(T \mid X)$. The Das model allows T to be correlated with e. Similar estimators of models with endogenous binary regressors include Newey and Powell (2003) and Florens and Malavolti (2003). The Das estimator cannot be applied in the present context, because here $E(Y \mid X) = h_0^*(X) + \tau^*(x)E(T^* \mid X)$ and, unlike the case with classically mismeasured variables, (see, e.g., Aigner 1973), with a mismeasure T of a binary regressor T^*, $E(T \mid X) \neq E(T^* \mid X)$.

Let $X = (Z, V)$. Mahajan (2005) obtains identification of $h^*(x, t)$ assuming that $E(Y \mid Z, V, T^*) = E(Y \mid Z, T^*)$ and that $E(T^* \mid Z, V) \neq E(T^* \mid Z)$, so V is essentially an instrument that correlates with the true probability of treatment but is conditionally mean independent of the outcome. Hu (2004) obtains identification of $h^*(x, t)$ under similar assumptions for multiple valued discrete treatments t, and weakens some of Mahajan's side conditions.

The present paper focuses on identifying and estimating $\tau^*(x)$, the effect on Y of changing T^*. I will refer to $\tau^*(x)$ as the conditional outcome treatment effect, or just the treatment effect, though formally it is just the change in the conditional mean of Y that results in changing the binary variable t^* from zero to one holding x fixed, and only equals a true treatment effect when an unconfoundedness condition holds, as discussed later.

Define the mismeasured (treatment) effect $\tau(x) = h(x, 1) - h(x, 0)$ where $h(X, T) = E(Y \mid X, T)$. Theorem 1 shows that under some general conditions $\tau(x) = m(x)\tau^*(x)$ where $0 \leq m(x) \leq 1$, so mismeasurement of the treatment indicator results in attenuation bias, that is, the mismeasured estimate $\tau(x)$ of the true treatment effect $\tau^*(x)$ is biased toward zero. Related results include Klepper (1988), Manski (1990), Bollinger (1996), and Hotz, Mullin, and Sanders (1997).

Theorem 2 then provides sufficient conditions for identifying and estimating the function $m(x)$, and hence the true treatment effect $\tau^*(x)$. As in Mahajan (2005), again let $X = (Z, V)$ and assume $E(T^* \mid Z, V) \neq E(T^* \mid Z)$, but replace Mahajan's assumption that $E(Y \mid Z, V, T^*) = E(Y \mid Z, T^*)$ with the weaker assumption that $E(Y \mid Z, V, T^* = 1) - E(Y \mid Z, V, T^* = 0) = E(Y \mid Z, T^* = 1) - E(Y \mid Z, T^* = 0)$. Equivalently, rewriting equation (2) as $h^*(X, T^*) = h_0^*(Z, V) + \tau^*(Z, V)T^*$, Mahajan assumes that both $h_0^*(Z, V)$ and $\tau^*(Z, V)$ do not depend on V, while this paper assumes only that $\tau^*(Z, V)$ does not depend on V. For example, a standard regression model where $h^*(x, t)$ is linear in its arguments with nonzero coefficients will satisfy this paper's assumption but violate Mahajan's assumption. Another example is relating wages Y to schooling T^* using a Card (1995, 1996) type distance to school instrument V. The present paper would only require that V be conditionally independent of the increase in wages associated with schooling, rather than conditionally independent of the level of wages. The latter, Mahajan requirement could be violated if, e.g., access to schooling is correlated with access to higher paying jobs on average (this example assumes other possible problems associated with distance to schooling such as those described by Carneiro and Heckman 2002 can be ignored).

Nonparametric identification will not require a continuous instrument, so V can be discrete, but it will be required that V take on at least three values. In contrast, Mahajan (2005) only requires a binary valued instrument. Some intuition for this result is that V affects the true probability of treatment, and under Mahajan's assumptions a change in V affects Y only through this effect on T^*. In contrast, this paper allows V to affect Y both through T^* and through $h_0^*(Z, V)$, so two different changes in V must be observable to separate these effects.

2 Relating True and Mismeasured Effects of Treatment

Let $Y(t)$ denote the outcome from treatment $T^* = t$ for t equal zero or one. If

$$E[Y(t) \mid T^*, X] = E[Y(t) \mid X] \tag{4}$$

then $\tau^*(x) = E[Y(1) - Y(0) \mid X = x]$, which is the conditional average treatment effect. Equation (4) is a weak version of the Rubin (1978) or Rosenbaum and Rubin (1984) unconfoundedness assumption with respect to the true treatment T^*. See, e.g., Heckman, Ichimura, and Todd (1998) for a discussion of this and similar results. In what follows, equation (4) is not imposed. If it does hold where T^* is a treatment indicator, then $\tau^*(x)$ is the conditional average treatment effect, otherwise $\tau^*(x)$ is just the effect on the mean of Y of changing a binary T^* while holding X fixed.

If T^* were observed without error, then equations (1) and (3) would provide an estimator for $\tau^*(x)$, by replacing expectations with nonparametric regressions. Other existing estimators, e.g. those based on matching or conditional propensity scores could also be used in that case.

ASSUMPTION A1: $E(Y \mid X, T^*, T) = E(Y \mid X, T^*)$.

Equivalently, Assumption A1 says that Y is mean independent of $T - T^*$, conditional on X, T^*, so misclassification does not affect the true expected outcome. This is analogous to the classical assumption of independent measurement errors. This can be a substantive assumption if misclassification is due to misperception or deceit on the part of the subject, for example, if T indicates treatment that a respondent believes he or she has received, then Assumption A1 rules out placebo effects. This assumption could also be violated if an individual's propensity to lie about T^* is related to outcomes, e.g., individuals who erroneously claim to have a college degree might also be more aggressive job or wage seekers in general.

Define $r^*(x)$, $b_0(x)$, $b_1(x)$ and $\tau(x)$ as

$$r^*(x) = E(T^* \mid X = x) = \Pr(T^* = 1 \mid X = x)$$

$$b_t(x) = \Pr(T = 1 - t \mid X = x, T^* = t).$$

$$\tau(x) = h(x, 1) - h(x, 0)$$

The function $r^*(x)$ is the conditional (on $X = x$) probability of receiving treatment, while $b_1(x)$ and $b_0(x)$ are the conditional probabilities of misclassifying the treated and the untreated, respectively.

ASSUMPTION A2: $b_0(x) + b_1(x) < 1$ and $0 < r^*(x) < 1$ for all $x \in supp(X)$.

Assumption A2 says first that the sum of misclassification probabilities is less than one, meaning that, on average, observations of T are more accurate than pure guesses (see, e.g., Bollinger 1996). In a binomial response model with misclassification, this assumption is what Hausman, Abrevaya, and Scott-Morton (1998) call the monotonicity condition. Without an assumption like this, by symmetry one could never

tell if the roles of $t = 0$ and $t = 1$ were reversed, and so for example one could not distinguish whether any estimate of $\tau^*(x)$ corresponded to the effect of changing T^* from zero to one or the negative of this effect. This assumption can be relaxed to $b_0(x) + b_1(x) \neq 1$ if we only wish to identify the magnitude but not the sign of $\tau^*(x)$, which may be useful in applications where the sign of the effect is not in doubt and large misclassification probabilities cannot be ruled out. See Hu (2004) for other possible ways to relax this assumption.

Assumption A2 also requires that for any value x we may condition on, there is a nonzero probability of treatment and a nonzero probability of nontreatment, which is needed because a conditional treatment effect cannot be identified if everyone is treated or if no one is treated. If this condition is violated, then we will only obtain identification for values of x where $r^*(x)$ is not zero or one.

Define the following functions

$$r(x) = E(T \mid X = x)$$

$$h(x,t) = E(Y \mid X = x, T = t)$$

$$\tau(x) = h(x,1) - h(x,0)$$

Note that $r(x)$ and $\tau(x)$ are the same as $r^*(x)$ and $\tau^*(x)$, except defined in terms of the observed treatment mismeasure T instead of the true treatment T^*, so if treatment were observed without error, then $r(x)$ would be the conditional probability of treatment and $\tau(x)$ would equal the conditional average treatment effect.

