

Estimation of Average Treatment Effects With Misclassification

Arthur Lewbel*
Boston College

May, 2003

Abstract

This paper provides conditions for identification, and an associated estimator, of the average effect of a binary treatment or policy on a scalar outcome in models where treatment may be misclassified. Misclassification probabilities and the true probability of treatment are also identified.

Misclassification occurs when treatment is measured with error, that is, some units are reported to have received treatment when they actually have not, and vice versa. Conditional outcomes, treatment probabilities, and misclassification probabilities are all nonparametric. The identifying assumption is an exclusion restriction, specifically, the existence of a variable that can take on at least three different values, affects the decision to treat, and is conditionally independent of the conditional misclassification probabilities and the average treatment effect.

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

This research was supported in part by the National Science Foundation through grant SES-9905010. The author wishes to thank Alberto Abadie for many helpful comments. Any errors are my own.

*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>

1 Introduction

This paper provides conditions for identification, and an associated estimator, of the average effect (conditioned on covariates) of a binary treatment, program, or policy on a scalar outcome in models where treatment may be misclassified. Misclassification occurs when treatment is measured with error, that is, some units are reported to have received treatment when they actually have not, and vice versa. The assumptions provided also identify conditional on covariates misclassification probabilities and the true probability of treatment, in addition to identifying the conditional average treatment effect.

Treatment is defined to be exogenous or unconfounded (also known as selection on observables) if the decision to treat or to enroll in a program is independent of potential outcomes conditional on covariates. Assuming unconfoundedness, when treatment is observed without error the average treatment effect can be estimated by matching, differencing within subpopulation averages of treated and untreated units, or by propensity score methods. Relevant models and estimators include Heckman (1974, 1976), Rubin (1974), Heckman and Robb (1985), Rosenbaum and Rubin (1985), Manski (1990, 1997), Robins, Mark, and Newey (1992), Angrist, Imbens, and Rubin (1996), Heckman, Ichimura, and Todd (1998), Hahn (1998), Abadie and Imbens (2002), and Hirano, Imbens, and Ridder (2002).

Assume that treatment is unconfounded, but the researcher has an imprecise measure of treatment, so some individuals who were actually treated are recorded as untreated and vice versa. For example, if treatment is schooling, some respondents may either lie or not know if the particular training or schooling they've had counts as higher education. In a medical context, some patients may fail to follow a therapy regimen that is assigned to them. In a study of pension plans, Gustman and Steinmeier (2003, table 6c) and Molinari (2002, table 5) found that 15% of respondents that actually had a defined benefit plan claimed to have a defined contribution plan, and 26% that actually had a defined contribution plan claimed to have a defined benefit plan. In this example, an analysis of treatment (plan type) on outcome (e.g., retirement income) would suffer from substantial misclassification bias if respondent's data on plan type were used.

Many structural treatment models, in particular some parametric or semiparametric latent variable selection models, violate unconfoundedness. See, Lewbel (2002) for an example, and more generally Vytlačil (2002) and Heckman and Vytlačil (2001) for relationships between latent variable selection models and treatment effect estimators. Apparent violations of unconfoundedness, or of other conditions used to identify and estimate treatment effects, could be due to misclassification, i.e., potential outcomes that are unconfounded with respect to true treatment could violate unconfoundedness with respect to treatments observed with error. For example, the assignment of individuals to either treatment or no treatment could

be completely determined by some unknown function of observables, but because of misclassification there will be apparent randomness in the assignments.

In linear outcome models, if observed treatment satisfied classical measurement error, then ordinary two stage least squares methods could be used. However, binary regressors cannot satisfy classical measurement error assumptions. Alternatives to two stage least squares for linear outcome models with mismeasured binary regressors include Card (1996) and Kane, Rouse, and Staiger (1999).

Choice or assignment of treatment can be interpreted as a binary choice or binomial response model. Examples of recent papers that consider estimation of binomial response misclassification model parameters or misclassification probabilities include McFadden (1984), Chua and Fuller (1987), Brown and Light (1992), Poterba and Summers (1995), Abrevaya and Hausman (1999), Hausman, Abrevaya, and Scott-Morton (1998), and Lewbel (2000). In binomial response models a distinction is made between misclassification that can happen to any unit with some probability, versus the case where some respondents (which are unknown to the researcher) are always correctly measured while others provide "natural responses," e.g., always claiming to be treated or untreated regardless of the truth. See, e.g., Finney (1964). For the purposes of the present paper the distinction between these two types of misclassification is irrelevant; they will be observationally equivalent.

