Econometric Theory, **16**, 2000, 603–609. Printed in the United States of America.

IDENTIFICATION OF THE BINARY CHOICE MODEL WITH MISCLASSIFICATION

Arthur Lewbel
Boston College

Misclassification in binary choice (binomial response) models occurs when the dependent variable is measured with error, that is, when an actual "one" response is sometimes recorded as a zero and vice versa. This paper shows that binary response models with misclassification are semiparametrically identified, even when the probabilities of misclassification depend in unknown ways on model covariates and the distribution of the errors is unknown.

1. INTRODUCTION

This paper shows that binary response models with misclassification of the dependent variable are semiparametrically identified, even when the probabilities of misclassification depend in unknown ways on model covariates and the distribution of the errors is unknown.

Let x_i be a vector of covariates that may affect both the response of observation i and the probability that the response is observed incorrectly. For identification, assume there exists a covariate v_i that affects the true response but does not affect the probability of misclassification. If more than one such covariate exists, let v_i be any one of the available candidates (that satisfies the regularity conditions listed subsequently), and the others can without loss of generality be included in the vector x_i.

Let y_i^* be an unobserved latent variable associated with observation i, given by

$$y_i^* = v_i \gamma + x_i \beta + e_i,$$

where the e_i are independently and identically distributed errors. The true response is given by

$$\tilde{y}_i = I(y_i^* \geq 0),$$

This research was supported in part by the National Science Foundation through grant SBR-9514977. I thank the associate editor and two referees for helpful comments and suggestions. Any errors are my own. Address correspondence to: Arthur Lewbel, Department of Economics, Boston College, 140 Commonwealth Ave., Chestnut Hill, MA 02467, USA; e-mail: lewbel@bc.edu.

 0266-4666/00 $9.50

https://doi.org/10.1017/S0266466600164060 Published online by Cambridge University Press

where $I(\cdot)$ equals one if $\cdot$ is true and zero otherwise. When $\tilde{y}_i$ is observed, this is the standard latent variable specification of the binary response model (see, e.g., McFadden, 1984).

Now permit the true response (i.e., classification of observation i) to be observed with error. Letting y_i denote the observed binary dependent variable, the misclassification probabilities are

$$a(x_i) = \Pr(y_i = 1 | \tilde{y}_i = 0, x_i),$$

$$a^*(x_i) = \Pr(y_i = 0 | \tilde{y}_i = 1, x_i).$$

So $a(x_i)$ is the probability that an actual zero response is misclassified (i.e., incorrectly recorded) as a one, and $a^*(x_i)$ is the probability that a one response is misclassified as a zero. These misclassification probabilities are permitted to depend in an unknown way on observed covariates x_i. This framework encompasses models where misclassification probabilities may also depend on variables that do not affect the true response, because any covariate x_{ji} that affects a or a^* but not y^* is just a covariate that has a coefficient β_j that equals zero.

Define $b(x)$ as

$$b(x_i) = [1 - a(x_i) - a^*(x_i)]$$

and define the function g to be the conditional expectation of y, which in this model is

$$g(v_i, x_i) = E(y_i | v_i, x_i) = a(x_i) + b(x_i) F(v_i \gamma + x_i \beta), \tag{1}$$

where F is the cumulative distribution function of the random variable $-e$.

Another model that corresponds to equation (1) is when a fraction $a(x)$ of respondents having characteristics x always answers one, a fraction $a^*(x)$ always answers zero, and the remainder respond with $I(v\gamma + x\beta + e \geq 0)$. In this interpretation some respondents give "natural responses" that are due to factors other than the latent variable, whereas the other respondents follow the latent variable model. Although this model is observationally equivalent to the misclassification model, the interpretation of the natural response model (in particular, the implied marginal effects) is quite different (see, e.g., Finney, 1964).

Examples of recent papers that consider estimation of misclassification model parameters or misclassification probabilities include Manski (1985), Chua and Fuller (1987), Brown and Light (1992), Poterba and Summers (1995), Abrevaya and Hausman (1997), and Hausman, Abrevaya, and Scott-Morton (1998). These last two papers provide parametric (maximum likelihood) estimators of the model when the function F is known and a semiparametric estimator for the case where F is unknown and the misclassification probabilities a and a^* are constants (independent of all covariates). They also show that when F is unknown, the coefficients of covariates that do not affect the misclassification probabilities can be estimated.

