ELSEVIER Journal of Econometrics 87 (1998) 239–269

JOURNAL OF Econometrics

Misclassification of the dependent variable in a discrete-response setting

J.A. Hausman[a,*], Jason Abrevaya[b], F.M. Scott-Morton[b]

[a] Department of Economics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[b] Graduate School of Business, University of Chicago, Chicago, IL 60637, USA

Received 1 September 1996; received in revised form 1 November 1997

Abstract

Misclassification of dependent variables in a discrete-response model causes inconsistent coefficient estimates when traditional estimation techniques (e.g., probit or logit) are used. A modified maximum likelihood estimator that corrects for misclassification is proposed. A semiparametric approach, which combines the maximum rank correlation estimator of Han (1987) (Journal of Econometrics 35, 303–316) with isotonic regression, allows for more general forms of misclassification than the maximum likelihood approach. The parametric and semiparametric estimation techniques are applied to a model of job change with two commonly used data sets, the Current Population Survey (CPS) and the Panel Study of Income Dynamics (PSID).

JEL classification: C25; C14

Keywords: Binary choice model; Response error; Isotonic regression

1. Introduction

Misclassification of dependent variables in a discrete-response model causes inconsistent coefficient estimates when traditional estimation techniques (e.g., probit or logit) are used. By 'misclassification', we mean that the response is reported or recorded in the wrong category; for example, a variable is recorded as a one when it should have the value zero. This mistake might easily happen in an interview setting where the respondent misunderstands the question or the

* Corresponding author. E-mail: jhausman@mit.edu

PII S0304-4076(98)00015-3

interviewer simply checks the wrong box. Other data sources where the researcher suspects measurement error, such as historical data, certainly exist as well. Also, researchers often will construct a dummy variable to serve as a proxy for some underlying 'true' variable; the extent of misclassification in this situation will depend on how good a proxy the constructed variable is.

We show that the researcher can correct the problem of misclassification by employing a modified maximum likelihood approach and, in doing so, can explicitly estimate the extent of misclassification in the data. We also discuss a semiparametric method of estimating the unknown slope coefficients which does not depend on an assumed error distribution. The semiparametric method, which combines the maximum rank correlation estimator of Han (1987) with isotonic regression, allows for more general forms of misclassification than the maximum likelihood approach. In addition, the semiparametric method allows for estimation of the observed response function with known asymptotic properties.

We apply our methodology to a model of job change with two commonly used data sets, the Current Population Survey (CPS) and the Panel Study of Income Dynamics (PSID). In both cases, we construct a job-change variable from responses about job tenure, a method quite common in the empirical literature. Unfortunately, tenure responses are well-known to be poorly measured (see, e.g., Brown and Light 1992) because respondents often misunderstand the question, have poor recall, or confuse a change in position with an actual job change. Both our parametric and semiparametric estimates demonstrate conclusively that significant misclassification of the job-change variable exists in both the CPS and PSID samples. Furthermore, the probability of misclassification is not the same across observed response classes. In the CPS sample, workers who change jobs are more likely to misreport than workers who do not change jobs. In the PSID sample, inferences about misclassification of the constructed job-change variable depend in part on the method used to construct the variable.

Section 2 of this paper introduces a basic model of misclassification for a binary dependent variable. Section 3 proposes the natural parametric estimators for this model. Section 4 presents the results from a Monte Carlo study. Section 5 describes the semiparametric approach to the problem of misclassification and relaxes the model of misclassification from the previous sections. Section 6 applies the estimation techniques to the CPS and PSID data. Finally, Section 7 concludes.

2. Binary choice model with misclassification

We start with the usual latent-variable specification of the binary response model (cf., Greene (1990) or McFadden (1984)). Let y_i^* be the latent variable (where i ranges from 1 to n, the sample size), given by

$$y_i^* = x_i'\beta + \varepsilon_i, \tag{1}$$

where ε_i is an i.i.d. error disturbance. Let F denote the (common) c.d.f. of $-\varepsilon_i$. Write $\tilde{y}_i$ for the true response,

$$\tilde{y}_i = 1(y_i^* \geqslant 0), \tag{2}$$

where $1(E)$ is the indicator function equal to one if E is true and zero otherwise. Without misclassification of the dependent variable, the true response $\tilde{y}_i$ is observed. This paper focuses on the situation in which the true response may be misclassified.

The model of misclassification that we consider is one in which the probability of misclassification depends on the value of $\tilde{y}_i$, but is otherwise independent of the covariates x_i. In particular, if y_i denotes the observed binary dependent variable, the misclassification probabilities are

$$\alpha_0 = \Pr(y_i = 1 | \tilde{y}_i = 0), \tag{3}$$

$$\alpha_1 = \Pr(y_i = 0 | \tilde{y}_1 = 1). \tag{4}$$

The probability that a zero is misclassified as a one is given by α_0; the probability that a one is misclassified as a zero is given by α_1. This model of misclassification is called *Model I*. A more flexible model of misclassification is considered in Section 5.

The expected value of the observed dependent variable is

$$\mathrm{E}(y_i|x_i) = \Pr(y_i = 1|x_i) = \alpha_0 + (1 - \alpha_0 - \alpha_1)F(x_i'\beta), \tag{5}$$

which collapses to the usual expression, $F(x_i'\beta)$, when there is no misclassification ($\alpha_0 = \alpha_1 = 0$).

We briefly mention the connection of this model of misclassification to a similar concept in the biometrics literature (e.g., Finney 1964). Biometricians are oftentimes concerned with 'natural responses' by experimental subjects which have nothing to do with the stimulus being tested. For instance, in testing the toxicity of a substance, some subjects may die of natural causes unrelated to the experiment. In our terminology, the fraction of those subjects who would have survived the experiment but die from unrelated natural causes would be α_0 (if 'death' corresponds to an observed y_i equal to one). Such responses are effectively misclassified. Similarly, there may be some subjects who are immune to the treatment. If the immunity is independent of the observables used in the empirical analysis, the fraction of subjects who always survive the treatment is given by α_1 in our notation.[1]

[1] Viewed in this way, notice that our model is indistinguishable from the following binary choice model with heterogeneity incorporated: a fraction α_0 of individuals always respond with a one, independent of the observed x_i; a fraction α_1 always respond with a zero, again independent of the observed x_i; and, the remaining individuals follow the traditional binary choice model. The parameters α_0 and α_1 would then have a different interpretation, as values describing the degree of heterogeneity rather than the extent of misclassification.

3. Parametric estimation

In this section, we assume that F is known (e.g., normal or logistic distribution) and consider parametric methods for estimating the binary-choice model with misclassification. One can estimate $(\alpha_0, \alpha_1, \beta)$ with nonlinear least squares (NLS), based on the moment condition in Eq. (5), minimizing

$$\sum_{i=1}^{n} (y_i - a_0 - (1 - a_0 - a_1)F(x_i'b))^2 \tag{6}$$

over (a_0, a_1, b). Significance tests on a_0 and a_1 can be used as tests of misclassification.

Alternatively, one can estimate $(\alpha_0, \alpha_1, \beta)$ with maximum likelihood estimation (MLE), maximizing the log likelihood function

$$\mathscr{L}(a_0, a_1, b) = n^{-1} \sum_{i=1}^{n} \{y_i \ln(a_0 + (1 - a_0 - a_1)F(x_i'b)) + (1 - y_i)\ln(1 - a_0 - (1 - a_0 - a_1)F(x_i'b))\} \tag{7}$$

over (a_0, a_1, b).

Conditions for identification of $(\alpha_0, \alpha_1, \beta)$ are similar to those for the traditional binary choice model (see, e.g., Example 1.2 of Newey and McFadden (1994)). In the linear probability model (where $F(v) = v$), however, the parameters are not separately identified since

$$\mathrm{E}(y_i|x_i) = \alpha_0 + (1 - \alpha_0 - \alpha_1)(x_i'\beta) = (\alpha_0 + \beta_0) + z_i'((1 - \alpha_0 - \alpha_1)\beta_1), \tag{8}$$

where $x_i = (1, z_i')'$ and $\beta = (\beta_0, \beta_1')'$. This example shows that identification of the model parameters stems from the nonlinearity of F, which partly motivates the semiparametric approach developed in Section 5.

The only additional assumption needed for identification is the following:[2]

Monotonicity condition # 1 (MC1): $\alpha_0 + \alpha_1 < 1$

To see why MC1 is needed, consider a symmetric F (where $F(v) = 1 - F(-v)$) like the normal or logistic distributions. Then, denoting $\tilde{\alpha}_0 = 1 - \alpha_1$, $\tilde{\alpha}_1 = 1 - \alpha_0$, and $\tilde{\beta} = -\beta$, notice that

$$\tilde{\alpha}_0 + (1 - \tilde{\alpha}_0 - \tilde{\alpha}_1)F(x_i'\tilde{\beta}) = \alpha_0 + (1 - \alpha_0 - \alpha_1)F(x_i'\beta). \tag{9}$$

Thus, estimators based on Eq. (5) (like NLS and MLE) cannot distinguish between the parameter values $(\alpha_0, \alpha_1, \beta)$ and $(1 - \alpha_1, 1 - \alpha_0, -\beta)$. MC1 rules out

[2] In cases of multiple interviews, Chua and Fuller (1987) demonstrate that identification of misclassification probabilities is possible given a sufficient number of interviews. However, generally they must make special assumptions on the form of misclassification to achieve identification.

such a situation since $\alpha_0 + \alpha_1 < 1$ implies $(1 - \alpha_1) + (1 - \alpha_0) > 1$. If misclassification is so problematic that $\alpha_0 + \alpha_1$ is larger than one (in which case the project should probably be abandoned!), imposing MC1 will result in estimates of β (and, in turn, marginal effects) of the wrong sign.