THEOREM 1: If Assumptions A1 is satisfied then there exists a function $m(x)$ with $|m(x)| \leq 1$ such that $\tau(x) = \tau^*(x)m(x)$. If in addition Assumption A2 is satisfied then $m(x) > 0$.

Proofs are in the Appendix. Three expressions for $m(x)$ are provided there. First, $m(x) = \Pr(T^* = 1 \mid X, T = 1) - \Pr(T^* = 1 \mid X, T = 0)$, which shows that $|m(x)| \leq 1$ because m is a difference of probabilities. Second is $[(1 - r(x)]r(x)m(x) = [1 - r^*(x)]r^*(x)[1 - b_0(x) - b_1(x)]$, which signs m, and third is $m(x) = M[b_0(x), b_1(x), r(x)]$ where M is defined by

$$m = M(b_0, b_1, r) = \left(\frac{1}{1 - b_1 - b_0}\right)\left(1 - \frac{(1 - b_1)b_0}{r} - \frac{(1 - b_0)b_1}{1 - r}\right). \tag{5}$$

which is used later for point identification and estimation.

Theorem 1 shows that, given just Assumption A1, the magnitude of the mismeasured treatment effect $\tau(x)$ provides a lower bound on the true treatment effect $\tau^*(x)$, and with the addition of Assumption A2, the sign of the mismeasured effect $\tau(x)$ equals the sign of the true effect $\tau^*(x)$. This is like the attenuation bias property of a classically mismeasured regressor coefficient in a linear bivariate regression, even though here the measurement (misclassification) error is nonclassical and the model is not parameterized.

Theorem 1 also shows that if one is only interested in testing whether $\tau^*(x) = 0$ then misclassification can be ignored, since given Assumptions A1 and A2, $\tau^*(x) = 0$ if and only if $\tau(x) = 0$.

3 Identification

ASSUMPTION A3: Assume $r(x)$ and $\tau(x)$ are identified.

Assumption A3 holds given any data set that permits consistent estimation of conditional expectations of observable data. Identification of the conditional outcome function $h(x,t)$ is not required, though an obvious estimator for $\tau(x)$ would be to difference estimates of $h(x,t)$. Identification of $\tau(x)$ provides identification of a lower bound on $\tau^*(x)$ by Theorem 1. Also, if the misclassification probabilities $b_0(x)$ and $b_1(x)$ can be identified, then given Assumptions A1, A2, and A3, the true treatment effect $\tau^*(x)$ can be identified using $\tau^*(x) = \tau(x)/m(x)$ and equation (5). Additional assumptions will now be made that suffice for this identification.

Partition X into two subvectors V and Z, so $X = (V, Z)$.

ASSUMPTION A4: For some set $\Omega \subset supp(V)$, for each $z \in supp(Z)$ there exists a set $\Omega_z \subset \Omega$ such that for all $v \in \Omega_z$ and $v' \in \Omega_z$, we have $b_t(v,z) = b_t(v',z)$, $\tau^*(v,z) = \tau^*(v',z)$, and for $v \neq v'$, $r^*(v,z) \neq r^*(v',z)$.

In a small abuse of notation, let $b_t(z)$ and $\tau^*(z)$ denote $b_t(v,z)$ and $\tau^*(v,z)$, respectively, for $v \in \Omega_z$. The distribution of V can be discrete, e.g., V could be a scalar that only takes on a few different values. Assumption A4 says that there exists a variable V that affects the true treatment probabilities r^*, but after conditioning on other covariates does not affect the measurement errors b_t and does not affect τ^* (at least for some values that V might take on). If T^* is a treatment and unconfoundedness holds, then Assumption A4 says that after conditioning on other covariates, V does not affect the conditional average treatment effect but is correlated with eligibility or selection for treatment.

Having a V that doesn't affect misclassification probabilities is sometimes used for identification in binomial response models with misclassification. See, e.g., Hausman, Abrevaya, and Scott-Morton (1998), Abrevaya and Hausman (1999), and Lewbel (2000). A typical assumption in misclassified binomial response is that b_0 and b_1 are constants, which would imply that any elements of X could serve as V for that part of Assumption A4.

Having V affect r^* but not τ^* is a weaker version of the type of exclusion assumption that is commonly used in the identification of selection models. See e.g., Heckman (1990) for a discussion. Close variants of this assumption are used by Manski (1990) to sharpen bounds on treatment effects and by Imbens and Angrist (1994) to identify local average treatment effects. The τ^* condition is satisfied if $E(Y \mid Z = z, V = v, T^* = t) = s_1(z,t) + s_2(z,v)$ for some functions s_1 and s_2.

Assumptions A1, A2, A3, and A4 are all equivalent to (or are implied by) assumptions made by Mahajan (2005), but as discussed in the introduction, Mahajan also requires that $h^*(v,z,t) = h^*(v',z,t)$ for v and v' as defined in Assumption A4. However, Mahajan only requires that Ω contain two elements, so V can be binary, while the next assumption here requires that Ω contain at least three elements.

ASSUMPTION A5: Each set $\Omega_z \subset \Omega$ contains three elements $v_k \in \Omega_z$, $k = 0, 1, 2$, such that

$$\left(\frac{\tau(v_0,z)}{r(v_1,z)} - \frac{\tau(v_1,z)}{r(v_0,z)}\right)\left(\frac{\tau(v_0,z)}{1-r(v_2,z)} - \frac{\tau(v_2,z)}{1-r(v_0,z)}\right) \neq \left(\frac{\tau(v_0,z)}{r(v_2,z)} - \frac{\tau(v_2,z)}{r(v_0,z)}\right)\left(\frac{\tau(v_0,z)}{1-r(v_1,z)} - \frac{\tau(v_1,z)}{1-r(v_0,z)}\right)$$

The main content of Assumption A5 is that V can take on at least three values. Assumption A5 is expressed in a form that can be empirically tested, because the $\tau(v,z)$ and $r(v,z)$ functions are conditional expectations of observable data, and so can be directly estimated (they are identified by Assumption A3). Assumption A5 can alternatively be written as requiring that $\tau^*(z) \neq 0$, $b_0(z) + b_1(z) \neq 1$, and a certain inequality holds among just the r and r^* functions (see the Appendix for details). Assumption A5 will therefore fail to hold only if the true treatment effect $\tau^*(z)$ is zero or if a complicated nonlinear equality relationship holds amongst the true and mismeasured conditional treatment probabilities. This would require a perfect coincidence regarding the levels of these probabilities across all the values V can take on.

If Ω has more than three elements, then Assumption A5 will hold as long as, for each value of z, there exists any one triplet v_0, v_1, v_2 of V values in Ω_z that satisfy the necessary inequality. Note that the triplets v_0, v_1, v_2 are permitted to vary by (i.e., depend upon) z.

THEOREM 2: Let Assumptions A1, A2, A3, A4, and A5 hold. Then the conditional misclassification probabilities $b_0(x)$ and $b_1(x)$, the conditional probability of treatment $r^*(x)$, and the effect $\tau^*(x)$ are all identified. Also, if the condition in Assumption A2 that $b_0(x)+b_1(x) < 1$ is replaced by $b_0(x)+b_1(x) \neq 1$, then $\tau^*(x)$ is identified up to sign.

A key component of Theorem 2 is that data on outcomes helps to identify misclassification probabilities. In particular, suppressing z for clarity, it follows from Theorem 1 and Assumption A4 that

$$\tau(v_k)M[b_0, b_1, r(v_0)] = \tau(v_0)M[b_0, b_1, r(v_k)]. \tag{6}$$

Equation (6) depends only on the identified functions τ and r and on the unknowns b_0 and b_1. For each z, evaluating this expression at $k = 1$ and $k = 2$ gives two equations in the two unknowns b_0 and b_1. These equations are nonlinear, but the proof of Theorem 2 shows that these equations have a unique solution and thereby identify b_0 and b_1. Identification of the true treatment effect $\tau^*(x)$ then follows from equation (5).

