Identification and Estimation of Regression Models with Misclassification

Aprajit Mahajan [1]

First Version: October 1, 2002
This Version: December 1, 2005

[1]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, Ernst Schaumburg and members of the Princeton Microeconometrics Reading Group. I would like to thank John Pepper for his comments at the SEA meetings in San Antonio 2003 and seminar participants at Princeton, Georgetown, Yale, UCLA, Stanford, University of Chicago, UW Madison, New York University, LSE, UC Berkeley, UC-San Diego, UC-Davis, Duke, MIT-Harvard, CEME and Northwestern. I also would like to thank the co-editor and three anonymous referees for detailed comments.

Abstract

This paper studies the problem of identification and estimation in nonparametric regression models with a misclassified binary regressor where the measurement error may be correlated with the regressors. We show that the regression function is non-parametrically identified in the presence of an additional random variable that is correlated with the unobserved true underlying variable but unrelated to the measurement error. Identification for semi-parametric and parametric regression functions follows straightforwardly from the basic identification result. We propose a kernel estimator based on the identification strategy and derive its large sample properties and also discuss alternative estimation procedures.

KEYWORDS: Non-Classical Measurement Error, Identification, Non-Linear Models, Misclassification.

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. Such 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[2] – 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 non-parametric regression 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 (1991) and Bollinger (1998).[3] In their article on measurement error in survey data, Bound, Brown, and Mathiowetz (2000) report similar dependencies for a number of variables including hours worked, union status, welfare program participation, and employment status.

A second assumption 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 (1997) note that the probability of misreporting AFDC status was highest amongst the poorest households. Mellow and Sider (1983) conclude that over-reporting of hours worked varied systematically with education and race. Summarizing work on unemployment data, Bound, Brown, and Mathiowetz (2000) report that misclassification rates tended to vary by age and sex.[4]

[2] 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.

[3] Also see the caveat therein.

[4] Their paper also provides a comprehensive overview of the current empirical evidence on the nature of measurement error in survey data.

This paper examines the effect of relaxing these assumptions on the identification and estimation of a non-parametric regression model. The object of interest is the effect of a binary random variable x^* on the conditional expectation of an outcome y while controlling for other explanatory variables z. The econometrician observes x, an error-ridden measure of x^*. Since the unobserved variable is binary, the measurement error is necessarily correlated with the true value x^*. Importantly, we also allow the measurement error to depend upon the other explanatory variables in the model.

This paper provides sufficient condition for point identification of the conditional expectation of y given x^* and z. In a related paper Mahajan (2003) obtains sharp bounds for a similar model. The bounds demonstrate that the regression function is in general not point identified in the presence of misclassified regressors. Therefore, further information is needed for identification. In this paper we introduce additional information in the form of what is referred to as an ILV (Instrument-Like Variable) because it is required to satisfy analogues of the conditions for an instrumental variable in a standard linear model. However, it is not equivalent to the standard IV as it is required to satisfy a further condition explained below. This additional information is enough to achieve point identification, and we propose simple estimation strategies for the parameters of interest.

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 describes the general model and presents the fundamental identification result. The section also provides identification results for semi-parametric and fully parametric regression functions which will be shown to follow straightforwardly from the basic identification result. Section 4 outlines an estimation strategy based on the identification result and details the large sample properties of the proposed estimator. Section 5 outlines alternative estimation strategies including one based on the method of sieves that is relatively straightforward to implement. Section 6 discusses testing for mislassification and proposes a simple test based on an exclusion restriction and Section 7 concludes.

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 (1995). Hausman, Newey, Ichimura, and Powell (1991) 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] Under a similar set of conditions Schennach (2004a) proposes an estimation strategy for a much more general class of non-linear models.

Amemiya (1985) considers instrumental variable estimation for non-linear models in the presence of classical measurement error and is able to show consistency and asymptotic normality for his estimator as long as the variance of the measurement error shrinks to zero fast enough (see also Chesher (1991)). Stefanski and Buzas (1995) consider instrumental variable estimation for the binary choice model with classical measurement error and derive "approximately consistent" methods whose accuracy depends upon the degree of non-linearity in the model. Buzas and Stefanski (1996) discuss estimation of generalized linear models with known link functions in the presence of classical measurement error. Identification in their paper rests on the existence of an instrumental variable with a parametrically specified distribution that is unrelated to the measurement error and the outcome. Estimation follows from a modification of the usual score equations to obtain a set of unconditional moment conditions. Newey (2001) also considers instrumental variable estimation for general nonlinear models with measurement error and uses the method of simulated moments while Lewbel (1996) considers instrumental variable estimation of Engel curves with measurement error and imposes distributional assumptions to identify and estimate the model. More recently, Schennach (2004b) considers instrumental variable estimation of non-linear models in the presence of classical measurement error.