This paper shows that (given some regularity) the entire model is identified even when the functions a, a^*, and F are unknown.

Assumption A1. Assume for all x that $0 \le a(x)$, $0 \le a^*(x)$, and $a(x) + a^*(x) < 1$. Assume that v, conditional on x, is continuously distributed. Assume that $F(w)$ is three times differentiable with $f(w) = dF(w)/dw \neq 0$ and $f'(w) = df(w)/dw$. Assume $|\gamma| = 1$ and, for all $\beta^* \neq \beta$, $\text{prob}([f'(v\gamma + x\beta)/f(v\gamma + x\beta)] \neq E[f'(v\gamma + x\beta)/f(v\gamma + x\beta)|v\gamma + x\beta^*]) > 0$.

The assumption that the sum of misclassification probabilities is less than one is what Hausman et al. (1998) call the monotonicity condition, and it holds by construction in the "natural response" form of the model. Letting $|\gamma| = 1$ is an arbitrary free normalization, as long as $\gamma \neq 0$. Only the covariate v is assumed to be continuous. The final condition in Assumption A1 is a parametric identification assumption that would provide identification of β from the score function if f was a known function and there was no misclassification.

Define the function $\phi(v,x)$ by

$$\phi(v,x) = \frac{\partial^2 g(v,x)/\partial v^2}{\partial g(v,x)/\partial v} \,\text{sign}\left[E\left(\frac{\partial g(v,x)}{\partial v}\right)\right]. \tag{2}$$

Let $r(v,x)$ be any function such that $r(v,x) \ge 0$, $\sup r(v,x)$ is finite, and $E[r(v,x)] = 1$.

LEMMA 1. *Given Assumption* A1, $\phi(v,x) = f'(v\gamma + x\beta)/f(v\gamma + x\beta)$, $\gamma = \text{sign}(E[r(v,x)\partial g(v,x)/\partial v])$, *and* $\beta = \arg\min_{\beta^*} E[(\phi(v,x) - E[\phi(v,x)|v\gamma + x\beta^*])^2]$. *Also*, $\beta = E(r(v,x)[\partial\phi(v,x)/\partial x]/[\partial\phi(v,x)/\partial v])\gamma$.

This lemma shows identification of the model coefficients. Estimation based on this lemma could proceed as follows. First, estimate $\hat{g}$ as a nonparametric regression of y on v and x. Next define $\hat{\phi}$ by equation (2), replacing g with $\hat{g}$ and the expectation with a sample average. Then let $\hat{\gamma}$ equal the sign of any weighted average derivative of $E(y|v,x)$ with respect to v (using, e.g., the estimator of Powell, Stock, and Stoker, 1989).

The lemma suggests two different estimators for β. Let $\xi(v\gamma + x\beta^*) = E[\phi(v,x)|v\gamma + x\beta^*]$ for any β^* and let $\hat{\xi}(v\hat{\gamma} + x\beta^*)$ be a nonparametric regression of $\hat{\phi}(v,x)$ on $v\hat{\gamma} + x\beta^*$. The estimate $\hat{\beta}$ is then the value of β^* that minimizes the sample average of $[\hat{\phi}(v,x) - \hat{\xi}(v\hat{\gamma} + x\beta^*)]^2$. This is essentially Ichimura's (1993) linear index model estimator, using $\hat{\phi}(v,x)$ as the dependent variable.

Another estimator for β suggested by the lemma is to let $\hat{\beta}$ equal the sample average of $r(v,x)[\partial\hat{\phi}(v,x)/\partial x]/[\partial\hat{\phi}(v,x)/\partial v]\hat{\gamma}$. This is an average derivative type estimator, which is only feasible for continuously distributed regressors because of the need to estimate the term $\partial\hat{\phi}(v,x)/\partial x$.

More generally, Lemma 1 shows that $\phi(v,x) = \xi(v\gamma + x\beta)$, so β can be estimated using any of a variety of linear index model estimators, treating $\hat{\phi}(v,x)$

as the dependent variable. For example, the method of Powell et al. (1989) could be used to estimate the coefficients of the continuous regressors and that of Horowitz and Härdle (1996) for the discrete regressors. The limiting distributions of these estimators will be affected by the use of an estimated dependent variable $\hat{\phi}(v,x)$ instead of an observed one. However, all of these estimators involve unconditional expectations, estimated as averages of functions of nonparametric regressions. With sufficient regularity (including judicious selection of the function r, e.g., having r be a density function that equals zero wherever ϕ might be small), such expectations can typically be estimated at rate root n (see, e.g., Newey and McFadden, 1994). Also, some relevant results on the uniform convergence and limiting distribution of nonparametric kernel estimators based on estimated (generated) variables include Andrews (1995) and Ahn (1997).