Equation (6) also shows why identification requires V to take on three values. This equation depends on v_0 and v_k, so evaluating it at $k = 1$ and $k = 2$ requires existence of a v_0, v_1, and v_2. Each additional value that V can take on provides another equation that b_0 and b_1 must satisfy, so the larger is the set of values Ω_z, the greater will be the number of overidentifying restrictions determining b_0 and b_1 at that z.

A binary V would suffice for identification if we had some additional equality restriction on the misclassification probabilities b_0 and b_1. For example, in some applications it may be known that one or the other of these probabilities is zero, such as when T^* is a job training program where we directly observe everyone who takes the offered program, but we do not always observe when someone who turns down the program

gets alternative training elsewhere. An example of a binary V might be a second mismeasured treatment indicator. Theorem 2 could be applied without additional information if we observed three mismeasured treatment indicators, by taking one of them as T and the other two (which together can take on a total of four values) as V.

4 Semiparametric and Nonparametric Estimation

To construct estimators, the previous identification conditions will now be expressed in the form of conditional moments. Assume the distribution of V is discrete, define $\Omega = supp(V) = \{v_0, ..., v_K\}$, for simplicity assume $\Omega_z = \Omega$, and let $r_k^*(z) = r^*(v_k, z)$. Let $W = (Y, T, V)$.

Define the vector valued function $q_0(z)$ as the vector of $K + 4$ elements

$$q_0(z) = \big(b_0(z), b_1(z), r_0^*(z), ...r_K^*(z), \tau^*(z)\big) \tag{7}$$

and define g as the vector valued function $g[q_0(z), w]$ consisting of the following $2K + 2$ elements

$$[b_0(z) + (1 - b_0(z) - b_1(z))r_k^*(z) - T]I(V = v_k), \qquad k = 0, ..., K \tag{8}$$

$$\tau^*(z)I(V = v_k) + \frac{YT - (1 - b_1(z))r_k^*(z)\tau^*(z)I(V = v_k)}{b_0(z) + (1 - b_0(z) - b_1(z))r_k^*(z)}$$
$$- \frac{Y(1 - T) + (1 - b_0(z))(1 - r_k^*(z))\tau^*(z)I(V = v_k)}{1 - [b_0(z) + (1 - b_0(z) - b_1(z))r_k^*(z)]}, \qquad k = 0, ..., K \tag{9}$$

COROLLARY 1: Define the function q_0 by equation (7) and the function g as the vector of functions (8) and (9). For any value of z in its support, if Assumptions A1, A2, A3, A4, and A5 hold then the only function $q(z)$ that satisfies $E[g(q(Z), W) \mid Z = z)] = 0$ and has first two elements that are nonnegative and sum to less than one, is $q(z) = q_0(z)$.

The objects we wish to estimate are elements of $q_0(z)$. Corollary 1 shows that the identification based on Theorem 3 can be expressed as the statement that the unknown functions $q_0(z)$ are the solutions to the vector of conditional moments $E[g(q(Z), W) \mid Z = z)] = 0$.

Based on Corollary 1, the functions of interest $q_0(z)$ can be nonparametrically estimated using nonparametric conditional moment GMM based estimators such as Carrasco and Florens (2000), Newey and Powell (2003), or Ai and Chen (2003). In particular, estimation can take the form of parametric conditional moment GMM, replacing the unknown functions with sieve approximations. Empirical likelihood based conditional moment estimation such as Otsu (2003) could also be used. Lewbel (2006a) provides a simple local GMM estimator for this $q_0(z)$.

To construct semiparametric estimators based on Corollary 1, assume we can write $q_0(z) = s(z, \beta_0)$ where s is a known function and β_0 is a finite vector of unknown parameters. In this case the conditional moments $E[g(q(Z), W) \mid Z = z)] = 0$ imply unconditional moments

$$E[\eta_j(Z)g(s(Z, \beta_0), W)] = 0, \qquad j = 1, ..., J \tag{10}$$

for bounded functions $\eta_j(Z)$ chosen by the econometrician. Given the unconditional moments of equation (10) for $j = 1, ..., J$, we may estimate β_0 using Hansen's (1982) Generalized Method of Moments (GMM). Asymptotic efficiency can be obtained by using estimated optimal $\eta_j(Z)$ functions as in Newey (1993), Donald, Imbens, and Newey (2003), or Dominguez and Lobato (2004).

With this semiparametric estimator only the dependence of probabilities and treatment effects on z is parameterized. The dependence of probabilities and treatment effects on v and on unobservables remains nonparametric. Identification of β_0 will depend on the specification of the function s and $\eta_j(Z)$, but as long as β_0 is identified from $q_0(z) = s(z, \beta_0)$, we can choose $\eta_j(Z)$ functions as above to identify β_0. This identification requires that $2K + 2$ (the dimension of s) times J be greater than or equal to the dimension of β_0.

If Z is discrete with a finite number of support points, then nonparametric estimation of all the functions $q_0(z)$ can be written as a special case of this semiparametric estimator by defining β_0 as the vector of elements $\beta_{j0} = q_0(z_j)$ and letting $\eta_j(Z) = I(Z = z_j)$ where j indexes every value of $z_j = z$ in the support of Z.

It may be of interest to directly estimate the misclassified treatment effect $\tau_k(z) = \tau(v_k, z)$ and misclassified treatment parameters $r_k(z) = r(v_k, z)$. By Theorem 1, τ^* is zero if and only if τ is zero, so estimates of misclassified treatment effects could be used to test if the true treatment effects are zero. Estimates of τ and r could also be used in the bounds calculation of Theorem 1 when validity of V as an instrument is in doubt, and they can be used to test Assumption A5. To express the misclassified functions as conditional moments, define $Q_0(z)$ as the $2K + 2$ vector

$$Q_0(z) = (r_0(z), ...r_K(z), \tau_0(z), ...\tau_K(z)) \tag{11}$$

and define $G[Q_0(z), w]$ as the vector valued function consisting of the $2K + 2$ elements

$$[r_k(z) - T]I(V = v_k), \qquad k = 0, ..., K \tag{12}$$

$$\left(\frac{YT}{r_k(z)} - \frac{Y(1-T)}{1 - r_k(z)} - \tau_k(z)\right) I(V = v_k), \qquad k = 0, ..., K \tag{13}$$

COROLLARY 2: Define the function Q_0 by equation (11) and the function G as the vector of functions (12) and (13). For any value of z in its support, if Assumptions A1 and A2 hold then the only function $Q(z)$ that satisfies the equations $E[G(Q(Z), W) \mid Z = z)] = 0$ is $Q(z) = Q_0(z)$.

Nonparametric or semiparametric estimation based on Corollary 2 proceeds exactly like estimation based on Corollary 1 as described above, replacing q and g with Q and G.

5 Conclusions

This paper provides bounds and conditions for nonparametric point identification of the effect on an outcome of changing a possibly mismeasured binary variable. Given a weak form of the unconfounding assumption, this provides identification of conditional average treatment effects when the treatment indicator may be mismeasured. Estimators that employ these identification conditions are provided, based on direct estimation of relevant conditional expections.

Lewbel (2006a) and an addendum to this paper, Lewbel (2006b), describe both semiparametric and local GMM nonparametric estimation of a vector of functions $q(Z)$ based on moments of the form $E[g(q(Z), W) \mid Z = z)] = 0$, (Corollaries 1 and 2 are examples of such functions and moments), and provides both a small Monte Carlo and an empirical application of Theorem 2 and Corollaries 1 and 2, relating wages to attainment of an undergraduate degree, allowing for misclassification errors in transcript reports.

It would useful to explore how other binary covariate effect estimators such as matching and propensity score based methods might be adapted to the present application where the binary covariate such as treatment is mismeasured.

6 References

Abrevaya, J. and J. A. Hausman, (1999), ""Semiparametric Estimation With Mismeasured Dependent Variables: An Application to Duration Models for Unemployment Spells", Annales d'Economie et de Statistique, 55/56, 243-275.