Another class of models achieves identification by supplementing the classical independence assumptions with a distributional assumption on the measurement error. Taupin (2001) proposes a consistent estimator for a nonlinear regression model by imposing normality on the measurement error, while Hong and Tamer (2001) derive estimators for a broader class of moment based models by assuming a Laplace 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 (1991) use validation data to estimate a logistic regression with measurement error, as do Lee and Sepanski (1995) 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 (1995). In their work, the observed data are a mixture of the variable of interest

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

and another random variable and the error is allowed to depend upon the variable of interest (see also the important work by Molinari (2004)). Chen, Hong, and Tamer (2002) 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 on non-linear unconditional moment restrictions. Imbens and Hyslop (2000) 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. Under a minimal set of assumptions, Molinari's (2004) work can be used to derive restrictions on the oberved probability distributions in the presence of misclassified data. If the misclassified variable is a response variable, Horowitz and Manski's (1995) work can be used to derive bounds for the unidentified parameters of the conditional distribution of interest. Abrevaya, Hausman, and Scott-Morton (1998) examine the effect of a mismeasured left hand side binary variable within a parametric maximum likelihood as well as a semiparametric framework (see also Lewbel (2000)). The issue of misclassified binary regressors was first addressed by Aigner (1973) and subsequently by Bollinger (1996) 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, Black, Berger, and Scott (2000) 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 (1999) to study the effect of mismeasured schooling on returns to education and Card (1996) 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] This paper is very closely related to Lewbel (2004) who discusses bounds and the estimation of the marginal effect of a misclassified treatment in a very similar context under various related sets of assumptions.

The present paper makes three contributions to the literature on measurement error in non-linear models. First and most importantly, it provides identification results for a large class of non-linear models with measurement error of a fundamentally non-classical nature. Second, we propose an estimator based on the identification argument and establish its large sample properties. Finally, we develop a relatively straightforward maximum likelihood estimator based on the method of sieves for the special case of misclassification in a parametric binary choice model.

[6] Although, Abrevaya, Hausman, and Scott-Morton (1998) discuss a case where this is not true.

3 The Model and Identification

In this section we provide identification results for the regression function in the presence of misclassified regressors in models of the form

$$\mathbb{E}\left[y - g(\tilde{z})|\tilde{z}\right] = 0 \tag{I}$$

where the conditional expectation function $g(\cdot)$ is unknown.

We assume that the random vector $\tilde{z}$ can be partitioned as (x^*, z) where x^* is a binary random variable and z is an observed $d_z \times 1$ random vector. Instead of observing x^* we observe x, a misclassified version of x^* (known as a "surrogate" in the literature). In addition to the surrogate we also observe another random variable v (with properties to be specified below) and the model satisfies

$$\mathbb{E}\left[y - g\left(\tilde{z}\right)|\tilde{z}, x, v\right] = 0 \tag{1}$$

The assumption that the conditional mean of y is unaffected by knowledge of x once x^* is known is referred to as the assumption of non-differential measurement error and asserts that the misclassification rates themselves are uninformative about the outcome of interest given the truth and the other covariates z. The conditional statement is important since the misclassification rates may be informative about responses through their correlation with the other explanatory variables in the model. In their survey paper on measurement error, Bound, Brown, and Mathiowetz (2000) provide instances when such an assumption is likely or unlikely to hold.[7] In the linear context with a convolution model for measurement error, this assumption is analogous to the assumption that the error term in the outcome equation is conditionally mean independent of the measurement error in the mismeasured regressor.

The specification (1) also requires that the variable v be excluded from the outcome equation conditional on all the other covariates (and the surrogate) and is analogous to the exclusion restriction in standard linear instrumental variable models. It is for this reason (and a further analogy explained below) that we refer to v as an ILV ("Instrument Like Variable").[8] It is related to the identification argument in Lewbel (2000) (see also Abrevaya and Hausman (1999)) which requires the existence of an included regressor that does not affect the misclassification probabilities. The differences between those papers and this one

[7] For instance, as a referee has pointed out, non-differential measurement error in the program evaluation context (where x^* denotes treatment status) rules out placebo effects.

[8] A referee has pointed out that it is also possible to interpret this variable exactly as an instrument by imposing the additional structure that $x = h\left(x^*, z, e\right)$ where $e \perp (x^*, z)$ and requiring v to be independent of e and conditionally correlated with x^*. In the paper, however, we do not impose this additional structure since it is not required at any other point.

are that v must be an excluded variable in the model (1) here and that misclassification in Lewbel (2000) is occurs in the left-hand side of a binary choice model.

