Misclassified Regressors in Binary Choice Models*

Aprajit Mahajan†

Department of Economics,

Stanford University.

October 1, 2003

Abstract

This paper examines the effect of mismeasured discrete regressors in binary choice models. I examine plausible scenarios for the nature of the measurement error and discuss identifiability and estimation under various sets of semiparametric assumptions. Under a minimal set of assumptions, the model is only partially identified and I derive bounds for some of the parameters of interest. If the probability of misclassification is conditionally independent of the other regressors, the model is point identified and I propose a $\sqrt{n}$ consistent, asymptotically normal semiparametric two-step estimator under this set of conditions. If, however, the misclassification rates are not independent of the other regressors, further information is required. When an additional measurement on the mismeasured regressor is available I develop a $\sqrt{n}$ consistent, asymptotically normal estimator using the method of sieves without specifying the relationship between the probability of misclassification and the other explanatory variables. Monte Carlo simulations suggest good finite sample properties of the estimators and the method is illustrated with a study on the effect of unionization on the receipt of health benefits using data from the Current Population Survey.

*I would like to thank Han Hong, Bo Honoré and Elie Tamer for their tireless encouragement and advice. I have also benefited from discussions with Xiaohong Chen, Henry Farber and Ernst Schaumburg. I would also like to thank the participants in the Princeton Microeconometrics Reading Group.

†Mailing Address: 1 Fisher Hall, Princeton, NJ 08544-1021. Phone: (609) 430-8483 Fax: (609) 258-6419. Email: amahajan@princeton.edu. Revisions available at www.princeton.edu/~amahajan.

1 Introduction

Measurement error in Non-Linear models introduces difficulties whose solutions require techniques that are quite distinct from those usually called for in linear models. Ordinary instrumental variable estimation is no longer feasible and estimation typically proceeds by making strong distributional assumptions and/or introducing further information. The evidence as is available on the validity of some of these assumptions suggests that they may not be a reasonable approximation of the true data generating process. In particular, the assumptions of the classical measurement error model – that the measurement error is independent of the true value and of other variables in the model[1] – have been shown not to hold in a number of studies. This paper attempts to shed light on the effect of relaxing these assumptions in non-linear models by examining in detail the case of misclassified regressors in binary choice models.

When a mismeasured variable is binary (or more generally has a known finite support) – commonly referred to as the problem of misclassification – the independence assumption between the measurement error and the true values of the variable, invoked by the classical model for measurement error, is particularly untenable. More generally, the phenomenon of negative correlation between the errors and the true values (referred to as "mean reversion") has been found to exist for a number of quantities of interest. Evidence of such a pattern has been found in earnings data by Bound and Krueger [13] and Bollinger [10].[2] In their article on measurement error in survey data, Bound, Brown, and Mathiowetz [12] also report similar dependencies for a number of variables including hours worked, union status, welfare program participation, and employment status.

A second assumption that is usually imposed on error models is the independence between the measurement error and the other explanatory variables in the model. Again, there is evidence from studies that suggests otherwise. For instance, Bollinger and David [11] noted that the probability of misreporting AFDC status was

[1] Unlike linear models where uncorrelatedness suffices for identification (instrumental variable approaches for instance), most non-linear models require some sort of independence assumption. The classical measurement error or "errors-in-variables" model in these contexts is usually given by $x = x^* + \varepsilon$ and ε is assumed independent of the unobserved x^* and the other regressors in the model.

[2] Also see the caveat therein.

highest amongst the poorest households. Mellow and Sider [40] concluded that over-reporting of hours worked varied systematically with education and race. Summarizing work on unemployment data Bound, Brown, and Mathiowetz [12] report that misclassification rates tended to vary by age and sex.[3]

This paper examines the effect of relaxing these assumptions on the identifiability and estimation of the parameters in a binary choice model. The object of interest is the effect of a binary random variable x^*on the probability of a dichotomous outcome y while controlling for other explanatory variables z. The econometrician observes a random sample $\{y_i, x_i, z_i\}_{i=1}^n \equiv \{w_i\}_{i=1}^n$ where x_i is an error-ridden measure of x_i^*. Since the unobserved variable is binary, the measurement error is necessarily correlated with the true value x^*. In addition, I also study the possibility that the measurement error is not independent of the other explanatory variables z in the model, something that has not been explored in any detail previously.

I first study the problem in a non-parametric context and then proceed to a semiparametric setting and also discuss fully parametric specifications. I begin with minimal assumptions and characterize bounds on the objects of interest before adding further information to the model that tightens the bounds and enables identification and in these cases I develop semiparametric estimation strategies for the parameters of interest.[4]

Section 2 provides a review of related work in the non-linear measurement error and misclassification literature following which we begin our first study of the problem. Section 3 constructs sharp bounds on the probabilities of interest under minimal assumptions. The subsequent sections add further information and discuss identification and estimation of the parameters of interest. Section 4 discusses identification and estimation when the probability of misclassification is independent of the other regressors. When this no longer true, I add further information to identify the model and develop a sieve based estimator for the parameters of interest in Section 5. Section 6 studies the finite sample properties of the proposed estimators using a series of Monte Carlo simulations and Section 7 illustrates the methods using data from the Current Population Survey. Section 8 concludes.

[3] Their paper also serves as an excellent summary of the current empirical evidence on the nature of measurement error in survey data.

[4] Most of the results in the paper also apply to discrete regressors in general with little modification.

2 Related Literature

There has been a fair amount of work on measurement error in non-linear models over the past two decades. A useful survey of the statistics literature can be found in Carroll, Ruppert and Stefanski [15]. Hausman, Newey, Ichimura and Powell [28] address the issue of measurement error in polynomial regression models by introducing auxiliary information either in the form of an additional indicator or an instrument. Their identification and estimation strategy relies on the classical measurement error model requiring independence between the measurement error and the true value of the mismeasured variable as well as independence between the error and the correctly measured covariates.[5] Tong Li [36] and Schennach [47] propose an estimation strategy under similar assumptions for more general classes of nonlinear models. Mahajan [38] adopts a similar estimation strategy while accounting for dependencies between the mismeasured variable and the exogenous regressors in parametric fashion.

Another class of models achieves identification by supplementing the classical independence assumptions with a distributional assumption on the measurement error. Taupin [51] proposes a consistent estimator for a nonlinear regression model by imposing normality on the measurement error while Hong and Tamer [29] derive estimators for a broader class of moment based models by assuming a Laplacean distribution for the error term. Another set of papers rely on the availability of a validation data set (that is, a sample where both the mismeasured as well as the true value of the variable are observed). Carroll and Wand [16] use validation data to estimate a logistic regression with measurement error, as do Lee and Sepanski [34] for a non linear regression model.

All these papers assume some variant of the classical additive measurement error model and independence between the correctly measured covariates and the measurement error. Work departing from the classical measurement error assumptions can be found in Horowitz and Manski [31]. In their work, the observed data are a mixture of the variable of interest and another random variable and the error is allowed to depend upon the variable of interest. Chen, Hong and Tamer [17] relax the independence assumptions between errors and the true values of the variables in the presence of an auxiliary data set and derive distribution theory for estimates based

[5] The full independence assumption is imposed, however, only on one of the error terms in the repeated measurements.

on non-linear unconditional moment restrictions. Imbens and Hyslop [32] interpret the problem within a prediction model in which the errors are correlated with the true values and discuss the bias introduced in estimation procedures if measurement errors followed their model.

Measurement error in binary variables is necessarily non-classical in nature since the error term is perforce negatively correlated with the true outcome. If the misclassified variable is a response variable, Horowitz and Manski's [31] work can be used to derive bounds for the unidentified parameters of the conditional distribution of interest. Hausman, Abrevaya and Scott-Morton [27] examine the effect of a mismeasured left hand side binary variable within a maximum likelihood as well as a semiparametric framework. The issue of misclassified binary regressors was first addressed by Aigner [2] and subsequently by Bollinger [9] in the context of a linear regression model. In the absence of further information, they show that the model is not identified and both papers obtain sharp bounds for the parameters of interest. With the addition of further information in the form of a repeated measurement, a recent paper by Black, Berger, and Scott [8] obtained point identification for the slope coefficient in a univariate regression using a method of moments approach. An essentially similar approach was used by Kane, Rouse, and Staiger [33] to study the effect of mismeasured schooling on returns to education and Card [14] studies the effect of unions on wages taking misclassification of union status explicitly into account using a validation data set. All these papers make the assumption that misclassification rates are independent of the other regressors in the model.[6]

3 Bounds

3.1 Bounds for the Non-parametric case

I first consider the problem in a non-parametric setting with no functional form assumptions for the binary choice model. I do, however, place two restrictions on the nature of the measurement error and these are maintained

[6] Although, Hausman et al. [27] discuss a case where this is not true.

throughout the paper. The first assumes that the outcome y is independent of x conditional on the correctly measured random variable x^* and the other explanatory variables z. Formally,

$$P(y = 1|x^*, x, z) = P(y = 1|x^*, z) \tag{A1}$$

In the literature this is referred to as the assumption of non-differential measurement error and x is known as a "surrogate" for x^*. In the more familiar linear context this is analogous to the assumption that the error term in the outcome equation is independent of the measurement error in the incorrectly measured regressor, conditional on all the regressors in the model. This implies that the measurement error itself is uninformative about the response given information on the truth and z. The conditional statement is important since the misclassification rates may in fact be informative about responses through their correlation with other variables in the model.

The second restriction limits the extent of the measurement error by requiring that the probability of a correct classification be greater than that of an incorrect one, i.e.,

$$P\left(x = 1|x^* = 1, z\right) > P\left(x = 1|x^* = 0, z\right) \quad \text{a.e. } z \tag{A2}$$

This ensures that the unobserved variable x^* is positively correlated with its surrogate x. This assumption ensures that the direction of the effect of the surrogate on the response is the same as the effect of the unobserved true regressor.

I place no further restrictions on the form of the measurement error in this section. In particular, I do not require, unlike most papers on the subject, that the measurement error be independent of the other explanatory variables in the model. It is therefore possible that the misclassification rates are systematically related to one or more of the z variables. I do, however, require that the relationship between them remains stable in the sense of (A2).

In the absence of further information, (A1) in not identified, however, we can bound it. In order to discuss

the bounds, some notation is helpful. Define

$$\begin{aligned} m_1^*(z) &\equiv P(y=1|x^*=1,z) \quad (1)\\ m_0^*(z) &\equiv P(y=1|x^*=0,z) \quad (2)\end{aligned}$$

$$\begin{aligned} m_1(z) &\equiv P(y=1|x=1,z) \quad (3)\\ m_0(z) &\equiv P(y=1|x=0,z) \quad (4)\end{aligned}$$

Lemma 1 *Let* $g_0(z) \equiv 1[m_0(z) \geq m_1(z)]$ *and* $g_1(z) \equiv 1[m_0(z) \leq m_1(z)]$ *Then, Under (A1) and (A2)*

$$m_1(z) \begin{matrix}\geq \\ \leq\end{matrix} m_0(z) \Rightarrow m_1^*(z) \begin{matrix}\geq \\ \leq\end{matrix} m_0^*(z) \quad (5)$$

$$m_1(z)g_1(z) + m_0(z)g_0(z) \leq m_1^*(z)g_1(z) + m_0^*(z)g_0(z) \leq 1 \quad (6)$$

and

$$0 \leq m_1^*(z)g_0(z) + m_0^*(z)g_1(z) \leq m_1(z)g_0(z) + m_0(z)g_1(z) \quad (7)$$

In the absence of other information the bounds in (6) and (7) are sharp.

The proof is in the appendix and the idea is quite straightforward. Each of $\{m_1, m_2\}$ is a linear combination of the unknown probabilities $\{m_1^*, m_2^*\}$. (A2) provides us with information on the relative weights attached to each of these and this in turn allows us to infer bounds on $\{m_1^*, m_2^*\}$ based on the observed $\{m_1, m_2\}$ and are reported above. Suppose for instance that $m_1(z) > m_0(z)$, then the bound for $m_1^*(z)$ will be $[m_1(z), 1]$ and for $m_0^*(z), [0, m_0(z)]$.

3.2 Binary Choice Model

I continue to maintain (A2) but specialize (A1) by parameterizing it as

$$P(y=1|x^*,z,x) = F(\beta_0 + \beta_1 x^* + \beta_2 z) \quad (A1')$$

where F is a known monotone increasing function and for the moment I assume that in addition to the intercept term there is only one other explanatory variable so that the object of interest $\beta = \{\beta_0, \beta_1, \beta_2\} \in \mathbb{R}^3$.

The identification results in this section depend importantly upon the assumptions on the support of z (which is henceforth denoted by S_z) with different assumptions yielding dramatically different conclusions on the bounds of interest. At one extreme, if z is binary there is no finite bound for β_2 at all, whereas if $S_z = \mathbb{R}$, β_2. is point identified. In contrast, as lemma (2) states the sign of β_1 is always identified regardless of the assumptions on S_z.

Lemma 2 *Under (A1′) and (A2) suppose for any $z \in S_z$*

$$m_1(z) \underset{\leq}{\geq} m_0(z) \tag{8}$$

then

$$m_1(z) \underset{\leq}{\geq} m_0(z) \ \forall \ z \ \in S_z \tag{9}$$

and

$$\beta_1 \underset{\leq}{\geq} 0 \tag{10}$$

The result follows straightforwardly from (1) and also has the virtue of providing a check on the validity of $(A1')$ as a model for (A1) since the sets $\{z : m_1(z) > m_0(z)\}$ and $\{z : m_1(z) < m_0(z)\}$ cannot both have positive probability under $(A1')$.

Consider first the simple case where $S_z \in \{0, 1\}$. Then, by using the bounds obtained in Lemma (1) I obtain bounds on the parameter vector β

Lemma 3 *Suppose (A2) and (A1′) hold and that $S_z \in \{0, 1\}$ Suppose w.l.o.g. that $m_1(z) > m_0(z)$. Then*

$$\beta_1 > \max_{S_z} \left\{ F^{-1}(m_1(z)) - F^{-1}(m_0(z)) \right\} \tag{11}$$

$$\beta_0 \leq F^{-1}(m_0(0)) \tag{12}$$

$$\beta_2 \in \mathbb{R} \tag{13}$$

In the absence of further information these bounds are sharp.

More generally with a richer support of z, the identification possibilities can be characterized using the method

proposed by Manski and Tamer [39]. Given $(A1')$

$$P(y=1|x,z) = F\left(\beta_0 + \beta_2 z\right) P\left(x^* = 0|x,z\right) + F\left(\beta_0 + \beta_1 + \beta_2 z\right) P\left(x^* = 1|x,z\right) \tag{14}$$

So that

$$\min\left\{F\left(\beta_0 + \beta_2 z\right), F\left(\beta_0 + \beta_1 + \beta_2 z\right)\right\} \leq P(y=1|x,z) \leq \max\left\{F\left(\beta_0 + \beta_2 z\right), F\left(\beta_0 + \beta_1 + \beta_2 z\right)\right\} \tag{15}$$

For any $b \in \Theta$ consider the set

$$\begin{aligned} V(b) &= \left\{(x,z) : \min\left\{F\left(b_0 + b_2 z\right), F\left(b_0 + b_1 + b_2 z\right)\right\} > P(y=1|x,z)\right\} \cup \\ &\quad \left\{(x,z) : \max\left\{F\left(b_0 + b_2 z\right), F\left(b_0 + b_1 + b_2 z\right)\right\} < P(y=1|x,z)\right\} \end{aligned}$$

Then any b such that $P\left(V(b)\right) > 0$ is clearly inadmissible since it violates (15). The set

$$B^* = \left\{b \in \Theta : P(V(b)) = 0\right\}$$

is the set of elements in Θ that are observationally equivalent to β and the aim is to characterize this set. Following Lemma 2 in Manski and Tamer [39], I define a objective function based on (15) such that every $b \in B^*$ minimizes this function. First, define

$$\begin{aligned} m(b,z) &= \max\left\{F\left(b_0 + b_2 z\right), F\left(b_0 + b_1 + b_2 z\right)\right\} \\ n(b,z) &= \min\left\{F\left(b_0 + b_2 z\right), F\left(b_0 + b_1 + b_2 z\right)\right\} \\ \eta(x,z) &= P(y=1|x,z) \end{aligned}$$

The objective function is given by

$$Q_0(b) = E\left[I\left(m(b,z) < \eta\left(x,z\right)\right)\left(m(b,z) - \eta\left(x,z\right)\right)^2 + I\left(n(b,z) > \eta\left(x,z\right)\right)\left(n(b,z) - \eta\left(x,z\right)\right)^2\right]$$

and by Lemma 2 in [39]of we obtain that no $b \notin B^*$ will minimize Q_0 and that every $b \in B^*$ will. This line of argument reveals that β_2 is identified when z has unbounded support. This result follows from Proposition 4 in Manski and Tamer [39] and is proved in Appendix A.0.1.

