working paper
department
of economics

MISCLASSIFICATION OF A DEPENDENT VARIABLE IN A DISCRETE RESPONSE SETTING

Jerry A. Hausman
Fiona M. Scott Morton

94-19 June 1994

massachusetts
institute of
technology

50 memorial drive
cambridge, mass. 02139

MISCLASSIFICATION OF A DEPENDENT VARIABLE IN A DISCRETE RESPONSE SETTING

Jerry A. Hausman
Fiona M. Scott Morton

94-19 June 1994

First Draft December 1993

Misclassification of a Dependent Variable in a Discrete Response Setting

Jerry A. Hausman[1]
Fiona M. Scott Morton

MIT and NBER

June 1994

[1] We would like to thank Y. Ait-Sahalia, G. Chamberlain, S. Cosslett, Z. Griliches, B. Honore, W. Newey, P. Phillips, T. Stoker, and L. Ubeda Rives, for helpful comments. Thanks to Jason Abrevaya for outstanding research assistance. Hausman thanks the NSF for research support.
Address for correspondence: E52-271A, Department of Economics, MIT, Cambridge, MA, 02139.

Introduction

A dependent variable which is a discrete response causes the estimated coefficients to be inconsistent in a probit or logit model when misclassification is present. 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 interviewer simply checks the wrong box. Other data sources where the researcher suspects measurement error, such as historical data, certainly exist as well. We show that when a dependent variable is misclassified in a probit or logit setting, the resulting coefficients are biased and inconsistent. However, the researcher can correct the problem by employing the likelihood function we derive below, and can explicitly estimate the extent of misclassification in the data. We also discuss a semi-parametric method of estimating both the probability of misclassification and the unknown slope coefficients which does not depend on an assumed error distribution and is also robust to misclassification of the dependent variable. Each of these departures from the usual qualitative response model specifications creates inconsistent estimates.

We apply our methodology to a commonly used data set, the Current Population Survey, where we consider the probability of individuals changing jobs. This type of question is well-known for its potential misclassification. Both our parametric and semiparametric estimates demonstrate conclusively that significant misclassification exists in the sample. Furthermore, the probability of misclassification is not the same across observed response classes. A much higher probability exists for misclassification of reported individual job changes than the probability of misclassification of individuals who are reported not to have changed jobs.

I. Qualitative Response Model with Misclassification

We use the usual latent variable specification of the qualitative choice model; for the present we consider the binomial response model, c.f. Greene (1990) or MacFadden (1984). Let y_i^* be the latent variable:

$$y_i^* = X_i\beta + \epsilon_i \qquad (1)$$

The observed data correspond to:

$$y_i = 1 \qquad \text{if } X_i\beta + \epsilon_i \geq 0 \tag{2}$$
$$y_i = 0 \qquad \text{if } X_i\beta + \epsilon_i < 0$$

For now, let π be the probability of correct classification. Assume that π is independent of X and constant in the sample for both types of responses. The model specification follows from the assumption that ϵ_i is distributed as N(0,1), the usual probit assumption:

$$pr(y_i=1|X) \quad = \pi \cdot pr(y_i^* > 0) + (1-\pi) \cdot pr(y_i^* \leq 0)$$

$$= \pi \cdot \Phi(X_i\beta) \; + \; (1-\pi) \cdot (1-\Phi(X_i\beta)) \tag{3}$$

where $\Phi(.)$ is the standard normal cumulative function. We then calculate the expectation of y_i:

$$E(y_i|X) \quad = 1-\pi + (2\pi-1) \cdot \Phi(X_i\beta)$$

This equation can be estimated consistently in the form,

$$y_i \quad = \alpha \; + \; (1-2\alpha) \cdot \Phi(X_i\beta) + \eta_i \tag{4}$$

where $\alpha = 1-\pi$.

Alternatively, we could approach the problem using the other response,

$$pr(y_i=0|X) \quad = \pi \cdot pr(y_i^* \leq 0) + (1-\pi) \cdot pr(y_i^* > 0) \tag{5}$$

$$= \pi - (2\pi-1) \cdot \Phi(X_i\beta)$$

In terms of α:

$$= (1-\alpha) - (1-2\alpha) \cdot \Phi(X_i\beta)$$

Note that as above we find:

$$E(y_i|X) \quad = 1 - \{\pi + (1-2\pi)\cdot\Phi(X_i\beta)\} \tag{6}$$

If one estimates a probit specification as if there were no measurement error when in fact misclassification is present, the estimates will be biased and inconsistent. This result is in contrast to a linear regression where classical measurement error in the dependent variable leaves the coefficient estimates consistent, but leads to reduced precision in coefficient estimates.

The log likelihood for the probit specification with misclassification is as follows:

$$L = \Sigma_i \{y_i\cdot \ln[\alpha+(1-2\alpha)\cdot\Phi(X_i\beta)] + (1-y_i)\cdot\ln[(1-\alpha)+(2\alpha-1)\cdot\Phi(X_i\beta)]\} \tag{7}$$

In the case of $\alpha=0$ there is no classification error and the log likelihood will collapse to the usual case. In order for the equation to be estimated, α must be less than one half.[2] Pratt (1981) shows that if $\alpha=0$ this log likelihood is everywhere concave, being the sum of two concave functions. Unfortunately, the result does not hold for our log likelihood for either the logit or probit functional forms. The appendix gives details of the conditions for concavity. We also demonstrate in the appendix that the expected Fisher information matrix is not block diagonal in the parameters (α,β).[3]

Notice that the linear probability model does not allow separate identification of the coefficients β and π. Instead the estimated coefficients are linear combinations of π and β.

$$\begin{aligned} pr(y_i=1|X) &= \pi\cdot pr(y_i^*>0) + (1-\pi)\cdot pr(y_i^*\leq 0) \\ &= \pi\cdot X_i\beta + (1-\pi)\cdot(1-X_i\beta) \end{aligned} \tag{8}$$

---

[2] In the case of $\alpha>.5$, the data are so misleading that the researcher might want to abandon the project.

[3] Thus, previous papers which assume they know α from exogenous sources, e.g., Poterba and Summers (1993) suffer from two defects. Their estimates are likely to be inconsistent unless their assumed α is correct. But even with a correct α the standard errors of their coefficient estimates are inconsistent. In certain special 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.

$$E(y_i|X) \quad = [1-\pi + (2\pi-1)\cdot\beta_0] + (2\pi-1)\cdot(X_i\beta_1)$$

This example show that identification of the true coefficients in probit and logit comes from their non-linearity. To address this concern, we shall introduce a semiparametric approach so other results do not depend on distribution assumptions.

An interesting feature of the misclassification problem is that inconsistency can be large even for a small amount of misclassification. The usual probit first order condition is written below. The first term is summed over observations where y=1, the second for those where y=0,

$$y\cdot\sum_i \frac{\phi_i X_i}{\Phi_i} - (1-y)\cdot\sum_j \frac{\phi_j X_j}{1-\Phi_j} = 0 \tag{9}$$

where $\Phi_i = \Phi(X_i\beta)$, and similarly for j. The intuitive reasoning for the inconsistency due to misclassification is as follows. When misclassified observations are present, they are added into the likelihood and the score through the incorrect term (because the dependent variable is misclassified). The large inconsistency in the estimated parameters arises if probit (or logit) is used because misclassified observations will predict the opposite result from that actually observed. For example, an observation with a large index value will predict a one with probability $\Phi(X_i\beta)$ close to one. However, if that observation is misclassified, its observed *y* value will be a zero. The observation will be included in the second term in equation (9); a $\Phi(X_i\beta)$ close to one will cause the denominator to approach zero, so the whole term will approach infinity. The same problem exists for observations with very low index values. Therefore, the sum of first order conditions in the case of misclassification can become large, and the estimated β's can be strongly inconsistent as a result. It is somewhat ironic that the inconsistency will increase as the model gets "better" because there will be more good fits that are misclassified and therefore more large terms in the log likelihood.[4]

Another way to think about the inconsistency caused by misclassification is to use the fact that maximum likelihood sets the score equal to zero under the situation of no misclassification.

---

[4] See Table Ib for a simulated demonstration of this property.

Taking expectations and using the probabilities Φ_i and $1\text{-}\Phi_i$ in the case of the normal probit causes the denominators to cancel and the expectation to equal zero. However, in the case of misclassification, the probabilities change. The "ones term" is used for observations that have y=1 and are not misclassified and also for observations that have y=0 and are misclassified:

$$\frac{\phi X}{\Phi}\ [(1-\alpha)\Phi + \alpha(1-\Phi)]\ -\ \frac{\phi X}{1-\Phi}\ [(1-\alpha)(1-\Phi) + \alpha\Phi]\ =\ 0 \tag{10}$$

When no misclassification is present, $\alpha = 0$ and the expectation of the equation above collapses to the usual probit case. If $\alpha \neq 0$, then the expectation of the score is, in general, not equal to zero.

We are particularly interested in the inconsistency in probit or logit coefficients when only small amounts of misclassification are present, since this might be the most common case facing a researcher. We can use the modified score above to evaluate the change in the estimated coefficients with respect to the extent of misclassification at zero misclassification. The full derivation is in the appendix. In the case of the probit, the derivative will equal:

$$\frac{\partial\beta_k}{\partial\alpha}\Big|_{\alpha=0,F(.)=\Phi} = \frac{1\ -\ 2\Phi(X_i\beta)}{\phi(X_i\beta)X_i} \tag{11}$$

Again, notice that in the case where an observation has $\Phi(X_i\beta)$ close to either zero or one, then $\phi(X_i\beta)$ goes to zero also. Therefore, even a single observation can theoretically add considerable inconsistency to the estimated β's. In general, the amount of inconsistency in the estimated β's when misclassification is present will depend on the distribution of X's.[5] Table A reports the value of this derivative for the job change dataset used later in this paper. The crucial feature of misclassification is immediately apparent. The "maximum derivative" row of the table shows that the bias in β caused by a specific misclassified observation can be very

[5] Distributions such as the uniform lead the average derivative to be relatively small, whereas a distribution with unbounded support will cause it to be larger.

large. Averaged across observations, the change in β is not nearly so large, indicating that particular observations contribute heavily to the inconsistency of the estimated β.

Table A:
Derivative of estimated β with respect to the fraction of misclassification in the dataset, evaluated at zero misclassification using *Job change* dataset from the CPS.[6]

N=5221	Married	Grade	Age	Union	Earnings	West
Average Derivative	519	65.5	19.5	535	0.961	839
Max Derivative	179,666	27,640	7,486	179,666	360	359,332
Min Derivative	-3.60	-0.200	-0.136	-3.60	-1.70	-3.48

II. Simulation

In order to assess the empirical importance of misclassifications, we create a sample using random number generators. The first right hand side variable, X_1, is drawn from a lognormal distribution, the second, X_2, is a dummy variable that takes the value one with a one third probability, and X_3 is distributed uniformly. ϵ is drawn from a normal (0,1) distribution. y_i^* and y_i are defined to be:

$$y_i^* = \beta_0 + X_{1i}\beta_1 + X_{2i}\beta_2 + X_{3i}\beta_3 + \epsilon_i \;=\; X_i\beta + \epsilon_i \tag{12}$$

$$y_i = \begin{cases} 1 & \text{if } y_i^* > 0 \\ 0 & \text{otherwise} \end{cases}$$

Next we create a misclassified version of y_i called $\tilde{y}_i$. A proportion α of the sample (recall that $\alpha=1-\pi$) is misclassified on a random basis.