The identification argument relies on four important assumptions: 1) identification of model (I) in the absence of misclassification, 2) restrictions on the extent of misclassification, 3) independence between the misclassification rates and the ILV conditional on the other regressors and 4) a dependency relationship between the unobserved regressor and the ILV. For simplicity, I assume that the ILV v is binary (taking on values (v_1, v_2)). This clarifies the arguments and is sufficient for identification. These four assumptions and an additional one are outlined below and remarks follow. Throughout, we assume that the econometrician observes an i.i.d. sample $\{y_i, x_i, z_i, v_i\}_{i=1}^n$. Let $z_a \in \mathbb{S}_z$ denote an arbitrary element of the support of z ($\mathbb{S}_z$)and we use the shorthand $\mathbb{P}(\cdot|z_a)$ to denote $\mathbb{P}(\cdot|z = z_a)$

ASSUMPTION 1 *(Identification in the absence of Misclassification) The regression function $g(x^*, z_a)$ in the model (I) is identified given knowledge of the population distribution of $\{y, x^*, z\}$.*

ASSUMPTION 2 *(Restriction on the extent of Misclassification)*

$$\underbrace{\mathbb{P}(x = 1|x^* = 0, z_a)}_{\eta_0(z_a)} + \underbrace{\mathbb{P}(x = 0|x^* = 1, z_a)}_{\eta_1(z_a)} < 1 \tag{2}$$

ASSUMPTION 3 *(Conditional Independence between the surrogate and the ILV)*

$$x \perp v \ \ |(x^*, z_a) \tag{3}$$

ASSUMPTION 4 *(Dependency Condition)* $v_1 \neq v_2$ *and*

$$\mathbb{P}(x^* = 1|z_a, v_1) \neq \mathbb{P}(x^* = 1|z_a, v_2) \tag{4}$$

ASSUMPTION 5 *(Relevance of x^*)*

$$g(1, z_a) \neq g(0, z_a) \tag{5}$$

Assumption 1 is straightforward and requires that the conditional expectation $g(x^*, z_a)$ in the model (I) is identified in the absence of misclassification. Assumption 2 ensures that the misclassification is not "too bad" in the sense that the surrogate is (conditionally) positively correlated with the unobserved x^*. While we allow for misclassification rates to differ across the explanatory variables (e.g. more educated people are less likely to misreport)

we require that the relationship is stable in the sense of Assumption 2. It is possible to weaken this assumption to non-zero correlation between the surrogate and x^* (conditional on z) and knowledge of the sign of the correlation. This information will allow the researcher to relabel the variables so as to satisfy Assumption 2.[9] In its unconditional form, Assumption 2 is referred to as the Monotonicity Assumption by Abrevaya, Hausman, and Scott-Morton (1998) and is crucial for identification as discussed below.

Under just assumptions 1 and 2 the regression function $g(x^*, z_a)$ is not identified. We can, however, obtain sharp bounds on the regression function which are detailed in Mahajan (2003).[10] The bounding results imply that we need further information to point identify the model and the additional information provided by assumptions 3 and 4 provides us with sufficient variation to identify the regression function.

Assumption 3 requires that the ILV be conditionally independent of the misclassification probabilities. This is a strong assumption but it is difficult to relax it and retain point identification. It is analogous to assumptions placed on the repeated measurement in models of measurement error in non-linear models (see for instance Hausman, Newey, Ichimura, and Powell (1991), Schennach (2004a), Li (1998) and for the linear case Kane, Rouse, and Staiger (1999)) and is similar in spirit to the requirement that the instrument be uncorrelated with the measurement error in the standard IV approach to dealing with classical measurement error in linear models. The statement is, however, conditional on the other regressors z so it is possible that the ILV and x are unconditionally correlated with each other.

Assumption 4 is particular to the model in this paper and requires that the ILV be informative about x^* even after conditioning on the other covariates. It is analogous to the requirement in IV models that the instrument be related to the endogenous regressor conditional on all the other regressors (although obviously the analogy is imperfect since x^* is unobserved) and is like the identification assumption in Theorem 2.1 of Das (2004). This points to a link between the proposed model above and endogenous regression models and is discussed in some detail below. Given the other assumptions, Assumption 4 can be tested in principle since it is equivalent to the statement $\mathbb{P}(x = 1|z_a, v_1) \neq \mathbb{P}(x = 1|z_a, v_2)$ and both quantities are directly identified.

Finally, Assumption 5 ensures that x^* is not redundant in the sense that $\mathbb{E}[y|x^*, z_a] \neq \mathbb{E}[y|z_a]$. This assumption can be verified in principle since its negation implies $\mathbb{E}[y|x = 1, z] = \mathbb{E}[y|x = 0, z]$ which is a testable implication since all quantities are directly identified. If

[9] Essentially, all that is required is that for a pair of misclassification probabilities, $(\eta_0(z_a), \eta_1(z_a))$, the pair $(1 - \eta_0(z_a), 1 - \eta_1(z_a))$ is not a permissible pair of misclassification rates. I thank two referees for these points.

[10] For instance in the case where $\mathbb{E}(y|x = 1, z) > \mathbb{E}(y|x = 0, z)$, the sharp bounds are $g(1, z) \in [\mathbb{E}(y|x = 1, z), \infty)$ and $g(0, z) \in (-\infty, \mathbb{E}(y|x = 0, z)]$.

Assumption 5 does not hold, the regression function is trivially identified although the misclassification rates are not. The assumptions above are very close to those in Lewbel (2004) for the case of a two-valued instrument except that we do not require his assumptions (B6) or the second assumption in (A3).

Under these conditions we can now state the central result of the paper.