The conclusion from the section on bounds is therefore that while certain features of the model may be identified further assumptions are required to identify all the parameters of interest in the model.

4 Distributional Assumption I

Section (3.2) demonstrated that the model $(A1')$ under only (A2) is only partially identified. This section and the next two explore the role of different types of additional information in achieving identification and related estimation strategies. In this section I add information by assuming that the probability of misclassification is independent of the other explanatory variables in the model.[7] Formally,

$$P\left(x = 1|x^*, z\right) = P\left(x = 1|x^*\right) \quad \text{a.e.} \tag{16}$$

This is assumed, for instance, in [27] and [9] and implies that the misclassification rates are completely characterized by two constants $\alpha_0 \equiv P\left(x = 1|x^* = 0\right)$ the "false positive" rate and $\alpha_1 \equiv P\left(x = 0|x^* = 1\right)$ the "false negative" rate which by (A2) must satisfy

$$\alpha_0 + \alpha_1 < 1 \tag{17}$$

[7] In a work in progress I explore identification and estimation possibilities when only one of the elements of the z vector is conditionally independent of x given all the other regressors and x^*.

Some papers also assume (for instance Card [14]) that $\alpha_0 = \alpha_1$ and the problem is referred to as one of symmetric misclassification but I shall not purse this simplification here. Under $(A1')$,(17) and (16) the expected log likelihood of $\{y, x|z\}$ where z is a r dimensional vector valued random variable is

$$Q_0(a_0, a_1, b) = E\left[\log\left(F_1^y (1-F_1)^{1-y} (1-a_1)^x a_1^{1-x} P^* + F_0^y (1-F_0)^{1-y} (1-a_0)^{1-x} a_0^x (1-P^*)\right)\right] \tag{18}$$

$$F_{x^*} = F(b_0 + b_1 x^* + b_2 z) \quad x^* \in \{0, 1\} \tag{19}$$

$$P^*(z, a_0, a_1) = \frac{P(x=1|z) - a_0}{1 - a_0 - a_1} = \frac{\xi(z) - a_0}{1 - a_0 - a_1} \tag{20}$$

I propose to estimate $\theta = (a_0, a_1, b)$ by maximizing the sample version of (18). The estimation technique is semiparametric (or perhaps more accurately, quasi-likelihood) in the sense that I do not impose a parametric form for the distribution of x conditional on z but instead approximate it by a smooth function using non-parametric methods. The sample objective function is given by

$$Q_n(\theta) = \frac{1}{n} \sum \log\left(F_1^{y_i} F_1^{1-y_i} (1-a_1)^{x_i} a_1^{1-x_i} \hat{P}_i^* + F_0^{y_i} F_0^{1-y_i} (1-a_0)^{1-x_i} a_0^{x_i} (1-\hat{P}_i^*)\right) w(z_i) \tag{21}$$

where

$$\hat{P}_i^* = \frac{\hat{P}(x=1|z_i) - a_0}{1 - a_0 - a_1} \tag{22}$$

$$\hat{\xi}(z_i) = \sum_{j=1}^{n} \frac{K_h(z_j - z_i)}{\sum_{k=1}^{n} K_h(z_k - z_i)} x_j \tag{23}$$

where $\hat{\xi}(z_i)$ is a kernel estimator of $P(x=1|z_i)$, h is a positive smoothing parameter that goes to zero as the sample size increases, $K_h(z) = \frac{1}{h^r} K\left(\frac{z}{h}\right)$ for a given kernel K and w is a weighting function (with compact support W_z) whose role is discussed below. The estimator $\hat{\theta}$ is then given by

$$\hat{\theta} = \arg\min_{\Theta} Q_n(\theta) \tag{24}$$

4.1 Identification

I state a sufficient condition for identification that essentially requires sufficient variation in the conditional probability of x^* given z. In addition, I also require that this relationship be sufficiently different from the relationship between the response y and z in a sense made clear below. I also state a separate set of more intuitive "Instrumental Variable"- like conditions that also guarantee identification.

Consider two points $\theta = (a, b), \bar{\theta} = \left(\bar{a}, \bar{b}\right) \in \Theta$ and $\theta \neq \bar{\theta}$[8] and let $\kappa = (1 - a_0 - a_1) / (1 - \bar{a}_0 - \bar{a}_1)$. The critical condition for identification is that for some $B \subseteq S_z$ with $P(B) > 0$ and $\forall z \in B$

$$\begin{aligned} &\xi(z)\left[F_1(b) - F_0(b) - \kappa\left(F_1\left(\bar{b}\right) - F_0\left(\bar{b}\right)\right)\right] \neq \\ &\kappa\left(\left(1 - \bar{a}_1\right) F_0\left(\bar{b}\right) - \bar{a}_0 F_1\left(\bar{b}\right)\right) - \left(\left(1 - a_1\right) F_0(b) - a_0 F_1(b)\right) \end{aligned} \tag{25}$$

The result is stated as follows

Lemma 4 *Suppose that (A1′), (16), (17), (25) hold, (i)* $\beta_1 \neq 0$*,(ii)* $Var[(P(x^* = 1|z)] > 0$ *and (iii)* $E[1, z]'[1, z] > 0$*. Then, the model is identified*

The necessity of (17) is easily seen. In its absence for any $\theta = (a_0, a_1, b_0, b_1, b_2)$

$$P(y = 1, x|z, a_0, a_1, b_0, b_1, b_2) = P(y = 1, x|z, 1 - a_1, 1 - a_0, b_0 + b_1, -b_1, b_2) \tag{26}$$

The necessity of the other conditions and the proof for the lemma are detailed in the appendix A.0.1..

There is a simpler set of conditions that guarantee identification that are also perhaps more intuitive since they are similar to the usual instrumental variable requirements. They rely upon the existence of a variable v that is unrelated to the response but is related to the misclassified regressor x^*. Specifically, suppose z can be

[8] At least one component of both a and b must differ, since it can be shown that as long as $\beta_1 \neq 0$ then α is identified if and only if β is identified. Note that in principle one can deduce whether β_1 equals 0 using the results from Lemma (2). In fact if z is discrete, we can use the standard tests for equality of means.

partitioned into (v, z_2) and b_2 into (b_v, b_{z_2})such that

C1 $\beta_v = 0$ so that $P(y = 1|x^*, v, z_2) = P(y = 1|x^*, z_2)$ (exclusion restriction)

C2 $P(x = 1|x^*, v, z_2) = P(x = 1|x^*)$ and that

C3 $Var(P(x^* = 1|z_2 = c, v)) > 0$ for some $c \in S_{z_2}$

Lemma 5 *Suppose that (A1′), (16), (17), C1-C3 hold,* $\beta_1 \neq 0$ *and* $E[1, z]'[1, z] > 0$ *. Then, the model is identified*

Identification is achieved here because the variation in $\xi(z_2, v)$ ensures (25) holds. One case of interest is when v is another surrogate for x^*. This occurs when there is a repeated (error ridden) measurement on x^*, for instance from a reinterview survey or some other data source. In this case $C2$ will be satisfied if, conditional on the truth and the other regressors, the two measurements are independent of each other. In the sequel I shall assume that the model is identified although note that identification will need to be established on a case by case basis for particular distributions of $\{y, x, z\}$ and choices of F.

4.2 Consistency

The consistent estimation of $\theta \in \Theta \subseteq \mathbb{R}^{r+4}$ (recall that r is the dimension of z) follows from a standard theorem on the consistency of M-estimates. The conditions imposed by (17) on the misclassification probabilities imply that

$$\alpha_0 < P(x = 1|z) < 1 - \alpha_1 \quad a.e. \tag{27}$$

Therefore $\alpha_0 \in [0, \inf_z \mu(z)]$ and $\alpha_1 \in [0, 1 - \sup_z \mu(z)]$ (where $\mu(z) \equiv E(x|z)$) almost everywhere. In the sample I impose the restriction that

$$\begin{aligned} 0 &\leq a_0 \leq \inf_{W_z} \hat{P}(x = 1|z) \\ 0 &\leq a_1 \leq 1 - \sup_{W_z} \hat{P}(x = 1|z) \end{aligned} \tag{28}$$

As long as $P(x=1|z)$ is consistently estimable and I incorporate this restriction into the maximization procedure the asymptotic analysis in this section should remain unaffected because the constraint will be satisfied in large samples. Therefore, in the subsequent analysis I assume that the constraint is satisfied.

With the parameter set specified to be a compact subset of $\mathbb{R}^{r+4}$ with the restrictions above we can deduce the consistency of $\hat{\theta}$ defined in (24).

Theorem 6 *Suppose that (A1′), (16), (17), (25), and D1-D3 hold. Suppose in addition that (i)* Θ *is compact and (ii)* $\alpha_0 < \sup_z P(x=1|z) < 1-\alpha_1$. *Then*

$$\hat{\theta} \to \theta_0$$

The result follows from Theorem 2.1 and Lemma 2.9 of Newey and McFadden [42] and the proof is detailed in Appendix A.0.4.

4.3 Asymptotic Distribution

The asymptotic distribution of the estimator $\hat{\theta}$ follows from an application of the distribution theory for two-step semiparametric estimators as covered for instance in Newey and McFadden [42]. I show that $\hat{\theta}$ converges at the usual parametric rate $\sqrt{n}$ so that the nonparametric estimation of $P(x=1|z)$ does not lead to a slower rate of convergence. Although the nonparametric estimation does not affect the rate of convergence I find that it does, however, affect the covariance matrix of the limiting distribution.

An important caveat is that the model with no measurement error cannot be analyzed using the asymptotics derived here since the first order conditions are not defined for $\alpha_0 = 0$ or for $\alpha_1 = 0$.[9] More generally, this is a consequence of the requirement that the true parameter values lie in the interior of Θ. Relaxing this assumption is possible and has been carried out by Andrews [3] and also in the fully parametric MLE setting by Geyer [25] but I do not consider this generalization here. With the assumption that the truth lies in the interior of Θ for

[9] To see this algebraically, note that the first order order conditions in (97) and (98) have α_0 and α_1 as denominator terms.

large enough sample sizes the constraint will not be binding with (in fact with probability approaching one) I shall subsequently ignore the constraints (28) in deriving the asymptotic distribution.

To fix ideas I state the first order conditions from (24).

$$\nabla_\theta Q_n\left(\hat{\theta}\right) = \sum_{i=1}^{n} q\left(w_i, \hat{\theta}, \hat{\gamma}\right) = 0 \tag{29}$$

$$\hat{\gamma}(z) = \frac{1}{n}\sum_{i=1}^{n} \tilde{x}_i K_h\left(z - z_i\right) \tag{30}$$

where $\tilde{x}_i = [1 \; x_i]'$ so that $\hat{\gamma}\left(z\right)$ is a kernel estimate of $\gamma_0(z) \equiv f_z(z)E\left[\tilde{x}|z\right]$. This notational change for the conditional expectation[10] makes for expositional ease in the results that follow . I also adopt the convention $\nabla_x f\left(\bar{x}\right) \equiv \nabla_x f\left(x\right)|_{x=\bar{x}}$.

The estimator $\hat{\theta}$ is quite naturally viewed as a two-step estimator where I first estimate γ and then estimate θ by maximizing (24) given the first step estimate $\hat{\gamma}$. The basic idea is to deduce the asymptotic distribution of $\hat{\theta}$ by obtaining its influence function representation, taking into account the first step estimation. Since γ is infinite dimensional, the asymptotic distribution is derived using large sample theory for two-step semiparametric estimators.

The first practical problem is that since the objective function contains the reciprocal of the density of z it may be ill behaved for extreme realizations of z. This is the well-known "random denominator" problem and various authors have dealt with it using differing approaches. The most common one is to introduce a weighting function that is zero outside of a bounded set over which the density is strictly positive so that the effect of the density is "trimmed" to be zero outside a given set. Robinson [46] avoids the problem by using nearest neighbor estimators, while for instance, [30] employs a variable trimming technique where the amount of trimming is allowed to increase with sample size. I follow the first approach by introducing a weight function $w\left(z\right)$ that is zero outside a closed and bounded set W_z.

[10] Now $E(x|z) = \gamma_{02}(z)/\gamma_{01}(z) = f\left(z\right)E\left[x|z\right]/f(z)$

As detailed in the appendix, I first obtain an appropriate linearization for the first-order condition $q(w, \theta_o, \gamma)$ in γ. In particular there exists a functional $D(w, \gamma)$ that is linear in γ such that for γ close enough to γ_0

$$\|q(w, \theta_0, \gamma) - q(w, \theta_0, \gamma_0) - D(w, \gamma - \gamma_0)\| \leq C(w) \|\gamma - \gamma_0\|^2$$

The conditions below ensure that $\hat{\gamma}$ is close enough to γ_0 for n large enough in the precise sense that $\sqrt{n} \|\hat{\gamma} - \gamma_0\|^2 \to 0$

D1 There is a version of $\gamma_0(z)$ that is continuously differentiable of order $m(> r)$ and $\gamma_{02}(z) = f_z(z)$ is bounded away from 0 on W_z

D2 $\int K(u) du = 1$, and for all $j < m$ $\int K(u) (\bigotimes_{l=1}^{j} u) du = 0$

D3 The bandwidth h satisfies $nh^{2r} / (\ln n)^2 \to \infty$ and $nh^{2m} \to 0$ (so that $m > r$)

D4 There exists a $b(w), \varepsilon > 0$ such that $\|\nabla_\theta q(w, \theta, \gamma) - \nabla_\theta q(w, \theta_0, \gamma_0)\| \leq b(w) [\|\theta - \theta_0\|^\varepsilon + \|\gamma - \gamma_0\|^\varepsilon]$ and $E[b(w)] < \infty$

D5 $\hat{\theta} \to \theta_0 \in int(\Theta)$

Theorem 7 *Assume (A1′),(17),(16),(D1),(D2), (D3),(D4) and (D5) and δ is as defined in Appendix A.0.5. Then*

$$\sqrt{n} \left(\hat{\theta} - \theta_0\right) \Rightarrow N\left(0, G_\theta^{-1} \Omega G_\theta^{-1\prime}\right)$$

where

$$\begin{aligned} \Omega &= Var[q(w, \gamma_0) + \delta(w)] \\ G_\theta &= E[\nabla_\theta q(w, \theta_0, \gamma_0)] \end{aligned} \tag{31}$$

The first two assumptions ensure that the bias is order h^m and in the case where $r > 2$ will require the use of higher order kernels. The third assumption ensures $n^{1/4}$ convergence for $\|\hat{\gamma} - \gamma_0\|$. Having obtained the linearization, I calculate the influence function representation for the linear functional $D(w, \gamma)$, combine the

results obtained to deduce a influence function representation for $\sqrt{n}\left(\hat{\theta}-\theta_0\right)$ and then apply a standard Central Limit Theorem to obtain the limiting distribution (see Appendix A.0.5 for the proof). The standard errors can be obtained analytically, by first constructing an estimate of δ, and then forming the sample versions of (31).

5 Distributional Assumption 2

The previous section achieved identification by assuming that the probability of misclassification did not depend upon the other explanatory variables (z) in the model. However, there is evidence, as reported in Section 2, that measurement errors are in fact correlated with individual characteristics, some of which may well be part of the z vector. As noted earlier, once we allow for such dependencies, the model is no longer identified without further assumptions. In this section we allow the misclassification probabilities to depend upon other explanatory variables but assume that we have available another surrogate for x^*. This additional information is enough to achieve point identification and we estimate the model using a (parametric and semi-parametric) likelihood framework.