[6] The dummy variables were rescaled to equal 1 and 2 and the regressions rerun in order to compute the derivative.

Table I reports the Monte Carlo results. We create the sample and estimate the MLE coefficients 146 times. For comparison, the first column reports the parameters that would be estimated if ordinary probit were run on these misclassified data. The sample design used in the first row has 5% misclassification; the second row has 20%. Even in the case of a small amount of misclassification, ordinary probit can produce very inconsistent coefficients. The problem is intensified as the amount of misclassification grows. As discussed above, this is likely to be a more severe problem when the model fits well, as it does here. The likelihood function we propose does very well at estimating the "true" coefficients and α. The variance of the estimates rises as the extent of misclassification in the sample increases. Table II reports estimates from one random sample generation only. It includes a column that estimates the β assuming α is already known. Again, MLE does well, and standard errors are lower than in the case where α is estimated. However, if α is not truly known, but treated as if it is known, e.g. Poterba and Summers (1993), these estimated standard errors will be downward biased. In the fourth column are the MLE estimates of the misclassification error parameter, α, as well as estimates of β. Notice again that the estimates of β are closer to the true parameters than the estimates obtained without correcting for likely misclassification error (column 2).

As previously noted, the standard errors in the probit specification using misclassified data are relatively small. The researcher depending on the incorrectly estimated model will think that the reported coefficients are precisely estimated, although misclassification is actually a problem in the data. Although the MLE method that accounts for misclassification may not be able to estimate very precise coefficients if the model does not fit well, it will still give standard errors that reflect the true imprecision of the estimates. Also, the estimates of α are quite precisely estimated in cases where the data fit the model well, as in Table II. Not just the absolute size of the β's, but their ratios vary from the true to the inconsistent case.[7]

We test our theoretical model two additional ways. When there is no misclassification

---

[7] The ratios of the betas will stay constant if the simulated Xs are drawn from normal distributions, even when each individual beta is biased. The appendix includes a table which reports the results of simulations run using normally distributed Xs. Ruud (1983) discusses reasons for this result. The ratio of the estimated betas differs between probit and MLE when the Xs are distributed uniformly or log normally.

in the sample, a regression of y_i on a constant and $\Phi(X\beta)$ leads to the coefficients 0 and 1 as expected (see equation 4). By contrast, the same regression with dependent variable $\tilde{y}_i$ created to have 10% misclassification error gives coefficients .1 and .8, as predicted. The model can also be estimated by nonlinear least squares. We minimize a quadratic error term which is altered somewhat from the MLE case. Instead of estimating α, the equation allows the constant term and the coefficient on $\Phi(X\beta)$ to differ from zero and one, respectively:

$$y_i = \delta_0 + \delta_1 \cdot \Phi(X_i\beta) + \eta_i \tag{13}$$

We use White standard errors here; they are calculated using the derivative of the nonlinear function with respect to the parameters in place of the usual regressors, X. The NLS method results in the following estimates for δ_0, δ_1, and the β's.

Table B: One sample generated.

The simulated X's are drawn from a standard normal distribution.
$\beta_0 = -2$
$\beta_1 = 1$

Design $\alpha=0.1$ N=2000	Probit: correct y		NLS y misclassified			
	β_0	β_1	δ_0	δ_1	β_0	β_1
	-1.99 (.141)	0.995 (.059)	.099 (.024)	.804 (.028)	-2.05 (.623)	1.03 (.278)

The results here are close to the MLE results and they satisfy the overidentifying restriction that

$\delta_1 = (1-2\delta_0)$: $\quad .802 = 1-(2 \cdot .099) \approx .804.$

Thus, NLS leads to consistent estimates in the presence of misclassification. It also permits a test of the restriction of independent misclassification error made in the basic specification.

In a situation requiring misclassification treatment, the researcher may suspect the probabilities of misclassification from one category to another are not symmetric.[8] Below is the basic model, modified to allow for this difference. Simulation results follow.

Let $\pi_0 \equiv$ probability of correct classification of 0's.

$\pi_1 \equiv$ probability of correct classification of 1's.

We can then evaluate the probability of observing given responses under different probabilities of misclassification:

$$\begin{aligned} pr(y_i=1|X) &= \pi_1 \cdot pr(y_i^* > 0) + (1-\pi_0) \cdot pr(y_i^* \leq 0) \\ &= (1-\pi_0) + (\pi_1 + \pi_0 - 1) \cdot \Phi(X\beta) \end{aligned} \tag{14}$$

$$pr(y_i=0|X) = \pi_0 - (\pi_1 + \pi_0 - 1) \cdot \Phi(X\beta).$$

The log likelihood function for the specification which allows for different probabilities of misclassification is:

$$L = \sum_i \{y_i \ln[(1-\pi_0) + (\pi_1-\pi_0-1)\ \Phi\ (X_i\beta)] + (1-y_i) \ln\ [\pi_0-(\pi_1+\pi_0-1)\ \Phi\ (X_i\beta)]\}$$

For this equation to be identified, $\pi_1 + \pi_0$ must be greater than one. If the condition does not hold, it is likely that the data are sufficiently "noisy" that model estimation is not warranted; in practice, the condition should not be a restrictive requirement.

---

[8] This problem displays a loose relationship to the switching regression problem, e.g. Porter and Lee (1984). However, the switching regression framework uses the discrete variable to represent a regime shift as a right hand side variable.

Table C: One Sample Generated The simulated X's are drawn from a standard normal distribution. β_0 = -2 β_1 = 1				
Design π's	Probit: correct y β_0 β_1	Probit: y misclassified β_0 β_1	MLE : using design π's β_0 β_1	MLE: solving for π_0 and π_1 π_0 π_1 β_0 β_1
π_0= .950 π_1= .975	-2.11 1.03 (.107)(.044)	-.770 .373 (.044)(.011)	-2.18 1.02 (.155) (.064)	.978 .970 -1.98 .969 (.007)(.004)(.145)(.061)

Thus, again MLE is able to estimate the parameters both accurately and quite precisely. However, note that we can no longer test the overidentifying restriction of the misclassification specification when we allow for different probabilities of misclassification.

It is also likely that a researcher could encounter a situation where he or she suspects that the probability of misclassification is correlated with the y_i^*'s in a continuous fashion:

$$\begin{aligned} pr(y_i=1|X) &= \pi(y^*)\cdot pr(y_i^*>0) + (1-\pi(y^*))\cdot pr(y_i^*\leq 0) \\ &= \pi(X_{1i}\beta_1)\cdot\Phi(X_{2i}\beta_2) + (1-\pi(X_{1i}\beta_1))\cdot(1-\Phi(X_{2i}\beta_2)) \\ E(y_i|X) &= E[1-\pi(X_{1i}\beta_1)|X_1] + E[(2\pi(X_{1i}\beta_1)-1)\cdot\Phi(X_{2i}\beta_2)|X_2] \\ &= 1-\pi(X_{1i}\beta_1) + (2\pi(X_{1i}\beta_1)-1)\cdot\Phi(X_{2i}\beta_2) \end{aligned} \tag{15}$$

Suppose logit were the monotonic function ranging from 0 to 1 that determined π. Then the expected value of y_i given X could be written,

$$E(y_i|X) = 1-\Lambda(X_i\delta) + (2\Lambda(X_i\delta)-1)\cdot\Phi(X_i\beta) \tag{16}$$

If we condition on the X's, all the previous results hold. This case is much more complicated to estimate; we do not provide simulations here.

Logit case

The logit functional form can also be analyzed in the case of suspected measurement error in the dependent variable. First, we calculate the probabilities of the observed data given misclassification:

$$\begin{aligned} pr(y_i=1|X) &= \pi\cdot pr(y_i^*>0) + (1-\pi)\cdot pr(y_i^*\leq 0) \\ &= \pi\cdot[\exp(X\beta)/(1+\exp(X\beta))] + (1-\pi)\cdot[1/(1+\exp(X\beta))] \end{aligned} \tag{17}$$

$$\begin{aligned} pr(y_i=0|X) &= \pi\cdot pr(y_i^*\leq 0) + (1-\pi)\cdot pr(y_i^*>0) \\ &= \pi\cdot[1/(1+\exp(X\beta))] + (1-\pi)\cdot[\exp(X\beta)/(1+\exp(X\beta))] \end{aligned}$$

We then calculate the expectation of the left hand side variable conditional on the X's and given the probabilities of misclassification:

$$\begin{aligned} E(y_i|X) &= [(1-\pi) + \pi\cdot\exp(X\beta)]/[1+\exp(X\beta)] \\ &= [\alpha + (1-\alpha)\cdot\exp(X\beta)]/[1+\exp(X\beta)] \end{aligned}$$

We now demonstrate that the case of misclassification leads to a very different outcome than the case of endogenous sampling, e.g., the papers in Manski and MacFadden (1981). In the case of endogenous sampling the slope coefficients are estimated consistently by MLE logit; only the constant is incorrectly estimated. Thus, estimation refinements are concerned with increasing asymptotic efficiency. Here, in the situation of misclassification, MLE logit leads to inconsistent estimates of both the slope coefficients and the constant term. This result follows easily from a comparison of the (Berkson) log odds form of the logit in the case of measurement error and endogenous sampling. Without misclassification we have:

$$\begin{aligned} pr(y_i=1|X) &= \exp(x_1\beta)/(\exp(X_0\beta + \exp(X_1\beta)) \\ pr(y_i=0|X) &= \exp(X_0\beta)/(\exp(X_0\beta)+\exp(X_1\beta)). \end{aligned} \tag{18}$$

Thus, the expectation of the log odds is:

$$E(\ln(s_1/s_0)) = E[\ln(\exp(X_1\beta)/\exp(X_0\beta))] = (X_1 - X_0)\beta.$$

With endogenous sampling where we redefine λ to be the weighting constant we find:

$$pr(y_i=1|X) = \lambda\exp(X_1\beta)/(\exp(X_0\beta) + \exp(X_1\beta)) \quad (19)$$

$$pr(y_i=0|X) = (1-\lambda)\cdot\exp(X_0\beta)/(\exp(X_0\beta) + \exp(X_1\beta))$$

$$E(\ln(s_1/s_0)) = \ln(\lambda\exp(X_1\beta)[(1-\lambda)\cdot\exp(X_0\beta)]^{-1})$$

$$= (\ln(\lambda/(1-\lambda)) + \beta_0) - (X_1-X_0)\beta_1 \quad (20)$$

Thus, only the constant term is inconsistent with endogenous sampling. Using the expressions derived above for the case of misclassification, where as before π is the probability of correct classification:

$$E(\ln(s_1/s_0)) = \ln[(\pi + (1-\pi)\cdot\exp(X_1\beta))/(1-\pi + \pi\cdot\exp(X_0\beta))]. \quad (21)$$

If we express the problem in a log odds form, it is easier to see that the misclassification will induce inconsistency in both the constant and the slope coefficient estimates. An interesting feature of misclassification is that it prevents one observation from causing unbounded inconsistency. The dependent variable for any one observation can become arbitrarily large or small as the proportion of ones approaches one or zero. When misclassification is present, the α terms prevent the limit from reaching infinity.