THEOREM 1 *Under Assumptions 1-5, $g(x^*, z_a)$ and the misclassification rates $\eta_0(z_a)$ and $\eta_1(z_a)$ in model (1) are identified.*

Remark 1 *If Assumptions 1-5 are modified to hold for almost all $z \in \mathbb{S}_z$ (by appending "a.e. $\mathbb{P}_z$" to the displays 2,4 and 5, then the entire regression function $g(\cdot)$ and the misclassification rates $\eta_0(\cdot)$ and $\eta_1(\cdot)$ in model (1) are identified.*

Note that the regression function is identified even though x^* is measured with error that does not conform to the classical error in variables assumptions. The proof is detailed in the appendix but we outline the main steps here. In the first step, we show that if the misclassification rates are identified, then the regression function at (x^*, z_a) is identified. In the second step, we show that the misclassification rates are identified.

To motivate the argument, consider

$$\underbrace{\mathbb{P}(x=1|z_a,v)}_{\eta_2(z_a,v)} = \sum_{s\in\{0,1\}} \mathbb{P}(x=1|x^*=s,z_a,v)\,\mathbb{P}(x^*=s|z_a,v)$$

Let $\eta_2^*(z_a,v) \equiv \mathbb{P}(x^*=1|z_a,v)$. Assumption 3 implies

$$\eta_2^*(z_a,v) = \frac{\eta_2(z_a,v) - \eta_0(z_a)}{1-\eta_0(z_a)-\eta_1(z_a)} \tag{6}$$

where the denominator above is a.e. strictly positive because of Assumption 2. The function $\eta_2(z_a,v)$ is directly identified since $\{x,z,v\}$ are observed. Therefore, if the functions $\{\eta_0(z_a), \eta_1(z_a)\}$ are identified, then $\eta_2^*(z_a,v)$ is also identified.

The first step of the argument concludes that if the misclassification rates $\{\eta_0(z_a), \eta_1(z_a)\}$ are identified, then assumptions 1, 4 and 5 ensure that $g(x^*,z)$ is identified. To see the argument, note that we can write the identified moment $\mathbb{E}[y|z_a,v]$ as

$$\mathbb{E}[y|z_a,v] = g(0,z_a)(1-\eta_2^*(z_a,v)) + g(1,z_a)\,\eta_2^*(z_a,v) \tag{7}$$

If the misclassification rates are identified, then we can use the variation in v and Assumptions 1 and 5 to deduce straightforwardly (indeed it is a linear system of equations) that $g(x^*,z_a)$

is identified. [11]

In the final part of the proof we show that the ILV and the directly observed moments ensure identification up to a "probability flip" – that is to say given a set of misclassification probabilities $(\eta_0(z_a), \eta_1(z_a))$, the ILV ensures that all elements of the identified set (of observationally equivalent regression functions and misclassification rates at the point z_a must have misclassification rates $\left(\tilde{\eta}_0(z_a), \tilde{\eta}_1(z_a)\right) = (1-\eta_1(z_a), 1-\eta_0(z_a))$. These probabilities, however, are ruled out by Assumption 2 (since then $\tilde{\eta}_0(z_a) + \tilde{\eta}_1(z_a) > 1$) and therefore the misclassification rates are identified.

In fact, Apppendix A.1 shows that we can directly identify the misclassification rates as a function of the observed moments of $w = (y, x, xy)$

$$\eta_1(z_a) = (1 - h_1(\mathbb{E}(w|z_a, v_1), \mathbb{E}(w|z_a, v_2)) + h_0(\mathbb{E}(w|z_a, v_1), \mathbb{E}(w|z_a, v_2)))\, 2^{-1} \tag{8}$$

$$\eta_0(z_a) = (1 + h_1(\mathbb{E}(w|z_a, v_1), \mathbb{E}(w|z_a, v_2)) + h_0(\mathbb{E}(w|z_a, v_1), \mathbb{E}(w|z_a, v_2)))\, 2^{-1} \tag{9}$$

for known smooth functions $h_1(\cdot)$ and $h_0(\cdot)$ whose precise form is given in (41) and (40) respectively.

The misclassification rates identified above then are used to identify $\eta_2^*(z, v)$ using (6), which in turn are then used to solve for $g(x^*, z)$ using (7). This yields (for a smooth well defined known function $q(\cdot)$ whose precise form is given by (46) (see Appendix A.4)

$$\begin{aligned} g(1, z_a) &= \frac{1}{2} q(\mathbb{E}(w|z_a, v_1), \mathbb{E}(w|z_a, v_2)) \\ &\quad + \frac{1}{2} h_0(\mathbb{E}(w|z_a, v_1), \mathbb{E}(w|z_a, v_2)) \frac{\mathbb{E}(y|z_a, v_1) - \mathbb{E}(y|z_a, v_2)}{\mathbb{E}(x|z_a, v_1) - \mathbb{E}(x|z_a, v_2)} \end{aligned}$$

$$\begin{aligned} g(0, z_a) &= \frac{1}{2} q(\mathbb{E}(w|z_a, v_1), \mathbb{E}(w|z_a, v_2)) \\ &\quad - \frac{1}{2} h_0(\mathbb{E}(w|z_a, v_1), \mathbb{E}(w|z_a, v_2)) \frac{\mathbb{E}(y|z_a, v_1) - \mathbb{E}(y|z_a, v_2)}{\mathbb{E}(x|z_a, v_1) - \mathbb{E}(x|z_a, v_2)} \end{aligned}$$