For treatment effects, misclassification is closely related to the problem of identification in the presence of imperfect compliance in an otherwise randomized experiment. See, e.g., Angrist, Imbens and Rubin (1996) and Balke and Pearl (1997). Also related are cases where an instrument used for identification is imperfect, as in Hotz, Mullin, and Sanders (1997), or when either covariates, treatment, or outcomes are not observed for some subjects, as in Robins (1997), Horowitz and Manski (2000), and Molinari (2002).

This paper provides a set of minimal conditions for identification of probability of treatment, misclassification probabilities, and of average treatment effects when treatment may be mismeasured. An estimator that employs these identification conditions is provided. The identifying condition is the existence of an instrument, i.e., a scalar or vector of variables that can take on at least three different values, affects the decision to treat, and does not effect the conditional misclassification probabilities or the average treatment effect.

This source of identification is an example of an exclusion restriction, that is, a variable that affects some relevant functions and not others. Exclusion restrictions are a common method of obtaining identification in econometric models. See, e.g., Powell's (1994, section 2.5) survey. However, what is perhaps peculiar or surprising about this result is that a collection of potentially high dimensional unknown functions are

identified even though the instrument can be discrete, with as few as three mass points. A related result is Abadie (2003), in which a binary instrument that affects treatment decisions in specific ways is considered.

2 Identification

Let Y be an observed outcome, T^* index the actual, unobserved treatment, and T index the reported treatment. The possible treatments are $t = 1$ corresponding to being treated or enrolling in a program, and $t = 0$ for no treatment. Let $Y(t)$ denote the outcome from treatment $T^* = t$. Let X be a vector of covariates. The goal is estimation of the conditional average treatment effect $E[Y(1) - Y(0) | X = x]$. Define

$$\tau^*(x) = E(Y | X = x, T^* = 1) - E(Y | X = x, T^* = 0) \quad (1)$$

ASSUMPTION A1: $E[Y(t) | T^*, X] = E[Y(t) | X]$

Assumption A1 is the conditional mean weakening of the standard unconfoundedness assumption, which is with respect to the true treatment T^* . Heckman, Ichimura, and Todd (1998) show that this version of unconfoundedness implies that the conditional average treatment effect satisfies

$$E[Y(1) - Y(0) | X = x] = \tau^*(x)$$

If T^* were observed without error, then equation (1) would provide an estimator for $\tau^*(x)$, by replacing expectations with nonparametric regressions. Other estimators, e.g. those based on the propensity score, would also be possible given unconfoundedness, as noted in the introduction.

ASSUMPTION A2: $E(Y | X, T^*, T) = E(Y | X, T^*)$.

Assumption A2 says that, conditional on X and on the actual treatment T^* , the measurement of treatment does not affect the expected outcome. This is analogous to the classical measurement error assumption that actual outcomes be independent of any measurement errors made by the researcher. This could be a substantive assumption if the misclassification is due to a misperception on the part of the subject, for example, if T indicates the treatment that the subject thinks he or she had, then Assumption A2 would rule out placebo effects.

Make the following definitions.

$$r^*(x) = E(T^* | X = x)$$

$$b_t(x) = E[I(T = 1 - t) | X = x, T^* = t] = \Pr(T = 1 - t | X = x, T^* = t)$$

Note that, conditioning on $X = x$, the function $r^*(x)$ is the probability of receiving treatment, while $b_1(x)$ is the probability of misclassifying the treated and $b_0(x)$ is the probability of misclassifying the untreated.

ASSUMPTION A3: $b_0(x) + b_1(x) < 1$, $E(T^* | X = x, T = 1] \neq E(T^* | X = x, T = 0]$, and $0 < r^*(x) < 1$ for all $x \in \text{supp}(X)$.

Assumption A3 says first that the sum of misclassification probabilities is less than one, so on average our observations of T are more accurate than pure guesses. In a binomial response model with misclassification, this assumption is what Hausman, Abrevaya, and Scott-Morton (1998) call the monotonicity condition. Given failure to observe T^* , 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 treatment effect or the negative of the treatment effect. Similarly, the second condition of Assumption A3 says that T provides some information beyond what x contains regarding the probability of treatment. Assumption A3 also requires that for any 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.