Let $\{x_1, x_2\}$ denote the two replicated measurements (or more generally surrogates) for x^*. As before, we require that, conditional on the truth and the other explanatory variables, the surrogates provide no further information about the response variable. Specifically, $(A1')$ is modified to

$$P\left(y=1|x^*,x_1,x_2,z\right)=F\left(\beta_0+\beta_1 x^*+\beta_2 z\right) \qquad (A1'')$$

We also limit the probability of misclassification, now conditional on the other regressors, by requiring that for each replicate

$$P(x_j=1|x^*=0,z)+P\left(x_j=0|x^*=1,z\right)<1 \quad \text{a.e.}P_z \quad j=1,2 \qquad (A2')$$

Finally, I require that the two surrogates are independent conditional on the truth (x^*) and the other explanatory

variables. Formally, it is sufficient to assume

$$P(x_1|x_2, x^*, z) = P(x_1|x^*, z) \tag{A5}$$

I do not assume that the surrogates are identically distributed conditional on $\{x^*, z\}$. While (A5) does restrict the nature of the dependency between the two surrogates, it is sufficiently weak so as to allow them to be unconditionally dependent and hence allows for correlation across surrogates. Under $(A1'')$,$(A2')$,(A5) the conditional likelihood is given by

$$\begin{aligned}
&P\left(y, x_1, x_2|z\right) = \\
&F\left(\beta_0 + \beta_2 z\right)^y \left(1 - F\left(\beta_0 + \beta_2 z\right)\right)^{1-y} P\left(x_1|x^* = 0, z\right) P\left(x_2|x^* = 0, z\right) P\left(x^* = 0|z\right) \\
&+F\left(\beta_0 + \beta_1 + \beta_2 z\right)^y \left(1 - F\left(\beta_0 + \beta_1 + \beta_2 z\right)\right)^{1-y} P\left(x_1|x^* = 1, z\right) P\left(x_2|x^* = 1, z\right) P\left(x^* = 1|z\right)
\end{aligned}$$

I next introduce some notation. Let η be a five dimensional vector of functions mapping from the support of z to the open unit interval and define the parameter as $\alpha \equiv (b, \eta)$ and as before $w = (y, x_1, x_2, z)$. The log likelihood for the model as a function of α is given by

$$\begin{aligned}
l\left(w, \alpha\right) = \log\{&\left(1 - F\left(b_0 + b_2 z\right)\right)^{1-y} \left(F\left(b_0 + b_2 z\right)\right)^y \\
&\left(\eta_1(z)^{x_1} \left(1 - \eta_1(z)\right)^{1-x_1}\right) \left(\eta_3(z)^{x_2} \left(1 - \eta_3(z)\right)^{1-x_2}\right) \left(1 - \eta_5(z)\right)\} \\
&+ \{\left(1 - F\left(b_0 + b_1 + b_2 z\right)\right)^{1-y} \left(F\left(b_0 + b_1 + b_2 z\right)\right)^y \\
&\left(\eta_2(z)^{1-x} \left(1 - \eta_2(z)\right)^{x_1}\right) \left(\eta_4(z)^{1-x_2} \left(1 - \eta_4(z)\right)^{x_2}\right) \eta_s(z)\}
\end{aligned} \tag{32}$$

The η are the nuisance parameters and are unknown functions of z. The parameters of interest b belong to B a compact subset of $\mathbb{R}^3$ and $\eta \in \Lambda$ where Λ is the (potentially infinite dimensional) nuisance parameter space (to be defined below). The true parameter vector is given by $(\beta, \eta_0) = (\beta, \alpha_0\left(z\right), \alpha_1\left(z\right), \gamma_0\left(z\right), \gamma_1\left(z\right), \mu\left(z\right))$ where

$\alpha_0(z) = P(x_1 = 1|x^* = 0, z)$, $\alpha_1 = P(x_1 = 0|x^* = 1, z)$, $\gamma_0(z)$, $\gamma_1(z)$ are defined similarly for the second surrogate and $\mu(z) = P(x^* = 1|z)$.

5.1 Identification

The approach followed is similar to Section 4.1. The usual identification condition for MLE requires that for any $\alpha \neq \alpha_0$

$$P\{(y, x_1, x_2, z) : P(y = 1, x_1, x_2|z, \alpha_0) \neq P(y = 1, x_1, x_2|z, \alpha)\} > 0 \tag{33}$$

Intuitively, since each element of η represents a probability, we expect that changing any element will lead to (33) holding, particularly if all outcomes occur with positive probability. We, however, need to rule out possible recombinations of (b, η) that yield the same observed probabilities $P(y, x_1, x_2|z)$. I state two conditions under which the model is identified which, as before, depend upon sufficient variation in the distribution of x^*conditional on the other explanatory variables. Another, perhaps more intuitive way to understand the result is to count equations and unknowns and observe that, for instance when z has k points of support, we have $7k$ equations in $3 + 5k$ unknowns.

I state two sets of conditions under which the model is identified. The first, more general, one requires that for any $\alpha \neq \bar{\alpha}$ $\exists\ B \subseteq S_z$, with $P(B) > 0$ such that $\forall z \in B$

$$F_0(b)(1 - \eta_5(z)) + F_1(b)\eta_5(z) \neq F_0\left(\bar{b}\right)(1 - \bar{\eta}_5(z)) + F_1\left(\bar{b}\right)\bar{\eta}_5(z) \tag{34}$$

Lemma 8 *Consider the model given by (A1″),(A2′) and (A5). Suppose in addition that (i)* $\beta_1 \neq 0$[11]*,(ii)* $E[zz'] > 0$ *and (34) holds. Then, the model is identified.*

I can also state a set of conditions related to the instrumental variable conditions discussed in Section 4.1. Assume that C1-C3 from that section hold for each measurement. Then the crucial condition is that for any $\alpha, \bar{\alpha}$

[11] *As noted before, we could use the results from Lemma 2 to learn whether* $\beta_1 = 0$.

there exist three points v_1, v_2, v_3 such that

$$\frac{\eta_5(z, v_1) - \eta_5(z, v_2)}{\bar{\eta}_5(z, v_1) - \bar{\eta}_5(z, v_2)} \neq \frac{\eta_5(z, v_1) - \mu(z, v_3)}{\bar{\eta}_5(z, v_1) - \bar{\eta}_5(z, v_3)} \tag{35}$$

Here, as previously, we are placing a restriction on the relationship between x^* and v that guarantees identification. This condition will hold for instance if v has unbounded support and μ are the logit or Probit functions.

Lemma 9 *Consider the model given by (A1″),(A2′) and (A5). Suppose $\beta_1 \neq 0$ and C1-C3 and (35) hold. Then, the model is identified.*

I define the estimator as the solution to

$$\min_{(b,\eta)\in B\times\Lambda} \frac{1}{n}\sum_{i=1}^{n} \log P(y_i, x_{1i}, x_{2i}|z_i, \alpha)$$

Consider first the case where z is scalar and has a finite support. In this case the parameter vector α is Euclidean ($\in \mathbb{R}^{3+5k}$ when z has k points of support) and the model can be estimated using standard maximum likelihood techniques and the large sample properties of the estimator can be derived from the asymptotic theory for fully parametric maximum likelihood estimation. I do not cover this case here, particularly since the next section generalizes this model considerably and I study the asymptotic properties of that model in some detail.

Once z is continuous, the η vector can no longer be fully characterized by a finite set of parameters without placing further restrictions on the nature of the misclassification so we can no longer follow the procedure outlined immediately above. One alternative is to consider various parametric assumptions on the form of misclassification and carry out fully parametric likelihood estimation. An alternative technique, explored here, that does not place such parametric restrictions is sieve estimation. This approach allows us to estimate β without specifying the effect of z on the misclassification probabilities or on x^*, and thus is in keeping with the spirit of the semiparametric assumptions of Section 4.

The basic idea behind sieve likelihood estimation can be summarized quite briefly. Consider a likelihood that depends upon a parameter belonging to an infinite dimensional parameter space (in our case these are the misclassification probabilities). Conventional maximum likelihood estimation in this model is often infeasible, inconsistent, or has extremely slow rates of convergence[12] because of the large parameter space over which maximization needs to be carried out. Sieve estimation carries out the maximization over (an increasing sequence of) much smaller spaces that approximate the large space as the sample size becomes large. Thus, the infinite dimensional problem is reduced to a finite dimensional one and under a suitable set of conditions, the estimates obtained are consistent at the parametric rate $\sqrt{n}$, asymptotically normal and achieve the semiparametric efficiency bound.

Sieve estimation is not, however, the only possible estimation scheme in a likelihood model with infinite dimensional parameters. I have also experimented with profile likelihood maximization following the method of Severini and Wong [48]. Their proposed estimation technique involves first concentrating out the unknown functions by maximizing the likelihood locally around each data point z_i and then maximizing the concentrated likelihood to obtain the estimate of β. The proposed method is intuitively appealing but computationally intensive in this instance. In order to carry out the routine, we need to carry out n maximizations for each evaluation of the likelihood function at a candidate point b. For this reason, I did not pursue this approach at any length and do not discuss its asymptotic properties.[13]

Sieve estimation[14] for likelihood functions (and extremum estimators in general) was explored in a series of papers beginning with Geman and Hwang [24], Gallant and Nychka [23], Birge and Massart [7], and Shen and Wong [50]. Shen [49] provides theory for the asymptotic normality and efficiency for smooth scalar functions of the sieve estimates ("plug in" estimators), while White and Wooldridge [54] and Chen and Shen [19] develop theory for weakly dependent data.

[12] Grenander [26] provides examples where the MLE is inconsistent or nonexistent. See Severini and Wong [55] for conditions under which MLE in this non-sieve context is consistent and efficient.

[13] See also Vaart [41] for a development of the large sample theory for this case. Neither paper, however, addresses sieve profile likelihood estimation.

[14] The term was introduced by Grenander [26].

I first define the nuisance parameter space Λ. It is given by H^5 so that each of the five nuisance parameters belongs to the set H which is given by

$$H = \left\{ f \in C^s(S_z) : 0 < f(z) < 1 \sup_{x,y \in S_Z, x \neq y} \frac{f^{(s)}(x)) - f^{(s)}(y))}{|x-y|^\gamma} \leq c_f < \infty \right\} \tag{36}$$

I assume for simplicity that z is a scalar and has as its support a closed interval S_z in $\mathbb{R}$. The scalar assumption is costless but the support simplification is needed to apply some of the approximation results in the literature. The relevant measure of the size of a function space here depends on its degree of smoothness. I shall consider functions that are s times continuously differentiable and whose s^{th} derivative is Lipschitz of order γ and $s + \gamma > .5$. Since I am not particularly concerned with nonsmoothness in the misclassification rates I allow for the functions to be twice continuously differentiable and Lipschitz of order 1 so that $s + \gamma = 3$. This is a sufficiently smooth class and is well approximated by a variety of sieves.

Finally, define a sieve $\Lambda_n = \Pi_{k=1}^5 H_{nk}$, a sequence of approximating spaces, that approximates Λ as the sample size gets large for some choice of H_{nk}. I have experimented with splines, trigonometric ('Fourier') series and neural network sieves (for their definition see Appendix A.0.3). Given our choice of sieve, I define the estimator[15]

$$\hat{\alpha} \equiv \left(\hat{b}, \hat{\eta}\right) = \arg \max_{B \times \Lambda_n} \frac{1}{n} \sum_{i=1}^{n} l(w_i, b, \eta) \tag{37}$$

5.2 Consistency

In this section I derive a rate of convergence result for $\hat{\alpha}$. The result will depend on the size of the sieve and how well it approximates the parameter space. I assume that the sieve approximates the parameter space in the sense that for every $\alpha \in B \times \Lambda$ there exists a $\Pi_n \alpha \in B \times \Lambda_n$ such that $\|\alpha - \Pi_n \alpha\| \to 0$ as $n \to \infty$ (for an appropriately chosen metric[16]). This approximation error $\|\alpha - \Pi_n \alpha\|$ is completely deterministic and has been calculated in the

[15] In a work in progress we also explore estimation using the log likelihood conditional on $\{z, x_2\}$ and using an approach similar to Section 4 to eliminate $P(x^*|z, x_2)$ from the likelihood.

[16] We use the Fisher metric, which is defined in Appendix A.0.7.

literature for various choice of sieves, norms and parameter spaces. As is intuitive, the greater the the number of terms in (88) the smaller the approximation error.

In order to measure the size of the sieve I use the L_2 metric entropy with bracketing. Let $F = \{q(\cdot, \theta) : \theta \in \Theta\}$ be a set of real valued functions defined over the support of a random variable z and having finite second moments. Define the norm of this space to be the $L_2(P)$ norm, i.e. $\|q(z,\theta_1) - q(z,\theta_2)\|_2^2 = E[f(z,\theta_1) - f(z,\theta_2)]^2$ for some probability measure P. Given any two functions l and u, the bracket $[l,u]$ is the set of all functions h such that for all z $l(z) \leq h(z,\theta) \leq u(z)$. An ε bracket is a bracket $[l,u]$ such that $\|u-l\|_2 < \varepsilon$. The bracketing number $N_{[]}(\varepsilon, F)$ is the minimum number of ε brackets needed to cover F by which we mean that for any $h \in F$ there exists a bracket $\{l_j, u_j\}$ such that $l_j(z) \leq h(z,\theta) \leq u_j(z)$ almost everywhere. The logarithm of the bracketing number is known as the L_2 metric entropy with bracketing.[17] As is intuitive, the L_2 metric entropy of the sieve increases with the number of terms r_n. Consequently, approximation errors can be made small only at the cost of increasing the entropy of the sieve class. This is analogous to the bias-variance trade-off familiar from the literature on nonparametric estimation.

I follow Chen and Shen [19] with the simplification that the data are independent and identically distributed. A linear approximation to the likelihood is used to study the asymptotic properties of the estimator. Let $dl_{\alpha_0}(w, \bar{\alpha})$ denote the pathwise derivative of l at the point α_0 in the direction $\bar{\alpha}$. More generally, for each w, $dl_{\alpha_0}(w,\cdot)$ is a linear mapping from $B \times \Lambda$ to the reals and when evaluated at a point $\bar{\alpha} = \left(\bar{b}, \bar{\eta}\right)$ is given

$$dl_{\alpha_0}(w,\bar{\alpha}) = \left.\frac{\partial l(w,b)}{\partial b}\right|'_{b=\beta_0}\left(\bar{b}\right) + \sum_{j=1}^{5} A_j(w,\beta_0,\alpha_0)\left(\bar{\eta}_j\right) \tag{38}$$

The specific form of the terms A_j is detailed in the appendix (they are the partial derivatives of the log likelihood treating η as the variable of differentiation). With some abuse of notation I shall use $dl_{\alpha_0}(w)$ to refer to the vector $\left(\left.\frac{\partial l(w,b)}{\partial b}\right|'_{b=\beta_0}, A(w,\alpha_0)'\right)' \equiv (dl_\beta(w)', dl_\eta(w)')'$ and refer to $l(w,\alpha) - l(w,\alpha_0)$ as the criterion difference

[17] See Vaart [52] , Pollard [45], or Vaart and Wellner [53] for a comprehensive treatment.

which is approximated in the sequel $dl_{\alpha_0}(w, \alpha - \alpha_0)$. I assume the following conditions which are sufficient for consistency of the parameter estimates $\left(\hat{b}, \hat{\eta}\right)$.

F1 Let $F_n = \{l(w,\alpha) - l(w,\alpha_0) : \|\alpha - \alpha_0\| < \delta, \alpha \in B \times \Lambda_n\}$ There exists a constant C_2 and a sequence $\delta_n \in (0,1)$ decreasing to 0 such that

$$\delta_n = \sup\left\{\delta > 0 : \int_{\delta}^{\delta^2} \sqrt{\log N_{[]}(\varepsilon, F_n)}d\varepsilon \leq C_2 n^{1/2}\right\}$$

F2 For all small $\varepsilon > 0$ there exists a C_1 such that

$$\sup_{\{\alpha \in B \times \Lambda_n : \|\alpha - \alpha_0\| < \delta\}} Var\left(l(w,\alpha) - l(w,\alpha_0)\right) \leq C_1 \varepsilon^2$$

F3 The matrix $E\left[dl_{\alpha_0}(w)\, dl_{\alpha_0}(w)'\right]$ is positive definite and $\lambda_{\max} < c_1$ and $\lambda_{\min} > c_2$ almost everywhere.

F4 $\mathrm{E}\left[\sup_{\{\alpha \in \Lambda_n : \|\alpha - \alpha_0\| < \delta\}} \|dl_{\alpha_0}(w)\|\right]^2 < \infty$

We can now state the basic result.

Theorem 10 *Consider the Model given by (A1″),(A2′), and (A5) and assume (33), F1-F4 hold. Then,*

$$\|\hat{\alpha} - \alpha_0\| = O_p\left(\max\{\|\alpha_0 - \Pi_n \alpha_0\|, \delta_n\}\right)$$

The result follows from a suitable adaptation for i.i.d. data of Theorem 1 of Chen and Shen [19] and the verification of the conditions for the proof are relegated to Appendix A.0.7. Condition $F1$ controls the size of the criterion difference over the the sieve and for the sieves we consider in this paper δ_n is of the order $r_n^{1/2} n^{-1/2}$. $F2$ places restrictions on how fast the variance of the log difference declines while $F3$ and $F4$ ensure that the log likelihood satisfies a continuity condition.