The formula for the marginal effect implied by the two equations above is instructive and provides a connection to the literature on estimation of endogenous regression models:

$$g(1, z) - g(0, z) = \frac{\mathbb{E}(y|z_a, v_1) - \mathbb{E}(y|z_a, v_2)}{\mathbb{E}(x|z_a, v_1) - \mathbb{E}(x|z_a, v_2)} h_0(\mathbb{E}(w|z_a, v_1), \mathbb{E}(w|z_a, v_2))$$

The first term on the right hand side is akin to the Wald estimator of the marginal effect of

[11] I thank a refree for pointing this out (see Appendix A.1 for details).

x on y using v as an instrument. The second term can be thought of as a correction term to account for the fact that in the presence of binary misclassification, the Wald-IV estimator does not identify the marginal effect.[12]

The form of the marginal effect also suggests that the model may be extended to allow for endogeneity of the true (unobserved) x^* in a regression model with additively separable errors. In particular, we can retain identification for the function $g^*(x^*, z)$ when

$$y = g^*(x^*, z) + \varepsilon \tag{10}$$

where $\mathbb{E}(\varepsilon|z, v) = 0$ which is the usual IV assumption so that x^* is endogenous as well as misclassified. To account for the mismeasurement, we need to impose the analogue of (I) which is

$$\mathbb{E}(y|x^*, z, x, v) = \mathbb{E}(y|x^*, z) \tag{11}$$

Inspection of the proofs of Theorem 1 reveals that the function $g^*(x^*, z_a)$ is identified even in this model (a formal argument is provided at the end of the proof Theorem 1 in Appendix A.2) without modification. Finally, under the modifications outlined in Theorem 1, the result can be extended to yield identification of the entire regression function $g^*(x^*, \cdot)$. These results are related to those in Das (2004) and the main difference is that for the model under consideration here the endogenous regressor is unobserved but restricted to be binary.

3.1 Semiparametric Models

The identification of semi-parametric regression models, such as single-index models follows in a straightforward manner from the identification result above. In this sub-section we provide identification results in the presence of misclassified regressors for the parameter θ_0 in models of the form

$$\mathbb{E}[y - g(\tau(\tilde{z}; \theta_0))|\tilde{z}] = 0 \tag{12}$$

where $g(\cdot)$ and the parameter vector θ_0 are unknown and $\tau(\cdot)$ is a known index function. The assumption of non-differential measurement error and the exclusion restriction that motivated the additional conditional moment restrictions in (1) can now be written as

$$\mathbb{E}[y - g(\tau(\tilde{z}; \theta_0))|\tilde{z}, x, v] = 0$$

The following modifications are sufficient for identification of the parameter θ_0.

[12] In fact, the marginal effect above is the nonparametric analogue to the formula in the linear regression case stated in Kane, Rouse, and Staiger (1999).

ASSUMPTION 6 *The parameter θ_0 and the function $g(\cdot)$ in the model (12) are identified in the model given knowledge of the population distribution of $\{y, x^*, z\}$.*[13]

ASSUMPTION 7 $\tau(1, z, \theta_0) \neq \tau(0, z, \theta_0)$ *a.e.* $\mathbb{P}_z$

Lemma 1 *If Assumptions 2-4,6 and 7 are modified to hold for almost all $z \in \mathbb{S}_z$ by appending "a.e. $\mathbb{P}_z$" to the displays (2) and (5), then the parameter vector θ_0 and the misclassification rates $\eta_0(\cdot)$ and $\eta_1(\cdot)$ in the model (12) are identified.*

3.2 Parametric Models

Identification for parametric models also follows from the identification result above. An important special case (see Zinman (2004) for an application of the model discussed here) is that of the parametric binary choice model. The appropriate modifications to the identification result yield the identification of the binary choice coefficients as well as the misclassification rates. In this context, the model is given by

$$\mathbb{P}(y = 1|x^*, z, x, v) = \mathcal{F}(\theta_{10} + \theta_{20}x^* + \theta_{30}z) \tag{13}$$

for some known strictly increasing function $\mathcal{F}(\cdot)$. If assumptions 2, 4 are modified to hold almost everywhere $\mathbb{P}_z$ and we modify assumptions 1 and 5 to

ASSUMPTION 8 $\mathbb{E}([1, x^*, z]'[1, x^*, z]) > 0$ *and*

ASSUMPTION 9 $\theta_{20} \neq 0$

then the parameter vector θ_0 and the misclassification rates are identified using similar arguments as in the main identification theorem. Assumption 8 is the usual identification assumption in a binary choice model while Assumption 9 ensures that x^* is not redundant in (13). Similar approaches can be applied to other conditional maximum likelihood models.