Define $w = v\gamma + x\beta$, which by Lemma 1 is identified. Let $f_w(w)$ denote the unconditional probability density function of w. Define h by $h(w,x) = E(y|w,x) = a(x) + b(x)F(w)$. Define the function ψ by the indefinite integral

$$\varphi(w) = E\left(\frac{\partial^2 h(w,x)/\partial w^2}{\partial h(w,x)/\partial w}\,\middle|\, w\right)$$
$$\psi(\varpi) = \exp\int \varphi(\varpi)\,d\varpi \tag{3}$$

Let Ω_w and Ω_e denote the supports of w and $-e$, respectively. Define the constant c by $c = \int_{\Omega_w} \psi(w)\,dw$.

LEMMA 2. *Given Assumption* A1, $f(w) = \psi(w)/c$ *and* $b(x) = E([\partial h(w,x)/\partial w]/\psi(w)|x)c$. *If* Ω_e *is a subset of* Ω_w, *then* $c = E[\psi(w)/f_w(w)]$.

This lemma shows that the density function $f(w)$ and the misclassification function $b(x)$ are identified up to the constant c, and the constant c is also identified (and can be estimated as a sample average), provided that the data generating process for w has sufficiently large support.

Estimators based directly on Lemma 2 would consist of the following steps. First, construct $\hat{w} = v\hat{\gamma} + x\hat{\beta}$ and let $\hat{h}(\hat{w},x)$ be a nonparametric regression of y on $\hat{w}$ and x. Next, let $\hat{\zeta}(\hat{w})$ be a nonparametric regression of $[\partial^2\hat{h}(\hat{w},x)/\partial\hat{w}^2]/[\partial\hat{h}(\hat{w},x)/\partial\hat{w}]$ on $\hat{w}$ and define the function $\hat{\psi}(w) = \exp\int\hat{\zeta}(w)\,dw$. The scalar $\hat{c}$ then equals the sample average of $\hat{\psi}(\hat{w})/\hat{f}_w(\hat{w})$, where $\hat{f}_w$ is a nonparametric estimator (e.g., a kernel estimator) of the density of $\hat{w}$. Finally, $\hat{f}_w(w) = \hat{\psi}(w)/\hat{c}$, and $\hat{b}(x)$ equals $\hat{c}$ times a nonparametric regression of $[\partial\hat{h}(\hat{w},x)/\partial\hat{w}]/\hat{\psi}(\hat{w})$ on x. The resulting estimates should be consistent, as long as uniformly consistent nonparametric estimators are used at each stage. Note that consistency may require trimming (possibly asymptotic trimming) to a compact subset of Ω_w, because of division by the density f_w.

The preceding lemmas show that the marginal effects $\partial \Pr(\tilde{y}=1|v,x)/\partial x = f(v\gamma + x\beta)\beta$ and $\partial \Pr(\tilde{y}=1|v,x)/\partial v = f(v\gamma + x\beta)\gamma$ are identified and that the misclassification error function $b(x)$ is also identified. If $a(x) = a^*(x)$, that is, if the probability of misclassification does not depend on $\tilde{y}$, then Lemma 2 implies that the misclassification probability $a(x) = a^*(x) = [1 - b(x)]/2$ is also identified.

Instead of using Lemma 2, log derivatives of $b(x)$ (and hence of $a(x)$ and $a^*(x)$ when they are equal) with respect to continuously distributed elements of x can be directly estimated, without requiring numerical integration, the "large w support" assumption, or the generated variable $\hat{w}$, by the following lemma.

LEMMA 3. *Let x_j be any continuously distributed element of x and let β_j be the corresponding element of β. Let Assumption* A1 *hold and assume that $b(x)$ is differentiable in x_j. Then*

$$\frac{\partial \ln b(x)}{\partial x_j} = E\left(\frac{\partial^2 g(v,x)/\partial v \partial x_j}{\partial g(v,x)/\partial v} - \phi(v,x)\beta_j | x\right)$$

By Lemma 3, a nonparametric regression of $[\partial^2 \hat{g}(v,x)/\partial v \partial x_j]/[\partial \hat{g}(v,x)/\partial v] - \hat{\phi}(v,x)\hat{\beta}_j$ on x is an estimator of $\partial \ln b(x)/\partial x_j$. Dividing this estimate by -2 yields an estimate of $\partial \ln a(x)/\partial x_j$ and $\partial \ln a^*(x)/\partial x_j$ when $a(x) = a^*(x)$.