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

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

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

Card, D. (1995), "Using Geographic Variations in College Proximity to Estimate the Returns to Schooling," in Aspects of Labor Market Behavior: Essays in Honor of John Vanderkamp," L. N. Christofides, E. K. Grand, and R. Swidinsky, eds., Toronto: University of Toronto Press.

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

Carneiro, P. and J. Heckman (2002), "The Evidence on Credit Constraints in Post Secondary Schooling," Economic Journal, 112, 705-734.

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

Das, M., (2004), "Instrumental Variables Estimators of Nonparametric Models With Discrete Endogenous Regressors.," Journal of Econometrics, 124, 335-361.

Dominguez, M. and I. Lobato, (2004), "Consistent Estimation of Models Defined by Conditional Moment Restrictions," Econometrica, 72, 1601-1615.

Donald, S. G., G. W. Imbens, and W. K. Newey (2003), "Empirical Likelihood Estimation and Consistent Tests With Conditional Moment Restrictions," Journal of Econometrics, 117, 55-93.

Florens, J.-P. and L. Malavolti (2003), "Instrumental Regression with Discrete Variables," unpublished manuscript..

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

Hausman, J. A., J. Abrevaya, and F. M. Scott-Morton (1998), "Misclassification of the Dependent Variable in a Discrete-Response Setting," Journal of Econometrics, 87, 239-269.

Heckman, J. (1990), "Varieties of Selection Bias," American Economic Review, Papers and Proceedings, 80, 313-338.

Heckman, J. H. Ichimura and P. Todd, (1998), "Matching as an Econometric Evaluations Estimator, Review of Economic Studies, 65, 261-294.

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.

Hu, Y., (2005), "Identification and estimation of nonlinear models with misclassification error using instrumental variables," unpublished manuscript.

Imbens, G. W. and J. D. Angrist, (1994), "Identification and Estimation of Local Average Treatment Effects," Econometrica, 62, 467-475.

Kane, T. J., and C. E. Rouse, (1995), "Labor market returns to two- and four- year college," American Economic Review, 85, 600-614

Kane, T. J., C. E. Rouse, and D. Staiger, (1999), "Estimating Returns to Schooling When Schooling is Misreported," NBER working paper #7235.

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

Lewbel, A., (2000), "Identification of the Binary Choice Model With Misclassification," Econometric Theory, 16, 603-609.

Lewbel, A., (2006a), "A Local Generalized Method of Moments Estimator," unpublished manuscript, Boston College.

Lewbel, A., (2006b), "Estimation of Average Treatment Effects With Misclassification - Addendum," available as http://www2.bc.edu/~lewbel/mistrea2addendum

Mahajan, A. (2005) "Identification and Estimation of Regression Models with Misclassification," Unpublished manuscript, Stanford University.

Manski, C. F. (1990) "Nonparametric Bounds on Treatment Effects," American Economic Review Papers and Proceedings, 80, 319-323.

Newey, W. K. (1993), "Efficient Estimation of Models With Conditional Moment Restrictions," in Handbook of Statistics, vol. 11, ed. by G. S. Maddala, C. R. Rao, and H. D. Vinod, Amsterdam: North Holland, chapter 16.

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

Otsu, T., (2003), "Penalized Empirical Likelihood Estimation of Conditional Moment Restriction Models with Unknown Functions," unpublished manuscript.

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.

Rubin, D. B. (1978), "Bayesian Inference for Causal Effects: The Role of Randomization," Annals of Statistics, 6, 34—58

7 Appendix

Define $p_t(X) = E(T^* \mid X, T = t) = \Pr(T^* = 1 \mid X, T = t)$. Suppressing the X argument for clarity, some relationships to be used later are, by Bayes theorem

$$p_0 = \frac{b_1 r^*}{(1-r)} \quad \text{and} \quad p_1 = \frac{(1-b_1)\, r^*}{r}. \tag{14}$$

Also,

$$r = E(T) = E(T \mid T^* = 1)\Pr(T^* = 1) + E(T \mid T^* = 0)\Pr(T^* = 0)$$

$$r = (1-b_1)r^* + b_0(1-r^*) \tag{15}$$

which gives $r = b_0$ when $b_0 + b_1 = 1$, otherwise

$$r^* = \frac{r - b_0}{1 - b_0 - b_1} \quad \text{and} \quad 1 - r^* = \frac{1 - b_1 - r}{1 - b_0 - b_1}. \tag{16}$$

PROOF OF THEOREM 1: Continuing to suppress the X argument, we have by equations (1) and (2) and Assumption A1 that $E(Y \mid T^*, T) = h_0^* + \tau^* T^*$. By the law of iterated expectations this gives $E(Y \mid T = t) = h_0^* + \tau^* p_t$. Then since $\tau = E(Y \mid T = 1) - E(Y \mid T = 0)$ we obtain $\tau = (p_1 - p_0)\tau^*$, so m in Theorem 1 equals $p_1 - p_0$, and $-1 \leq m \leq 1$ follows from m equalling the difference between two probabilities.

Using equation (14)

$$m = p_1 - p_0 = \frac{(1-b_1)\,r^*}{r} - \frac{b_1 r^*}{(1-r)} \tag{17}$$

$$m = \frac{r^*}{(1-r)r}\,(1-b_1-r)$$

and using equation (16) for $1-r^*$,

$$(1-r)rm = (1-r^*)r^*\,(1-b_0-b_1)\,. \tag{18}$$

Since probabilities r and r^* lie between zero and one, this shows that $m > 0$ when Assumption A2 holds, $m = 0$ when $b_0 + b_1 = 1$, and that $m < 0$ when the sum of the misclassification probabilities $b_0 + b_1$ is greater than one. Also, substituting equation (16) for r^* into equation (17) yields, after some algebraic simplification, equation (5).

PROOF OF THEOREM 2: For clarity suppress z, and for the functions r and τ denote $r_k = r(v_k)$ and $\tau_k = \tau(v_k)$. For any given z, we have for all $v_0 \in \Omega_z$ and $v_k \in \Omega_z$ that, by Theorem 1 and equation (5),

$$0 = M(b_0, b_1, r_k)\tau_0 - M(b_0, b_1, r_0)\tau_k$$

$$0 = \left(1 + \frac{(b_1-1)b_0}{r_k} + \frac{(b_0-1)b_1}{1-r_k}\right)\tau_0 - \left(1 + \frac{(b_1-1)b_0}{r_0} + \frac{(b_0-1)b_1}{1-r_0}\right)\tau_k \tag{19}$$

$$0 = (1-b_1)b_0\left(\frac{\tau_0}{r_k} - \frac{\tau_k}{r_0}\right) + (1-b_0)b_1\left(\frac{\tau_0}{1-r_k} - \frac{\tau_k}{1-r_0}\right) + \tau_k - \tau_0 \tag{20}$$

Write this equation as

$$0 = B_0 w_{0k} + B_1 w_{1k} + w_{2k}$$

where $B_t = (1-b_{1-t})b_t$ and each w_{jk} is a function of r_0, r_k, τ_0, and τ_k. Given that $\Omega_z \in \Omega$ contains three elements v_0, v_1, and v_2, we have two equations $0 = B_0 w_{0k} + B_1 w_{1k} + w_{2k}$ for $k = 1, 2$ that are linear in the two unknowns B_0 and B_1, and so can be uniquely solved as long as the matrix of elements w_{jk}, $j = 0, 1$, $k = 1, 2$, is nonsingular. The determinant of this matrix is

$$\left(\frac{\tau_0}{r_1} - \frac{\tau_1}{r_0}\right)\left(\frac{\tau_0}{1-r_2} - \frac{\tau_2}{1-r_0}\right) - \left(\frac{\tau_0}{r_2} - \frac{\tau_2}{r_0}\right)\left(\frac{\tau_0}{1-r_1} - \frac{\tau_1}{1-r_0}\right) \tag{21}$$

and the inequality in Assumption A5 makes this determinant nonzero, as required.