Define the following functions.

$$r(x) = E(T | X = x)$$

$$\tau(x) = E(Y | X = x, T = 1) - E(Y | X = x, T = 0)$$

ASSUMPTION A4: Assume $r(x)$ and $\tau(x)$ are identified.

The functions $r(x)$ and $\tau(x)$ are conditional expectations of observable data, so Assumption A4 will hold given any data set that permits consistent estimation of these conditional expectations. If X is discretely distributed, then only consistency of sample averages is required.

Note that $r(x)$ and $\tau(x)$ are the same as $r^*(x)$ and $\tau^*(x)$, except defined in terms of the observed treatment 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, by Assumption A1, $\tau(x)$ would equal the conditional average treatment effect

Define

$$m(x) = \left(\frac{1}{1 - b_1(x) - b_0(x)} \right) \left(1 - \frac{[1 - b_1(x)]b_0(x)}{r(x)} - \frac{[1 - b_0(x)]b_1(x)}{1 - r(x)} \right). \quad (2)$$

THEOREM 1: Let Assumptions A1, A2, A3, and A4 hold. Then

$$r^*(x) = \frac{r(x) - b_0(x)}{1 - b_0(x) - b_1(x)}, \quad (3)$$

$$\tau(x) = \tau^*(x)m(x), \quad (4)$$

and if $b_0(x)$ and $b_1(x)$ are identified, then the probability of treatment $r^*(x)$ and conditional average treatment effect $E[Y(1) - Y(0) | X = x]$ are identified.

Theorem 1 shows that if the misclassification probabilities $b_t(x)$ are known to the researcher or can be identified (for example from a validation sample), then the true conditional average treatment effect $\tau^*(x)$ is identified. Results similar to those of Theorem 1 have been used to construct bounds on treatment effects. See, e.g., Hotz, Mullin, and Sanders (1997).

Now consider identification of the misclassification probabilities $b_t(x)$ without outside data. An interesting implication of Theorem 1 is that the true conditional average treatment effect $\tau^*(x)$ equals the estimated treatment effect $\tau(x)$ divided by a function $m(x)$ where $m(x)$ is determined entirely by treatment and classification probabilities, not outcomes. This fact is exploited to obtain identification in Theorem 2 below.

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

ASSUMPTION A5: For some set $\Omega \subset \text{supp}(V)$, for all $v \in \Omega$, $v_0 \in \Omega$, and $z \in \text{supp}(Z)$, we have $b_t(v, z) = b_t(v_0, z)$, $\tau^*(v, z) = \tau^*(v_0, z)$, and $r^*(v, z) \neq r^*(v_0, z)$.

In a small abuse of notation, let $b_t(z)$ and $\tau^*(z)$ denote $b_t(v, z)$ and $\tau^*(v, z)$, respectively, for $v \in \Omega$. The distribution of V can be discrete; V could be a scalar that only takes on a few different values. Assumption A5 says that there exists a variable V that affects r^* , and hence the true treatment probabilities, but after

conditioning on other covariates Z does not affect either the measurement errors b_t or the conditional average treatment effect τ^* (at least for some values that V might take on).

Having a V that doesn't affect the misclassification probability is a commonly employed assumption for identification in binomial response models with misclassification. See, e.g., Abrevaya and Hausman (1999), Hausman, Abrevaya, and Scott-Morton (1998), 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 A5.

Having V affect r^* but not τ^* is a weaker version of the type of exclusion of assumption that is commonly used in the identification of selection models. For a job training program, an example of V might be nonwage related income or benefits, or more generally any variable that, after conditioning on other covariates, does not affect the average effectiveness of the program but is correlated with eligibility or selection, such as distance to the school as employed by Card (1995) and others. Another possibility is that V could be a second mismeasured observation of T^* , with an independent source of classification error, as in Kane, Rouse, and Staiger (1999).

ASSUMPTION A6: There exists three elements $v_k \in \Omega$, $k = 0, 1, 2$, 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 A6 is that V can take on at least three values. The required inequality in Assumption A6 will only fail to hold if $\tau(v_0, z) = 0$, or if $r(v_0, z) = r(v_1, z) = r(v_2, z)$, or if a complicated equality relationship holds amongst the three conditional outcomes and conditional treatment probabilities, which would require a perfect coincidence between probabilities and outcomes. It is possible to directly test this assumption, because these $\tau(v_k, z)$ and $r(v_k, z)$ functions are conditional expectations of observable data, and so can be directly estimated (they are identified by Assumption A4). Finally, note that if V can take on more than three values, then Assumption A6 will hold as long as there exists any one triplet of V values that satisfies the necessary inequality.