Simulation results analogous to those using a probit functional form appear in Tables III and IV. Similar results hold if the probabilities of misclassification are not equal across both responses. Furthermore, if we use the log odds specification, we can extend our approach to the case of three or more categories of qualitative responses where the misclassification probabilities differ both across response category and within response categories. The appendix holds a simple treatment of this topic.

III. Semiparametric Analysis

The assumption of normally distributed (or extreme value) disturbances required by the probit (logit) specifications is not necessary to solve for the amount of misclassification. One can use semiparametric methods which estimate the extent of misclassification in the dependent variable and estimate the slope coefficients consistently. We employ a kernel regression technique here. The kernel regression uses an Epinechnakov kernel, as it has desirable efficiency properties.[9] The window width, h, defines the intervals over which the kernel, or weighted average, operates; we use Silverman's rule-of-thumb method as a first approximation:

$$h^* = 1.06\sigma n^{-1/5}$$

One could also determine h^* by cross-validation, a computationally intensive procedure which minimizes the integrated square error over h. However, past experience leads us to expect very similar results.

Our kernel regression is of the simplest form as it uses an index function to make the kernel one-dimensional. We calculate the index $I=X\beta$, where the β's are found in a first-stage procedure which will be discussed subsequently. Then a set of evaluation points covering the range of the index is chosen. The kernel estimator is the weighted average of the y's associated with each observation falling in the window centered at the point of evaluation:

$$\hat{f}(i) = \frac{1}{nh} \sum_{j=1}^{n} K\left(\frac{i-I_j}{h}\right) \tag{22}$$

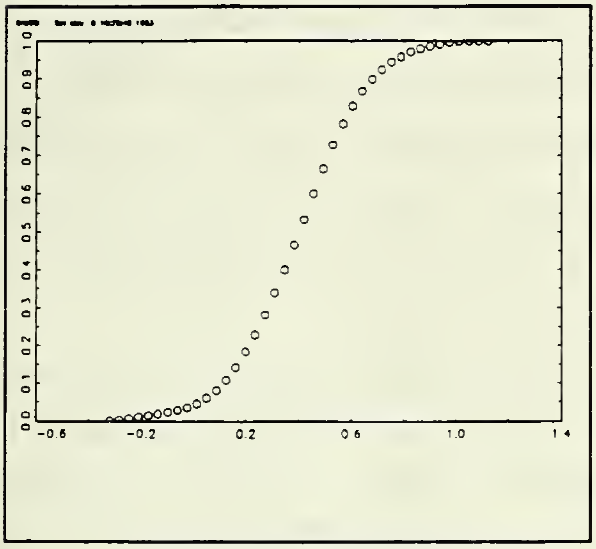
where I is the index, h is the window width, n is the number of observations, and K(.) is the Epinechnakov kernel. The cumulative distribution of the response function (cdf) is formed by dividing the weighted sum of y's by the total number of points falling in the window of the

---

[9] Silverman (1986), p. 42, describes how the Epinechnakov kernel is the most "efficient" kernel in the sense of minimizing the mean integrated square error if the window width is chosen optimally.

kernel. We plot the cdf using the probit coefficients from above to construct the index. We use simulated data with a misclassification probability, α, of 0.07, which approximately corresponds to the estimated probability we find in our empirical example below with the CPS data:[10]

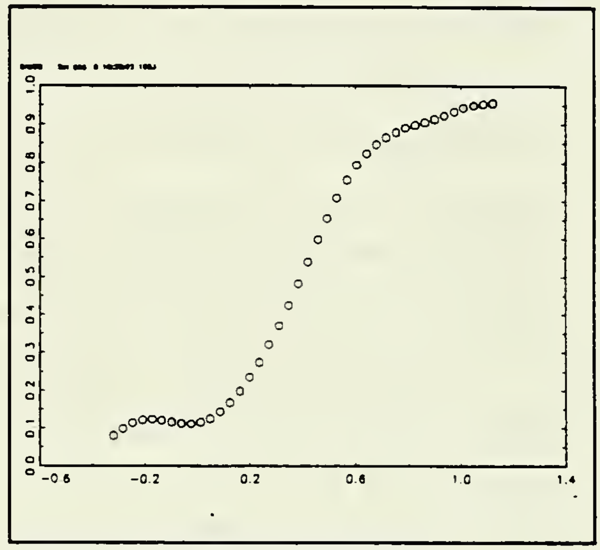
Figure 1a. Standard CDF



index

true y: simulated data

Figure 1b. CDF with Misclassification



index

.07 y misclassified: simulated data

Above are two graphs, one using misclassified data and one using true data. The method produces the smooth curves of Figures 1a and 1b, where the misclassification problem shows up clearly. For example, a .07 misclassification rate would cause the cdf of the dependent variable to begin at .07 and asymptote at a value of only .93. No matter how small $X\beta$ is, there is always a seven percent chance that the observation will be misclassified and observed y will be one. A kernel regression performed on simulated data, demonstrates the difference in the cdf's. Note that in Figure 1a the cdf goes between the values of 0.0 and 1.0 as the index increases. However, in Figure 1b the beginning value is about .07 while the cdf asymptotes to

[10] The simulated data consist of 2,000 observations of the sum of normally distributed X's and ϵ's.

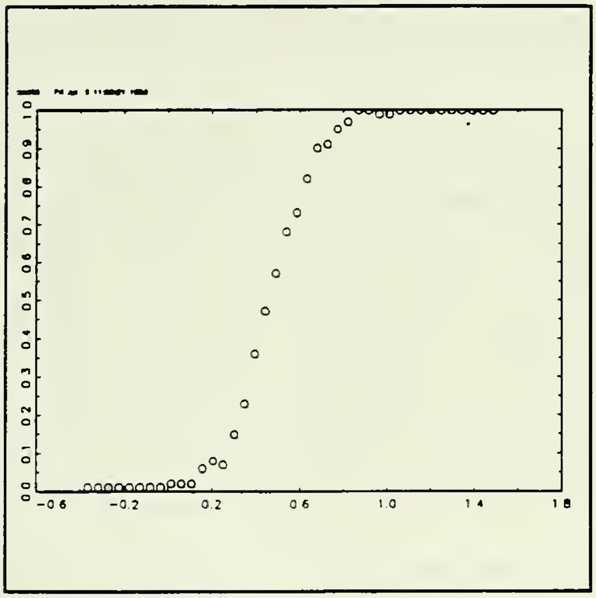
about .93. Using the cdf in this manner allows for identification of the probability of misclassification using semiparametric or nonparametric techniques so long as the probability of misclassification is independent of the X's.

We experienced two major problems with the kernel regression. The first is the classic problem of how to handle the tails of the distribution. In this case, the behavior of the tails is critical to the discussion of misclassification. Yet, is it misleading to continue the kernel regression as far out into the tails as possible because the number of observations falling within the window width becomes very low. In fact, the asymptotic behavior of the cdf becomes obscured if estimation continues to the extreme points. In Figure 1b above we use the points one bandwidth from the extreme index values as truncation points for the regression. Other solutions include variable kernel and nearest neighbor methods.

We use the nearest neighbor algorithm as an alternative method to investigate the behavior of the cdf in the tails. The nearest neighbor method is a weighted average of the k closest points to the evaluation point. As observations begin to thin out, nearest neighbors become further away rather than fewer in number. We looked at the effect nearest neighbor methods have on our particular tail problem using several different k values. The graphs below use k=100 and exhibit similar tail behavior as the truncated kernel regressions which we considered above. Again, the probit coefficients are used to construct the index.

Nearest Neighbor Method CDFs

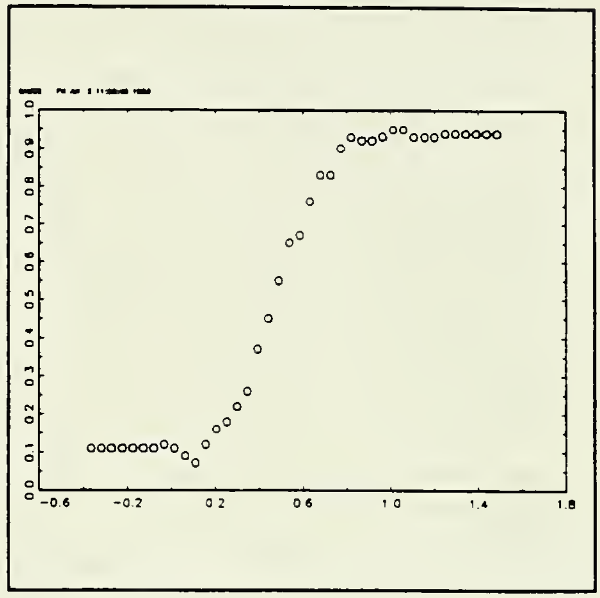
Figure 2a: Standard CDF



index

true y: simulated data

Figure 2b: Misclassified CDF



index

.07 y misclassified: simulated data

The second problem is more severe: how to find the starting β's. One method is simply to use the probit regression coefficients in the construction of the index. While illustrating the problem, we are clearly not freeing the data from an assumed error structure. More importantly, while the results will demonstrate that misclassification is present because of the behavior of the cdf, the estimates of misclassification will not, in general, be consistent estimates.

IV. 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 the probability of individuals changing jobs over the past year. Because of the possible confusion of changing positions versus changing employers, respondents may well make a mistake in their answers. We would expect that job changers have a higher probability of misclassification than do non-job changers. Thus, we consider models both with the same and different probabilities of misclassification, and we test for equality of the probabilities.

Our data come from the January 1987 Current Population Survey of the Census Bureau. We extracted the approximately 5,000 complete personal records of men between the ages of 25 and 55 whose wages are reported. Among the questions in the survey is one asking for the respondent's job tenure. This type of question is well-known for its misclassification; people confuse position with job or may have re-entered the labor force. 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 means of the data are in Table V below.

We estimate a probit specification on the variable for changing jobs, which we call *Jobchange*. Our specification is quite similar to specifications previous used in the applied labor economics literature, c.f. the specification of Freeman (1984). We maximize the log likelihood function presented initially in equation (7) using the same explanatory variables but allowing α, the probability of misclassification, to be non-zero. We present the results in Table VI. Note that when we allow for the probability of misclassification we estimate the probability to be 0.058, with an asymptotic t-statistic of about 8.3. Thus, we find quite strong evidence of misclassification. Many of the probit coefficients also change by substantial amounts. For instance, the effect of unions in deterring job changes is found to be much higher when possible misclassification of the response is allowed for. Likewise, the effect of higher earnings in affecting job change also becomes much greater in the misclassification specification.