Now let $s = 1-b_1-b_0$. It follows from $B_t = (1-b_{1-t})b_t$ that $(s+b_0)b_0 = B_0$ and $2b_0 = B_0 - B_1 + 1 - s$. Substituting the second of these equations into the first and solving for s gives

$$1 - b_1 - b_0 = s = \pm\left[(B_0 - B_1 + 1)^2 - 4B_0\right]^{1/2}$$

if the assumption regarding s is $s \neq 0$, then we have that s is identified up to sign. By Theorem 1 $\tau^* = \tau(v)/M[b_0, b_1, r(v)]$ and

$$M[b_0, b_1, r(v)] = \left(\frac{1}{s}\right)\left(1 - \frac{B_0}{r(v)} - \frac{B_1}{1 - r(v)}\right)$$

so it follows that τ^* is identified up to sign. Making the stronger assumption that $s > 0$, we have s is identified, so b_0 and b_1 are now identified by $b_0 = (B_0 - B_1 + 1 - s)/2$ and $b_1 = -(B_0 - B_1 + 1 + s)/2$, and by equations (16), (5) and Theorem 1, identification of these misclassification probabilities means that r^* and τ^* are also identified.

Define $R_k = \left[(1 - r_k^*)r_k^*\right] / \left[(1 - r_k)r_k\right]$. Using equation $\tau = m\tau^*$ and equation (17), the determinant (21) can be rewritten as

$$\left[\left(\frac{R_0}{r_1} - \frac{R_1}{r_0}\right)\left(\frac{R_0}{1 - r_2} - \frac{R_2}{1 - r_0}\right) - \left(\frac{R_0}{r_2} - \frac{R_2}{r_0}\right)\left(\frac{R_0}{1 - r_1} - \frac{R_1}{1 - r_0}\right)\right](1 - b_0 - b_1)\,\tau^*$$

So the nonzero determinant condition can equivalently be written as requiring that $\tau^* \neq 0$, $b_0 + b_1 \neq 1$, and the above described function of r_k^* and r_k for $k = 0, 1, 2$ not equal zero.

PROOF OF COROLLARY 1: To ease notation, drop the argument z everywhere and let all expectations below be conditional on $Z = z$. Let $I_k = I(V = v_k)$. Having the mean of equation (8) equal zero makes $b_0 + (1 - b_0 - b_1)r_k^* = E(I_k T)/E(I_k)$, which equals the true r_k by definition of r_k. Solving the resulting equation $b_0 + (1 - b_0 - b_1)r_k^* = r_k$ for r_k^* and substituting the result into equation (9) gives

$$\left(\frac{YT}{r_k} - \frac{(1 - b_1)\tau^*}{r_k}\frac{r_k - b_0}{(1 - b_0 - b_1)} - \frac{Y(1 - T)}{1 - r_k} - \frac{(1 - b_0)\tau^*}{1 - r_k}\frac{1 - b_1 - r_k}{(1 - b_0 - b_1)} + \tau^*\right) I_k$$

Setting the mean of this result to zero and dividing by $E(I_k)$ gives

$$\frac{E(YTI_k)}{r_k E(I_k)} - \frac{(1 - b_1)\tau^*}{r_k}\frac{r_k - b_0}{(1 - b_0 - b_1)} - \frac{E[Y(1 - T)I_k]}{(1 - r_k)E(I_k)} - \frac{(1 - b_0)\tau^*}{1 - r_k}\frac{1 - b_1 - r_k}{(1 - b_0 - b_1)} + \tau^* = 0$$

which, using $r_k = E(TI_k)/E(I_k)$ simplifies to

$$\frac{E(YTI_k)}{E(TI_k)} - \frac{(1 - b_1)\tau^*}{r_k}\frac{r_k - b_0}{(1 - b_0 - b_1)} - \frac{E[Y(1 - T)I_k]}{E[(1 - T)I_k]} - \frac{(1 - b_0)\tau^*}{1 - r_k}\frac{1 - b_1 - r_k}{(1 - b_0 - b_1)} + \tau^* = 0.$$

which, after rearranging terms and using $E(TI_k) = prob(T = 1, V = v_k)$ gives

$$\begin{aligned} E(Y \mid T = 1, V = v_k) - E(Y \mid T = 0, V = v_k) \\ = \left(\frac{(1 - b_1)}{r_k}\frac{r_k - b_0}{(1 - b_0 - b_1)} + \frac{(1 - b_0)}{1 - r_k}\frac{1 - b_1 - r_k}{(1 - b_0 - b_1)} - 1\right)\tau^* \end{aligned}$$

which, by the definitions of the function τ and m reduces to $\tau(v_k) = M(b_0, b_1, r_k)\tau^*$. It has now been shown that the conditional mean of g equalling zero is equivalent to $r(v_k) = b_0 + (1 - b_0 - b_1)r_k^*$ and $\tau(v_k) = M[b_0, b_1, r(v_k)]\tau^*$ with the true functions $r(v_k)$ and $\tau(v_k)$, and by Theorem 2 the only solutions

to these equations for $k = 0, ..., K$ that also satisfy $b_0 \geq 0$, $b_1 \geq 0$, and $b_0 + b_1 < 1$ are the true values of $r_0^*, ..., r_K^*, b_0, b_1$, and τ^*.

PROOF OF COROLLARY 2: Setting the conditional mean of equation (12) equal to zero and solving for $r_k(z)$ yields the definition of $r_k(z)$, and setting the conditional mean of equation (13) equal to zero and solving for $\tau_k(z)$ yields the definition of $\tau_k(z)$.

Estimation of Average Treatment Effects With Misclassification - Addendum

Arthur Lewbel*
Boston College

revised June 2006

This addendum provides additional material regarding the paper, "Estimation of Average Treatment Effects With Misclassification." This additional material consists of

1. Generic parametric, semiparametric, and nonparametric estimators for vectors of functions $q_0(z)$ based on the conditional moment restriction $E[g(q_0(Z), W) \mid Z = z)] = 0$. The treatment effects paper provides one example of conditional moment restrictions of this form.

2. A small Monte Carlo study.

3. An empirical application attempting to estimate the effect on wages of attaining a college degree, where completion of this undergraduate degree may be misreported.

See also Lewbel (2006), "A Local Generalized Method of Moments Estimator."

1 Estimation

This section describes general estimators for vectors of functions $q_0(z)$ based on the conditional moment restriction $E[g(q_0(Z), W) \mid Z = z)] = 0$. While the application in this paper is to treatment effect estimation using the Corollaries in the previous section, many other econometric models can also be cast as conditional moment restrictions of this type. For example, consider the nonparametric probit model

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

$W = I[q_0(z) + e \geq 0]$, where $q_0(z)$ is an unknown function and e is a standard normal independent of Z, or has some other known distribution. Then $g(q_0(Z), W) = W - F_e[q_0(Z)]$ where F_e is the known CDF of $-e$. Nonparametric censored or truncated regression would have a similar form. The estimators could also be used for Euler equations, which are mean zero conditional on information in a given time period, and may have parameters $q_0(z)$, such as preference parameters, that are unknown functions of observables.

Three estimators for conditional moment restrictions are considered. The first is standard GMM, for use when $q_0(z)$ can be finitely parameterized. The second and third are for nonparametric estimation of $q_0(z)$ when Z is discretely or continuously distributed. For these estimators it is assumed that we have data consisting of Z_i, W_i for $i = 1, ..., n$. Limiting distributions are provided assuming these observations are independent and identically distributed.

1.1 Parameterized Estimation

Suppose we can write $q_0(z) = s(z, \beta_0)$ where s is a known function and β_0 is finite vector of unknown parameters. In this case the conditional moments $E[g(q(Z), W) \mid Z = z)] = 0$ imply unconditional moments

$$E[\eta_j(Z)g(s(Z, \beta_0), W)] = 0, \qquad j = 1, ..., J \tag{1}$$

for any J bounded functions $\eta_j(Z)$ chosen by the econometrician. Given the unconditional moments of equation (1) for $j = 1, ..., J$, we may estimate β_0 using Hansen's (1982) Generalized Method of Moments (GMM). Asymptotic efficiency can be obtained by using estimated optimal $\eta_j(Z)$ functions as in Newey (1993), Donald, Imbens, and Newey (2003), or Dominguez and Lobato (2004).