5.3 Asymptotic Distribution for $\hat{b}$

The asymptotic distribution of the finite dimensional part of the sieve estimator $\hat{\alpha}$ is obtained by studying a linear approximation to the log likelihood. This approximation is shown to satisfy an essential equicontinuity property. I

then characterize the asymptotic distribution of $\sqrt{n}\left(\hat{b}-\beta_0\right)$ using the Cramer-Wold device (i.e. by characterizing the limit distribution of $\sqrt{n}\lambda'\left(\hat{b}-\beta_0\right)$ for an arbitrary nonzero vector λ) and the Riesz Representation Theorem. Let v^* denotes the Riesz representer[18] for the linear functional $f(b,\eta)=\lambda' b$, then I verify conditions such that $\sqrt{n}\lambda'\left(\hat{b}-\beta_0\right)$can be expressed as a normalized average of i.i.d random variables and an asymptotically negligible term. In particular, I show that

$$\sqrt{n}\lambda'\left(\hat{b}-\beta_0\right)=n^{-1/2}\sum_{i=1}^{n}dl_{\alpha_0}\left(w_i,v^*\right)+o_p(1) \tag{39}$$

Asymptotic normality will then follow under standard conditions from the application of a central limit theorem for i.i.d variables. In order to state the necessary conditions for the result I introduce some notation. Define the remainder term

$$r\left(w,\alpha-\alpha_0\right)\equiv l\left(w,\alpha\right)-l\left(w,\alpha_0\right)-dl_{\alpha_0}\left(w,\alpha-\alpha_0\right) \tag{40}$$

Consider a perturbation around a point $\alpha\in B\times\Lambda_n$ as $\alpha^*(\alpha)=\alpha+\varepsilon_n u^*$ where $\varepsilon_n=o\left(n^{-1/2}\right)$ and $u^*=\pm v^*$. The Riesz representer v^* satisfies[19]

$$\left\|v^*\right\|^2=\sup_{\{\alpha\in B\times\Lambda:\|\alpha-\alpha_0\|>0\}}\frac{\left(\lambda'\left(b-\beta_0\right)\right)^2}{\left\|\alpha-\alpha_0\right\|^2} \tag{41}$$

and an explicit formula for v^* results from an appropriate characterization of this maximization problem (see below). In what follows denote $\mathbb{P}_n(f)=n^{-1}\left(\Sigma_i f(x_i)-E\left[f(x)\right]\right)$ and $K\left(\alpha_0,\alpha\right)=E\left[l\left(w,\alpha_0\right)-l\left(w,\alpha\right)\right]$

I now state conditions sufficient for $\left(\hat{b}-\beta_0\right)$ to converge at the $\sqrt{n}$ rate and have a limiting normal distribution. Let the convergence rate of $\|\hat{\alpha}-\alpha_0\|$ be $o_p\left(\delta_n\right)$.

[18] If $\{\Lambda, <,>\}$ is a Hilbert space and f is a bounded linear function from Λ to $\mathbb{R}$, then by the Riesz Representation Theorem, there is a unique member v^* of Λ such that $f(\alpha)=<\alpha,v^*>$ for each α in Λ. v^* is referred to here as the Riesz representer (see for instance Debnath and Mikusinski [20]).

[19] See Ai and Chen [1].

G1

$$\sup_{\{\alpha\in B\times\Lambda_n:\|\alpha-\alpha_0\|<\delta_n\}} \mathbb{P}_n\left(r\left(w,\alpha-\alpha_0\right)-r\left(\Pi_n\alpha^*\left(\alpha\right)-\alpha_0\right)\right)=o_p\left(n^{-1}\right)$$

G2

$$\sup_{\{\alpha\in B\times\Lambda_n:\|\alpha-\alpha_0\|<\delta_n\}} \left[K\left(\alpha_0,\Pi_n\alpha^*\left(\alpha\right)\right)-K\left(\alpha_0,\alpha\right)\right]-$$
$$(1/2)\left[\left\|\Pi_n\alpha^*\left(\alpha\right)-\alpha_0\right\|^2-\left\|\alpha-\alpha_0\right\|^2\right]=o\left(n^{-1}\right)$$

G3

$$\sup_{\{\alpha\in B\times\Lambda_n:\|\alpha-\alpha_0\|<\delta_n\}} \left\|\alpha^*\left(\alpha\right)-\Pi_n\alpha^*\left(\alpha\right)\right\|=O\left(\delta_n^{-1}\varepsilon_n^2\right)$$

G4

$$\sup_{\{\alpha\in B\times\Lambda_n:\|\alpha-\alpha_0\|<\delta_n\}} \mathbb{P}_n dl_{\alpha_0}\left(w,\alpha^*\left(\alpha\right)-\Pi_n\alpha^*\left(\alpha\right)\right)=o_p\left(n^{-1}\right)$$

G5

$$\sup_{\{\alpha\in B\times\Lambda_n:\|\alpha-\alpha_0\|<\delta_n\}} \mathbb{P}_n dl_{\alpha_0}\left(w,\alpha-\alpha_0\right)=o_p\left(n^{-1/2}\right)$$

Theorem 11 *Consider the Model given by (A1″),(A2′), and (A5) and suppose the sieve estimate has a rate of convergence* $\|\hat{\alpha}-\alpha_0\|=o_p\left(\delta_n\right)$. *Then,*

$$\sqrt{n}\left(\hat{b}-\beta_0\right)\Rightarrow N\left(0,Var\left(l_{\alpha_0}\left(w,v^*\right)\right)\right) \tag{42}$$

The result is an application of Theorem 1 of Shen [49]. $G1$ is very much like an equicontinuity condition that ensures that the linear approximation is small enough to ensure $\sqrt{n}$ convergence while the next condition stipulates that the limiting objective function is roughly quadratic (in the Fisher metric) around the truth. The other conditions are suitable generalizations of the requirement in the finite dimensional case that $\alpha_0$be an interior point of the parameter space. To see this, note that if α_0 is an interior point of the Euclidean space Λ then for ε_n small enough α^* will also be in the sieve Λ_n so that the difference in (G3) and (G4) will be identically zero. The details of the proof are in Appendix A.0.8 but I discuss below the calculation of the efficient score function and the Riesz representer.

5.3.1 Calculation of the Riesz representer

I calculate the Riesz representer v^* in two steps. I first compute the least favorable directions for the model using the results in Begun, Hall, Huang and Wellner [4] as stated in Severini and Wong [48]. I then characterize the maximization problem in (41) using the least favorable directions to obtain v^* following the method outlined in Ai and Chen [1].

The calculation of an explicit formula for v^* in general can be quite difficult. However, the likelihood in (32) falls within the category of what are called conditionally parametric models for which an explicit formula is available. Denote by $\bar{B} \times \bar{\Lambda}$ the linear completion of $B \times \Lambda$ under the Fisher norm (we need the linear completion in order to be able to apply the projection theorem for Hilbert spaces). Following Begun et al. [4][20] I can compute the least favorable direction for β (component by component $j = 1, 2, 3$) as

$$\delta^{j*}(z) = E\left(\left[\frac{\partial^2 l\left(w, \alpha_0\right)}{\partial \eta d\eta'}\middle| z\right]\right)^{-1} E\left[\frac{\partial^2 l\left(w, \alpha_0\right)}{\partial \eta \partial b_j}\middle| z\right] \tag{43}$$

The intuition for this result comes from examining the nature of the problem of projecting the scores for b_j onto the space spanned by the components of the scores for η.

Using $\delta^*(z) = \left(\delta^{1*}, .., \delta^{3*}\right)$ to represent the 5×3 matrix of the least favorite directions I next characterize (41). After a few calculations (see Appendix) we can show

$$\begin{aligned} \|v^*\|^2 &= \lambda' E\left[\tilde{l}_{\beta_0}\left(w\right)\tilde{l}_{\beta_0}\left(w\right)'\right]^{-1}\lambda \\ &\equiv \lambda' \tilde{I}^{-1}\lambda \end{aligned} \tag{44}$$

where

$$\tilde{l}_{\beta_0}\left(w\right) = \left(dl_\beta(w) - \delta^*(z)' dl_\eta(w)\right) \tag{45}$$

[20] A detailed exposition of these calculations (and much more) is also available in Bickel et al. [6].

is the efficient score function for β. This implies that

$$v^* = \left(\tilde{I}^{-1}\lambda, -\delta^*(z)\tilde{I}^{-1}\lambda\right) \tag{46}$$

so that under the conditions above we can express

$$\begin{aligned}
\sqrt{n}\lambda'\left(\hat{b} - \beta_0\right) &= n^{-1/2}\sum_{i=1}^{n}\left[dl_\beta(w_i)'\tilde{I}^{-1}\lambda - dl_\eta(w)'\delta^*(z)\tilde{I}^{-1}\lambda\right] + o_p(1) \\
&= \left(\lambda'\tilde{I}^{-1}\right)n^{-1/2}\sum_{i=1}^{n}\left[dl_\beta(w_i) - \delta^*(z)'dl_\eta(w)\right] + o_p(1) \\
&= \left(\lambda'\tilde{I}^{-1}\right)n^{-1/2}\sum_{i=1}^{n}\tilde{l}_{\beta_0}(w_i) + o_p(1)
\end{aligned} \tag{47}$$

so that

$$\sqrt{n}\lambda'\left(\hat{b} - \beta_0\right) \Rightarrow N\left(0, \lambda'\tilde{I}^{-1}\lambda\right) \tag{48}$$

since this holds for any non-zero vector λ,we have that

$$\sqrt{n}\left(\hat{b} - \beta_0\right) \Rightarrow N\left(0, \tilde{I}^{-1}\right) \tag{49}$$

A consistent estimate of the efficient Fisher information can be calculated by computing the sample covariance of the estimated efficient score. The efficient score in turn can be obtained as the residual of the projection of the score for the β parameters onto the scores for the sieve coefficients.

6 Monte Carlo Results

6.1 Distributional Assumption I

In this sub-section I use a set of Monte Carlo simulations to illustrate the estimator and study its small sample properties. The model consists of

$$P\left(y=1|x^{*},z\right)=\Phi\left(\beta_{0}+\beta_{1}x^{*}+\beta_{2}z\right) \tag{50}$$

I specify $P(x^{*}=1|z)$ as a logit (I also experimented with a Probit and a simple distribution function for discrete z). I carry out numerical optimization of (21) in MATLAB using a sequential quadratic programming method. Throughout, $(\beta_{0},\beta_{1},\beta_{2})$ is set to $(0,-1,1)$. Table (3) displays the results for different sample sizes as the misclassification problem increases. Table (4) displays the same statistics for the case when there exists a variable v that satisfies assumptions C1-C3. In both tables, for the misclassification rates considered and for moderate sample sizes, the estimator seems well behaved. The standard errors roughly halve when we quadruple the sample size reflecting the $\sqrt{n}$ convergence rate of the estimator. We also obtain similar results with different choices for the support of z. For the first two tables z is discrete (taking on three values) while for the next two it is continuous (uniform on $[-1,1]$ and standard normal respectively). In the case where z is continuous we need to use a first step kernel estimator to estimeate $P\left(x=1|z\right)$ and we use a locally linear regression for this purpose. The results are somewhat more favorable to the estimator when z has a rich support which jibes with the intuition that a bigger support of z in essence adds more information (or moment conditions in the naive counting equations and unknowns approach).

6.2 Distributional Assumption 2

In this sub-section I illustrate the performance of the estimator under the second set of distributional assumptions and study its small sample properties. As before, the binary response model is given by (50) and I specify z to

be uniform over $[-2,2]$. Each of the functions $(\eta_1(z),...,\eta_5(z))$ is a logit (we also experimented with a Probit specification and obtained similar results) and use as our sieve a spline which is locally cubic over $[-2,2]$. The theory does not provide an exact number of terms for the approximation and for the samples considered here I use 1 to 3 terms. As before, $(\beta_0,\beta_1,\beta_2)=(0,-1,1)$. The results are in Table (7). As might be expected, they are less precise than in the previous set of simulations although for moderately large sample sizes and moderate misclassification rates the estimator performs reasonably well. [21]

7 Empirical Illustration

As an empirical illustration I examine the effect of union status on the probability of receiving health insurance[22] using data from the Current Population Survey (CPS). There is a substantial literature on the prevalence of measurement error in union status (see for instance Card [14] and the review in Bound, Brown and Mathiowetz [12]) although somewhat less is known about the effect of this misclassification on estimated parameters.[23]

The February 1999 CPS Basic Monthly Questionnaire (BMS) asked information on union status for all respondents in their fourth and eighth months (the outgoing rotation groups) while the Contingent Worker Supplement to the questionnaire recorded information on the receipt of health insurance and whether or not it was employer provided. I restrict attention to employed individuals between the ages of 18 and 60 who are not self-employed or engaged in agricultural work and the summary statistics for the data set are given in Table (9). Overall about 18% of the sample belonged to a union or employee association[24] and 70% had health insurance at the time of the survey. The results from a direct Probit estimation are given in Table (10) and those for the estimation method proposed above are given in Table (11) with the corresponding results for marginal effects for both methods given

[21] All standard errors for the simulations are bootstrapped with the number of bootstrap replications set to 1000.

[22] For work on the effect of unionization on the provision of health insurance and other fringe benefits see for instance Freeman [22] and Belman and Heywood [5].

[23] Card [14] studies the effect of union status on wages that explicitly accounts for the presence of measurement error in union status in the context of a linear model. The study assumes that misclassification rates are symmetric and independent of other regressors in the model.

[24] The sample also includes some workers (less than 1.5% of the sample) who do not belong to a union but whose jobs are covered by a union or employee association contract. The results are robust to their inclusion in the union category.

in Table (14). The point estimate for union status is about 26% higher under the estimation method discussed above, although the confidence interval is considerably wider than that for the usual Probit. The point estimates and standard errors for the remaining coefficients are roughly comparable across the two methods. The point estimate of the marginal effect of union status is about a third higher under the method that takes the misclassification rates into account, although again here the corresponding standard errors are approximately twice as large, thereby reducing the t-statistic (for testing the null of no effect) by about a fifth.

I also implement the estimator obtained under the "instrumental variable" conditions by constructing another surrogate for union status by matching two consecutive February CPS data sets. The February 2000 CPS BMS again asked the union status question to households in the outgoing rotation groups and the Job Tenure Supplement asked respondents how long they had been working at their current job and whether they had changed occupations. I then have plausibly two self-reported measures of union status for those individuals who were in their fourth month of the survey in February 1999 and who did not change jobs during the next twelve months. Individuals were matched using the household identifier, line number, household number, race and sex and we obtained a match rate of about 62% which is line with what we would expect from previous work. We also imposed age consistency criteria without altering the match rate to any significant degree. [25] The results from the estimation are quite similar to the results from the first estimation method and are displayed in Table (12). Table (13) illustrates the implementation of the estimator derived in Section 5 using a neural network sieve. Here, unlike the previous tables, I allow for arbitrary relationships between the probability of misclassification and the other explanatory variables. The point estimate for union status is almost 25-50% higher than obtained from the previous estimation methods and the remaining coefficients also differ from previous estimates by an average of 30%. The standard errors for union status differ from the Probit standard errors by 18-60% under the various estimation methods considered. Finally, the marginal effect of union status on health benefits for the last method is reported in Table (14) and is roughly 60% higher than that from the Probit with approximately the

[25] For more information on matching individuals across CPS surveys see Madrian and Lefgren [37].

same standard error.

8 Conclusion

The evidence on measurement error in typical data sets suggests that it does not satisfy the independence assumptions usually required by error correction techniques in non-linear models. This paper looks at the effect of relaxing these assumptions in the context of a simple non-linear model with a simple type of measurement error. Section 3 shows that the model is only partially identified under a minimal set of assumptions and derives sharp bounds for the parameters of interest. I then show that the model is identified without further information the misclassification is independent of the other regressors in the model and develop $\sqrt{n}$ consistent and asymptotically normal semiparametric estimators in this situation. However, when the misclassification probabilities to depend on the other regressors, as is suggested by the empirical evidence on the issue, Section 5 shows that the model can still be estimated as long as we have another surrogate for x^*. I develop a semiparamatric estimator using the method of sieves that allows the misclassification probabilities to depend arbritrarily (albeit smoothly) on the other regressors in the model but still attains the parametric rate of convergence.

In future work, it would be interesting to extend the framework to allow for unknown functions F and to models other than the binary choice model such as quantile regression. It is also of independent interest to formulate tests to detect the presence of misclassification (i.e. testing when the parameters may lie on the boundary of the parameter space) and more importantly, to apply this framework to answer empirical questions of practical import.

A Appendix

A.0.1 Proofs for Non-Parametric Bounds

In the subsequent arguments all inequalities hold conditional on z. We first collect some results that will prove useful in the sequel.