THEOREM 2: Let Assumptions A1, A2, A3, A4, A5, and A6 hold. Then the conditional misclassification probabilities $b_0(x)$ and $b_1(x)$, the conditional probability of treatment $r^*(x)$, and the conditional average treatment effect $E[Y(1) - Y(0) | X = x]$ are all identified.

It follows from Theorem 1 that $\tau(v_k, z)m(v_0, z) = \tau(v_0, z)m(v_k, z)$. Substituting equation (2) into this expression yields an equation that depends only on the identified functions τ and r and on the two unknowns b_0 and b_1 . Evaluating this expression for $k = 1$ and $k = 2$ gives two equations in the two unknowns. These equations are nonlinear, but the proof of Theorem 2 shows that these equations still uniquely define and thereby identify b_0 and b_1 . Each value that V can take on provides another equation that b_0 and b_1 must satisfy, so in general the larger is the set Ω of values that V can take on (which satisfy Assumption A5), the greater will be the number of overidentifying restrictions determining $b_0(z)$ and $b_1(z)$.

Here V was required to take on at least three different values. These results may be immediately extended to show that only a binary V would be required 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, or that these probabilities are equal to each other.

3 Estimation

For simplicity, assume that the distribution of X is discrete, and in particular that the number of observations of each element of $\text{supp}(X)$ goes to infinity with the sample size. Also for simplicity let $\Omega = \text{supp}(V) = \{v_0, \dots, v_K\}$. To estimate $r^*(v, z)$, $b_0(z)$, $b_1(z)$, and $\tau^*(z)$ for any given z , restrict attention to the set of $n(z)$ observations Y_i, T_i, V_i having $Z_i = z$. All further mention of z may now be dropped, with the understanding that all defined and estimated parameters are implicitly functions of z .

Define

$$a_{tk} = 1/E[I(T = t, V = v_k)],$$

$$\tau_k = E [[TYa_{1k} + (1 - T)Ya_{0k}]I(V = v_k)]$$

$$q(b_0, b_1, a_{0k}, a_{1k}) = 1 + (b_1 - 1)b_0 \left(1 + \frac{a_{1k}}{a_{0k}}\right) + (b_0 - 1)b_1 \left(1 + \frac{a_{0k}}{a_{1k}}\right).$$

COROLLARY 1: If Assumptions A1, A2, A3, A4, A5, and A6 hold then

$$q(b_0, b_1, a_{0k}, a_{1k})\tau_0 = q(b_0, b_1, a_{00}, a_{10})\tau_k, \tag{5}$$

$$r^*(v_k) = \frac{a_{0k}}{a_{1k} + a_{0k}}, \tag{6}$$

and

$$\tau^* = \frac{(1 - b_1 - b_0)}{q(b_0, b_1, a_{00}, a_{10})} \tau_0. \quad (7)$$

Note that equation (5) comes from $m(v)\tau(v_0) = m(v_0)\tau(v) = 0$, which is the equation that provides generic identification of b_0, b_1 in Theorem 2.

Define

$$\theta = (a_{00}, \dots, a_{0K}, a_{10}, \dots, a_{1,K-1}, \tau_0, b_0, b_1)'$$

and define $g(\theta, Y, T, V)$ as the vector consisting of the following $3K + 2$ functions

$$I(T = t, V = v_k) - (1/a_{tk}), \quad t = 0, 1; \quad k = 1, \dots, K - 1$$

$$I(T = 0, V = v_0) - (1/a_{0K})$$

$$[TYa_{10} + (1 - T)Ya_{00}]I(V = v_0) - \tau_0$$

$$q(b_0, b_1, a_{0k}, a_{1k})\tau_0 - q(b_0, b_1, a_{00}, a_{10})[TYa_{1k} + (1 - T)Ya_{0k}]I(V = v_k), \quad k = 1, \dots, K - 1$$

$$q\left(b_0, b_1, a_{0K}, \left(1 - \sum_{t=0}^1 \sum_{k=0}^{K-1} 1/a_{tk} + 1/a_{0K}\right)^{-1}\right)\tau_0 -$$

$$q(b_0, b_1, a_{00}, a_{10})\left(TY\left(1 - \sum_{t=0}^1 \sum_{k=0}^{K-1} 1/a_{tk} + 1/a_{0K}\right)^{-1} + (1 - T)Ya_{0K}\right)I(V = v_k)$$