MC1 implies that $\alpha_0 + (1 - \alpha_0 - \alpha_1)F(v)$ is strictly increasing in v if F is strictly increasing (i.e., ε has positive density everywhere). In combination with the other identification conditions needed for a given F, MC1 should yield identification. For instance, the identification result for probit estimation is given in the following theorem:

Theorem 1. If F is the normal c.d.f., $\mathrm{E}[xx']$ *exists and is nonsingular, and MC1 holds, then the parameters* $(\alpha_0, \alpha_1, \beta)$ *are identified by NLS or MLE.*

The proof follows immediately from Newey and McFadden (1994) (pp. 2125, 2126).

For the remainder of this section, we focus on MLE estimation since it is efficient when F is known. If misclassification is ignored and the log-likelihood

$$\mathscr{L}(b) = n^{-1} \sum_{i=1}^{n} \{y_i \ln F(x_i'b) + (1 - y_i)\ln(1 - F(x_i'b))\} \tag{10}$$

is maximized instead, the resulting estimate of β will generally be inconsistent (see Hausman et al. (1996) for a complete discussion). We are particularly interested in the inconsistency of estimates of β (using Eq. (10)) when only small amounts of misclassification are present since this might be the most common case facing a researcher. With no misclassification, the true value β maximizes the expected log-likelihood (i.e., the expectation of Eq. (10)). Let $\beta_{\mathrm{E}}(\alpha_0, \alpha_1)$ denote the maximand of the expected log-likelihood when there is misclassification (so that $\beta_{\mathrm{E}}(0,0) = \beta$). The partial derivatives of $\beta_{\mathrm{E}}(\alpha_0, \alpha_1)$ with respect to the misclassification probabilities, evaluated at $\alpha_0 = \alpha_1 = 0$, are given by[3]

$$\left.\frac{\partial \beta_{\mathrm{E}}}{\partial \alpha_0}\right|_{\alpha_0=\alpha_1=0} = -\left[\mathrm{E}\left(\frac{f(x'\beta)^2}{F(x'\beta)(1 - F(x'\beta))}xx'\right)\right]^{-1} \mathrm{E}\left(\frac{f(x'\beta)}{F(x'\beta)}x\right), \tag{11}$$

$$\left.\frac{\partial \beta_{\mathrm{E}}}{\partial \alpha_1}\right|_{\alpha_0=\alpha_1=0} = \left[\mathrm{E}\left(\frac{f(x'\beta)^2}{F(x'\beta)(1 - F(x'\beta))}xx'\right)\right]^{-1} \mathrm{E}\left(\frac{f(x'\beta)}{1 - F(x'\beta)}x\right). \tag{12}$$

In general, the degree of inconsistency will depend on the distributions of the index $x'\beta$ and the covariate vector x.

[3] A derivation is given in the appendix of Hausman et al. (1996).

The Fisher information matrix associated with maximization of the log-likelihood in Eq. (7) is not block diagonal in the parameters $(\alpha_0, \alpha_1, \beta)$ since

$$-\mathrm{E}\begin{bmatrix} \dfrac{\partial^2 \mathscr{L}}{\partial \alpha_0^2} & \dfrac{\partial^2 \mathscr{L}}{\partial \alpha_0 \partial \alpha_1} & \dfrac{\partial^2 \mathscr{L}}{\partial \alpha_0 \partial \beta'} \\ \dfrac{\partial^2 \mathscr{L}}{\partial \alpha_1 \partial \alpha_0} & \dfrac{\partial^2 \mathscr{L}}{\partial \alpha_1^2} & \dfrac{\partial^2 \mathscr{L}}{\partial \alpha_1 \partial \beta'} \\ \dfrac{\partial^2 \mathscr{L}}{\partial \beta \partial \alpha_0} & \dfrac{\partial^2 \mathscr{L}}{\partial \beta \partial \alpha_1} & \dfrac{\partial^2 \mathscr{L}}{\partial \beta \partial \beta'} \end{bmatrix}$$

$$= \mathrm{E}\begin{bmatrix} \dfrac{(1-F)^2}{P(1-P)} & -\dfrac{F(1-F)}{P(1-P)} & \dfrac{(1-\alpha_0-\alpha_1)f(1-F)}{P(1-P)}x' \\ -\dfrac{F(1-F)}{P(1-P)} & \dfrac{F^2}{P(1-P)} & -\dfrac{(1-\alpha_0-\alpha_1)fF}{P(1-P)}x' \\ \dfrac{(1-\alpha_0-\alpha_1)f(1-F)}{P(1-P)}x & -\dfrac{(1-\alpha_0-\alpha_1)\,fF}{P(1-P)}x & \dfrac{(1-\alpha_0-\alpha_1)^2\,f^2}{P(1-P)}xx' \end{bmatrix},$$

where $f \equiv f(x'\beta)$, $F \equiv F(x'\beta)$, and $P \equiv \alpha_0 + (1-\alpha_0-\alpha_1)F(x'\beta)$. Thus, previous papers which assume they know the misclassification probabilities from exogenous sources (e.g., Poterba and Summers 1995) suffer from two defects. First, their estimates are likely to be inconsistent unless their assumed misclassification probabilities are consistent estimates of the true misclassification probabilities. Second, even with consistent estimates of the misclassification probabilities, the standard errors of their coefficient estimates are understated since these probabilities are not known with certainty.

Without misclassification, the information matrix simplifies to

$$\mathrm{E}\begin{bmatrix} \dfrac{1-F}{F} & -1 & \dfrac{f}{F}x' \\ -1 & \dfrac{F}{1-F} & -\dfrac{f}{1-F}x' \\ \dfrac{f}{F}x & -\dfrac{f}{1-F}x & \dfrac{f^2}{F(1-F)}xx' \end{bmatrix}. \tag{13}$$

The bottom-right block is the usual information matrix for β in a binary choice model. Since the information matrix is still not block diagonal, the modified log-likelihood in Eq. (7) yields less efficient estimates of β than the log-likelihood in Eq. (10).

Researchers often care about marginal effects rather than estimates of β itself. The marginal effects of interest here are

$$\frac{\partial \Pr(\tilde{y}=1|x)}{\partial x} = f(x'\beta)\beta, \tag{14}$$

which can be estimated by plugging in an estimate $\hat{\beta}$ and evaluating at different levels of the estimated index $x'\hat{\beta}$. The marginal effect on the observed response is

$$\frac{\partial \Pr(y=1|x)}{\partial x} = (1-\alpha_0-\alpha_1)f(x'\beta)\beta, \tag{15}$$

which is always less than the marginal effect on the true response. The marginal effects on the observed and true reponses will differ by a factor of $(1-\alpha_0-\alpha_1)$ regardless of the x value at which they are evaluated.

The MLE method described in this section can be easily extended to discrete-response models with more than two categories; see Hausman et al. (1996) and Abrevaya and Hausman (1997) for a systematic treatment of this topic. Furthermore, if the researcher thinks that misclassification depends on the covariates (or other observables), the dependence can be modeled explicitly and incorporated into the log-likelihood function. This extension is briefly discussed in Section 7.

4. Monte Carlo simulations

In order to assess the empirical importance of misclassification, we examine some Monte Carlo simulation results. The Monte Carlo design has three covariates: the first variable, x_1, is drawn from a lognormal distribution; the second, x_2, is a dummy variable equal to one with probability 1/3; the third, x_3, is distributed uniformly. The error disturbance, ε, is drawn from a standard normal distribution. The latent dependent variable is given by

$$y_i^* = -1 + 0.2x_{i1} + 1.5x_{i2} - 0.6x_{i3} + \varepsilon_i. \tag{16}$$

The observed dependent variable is generated using symmetric misclassification (i.e., $\alpha_0 = \alpha_1$).

Table 1 reports the results of the Monte Carlo simulations. Misclassification probabilities of 2%, 5%, and 20% are considered. A sample size of 5000 is used for each simulation. Results from ordinary probit estimation and modified MLE (where we restrict $\alpha_0 = \alpha_1$) are reported. Even in the case of a small amount of misclassification, ordinary probit produces estimates that are biased by 15–25%. The problem worsens as the amount of misclassification grows. Notice that the estimated standard errors for the modified MLE increase with the level of misclassification, whereas the standard errors for the ordinary probit do not.

Table 1
Monte Carlo simulation results (n = 5000)

	True	Probit Coefficient estimates	Ratio to constant	MLE ($\alpha = \alpha_0 = \alpha_1$) Coefficient estimates	Ratio to constant
α	0.02	—	—	0.0192 (0.0054)	—
β_0	− 1.0	− 0.787 (0.069)	—	− 0.990 (0.068)	—
β_1	0.2	0.158 (0.001)	0.20	0.199 (0.008)	0.20
β_2	1.5	1.27 (0.06)	1.61	1.49 (0.08)	1.51
β_3	− 0.6	− 0.158 (0.023)	0.66	− 0.598 (0.026)	0.60
α	0.05	—	—	0.0497 (0.0076)	—
β_0	− 1.0	− 0.567 (0.073)	—	− 1.007 (0.084)	—
β_1	0.2	0.114 (0.010)	0.20	0.201 (0.010)	0.20
β_2	1.5	1.06 (0.05)	1.87	1.50 (0.08)	1.50
β_3	− 0.6	− 0.431 (0.019)	0.76	− 0.599 (0.032)	0.60
α	0.02	—	—	0.198 (0.014)	—
β_0	− 1.0	− 0.163 (0.061)	—	− 0.991 (0.168)	—
β_1	0.2	0.037 (0.005)	0.23	0.198 (0.023)	0.20
β_2	1.5	0.554 (0.045)	3.40	1.48 (0.18)	1.49
β_3	− 0.6	− 0.228 (0.018)	1.40	− 0.592 (0.072)	0.60

Note: See text for Monte Carlo design. Results are from 146 Monte Carlo simulations. The standard deviations of the simulation results are reported in parantheses.

Thus, not only does probit yield inconsistent estimates, but it can also overstate the precision of the estimates. The ratios of the estimated coefficients are also reported, and ordinary probit yields inconsistent estimates of these ratios as well.[4]

[4] The estimated ratios remain consistent if the simulated covariates are drawn from normal distributions, even if the estimates of the individual parameters are biased. See Ruud (1983) for a discussion of this result.