Lemma 12 *Define*

$$P(x=1|x^*=0) \equiv \alpha_0(z) \tag{51}$$
$$P(x=0|x^*=0) \equiv \alpha_1(z) \tag{52}$$
$$P(x^*=1|z) = p(z) \tag{53}$$

$$\begin{aligned} \gamma(z) &\equiv P(x^*=1|x=1,z) = \frac{(1-\alpha_1)p}{\alpha_0(1-p)+(1-\alpha_1)p} \\ \delta(z) &\equiv P(x^*=1|x=0,z) = \frac{\alpha_1 p}{(1-\alpha_0)(1-p)+\alpha_1 p} \end{aligned} \tag{54}$$

Then under (A2)

$$\gamma(z) > \delta(z) \quad a.e.\ z \tag{55}$$

Proof. From (A2) and suppressing the dependence on z

$$\begin{aligned} & 1-\alpha_1 > \alpha_0 \\ \Rightarrow\ & (1-p)(1-\alpha_1) > (1-p)\alpha_0 \\ \Rightarrow\ & (1-\alpha_1) > (1-p)\alpha_0 + (1-\alpha_1)p \\ \Rightarrow\ & 1-\alpha_1 > \eta \end{aligned} \tag{56}$$

where

$$\eta \equiv P(x=1) \tag{57}$$

so that

$$\begin{aligned} & 1-\alpha_1-\eta+\alpha_1\eta > \alpha_1\eta \\ \Rightarrow\ & (1-\alpha_1)(1-\eta) > \alpha_1\eta \\ \Rightarrow\ & \gamma = \frac{(1-\alpha_1)p}{\eta} > \frac{\alpha_1 p}{(1-\eta)} = \delta \end{aligned} \tag{58}$$

■

Proof. Proof of (5)

Using Bayes Rule

$$m_1(z) = \gamma(z)\, m_1^*(z) + (1-\gamma(z))\, m_0^*(z) \tag{59}$$
$$m_0(z) = \delta(z)\, m_1^*(z) + (1-\delta(z))\, m_0^*(z) \tag{60}$$

First consider the case where $g_1(z) = 1$. Both (59) and (60) are linear combinations of $\{m_1^*, m_0^*\}$ and from Lemma 12 we know that $\gamma > \delta$. Then we must have that

$$m_1^*(z) > m_0^*(z) \tag{61}$$

This in turn implies that

$$m_1^* \geq \max\{m_1(z), m_0(z)\} = m_1(z) \tag{62}$$
$$m_0^*(z) \leq \min\{m_0(z), m_1(z)\} = m_0(z) \tag{63}$$

A similar argument applies to the case where $g_0(z) = 1$. ■

In order to show the bounds are sharp we need to characterize the solutions $\{m_1^*, m_0^*\}$ to the equations

$$\begin{bmatrix} \gamma & 1-\gamma \\ \delta & 1-\delta \end{bmatrix}^{-1} \begin{bmatrix} m_1 \\ m_0 \end{bmatrix} = \begin{bmatrix} m_1^* \\ m_0^* \end{bmatrix} \tag{64}$$

as a function of $\{\alpha_0, \alpha_1, p, m_1, m_0\}$ (we suppress dependency on z but note that the bounds are conditional on z). The solution to (64) under (A2) and $0 < p < 1$ is given by

$$m_1^*(\alpha_0, \alpha_1, p) = (1-\alpha_0)\left(1 + \frac{\alpha_0}{(1-\alpha_0-\alpha_1)\,p}\right) m_1 + \alpha_0\left(1 - \frac{1-\alpha_0}{(1-\alpha_0-\alpha_1)\,p}\right) m_0 \tag{65}$$
$$m_0^*(\alpha_0, \alpha_1, p) = \frac{(1-\alpha_0-p\,(1-\alpha_0-\alpha_1))\,(1-\alpha_1)\,m_0}{(1-p)\,(1-\alpha_0-\alpha_1)} - \frac{(\alpha_0\,(1-p) + (1-\alpha_1)\,p)\,\alpha_1 m_1}{(1-p)\,(1-\alpha_0-\alpha_1)} \tag{66}$$

We next collect a few useful results.

Lemma 13 *Under (A2) and $0 < p < 1$ (65) and (66) are well defined and continuous in $\{\alpha_0, \alpha_1, p\}$. Assume $1 > m_1 > 0, 1 > m_0 > 0$ Then,*

1. $m_1^*(0, \alpha_1, p, m_1, m_0) = m_1,\ m_0^*(\alpha_0, 0, p, m_1, m_0) = m_0$
2. $sgn(\frac{\partial}{\partial p} m_1^*(\alpha_0, \alpha_1, p, m_1, m_0)) = sgn\left(\frac{\alpha_0(1-\alpha_0)}{-(1-\alpha_0-\alpha_1)p^2}\,(m_1 - m_0)\right) = -sgn\,(m_1 - m_0)$ *for* $\alpha_0 > 0$
3. $sgn(\frac{\partial}{\partial p} m_0^*(\alpha_0, \alpha_1, p, m_1, m_0)) = sgn\left(\frac{\alpha_1(\alpha_1-1)}{(1-\alpha_0-\alpha_1)(1-p)^2}\,(m_1 - m_0)\right) = -sgn\,(m_1 - m_0)$ *for* $\alpha_1 > 0$
4. $sgn\left(\frac{\partial}{\partial \alpha_1} m_1^*\,(\alpha_0, \alpha_1, p, m_1, m_0)\right) = sgn\left((1-\alpha_0)\,\alpha_0 \frac{(m_1-m_0)}{(1-\alpha_0-\alpha_1)^2 p}\right) = sgn\,(m_1 - m_0)$
5. $sgn\left(\frac{\partial}{\partial \alpha_0} m_1^*\,(\alpha_0, \alpha_1, p, m_1, m_0)\right) = sgn\left(\left(1 + \frac{(1-2\alpha_0)(\alpha_1-1)-\alpha_0^2}{(1-\alpha_0-\alpha_1)^2 p}\right)(m_0 - m_1)\right) = sgn\,(m_1 - m_0)$

Proof. (1) We fix z and subsequently suppress it as an argument. Consider the case where $m_1 > m_0 > 0$ and consider first the bound $[m_1, 1)$. Pick any $\bar{m} \in [m_1, 1)$. We will show that there exists a choice of $\{\alpha_0, \alpha_1, p\}$ and a corresponding $m_0^*\,(\alpha_0, \alpha_1, p, m_1, m_0) \in (0, m_0]$ such that $m_1^*\,(\alpha_0, \alpha_1, p, m_1, m_0) = \bar{m}$. We shall suppress the dependence of the functions on $\{m_0, m_1\}$ from now on

First choose $\bar{\varepsilon} \in (0,1)$ and $\bar{p} \in (0,1)$ such that

$$\frac{\bar{p} + \bar{\varepsilon} - 1}{1 - \bar{p}\bar{\varepsilon}} > \frac{m_1 - m_0}{m_0} \tag{67}$$

This ensures that there exists some $x \in (0, 1-\bar{\varepsilon})$ such that

$$m_0^*\,(x, \bar{\varepsilon}, \bar{p}) > 0 \tag{68}$$

Such an x satisfies

$$x\bar{p} + (1-\bar{p})\,(1-\bar{\varepsilon}) < \frac{m_0\,(1-\bar{\varepsilon})}{(1-\bar{\varepsilon})\,m_0 + \varepsilon m_1} \tag{69}$$

Consider the function

$$g(\varepsilon,p)=\frac{1}{p}\left(\frac{m_0(1-\varepsilon)}{(1-\varepsilon)m_0+\varepsilon m_1}-(1-p)(1-\varepsilon)\right) \tag{70}$$

Note that g is decreasing in p for any $\varepsilon\in(0,1)$. For any (ε,p) we then have that

$$m_0^*(x,\varepsilon,p)>0 \Longleftrightarrow x<g(\varepsilon,p) \tag{71}$$

Next, observe that $m_1^*(0,\varepsilon,p)=m_1$ for any (ε,p) and $m_1^*(\cdot,\varepsilon,p)$ is continuous and increasing and is onto $[m_1,1)$ for $0<\alpha_0<1-\varepsilon$ (see previous lemma). Then, there exists an $c(\bar{\varepsilon},\bar{p})$ such that

$$m_1^*(c(\bar{\varepsilon},\bar{p}),\bar{\varepsilon},\bar{p})=\bar{m} \tag{72}$$

If

$$g(\bar{\varepsilon},\bar{p})>c(\bar{\varepsilon},\bar{p}) \tag{73}$$

then by (71)

$$m_0^*\left(c(\bar{\varepsilon},\bar{p}),\bar{\varepsilon},\bar{p}\right)\in(0,m_0) \tag{74}$$

and we are done.

Consider, however the case where $c(\bar{\varepsilon},\bar{p})>g(\bar{\varepsilon},\bar{p})$. Since for any $a_0\in(0,1-\bar{\varepsilon})$ there exists a p such that $m_1^*(\alpha_0,\bar{\varepsilon},p)=\bar{m}$ there must exist a $\tilde{p}$ such that $m_1^*(g(\bar{\varepsilon},\bar{p}),\bar{\varepsilon},\tilde{p})=\bar{m}$ and since m_1^* is decreasing in p (for fixed a_0,α_1) $\tilde{p}<\bar{p}$. This in turn implies that $g(\bar{\varepsilon},\tilde{p})>g(\bar{\varepsilon},\bar{p})=c(\bar{\varepsilon},\tilde{p})$ so that $m_1^*(c(\bar{\varepsilon},\tilde{p}),\bar{\varepsilon},\tilde{p})=\bar{m}$ and $m_0^*(c(\bar{\varepsilon},\tilde{p}),\bar{\varepsilon},\tilde{p})\in(0,m_0)$

The proof for the other interval $[0,m_0]$ follows analogously. If $0<m_1<m_0$ then the bound on m_1^* would be $(0,m_1]$ and for m_0^* would be $[m_0,1)$ by similar arguments as above.

Finally, note that these bounds cannot be improved even if we strengthen (A2) to

$$P(x=1|x^*=0,z)\ +P(x=0|x^*=1,z)<\kappa \text{ a.e. } z \tag{75}$$

where $\kappa\in(0,1]$ ((A2) corresponds to $\kappa=1$) ■

Proof of Lemma (3)

Proof. Pick any $\bar{b}\in\mathbb{R}$. Choose any $\bar{b}_0<\min_{S_z}\left(F^{-1}(m_0(z))-\bar{b}z\right)$ and pick any

$$\bar{b}_1>\max_{S_z}\left\{F^{-1}(m_1(z))-\bar{b}z\right\}-\min_{S_z}\left(F^{-1}(m_0(z))-\bar{b}z\right)$$

This choice of $\bar{b}_1$ satisfies (11). It is easy to see this when the same value of z maximizes and minimizes the two objects on the right hand side. Now suppose that $z_1=\arg\max_{S_z}\left\{F^{-1}(m_1(z))-\bar{b}z\right\}$ and $z_2=\arg\min_{S_z}\left(F^{-1}(m_0(z))-\bar{b}z\right)$. Then, we must have

$$F^{-1}(m_1(z_1))-\bar{b}z_1\geq F^{-1}(m_1(z_2))-\bar{b}z_2$$

so that $\bar{b}_1>F^{-1}(m_1(z_2))-F^{-1}(m_1(z_2))$. Also, since $z_2=\arg\min_{S_z}\left(F^{-1}(m_0(z))-\bar{b}z\right)$

$$F^{-1}(m_0(z_2))-\bar{b}z_2\leq F^{-1}(m_0(z_1))-\bar{b}z_1$$

so that

$$\left(F^{-1}(m_1(z_1))-\bar{b}z_1\right)-F^{-1}(m_0(z_2))-\bar{b}z_2\geq\left(F^{-1}(m_1(z_1))-\bar{b}z_1\right)-\left(F^{-1}(m_0(z_1))-\bar{b}z_1\right)$$

so that $\bar{b}_1>F^{-1}(m_1(z_1))-F^{-1}(m_1(z_1))$. This shows that $\beta_2\in\mathbb{R}$.

Next, pick any $b_1 > \max_{S_z}\left(F^{-1}(m_1(z)) - F^{-1}(m_1(z))\right)$. Then we can always pick (b_0, b_2) such that

$$\min_{S_z} F^{-1}(m_0(z)) \geq b_0 + b_2 z \geq \max_{S_z} F^{-1}(m_1(z)) - b_1$$

is satisfied (the acceptable region is given by the parallelogram formed by the four lines with slopes (z_1, z_2) and intercepts $\left\{\min_{S_z} F^{-1}(m_0(z)), \max_{S_z} F^{-1}(m_1(z)) - b_1\right\}$ Finally, consider $\bar{b}_0 \leq F^{-1}(m_0(0))$ Then, choose a b_2 $\leq F^{-1}(m_0(1)) - \bar{b}_0$ and pick b_1 such that

$$b_1 \geq \max\left\{F^{-1}(m_1(1)) - F^{-1}(m_0(1)), F^{-1}(m_1(0)) - \bar{b}_0\right\}$$

to see that any $\bar{b}_0 \leq F^{-1}(m_0(0))$ is acceptable. Note that only the last part of the proof relied upon 0 being in the support of z. ■

Lemma 14 *Suppose z in (A1′) has unbounded support. Then β_2 is identified*
Suppose that $b \in B^*$ but $b_2 \neq \beta_2$. Then

$$\begin{aligned} P(V(b)) &\geq P[E(y|x,z) < F(b_0 + b_2 z)] \\ &\geq P[F(\beta_0 + \beta_1 + \beta_2 z) < F(b_0 + b_2 z)] \\ &= P[(\beta_2 - b_2)z < (b_0 - \beta_0) - \beta_1] > 0 \end{aligned}$$

because z has unbounded support. But this implies $b \notin B^*$. Therefore $b_2 = \beta_2$ for all $b \in B^*$

A.0.2 Identification Results

Lemma 15 *Suppose that (A1′), (16), and (17) hold, (i) $\beta_1 \neq 0$,(ii) $Var[(P(x^* = 1|z)] > 0$ and (iii) $E[1,z]'[1,z] > 0$. Then, $\{\alpha_0, \alpha_1\}$ is identified if and only if β is identified*

Proof. " $\Rightarrow$ "Suppose $\{\alpha_0, \alpha_1\}$ are identified. Then it follows that $\{\gamma,\ \delta\}$ as defined in (54) are identified. Next, suppose there exist $\bar{b}$ such that

$$P(y = 1|x, z, \alpha_0, \alpha_1, \beta) = P(y = 1|x, z, \alpha_0, \alpha_1, \bar{b}) \tag{76}$$

Then, using the notation introduced in Lemma (12) and using $\bar{F}_s \equiv F\left(\bar{b}_0 + \bar{b}_1 s + \bar{b}_2 z\right)$ as shorthand

$$(F_1 - \bar{F}_1)\gamma(z) + (F_0 - \bar{F}_0)(1 - \gamma(z)) = 0 \tag{77}$$
$$(F_1 - \bar{F}_1)\delta(z) + (F_0 - \bar{F}_0)(1 - \delta(z)) = 0 \tag{78}$$

which in turn imply

$$(F_1 - \bar{F}_1)(\gamma(z) - \delta(z)) - (F_0 - \bar{F}_0)(\gamma(z) - \delta(z)) = 0 \tag{79}$$

so that

$$(\gamma(z) - \delta(z))\left[(F_1 - \bar{F}_1) - (F_0 - \bar{F}_0)\right] = 0 \tag{80}$$

By Lemma (12) we know $\gamma(z) > \delta(z)$ so that we must have

$$(F_1 - \bar{F}_1) = (F_0 - \bar{F}_0) \tag{81}$$

which implies (via 77) that

$$F_0 = \bar{F}_0 \tag{82}$$

so that under (iii) we must have

$$b_0 = \beta_0 \text{ and } b_2 = \beta_2 \tag{83}$$

Finally, we also have from (81)

$$F_1 = \bar{F}_1 \tag{84}$$

so that in conjunction with (83) we must have

$$b_1 = \beta_1 \tag{85}$$

" $\Leftarrow$ "