COROLLARY 2: If Assumptions A1, A2, A3, A4, A5, and A6 hold then $E[g(Y, T, V, \theta)] = 0$

Now $E[g(Y, T, V, \theta)] = 0$. This is a total of $3K + 2$ moment conditions, and θ has $2K + 4$ elements, so assuming identification (which by Theorem 2 holds for $K \geq 2$), we may apply Hansen's (1982) Generalized Method of Moments (GMM) to obtain a consistent, asymptotically normal estimate $\hat{\theta}$ of θ at rate root n under standard conditions. In particular, if we have n independently, identically distributed draws Y_i, T_i, V_i , efficient GMM gives

$$\sqrt{n}(\hat{\theta} - \theta) \rightarrow^d N(0, (S'W^{-1}S)^{-1})$$

where $S = E[\partial g(Y, T, V, \theta)/\partial \theta']$ and $W = E[g(Y, T, V, \theta)g(Y, T, V, \theta)']$. A technicality is that GMM assumes parameters have compact support. This could be imposed, consistent with (and slightly strengthening) Assumption A3 by assuming that $\delta \leq 1/a_{jk} \leq 1 - \delta$, $\delta \leq b_t$, and $b_0 + b_1 \leq 1 - \delta$ for some small $\delta > 0$.

Equation (7) defines $\tau^*(\theta)$, so the estimate of the conditional average treatment effect is $\tau^*(\hat{\theta})$. Applying the delta method yields the limiting distribution

$$\sqrt{n}[\tau^*(\hat{\theta}) - \tau(\theta)] \rightarrow^d N\left(0, \frac{\partial \tau(\theta)'}{\partial \theta} (S'W^{-1}S)^{-1} \frac{\partial \tau(\theta)}{\partial \theta}\right)$$

Similarly, the estimate of the probability of treatment is $r^*(\hat{\theta})$ as defined by equation (6), and the estimated misclassification probabilities are \hat{b}_0 and \hat{b}_1 .

4 Conclusions

This paper provides a set of minimal conditions for identification of treatment probabilities, misclassification probabilities, and average treatment effects when treatment may be mismeasured. An estimator that employs these identification conditions is provided, based on direct estimation of relevant conditional expectations. It would be useful to explore whether other treatment effect estimators such as matching and propensity score based methods could be adapted to the present application where treatment may be mismeasured.

References

- [1] ABADIE, A. (2003), "Semiparametric Instrumental Variable Estimation of Treatment Response Models," *Journal of Econometrics*, 113, 231-263.
- [2] ABADIE, A. AND G. IMBENS, (2002), "Simple and Bias Corrected Matching Estimators for Average Treatment Effects," NBER working paper.
- [3] ABREVVAYA, 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.

- [4] ANGRIST, J., G. IMBENS, AND D. B. RUBIN, (1996), "Identificaiton of Causal Effects Using Instrumental Variables," *Journal of the American Statistical Association*, 91, 444-455.
- [5] BALKE, A. AND J. PEARL, (1997), "Bounds on Treatment Effects From Studies With Imperfect Compliance," *Journal of the American Statistical Association*, 92, 1171-1176.
- [6] BROWN, J. N., AND A. LIGHT, (1992), "Interpreting Panel Data on Job Tenure," *Journal of Labor Economics*," 10, 219-257.
- [7] 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.
- [8] CARD, D. (1996), "The Effect of Unions on the Structure of Wages: A Longitudinal Analysis," *Econometrica*, 64, 957-979.
- [9] CHUA, T. C. AND W. A FULLER, (1987) A Model For Multinomial Response Error Applied to Labor Flows," *Journal of the American Statistical Association*, 82, 46-51.
- [10] FINNEY, D. J. (1964) *Statistical Method in Biological Assay*. Havner: New York.
- [11] GUSTMAN, A. L. AND T. L. STEINMEIER, (2003), "What People Don't Know About Their Pension and Social Security," in Gale, Shoven, and Warshawsky, eds., *The Evolving Pension System: Trends, Effects, and Proposals for Reform*, Washington: Brookings Institution Press.
- [12] HAHN, J., (1998), "On the Role of Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects," *Econometrica*, 66, 315-331.
- [13] HANSEN, L., (1982), "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, 50, 1029-1054.
- [14] 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.
- [15] HECKMAN, J. (1974), "Shadow Prices, Market Wages, and Labor Supply," *Econometrica*, 42, 679-693.