Additional simulations, which consider the logit specification and NLS estimation, are reported in Hausman et al. (1996). The results are qualitatively similar to those reported here.

5. Semiparametric analysis

The assumption of normally distributed (or extreme value) disturbances required by probit (or logit) specifications is not necessary for estimating the binary choice model with misclassification. One can instead use semiparametric methods which do not require distributional assumptions.[5] Recall from Section 3 that the MC1 condition implies that the expected value of the observed dependent variable is an increasing function of the underlying index. This latter condition will be the underlying identification condition for the semiparametric method that we propose. The proposed method involves two stages of estimation, each of which is based on this identification condition.

The first stage estimates β up-to-scale using the *maximum rank correlation* (MRC) estimator of Han (1987). Unlike some other alternatives (e.g., Powell et al. 1989), this binary response estimator is straightforward to calculate even in the presence of multiple explanatory variables and dummy variables; in addition, no bandwidth selection or trimming is required for estimation of β. Unlike the parametric method, a more flexible model of misclassification is sufficient for consistency; the exact form of misclassification need not be specified (nor estimated, as in the parametric case). As a result, the MRC estimates are more robust, both with respect to distributional assumptions and the misclassification mechanism.

The second stage estimates the expectation of the observed response y as a function of the estimated index $x'\hat{\beta}_{mrc}$ using *isotonic regression* (IR). Unlike kernel regression methods, IR does not require the researcher to make any decisions about window widths or kernel weights. As shown by Groeneboom (1993), IR is pointwise consistent and, like kernel regression methods, is $\sqrt[3]{n}$-consistent. The key insight here is that the $\sqrt{n}$-consistent estimate of β can be treated as the true β in deriving the asymptotic distribution of the IR estimates since the MRC convergence rate is faster than the IR convergence rate. As a result, the asymptotic distribution results derived by Groeneboom (1993) can be used. This approach differs from Cosslett (1983), in which β and the response function are estimated *jointly* and, to date, have no known asymptotic distributions.

[5] Manski (1985) discusses use of the maximum score estimator when there is symmetric misclassification ($\alpha_0 = \alpha_1$) in the data. The maximum score estimator is not $\sqrt{n}$-consistent, however.

The key condition needed for the semiparametric analysis is the following:

Semiparametric identification condition (SIC): $\mathrm{E}[y|x] = G(x'\beta)$, *where* $G\colon \mathscr{R} \to (0,1)$ *is a strictly increasing function.*

This condition states that the observable dependent variable is a function of x only through the index value $x'\beta$, sometimes called 'index sufficiency' (see Manski 1988). SIC is stronger than index sufficiency, though, since strict monotonicity is also required. The following theorem formally states that SIC is a weaker condition than MC1 in Model I:

Theorem 2. If Model I and MC1 hold and F is strictly increasing (*i.e., everywhere positive density*), *then SIC holds.*

In this case, the response function is given by $G(v) = \alpha_0 + (1 - \alpha_0 - \alpha_1)F(v)$.

SIC allows for more flexible forms of misclassification than described by Model I. For instance, the probability of misclassification might depend on the level of the latent variable y^* rather than just its sign. One might expect lower misclassification probabilities for extreme negative or positive values of y^* and higher misclassification probabilities for values of y^* closer to zero.

The more flexible model of misclassification is called *Model II*. Like Model I, Model II restricts misclassification to depend on x only through y^*. Unlike Model I, however, misclassification probabilities are not restricted to be constant for all negative y^* and for all positive y^*. In particular, Model II replaces Eq. (3) and Eq. (4) with the following:[6]

$$\mathrm{E}(y|x, y^*) = \mathrm{E}(y|y^*). \tag{17}$$

The analogue to MC1 in Model II is the following monotonicity condition:

Monotonicity condition #2 (MC2): $Y(y^*) \equiv \mathrm{E}(y|y^*)$ *is an increasing function of* y^*.

SIC is a direct consequence of MC2:

Theorem 3. If Model II and MC2 hold and F is strictly increasing (*i.e., everywhere positive density*), *then SIC holds.*

In fact, condition MC2 is a special case of a stochastic-dominance condition used in Abrevaya and Hausman (1997) to show consistency of rank estimators in a more general linear index model than the binary-choice model considered here. The stochastic-dominance condition, briefly stated, is that the random

[6] Alternatively, the misclassification probability can be a function of $x'\beta$ as long as the expectation of the observed y is increasing in the index $x'\beta$.

variable associated with a larger latent dependent variable first-order stochastically dominates the random variable associated with a smaller latent dependent variable. Although this condition describes the complete distribution of y conditional on y^*, it is equivalent to MC2 in the binary-choice case since y can take on only the values of zero and one.

5.1. First-stage estimation: maximum rank correlation

Throughout the remainder of this section, the $k \times 1$ covariate vector x is understood not to contain a constant term since MRC estimation does not identify the location parameter in β. MRC estimation only identifies the remaining components of β up-to-scale, so a normalization of the parameter vector is required as well. Like Sherman (1993), we fix the last parameter so that the parameter space, denoted $\mathscr{B}$, is taken to be a compact subset of $\{\beta \in \mathscr{R}^k\colon |\beta_k| = 1\}$.

The MRC estimator is the value $\hat{\beta}$ that maximizes the objective function

$$S(b) = \sum_{i=1}^{n} \text{Rank}(x_i'b) \cdot y_i \tag{18}$$

over $\mathscr{B}$, where Rank$(\cdot)$ gives the rank of the associated index value (1 for the lowest, n for the highest).[7]

Han (1987) proves strong consistency of $\hat{\beta}$, and Sherman (1993) proves asymptotic normality of $\hat{\beta}$. Since SIC ensures that their arguments apply to the binary-choice model with misclassification, we direct the reader to Sherman (1993) for the regularity conditions and the asymptotic distribution.[8] The result is stated in the following theorem:

Theorem 4. If SIC holds and Assumptions A1, A3, and A4 of Sherman (1993) hold, then the MRC estimator $\hat{\beta}$ is $\sqrt{n}$-consistent and asymptotically normal.

Assumption A1 restricts β to the interior of the compact parameter space $\mathscr{B}$. Assumption A3 has two parts, a full-rank condition on x and a condition that one component of x is continuous (everywhere positive Lebesgue density) conditional on the other components of x. Assumption A4 consists of regularity conditions that allow for a Taylor-expansion argument in the asymptotic normality proof.

---

[7] The fastest sorting algorithms, used to evaluate the Rank$(\cdot)$ function, require $O(n \log n)$ computer instructions. Thus, it takes $O(n \log n)$ instructions to evaluate the objective function $S(b)$ for any given b.

[8] Note that Eq. (20) of Sherman (1993) should read $V = \cdots$ rather than $2V = \cdots$. Likewise, the asymptotic distribution for the binary choice model that follows (p. 134) should not have a 2 multiplying V.

The MRC and MLE estimates of β can be compared using a Hausman (1978) specification test. The MLE estimates are consistent and efficient under Model I and the correct specification of the distribution F. The MRC estimates remain consistent under the more flexible Model II and arbitrary (i.i.d.) specification of F. The Hausman test may reject from either a misspecification of functional form or a misspecification of the model of misclassification.

Whereas MLE estimation explicitly incorporates the misclassification (through the parameters α_0 and α_1) into the objective function, MRC estimation essentially ignores the presence of misclassification. As such, another method must be used in order to investigate the underlying misclassification mechanism. The MRC estimate $\hat{\beta}$ is used to form an estimated index $x'\hat{\beta}$ for use in the second stage of the estimation procedure. The second stage estimates the response function $G(\cdot)$ using the observed dependent variables and the estimated index values.

5.2. Second-stage estimation: isotonic regression

Estimation of the response function $G(\cdot)$ requires a nonparametric regression of y on the estimated index $x'\hat{\beta}$. The standard technique used in the literature is kernel regression (see, e.g., Cavanagh and Sherman 1997). In this section, we instead use isotonic regression. One advantage of isotonic regression is that, like MRC estimation, no bandwidth selection is needed for estimation. Isotonic regression also restricts the response function to be monotonic, which is simply condition SIC used for consistency of MRC estimation. Kernel regression, on the other hand, does not impose monotonicity even though monotonicity has already been used for the first stage. Another advantage for the application considered here is that the resulting estimate is in the form of a step function, so that behavior in the tails is less erratic than kernel estimation.

Denote the index value for observation i by $\hat{v}_i = x_i'\hat{\beta}$. To simplify notation, re-order the observations so that $\hat{v}_1 \leqslant \hat{v}_2 \leqslant \cdots \leqslant \hat{v}_n$. An isotonic function is any nondecreasing function defined at the n index values. The functional estimate, $\hat{G}$, is an isotonic regression (IR) of y on the index values if it minimizes the objective function

$$\sum_{i=1}^{n} (y_i - \hat{G}(\hat{v}_i))^2 \tag{19}$$

over the set of isotonic functions. Eq. (19) determines $\hat{G}(v)$ only at the n index values; at other values, $\hat{G}(v)$ is given by

$$\hat{G}(v) = \begin{cases} 0 & \text{if } v < \hat{v}_1, \\ \hat{G}(\hat{v}_i) & \text{if } v \in (\hat{v}_i, \hat{v}_{i+1}), \\ 1 & \text{if } v > \hat{v}_n. \end{cases} \tag{20}$$

The resulting functional estimate is in the form of a step function.

The algorithm for IR is quite straightforward.[9] The idea is to organize the index values into 'pools', where each pool is assigned a 'best guess' for the value of y conditional on the index being in the pool. The 'best guess' is just the average of the y_i values associated with the index values $\hat{v}_i$ in a pool. The initial set of pools has each pool corresponding to a distinct index value. Then, the algorithm compares the lowest-indexed pool with the next lowest-indexed pool. If the guess for the first pool is less than the guess for the second pool, the pools are left intact; the second pool is used for the next comparison. Otherwise, the pools are combined, and the combined pool is used for the next comparison. This process of comparing adjacent pools is continued until the pools are exhausted. Finally, if any combinations of pools occurred during the last pass-through, the process is repeated. Once the pools are in nondecreasing order, the IR is complete. For each index value $\hat{v}_i$, the estimate $\hat{G}(\hat{v}_i)$ is the guess associated with $\hat{v}_i$'s pool.