4 Estimation

We next discuss estimation of the regression function $g(x^*, z_a)$ non-parametrically and also outline semi-parametric estimation procedures for some of the other models discussed above.

[13] Sufficient conditions are standard in the literature and detailed in Ichimura (1993) and collected for instance in Theorem 2.1 of Horowitz (1998). The assumptions require (roughly) non-constant, non-periodic and differentiable index function $g(\cdot)$, continuously distributed z with $\mathbb{E}(zz') > 0$, parameter vector normalizations and conditions on the support of the index.

Detailed estimation strategies for the semi-parametric (single-index) and parametric (maximum likelihood binary choice) models are discussed in Mahajan (2004).

The estimation strategy follows directly from the identification argument. In particular, Theorem 1 showed that the regression function $(g(0,z), g(1,z))$ is a smooth function of the observed coditional moments $\mathbb{E}(w|z_a, v_a)$ where $w = (x, y, xy)$. We estimate these conditional expectations non-parametrically and apply the continuous mapping theorem to study the consistency and asymptotic distribution of the proposed estimator. It is convenient to introduce some notation to describe the kernel smooths that will be useful in the sequel. Let $f(z|v_a)$ denote the density of z conditional on $v = v_a$ and let $\gamma_a^1(z) \equiv f(z|v_a)\mathbb{P}(v = v_a)$. For $a \in \{1, 2\}$ define

$$\gamma_a^x(z) \equiv \mathbb{P}(v = v_a) f(z|v_a) \mathbb{E}[x|z, v_a] \tag{14}$$

$$\gamma_a^y(z) \equiv \mathbb{P}(v = v_a) f(z|v_a) \mathbb{E}[y|z, v_a] \tag{15}$$

$$\gamma_a^{xy}(z) \equiv \mathbb{P}(v = v_a) f(z|v_a) \mathbb{E}[xy|z, v_a] \tag{16}$$

and define their empirical counterpart (for $s \in \{1, x, y, xy\}$)

$$\hat{\gamma}_a^s(z_a) = \frac{1}{nh^{d_z}} \sum_{i=1}^n s_i K\left(\frac{z_i - z_a}{h_n}\right) \mathbb{I}\{v_i = v_a\}$$

for a kernel function $K(\cdot)$ and a choice of bandwidth sequence $\{h_n\}$ with properties specified below. The estimator for the conditional expectations of interest is then

$$\begin{aligned}
\widehat{\mathbb{E}}(w|z_a, v_a) &= \frac{1}{\hat{\gamma}_a^1(z_a)} \begin{bmatrix} \hat{\gamma}_a^x(z_a) \\ \hat{\gamma}_a^y(z_a) \\ \hat{\gamma}_a^{xy}(z_a) \end{bmatrix} \\
&= \frac{1}{\sum_{j=1}^n K\left(\frac{z_j - z_a}{h_n}\right) \mathbb{I}\{v_j = v_a\}} \sum_{i=1}^n \begin{bmatrix} x_i \\ y_i \\ x_i y_i \end{bmatrix} K\left(\frac{z_i - z_a}{h_n}\right) \mathbb{I}\{v_i = v_a\}
\end{aligned}$$

The regression function can then be estimated as

$$\hat{g}(1, z_a) = \frac{1}{2} q\left(\hat{\mathbb{E}}(w|z_a, v_1), \hat{\mathbb{E}}(w|z_a, v_2)\right) + \tag{17}$$

$$\frac{1}{2} h_0\left(\hat{\mathbb{E}}(w|z_a, v_1), \hat{\mathbb{E}}(w|z_a, v_2)\right) \frac{\hat{\mathbb{E}}(y|z_a, v_1) - \hat{\mathbb{E}}(y|z_a, v_2)}{\hat{\mathbb{E}}(x|z_a, v_1) - \hat{\mathbb{E}}(x|z_a, v_2)} \tag{18}$$

$$\hat{g}(0,z_a) = \frac{1}{2}q\left(\hat{\mathbb{E}}(w|z_a,v_1),\hat{\mathbb{E}}(w|z_a,v_2)\right) - \tag{19}$$

$$\frac{1}{2}h_0\left(\hat{\mathbb{E}}(w|z_a,v_1),\hat{\mathbb{E}}(w|z_a,v_2)\right)\frac{\hat{\mathbb{E}}(y|z_a,v_1)-\hat{\mathbb{E}}(y|z_a,v_2)}{\hat{\mathbb{E}}(x|z_a,v_1)-\hat{\mathbb{E}}(x|z_a,v_2)} \tag{20}$$

and the marginal effect

$$\hat{g}(1,z)-\hat{g}(0,z) = \frac{\hat{\mathbb{E}}(y|z_a,v_1)-\hat{\mathbb{E}}(y|z_a,v_2)}{\hat{\mathbb{E}}(x|z_a,v_1)-\hat{\mathbb{E}}(x|z_a,v_2)}h_0\left(\hat{\mathbb{E}}(w|z_a,v_1),\hat{\mathbb{E}}(w|z_a,v_2)\right)$$

The misclassification rates are similarly estimable by substituting the estimator for $\mathbb{E}(w|z_a,v_a)$ into the formulae in (9) and (8).