- [16] HECKMAN, J. (1976), "Sample Selection Bias as a Specification Error," *Econometrica*, 47, 153-161.
- [17] HECKMAN, J. H. ICHIMURA AND P. TODD, (1998), "Matching as an Econometric Evaluations Estimator," *Review of Economic Studies*, 65, 261-294.
- [18] HECKMAN, J. AND R. ROBB, (1985), "Alternate Methods for Evaluating the Impact of Interventions," in J. Heckman and B. Singer, eds., *Longitudinal Analysis of Labor Market Data*, New York: Cambridge University Press.
- [19] HECKMAN, J. AND E. VYTLACIL, (2001), "Structural Equations, Treatment Effects and Econometric Policy Evaluation," unpublished manuscript.
- [20] HIRANO, K., G. IMBENS, AND G. RIDDER, (2002), "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," Unpublished manuscript.
- [21] HOROWITZ, J. L. AND C. F. MANSKI, (2000), "Nonparametric Analysis of Randomized Experiments With Missing Covariate and Outcome Data," *Journal of the American Statistical Association*, 95, 77-84.
- [22] KANE, T. J., C. E. ROUSE, AND D. STAIGER, (1999), "Estimating Returns to Schooling When Schooling is Misreported" NBER working paper #7235.
- [23] LEWBEL, A., (2000), "Identification of the Binary Choice Model With Misclassification," *Econometric Theory*, 16, 603-609.
- [24] LEWBEL, A., (2002), "Endogeneous Selection or Treatment Model Estimation," Unpublished manuscript.
- [25] MANSKI, C. F. (1990) "Nonparametric Bounds on Treatment Effects," *American Economic Review Papers and Proceedings*, 80, 319-323.
- [26] MANSKI, C. F. (1997) "Monotone Treatment Response," *Econometrica*, 65, 1311-1334.
- [27] MCFADDEN, D., (1984), "Econometric Analysis of Qualitative Response Models," In: Griliches, Z., Intriligator, M.D.(Eds.), *Handbook of Econometrics*, vol. 2. North-Holland, Amsterdam.

- [28] MOLINARI, F. (2001) "Identification of Probability Distributions With Misclassified Data" Unpublished manuscript, Northwestern University.
- [29] MOLINARI, F. (2002) "Missing Treatments" Unpublished manuscript, Northwestern University.
- [30] POTERBA, J. M. AND L. H. SUMMERS (1995) "Unemployment Benefits and Labor Market Transitions: A Multinomial Logit Model With Errors in Classification," *Review of Economics and Statistics*, 77, 207-216.
- [31] POWELL, J. L., (1994), "Estimation of Semiparametric Models," in Handbook of Econometrics, vol. iv, ed. by R. F. Engle and D. L. McFadden, pp. 2444-2521, Amsterdam: Elsevier.
- [32] ROBINS, J., (1997), "Non-Response Models for the Analysis of Non-Monotone Non-Ignorable Missing Data," *Statistics in Medicine*, 16, 21-37.
- [33] ROBINS, J., S. MARK AND W. NEWEY, (1992), "Estimating Exposure Effects by Modeling the Expectations of Exposure Conditional on Confounders," *Biometrics*, 48, 479-495.
- [34] ROSENBAUM, P. AND D. RUBIN, (1985), "Reducing Bias in Observational Studies Using Subclassification on the Propensity Score," *Journal of the American Statistical Association*, 79, 516-524.
- [35] RUBIN, D. (1974), "Estimating Causal Effects of Treatments in Randomized and Non-Randomized Studies," *Journal of Educational Psychology*, 76, 688-701.
- [36] VYTLACIL, E. (2002), "Independence, Monotonicity, and Latent Index Models: An Equivalence Result," *Econometrica*.

5 Appendix

PROOF OF THEOREM 1: Recall $r^*(x) = E(T^* | X = x)$.