When the index values are *known* (rather than estimated), Groeneboom (1985, 1993) proves that the point estimates from isotonic regression are $\sqrt[3]{n}$-consistent. Let $\hat{G}_\beta(v)$ denote the point estimate from an isotonic regression of y on the true index $x'\beta$. Then, Groeneboom (1985, 1993) shows that for $\hat{G}_\beta(v) \in (0,1)$,

$$\frac{n^{1/3}(\hat{G}_\beta(v) - G(v))}{\left(\frac{1}{2}G(v)(1 - G(v))\frac{g(v)}{h(v)}\right)^{1/3}} \xrightarrow{d} 2Z, \tag{21}$$

where g is the derivative of the true function G, h is the density of the index, and the random variable Z is the last time where two-sided Brownian motion minus the parabola u^2 reaches its maximum. The distribution of Z can be written Groeneboom 1985) as

$$f_Z(u) = \tfrac{1}{2}s(u)s(-u), \quad u \in \mathscr{R}, \tag{22}$$

where the function $s(\cdot)$ has a Fourier transform

$$\hat{s}(w) = \frac{2^{1/3}}{\mathrm{Ai}(2^{-1/3}\; wi)}, \tag{23}$$

and where $\mathrm{Ai}(\cdot)$ is the 'Airy function' (as defined in Abramowitz and Stegun (1964), for example) and $i = \sqrt{-1}$.

Due to the rate of convergence of $\hat{\beta}$, the use of the estimated index $x'\hat{\beta}$ in place of the true index $x'\beta$ has no effect on the asymptotic distribution in Eq. (21). The key idea is that $\hat{\beta}$ converges at a faster rate ($n^{-1/2}$) than isotonic regression

[9] See Barlow et al. (1972), Cosslett (1983), or Robertson et al. (1988) for detailed discussions of IR and estimation algorithms.

$(n^{-1/3})$. Note that

$$n^{1/3}(\hat{G}(v) - G(v)) = n^{1/3}(\hat{G}(v) - \hat{G}_\beta(v)) + n^{1/3}(\hat{G}_\beta(v) - G(v)). \tag{24}$$

The second term has an asymptotic distribution given by Eq. (21). The first term converges in probability to zero, as shown in the appendix. The main result is

Theorem 5. If $\hat{\beta}$ is a $\sqrt{n}$-consistent estimate of β and SIC holds, then for any v s.t. $\hat{G}(v) \in (0,1)$ and the index $x'\beta$ has positive density (w.r.t. Lebesgue measure) in a neighborhood around v,

$$\frac{n^{1/3}(\hat{G}(v) - G(v))}{\left(\frac{1}{2}G(v)(1 - G(v))\frac{g(v)}{h(v)}\right)^{1/3}} \xrightarrow{d} 2Z. \tag{25}$$

Note that the assumption of a positive density is implied by Assumption A3 of Sherman (1993), so the conditions of this theorem are no stronger than those for consistency of MRC.

From the symmetry of f_Z, we have $\mathrm{E}(Z) = 0$. Also, via numerical approximation of the Fourier transform and numerical integration, we estimate $\mathrm{Var}(Z) \approx 0.26$. Eq. (25) then yields

$$\mathrm{Var}(\hat{G}(v) - G(v)) \approx 1.04\left(\frac{\hat{G}(v)(1 - \hat{G}(v))\hat{g}(v)}{2n\hat{h}(v)}\right)^{2/3}, \tag{26}$$

where we use the kernel estimate of the index density for $\hat{h}(v)$. Unfortunately, we cannot use the numerical derivative of $\hat{G}(v)$ for $\hat{g}(v)$ since the derivative of a step function is zero except at a finite number of points (where it is infinite), so instead we use another kernel estimate for $\hat{g}(v)$.

5.3. Inferences about misclassification under Model I

The attractivenes of the semiparametric approach is that consistency of MRC and isotonic regression requires only the SIC condition. Without a particular model of misclassification, however, it is difficult to draw useful inferences about the misclassification or marginal effects. In this section, we discuss the usefulness of the semiparametric approach when Model I holds.

The marginal effects are given by

$$\frac{\partial \Pr(\tilde{y} = 1 | x)}{\partial x} = \frac{g(x'\beta)\beta}{1 - \alpha_0 - \alpha_1}. \tag{27}$$

Even without information on α_0 and α_1, a lower bound (in absolute values) for the marginal effects is given by $g(x'\beta)\beta$. If the misclassification probabilities

α_0 and α_1 are known (or consistent estimates are available from an outside source), marginal effects can be consistently estimated by the semiparametric approach even though the underlying c.d.f. F is unknown.

Inference about α_0 and α_1 is based on the behavior of the response function in its tails. Since $\lim_{v\to-\infty} G(v) = \alpha_0$ and $\lim_{v\to+\infty} G(v) = 1 - \alpha_1$, the left-tail asymptote gives information about α_0 and the right-tail asymptote gives information about α_1. If the data take on extreme values, the step levels from the isotonic regression in the tails can suggest the extent of misclassification. Unfortunately, for any given dataset, one can never know for sure if the index values extend far enough into the tails since the true c.d.f. F is unknown. Response-function estimates near zero in the left tail and near one in the right tail may be evidence against misclassification.[10] Asymptotes above zero in the left tail or below one in the right tail are consistent with misclassification. Estimates of α_0 and α_1 derived from these asymptotes can be thought of as upper bounds on the true misclassification probabilities since the data may not extend far enough into the tails. In the application in Section 6, we use the first and last steps of the isotonic regression estimate to yield information about α_0 and α_1. Finally, we note that if the resulting estimates of the misclassification probabilities are actually upper bounds, call them $\bar{\alpha}_0$ and $\bar{\alpha}_1$, then $g(x'\beta)\beta/(1-\bar{\alpha}_0-\bar{\alpha}_1)$ is an upper bound (in absolute values) for the marginal effects. In practice, it may be worthwhile to report the marginal effects from Eq. (27) for several plausible values of $\alpha_0 + \alpha_1$.

5.4. Other semiparametric estimators

The approach described in this section utilizes monotonicity in both stages of estimation. We believe that the monotonicity condition is weak enough to be applicable in most situations in which misclassification is a concern. If one is worried that condition MC2 does not hold, however, one can still estimate β and the observed response function in Model II semiparametrically. In Model II, y depends on x only through the index $x'\beta$, so that any semiparametric single-index estimator (e.g., the semiparametric least-squares (SLS) estimator of Ichimura (1993)) will estimate β consistently up-to-scale. A specification test between the MRC and SLS estimates could detect failure of condition MC2. For

[10] Of course, even with misclassification, there is the possibility that the lower and upper steps estimated by isotonic regression will be at zero or one, respectively. If the lowest-ranked index values have responses of zero, the initial step will be estimated at zero. Likewise, if the highest-ranked index values have responses of one, the final step will be estimated at one. With misclassification, the probability of seeing k observations in the left tail with a response of zero is at most $(1-\alpha_0)^k$. In fact, this observation could be used to construct a test of the misclassification model. As a practical matter, a few points in either a lower tail at zero or an upper tail at one is not proof against misclassification.

estimation of the response function, a kernel regression of y on the estimated index would suffice.

5.5. Covariate-dependent misclassification

The methods of this paper rely on misclassification being independent of the covariates. Abrevaya and Hausman (1997) discuss estimation techniques when misclassification is possibly covariate dependent. As a simple example, consider a situation where misclassification depends on a single covariate which is a dummy variable. For instance, in a given application, one might expect misclassification probabilities to differ for union and nonunion workers or for males and females. Denote the covariate of interest as x_1, and write $x \equiv (x_1, x_{-1})$ (i.e., x_{-1} are the components of x excluding x_1). In the MLE framework, the likelihood function now has four misclassification parameters (instead of two) since α_0 and α_1 are now functions of x_1. Under mild restrictions on the misclassification parameters (akin to MC1 of Section 3), the entire parameter vector β remains identified. In the semiparametric framework, identification of the coefficient on x_1 must be sacrificed in order to estimate the remaining components of β. In particular, one can estimate the remaining components of β using MRC on those observations having $x_1 = 0$ or on those observations having $x_1 = 1$. The two resulting MRC estimates can be combined for greater efficiency. Specification testing can be used in order to determine whether there is a significant difference between the estimates when misclassification is assumed to be independent of x_1 and the estimates when misclassification is allowed to depend on x_1.

6. Application to a model of job change

We now consider an application where misclassification has previously been considered to be a potentially serious problem. We estimate a model of job change with two widely used datasets, the Current Population Survey (CPS) and the Panel Study of Income Dynamics (PSID). For the cross-sectional CPS, we look at the probability of individuals changing jobs over the past year. For the longitudinal PSID, there are multiple interviews for individuals, allowing us to look at the probability of individuals changing jobs between adjacent interviews. In both cases, questions concerning job tenure are used to construct the relevant job-change binary variable. Since tenure questions are often misunderstood and respondents lack perfect recall, misclassification is a potentially serious problem. We consider parametric models of misclassification, looking at both symmetric and asymmetric probabilities of misclassification. There is evidence of misclassification for both datasets. We also apply the semiparametric techniques discussed in the previous section, finding results quite similar to those found by parametric methods.

6.1. Current Population Survey (CPS)

Our data come from the January 1987 CPS wave from the Census Bureau. We extracted 5221 complete personal records of men between the ages of 25 and 55 whose wages were reported. Among the questions in the survey is one asking for the respondent's job tenure. Those respondents who give tenure as 12 months or fewer are classified as having changed jobs in the last year. Those individuals who answer more than one year are classified as not having changed jobs. The sample statistics are reported in Table 2.