Suppose β is identified but the model is not completely identified. Then, it must be that there exists an $(a_0, a_1) \neq (\alpha_0, \alpha_1)$ such that

$$P(y = 1|z, \alpha_0, \alpha_1, \beta) = P(y = 1|z, a_0, a_1, \beta)$$

so that

$$F_0(1 - \bar{\gamma}) + F_1\bar{\gamma} = F_0(1 - \gamma) + F_1\gamma$$

almost everywhere. This, given (i) implies

$$\gamma(z) = \bar{\gamma}(z) \tag{86}$$

and given (ii) there exist at least two values of z with $\xi(z_1) \neq \xi(z_2)$ and for each z

$$\frac{\xi(z) - a_0}{1 - a_0 - a_1} = \frac{\xi(z) - \alpha_0}{1 - \alpha_0 - \alpha_1}$$

so that

$$\frac{\xi(z_1) - a_0}{\xi(z_2) - a_0} = \frac{\xi(z_1) - \alpha_0}{\xi(z_2) - \alpha_0}$$

which in turn implies that $\alpha_0 = a_0$ and $a_1 = \alpha_1$ follows. ■

Proof of Lemma 4

Proof. Suppose that the model is not identified, then there exist $\theta, \bar{\theta}$ and $\theta \neq \bar{\theta}$ such that

$$P\left(y, x|z, \bar{\theta}\right) = P\left(y, x|z, \bar{\theta}\right) \quad a.e.$$

which in turn implies that

$$P(y = 1|z, \theta) = P\left(y = 1|z, \bar{\theta}\right) \quad a.e. \tag{87}$$

Recall that

$$P(y = 1|z, \theta) = \left(F_0(b)(1 - \xi(z) - a_1) + F_1(b)(\xi(z) - a_0)\right) / (1 - a_0 - a_1)$$

so that (87) implies that

$$\frac{\left(F_0(b)(1 - \xi(z) - a_1) + F_1(b)(\xi(z) - a_0)\right)}{\left(F_0\left(\bar{b}\right)(1 - \xi(z) - \bar{a}_1) + F_1\left(\bar{b}\right)(\xi(z) - \bar{a}_0)\right)} = \frac{1 - a_0 - a_1}{1 - \bar{a}_0 - \bar{a}_1}$$

so that

$$\begin{aligned}&\xi(z)\left[F_1(b) - F_0(b) - \kappa\left(F_1(b) - F_0(b)\right)\right] = \\ &\kappa\left[(1 - \bar{a}_1)F_0\left(\bar{b}\right) - \bar{a}_0F_1\left(\bar{b}\right)\right] - \left[(1 - a_1)F_0(b) - a_0F_1(b)\right] \quad a.e.\end{aligned}$$

which cannot hold by assumption (25). ■

Proof of Lemma 5

Proof. C3 ensures that B1 holds. ■

A.0.3 Sieve Definitions

The Fourier space of approximating functions is given by

$$H_{nk} = \left\{ f(x) = \sum_{j=1}^{r_n} a_{j,k} \cos(2\pi j x) + b_{j,k} \sin(2\pi j x), \Sigma_j j^{2q} \left(a_{j,k}^2 + b_{j,k}^2\right) \leq c_{n,k}^2 \right\} \tag{88}$$

where q is some number slightly bigger than $s + \gamma$ (defined in the text). The sigmoid neural network sieve with the logit as the sigmoid function is given by

$$H_{nk} = \left\{ f(x) = a_{0k} + \sum_{j=1}^{r_n} b_{jk} \frac{\exp(a_{0jk} + a_{1jk} x)}{1 + \exp(a_{0jk} + a_{1jk} x)}, \Sigma_j |b_{jk}| \leq c_n, \max_{j \in \{1,..,r_n\}} (a_{0jk} + a_{1jk}) \leq \tilde{c}_n \right\}$$

A sieve based on B-splines is given by

$$H_{nk} = \left\{ f(x) = \sum_{j=1}^{r_n + [p] + 1} b_{jk} \phi_{jk}, , \max_{j \in \{1,..,r_n + [p]+1\}} (b_{jk}) \leq l_n \right\}$$

where $\left(\phi_i, ..., \phi_{r_n + [p]+1}\right)$ are B-splines of order [p]+1 on the support of x.

A.0.4 Consistency for Distributional Assumption 1

Proof of Lemma 6. We obtain consistency by checking the conditions for Theorem 2.1 in Newey and McFadden [42].

Proof. The proof follows from checking condition (iv) for Theorem 2.1 in Newey and McFadden [42] since the remaining conditions are satisfied. [26] ■

$$\sup_{\Theta} |Q_n(\theta, \hat{\gamma}) - Q_0(\theta, \gamma_0)| \leq \sup_{\Theta} |Q_n(\theta, \hat{\gamma}) - Q_n(\theta, \gamma_0)| + \sup_{\Theta} |Q_n(\theta, \gamma_0) - Q_0(\theta, \gamma_0)| \tag{89}$$

We can use a point-wise Mean Value Expansion to obtain

$$Q_n(\theta, \hat{\gamma}) - Q_n(\theta, \gamma_0) \leq \|\hat{\gamma} - \gamma_0\|_\infty \left\| \frac{1}{n} \Sigma_i \nabla_\gamma q(w_i, \theta, \gamma^*) \right\| \tag{90}$$

where $\nabla_\gamma m(w_i, \theta, \gamma)$ is the ordinary derivative w.r.t. γ computed in the usual fashion by treating γ as the variable of interest. The first term is $o_p(1)$ under D1-D3. B3 ensures that

$$E\left[\sup_{\Theta} \sup_{\|\gamma - \gamma_0\| \leq \delta_n} \|\nabla_\gamma q(w_i, \theta, \gamma)\| \right] < \infty$$

which ensures that the last term in the above expression is $O_p(1)$. This gives us

$$\sup_{\Theta} |Q_n(\theta, \hat{\gamma}) - Q_n(\theta, \gamma_0)| \to 0$$

[26] We define the norm for a matrix M as $\|M\| = \sqrt{Trace(M'M)}$ and for the function γ as $\|\gamma\| = \sup_{z \in S_z} \|\gamma(z)\|$

The second term in (89) is much easier to deal with since $q(w,\theta,\gamma_0)$ is uniformly bounded in θ by a constant so that by Lemma 2.4 in Newey and McFadden

$$\sup_{\Theta} \left| Q_n\left(\theta,\hat{\gamma}\right) - Q_0\left(\theta,\gamma_0\right)\right| \to 0 \tag{91}$$

A.0.5 Asymptotic Distribution for Distributional Assumption 1

The proof Theorem 7 will follow from checking the conditions for Theorem 8.12 in [42]. We assume $(A1')$, (17), (16), (D1), (D2), (D3), (D4) and (D5). For the sake of exposition consider first a standard Taylor Series Expansion for $\hat{\theta}$ which gives

$$\sqrt{n}\left(\hat{\theta}-\theta_0\right) = -\left[\Sigma \nabla_\theta q\left(w_i,\bar{\theta},\hat{\gamma}\right)\right]^{-1}\left[\frac{1}{\sqrt{n}}\Sigma_i q(w_i,\hat{\gamma})\right] \tag{92}$$

where we adopt the convention $q(w_i,\gamma) = q(w_i,\theta_0,\gamma)$. The idea is to show convergence in distribution for the term $n^{-1/2}\Sigma_i q(w_i,\hat{\gamma})$.

We begin with a study of the properties of the first order conditions which after some manipulation can be expressed as

$$q\left(w,\theta,\gamma\right) = \left(sv_1 + (1-s)\,v_2\right) w(z) \tag{93}$$

where

$$s\left(w,\theta,\gamma\right) = \frac{d_0\left(w,\theta,\gamma\right)}{d_0\left(w,\theta,\gamma\right) + d_1\left(w,\theta,\gamma\right)} \tag{94}$$

In the interest of brevity we suppress the arguments (w,θ,γ) from now on. The remaining terms are

$$d_0 = F_0^y\left(1-F_0\right)^{1-y}\alpha_0^x\left(1-\alpha_0\right)^{1-x}\frac{1-\alpha_1-\gamma}{1-\alpha_0-\alpha_1} \tag{95}$$

$$d_1 = F_1^y\left(1-F_1\right)^{1-y}\alpha_1^{1-x}\left(1-\alpha_1\right)^{x}\frac{\gamma-\alpha_0}{1-\alpha_0-\alpha_1} \tag{96}$$

$$v_1 = \left[\left(\frac{y-F_0}{F_0\left(1-F_0\right)}f_0\right)\left[1\ z\right]', 0, \left(\frac{x-\alpha_0}{\alpha_0\left(1-\alpha_0\right)}+\frac{1}{1-\alpha_1-\alpha_0}\right), \left(\frac{\alpha_0-\gamma}{\left(1-\alpha_0-\alpha_1\right)\left(1-\alpha_1-\gamma\right)}\right)\right]' \tag{97}$$

$$v_2 = \left[\left(\frac{y-F_1}{F_1\left(1-F_1\right)}f_1\right)\left[1\ z\right]', \left(\frac{y-F_1}{F_1\left(1-F_1\right)}f_1\right), \left(\frac{\gamma-1+\alpha_1}{\left(1-\alpha_0-\alpha_1\right)\left(\gamma-\alpha_0\right)}\right), \left(\frac{1}{1-\alpha_1-\alpha_0}+\frac{1-\alpha_1-x}{\alpha_1\left(1-\alpha_1\right)}\right)\right]' \tag{98}$$

We first check the various finiteness conditions required by the conditions of the theorem. Each element of the vector $q(\cdot,\theta_0,\gamma_0)$ is bounded almost surely so that $E\left[\left\|q(w,\theta_0,\gamma_0)\right\|^2\right] < \infty$. The boundedness of each element follows from the binary nature of $\{y,x\}$ and the imposition that the weight function be zero outside of a compact set. We can similarly show by inspection that $q\left(w,\theta,\gamma_0\right)$ is continuously differentiable in $\theta \in int(\Theta)$ and that $\nabla_\theta q\left(\cdot,\theta_0,\gamma_0\right)$ is bounded element by element so that $E\left[\nabla_\theta q\left(w,\theta_0,\gamma_0\right)\right] < \infty$. In addition, $\nabla_{\gamma\gamma} q\left(\cdot,\gamma_0\right)$ is also bounded so that $E\left[\left\|\nabla_{\gamma\gamma} q(w,\gamma_0)\right\|\right] < \infty$.

Next, consider a pointwise Taylor expansion for the ith element of q

$$\begin{aligned} q_i(w,\gamma) &= q_i\left(w,\gamma_0\right) + \nabla_\gamma q_i\left(w,\gamma_0\right)\left(\gamma\left(z\right)-\gamma_0\left(z\right)\right) + \left(\gamma\left(z\right)-\gamma_0\left(z\right)\right)'\nabla_{\gamma\gamma} q_i\left(w,\gamma_0\right)\left(\gamma\left(z\right)-\gamma_0\left(z\right)\right) \\ &\quad + o\left(\left\|\gamma-\gamma_0\right\|^2\right) \end{aligned} \tag{99}$$

where the derivative ∇_γ is taken in the usual fashion by treating γ as the variable of interest and the norm over the γ space is chosen appropriately (say the sup-norm). Note that the problem of finding the linearization is

simplified in this problem since γ affects $q(w,\gamma)$ only through its value at one point z. Next, note that

$$\begin{aligned} |q_i(w,\gamma) - q_i(w,\gamma_0) + \nabla_\gamma q_i(w,\gamma_0)(\gamma(z) - \gamma_0(z))| &\leq \left\|(\gamma(z) - \gamma_0(z))' \nabla_{\gamma\gamma} q_i(w,\gamma_0)(\gamma(z) - \gamma_0(z))\right\| \\ &\quad + o(\|\gamma - \gamma_0\|^2) \\ &\leq \|\gamma - \gamma_0\|^2 \|\nabla_{\gamma\gamma} q_i(w,\gamma_0)\| + o(\|\gamma - \gamma_0\|^2) \end{aligned} \tag{100}$$

using the triangle inequality and the Cauchy-Schwarz inequality. Therefore for $\|\gamma - \gamma_0\|$ small

$$|q_i(w,\gamma) - q_i(w,\gamma_0) + \nabla_\gamma q_i(w,\gamma_0)(\gamma(z) - \gamma_0(z))| \leq \|\gamma - \gamma_0\|^2 \|\nabla_{\gamma\gamma} q_i(w,\gamma_0)\| \tag{101}$$

so that

$$\begin{aligned} \|q(w,\gamma) - q(w,\gamma_0) + \nabla_\gamma q(w,\gamma_0)(\gamma - \gamma_0)\| &\leq \|\gamma - \gamma_0\|^2 \|\nabla_{\gamma\gamma} q(w,\gamma_0)\| \\ \|q(w,\gamma) - q(w,\gamma_0) + D(w,\gamma - \gamma_0)\| &\leq \|\gamma - \gamma_0\|^2 \|\nabla_{\gamma\gamma} q(w,\gamma_0)\| \end{aligned} \tag{102}$$

where $D(w,\gamma) = [\nabla_\gamma q(w,\gamma_0)]_{5X2} [\gamma]_{2X1}$ Next, note that

$$|D(w,\gamma)| \leq \|\nabla_\gamma q(w,\gamma_0)\| \|\gamma - \gamma_0\| \tag{103}$$

and we note that since each element of $\nabla_\gamma q(w,\gamma_0)$ is bounded it follows that $E\left[\|\nabla_\gamma q(w,\gamma_0)\|^2\right] < \infty$

Next, we establish the form of the influence function

$$\begin{aligned} \int D(w,\gamma) F_0(dw) &= \int f_z(z) E[\nabla_\gamma q(w,\gamma_0) | z] \gamma(z) dz \\ &= \int v(z) \gamma(z) dz \end{aligned} \tag{104}$$

so that by the arguments on p.2208 of [42] we have the influence function for $g(w,\hat{\gamma})$

$$\delta(w) = v(z)\tilde{x} - E[v(z)\tilde{x}] \tag{105}$$

and again by the boundedness of $\nabla_\gamma q(w,\gamma_0)$ it follows that $\int \|v(z)\| dz < \infty$. Also note that $E\left[\|q(w,\gamma_0\|^2\right] < \infty$

Finally, in order to apply Theorem 8.12 we need to guarantee the convergence of the Jacobian term which is ensured by D5.

A.0.6 Identification for Distributional Assumption 2

Identification (Lemma 8)follows since (34) implies (33) holds. Similarly, Lemma 9 also follows as (35) implies (33) holds.

A.0.7 Consistency for Distributional Assumption 2

Consider the model (32)under $F1 - F4$. Then by a modification for i.i.d. data by Theorem 1 in [19] we have that

$$\|\hat{\alpha} - \alpha_0\| = O_p(\max(\delta_n, \|\alpha_0 - \Pi_n \alpha_0\|)) \tag{106}$$

In order to check the conditions for the theorem we first need to impose a metric on the space $B \times \Lambda$. A convenient metric (which will be useful in the normality proof) that is referred to as the Fisher metric:

$$\|\alpha_1 - \alpha_2\| = \sqrt{E[dl_{\alpha_0}(w, \alpha_1 - \alpha_2)]^2} \tag{107}$$

where dl_{α_0} is the pathwise derivative of l at α_0 as detailed in the text. In order to check $(F1)$ we apply Ossiander's [44] result which bounds $N_{[]}(\varepsilon, F_n, L_2)$ by $N_{[]}\left(\varepsilon, \{\alpha \in \Lambda_n : \|\alpha - \Pi_n \alpha_0\| < \delta\}, \|\|\|_{\sup}\right)$and which under smoothness conditions satisfied here gives the bound as $Kr_n \log(\delta/\varepsilon)$. This implies that we can choose $\delta_n = r_n^{1/2} n^{-1/2}$.