By the definition of $r(x)$,

$$\begin{aligned} r(x) &= \Pr(T = 1 | X = x, T^* = 1) \Pr(T^* = 1 | X = x) + \Pr(T = 1 | X = x, T^* = 0) \Pr(T^* = 0 | X = x) \\ &= [1 - b_1(x)]r^*(x) + b_0(x)[1 - r^*(x)] \end{aligned}$$

So $r^*(x) = [r(x) - b_0(x)]/[1 - b_0(x) - b_1(x)]$.

Next, define

$$\begin{aligned} p_t(x) &= E[I(T^* = t) | X = x, T = t] \\ r_t^*(x) &= E[I(T^* = t) | X = x] \end{aligned}$$

so $r_1^*(x) = r^*(x)$ and $r_0^*(x) = 1 - r^*(x)$. By Bayes rule

$$\begin{aligned} p_t(x) &= \frac{\Pr(T = t | X = x, T^* = t) \Pr(T^* = t | X = x)}{\Pr(T = t | X = x)} \\ &= \frac{\Pr(T = t | X = x, T^* = t) \Pr(T^* = t | X = x)}{\Pr(T = t | X = x, T^* = t) \Pr(T^* = t | X = x) + \Pr(T = t | X = x, T^* = 1 - t) \Pr(T^* = 1 - t | X = x)} \\ &= \frac{[1 - b_t(x)]r_t^*(x)}{[1 - b_t(x)]r_t^*(x) + b_{1-t}(x)[1 - r_t^*(x)]} = \frac{[1 - b_t(x)]r_t^*(x)}{[1 - b_t(x) - b_{1-t}(x)]r_t^*(x) + b_{1-t}(x)} \end{aligned}$$

So

$$\begin{aligned} p_1(x) &= \frac{[1 - b_1(x)]r^*(x)}{[1 - b_1(x) - b_0(x)]r^*(x) + b_0(x)} \\ &= \frac{[1 - b_1(x)][r(x) - b_0(x)]}{[1 - b_1(x) - b_0(x)]r(x)} \\ p_0(x) &= \frac{[1 - b_0(x)][1 - r^*(x)]}{[1 - b_0(x) - b_1(x)][1 - r^*(x)] + b_1(x)} \\ &= \frac{[1 - b_0(x)][1 - r(x) - b_1(x)]}{[1 - b_1(x) - b_0(x)][1 - r(x)]} \end{aligned}$$

Now define $h_t(x) = E(Y | X = x, T^* = t)$. By Assumption A2, $h_t(x) = E(Y | X = x, T^* = t, T)$ and so

$$\begin{aligned} E(Y | X = x, T = t) &= \sum_{j=0}^1 E(Y | X = x, T = t, T^* = j) \Pr(T^* = j | X = x, T = t) \\ &= h_t(x)p_t(x) + h_{1-t}(x)[1 - p_t(x)] \end{aligned}$$

Substituting this expression into the definition of $\tau(x)$ gives

$$\begin{aligned} \tau(x) &= h_1(x)p_1(x) + h_0(x)(1 - p_1(x)) - [h_0(x)p_0(x) + h_1(x)(1 - p_0(x))] \\ &= [h_1(x) - h_0(x)][p_1(x) + p_0(x) - 1] \\ &= \tau^*(x)[p_1(x) + p_0(x) - 1] \end{aligned}$$

so $\tau(x) = \tau^*(x)m(x)$ where

$$\begin{aligned} m(x) &= [p_1(x) + p_0(x) - 1] \\ &= \frac{[1 - b_1(x)][r(x) - b_0(x)]}{[1 - b_1(x) - b_0(x)]r(x)} + \frac{[1 - b_0(x)][1 - r(x) - b_1(x)]}{[1 - b_1(x) - b_0(x)][1 - r(x)]} - 1 \\ &= \left(\frac{1}{1 - b_1(x) - b_0(x)} \right) \left(1 - \frac{[1 - b_1(x)]b_0(x)}{r(x)} - \frac{[1 - b_0(x)]b_1(x)}{1 - r(x)} \right). \end{aligned}$$

For identification, $E[Y(1) - Y(0) \mid X = x] = \tau^*(x)$ by Assumption A1, and $\tau^*(x) = \tau(x)/m(x)$, where $m(x)$ is identified by $b_0(x)$, $b_1(x)$, and $r(x)$. Assumption A3 ensures that $m(x)$ is finite and nonzero.