In the first column of Table 3, we report the results from a probit specification. This specification is quite similar to specifications previously used in the applied labor economics literature (e.g., Freeman 1984). In the second column of the table, we estimate a model of symmetric misclassification, restricting $\alpha_0 = \alpha_1$. The estimate of the misclassification probability is 5.8%, with an asymptotic t-statistic of about 8.3. Thus, we reject the probit specification without misclassification. Many of the estimated coefficients also change substantially.

Table 2
Sample statistics for the CPS sample

		mean	std dev
Married	Full sample	0.7293	0.4443
	$y=0$	0.7468	0.4349
	$y=1$	0.6253	0.4844
Grade	Full sample	14.38	2.823
	$y=0$	14.40	2.834
	$y=1$	14.30	2.760
Age	Full sample	37.43	8.526
	$y=0$	37.98	8.535
	$y=1$	34.17	7.712
Union	Full sample	0.2454	0.4303
	$y=0$	0.2668	0.4424
	$y=1$	0.1173	0.3220
Earn per week	Full sample	488.9	240.2
	$y=0$	507.9	235.7
	$y=1$	375.1	235.6
West	Full sample	0.2015	0.4012
	$y=0$	0.1946	0.3960
	$y=1$	0.2427	0.4290

Table 3
CPS coefficient estimates

	Probit	MLE $\alpha_0 = \alpha_1$	MLE $\alpha_0 \neq \alpha_1$	MRC/IR
α_0	—	0.058	0.061	0.035
		(0.007)	(0.007)	(0.015)
α_1	—	0.058	0.309	0.395
		(0.007)	(0.174)	(0.091)
Married	− 0.108	− 0.073	− 0.103	− 0.161
	(0.049)	(0.077)	(0.100)	(0.191)
Last grade attended	0.026	0.063	0.080	0.052
	(0.009)	(0.015)	(0.026)	(0.043)
Age	− 0.022	− 0.028	− 0.033	− 0.035
	(0.003)	(0.005)	(0.007)	(0.021)
Union membership	− 0.434	− 0.707	− 0.811	− 0.794
	(0.061)	(0.148)	(0.199)	(0.503)
Earnings per week	− 0.001	− 0.003	− 0.004	− 0.003
	(0.0001)	(0.0004)	(0.0009)	(0.0015)
Western region	0.214	0.301	0.367	0.367
	(0.054)	(0.086)	(0.127)	(—)
Constant	0.051	0.171	0.581	—
	(0.162)	(0.259)	(0.495)	
Log likelihood	− 1958.1	− 1941.4	− 1940.9	—
Number of obs.	5221	5221	5221	

Note: Standard errors are in parentheses. The MRC coefficient estimates have been normalized to have the same value for western region as the MLE with $\alpha_0 \neq \alpha_1$. There is no associated standard error on 'western region' due to the normalization. The IR estimates are the point estimates from the first and last steps of the isotonic regression step-function estimate.

Next, we allow for asymmetric misclassification since we would expect a priore that non-job changers are less likely to misreport their status. In the third column of Table 3, α_0, the probability of misclassification for non-job changers, is allowed to differ from α_1, the probability of misclassification for job changers. Freeing the misclassification parameters produces a markedly different value for α_1 than in the case where the two are constrained to be equal. α_1 jumps to 0.31 while α_0 remains at 0.06. The difference $(\alpha_1 - \alpha_0)$ is 0.248 with a standard error of 0.164, which is not quite statistically significant.

We now apply the two-step MRC/IR approach detailed in the previous section. The results of using the estimators on the job change data are reported in the last column of Table 3. Since MRC only identifies the ratios of the coefficients, we have scaled the MRC coefficients so that the coefficient estimate on 'western region' is identical to the coefficient estimate for the unconstrained MLE of the third column. A quick comparison of the third and fourth columns shows very little difference between the coefficient estimates found by our adjusted likelihood method and the semiparametric method. The reported MRC/IR estimates of α_0 and α_1 are inferred from the heights of the first and last steps of the estimated step function from the isotonic regression. The standard errors are computed using the method described in the previous section. We find the estimated probabilities of misclassification to be 0.035 and 0.395, not far off from the estimates from the likelihood method. As mentioned above, the accuracy of the upper asymptote depends on the amount of data we have in that index range. Since the number of datapoints in the upper range is small, the estimates of α_1 should perhaps be viewed with some caution. However, it seems to us that the lower asymptote is well established. There are plenty of observations at low values of the index and the lowest step is relatively long.

In Table 4, we investigate the marginal effects associated with the estimates from Table 3. The marginal effects are reported for the first quartile, the mean, and the third quartile of the estimated index. The first column contains the marginal effects for the asymmetric-misclassification MLE estimates, using Eq. (14). The second and third columns contain marginal effects for the MRC/IR estimates, using Eq. (27). The second column uses the estimates of α_0 and α_1 obtained in Table 3. The discussion from the previous section suggests that these marginal-effect estimates can be thought of as upper bound estimates (in magnitude). For a lower bound, the third column uses $\alpha_0 = \alpha_1 = 0$ in

Table 4
CPS marginal effects

	Quartile	MLE $\alpha_0 \neq \alpha_1$	MRC/IR $\alpha_0 + \alpha_1 = 0.43$	MRC/IR $\alpha_0 + \alpha_1 = 0$
Married	1st	− 0.0012	− 0.0032	− 0.0018
	mean	− 0.0073	− 0.0080	− 0.0046
	3rd	− 0.0184	− 0.0389	− 0.0211
Grade	1st	0.0009	0.0010	0.0006
	mean	0.0057	0.0026	0.0015
	3rd	0.0143	0.0127	0.0072
Age	1st	− 0.0004	− 0.0007	− 0.0004
	mean	− 0.0023	− 0.0018	− 0.0010
	3rd	− 0.0058	− 0.0085	− 0.0048
Union	1st	− 0.0092	− 0.0156	− 0.0089
	mean	− 0.0575	− 0.0395	− 0.0225
	3rd	− 0.1447	− 0.1920	− 0.1094
Earnings per week	1st	− 0.0005	− 0.0001	− 0.0003
	mean	− 0.0003	− 0.0001	− 0.0006
	3rd	− 0.0007	− 0.0007	− 0.0003
Western region	1st	0.0042	0.0072	0.0041
	mean	0.0260	0.0182	0.0104
	3rd	0.0655	0.0887	0.0506

Eq. (27). The parametric and semiparametric marginal-effect estimates are relatively similar. Interestingly, many of the parametric estimates fall in between the two reported semiparametric estimates.

In Fig. 1, we plot the estimated response function from the MLE model with asymmetric misspecification and also the MRC/IR estimate of the response function. The results are reasonably similar. Fig. 2 shows the MRC/IR estimate of the c.d.f. with a (pointwise) confidence interval of two standard errors in either direction. The confidence interval demonstrates that the IR step function estimate of the c.d.f. is estimated accurately, except when the size of the step function becomes small. In these situations, we smooth the estimation of the confidence interval. We also include a comparison of the MRC/IR estimate with a standard kernel estimate. Fig. 3a uses a fixed-window kernel; this method is problematic for observations at the ends of the distribution because there are only observations on one side of the point. One can see that the kernel estimate becomes nonmonotonic at the upper tail of our data. Fig. 3b tries an alternative approach by using the 200 nearest neighbors to construct the kernel estimate. Again, the upper tail of the c.d.f. is quite different from the MRC/IR estimate. Kernel regression techniques do not seem well-suited to estimating the asymptotes of a c.d.f. where data become sparse.

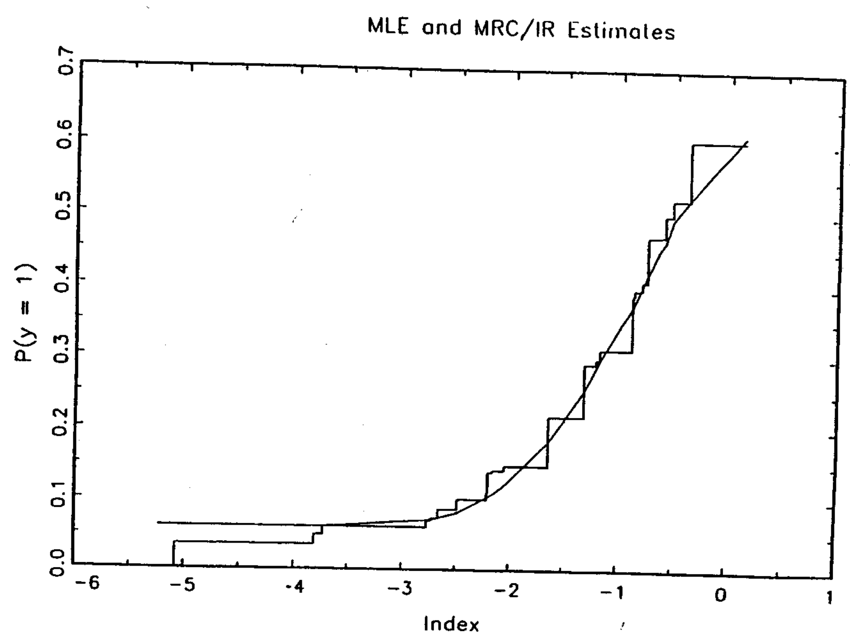


Fig. 1. MLE and MRC/IR estimates.