Consistency of the estmators $\hat{g}(1,z_a)$ and $\hat{g}(0,z_a)$ follows from the continuous mapping theorem and the consistency of the estimator $\hat{\mathbb{E}}(w|z_a,v_a)$ which can be shown using results familiar from the non-parametric estimation literature. Similarly, the large sample distribution of the proposed estimator follows from a standard asymptotic normality result for $\hat{\mathbb{E}}(w|z_a,v_a)$ and the "delta" method. We state a set of sufficient conditions for the asymptotic normality result as detailed in Bierens (1987) for the case of regression estimation with discrete and continuous regressors. We assume, as before, that the IVL v only takes on 2 values. Let $\sigma_k^s(z_a,v_a) = \mathbb{E}(w_k - \mathbb{E}(w_k|z_a,v_a))^s$ and $f(z|v_a)$ denote the density of z conditional on $v = v_a$

For each of $k = 1,2,3$ and both values of v, the following conditions hold

ASSUMPTION 10 *There exists a $\delta > 0$ such that $\sigma_k^{2+\delta}(z,v_a)f(z|v_a)$ is continuous and uniformly bounded in z*

ASSUMPTION 11 *The functions $(\mathbb{E}(w_k|z,v_a))^2 f(z|v_a)$,$\sigma_k^2(z,v_a)f(z|v_a)$ are continuous and uniformly bounded in z.*

ASSUMPTION 12 *The functions $f(z|v_a),(\mathbb{E}(w_k|z,v_a)f(z|v_a))$ have first and $m \geq 2$ partial derivatives that are continuous and uniformly bounded in z*

ASSUMPTION 13 *The Kernel function $K:\mathbb{R}^{d_z}\to\mathbb{R}$ satisfies for some $m \geq 2$*

$$\int (z_1)^{i_1}...(z_{d_z})^{i_{d_z}}K(z)\,dz = \begin{cases} 1 & if \quad i_1 = .. = i_{d_z} = 0 \\ 0 & if \quad 0 < i_1 + ... + i_{d_z} < m \end{cases}$$

for non-negative integers $(i_1,...,i_{d_z})$ and $\int \|z\|^i K(z)\,dz < \infty$ for $i = m$.

ASSUMPTION 14 *$h_n \to 0$, $nh_n^{d_z} \to \infty$ and $\sqrt{nh_n^{d_z}}h^m \to 0$*

These are standard conditions for deriving large sample properties of kernel estimators and are for instance expounded in Bierens (1987). Under these conditions

Lemma 2 *Suppose that 10-14 hold. Then for each* $v_a \in \{v_1, v_2\}$

$$\sqrt{nh^{d_z}}\left(\widehat{\mathbb{E}}\left(w|z_a, v_a\right) - \mathbb{E}\left(w|z_a, v_a\right)\right) \Rightarrow \mathcal{N}\left(0, V_a\right)$$

where

$$V_a = \left[\begin{array}{ccc} Var\left(x|z_a, v_a\right) & Cov\left(x, y|z_a, v_a\right) & Cov\left(x, xy|z_a, v_a\right) \\ Cov\left(x, y|z_a, v_a\right) & Var\left(y|z_a, v_a\right) & Cov\left(y, xy|z_a, v_a\right) \\ Cov\left(x, xy|z, v\right) & Cov\left(y, xy|z_a, v_a\right) & Var\left(xy|z_a, v_a\right) \end{array}\right] \frac{\int K\left(z\right)^2 dz}{f\left(z|v_a\right)\mathbb{P}\left(v = v_a\right)}$$

and the vectors

$$\sqrt{nh^{d_z}}\left(\widehat{\mathbb{E}}\left(w|z_a, v_1\right) - \mathbb{E}\left(w|z_a, v_1\right)\right)$$

and

$$\sqrt{nh^{d_z}}\left(\widehat{\mathbb{E}}\left(w|z_a, v_2\right) - \mathbb{E}\left(w|z_a, v_2\right)\right)$$

are asymptotically independent.