PROOF OF THEOREM 2: For a given z , we have for all $v \in \Omega$, using Theorem 1,

$$\frac{\tau(v, z)}{\tau(v_0, z)} = \frac{m(v, z)}{m(v_0, z)}$$

To ease notation further, drop z . Then $m(v)\tau(v_0) - m(v_0)\tau(v) = 0$ and substituting in for m gives

$$0 = \left(1 + \frac{(b_1 - 1)b_0}{r(v)} + \frac{(b_0 - 1)b_1}{1 - r(v)} \right) \tau(v_0) - \left(1 + \frac{(b_1 - 1)b_0}{r(v_0)} + \frac{(b_0 - 1)b_1}{1 - r(v_0)} \right) \tau(v) \quad (8)$$

$$0 = (1 - b_1)b_0 \left(\frac{\tau(v_0)}{r(v)} - \frac{\tau(v)}{r(v_0)} \right) + (1 - b_0)b_1 \left(\frac{\tau(v_0)}{1 - r(v)} - \frac{\tau(v)}{1 - r(v_0)} \right) + \tau(v) - \tau(v_0)$$

Evaluate this equation at $v = v_k$, and rewrite it 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 $r(v_0)$, $r(v_k)$, $\tau(v_0)$, and $\tau(v_k)$. Given that Ω 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 inequality in Assumption 6 makes the determinant of this matrix nonzero, as required. Let $s = 1 + B_1 - B_0$. Given B_0 and B_1 , the equations $B_t = (1 - b_{1-t})b_t$ imply

$$0 = b_1^2 - s b_1 + B_1$$

$$b_0 = 1 + b_1 - s$$

This pair of equations has two possible solutions

$$b_1 = [s \pm (s^2 - 4B_1)^{1/2}]/2$$

$$b_0 = 1 - [s \mp (s^2 - 4B_1)^{1/2}]/2$$

Summing these two equations gives

$$b_0 + b_1 = 1 \pm (s^2 - 4B)^{1/2}$$

so the restriction that $b_0 + b_1 < 1$ means that b_1 is uniquely determined by the negative root, and so b_0 and b_1 are uniquely determined. Then, by applying Theorem 1, the conditional average treatment effect $E[Y(1) - Y(0) | X = v, z]$ is identified.

PROOF OF COROLLARY 1: By the definitions of r and τ ,

$$\frac{1}{r(v_k)} = 1 + \frac{a_{1k}}{a_{0k}}, \quad \frac{1}{1 - r(v_k)} = 1 + \frac{a_{0k}}{a_{1k}} \quad (9)$$

$$\tau(v_k) = E[(Ta_{1k} + (1 - T)a_{0k})I(V = v_k)Y]. \quad (10)$$

Substitute equations (9) and (10) into equation (8) with $v = v_k$ to get

$$\begin{aligned} 0 = & \left[1 + (b_1 - 1)b_0 \left(1 + \frac{a_{1k}}{a_{0k}} \right) + (b_0 - 1)b_1 \left(1 + \frac{a_{0k}}{a_{1k}} \right) \right] E[(Ta_{10} + (1 - T)a_{00})I(V = v_0)Y] \\ & - \left[1 + (b_1 - 1)b_0 \left(1 + \frac{a_{10}}{a_{00}} \right) + (b_0 - 1)b_1 \left(1 + \frac{a_{00}}{a_{10}} \right) \right] E[(Ta_{1k} + (1 - T)a_{0k})I(V = v_k)Y] \end{aligned}$$

which gives equation (5). Next, from Theorem 1, $\tau^* = \tau(v_0)/m(v_0) = \tau(v_0)(1 - b_1 - b_0)/q_0(\theta)$, which gives equation (7).

PROOF OF COROLLARY 2 It follows from Corollary 1 and the definitions of a_{tk} , τ_k , and q that

$$E[I(T = t, V = v_k) - (1/a_{tk})] = 0, \quad t = 0, 1; \quad k = 0, \dots, K$$

$$E[\tau_0 - [TYa_{10} + (1 - T)Ya_{00}]I(V = v_0)] = 0$$

$$E[q(b_0, b_1, a_{0k}, a_{1k})\tau_0 - q(b_0, b_1, a_{00}, a_{10})[TYa_{1k} + (1 - T)Ya_{0k}]I(V = v_k)] = 0, \quad k = 1, \dots, K$$

Also, by construction $\sum_{t=0}^1 \sum_{k=0}^K 1/a_{tk} = 1$. The corollary follows after using this equality to substitute out for a_{1K} .