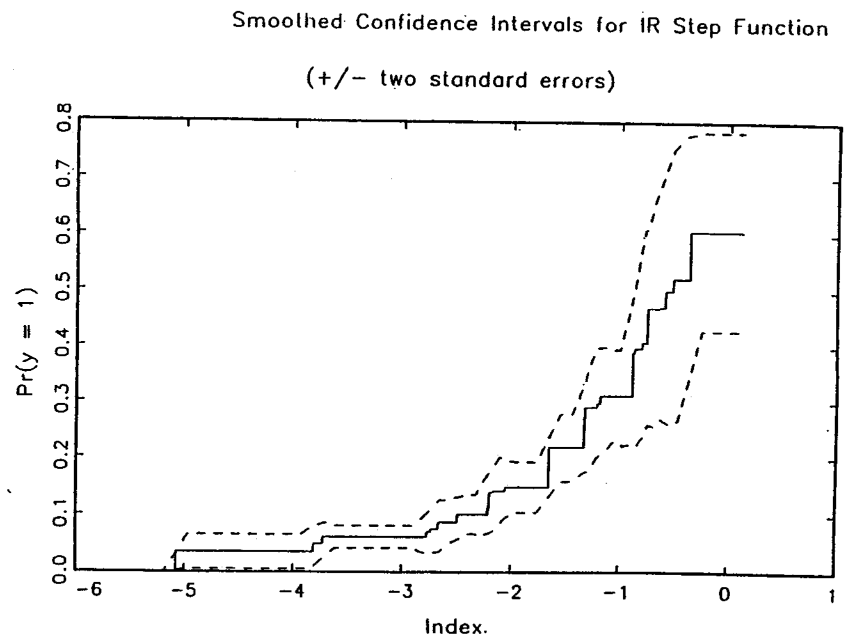


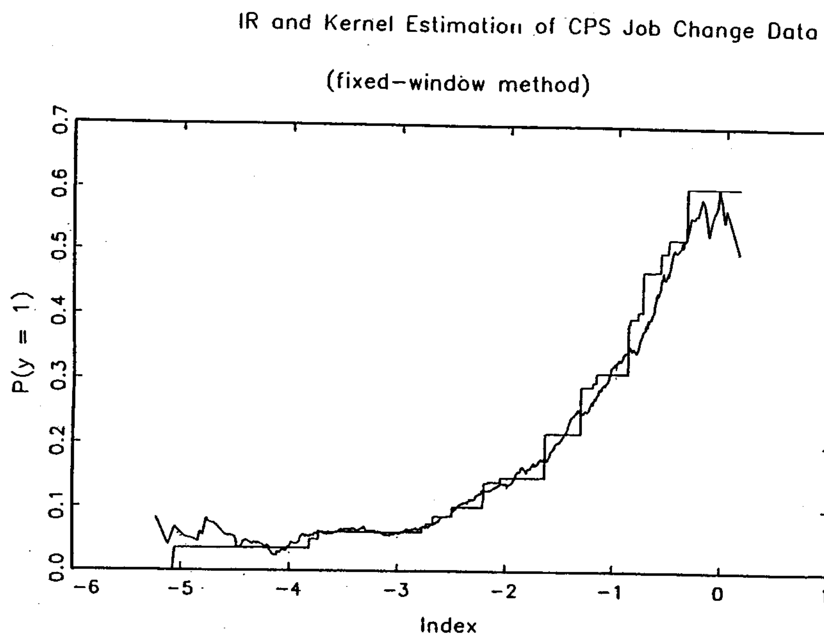
Fig. 2. Smoothed confidence intervals for IR step function (+ / − two standard errors).

6.2. Panel Study of Income Dynamics (PSID)

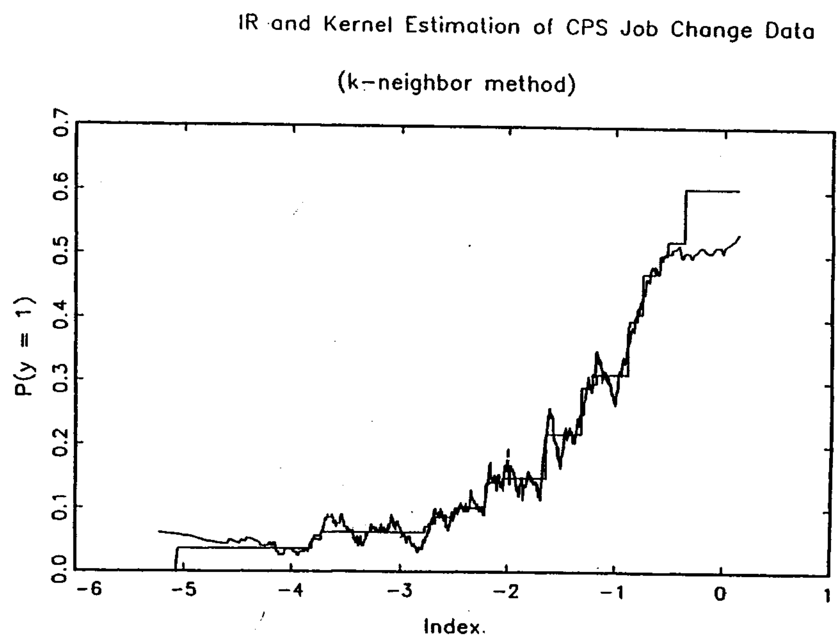
Brown and Light (1992) document the internal inconsistency of tenure responses using information from successive interviews of the PSID. Fewer than 10% of the observations in their sample contain responses which are in complete concordance over successive interviews on questions concerning job tenure. The authors suggest several different definitions of 'job separation' and (in Table 6 of their paper) report logit estimates of the probability of job separation (using demographic and work-related variables as covariates).

We follow the same methodology of Brown and Light (1992) to analyze 1981–87 PSID data.[11] Two different methods are used to construct the 'job separation' dummy, which takes on a value of one if a job separation has

[11] We were unable to obtain the 1976–85 abstract used by Brown and Light (1992). Since the 1979 and 1980 questionnaires do not contain questions pertaining to job or employer tenure, we decided to use only post-1980 responses so that, for instance, the job separation variable would not be constructed using 'adjacent' 1978 and 1981 interviews.



(a)



(b)

Fig. 3. (a). IR and kernel estimation of CPS job change data (fixed-window method). (b). IR and kernel estimation of CPS job change data (*k*-neighbor method).

Table 5
Sample statistics for the PSID sample

	mean	std dev
White	0.65	0.48
Male	0.81	0.39
Education	12.34	2.52
Age	36.57	11.28
Married	0.69	0.46
Poor health	0.07	0.25
Smsa	0.63	0.48
Union	0.24	0.43
ln(weekly wage)	6.73	0.51

occurred between interviews and zero otherwise.[12] The first method (called 'partition *T*') assumes that a job separation has occurred whenever reported tenure is less than the elapsed time since the previous interview. This approach, then, is quite similar to the approach employed above for the CPS. The second method (called 'partition 6') assumes that job separation has occurred if the change in reported tenure from one interview to the next differs from elapsed calendar time by more than six months (in either direction). For ease of notation, let y_T and y_6 denote the dummies created by 'partition *T*' and 'partition 6', respectively.

Our sample consists of household heads between 1981 and 1987 who are either employed or on temporary layoff at the time of interview. Part-time workers, government workers, workers with multiple jobs, and self-employed workers were discarded. Also, interviews in which any of the relevant data were missing were dropped. Since job separation is defined using adjacent interviews, the relevant observational unit is an adjacent-interview pair. The sample consists of 8674 such observations. Descriptive statistics for the sample are given in Table 5. Most of the variables are self-explanatory. The 'poor health' variable is equal to one if the individual has a 'physical or nervous condition' that limits their amount of work. The 'smsa' variable is equal to one if the individual lives in an SMSA whose largest city has a population greater than 50 000.

A comparison of the two methods for defining job separation is shown in Table 6. 'Partition 6' results in far more instances of job separation than

[12] Brown and Light (1992) use seven different methods. We report the results from two of these methods ('partition *T*' and 'partition 6' in their terminology). The results from the other five methods are available upon request from the authors, but they do not contribute anything extra to the discussion of misclassification. In particular, the 'partition *P*' results are quite similar to the 'partition *T*' results, and the results from the remaining four partitions are quite similar to the 'partition 6' results.

Table 6
Comparison of job-change variables for the PSID sample ('partition T' vs. 'partition 6')

	y_T		
		0	1
y_6	0	4954	52
	1	2655	1013

'partition T' (3668 vs. 1065). The two methods agree for 68.8% of the observations. The vast majority of disagreement occurs for observations having $y_T = 0$ and $y_6 = 1$ rather than vice versa.

The parametric estimation results are reported in Table 7. For each partition, we again estimate a probit model, a symmetric misclassification MLE, and an asymmetric misclassification MLE. The first three columns report the results for 'partition T' and the last three columns for 'partition 6'. Unlike our finding for the CPS, we strongly reject the hypothesis that $\alpha_0 = \alpha_1$ using a likelihood ratio test of the symmetric vs. asymmetric misclassification models. Misclassification for nonjob changers is estimated to be over 20% for both partitions, whereas misclassification for job changers is around 1% for 'partition T' as compared to 29% for 'partition 6'.

To check the specification of the asymmetric misclassification probit model against a semiparametric alternative, we report the results of MRC estimation for both partitions in Table 8. We have normalized the coefficient estimates so that ln(wage) has a coefficient of -1 across the columns. The first and second columns show the MLE and MRC estimates for 'partition T', and the third and fourth columns show the MLE and MRC estimates for 'partition 6'. The standard errors for the MLE estimates are computed using the delta method, whereas the standard errors for the MRC estimates are computed using the method of Cavanagh and Sherman (1997). As can be readily seen, the MLE and MRC estimates do not differ by much. In fact, a Hausman test of their difference fails to reject the parametric specification for both partitions (test statistic of 2.2 for 'partition T' and 1.5 for 'partition 6' with 13 degrees of freedom).