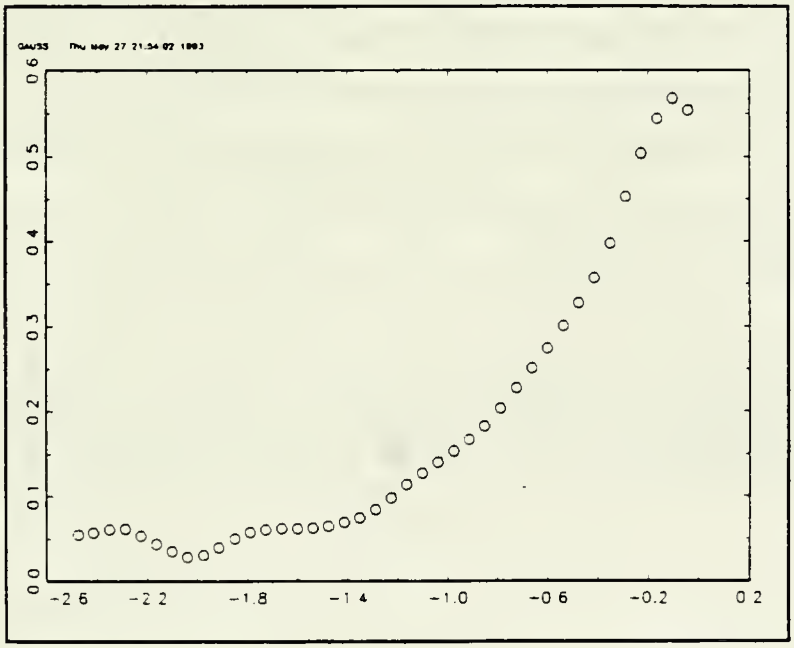
Next, we allow for different probabilities of misclassification as in equation (15) since non-job changers are less likely to misreport their status. In the third column of the table, α_0, the probability of misclassification if the observed response is no job change, is allowed to differ from α_1, the probability of misclassification if the observed response is a job change. Thus, we allow the misclassification probabilities to differ across the responses. Freeing the α 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 between the two estimates of

α in column 3 is .248 and has a standard error of .164, giving it an asymptotic t-statistic of 1.51, which is not quite significant. However, the effect of both union and earnings in deterring job change are even greater than previously estimated. Thus, allowing for different probabilities of misclassification again leads to quite different results than in the original probit specification.

We now perform kernel regression to determine if it also demonstrates that response misclassification exists in the responses. The kernel regression results in Figure 3 demonstrate the situation clearly. The graph below uses the probit coefficients to create an index function. We find the minimum probability of job change to be around .06 with an increase up to about 0.6. These results are extremely close to the MLE probit results of Table VI where we estimate the respective probabilities to be .06 and 0.7 with an asymptotic standard error of 0.17.

Semiparametric estimation of the cdf can reveal the α's by showing different asymptotes at each end of the function. In our case, the lower asymptote is 6-7% and the upper one, though much less well defined, is in the 50-60% range. However, semiparametric estimation of the probability of misclassification by visual inspection is not totally satisfactory. We cannot achieve consistent estimates of the job change model parameters nor can we estimate the sampling distribution of the misclassification probabilities. Thus, we now develop a semi-parametric methodology which allows us to overcome the shortcomings of our previous approach. Nevertheless, our results up to this point do demonstrate quite convincingly that misclassification of job changes does exist in the CPS data.

Figure 3: Probit Coefficients; Jobchange Data



V. Consistent Semiparametric Estimates

We now derive a methodology which permits consistent estimates of the misclassification probabilities and the unknown slope coefficients.[11] We first use the Maximum Rank Correlation (MRC) estimator described in Han (1987). This binary response estimator is straightforward to calculate even in the instance of multiple explanatory variables and dummy variables. MRC operates on the principle that for observations where the index $X_i\beta > X_j\beta$ for a given estimate of β, we would expect that $y_i \geq y_j$. The necessary monotonicity property for MRC holds under misclassification so long as $\pi_0 + \pi_1 > 1$; see equation (14). The observations are ordered by the index $I_i = X_i\beta$ and then a sum is constructed by giving an indicator variable a "one" in cases where the dependent variables are also in the "correct" order. Thus, the MRC function which is maximized is:

$$S_N(\beta) = \begin{bmatrix} N \\ 2 \end{bmatrix}^{-1} \sum \left[l(y_i > y_j)\, l(x_i'\beta > x_j'\beta) + l(y_i < y_j)\, l(x_i'\beta < x_j'\beta) \right] \tag{23}$$

where l(.) is an indicator function with l(.) = 1 if (.) is true, while l(.) = 0 otherwise and the summation is over all distinct pairs of elements (i,j) from the sample i = (1,...,N).

An advantage of using MRC instead of probit (or logit) is that the normality assumption is not required. Violation of the distributional assumption may lead to inconsistent results when probit (or logit) is used.

The MRC was designed to allow estimation of the β's in the index in the case of any monotonic transformation of the index. Misclassification is just such a case; the cdf no longer runs from 0 to 1, but instead is bounded by the probabilities of misclassification. However, the problem remains monotonic in the indicator variable, $I_j = X_j\beta$, so long as the probabilities of correct classification, π_0 and π_1, exceed 0.5 each and do not depend on the X's. Thus the β's

---

[11] Manski (1985) discusses use of the maximum score estimator when a single misclassification probability exists for the data.

from MRC estimation are consistent in the case of misclassification. The consistency is proved by Han (1987) and asymptotic normality of the MRC coefficients is proved by Sherman (1992) and in Cavanaugh and Sherman (1992). However, we still need a semiparametric, consistent estimate of the misclassification probabilities, α_0 and α_1.

Misclassification probabilities are given by the asymptotes of the cdf; the chosen technique gives estimates of those asymptotes. The first method that a researcher might choose would be kernel regression on each tail of the cdf. (Recall that we know the estimated β's so we can construct such a cdf.) However, it is not obvious how to weight observations in the tails of a distribution. Where does the "tail" begin, for example. The data become very thin and the last observation becomes disproportionately important.

Instead, in order to solve for the α's semiparametrically, we use the technique of isotonic regression (IR). IR nonparametrically estimates a cdf using ranked index values. The technique is useful for our problem because the final estimate is in the form of a step function; the lowest and highest "steps" provide consistent estimates of α_0 and $1-\alpha_1$. The combination of MRC and IR techniques achieves consistent estimates of both the coefficients on the explanatory variables and an estimate of the amount of misclassification in the data. The Isotonic Regression technique involves the following basic steps.

First, we use the coefficient estimates from the MRC procedure of Han (1983) which are $N^{1/2}$ consistent. We want to estimate the probability distribution of the dependent variable conditional on an index constructed from the MRC estimated β's. The index is defined $v_i = X_i \hat{\beta}$. Next, the indices are ordered so that $v_1 \leq v_2 \leq v_N$. An isotonic function, F, with respect to the index v is any non-decreasing function defined on the N index values. $\hat{F}$ is an isotonic regression of y if it minimizes

$$\sum_{i=1}^{N} (y_i - \hat{F}(v_i))^2$$

In addition, $\hat{F}(v) = 0$ if $v < v_1$ and $\hat{F}(v) = 1$ if $v > v_N$. F(.) must also be non-decreasing, so $\hat{F}(v_{i-1}) < \hat{F}(v_i)$. Under this specification for <u>known</u> index values v, Groeneboom (1993) proves that the point estimates, $\hat{F}(v)$ are $N^{1/3}$ consistent for $\hat{F}(v) \in$

(0,1). We will demonstrate that using Han's estimator of v leads to the same result.

Performing isotonic regression is quite straightforward.[12] The idea is to organize the index values into pools, where each pool is assigned a "best guess" for the value of y conditional upon the index being in the pool. The "guess" is just the average of the y_i values corresponding to the index values v_i in the pool.[13] 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 is continued until the pools are exhausted. Finally, if any combinations of pools occurred during the last sequential comparison of the pools, repeat the process. Otherwise, the regression is complete. The isotonic regression of y with respect of v is described completely by the guesses associated with the final set of pools; (v_i) is the guess for the pool containing v_i at the end of the algorithm.

The lowest and highest steps resulting from the isotonic regression are consistent estimates of α_0 and α_1. However, it is critical to realize that the estimated α's can only be identified using nonparametric methods. In order to observe the lower asymptote of the cdf, the researcher must have observations with very low index values. Similarly for the upper asymptote; if there are no observations at high index values, the estimated cdf will end before those values are reached. The researcher must ask the question, does the cdf end because the final value is the "true" asymptote or because there are no data points in that area? In order to solve the problem, the data would have to contain observations with index values which approach negative or positive infinity. However, if the data display a long tail that is approximately constant at some non-zero α, that is good evidence that misclassification could be a problem in the data. Certainly further investigation would be warranted; perhaps resampling techniques could be employed to find out more about the problem.

Once we have the MRC estimates and the IR estimate of the cdf, we find the asymptotic

---

[12] See Barlow et. al. (1972) or Robertson et. al. (1988) for detailed discussions of IR and the algorithms used for estimation.

[13] The non-decreasing property constrains pools to contain only adjacent index values.

distribution of the IR estimator. To date no asymptotic distribution has been found for methods which estimate both the unknown slope coefficients β and the cdf simultaneously in semiparametric models of discrete response. However, we are able to derive the distribution of both sets of coefficients here. We use known results combined with a basic insight. The insight follows from the fact that MRC estimates are $N^{1/2}$ consistent while isotonic regression of the form we use with step functions is $N^{1/3}$ consistent, so we can treat the MRC coefficients as known in deriving the asymptotic distribution for the isotonic regression in applying Groeneboom's (1993) results. (The proof is given in the Appendix V.).

Estimation of the Asymptotic Distribution for the Isotonic Regression Step Function

We use the method of isotonic regression to estimate the c.d.f. $F(I) = \Pr(y=1|I)$ with a step function where I is the index, $I = X\beta$. We use the $N^{1/2}$ consistent estimates of these parameters which come from estimating the model by the MRC technique in Han (1987) and then do an isotonic regression using these estimates for the index value. If we did not use the MRC estimator (or other $N^{1/2}$ estimator), we would be unable to derive the asymptotic distribution of the misclassification probabilities. Using the index observations $I_1,\ldots,I_n$ (derived from $x_1, \ldots, x_n$ and the $N^{1/2}$ consistent parameter estimates of β), we find the IR functional estimate $\hat{F}_n$. We are interested in the asymptotic standard errors associated with the height of the steps of $\hat{F}_n$. The asymptotic distribution was found by Groeneboom (1993): *For any* *<u>known</u>* *v s.t. F(v) ∈ (0,1),*

$$\frac{n^{1/3}(F_n(v) - F(v))}{\left(\frac{1}{2}F(v)(1-F(v))\frac{f(v)}{g(v)}\right)^{1/3}} \to_d 2Z \tag{24}$$

where f is the derivative of the true distribution F, g is the derivative of the distribution of the index, and Z is the last time where two-sided Brownian motion minus the parabola v^2 reaches its maximum. Thus, the Groeneboom technique to estimate the asymptotic distribution cannot be applied to the Cosslett-type IR estimator because the slope coefficients β are being estimated along with the cdf. However, our approach which utilizes an $N^{1/2}$ estimator for β allows the parameter vector to be treated as known for purposes of estimating the cdf.

The distribution of Z can be written (Groeneboom (1984)) as

$$h_z(v) = \frac{1}{2} s(v) s(-v), \quad v \in \Re \tag{25}$$

where the function s has a Fourier transform

$$\hat{s}(w) = \frac{2^{1/3}}{Ai(2^{-1/3} wi)} \tag{26}$$

and where Ai is the Airy Function and I $= \surd$-1.