When applying this estimator to Corollary 1, only the dependence of probabilities and treatment effects on z is parameterized. The dependence of probabilities and treatment effects on v and on unobservables remains nonparametric. Identification of β_0 will depend on the specification of the function s and $\eta_j(Z)$, but Corollary 1 implies that as long as β_0 is identified from $q_0(z) = s(z, \beta_0)$, it should be possible to choose $\eta_j(Z)$ functions to identify β_0. This identification requires that $2K + 2$ (the dimension of s) times J be greater than or equal to the dimension of β_0.

1.2 Estimation With Discrete Covariates

Now assume we do not have a parameterization for $q_0(z)$, but Z is discretely distributed, or more specifically, that Z has one or more mass points and we only wish to estimate $q_0(z)$ at those points.

Let $\theta_{z0} = q_0(z)$. If the distribution of Z has a mass point with positive probability at z, then

$$E[g(\theta_z, W) \mid Z = z] = \frac{E[g(\theta_z, W)I(Z = z)]}{E[I(Z = z)]}$$

so $E[g(q_0(z), W) \mid Z = z] = 0$ if and only if $E[g(\theta_{z0}, W)I(Z = z)] = 0$. It therefore follows from Corollary 1 that θ_{z0} is identified from these moment conditions, and we may estimate parameters θ_{z0} by the ordinary GMM estimator

$$\widehat{\theta}_z = \arg\min_{\theta_z} \sum_{i=1}^{n} g(\theta_z, W_i)' I(Z_i = z) \Omega_n \sum_{i=1}^{n} g(\theta_z, W_i)' I(Z_i = z) \tag{2}$$

for some sequence of positive definite Ω_n. If Ω_n is a consistent estimator of $\Omega_{z0} = E[g(\theta_{z0}, W)g(\theta_{z0}, W)'I(Z = z)]^{-1}$, then efficient GMM gives

$$\sqrt{n}(\widehat{\theta}_z - \theta_{z0}) \to^d N\left(0, \left[E\left(\frac{\partial g(\theta_{z0}, W)I(Z = z)}{\partial \theta_z'}\right)\Omega_{z0} E\left(\frac{\partial g(\theta_{z0}, W)I(Z = z)}{\partial \theta_z'}\right)'\right]^{-1}\right)$$

Now consider applying this to the moments in Corollary 1. Standard GMM assumes parameters have compact support. This could be imposed, consistent with Assumptions A2 and A5 by assuming that $\delta \leq r_k^*(z) \leq 1 - \delta$, $0 \leq b_t(z)$, $b_0(z) + b_1(z) \leq 1 - \delta$, and $\delta \leq |\tau^*(z)| \leq 1/\delta$ for some small $\delta > 0$.

Let $\widehat{\tau}_z^*$ denote the element of $\widehat{\theta}_z$ that corresponds to the marginal effect $\tau^*(z)$. This $\widehat{\tau}_z^*$ is a consistent estimator of $\tau^*(z)$ provided that this effect is nonzero. However, $\tau^*(z) = 0$ violates Assumption A5, so one cannot use an ordinary Wald t-statistic to test for a zero treatment effect (the ordinary t statistic will be valid for testing other values, such as whether $\tau^*(z)$ equals a given small, nonzero value). But, by Theorem 1, $\tau^*(z) = 0$ if and only if $\tau(v_k, z) = 0$, so we may estimate $\tau(v_k, z)$ by applying the above GMM estimator with q and g as defined by Corollary 2, and then perform an ordinary Wald test of the hypothesis that $\tau(v_k, z)$ is zero for $k = 0, ..., K$.

More generally, the moments in Corollary 2 may be estimated either separately, or (if $\tau^*(z)$ is nonzero) together with those of Corollary 1 to test differences between true and misclassified treatment probabilities or effects.

If $K > 3$ using Corollary 2 then θ_{z0} is overidentified and standard tests of moment validity such as Hansen's J test may be applied. Additional moments for estimating $\widehat{\theta}_z$ could also be constructed given more information about the misclassification probabilities $b_0(z)$ and $b_1(z)$. For example, in some applications it may be known that one or the other of these probabilities is zero, or that these probabilities are equal to each

other. Given either of these constraints, only a binary V would be required for identification and estimation based on Theorem 3.

GMM based on Corollary 1 or 2 provides estimates conditional on a given $z \in supp(Z)$. The unconditional marginal effect, corresponding to the unconditional treatment effect given the weak form of unconfounding, is

$$E(Y \mid T^* = 1) - E(Y \mid T^* = 0) = \sum_{z \in supp(Z)} \tau_z^* E[I(Z = z)]$$

which could be estimated by the sample average $\sum_{i=1}^{n} \sum_{z \in supp(Z)} \widehat{\tau}_z^* I(Z_i = z)/n$. Note that $cov(\widehat{\tau}_z^*, \widehat{\tau}_{\widetilde{z}}^*) = 0$ for $z \neq \widetilde{z}$ because $\widehat{\tau}_z^*$ and $\widehat{\tau}_{\widetilde{z}}^*$ are estimated using different subsets of data.

If constraints are known to exist on the parameters across values of z, then the GMM estimates for each z can be stacked into one large GMM to improve efficiency. For example, if misclassification probabilities b_t are known to be constant, or more generally independent of some elements of z, then that restriction could be imposed in the collection of moments $E[g(\theta_{z0}, W)I(Z = z)] = 0$ for all $z \in supp(Z)$.

1.3 Local GMM Estimation For Continuous Covariates

Continue to assume that $E[g(q_0(Z), W) \mid Z = z] = 0$, where g is known, q_0 is unknown and not parameterized, and now Z is continuously distributed. A local GMM estimator is proposed. The idea is to apply equation (2) to the case of continuous Z by replacing averaging over just observations $Z_i = z$ with local averaging over observations Z_i in the neighborhood of z.

Assumption C1. Let Z_i, W_i, $i = 1, ..., n$, be an iid random sample of observations of the random vectors Z, W. The d vector Z is continuously distributed with density function $f(Z)$. For given point z in the interior of $supp(Z)$ having $f(z) > 0$ and a given vector valued function $g(q, w)$ where $g(q(z), w)$ is twice differentiable in the vector $q(z)$ for all $q(z)$ in some compact set $\Theta(z)$, there exists a unique $q_0(z) \in \Theta(z)$ such that $E[g(q_0(z), W) \mid Z = z] = 0$. Let Ω_n be a finite positive definite matrix for all n, as is $\Omega = plim_{n \to \infty} \Omega_n$.

Assumption C1 provides the required moment condition structure for the model, and Assumption C2

below provides conditions for local averaging. Define $\varepsilon[q(z), W]$, $V(z)$, and $R(z)$ by

$$\begin{aligned} \varepsilon[q(z), W] &= g(q(z), W)f(z) - E[g(q(z), W)f(Z) \mid Z = z] \\ V(z) &= E\left[\varepsilon(q_0(z), W)\varepsilon(q_0(z), W)^T \mid Z = z\right] \\ R(z) &= E\left(\frac{\partial g[q_0(z), W]}{\partial q_0(z)^T} f(Z) \mid Z = z\right) \end{aligned}$$

Assumption C2. 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 $q(z) \in \Theta(z)$, $E[||g(q(z), W)f(Z)||^\eta \mid Z = z]$, $V(z)$, $R(z)$, and $Var[[\partial g(q(z), W)/\partial q(z)]f(Z) \mid Z = z]$ are finite and continuous at z and $E[g(q(z), W)f(Z) \mid Z = z]$ is finite and twice continuously differentiable in z.