We perform an isotonic regression for each partition as we did for the CPS sample. Fig. 4a and b graph the step functions with confidence intervals for 'partition T' and 'partition 6', respectively. Both step functions appear to flatten out at around 80%. And the step function for 'partition 6' seems to have a long flat region near 30% in the left tail. To further investigate this 'flattening', we find it extremely helpful to re-draw the step functions estimates in a different way, putting the rank of the observation rather than the index value of the observation on the x-axis. This approach allows the researcher to see exactly how many datapoints are along each particular step of the step function estimate. Fig. 5a and b display the step functions in this manner. Immediately,

Table 7
Parametric coefficient estimates for the PSID sample

	Partition T			Partition 6		
	Probit	MLE $\alpha_0 = \alpha_1$	MLE $\alpha_0 \neq \alpha_1$	Probit	MLE $\alpha_0 = \alpha_1$	MLE $\alpha_0 \neq \alpha_1$
α_0		0.0068	0.2533		0.2413	0.2017
		(0.0062)	(0.0702)		(0.0331)	(0.0309)
α_1		0.0068	0.0135		0.2413	0.2895
		(0.0062)	(0.0065)		(0.0331)	(0.0209)
White	0.2343	0.2487	0.3207	– 0.1651	– 0.3895	– 0.4372
	(0.0440)	(0.0490)	(0.0680)	(0.0320)	(0.1046)	(0.1047)
Male	0.2861	0.2979	0.3324	0.3426	0.8293	0.9901
	(0.0665)	(0.0707)	(0.0880)	(0.0525)	(0.2176)	(0.2081)
Educ	0.0198	0.0226	0.0307	– 0.0184	– 0.0470	– 0.0573
	(0.0098)	(0.0108)	(0.0137)	(0.0069)	(0.0204)	(0.0228)
Age	– 0.0252	– 0.0271	– 0.0317	0.0039	0.0082	0.0071
	(0.0021)	(0.0029)	(0.0039)	(0.0013)	(0.0034)	(0.0041)
Married	– 0.1418	– 0.1505	– 0.1631	0.0060	0.0526	0.0568
	(0.0548)	(0.0580)	(0.0711)	(0.0436)	(0.1026)	(0.1181)
Poorhlth	0.0720	0.0875	0.0772	0.0290	0.0285	0.0029
	(0.0778)	(0.0831)	(0.1013)	(0.0563)	(0.1377)	(0.1675)
Smsa	0.1215	0.1296	0.1701	0.0118	0.0236	0.0384
	(0.0414)	(0.0452)	(0.0583)	(0.0302)	(0.0720)	(0.0874)
Union	– 0.3573	– 0.4012	– 0.5163	– 0.0996	– 0.2365	– 0.2726
	(0.0551)	(0.0735)	(0.1071)	(0.0346)	(0.0958)	(0.1145)
Wage	– 0.4946	– 0.5301	– 0.6533	– 0.2877	– 0.6883	– 0.8472
	(0.0495)	(0.0621)	(0.0934)	(0.0368)	(0.1715)	(0.1732)
Const	1.6623	1.8425	2.5150	1.0888	2.6085	3.2729
	(0.2756)	(0.3371)	(0.4886)	(0.2052)	(0.8341)	(0.9232)
Log likelihood	– 2698.6	– 2679.9	– 2693.7	– 5637.5	– 5630.9	– 5626.9
LR test vs. probit		1.30	9.71		13.08	21.09
(d.f.)		(1)	(2)		(1)	(2)

Note: Four year dummies and elapsed time between adjacent interviews were also included as independent variables. The estimates are not reported in the interest of saving space. Standard errors are report in parentheses.

we see that the flattening at 80% is not as severe as one might think at first glance of Fig. 4 a and b. However, the flat region at 30% for 'partition 6' is highlighted even more by this method. Nearly 2000 observations (accounting for almost a quarter of the entire sample) fall on this step. The height of this step is remarkably close to the estimate of 0.29 found by our likelihood method (see Table 8).

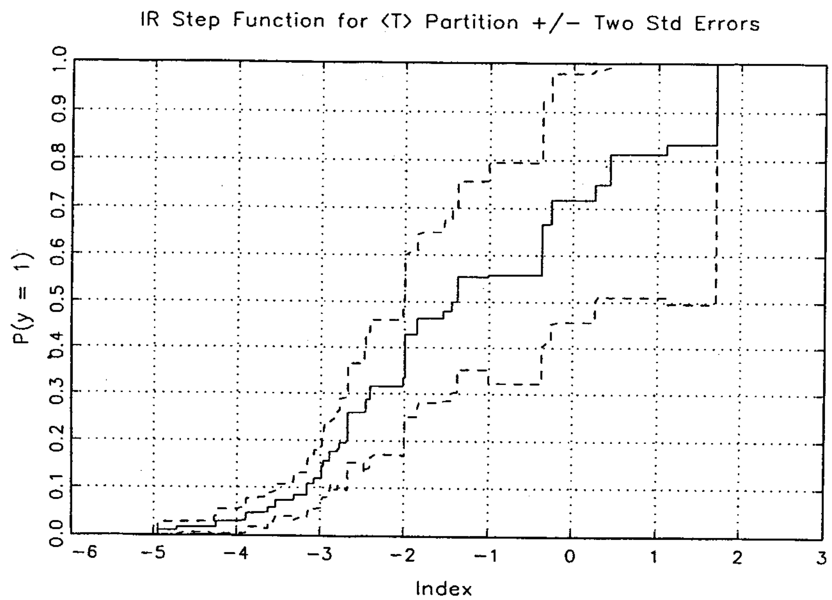
Table 8
Comparison of MLE and MRC coefficients for the PSID sample

	Paritition T MLE	MRC	Partition 6 MLE	MRC
White	0.4894	0.5570	− 0.5161	− 0.7497
	(0.0907)	(0.2049)	(0.1303)	(0.4331)
Male	0.5073	0.4948	1.1687	1.5327
	(0.1316)	(0.2892)	(0.2166)	(0.6282)
Educ	0.0468	0.0472	− 0.0676	− 0.0876
	(0.0182)	(0.0418)	(0.0279)	(0.0763)
Age	− 0.0485	− 0.0445	0.0084	0.0184
	(0.0073)	(0.0169)	(0.0050)	(0.0120)
Married	− 0.2489	− 0.2375	0.0670	− 0.0783
	(0.1122)	(0.2406)	(0.1383)	(0.3654)
Poorhlth	0.1178	0.1204	0.0034	0.1406
	(0.1557)	(0.3519)	(0.1978)	(0.5033)
Smsa	0.2596	0.3097	0.0453	0.0008
	(0.0841)	(0.1845)	(0.1025)	(0.2592)
Union	− 0.7880	− 0.8770	− 0.3218	− 0.3770
	(0.1624)	(0.3994)	(0.1458)	(0.3471)
ln(wage)	− 1.0000	− 1.0000	− 1.0000	− 1.0000
	—	—	—	—
Hausman test		2.2033		1.5206
(d.f.)		(13)		(13)

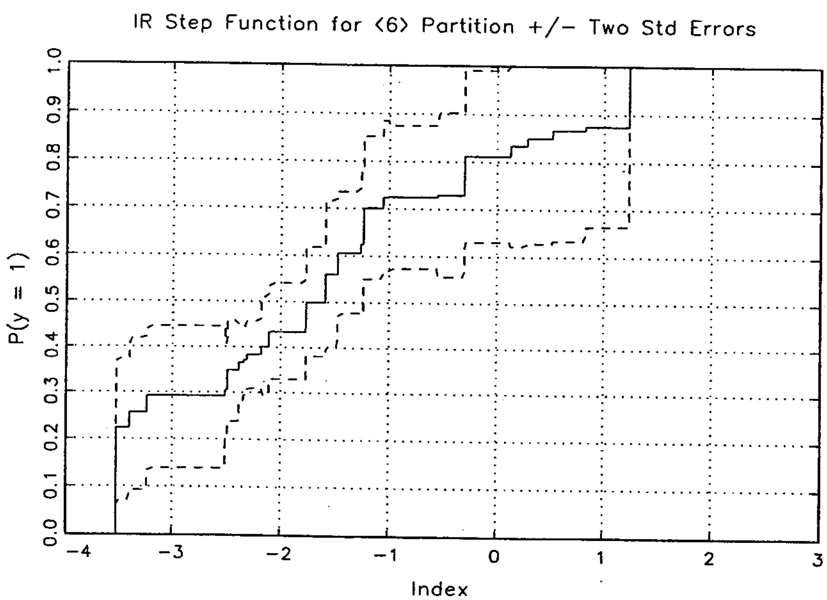
Note: Standard errors are reported in parentheses. Coefficients are scaled so that ln(wage) has a coefficient of − 1. The standard errors have been adjusted appropriately.

7. Conclusion

This paper has shown that ignoring potential misclassification of a dependent variable can result in biased and inconsistent coefficient estimates when using standard parametric specifications. The researcher can use the adjusted maximum likelihood procedure described in Section 3 to consistently estimate the extent of misclassification and the coefficients. However, should the error disturbances in the data not have the assumed parametric distribution, these coefficient estimates may nevertheless be inconsistent. Semiparametric regression using the MRC estimator of Han (1987) yields consistent estimates of the coefficients without specifying the error distribution. The MRC estimates are also consistent for a more flexible model of misclassification than the parametric estimates. Furthermore, the IR techniques detailed above provide pointwise consistent estimates of the response function. Due to the different convergence rates of MRC and IR estimation, we are able to derive and estimate the asymptotic distribution for both the slope parameters and the response function estimates.