Solving for Z will allow us to estimate the variance of F_n HAT. We have E(Z) = 0 from symmetry of h_z. Also, via numerical approximation of the Fourier transform and numerical integration, we estimate Var(Z) $\approx$ 0.26.

Using (23), we then have

$$Var(\hat{F}_n(v) - F(v)) \approx 1.04 \left(\frac{\hat{F}_n(v)(1-\hat{F}_n(v))\hat{f}(v)}{2n\hat{g}(v)} \right)^{2/3} \tag{27}$$

where we use the numerical estimate of the index density for $g(v)$. Unfortunately, we cannot use the numerical derivative of $F_n(v)$ for $f(v)$ since the derivative of a step function is zero except at a finite number of points (where it is infinity). Instead, for a given step level, we use a "slope approximation" arrived at by looking at the jumps from or to the adjacent steps. $f(v)$ is defined to be the slope of the straight line from the point half way up the vertical rise to the current step and the point half way up the rise to the next step. Because this definition of the slope does not exist for the highest step, we calculate it assuming that the next step level is one.

The density of the index is estimated using a window width of 0.1; the results are not sensitive to the choice of window width. The point at which $g(v)$ is estimated is the middle of the step. The process that estimates the standard errors of the β's uses a kernel window of σ*n

$^{1/5}$ where σ is the standard deviation of the index function and n is the number of observations. The standard errors for the α's in Table IX use a kernel window of 0.05. Again, the estimates are not greatly affected by the choice of window width.

Consistent Semiparametric Estimates of the Job Change Model

We now apply the combined MRC and isotonic regression estimators to the job change data from the CPS. The results of using the estimators on the job change data are reported in Table VIII. MRC/IR produces coefficients which are scaled so that each coefficient vector has length one. Thus, the absolute magnitude of the coefficients is indeterminate; only the ratios can be found from MRC/IR. In order to compare the results from probit and MLE with MRC/IR it is necessary to scale the MRC/IR coefficients. We assume that the coefficient on *Western* stays fixed across methods; it is also the coefficient which is assumed to be fixed in calculating the asymptotic semiparametric standard errors. In this way we obtain a scaling factor which will convert the MRC/IR coefficient into the probit coefficient and multiply the remaining coefficients by the same factor.

Table VIII restates the earlier results (Table VI) for ease of comparison. MLE produces point estimates which differ from those of probit as was seen before in Table VI. However, estimating the model using the MRC/IR procedure gives results that always lie in between the estimates obtained from ordinary probit, ignoring misclassifications, and MLE correcting for misclassification. Thus, when we use the combined MRC/IR approach we find that the estimated probabilities of misclassification are 0.035 and 0.395. Both are estimated quite precisely, and we can now reject the hypothesis that the probabilities of misclassification are equal. Freeing the misclassification model from the assumption of a normal error distribution significantly changes the estimate of α_0; only 3.5% of observed non-jobchangers have a misclassified response. 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 be 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 Figure 4 we plot the estimated cdf from the MLE model which allows for misspecification and also the MRC/IR estimate of the cdf. The results are reasonably similar. Figure 5 shows the MRC/IR estimate of the cdf with a confidence interval of two standard errors in either direction. The confidence interval demonstrates that the IR step function estimate of the cdf 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 estimated cdf with a standard kernel estimate of the cdf. Figure 6a uses a fixed-window kernel; this 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 non-monotonic at the upper tail of our data. Figure 6b tries an alternative approach by using the 200 nearest neighbors to construct the kernel estimate. Again, the upper tail of the cdf is quite different from the MRC/IR estimate. Kernel regression techniques are basically not suited to estimating the asymptotes of a cdf where data become sparse.

If the assumption of normally distributed errors were correct for our data set, we would expect to see the MRC coefficients come very close to the MLE coefficients. Some coefficients, such as *Union*, match closely, but others such as *Last Grade Attended* and *Married*, do not. The differences could well arise from a failure of the underlying probit model assumptions of normality.

To explore further the accuracy of the MRC/IR approach compared to probit, we conducted another simulation analysis. We generated 5,000 observations of simulated data with three explanatory variables (none distributed normally), a constant, and a normally distributed error. Ten percent of the binary response dependent variable was purposefully misclassified. The results of the three estimation techniques are reported in Table IX. MLE correcting for misclassification should return the coefficients used to construct the data, as should MRC with no special correction. The conclusion of this exercise is that the assumptions required for probit are not met by our job change data, and that the MRC estimator provides superior estimates. It is also interesting to note from Table VIII that the MRC/IR estimator is able to estimate the misclassification probabilities quite accurately and their values are close to our expectations formed by the earlier semiparametric and MLE analysis.

VI. Conclusions

The discussion above shows that ignoring potential misclassification of a dependent variable can result in substantially biased and inconsistent coefficient estimates when using probit or logit specifications. The researcher can use our maximum likelihood procedure described above to estimate the extent of misclassification and estimate consistent coefficients. If the misclassification probabilities are asymmetric across groups, they may still be estimated easily. However, should the errors in the data not be normally distributed, these coefficients may nevertheless be inconsistent. Semiparametric regression using the MRC technique of Han (1987) will yield consistent estimates of the coefficients regardless of the amount of misclassification in the data. Furthermore, the isotonic regression techniques detailed above provide a procedure which will give consistent estimates of the misclassification percentages in the data. We are able to derive and estimate the asymptotic distribution for both the slope parameters and for the misclassification probabilities.

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

We apply our econometric techniques to job change data from the CPS. We find that misclassification exists in the data. Furthermore, the probabilities of misclassification differ depending on the response. We are able to estimate the parameters by the MLE parametric approach and by the distribution free semi-parametric approach; the estimated parameters are reasonably similar although they are estimated with only moderate levels of precision. 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 --especially in probit and logit models.

References

Barlow, R. et.al. (1972), Statistical Inference under Order Restrictions, (New York, John Wiley).

Cavanagh, Chris and Sherman, Robert (1992), "Rank Estimators for Monotone Index Models," Bellcore mimeo.

Chua, Tin Chiu and Wayne Fuller (1987), "A Model for Mulitnomial Response Error Applied to Labor Flows," *Journal of the American Statistical Association* 82:397:46-51.

Cosslett, Stephen, R. (1983), "Distribution-Free Maximum Likelihood Estimator of the Binary Choice Model," *Econometrica* 51:3:765-782.

Freeman, Richard, (1984), " Longitudinal Analyses of the Effects of Trade Unions" *Journal of Labor Economics*, vol. 2, no. 1, 1-26.

Greene, William (1990), Econometric Analysis, (New York, Macmillan).

Groeneboom, Piet (1985), "Estimating Monotone Density", in L.M. Le Cam and R.A. Olshen, eds., Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Keifer, vol II, (Hayward, CA: Wadsworth).

Groeneboom, Piet (1993), "Asymptotics for Incomplete Censored Observations", mimeo.

Han, Aaron (1987), "Non-Parametric Analysis of a Generalized Regression Model: The Maximum Rank Correlation Estimator," *Journal of Econometrics* 35:303-316.

Han, Aaron, and Hausman, Jerry A., (1990), "Flexible Parametric Estimation of Duration and Competing Risk Models," *Journal of Applied Econometrics*, 5, 1-28.

Hardle, W., (1990), Applied Nonparametric Regression, Cambridge University Press, Cambridge.

Klein, R., and Spady, R., (1993), "An Efficient Semiparametric Estimator for Binary Response Models," *Econometrica* 61:2.

Manski, Charles F., (1985), "Semiparametric Analysis of Discrete Response: Asymptotic Properties of the Maximum Score Estimator," *Journal of Econometrics* 27:313-333.

Manski, Charles and Daniel McFadden (1981), Structural Analysis of Discrete Data with Econometric Applications, (Cambridge, MIT Press).

MacFadden, Daniel (1984), "Econometric Analysis of Qualitative Response Models,", in Z. Griliches and M. Intrilligator, eds., Handbook of Econometrics, vol. 2, (Amsterdam: North Holland).

Porter, Robert and Lee, Lung Fei (1984), "Switching Regression Models with Imperfect Sample Separation Information with an Application on Cartel Stability" *Econometrica* 52(2):391-418.

Poterba, J. and L. Summers (1993), "Unemployment Benefits, Labor Market Transitions, and Spurious Flows: A Multinomial Logit Model with Errors in Classification" MIT mimeo.

Pratt, John W., (1981), "Concavity of the Log Likelihood," *Journal of the American Statistical Association*, 76, No. 373, 103-106.

Robertson, T. et. al., (1988), Order Restricted Statistical Inference, (New York, John Wiley).

Ruud, Paul, (1983), "Sufficient Conditions for the Consistency of MLE Despite Misspecifications of Distribution in Multinomial Discrete Choice Models," *Econometrica* 51:1:225-28.

Sherman, Robert P., (1993), "The Limiting Distribution of the Maximum Rank Correlation Estimator" mimeo.

Stock, James, Thomas Stoker, and James Powell (1989), "Semiparametric Estimation of Index Coefficients," *Econometrica* 57:6:1403-30.

Silverman, B.W., (1986), Density Estimation for Statistics and Data Analysis, Chapman & Hall, London.

Table Ia: Simulations

X_1 is drawn from a lognormal distribution
X_2 is a dummy variable drawn from a uniform distribution
X_3 is drawn from a uniform distribution

Repetitions = 146		Sample Design	Probit: y misclassified	ratio to constant	MLE: solving for α	ratio to constant
n=5000	α	0.02	---	---	0.0192 (.0054)(.0051)	---
	β_0	-1.0	-0.787 (.069)(.066)	1	-0.990 (.068)(.077)	1
	β_1	0.20	0.158 (.001)(.005)	0.20	0.199 (.008)(.009)	0.20
	β_2	1.5	1.27 (.063)(.054)	1.61	1.49 (.075)(.073)	1.51
	β_3	-0.60	-0.518 (.023)(.021)	0.66	-0.598 (.026)(.028)	0.60
n=5000	α	0.05	---	---	0.0497 (.0076)(.0075)	---
	β_0	-1.0	-0.567 (.073)(.058)	1	-1.007 (.084)(.090)	1
	β_1	0.20	0.114 (.010)(.004)	0.20	0.201 (.010)(.011)	0.20
	β_2	1.5	1.06 (.051)(.048)	1.87	1.504 (.082)(.089)	1.50
	β_3	-0.60	-0.431 (.019)(.019)	0.76	-0.599 (.032)(.034)	0.60
n=5000	α	0.20	---	---	0.198 (.014)(.013)	---
	β_0	-1.0	-0.163 (.061)(.050)	1	-0.991 (.168)(.159)	1
	β_1	0.20	0.037 (.005)(.002)	0.23	0.198 (.023)(.022)	0.20
	β_2	1.5	0.554 (.045)(.041)	3.40	1.48 (.182)(.169)	1.49
	β_3	-0.60	-0.228 (.018)(.016)	1.40	-0.592 (.072)(.064)	0.60

(The left hand parentheses contain the standard deviation of the coefficient across 146 Monte Carlo simulations. The right hand parentheses contain the average (n=146) standard error calculated from MLE. The similarity of the two show that the maximum likelihood asymptotics work well.)