Next, consider identification of $a(x)$ and $a^*(x)$ when they are not equal. Let $F_w(w|x)$ denote the conditional cumulative distribution function of w given x, let $f_2(w|x) = \partial F_w(w|x)/\partial w$ be the conditional probability density function of w given x, and let $\Omega_{w|x}$ denote the support of w given x. Let $\theta(x) = 1 - E[f(w)F_w(w|x)/f_w(w|x)|x]$.

LEMMA 4. *Let Assumption* A1 *hold and assume that Ω_e is a subset of $\Omega_{w|x}$. Then $a(x) = E[h(w,x)|x] - b(x)\theta(x)$, $a^*(x) = b(x) - 1 + a(x)$, and $F(w) = E([h(w,x) - a(x)]/b(x)|w)$.*

Estimation of $\theta(x)$ requires extreme values of w given x, and hence of v, to be observable. Some intuition for this result comes from the observation that $g(v,x) \approx a(x)$ for very large v and $g(v,x) \approx 1 - a^*(x)$ for very small v. Hence, analogous to the estimation of c, data in the tails are required for estimation of $a(x)$ and $a^*(x)$. Estimation proceeds as in the previous lemmas, that is, employing $\hat{w}$ in place of w, nonparametric estimation of the density function $f_w(w|x)$, and nonparametric regression to estimate conditional expectations.

Taken together, these lemmas show that the entire model is identified. The parameters γ and β can be consistently estimated (with regularity, at rate root n), and the functions $a(x)$, $a^*(x)$, and $F(w)$ can be consistently estimated nonparametrically. The estimators provided here are not likely to be very practical, because they involve up to third-order derivatives and repeated applications of nonparametric regression, and they do not exploit some features of the model such as monotonicity of F. However, the demonstration that the entire model is identified suggests that the search for better estimators would be worthwhile.

REFERENCES

Abrevaya, J. & J.A. Hausman (1997) Semiparametric Estimation with Mismeasured Dependent Variables: An Application to Panel Data on Employment Spells. Manuscript.

Ahn, H. (1997) Semiparametric estimation of single index model with nonparametric generated regressors. *Econometric Theory* 13, 3–31.

Andrews, D.W.K. (1995) Nonparametric kernel estimation for semiparametric models. *Econometric Theory* 11, 560–596.

Brown, J.N. & A. Light (1992) Interpreting panel data on job tenure. *Journal of Labor Economics* 10, 219–257.

Chua, T.C. & W.A. Fuller (1987) A model for multinomial response error applied to labor flows. *Journal of the American Statistical Association* 82, 46–51.

Finney, D.J. (1964) *Statistical Method in Biological Assay.* New York: Havner.

Hausman, J.A., J. Abrevaya, & F.M. Scott-Morton (1998) Misclassification of the dependent variable in a discrete-response setting. *Journal of Econometrics* 87, 239–269.

Horowitz, J.L. & W. Härdle (1996) Direct semiparametric estimation of single-index models with discrete covariates. *Journal of the American Statistical Association* 91, 1632–1640.

Ichimura, H. (1993) Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *Journal of Econometrics* 58, 71–120.

McFadden, D. (1984) Econometric analysis of qualitative response models. In Z. Griliches & M.D. Intriligator (eds.), *Handbook of Econometrics*, vol. 2. Amsterdam, North-Holland.

Manski, C.F. (1985) Semiparametric analysis of discrete response: Asymptotic properties of the maximum score estimator. *Journal of Econometrics* 82, 46–51.

Newey, W.K. & D.L. McFadden (1994) Large sample estimation and hypothesis testing. In R.F. Engle, D.L. McFadden (eds.), *Handbook of Econometrics*, vol. 4. Amsterdam: North-Holland.

Poterba, J.M. & 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.

Powell, J.L., J.H. Stock, & T.M. Stoker (1989) Semiparametric estimation of index coefficients. *Econometrica* 57, 1403–1430.