Define

$$S_n(q(z)) = \frac{1}{nb^d} \sum_{i=1}^{n} g[q(z), W_i] K\left(\frac{z - Z_i}{b}\right)$$

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

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

THEOREM 3: Given Assumptions C1 and C2, if the bandwidth b satisfies $nb^{d+4} \to 0$ and $nb^d \to \infty$, then $\widehat{q}(z)$ is a consistent estimator of $q_0(z)$ with limiting distribution

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

Theorem 3 assumes a bandwidth rate that makes bias shrink faster variance, and so is not mean square optimal. One could instead choose the mean square optimal rate where nb^{d+4} goes to a constant, but the resulting bias term would then have a complicated form that depends on the kernel regression biases in both $S_n(q_0(z))$ and its derivative with respect to $q_0(z)$, among other terms.

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

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

and $R(z)$ can be estimated using

$$R_n(z) = \frac{1}{nb^d} \sum_{i=1}^{n} \frac{\partial g[\widehat{q}(z), W_i]}{\partial \widehat{q}(z)^T} K\left(\frac{z - Z_i}{b}\right)$$

This local GMM is closely related to the local nonlinear least squares estimator of Gozalo and Linton (2000). Alternative nonparametric conditional moment GMM based estimators include Carrasco and Florens (2000), Newey and Powell (2003), and Ai and Chen (2003). Empirical likelihood based estimators might also be used. Advantages of Theorem 3 are that it is relatively simple, and in the present context is a natural extension of the discrete Z estimator. In particular, the two estimators could be combined 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.

2 Monte Carlo

For each simulated observation, a V is drawn with $prob(V = v_k) = 1/3$ from $\Omega = \{0, 1, 2\}$. Next a T^* is drawn from $\{0, 1\}$ with $prob(T^* = 1 \mid V = v_k) = r^*(v_k)$ with $r^*(0) = 3/4$, $r^*(1) = 1/2$, and $r^*(2) = 1/4$. Next an outcome Y is drawn from a normal $N(T^*, 1)$ distribution, which makes the average treatment effect be $\tau^* = 1$, and finally T is randomly drawn from $\{0, 1\}$ with $prob(T = T^*) = .8$, so the misclassification probabilities are $b_0 = b_1 = .2$. This model is exactly identified using Theorem 2 and Corollary 1, so the GMM weighting matrix is taken to be the identity matrix. The sample size is $n = 1000$ and the number of simulations is 10, 000.

This design makes $\tau(0) = \tau(2) = .4945$ and $\tau(1) = .6$ so the limiting value of the estimated average treatment that would be obtained if one did not correct for misclassification error is $E[\tau(V)] = .530$, that is, a bias of 47% of the true effect $\tau^* = 1$.

Estimation was done using constrained GMM, with the constraints being that the estimated probabilities of treatment r^* lie between .01 and .99, and that the misclassification probabilities b_0 and b_1 be nonnegative with $b_0 + b_1 \leq .99$.

Table 1 reports the results. Estimation of the treatment effect τ^* shows a 5% mean bias and 1% median bias, compared to the 47% bias that would result from failure to account for misclassification. The estimates of treatment and misclassification probabilities display similar levels of accuracy.

3 Wages and Schooling

This section uses Theorem 2 to identify and estimate the effect on wages of attaining a college degree, where completion of this undergraduate degree may be misreported. The data, from Kane and Rouse (1995), is constructed using the National Longitudinal Survey of the High School Class of 1972 (NLS-72). The outcome Y is logged hourly wage, collected 14 years after the respondents graduated from high school. As in Kane and Rouse (1995), respondents with $Y < \$1.67$ (half the minimum wage at the time) or $Y > \$60$ are excluded.

The true T^* is a the binary indicator of obtaining an undergraduate degree. The reported indicator T is constructed using the Post-secondary Education Transcript Survey (PETS), which contains transcript information on all post-secondary schools reported by the respondents through 1979. Following Card (1995, 1996) and others, the instrument V is distance of the respondent's high school to the nearest 4-year college (as reported by high school counselors in the NLS-72) in quintiles, so $V = 1$ for the 20% of respondents that lived closest to a college, $V = 2$ for those in the second quintile of distance, etc.,.

Serious objections to the use of schooling distance have been raised in the literature (see, e.g., Carneiro and Heckman 2002), though as noted in the introduction, distance may satisfy this paper's assumptions even if it is not valid as an ordinary instrument. Also, even if the effect τ^* is estimated consistently, it may not equal a true average treatment effect because of possible confoundedness.

The vector of covariates Z, from the NLS-72, consists of an indicator of whether the respondent is African American (black), the log of the family income of the respondent in 1972 (income), the respondent's high school class rank (phsrank) expressed as a percentage, an index of the respondent's standarized test scores (stotal), the respondent's total number of years of work experience (totexp), an indicator of whether the respondent is female (female), an indicator of whether the respondent lived in a large city in 1972 (large city), and an indicator of whether the respondent worked part time (parttime).

This set of covariates, a subset of those considered by Kane and Rouse (1995), is selected because these variables appear to have the largest effects on Y. Less influential variables are omitted to facilitate model estimation or, in some cases, because their inclusion resulted in a failure to converge. The total sample size

is 6,763 respondents who were working as of 1986, of which 2,073 have $T = 1$, that is, a transcript reported undergraduate degree.

Nonparametric estimation is not practical given the number of relevant covariates, so semiparametric estimation is used. Let Φ denote the cumulative standard normal density and let k index values V can take on, with $v_k = k + 1$ for $k = 0, ..., 4$. The functions comprising $s(z, \beta)$, where $q_0(z) = s(z, \beta_0)$, are parameterized as

$$b_0(z) = 0, \quad b_1(z) = \Phi\left(\gamma_1 + z'\beta_1\right), \quad \tau^*(z) = \delta + z'\beta_3, \tag{3}$$

$$r_k^*(z) = \Phi\left(\alpha_k + z'\beta_2\right), \quad k = 0, ..., 4 \tag{4}$$

and β denotes the set of all the parameters α_j, β_k, γ_1, etc.,. Note that only the dependence of probabilities and the marginal effect on Z is parameterized here. Conditional outcomes $h^*(x, t)$ are not parameterized, and conditional on Z, the effects of the instrument V and the roles of errors or unobservables are nonparametric

The assumption that $b_0(z) = 0$ was made after first specifiying this conditional probability either as $b_0(z) = \Phi\left(\gamma_0 + z'\beta_0\right)$ or as $b_0(z) = \gamma_0$, and finding both that $b_0(z)$ was not significantly different from zero in the estimated models, and that its inclusion resulted in difficulties in obtaining numerical convergence. This assumption implies a zero probability of transcript reported education indicating that a respondent did not receive a college degree when, in fact, he or she did receive a degree.

The unconditional moments of equation (1) used for GMM estimation are $E\left[g\left(s(z, \beta_0), W\right) \otimes Z\right] = 0$. For comparative purposes, the misclassified treatment effect $\tau(z)$ and the misclassified treatment parameters $r_k(z)$ for $k = 0, ..., 4$, are also estimated based on the moments used in Corollary 2 with the above described Z parameterizations. These naive results are then compared to the estimates that allow for misclassification.

Tables 2 and 3 report estimation results. The columns labeled "True Effect" display results from the models that allow for misclassification, while the columns labeled "Mismeasured Effect" report the comparative naive estimates that ignore possible misclassification. Table 2 reports coefficient estimates, specifically, the first two columns of this table are estimates of β_{20}, that is, the covariate coefficients in the true and mismeasured treatment probability functions $r_k^*(z)$ and $r_k(z)$ respectively. The third column is β_{10}, the coefficients in the misclassification probability function $b_1(z)$, and the last two columns are β_{30}, the coefficients in the true and mismeasured marginal effect functions $\tau^*(z)$ and $\tau(z)$, respectively. To simplify the presentation, estimates of the constant terms γ_{10}, δ_0, ζ_0, and α_{k0} for $k = 0, ..., 4$ are not reported. Instead, more interpretable summary statistics are provided in Table 3. These tables report the mean and

standard deviation across the sample of estimated treatment probabilities for each value of V, estimated misclassification probabilities, and the estimated marginal wage effects.