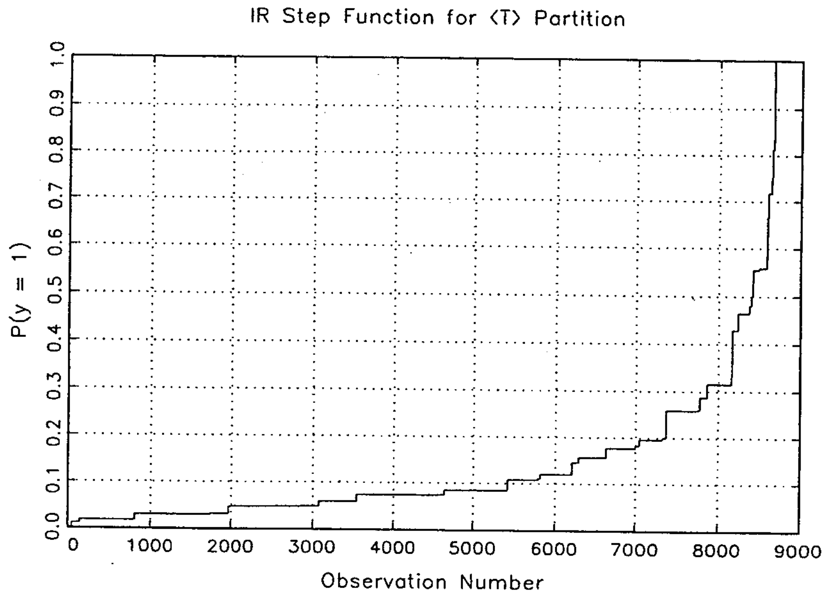


(a)

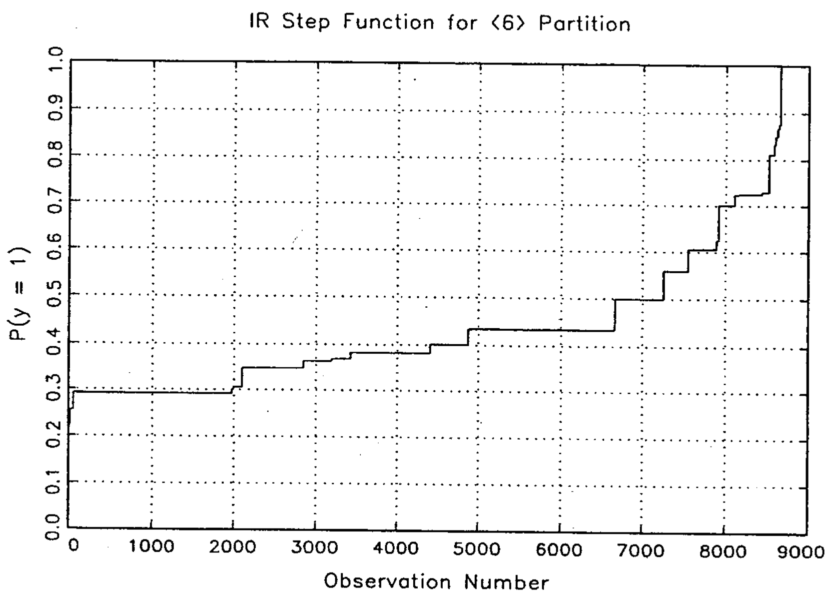


(b)

Fig. 4. (a). IR step function for ⟨T⟩ partition + / − two standard errors. (b). IR step function for ⟨6⟩ partition + / − two standard errors.



(a)



(b)

Fig. 5. (a). IR step function for $\langle T \rangle$ partition. (b). IR step function for $\langle 6 \rangle$ partition.

Applying our econometric techniques to job-change data from the CPS and PSID, we find that serious misclassification exists when we construct the job-change variable in very natural ways. Furthermore, the probabilities of misclassification differ depending on the response. We find reasonably similar results using both the MLE parametric approach and the distribution-free semiparametric approach. We find the isotonic regression to be quite useful in allowing the researcher to view features of the underlying response function. Our approach is quite straightforward to use on discrete-response models that are commonly used in applied research. Thus, we hope our approach will be useful to others working with discrete data for dependent variables, especially in probit and logit models.

Other model specification problems may exist besides misclassification, e.g., heterogeneity or heteroscedasticity. Misclassification in particular need not be the problem in a case where a probit or logit model does not fit. However, the types of results we achieve here suggest that misclassification can be a serious problem. Results that suggest the existence of misclassification certainly justify looking more closely at the data to determine what error structure does exist.

Acknowledgements

We would like to thank Yacine Ait-Sahalia, Gary Chamberlain, Stephen Cosslett, Zvi Griliches, Bo Honoré, Whitney Newey, Peter Phillips, Tom Stoker, Luis Ubeda-Rivas, and participants of the Harvard–MIT Econometrics Workshop for helpful comments. The suggestions of two anonymous referees and an associate editor greatly improved this paper. Hausman and Abrevaya thank the NSF for research support. Abrevaya also acknowledges financial support from the University of Chicago GSB faculty research fund.

Appendix A.

Proof of Theorem 3. Some additional notation is required. Let $\hat{G}_b(\cdot)$ denote the functional estimate from an isotonic regression of y on $x'b$ (where b is normalized to have its last component equal to one or minus one). Then, from the notation in Section 3, we have $\hat{G}(\cdot) = \hat{G}_\beta(\cdot)$. Let $H_b(v)$ be the cumulative sum of responses for index values less than v, defined as follows:

$$H_b(v) \equiv \sum_{i=1}^{n} 1(x_i'b < v)y_i. \tag{A.1}$$

Finally, let $U_b(a)$ be defined as follows:

$$U_b(a) \equiv \sup\{v\colon H_b(v) - av \text{ is minimal}\}. \tag{A.2}$$

Then, Groeneboom (1985) shows that, with probability one,

$$G_b(v) \leqslant a \Leftrightarrow U_b(a) \leqslant v. \tag{A.3}$$

The usefulness of this relationship is that the (inverse) process $U_b(a)$ is more tractable than $G_b(v)$.

We have

$$n^{1/3}(\hat{G}_{\beta}(v) - G(v)) = n^{1/3}(\hat{G}_{\beta}(v) - \hat{G}_{\beta}(v)) + n^{1/3}(\hat{G}_{\beta}(v) - G(v)). \tag{A.4}$$

The second term has known asymptotic distribution, given in Section 3. To show that the first term converges in probability to zero, we show that $n^{1/3}(U_{\beta}(a) - U_{\beta}(a))$ converges in probability to zero. Note that

$$U_{\beta}(a) = \sup\{v\text{: } (H_{\beta}(v) - H_{\beta}(v)) + (H_{\beta}(v) - av) \text{ is minimal}\}. \tag{A.5}$$

For a given v, we have

$$\begin{aligned} |H_{\beta}(v) - H_{\beta}(v)| &= \left|\sum_{i=1}^{n} (1(x_i'\hat{\beta} < v) - 1(x_i'\beta < v))y_i\right| \\ &= \left|\sum_{i=1}^{n} (1(x_i'\hat{\beta} < v < x_i'\beta) - 1(x_i'\beta < v < x_i'\hat{\beta}))y_i\right| \\ &\leqslant \sum_{i=1}^{n} 1(\min(x_i'\hat{\beta}, x_i'\beta) < v < \max(x_i'\hat{\beta}, x_i'\beta))y_i. \end{aligned}$$

Since $\sqrt{n}|\hat{\beta} - \beta| = \mathrm{O_p}(1)$ (for the components of the parameter vector not fixed by the normalization), it follows that $|H_{\beta}(v) - H_{\beta}(v)| = \mathrm{o_p}(n^{-1/3})$ at any v for which the index $x'\beta$ has a positive density in a neighborhood around v. Thus, the $(H_{\beta}(v) - H_{\beta}(v))$ term in Eq. (A.5) is negligible in determining the asymptotic behavior of $n^{1/3}(U_{\beta}(a) - U_{\beta}(a))$, which implies that $n^{1/3}(U_{\beta}(a) - U_{\beta}(a)) = \mathrm{o_p}(1)$ and completes the proof. □

References

Abramowitz, M., Stegun, I.A., 1964. Handbook of Mathematical Functions. U.S. Government Printing Office, Washington, DC.

Abrevaya, J., Hausman, J.A., 1997. Semiparametric estimation with mismeasured dependent variables: an application to panel data on unemployment spells, mimeo.

Barlow, R.E., Bartholomew, D.J., Bremner, J.M., Brunk, H.D., 1972. Statistical Inference under Order Restrictions. Wiley, London.

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

Cavanagh, C., Sherman, R.P., 1997. Rank estimators for monotonic index models. Journal of Econometrics, forthcoming.

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

Cosslett, S.R., 1983. Distribution-free maximum likelihood estimator of the binary choice model. Econometrica 51, 765–782.

Finney, D.J., 1964. Statistical Method in Biological Assay. Havner, New York.

Freeman, R.B., 1984. Longitudinal analyses of the effects of trade unions. Journal of Labor Economics 2, 1–26.

Greene, W.H., 1990. Econometric Analysis. Macmillan, New York.

Groeneboom, P., 1985. Estimating a monotone density. In: LeCam, L.M., Olshen, R.A. (Eds.), Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer, vol. II. Wadsworth, Hayward, CA.

Groeneboom, P., 1993. Asymptotics for incomplete censored observations, mimeo.

Han, A.K., 1987. Non-parametric analysis of a generalized regression model: the maximum rank correlation estimator. Journal of Econometrics 35, 303–316.

Hausman, J.A., 1978. Specification tests in econometrics. Econometrica 46, 1251–1272.

Hausman, J.A., Scott-Morton, F.M., Abrevaya, J., 1996. Misclassification of the dependent variable in a discrete-response setting, mimeo.

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

Manski, C.F., 1985. Semiparametric analysis of discrete response: asymptotic properties of the maximum score estimator. Journal of Econometrics 27, 313–333.

Manski, C.F., 1988. Identification of binary response models. Journal of the American Statistical Association 83, 729–738.

McFadden, D., 1984. Econometric analysis of qualitative response models. In: Griliches, Z., Intriligator, M.D. (Eds.), Handbook of Econometrics, vol. 2. North–Holland, Amsterdam.

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

Poterba, J.M., Summers, L.H., 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., Stock, J.H., Stoker, T.M., 1989. Semiparametric estimation of index coefficients. Econometrica 57, 1403–1430.

Robertson, T., Wright, F.T., Dykstra, R.S., 1988. Order Restricted Statistical Inference. Wiley, London.

Ruud, P.A., 1983. Sufficient conditions for the consistency of MLE despite misspecification of distribution in multinomial discrete choice models. Econometrica 51, 225–228.

Sherman, R.P., 1993. The limiting distribution of the maximum rank correlation estimator. Econometrica 61, 123–137.