Table Ib: Simulations from Table Ia in Ratios

X_1 is drawn from a lognormal distribution
X_2 is a dummy variable drawn from a uniform distribution
X_3 is drawn from a uniform distribution

Repetitions = 146		Design Ratios	Probit: y misclassified		MLE: solving for α	
			ϵ small	ϵ large	ϵ small	ϵ large
n=5000	β_0	1	1	1	1	1
α = .02	β_1	0.2	0.20	0.20	0.20	0.20
	β_2	1.5	1.61	1.58	1.51	1.48
	β_3	0.6	0.66	0.63	0.60	0.58
n=5000	β_0	1	1	1	1	1
α = .05	β_1	0.2	0.20	0.20	0.20	0.19
	β_2	1.5	1.87	1.67	1.50	1.47
	β_3	0.6	0.76	0.66	0.60	0.57
n=5000	β_0	1	1	1	1	1
α = .20	β_1	0.2	0.23	0.24	0.20	0.21
	β_2	1.5	3.40	2.49	1.49	1.50
	β_3	0.6	1.40	1.04	0.60	0.63

Table II: Simulations: One Sample Generated

Simulation β vector = (-3, 0.3, 0.2)
X_1 drawn from a normal distribution
X_2 a dummy variable drawn from a uniform distribution

true α sample size		True Probit X and y	Probit: y misclassified	MLE using true α	MLE: solving for α	α
α=.025	β_0	-2.958	-2.443	-2.937	-3.170	.0397
n=2000		(.138)	(.119)	(.166)	(.289)	(.013)
	β_1	.295	.240	.290	.314	
		(.012)	(.010)	(.015)	(.028)	
	β_2	.204	.208	.224	.236	
		(.080)	(.075)	(.088)	(.095)	
α=.05	β_0	-3.048	-2.396	-3.266	-3.210	.046
n=2000		(.137)	(.114)	(.200)	(.291)	(.014)
	β_1	.296	.233	.316	.311	
		(.012)	(.009)	(.018)	(.027)	
	β_2	.296	.168	.290	.282	
		(.079)	(.074)	(.097)	(.100)	
α=.1	β_0	-3.007	-1.901	-3.191	-2.961	.085
n=2000		(.137)	(.101)	(.243)	(.359)	(.020)
	β_1	.296	.187	.316	.293	
		(.012)	(.008)	(.022)	(.054)	
	β_2	.235	.143	.238	.218	
		(.077)	(.069)	(.110)	(.105)	
α=.2	β_0	-3.270	-1.186	-3.480	-3.282	.190
n=2000		(.151)	(.086)	(.401)	(.532)	(.019)
	β_1	.328	.122	.370	.348	
		(.013)	(.006)	(.040)	(.056)	
	β_2	.164	.051	-.0379	-.020	
		(.082)	(.065)	(.173)	(.163)	

Note: The coefficients are different in each row of column one because only one random sample is generated for each row; thus it is unlikely that the β's would match exactly.

Table III: Logit Simulations : One Sample Generated

β vector=(-2, 1)
X_1 drawn from a normal distribution

True α Sample size	True Logit X and y β_0 β_1	Logit: y misclassified β_0 β_1	MLE with true α known β_0 β_1	MLE solving for α and β β_0 β_1 α
α= .05 n= 2000	-1.990 1.012 (.153) (.054)	-.967 .504 (.095)(.023)	-1.897 .933 (.192) (.070)	-1.830 .901 .045 (.207)(.080)(.008)
α= .1 n= 2000	-2.131 1.065 (.159) (.058)	-.612 .342 (.081) (.017)	-2.100 1.071 (.257) (.107)	-2.115 1.079 .101 (.287)(.125)(.010)
α= .2 n= 2000	-2.057 1.045 (.154)(.056)	-.399 .201 (.071) (.012)	-2.338 1.146 (.401) (.178)	-2.425 1.192 .205 (.479)(.223)(.012)

Standard errors are in parentheses.

Table IV: Logit Simulations: One Sample Generated

Simulation β vector = (-2, 0.5, 0.2)
X_1 is drawn from a normal distribution
X_2 is a dummy variable

true α sample size		True Logit X and y	Logit: y misclassified	MLE using true α	MLE: solving for α	α
α=.025 n=2000	β_0	-2.08	-1.63	-2.14	-2.18	0.029
		(.151)	(.133)	(.172)	(.207)	(.011)
	β_1	0.504	0.407	0.511	0.520	
		(.023)	(.019)	(.027)	(.038)	
	β_2	.238	0.128	0.232	0.235	
		(.137)	(.127)	(.149)	(.152)	
α=.05 n=2000	β_0	-1.895	-1.56	-2.03	-1.917	0.037
		(.146)	(.130)	(.184)	(.215)	(.015)
	β_1	0.506	0.398	0.530	0.497	
		(.023)	(.018)	(.032)	(.047)	
	β_2	0.117	0.142	0.120	0.129	
		(.136)	(.125)	(.163)	(.153)	
α=.1 n=2000	β_0	-2.26	-1.17	-2.08	-2.40	0.124
		(.155)	(.117)	(.229)	(.389)	(.019)
	β_1	0.509	0.281	0.494	0.565	
		(.023)	(.014)	(.038)	(.076)	
	β_2	0.420	0.110	0.292	0.373	
		(.136)	(.115)	(.183)	(.220)	
α=.2 n=2000	β_0	-2.25	-0.903	-2.271	-2.29	0.201
		(.158)	(.107)	(.330)	(.540)	(.027)
	β_1	0.526	0.189	0.509	0.515	
		(.024)	(.012)	(.058)	(.118)	
	β_2	0.400	0.194	0.398	0.401	
		(.140)	(.107)	(.259)	(.268)	

Note: The coefficients are different in each row of column one because only one random sample is generated for each row; thus it is unlikely that the β's would match exactly. Standard errors are in parentheses.

Table V: Means of the CPS job change data

The first row of each cell is the sample statistic.
The second row contains the statistic for *jobchange*=0 observations
The third row contains the statistic for *jobchange*=1 observations

	mean	std dev	min	max	obs
married	.7293	.4443	0	1	5221
	.7468	.4349	0	1	4471
	.6253	.4844	0	1	750
grade	14.38	2.823	1	19	5221
	14.40	2.834	1	19	4471
	14.30	2.760	1	19	750
age	37.43	8.526	25	55	5221
	37.98	8.535	25	55	4471
	34.17	7.712	25	55	750
union	.2454	.4303	0	1	5221
	.2668	.4424	0	1	4471
	.1173	.3220	0	1	750
earn per	488.9	240.2	0	999	5221
week	507.9	235.7	2	999	4471
	375.1	235.6	0	999	750
west	.2015	.4012	0	1	5221
	.1946	.3960	0	1	4471
	.2427	.4290	0	1	750

Table VI: Determinants of Job Change

	Probit	MLE $\alpha_0=\alpha_1>0$	MLE $\alpha_0 \neq \alpha_1$
α_0	---	0.0579 (.007)	0.061 (.007)
α_1	---	0.0579 (.007)	0.309 (.174)
married	-0.108 (.049)	-0.073 (.077)	-0.103 (.100)
last grade attended	0.026 (.009)	0.063 (.015)	0.080 (.026)
age	-0.022 (.003)	-0.028 (.005)	-0.033 (.007)
union membership	-0.434 (.061)	-0.707 (.148)	-0.811 (.199)
earnings per week	-0.001 (.0001)	-0.003 (.0004)	-0.004 (.0009)
western region	0.214 (.054)	0.301 (.086)	0.367 (.127)
constant	0.051 (.162)	0.171 (.259)	0.581 (.495)
log likelihood number of obs.	-1958.1 5221	-1941.4 5221	-1940.9 5221

Standard errors are in parentheses.

Table VIIa: Marginal Effects

at the mean, first, and third quartiles of the data

	quartile	regular probit	MLE $\alpha_0=\alpha_1>0$	MLE $\alpha_0 \neq \alpha_1$
married	1st	0	0	0
	mean	-.13	-.04	-.05
		(.08)	(.07)	(.12)
	3rd	-.20	-.04	-.04
		(.12)	(.10)	(.23)
grade	1st	.41	.86	.93
		(.21)	(.25)	(.35)
	mean	.61	.73	.79
		(.34)	(.31)	(.60)
	3rd	.80	.54	.49
		(.51)	(.38)	(.97)
age	1st	-.81	-.88	-.87
		(.13)	(.17)	(.36)
	mean	-1.35	-.85	-.83
		(.29)	(.25)	(1.16)
	3rd	-1.76	-.62	-.51
		(.51)	(.37)	(2.07)
union	1st	0	0	0
	mean	-.18	-.14	-.14
		(.06)	(.07)	(.14)
	3rd	0	0	0
earnings per week	1st	-.49	-1.05	-1.09
		(.11)	(.16)	(.34)
	mean	-1.05	-1.32	-1.37
		(.31)	(.25)	(1.05)
	3rd	-1.49	-1.04	-.92
		(.54)	(.34)	(1.89)
western region	1st	0	0	0
	mean	.07	.05	.05
		(.02)	(.02)	(.07)
	3rd	0	0	0

(Asymptotic standard errors are in parentheses.)

Table VIIb: Semiparametric Marginal Effects at the mean, first, and third quartiles

	quartile	MRC/IR	MLE $\alpha_0 \neq \alpha_1$
married	1st	-.0032	-.0012
		(.0196)	(.0013)
	mean	-.0080	(-.0073
		(.0261)	(.0077)
	3rd	-.0389	-.0184
		(.0415)	(.0200)
grade	1st	.0010	.0009
		(.0065)	(.0005)
	mean	.0026	.0057
		(.0090)	(.0032)
	3rd	.0127	.0143
		(.0122)	(.0069)
age	1st	-.0007	-.0004
		(.0042)	(.0002)
	mean	-.0018	-.0023
		(.0055)	(.0011)
	3rd	-.0085	-.0058
		(.0026)	(.0021)
union	1st	-.0156	-.0092
		(.0965)	(.0042)
	mean	-.0395	-.0575
		(.1281)	(.0216)
	3rd	-.1920	-.1447
		(.1040)	(.0553)
earnings per week	1st	-.0001	-.00005
		(.0003)	(.00002)
	mean	-.0001	-.0003
		(.0004)	(.0001)
	3rd	-.0007	-.0007
		(.0003)	(.0003)
western region	1st	.0072	.0042
		(---)	(.0026)
	mean	.0182	.0260
		(---)	(.0162)
	3rd	.0887	.0655
		(---)	(.0299)

(Asymptotic standard errors are in parentheses.)

Table VIII. Comparison of Probit, MLE, and MRC/IR Results
The dependent variable is *Jobchange.*

	Probit	MLE $\alpha_0=\alpha_1>0$	MLE $\alpha_0 \neq \alpha_1$	MRC/IR
α_0		.058 (.007)	.061 (.007)	.035 (.015)
α_1		.058 (.007)	.309 (.174)	.395 (.091)
married (0 if unmarried)	-.108 (.049)	-.073 (.077)	-.103 (.100)	-.161 (.191)
last grade attended	.026 (.009)	.063 (.015)	.080 (.026)	.052 (.043)
age	-.022 (.003)	-.028 (.005)	-.033 (.007)	-.035 (.021)
union membership	-.434 (.061)	-.707 (.148)	-.811 (.199)	-.794 (.503)
earnings per week	-.001 (.0001)	-.003 (.0004)	-.004 (.001)	-.003 (.0015)
western region	.214 (.054)	.301 (.086)	.367 (.127)	.367 (---)
constant	.051 (.162)	.171 (.259)	.581 (.495)	---
log likelihood number of obs.	-1958.1 5221	-1941.4 5221	-1940.9 5221	---

Notes:

1. MRC/IR coefficient estimates and their standard errors have been rescaled so as to be comparable to the other coefficients.