Finally, in order to show consistency using Theorem 1 in Chen and Shen [19] we need to check the condition $(A4)$in their paper which states that

$$\sup_{\{\alpha \in B \times \Lambda_n : \|\alpha - \alpha_0\| < \delta\}} |l(w,\alpha) - l(w,\alpha_0)| \leq U_n(w)\,\delta^s \tag{108}$$

with $\sup_n E[U_n(w)^\gamma] \leq C_3$ for some constant C_3 and some $\gamma > 2$. We will use the pathwise derivative of the map $\alpha \to l(\alpha, w)$ which we assume is differentiable over all α for each w. The pathwise derivative at a point $\bar{\alpha}$ is a linear function mapping Λ to $\mathbb{R}$ and is given by

$$dl_{\bar{\alpha}}(w,\alpha) = \left.\frac{\partial l(w,b)}{\partial b}\right|_{b=\bar{b}}(b) + \sum_{j=1}^{5} A_j(w,b,\bar{\eta})\left(\eta_j\right) \tag{109}$$

where

$$A(w,b,\eta) = \tag{110}$$

$$\left[s\frac{(x_1-\eta_1)}{\eta_1(1-\eta_1)}, (1-s)\frac{(1-x_1-\eta_2)}{\eta_2(1-\eta_2)}, s\frac{(x_2-\eta_3)}{\eta_3(1-\eta_3)}, (1-s)\frac{(1-x_2-\eta_4)}{\eta_4(1-\eta_4)}, \frac{(1-x_2-\eta_5)}{\eta_5(1-\eta_5)}\right] \tag{111}$$

and

$$\begin{aligned}
s &= d_0/(d_0+d_1) \\
d_0 &= (1-F(b_0+b_2z))^{1-y}(F(b_0+b_2z))^y\left(\eta_1(z)^{x_1}(1-\eta_1(z))^{1-x_1}\right)\left(\eta_3(z)^{x_2}(1-\eta_3(z))^{1-x_2}\right)(1-\eta_5(z)) \\
d_1 &= (1-F(b_0+b_1+b_2z))^{1-y}(F(b_0+b_1+b_2z))^y\left(\eta_2(z)^{1-x_1}(1-\eta_2(z))^{x_1}\right)\left(\eta_4(z)^{1-x_2}(1-\eta_4(z))^{x_2}\right)\eta_5(z)
\end{aligned}$$

We assume that for each w the pathwise derivative exists for all points along the closed line segment with endpoints $\alpha_0$and α. Then, there exists (see for instance Flett [21] p.214, 254) at least one point $\bar{\alpha}$ along the line segment such that

$$\begin{aligned}
|l(w,\alpha) - l(w,\alpha_0)| &\leq |dl_{\bar{\alpha}}(w, \alpha(z) - \alpha_0(z))| \\
&\leq \|dl_{\bar{\alpha}}(w)\|_E \|(\alpha(z) - \alpha_0(z))\|_E \\
&\leq \|dl_{\bar{\alpha}}(w)\|_E \|(\alpha - \alpha_0)\|_\infty
\end{aligned} \tag{112}$$

where $\|\|_E$ denotes the Euclidean norm and $\|\alpha - \alpha\|_\infty^2 = \|b - \beta_0\|^2 + \Sigma_{i=1}^5 \left(\sup_{z \in S_Z} \left|\eta_j(z) - \eta_0(z)\right|\right)^2$. We next use an interpolation inequality result of Lemma 2 in Shen and Wong [50] to translate the result about the sup norm into one for the L_2 norm to obtain

$$|l(w,\alpha) - l(w,\alpha_0)| \leq \|dl_{\bar{\alpha}}(w)\|_E \|(\alpha - \alpha_0)\|_2^{\frac{2(s+\gamma)}{2(s+\gamma)+1}} \tag{113}$$

where under $(F3)$ we can replace the last term on the right hand side with the Fisher metric so that

$$|l(w,\alpha) - l(w,\alpha_0)| \leq \|dl_{\bar{\alpha}}(w)\|_E \|(\alpha - \alpha_0)\|^{\frac{2(s+\gamma)}{2(s+\gamma)+1}} \tag{114}$$

and given $(F4)$ we see that conditions $(A4)$ in [19] holds.

A.0.8 Asymptotic Distribution for Distributional Assumption 2

In this section we provide sufficient conditions for Theorem 1 of Shen [49] to hold. The first condition is essentially a stochastic equicontinuity condition.

We First write a pointwise (in w) Taylor series Expansion of l around $\alpha_0(z)$. Since, since l depends on α only through its value at the point $z\ \alpha(z)$, we can calculate a direct series expansion element by element to obtain

$$l(w,\alpha) = l(w,\alpha_0) + dl_{\alpha_0}(w)(\alpha-\alpha_0) + (\alpha-\alpha_0)' dl^2_{\bar{\alpha}}(w)(\alpha-\alpha_0) \tag{115}$$

where $dl^2_{\bar{\alpha}}$ is the matrix of second partial derivatives of l (treating b and η as the variables of interest) evaluated at some point $\bar{\alpha}$. Then,

$$r(w,\alpha-\alpha_0) \equiv (\alpha-\alpha_0)' dl^2_{\bar{\alpha}}(w)(\alpha-\alpha_0) \tag{116}$$

$$\begin{aligned} r(w,\alpha-\alpha_0) - r(w,\Pi_n\alpha^*-\alpha_0) &= (\alpha-\alpha_0)' dl^2_{\bar{\alpha}}(w)(\alpha-\alpha_0) \\ &\quad - (\Pi_n\alpha^*-\alpha_0)' dl^2_{\bar{\alpha}_2}(w)(\Pi_n\alpha^*-\alpha_0) \end{aligned} \tag{117}$$

If we assume that both $dl^2_{\bar{\alpha}}$and $dl^2_{\bar{\alpha}}$ are both equal to $dl^2_{\alpha_0}(w) \equiv V$, then we can rewrite the difference as

$$\begin{aligned} & -\Pi_n u'\varepsilon_n V(2(\alpha-\alpha_0)+\Pi_n u\varepsilon_n) \\ = \ & -\Pi_n u'\varepsilon_n V 2(\alpha-\alpha_0) - \Pi_n u\varepsilon_n V\Pi_n u\varepsilon_n \\ = \ & -\varepsilon_n^2\Pi_n u'V\Pi_n u - 2\varepsilon_n\Pi_n uV(\alpha-\alpha_0) \end{aligned} \tag{118}$$

Consider the class of functions $S_n^{s,j} = \left\{[\Pi_n u(w)]_s [V(w)]_{sj}\eta_j(z) - \eta_{0j}(z) : \eta_j \in H\right\}$. If we can show that this class is stochastically equicontinuous (for instance by showing that it is P-Donsker), then for any sequence δ_n going to 0 and an appropriate norm $\|\|$on H

$$\sup_{\{s\in H:\|s-\eta_{0j}\|<\delta_n\}} \mathbb{P}_n\left([\Pi_n u(w)]_s [V(w)]_{sj}\eta_j(z) - \eta_{0j}(z)\right) = o_p\left(n^{-1/2}\right) \tag{119}$$

A sufficient condition for the class to be Donsker given our assumptions on H is that $E\left[[\Pi_n u(w)]_s [V(w)]_{sj}\right]^2 < \infty$. This is enough to ensure that $(G1)$ holds.

Next,

$$\begin{aligned} & K(\alpha_0,\Pi_n\alpha^*(\alpha)) - K(\alpha_0,\alpha) = \\ & E\left[dl_{\alpha_0}(w,\alpha-\Pi_n\alpha^*(\alpha))\right] \\ & +E\left[(\alpha-\alpha_0)' dl^2_{\bar{\alpha}}(w)(\alpha-\alpha_0)\right] - E\left[(\alpha^*-\alpha_0)' dl^2_{\bar{\alpha}_2}(w)(\alpha^*-\alpha_0)\right] \end{aligned} \tag{120}$$

and as before we replace the dl^2 terms with the corresponding term evaluated at α_0 and assume that the matrix $E\left[dl^2_{\alpha_0}(w)\right]$ is positive definite and its biggest and smallest eigenvalues are bounded by constants c_1 and c_2 almost everywhere, then

$$c_2\|\alpha-\alpha_0\|_2^2 \leq E\left[(\alpha-\alpha_0)' dl^2_{\bar{\alpha}}(w)(\alpha-\alpha_0)\right] \leq c_1\|\alpha-\alpha_0\|_2^2 \tag{121}$$

and under $(F3)$ note that for the Fisher metric

$$c_1\|\alpha-\alpha_0\|_2^2 \leq \|\alpha-\alpha_0\|^2 \leq c_2\|\alpha-\alpha_0\|_2^2 \tag{122}$$

Further under $(G4)$ and the condition that

$$\mathbb{P}_n\left(dl_{\alpha_0}\left(w, \Pi_n u\right)\right) = O_p\left(n^{-1/2}\right) \tag{123}$$

$(G2)$ will hold.

A sufficient condition for $(G4)$ is

$$\mathbb{P}_n\left(dl_{\alpha_0}\left(w, u - \Pi_n u\right)\right) = O_p\left(n^{-1/2}\right) \tag{124}$$

$(G3)$ will be satisfied as long as $\delta_n^2 = o\left(n^{-1/2}\right)$ and $(G5)$ will hold if the class of functions $M = \{dl_{\alpha_0}\left(w, \alpha - \alpha_0\right) : \alpha \in \Lambda\}$ is Donsker. If Λ is Donsker and the elements of $dl_{\alpha_0}\left(w\right)$ are square P-integrable, then M is Donsker (see for instance p.193 in Vaart and Wellner [53]).

References

[1] C. Ai and X. Chen, Efficient Sieve Minimum Distance Estimation of Semiparametric Conditional Moment Models. Working Paper, May 2001.

[2] D. Aigner, Regression with a Binary Independent Variable subject to Errors of Observations, *Journal of Econometrics*, 1 (1973), pp. 49–60.

[3] D. Andrews, Estimation when a Parameter is on a Boundary, *Econometrica*, 67 (1999), pp. 1341–1383.

[4] J. Begun, W. Hall, W. Huang, and J. Wellner, Information and Asymptotic Efficiency in Parametric-Nonparametric Models, *Annals of Statistics*, 11 (1983), pp. 432–452.

[5] D. Belman and J. Heywood, Direct and Indirect Effects of Unionization and Government Employment on Fringe Benefit Provision, *Journal of Labor Research*, XII (1991), pp. 111–122.

[6] P. Bickel, C. Klaassen, Y. Ritov, and J. Wellner, *Efficient and Adaptive Estimation for Semiparametric Models*, Johns Hopkins University Press: Baltimore, 1993.

[7] L. Birge and P.Massart, Rate of Convergence for Minimum Contrast Estimators, *Probability Theory and Related Fields*, 97 (1993), pp. 115–133.

[8] D. Black, M. C. Berger, and F. A. Scott, Bounding Parameter Estimates with Nonclassical Measurement Error, *Journal of the American Statistical Association*, 95 (2000), pp. 739–748.

[9] C. Bollinger, Bounding Mean Regressions when a Binary Regressor is Mismeasured, *Journal of Econometrics*, 73 (1996), pp. pp 387–399.

[10] ——, Measurement Error in the Current Population Survey: A Nonparametric Look, *Journal of Labor Economics*, 16 (1998), pp. 576–594.

[11] C. Bollinger and M.H.David, Measuring Discrete Choice with Response Error: Food Stamp Participation, *Journal of the American Statistical Association*, 92 (1997), pp. 827–835.

[12] J. Bound, C. Brown, and N. Mathiowetz, Measurement Error in Survey Data. Institute for Social Research, University of Michigan, April 2000.

[13] J. Bound and A. Krueger, The Extent of Measurement Error in Longitudnal Earnings Data: Do Two Wrongs Make a Right, *Journal of Labor Economics*, 12 (1991), pp. 345–368.

[14] D. Card, The Effect of Unions on the Structure of Wages: A Longitudnal Analysis, *Econometrica*, 64 (1996), pp. 957–979.

[15] R. Carroll, D. Ruppert, and L. Stefanski, *Measurement Error in Non-Linear Models*, Chapman and Hall, 1995.

[16] R. Carroll and M. Wand, Semiparametric Estimation in Logistic Regression Models, *Journal of the Royal Statistical Society*, 53 (1991), pp. 573–585.

[17] X. Chen, H. Hong, and E. Tamer, Measurement Error Models with Auxiliary Data. Princeton University, April 2002.

[18] X. Chen and J. Huang, Semiparametric and Nonparametric Estimation via the Method of Sieves. November 2002.

[19] X. Chen and X. Shen, Sieve Extremum Estimates for Weakly Dependent Data, *Econometrica*, 66 (1998), pp. 289–314.

[20] L. Debnath and P. Mikusinski, *Hilbert Spaces with Applications*, Academic Press, 2nd ed., 1999.

[21] T. Flett, *Differential Analysis*, London: Cambridge University Press, 1980.

[22] R. Freeman, The Effect of Unionism on Fringe Benefits, *Industrial Relations Review*, 34 (1981), pp. 489–509.

[23] A. R. Gallant and D. W. Nychka, Semi-Nonparametric maximum likelihood estimation, *Econometrica*, 55 (1987), pp. 363–390.

[24] S. Geman and C. Hwang, Nonparametric Maximum Likelihood Estimation by the Method of Sieves, *Annals of Statistics*, 10 (1982), pp. 401–414.

[25] C. Geyer, On the Asymptotics of Constrained M-Estimation, *Annals of Statistics*, 22 (1994), pp. 1993–2010.

[26] U. Grenander, *Abstract Inference*, Wiley, Wiley, 1981.

[27] J. Hausman, J. Abreveya, and F. Scott-Morton, Misclassification of the Dependent Variable in a Discrete Response setting, *Journal of Econometrics*, 87 (1998), pp. 239–269.

[28] J. Hausman, W. Newey, H. Ichimura, and J. Powell, Identification and Estimation of Polynomial Errors-in-Variables Models, *Journal of Econometrics*, 50 (1991), pp. 273–296.

[29] H. Hong and E. Tamer, A Simple Estimator for Non Linear Errors in Variables Models. Princeton University, 2001.

[30] J. Horowitz, *Semiparametric Methods in Econometrics*, Springer-Verlag, 1998.

[31] J. Horowitz and C. Manski, Identification and Robustness with Contaminated and Corrupt data, *Econometrica*, 63 (1995), pp. 281–302.

[32] G. Imbens and D. Hyslop, Bias from Classical and other forms of Measurement Error, Tech. Report 257, National Bureau of Economic Research, August 2000.

[33] T. Kane, C. E. Rouse, and D. Staiger, Estimating Returns to Schooling when Schooling is Misreported, Tech. Report 7235, National Bureau of Economic Research, 1999.

[34] L. Lee and J. Sepanski, Estimation of Linear and Nonlinear Errors-in-Variables Models using Validation Data, *Journal of the American Statistical Association*, 90 (1995), pp. 130–140.

[35] A. Lewbel, Identification of the Binary Choice Model with Misclassification, *Econometric Theory*, 16 (2000), pp. 603–60.

[36] T. Li, Estimation of Non Linear Errors-in-Variables Models: A Semiparametric Minimum Distance Estimator. Working Paper, Washington State University, 1998.

[37] B. Madrian and L. Lefgren, A Note on Longitudnally Matching Current Population Survey (CPS) Respondents, Tech. Report 247, National Bureau of Economic Research, November 1999.

[38] A. Mahajan, Estimating Price Elasticities with Non-Linear Errors-in-Variables. mimeo, Princeton University, 2002.

[39] C. Manski and E. Tamer, Inference on Regressions with Interval Data on a Regressor or Outcome, *Econometrica*, 70 (2002), pp. 519–546.

[40] W. Mellow and H. Sider, Accuracy of Response in Labor Market Surveys: Evidence and Implications, *Journal of Labour Economics*, 1 (1983), pp. 331–44.

[41] S. Murphy and A. van der Vaart, On Profile Likelihood, *Journal of the American Statistical Association*, 95 (2000), pp. 449–465.

[42] W. Newey and D. McFadden, Large Sample Estimation and Hypothesis Testing, *Handbook of Econometrics*, R. Engle and D. McFadden, eds., vol. IV, Elsevier Science, 1994, ch. 36, pp. 2111–2245.

[43] W. L. Newey, Kernel Estimation of Partial Means and a General Variance Estimator, *Econometric Theory*, 10 (1994), pp. 233–253.

[44] M. Ossiander, A Central Limit Theorem under Metric Entropy with L2 Bracketing, *Annals of Probability*, 17 (1987), pp. 897–919.

[45] D. Pollard, *Convergence of Stochastic Processes*, New York: Springer-Verlag, 1984.

[46] P. Robinson, Asymptotically Efficient Estimation in the Presence of Heteroscedasticity of Unknown Form, *Econometrica*, 55 (1987), pp. 875–891.

[47] S. Schennach, Estimation of Non Linear Models with Measurement Error. University of Chicago, August (2001).

[48] T. Severini and W. Wong, Profile Likelihood and Conditionally Parametric Models, *Annals of Statistics*, 20 (1992), pp. 1768–1802.

[49] X. Shen, On Methods of Sieves and Penalization, *Annals of Statistics*, 25 (1997), pp. 2555–2591.

[50] X. Shen and W. Wong, Convergence Rates of Sieve Estimates, *Annals of Statistics*, 22 (1994), pp. 580–615.

[51] M. Taupin, Semiparametric Estimation in the Nonlinear Structural Errors-in-Variables Model, *Annals of Statistics*, 29 (2001).

[52] A. van der Vaart, *Asymptotic Statistics*, Cambridge University Press, 1998.

[53] A. van der Vaart and J. Wellner, *Weak Convergence and Empirical Processes With Applications to Statistics*, Springer, 1996.