For the model estimated under Theorem 2, the Hansen J statistic is 68 with 58 degrees of freedom, while the corresponding naive model (assuming no misclassification) has a J statistic of 369 with 83 degrees of freedom, so by this measure allowing for misclassification greatly improves fit. However, the parameterizations of equations (3) and (4) are rather simplistic, and all of the models are rejected by the J statistic, so caution is advised in interpreting the results. Ideally, one would like to include more of the covariates considered by Kane and Rouse (1995), and use the nonparametric estimators, but that would only be practical with a far larger data set. The interpretation of the estimated marginal effect may not equal a treatment effect due to confounding, and one may also question the validity of V, though see the discussions above.

With these caveats, it appears that allowing for misclassification has a substantial impact on the estimates of treatment probabilities and marginal effects. For example, in Table 3 the estimated average increase in wages associated with having an undergraduate degree, $E\left[\tau^*(z)\right]$, is 38% while the corresponding naive estimate that does not allow for misclassification is only 11%. The estimated average probability of misclassification from Table 3 is 31% which is implausibly high, however, the estimated effect on wages of getting a degree is more consistent with the findings of other researchers than the low naive estimator that ignores misclassification. For example, after combining transcript and self report data, Kane, Rouse, and Staiger (1999) report a marginal wage effect of an undergraduate education of approximately 25%.

PROOF OF THEOREM 3: Define

$$\begin{aligned} S_n'(q(z)) &= \frac{\partial S_n(q(z))}{\partial q(z)^T} = \frac{1}{nb^d}\sum_{i=1}^{n}\frac{\partial g[q(z), W_i]}{\partial q(z)^T}K\left(\frac{z-Z_i}{b}\right) \\ Q_n(q(z)) &= S_n(q(z))^T \Omega_n S_n(q(z)) \end{aligned}$$

Let $S_0(q(z)) = plim_{n\to\infty} S_n(q(z))$ and similarly for S_n' and Q_n. Assumptions C1 and C2 give sufficient conditions for consistency of these kernel estimators, so these probability limits exist and

$$\begin{aligned} S_0(q(z)) &= E[g(q(z), W) f(z) \mid Z = z] \\ Q_0(q(z)) &= S_0(q(z))^T \Omega S_0(q(z)). \end{aligned}$$

Now consider consistency. We have pointwise convergence of $S_n(q(z))$ to $S_0(q(z))$and compactness of $\Theta(z)$. It is also the case that $|S_n'(q(z))| = O_p(1)$, since $|S_n'(q(z))|$ is a kernel estimator, and standard

conditions have been provided for its consistency, that is, $plim|S_n'(q(z))| = E[|\partial g(q(z), W)/\partial q(z)|f(Z) \mid Z = z]$. This suffices for stochastic equicontinuity, and therefore we have the uniform convergence

$$plim \sup_{q(z)\in\Theta(z)} |S_n(q(z)) - S_0(q(z))| = 0.$$

It follows that $Q_n(q(z))$ also converges uniformly to $Q_0(q(z))$. The assumptions provide compactness of $\Theta(z)$ and imply continuity of $Q_0(q)$. The quadratic form of Q_0 is uniquely maximized at $S_0(q_0(z)) = 0$ and hence at $q(z) = q_0(z)$, so the standard conditions for consistency $plim\widehat{q}(z) = q_0(z)$ are satisfied.

For the limiting distribution, Taylor expanding the first order conditions as in standard GMM gives

$$S_n'(\widehat{q}(z))^T \Omega_n \left[S_n(q_0(z)) + S_n'(\widetilde{q}(z))(\widehat{q}(z) - q_0(z))\right] = 0$$

where $\widetilde{q}(z)$ lies between $\widehat{q}(z)$ and $q_0(z)$. By consistency of $\widehat{q}$, the uniform convergence of S_n, and using $R(z) = S_0'(q_0(z))$, this simplifies to

$$R(z)^T \Omega \left[S_n(q_0(z)) + R(z)(\widehat{q}(z) - q_0(z))\right] = o_p(1)$$

Solving for $\widehat{q}(z) - q_0(z)$ and multiplying by $(nb)^{1/2}$ gives

$$(nb)^{1/2}(\widehat{q}(z) - q_0(z)) = (R(z)^T \Omega R(z))^{-1} R(z)^T \Omega (nb)^{1/2} S_n(q_0(z)) + o_p((nb)^{1/2}).$$

Now $S_0(q_0(z)) = 0$ and standard kernel regression limiting distribution theory gives

$$(nb)^{1/2} S_n(q_0(z)) \to^d N[0, V(z) \int K(u)^2 du]$$

and the theorem follows.

Table 1. Simulation Results

	TRUE	MEAN	SD	LQ	MED	UQ	RMSE	MAE	MDAE
$r^*(0)$	.750	.722	.105	.662	.734	.797	.109	.083	.067
$r^*(1)$	.500	.481	.100	.414	.484	.553	.102	.083	.071
$r^*(2)$	.250	.243	.115	.157	.231	.323	.115	.095	.085
b_0	.200	.193	.097	.133	.213	.268	.097	.077	.068
b_1	.200	.180	.074	.124	.190	.237	.076	.061	.053
τ^*	1.00	1.05	.298	.825	1.01	1.24	.302	.239	.203

Notes: The reported statistics are as follows. TRUE is the true value of the parameter, MEAN and SD are the mean and standard deviation of the estimates across the simulations. LQ, MED, and UQ are the 25% (lower) 50% (median) and 75% (upper) quartiles. RMSE, MAE, and MDAE are the root mean squared error, mean absolute error and median absolute error of the estimates.

Table 2. Coefficient Estimates

	β_2		β_1	β_3	
	True Effect	Mismeasured Effect		True Effect	Mismeasured Effect
black	0.0448948	−0.1207730	−1.0353728	−0.1833281	−0.7087732
	(0.20746339)	(0.06393301)	(0.41059487)	(0.37992196)	(0.14792043)
income	**0.4724051**	**0.1426148**	**0.5633254**	**−0.1159633**	**−0.0836258**
	(0.12498693)	(0.03283066)	(0.18986917)	(0.26671178)	(0.05702173)
phsrank	**0.0155228**	**0.0112585**	**−0.0082122**	**0.0032316**	**0.0086985**
	(0.00296229)	(0.00084848)	(0.00612383)	(0.00687502)	(0.00140828)
stotal	0.1424727	0.3153344	−0.9172876	0.3225476	0.3162301
	(0.11269463)	(0.02600598)	(0.1838158)	(0.24166599)	(0.04188189)
totexp	**−0.0099431**	**−0.0052219**	**−0.0051576**	**−0.0146081**	**−0.0126890**
	(0.00226798)	(0.00050814)	(0.00282731)	(0.00752229)	(0.00090517)
female	0.5196376	−0.4260637	2.7281916	0.4472213	−0.4118640
	(0.28698793)	(0.03898326)	(1.3915567)	(0.45181952)	(0.06482936)
large city	−0.1663922	0.1084376	−0.6641598	−0.5281193	−0.4202254
	(0.17369701)	(0.05448494)	(0.40283449)	(0.23284787)	(0.08174698)
parttime	**0.9767346**	**0.5343357**	**0.6336924**	**2.4864880**	**1.8791449**
	(0.34799942)	(0.05293099)	(0.24031441)	(1.7755474)	(0.09464261)

Table 3. Function Estimates

	Mean		Standard Deviation	
	True Effect	Mismeasured Effect	True Effect	Mismeasured Effect
$r_0^*(z)$	0.4406517	0.3409593	0.2701055	0.1867520
$r_1^*(z)$	0.4182481	0.3263341	0.2687167	0.1838969
$r_2^*(z)$	0.3881061	0.2727147	0.2656879	0.1709973
$r_3^*(z)$	0.3796998	0.2820059	0.2646002	0.1735172
$r_4^*(z)$	0.4071389	0.3064211	0.2677571	0.1795623
$b_1(z)$	0.3136727		0.3392644	
$\tau^*(z)$	0.3787436	0.1108377	1.215702	0.9542675