2. The standard error for the variable *Western* can not be estimated because one coefficient must be held fixed when estimating the standard errors for the MRC estimator.

Table IX: Simulation
ϵ drawn from a standard normal

X_1 is drawn from a lognormal distribution
X_2 is a dummy variable
X_3 is drawn from a uniform distribution

Sample Design		Sample generated with 10% misclassified observations		
		Probit	MLE	MRC/IR scaled
-0.80	β_0	-0.539 (.052)	-1.01 (.137)	---
0.20	β_1	0.139 (.011)	0.235 (.026)	0.200 ---
1.50	β_2	0.952 (.046)	1.69 (.164)	1.44 (.036)
-0.70	β_3	-0.446 (.023)	-0.729 (.060)	-0.594 (.203)
0.106	α_0	assume α_0=0	0.117 (.015)	0.116 (.020)
0.096	α_1	assume α_1=0	0.111 (.037)	0.128 (.108)
	N	5000	5000	5000

Appendix

I. Concavity

The conditions for a probit log likelihood function, with argument z, to have a negative second derivative are:

$$\alpha(z^2-1) - (1-2\alpha)\exp(-.5z^2)\sqrt{2\pi} < 0$$

and,

$$(1-\alpha)(z^2-1) - (2\alpha-1)\exp(-.5z^2)\sqrt{2\pi} < 0$$

For the logit function:

$$\alpha-1 + \alpha\exp(-2z) < 0$$

and,

$$-\alpha + (1-\alpha)\exp(-2z) < 0$$

II. Derivative of Estimated Coefficients with respect to the Misclassification Parameter

The log likelihood can be transformed:

$$\begin{aligned} &E[\,E[\,l(\beta,x,y)\,|X]\,] = \\ &E[\ln F(x\beta)\cdot pr(y=1\,|x,\beta_0,\alpha_0) + \ln(1-F(x\beta))\cdot pr(y=0\,|x,\beta_0,\alpha_0)] \end{aligned} \tag{1}$$

Notice that the probability of observing a particular y value depends on the true parameters of the model, β_0 and α_0. By collecting terms, the expression above can be rewritten,

$$= E_{\alpha=0,\beta}l(\beta,x,y) + \alpha E[(\ln F(x\beta) - \ln(1-F(x\beta)]\,(1-2F(x\beta_0)) \tag{2}$$

The derivative with respect to β is then set equal to zero to form the usual first order condition. Since we want to know how the optimal estimated β varies with α, we take another derivative with respect to α. Call the first term of equation (2) J_1, and the second term, αJ_2, and use the product rule to calculate,

$$\frac{\partial^2 J_1}{\partial\beta\partial\beta}\Big|_{\alpha=0}\cdot\frac{\partial\beta}{\partial\alpha}\Big|_{\alpha=0} + \frac{\partial J_2}{\partial\beta}\Big|_{\alpha=0} + \alpha\frac{\delta J_2\,d\beta}{\partial\beta\,d\alpha}\Big|_{\alpha=0} = 0 \tag{3}$$

Since we are interested in evaluating this expression at $\alpha=0$, the last term drops out and we can isolate the expression of interest:

$$\frac{\partial \beta}{\partial \alpha}\Big|_{\alpha=0} = -\left[\frac{\partial^2 J_1(\beta(\alpha))}{\partial \beta\, \partial \beta}\right]^{-1} \cdot \frac{\partial J_2}{\partial \beta} \tag{4}$$

Note that the first term in equation (4) is simply the inverted information matrix for the likelihood without misclassification. In the case where F(.) is the probit functional form, equation (4) will equal:

$$\begin{aligned} &= -\left[\frac{\phi^2 X_i' X_i}{\Phi(1-\Phi)}\right]^{-1} \frac{\phi X_i \cdot (1-2\Phi_0)}{\Phi(1-\Phi)} \\ &= \frac{1-2\Phi(X_i\beta)}{\phi(X_i\beta)X_i} \end{aligned} \tag{5}$$

which can be evaluated for different X distributions.

Additionally, this expression is bounded if X has bounded second moments, a standard condition for consistency of probit estimates. Use a condition that (F'/F) is bounded by $c(1+|z|)$ to show that the estimates of β will change at a finite rate as α increases from zero.

III. Non-Block Diagonality of the Probit Information Matrix

(symmetric misclassification probabilities)

The true log likelihood is equation (7) from page four of the paper:

$$L = y \ln[\alpha + (1-2\alpha)\Phi(X\beta)] + (1-y)\ln[(1-\alpha) + (2\alpha-1)\Phi(X\beta)] \tag{1}$$

The information matrix off-diagonals are:

$$\frac{\partial^2 L}{\partial \beta \partial \alpha} = \frac{\phi(X\beta)X}{[(1-\alpha) + (2\alpha-1)\Phi(X\beta)]^2} - \frac{\phi(X\beta)X}{[\alpha + (1-2\alpha)\Phi(X\beta)]^2} \tag{2}$$

In expectation these off-diagonals are not equal to zero:

$$E\left[\frac{\partial^2 L}{\partial \beta \partial \alpha}\right] = \frac{\phi(X\beta)X}{(1-\alpha) + (2\alpha-1)\Phi(X\beta)} - \frac{\phi(X\beta)X}{\alpha + (1-2\alpha)\Phi(X\beta)} \tag{3}$$

IV. Identification of Berkson Log Odds with Three Classes, Misclassification, and the Logit Functional Form

Define the following:

π_0 = the probability of correct classification of an observed 0.
- α_{20} = probability a 2 is misclassified to be a 0.
- α_{10} = probability a 1 is misclassified to be a 0.

π_1 = the probability of correct classification of an observed 1.
- α_{21} = probability a 2 is misclassified to be a 1.
- α_{01} = probability a 0 is misclassified to be a 1.

π_2 = the probability of correct classification of an observed 2.
- α_{02} = probability a 0 is misclassified to be a 2.
- α_{12} = probability a 1 is misclassified to be a 2.

The probabilities of observing any particular outcome must add to one, resulting in the following restriction:

$$\pi_i + \alpha_{ij} + \alpha_{ik} = 1 \qquad \forall\, i \neq j \neq k \tag{1}$$

Then the probability of observing the outcome 0 can be expressed:

$$Pr(y=0) = \frac{\pi_0 e^{X_0\beta}}{\sum_{j=0}^{2} e^{X_j\beta}} + \frac{\alpha_{20} e^{X_2\beta}}{\sum_{j=0}^{2} e^{X_j\beta}} + \frac{\alpha_{10} e^{X_1\beta}}{\sum_{j=0}^{2} e^{X_j\beta}} \tag{2}$$

Similarly, the probabilities of observing outcomes 1 and 2 can be expressed:

$$Pr(y=1) = \frac{\pi_1 e^{X_1\beta}}{\sum_{j=0}^{2} e^{X_j\beta}} + \frac{\alpha_{21} e^{X_2\beta}}{\sum_{j=0}^{2} e^{X_j\beta}} + \frac{\alpha_{01} e^{X_0\beta}}{\sum_{j=0}^{2} e^{X_j\beta}} \tag{3}$$

$$Pr(y=2) = \frac{\pi_2 e^{X_2\beta}}{\sum_{j=0}^{2} e^{X_j\beta}} + \frac{\alpha_{12} e^{X_1\beta}}{\sum_{j=0}^{2} e^{X_j\beta}} + \frac{\alpha_{02} e^{X_0\beta}}{\sum_{j=0}^{2} e^{X_j\beta}} \tag{4}$$

The probabilities can be reinterpreted as shares and any two can be written in ratio form. The log odds of ups 0 and 1 will be:

$$\ln(\frac{S_0}{S_1}) = \ln(\pi_0 e^{X_0\beta} + \alpha_{20} e^{X_2\beta} + \alpha_{10} e^{X_1\beta}) - \ln(\pi_1 e^{X_1\beta} + \alpha_{01} e^{X_0\beta} + \alpha_{21} e^{X_2\beta}) \tag{5}$$

It is straightforward to write out the expression for other pairs of ups, so they are not included here. The condition that the probabilities of seeing any particular outcome must add to one results in overidentification of equation (5). Condition (1) places two restrictions on the six coefficients one would estimate in equation (5).

V. Proof of $N^{1/3}$ Convergence for Isotonic Regression when Index Parameters are Estimated with $N^{1/2}$ Convergence

Assume the regularity conditions of Groeneboom (1993). For each n and index value ν define $\ln(\nu)$ as the positive distance to the closest jump and define $j_n(\nu)$ as the (vertical) size of the jump. We now use two facts to establish the proof: (1) $\nu - \hat{v} = O_p(N^{-1/2})$ from the property of the MRC estimator and (2) the total cumulative size of the jumps cannot exceed 1.0 because we are estimating a cumulative distribution by IR. We can then calculate:

$$\begin{aligned} & P_r\left(\mid F_n(v) - F_n(\hat{v}) \mid \geq \epsilon\, n^{-1/3} \right) \\ \leq & \int \frac{1}{ln(v)\, n^{1/2}}\, j_n(v)\, \frac{1}{\epsilon}\, n^{1/3}\, w(v) dv \\ \leq & \frac{K}{\epsilon}\, n^{-1/6} \end{aligned} \tag{1}$$

where $w(v)$ is the density of v and $\int \frac{j_n(v)}{ln(v)} w(v) dv \leq K.$

Thus, because the total cumulative size of the jumps cannot exceed 1.0, either jumps are relatively large but far apart, or jumps are frequent but small. The net result is we can derive the bound. To complete the proof we calculate

$$n^{1/3}\,[F_n(\hat{\nu}) - F(\nu)] = n^{1/3}\,[F_n(\hat{\nu}) - F_n(\nu)] + n^{1/3}[F_n(\nu) - F(\nu)] \tag{2}$$

The first term on the right hand side, $n^{1/3}\,[F_n(\hat{\nu}) - F_n(\nu)]$ is order $o_p(1)$ by equation (1) while the second term $n^{1/3}\,[F_n(\nu) - F(\nu)]$ is order $O_p(1)$ with the asymptotic distribution as proven by Groeneboom (1993). Thus, we can treat the estimated $\hat{\beta}$ parameter as known in computing the asymptotic distribution of the IR.