[54] H. White and J. Wooldridge, Some Results in Sieve Estimation with Dependent Observations, *Nonparametric and Semi-Parametric Methods in Econometrics and Statistics*, J. W.A. Barnett and G.Tauchen, eds., W.A. Barnett, J.Powell and G.Tauchen, Cambridge University Press, New York, 1991, pp. 459–493.

[55] W. Wong and T. Severini, On Maximum Likelihood Estimation in Infinite Dimensional Parameter Spaces, *Annals of Statistics*, 19 (1991), pp. 603–632.

Table 1: Results from Naive Probit Estimation on Misclassified Data

	$\beta_0 = 0$	$\beta_1 = 1$	$\beta_2 = -1$
N=2000			
Mean	.21	5.19	-10.76
Median	.20	4.86	-10.15
Std. Dev	.19	1.39	2.77
IQR	.25	2.20	4.37
N=4000			
Mean	.20	5.26	-10.81
Median	.21	4.87	-10.11
Std. Dev	.13	1.25	2.35
IQR	.17	1.88	3.75
N=8000			
Mean	.20	5.02	-10.43
Median	.20	4.86	-10.12
Std. Dev	.09	.95	1.87
IQR	.12	1.34	2.65

Model: $P(y = 1) = \Phi(x^* - z)$
x^* binary unobserved
Observe $\{y, x, z\}$

Table 2: Results from Method 1 Estimation when A2 is violated

	$\beta_0 = 0$	$\beta_1 = 1$	$\beta_2 = -1$
N=2000			
Mean	.89	-.84	-.91
Median	.86	-.79	.-.91
Std. Dev	.10	.16	.15
IQR	.11	.20	.21
N=4000			
Mean	.89	-.82	-.89
Median	.87	-.80	-.89
Std. Dev	.09	.12	.09
IQR	.12	.12	.10
N=8000			
Mean	.91	-.85	-.91
Median	.88	-.85	-.90
Std. Dev	.08	.09	.07
IQR	.10	.15	.09

$P(y = 1) = \Phi(x^* - z)$
$P(x = 1|x^* = 0, z) = \alpha_0$
$P(x = 0|x^* = 1, z) = \alpha_1$
$z \sim Uniform[-1, 1]$
but $\alpha_0 + \alpha_1 > 1$

Table 3: Monte Carlo Simulations: Distributional Assumption 1

	$\alpha_0 + \alpha_1 = .15$			$\alpha_0 + \alpha_1 = .30$			$\alpha_0 + \alpha_1 = .45$		
	β_0	β_1	β_2	$\beta_0 = 0$	$\beta_1 = 1$	$\beta_2 = -1$	β_0	β_1	β_2
N=200									
Mean	1.18	2.33	-2.25	.94	2.23	-2.1	.25	2.24	-1.78
Median	1.53	2.56	-2.64	1.25	2.32	-2.38	-.81	1.91	-1.28
Std. Dev	1.25	1.32	1.18	1.33	1.53	1.17	1.90	2.37	1.17
IQR	2.13	2.15	2.20	1.71	2.22	2.19	2.27	3.30	-2.14
N=400									
Mean	.80	1.86	-1.83	.773	1.84	-1.80	.32	1.85	-1.60
Median	-.44	1.35	-1.18	.23	1.37	-1.16	-.42	1.31	-1.03
Std. Dev	1.04	1.07	1.03	1.08	1.15	1.01	1.32	1.66	1.05
IQR	1.97	2.00	-1.95	1.93	1.97	1.93	1.80	2.06	-1.80
N=800									
Mean	.41	1.40	-1.2	.37	1.40	-1.39	.29	1.56	-1.42
Median	.07	1.12	-1.07	.03	1.06	-1.03	.03	1.15	-1.04
Std. Dev	.81	.83	.80	.72	.87	.81	1.03	1.24	.90
IQR	.47	.61	.38	.46	.60	.35	.53	1.02	.56
N=1600									
Mean	.12	1.10	-1.05	.12	1.14	-1.13	.09	1.17	-1.13
Median	.006	1.02	-1.01	.02	1.04	-1.01	.006	1.07	-1.01
Std. Dev	.44	.47	.43	.48	.51	.48	.52	.56	.48
IQR	.24	.31	.24	.28	.36	.27	.31	.44	.29

Model: $P(y = 1) = \Phi(x^* - z)$
$P(x = 1|x^* = 0, z) = \alpha_0 \quad P(x = 0|x^* = 1, z) = \alpha_1$
z Discrete

Table 4: Monte Carlo Simulations: Distributional Assumption 1 ("IV")

	$\alpha_0 + \alpha_1 = .15$			$\alpha_0 + \alpha_1 = .30$			$\alpha_0 + \alpha_1 = .45$		
	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2
N=200									
Mean	.37	1.54	-1.47	.18	1.43	-1.35	-.13	1.07	-1.09
Median	0	1.24	-1.14	0	1.12	-1.07	-.32	.84	-.88
Std. Dev	1.03	1.09	.97	1.02	1.34	.87	1.00	1.41	.72
IQR	.74	.86	.73	.96	.85	.62	..64	1.01	.44
N=400									
Mean	.24	1.36	-1.30	.18	1.26	-1.23	-.09	1.07	-1.03
Median	0	1.16	-1.10	0.17	1.02	1.00	-.26	.88	-.89
Std. Dev	.71	.83	.67	.74	.88	.74	.75	.92	.57
IQR	.11	.61	.42	.34	.61	.40	.48	.77	.35
N=800									
Mean	.12	1.16	-1.14	.07	1.10	-1.10	-.07	1.02	-1.00
Median	-.16	1.08	-1.05	0.10	1.00	-1.01	.01	.95	-.93
Std. Dev	.44	.48	.44	.48	.53	.46	.47	.62	.36
IQR	.20	.38	.28	.34	.42	.30	.25	.56	.28
N=1600									
Mean	.02	1.05	1.04	.01	1.02	-1.03	-.05	.98	-.97
Median	0	1.02	-1.01	-.02	1.00	-.99	.14	.98	-.96
Std. Dev	.19	.28	.32	.31	.27	.20	.27	.40	.20
IQR	.17	.27	.18	.24	.32	.21	.15	.39	.20

Model: $P(y = 1) = \Phi(x^* - z)$

$P(x = 1|x^* = 0, z, v) = \alpha_0 \quad P(x = 0|x^* = 1, z, v) = \alpha_1$

z Discrete

Table 5: Monte Carlo Simulations: Distributional Assumption 1

	$\alpha_0+\alpha_1=.15$			$\alpha_0+\alpha_1=.30$			$\alpha_0+\alpha_1=.45$		
	β_0	β_1	β_2	$\beta_0=0$	$\beta_1=1$	$\beta_2=-1$	β_0	β_1	β_2
N=200									
Mean	.004	1.05	-1.02	.07	.95	-1.01	.15	.77	-.94
Median	.006	1.00	-1.01	.07	.82	-.98	.20	.66	-.89
Std. Dev	.27	.51	.18	.28	.52	.18	.46	.84	.22
IQR	.35	.58	.20	.37	.60	.20	.52	.90	.22
N=400									
Mean	.02	.96	-1.01	.039	.94	-.97	.19	.68	-.91
Median	.02	.92	-.99	.056	.90	-.96	.24	.60	-.92
Std. Dev	.18	.35	.112	.21	.37	.12	.29	.52	.15
IQR	.24	.43	.15	.23	.43	.15	.38	.75	.19
N=800									
Mean	.01	.98	-1.00	.04	.90	-.96	.09	.79	-.93
Median	.01	.95	-.97	.06	.88	-.96	.06	.73	-.90
Std. Dev	.13	.25	.08	.16	.27	.08	.24	.49	.13
IQR	.17	.40	.13	.21	.32	.09	.34	.69	.19
N=1600									
Mean	.004	1.01	-1.01	.01	.98	-.99	.09	.80	-.94
Median	.009	1.00	-1.01	.002	.98	-.98	.08	.81	-.91
Std. Dev	.10	.18	.06	.14	.26	.08	.20	.34	.10
IQR	.15	.25	.07	.16	.30	.12	.32	.48	.16

Model: $P(y=1)=\Phi(x^*-z)$
$P(x=1|x^*=0,z)=\alpha_0 \quad P(x=0|x^*=1,z)=\alpha_1$
$z\sim$ Uniform$[-1,1]$

Table 6: Monte Carlo Simulations: Distributional Assumption 1

	$\alpha_0+\alpha_1=.15$			$\alpha_0+\alpha_1=.30$			$\alpha_0+\alpha_1=.45$		
	β_0	β_1	β_2	$\beta_0=0$	$\beta_1=1$	$\beta_2=-1$	β_0	β_1	β_2
N=200									
Mean	-.05	1.13	-1.07	-.003	1.04	-1.04	.11	.77	-.91
Median	-.02	1.05	-1.03	.04	.88	-.98	.15	.65	-.89
Std. Dev	.24	.46	.26	.30	.58	.27	.29	.48	.24
IQR	.27	.42	.30	.35	.58	.30	.25	.52	.25
N=400									
Mean	-.03	1.07	-1.02	.03	.92	-.96	.08	.86	-.93
Median	-.03	1.05	-.99	.06	.86	-.95	.13	.78	-.89
Std. Dev	.15	.29	.15	.18	.44	.16	.22	.39	.18
IQR	.22	.35	.17	.22	.37	.19	.27	.53	.20
N=800									
Mean	-.03	1.03	-1.01	.01	.98	-1.00	.06	.90	-.95
Median	.01	.99	-.1.00	.02	.94	-.98	.10	.81	-.92
Std. Dev	.12	.25	.13	.15	.30	.13	.18	.30	.16
IQR	.17	.31	.16	.23	.44	.19	.24	.48	.21
N=1600									
Mean	0.01	1.01	-1.02	.002	.99	-1.00	.02	.92	-.96
Median	0.01	.99	-1.00	.01	.97	-.98	.01	.87	-.93
Std. Dev	.08	.18	.10	.12	.23	.11	.14	.26	.13
IQR	.09	.22	.10	.15	.35	.16	.23	.43	.18

Model: $P(y=1)=\Phi(x^*-z)$
$P(x=1|x^*=0,z)=\alpha_0 \quad P(x=0|x^*=1,z)=\alpha_1$
$z\sim \text{Normal}(0,1)$

Table 7: Monte Carlo Simulations: Distributional Assumption 2

	$Corr(x_i, x^*) = .85$			$Corr\,(x_i, x^*) = .70$			$Corr\,(x_i, x^*) = .55$		
	β_0	β_1	β_2	$\beta_0 = 0$	$\beta_1 = 1$	$\beta_2 = -1$	β_0	β_1	β_2
N=200									
Mean	.07	.96	-1.08	.27	.73	-1.05	.34	.75	-1.03
Median	.07	.90	-1.02	.26	.72	-.99	.32	.68	-1.00
Std. Dev	.15	.31	.47	.15	.27	.16	.15	.43	.24
IQR	.21	.35	.23	.20	.32	.21	.21	.36	.19
N=400									
Mean	.09	.95	-1.01	.26	.78	-.99	.32	.69	-.99
Median	.10	.92	-.98	.27	.80	-.99	.32	.67	-.97
Std. Dev	.11	.23	.20	.09	.17	.11	.10	.29	.18
IQR	.16	.19	.15	.12	.25	.14	.15	.28	.12
N=800									
Mean	.08	.95	-.97	.16	.79	-.98	.29	.72	-.99
Median	.07	.98	-.98	.18	.80	-.97	.31	.70	.98
Std. Dev	.07	.13	.18	.06	.11	.08	.07	.18	.10
IQR	.10	.16	.10	.09	.13	.11	.08	.20	.10
N=1600									
Mean	.03	.98	-.99	.15	.88	-1.00	.26	.75	-.97
Median	.07	.96	-.98	.14	.86	-.98	.32	.72	-.97
Std. Dev	.05	.10	.09	.06	.10	.12	.04	.13	.07
IQR	.06	.08	.07	.05	.11	.08	.05	.12	.08

Model: $P\,(y = 1) = \Phi\,(x^* - z)$

$P\,(x_1 = 1|x^* = 0, z) = \alpha_0(z) \quad P\,(x_1 = 0|x^* = 1, z) = \alpha_1(z)$

$P\,(x_2 = 1|x^* = 0, z) = \gamma_0(z) \quad P\,(x_2 = 0|x^* = 1, z) = \gamma_1(z)$

$z \sim$ Uniform$[-2, 2]$

Table 8: Monte Carlo Simulations: Distributional Assumption 2

	$Corr(x_i, x^*) = .97$			$Corr\,(x_i, x^*) = .95$			$Corr\,(x_i, x^*) = .90$		
	β_0	β_1	β_2	$\beta_0 = 0$	$\beta_1 = 1$	$\beta_2 = -1$	β_0	β_1	β_2
N=200									
Mean	.024	.94	-1.0	.05	.97	-1.04	.06	.96	-1.02
Median	.023	.934	-.97	.04	.99	-1.03	.05	.92	-1.02
Std. Dev	.16	.23	.16	.17	.24	.19	.17	.26	.17
IQR	.20	.32	.19	.22	.29	.23	.16	.32	.21
N=400									
Mean	.02	.96	-1.0	.05	.97	-.997	.06	.95	-1.02
Median	.01	.95	-.98	.05	.98	-.98	.05	.95	-1.01
Std. Dev	.10	.14	.09	.12	.18	.11	.12	.17	.11
IQR	.13	.18	.13	.19	.22	.14	.15	.17	.14
N=800									
Mean	.007	.991	-.994	.06	.98	-.999	.05	.95	-.98
Median	-.002	.998	-.987	.07	.97	-.990	.05	.94	-.98
Std. Dev	.08	.11	.07	.08	.10	.09	.07	.12	.06
IQR	.10	.16	.11	.09	.13	.11	.10	.15	.09
N=1600									
Mean	.007	.989	-.995	.05	.99	-.98	.04	.97	-.98
Median	.004	.976	-.997	.05	.97	-.98	.04	.96	-.98
Std. Dev	.05	.06	.05	.054	.07	.04	.06	.07	.05
IQR	.07	.08	.07	.06	.10	.07	.06	.09	.08

Model: $P\,(y = 1) = \Phi\,(x^* - z)$
$P\,(x_1 = 1|x^* = 0, z) = \alpha_0(z) \quad P\,(x_1 = 0|x^* = 1, z) = \alpha_1(z)$
$P\,(x_2 = 1|x^* = 0, z) = \gamma_0(z) \quad P\,(x_2 = 0|x^* = 1, z) = \gamma_1(z)$
$z \sim$ Uniform$[-2, 2]$

Table 9: Summary Characteristics for CPS data Feb 1999

Sample Characteristics	
Sample Size	3000
Average Age	41.3
Fraction Male	.498
Fraction White	.919
Occupational Categories	
Managerial/Professional	.65
Service	.11
Production, Repair,Craft	.24
Education	
Upto 11th Grade	.06
11th Grade - High School	.33
Some College	.29
BA and above	.31
Health Insurance	.70
Union Membership	.18

Table 10: Results: Probit Estimation (n=3000)

Variable	*Coefficient*	*(Std. Error)*
Union Status	.60	.074
Age	.010	.003
Sex	-.451	.054
Race	.369	.098
Occupation	-.032	.036
Education	.204	.040
Intercept	-.02	.20

Table 11: Results: Distributional Assumption 1 (n=3000)

Variable	*Coefficient*	*(Std. Error)*
Union Status	.760	.122
Age	.010	.002
Sex	-.430	.055
Race	.364	.104
Occupation	-.050	.038
Education	.215	.028
Intercept	.014	.197
Misclassification Probabilities		
False Positive (α_0)	.033	.009
False Negative (α_1)	.139	.052

Table 12: Results: Distributional Assumption 1 ("IV" with additional Surrogate) (n=3000)

Variable	*Coefficient*	*(Std. Error)*
Union Status	.620	.088
Age	.010	.003
Sex	-.450	.055
Race	.358	.097
Occupation	-.037	.035
Education	.205	.030
Intercept	.010	.199
Misclassification Probabilities		
False Positive (α_0)	.032	.002
False Negative (α_1)	.099	.039

Table 13: Results: Distributional Assumption 2 (n=3000)

Variable	*Coefficient*	*(Std. Error)*
Union Status	.949	.090
Age	.008	.003
Sex	-.564	.069
Race	.490	.095
Occupation	-.053	.048
Education	.121	.030
Intercept	-.09	.154

Table 14: Comparison of Marginal Effects of Union Status

Estimation Method	*Point Estimate*	*Std. Error*
Probit	.172	.018
Distribution Assn 1	.233	.033
Distribution Assn 1 ("IV")	.180	.019
Distribution Assn 2	.288	.016