APPENDIX

Proof of Lemma 1. $\partial g(v,x)/\partial v = b(x) f(v\gamma + x\beta)\gamma$, $b(x) > 0$, and $f(v\gamma + x\beta) > 0$, so $\gamma = \text{sign}[\partial g(v,x)/\partial v]$. Here $\partial^2 g(v,x)/\partial v^2 = b(x) f'(v\gamma + x\beta)\gamma^2$, and $\gamma^2 = 1$, so $\phi(v,x) = f'(v\gamma + x\beta)/f(v\gamma + x\beta)$. Let $\xi(v\gamma + x\beta^*) = E[\phi(v,x)|v\gamma + x\beta^*]$. It follows from the previous expression for ϕ that $\phi(v,x)$ and the final equality in Assumption A1 that $\text{prob}[\phi(v,x) = \xi(v\gamma + x\beta^*)] > 0$ for all $\beta \neq \beta^*$, and $\phi(v,x) = \xi(v\gamma + x\beta)$, so $\beta = \arg\min_{\beta^*} E[(\phi(v,x) - E[\phi(v,x)|v\gamma + x\beta^*])^2]$.

The alternative expression $\beta/\gamma = [\partial\phi(v,x)/\partial x]/[\partial\phi(v,x)/\partial v]$ follows because ϕ depends on x and v only through $v\gamma + x\beta$, so $E(r(v,x)[\partial\phi(v,x)/\partial x]/[\partial\phi(v,x)/\partial v])\gamma = E[r(v,x)\beta/\gamma]\gamma = \beta$. ■

Proof of Lemma 2. $\partial h(w,x)/\partial w = b(x)f(w)$, $\partial^2 h(w,x)/\partial w^2 = b(x)f'(w)$, so $[\partial^2 h(w,x)/\partial w^2]/[\partial h(w,x)/\partial w] = f'(w)/f(w) = E([\partial^2 h(w,x)/\partial w^2]/[\partial h(w,x)/\partial w]|w)$. Then $\psi(\varpi) = \exp[\int f'(\varpi)/f(\varpi)\,d\varpi] = f(\varpi)c$, where $\ln c$ is the constant of integration.

$$E([\partial h(w,x)/\partial w]/\psi(w)|x)/c = E([b(x)f(w)]/\psi(w)|x)/c$$
$$= E([b(x)f(w)]/[f(w)c]|x)/c = b(x).$$

Here $E[\psi(w)/f_w(w)] = \int_{\Omega_w}[\psi(w)/f_w(w)]\,f_w(w)\,dw = \int_{\Omega_w}\psi(w)\,dw = \int_{\Omega_w} f(w)c\,dw = c$, where the last equality holds as long as Ω_w contains every value of e for which $f(e)$ is nonzero.

Proof of Lemma 3. $\partial^2 g(v,x)/\partial v\partial x_j = f(\gamma v + \beta x)\partial b(x)/\partial x_j + b(x)f'(\gamma v + \beta x)\beta_j$, so $[\partial^2 g(v,x)/\partial v\partial x_j]/[\partial g(v,x)/\partial v] = [\partial b(x)/\partial x_j]/b(x) + [f'(\gamma v + \beta x)/f(\gamma v + \beta x)]\beta_j = \partial \ln b(x)/\partial x_j + \phi(v,x)\beta_j$. The lemma then follows immediately.

Proof of Lemma 4. Let $\partial\Omega_{w|x}$ denote the boundary of the support $\Omega_{w|x}$. Applying an integration by parts gives $E[F(w)|x] = \int_{\Omega_{w|x}} F(w)f_w(w|x)\,dw = F(w)F_w(w|x)|_{w=\partial\Omega_{w|x}} - \int_{\Omega_{w|x}} f(w)F_w(w|x)\,dw$. Having Ω_e be a subset of $\Omega_{w|x}$ ensures that $F(w)F_w \times (w|x)|_{w=\partial\Omega_{w|x}} = 1$, and so $\theta(x) = 1 - \int_{\Omega_{w|x}}[f(w)F_w(w|x)/f_w(w|x)]f_w(w|x)\,dw = E[F(w)|x]$. Therefore, $E[h(w,x)|x] = a(x) + b(x)E[F(w)|x] = a(x) + b(x)\theta(x)$, which gives the identification of $a(x)$. The expression $a^*(x) = b(x) - 1 + a(x)$ then follows from the definition of $b(x)$, and $h(w,x) = a(x) + b(x)F(w)$ is then used to obtain $F(w)$.

https://doi.org/10.1017/S0266466600164060 Published online by Cambridge University Press