Figure 4: MLE and MRC/IR Estimates

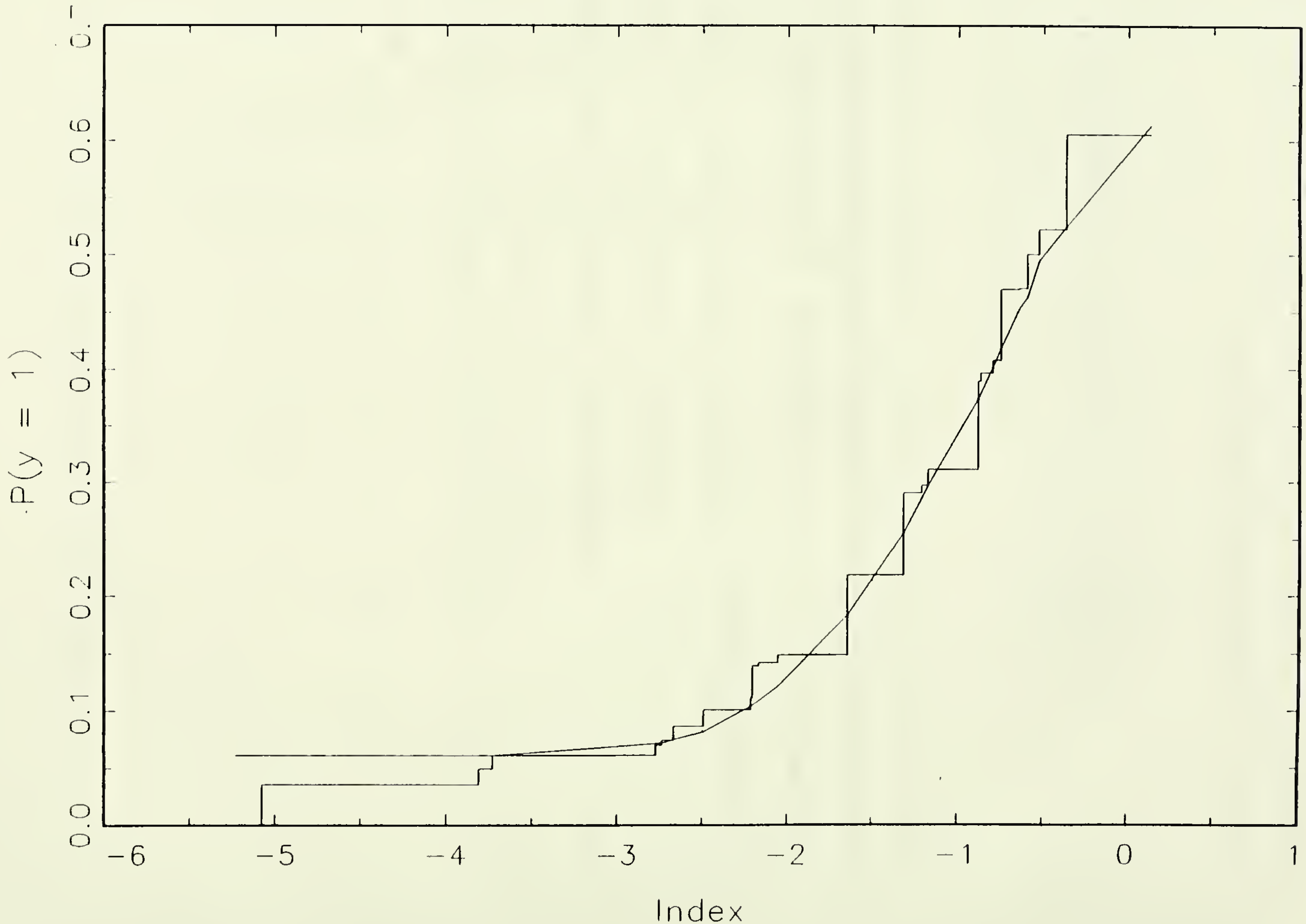


Figure 5: Smoothed Confidence Intervals for IR Step Function

(+/− two standard errors)

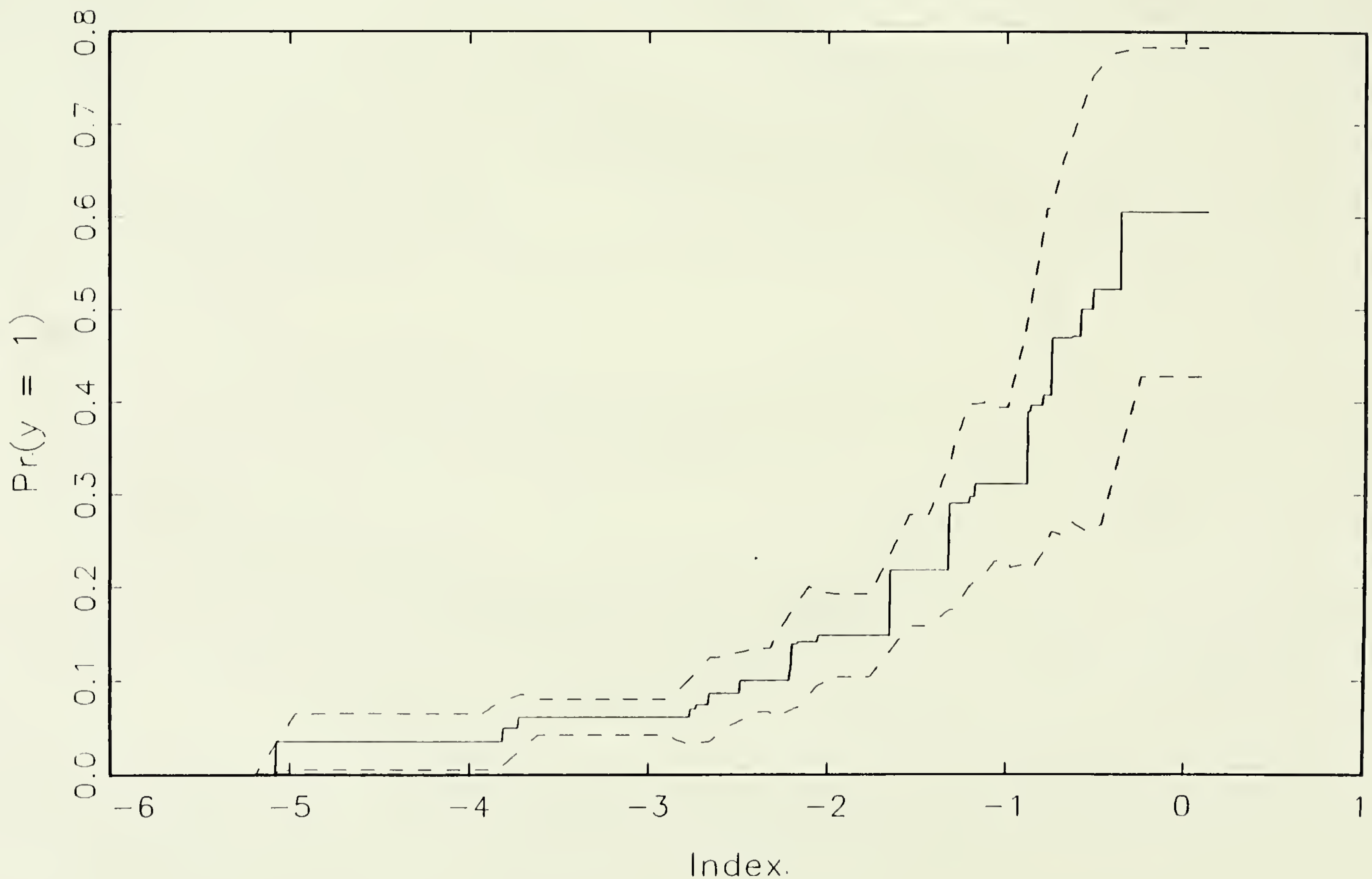


Figure 6A: IR and Kernel Estimation of CPS Job Change Data

(fixed–window method)

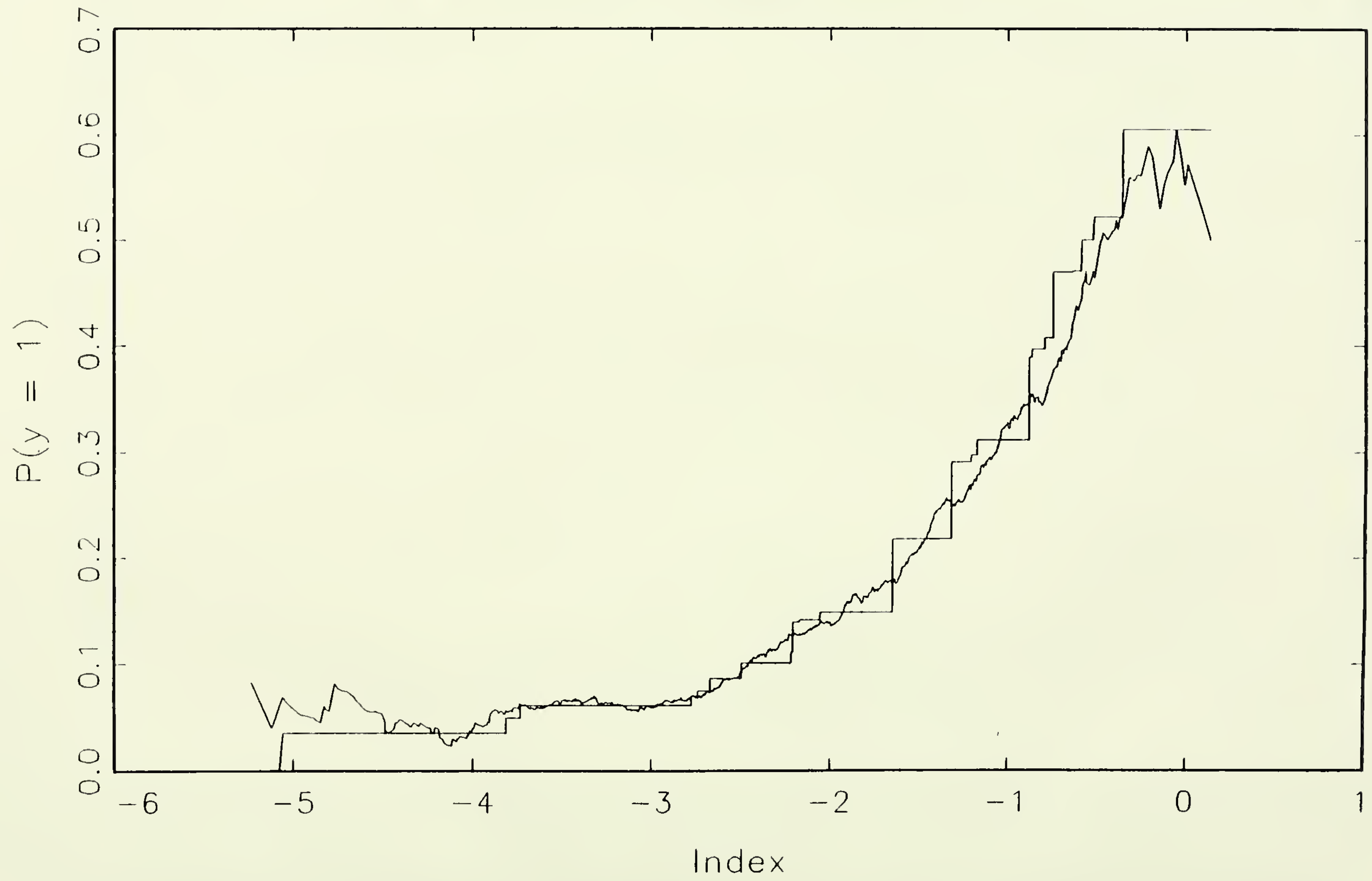


Figure 6B: IR and Kernel Estimation of CPS Job Change Data

(k–neighbor method)