The result follows directly from Theorem 3.2.1 in Bierens (1987) and the Cramer-Wold device. We can then apply the "delta" method to conclude

Lemma 3 *Suppose that Assumptions 1-5 and 10-14 hold. Then, the estimators* $\hat{g}\left(1, z\right)$ *and* $\hat{g}\left(0, z\right)$ *defined in (17) and (19) above converge weakly as follows:*

$$\sqrt{nh^{d_z}}\left(\hat{g}\left(1, z_a\right) - g\left(1, z_a\right)\right) \Rightarrow \mathcal{N}\left(0, \Omega_1\right)$$

$$\sqrt{nh^{d_z}}\left(\hat{g}\left(0, z_a\right) - g\left(0, z_a\right)\right) \Rightarrow \mathcal{N}\left(0, \Omega_0\right)$$

and the marginal effect

$$\sqrt{nh^{d_z}}\left(\left(\hat{g}\left(1, z_a\right) - \hat{g}\left(0, z_a\right)\right) - \left(g\left(1, z_a\right) - g\left(0, z_a\right)\right)\right) \Rightarrow \mathcal{N}\left(0, \Omega_M\right)$$

where

$$\Omega_1 = \frac{f_1\left(\mathbb{E}\left(w|z_a, v_1\right), \mathbb{E}\left(w|z_a, v_2\right), V\left(w|z_a, v_1\right), V\left(w|z_a, v_2\right)\right)}{2\left[\left(\mathbb{E}\left(y|z_a, v_1\right) - \mathbb{E}\left(y|z_a, v_2\right)\right)\left(\mathbb{E}\left(x|z_a, v_1\right) - \mathbb{E}\left(x|z_a, v_2\right)\right)\right]^2\left(1 - \eta_0\left(z_a\right) - \eta_1\left(z_a\right)\right)}$$

$$\Omega_0 = \frac{f_0\left(\mathbb{E}\left(w|z_a, v_1\right), \mathbb{E}\left(w|z_a, v_2\right), V\left(w|z_a, v_1\right), V\left(w|z_a, v_2\right)\right)}{2\left[\left(\mathbb{E}\left(y|z_a, v_1\right) - \mathbb{E}\left(y|z_a, v_2\right)\right)\left(\mathbb{E}\left(x|z_a, v_1\right) - \mathbb{E}\left(x|z_a, v_2\right)\right)\right]^2\left(1 - \eta_0\left(z_a\right) - \eta_1\left(z_a\right)\right)}$$

$$\Omega_M = \frac{f_M\left(\mathbb{E}\left(w|z_a,v_1\right),\mathbb{E}\left(w|z_a,v_2\right),V\left(w|z_a,v_1\right),V\left(w|z_a,v_2\right)\right)}{2\left[\left(\mathbb{E}\left(y|z_a,v_1\right)-\mathbb{E}\left(y|z_a,v_2\right)\right)\left(\mathbb{E}\left(x|z_a,v_1\right)-\mathbb{E}\left(x|z_a,v_2\right)\right)\right]^2\left(1-\eta_0\left(z_a\right)-\eta_1\left(z_a\right)\right)}$$

for positive functions $f_1(\cdot), f_0(\cdot)$ *and* $f_M(\cdot)$ *and where* $V\left(w|z_a,v_k\right)$ *denotes the conditional variance-covariance matrix of the vector* (x,y,xy).

The proof follows from an application of the "Delta" method, as expounded for instance in van der Vaart (1998). The denominator terms in the asymptotic variances are of some independent interest because they highlight the relationship between the identifying assumptions (which ensure that none of the three terms in the denominator is zero) and the weak convergence result. The variances can be estimated consistently for instance by the bootstrap (see e.g. Jones and Wand (1995) or Shao and Tu (1996)) .

We now briefly discuss estimation of the average marginal effect. Recall that the marginal effect (conditional on z) is consistently estimable by

$$\hat{g}(1,z)-\hat{g}(0,z)=\frac{\hat{\mathbb{E}}\left(y|z,v_1\right)-\hat{\mathbb{E}}\left(y|z,v_2\right)}{\hat{\mathbb{E}}\left(x|z,v_1\right)-\hat{\mathbb{E}}\left(x|z,v_2\right)}h_0\left(\hat{\mathbb{E}}\left(w|z.v_1\right),\hat{\mathbb{E}}\left(w|z.v_2\right)\right)$$

Here we outline the estimation of the marginal effect averaged over (a fixed subset of) the support of z.[14] The object of interest for this sub-section is the averaged marginal effect

$$\beta_m=\mathbb{E}\left(l(z)\left(g(1,z)-g(0,z)\right)\right)$$

where $l(z)$ is a fixed trimming function. A consistent estimator of this parameter is

$$\hat{\beta}_m=\frac{1}{n}\sum_{i=1}^{n}l(z_i)\left(\hat{g}(1,z_i)-\hat{g}(0,z_i)\right) \tag{21}$$

and asymptotic normality at the$\sqrt{n}$ rate can be obtained by verifying conditions for Theorem 8.11 of Newey and McFadden (1994) and we omit details here.

5 Semiparametric Estimation

We briefly discuss estimation strategies for the single-index model (12) in the case where $\tau\left(\tilde{z},\theta_0\right)=\tilde{z}'\theta_0=x^*\theta_{10}+z'\theta_{z0}$ (details can be found in Mahajan (2004)). Estimation of the single index model in this context can in principle be carried out in different ways. One is to use series approximations for each of the functions $\left(g(\cdot),\eta_0(\cdot),\eta_1(\cdot),\eta_2(\cdot)\right)$ and carry

[14]The choice of a variable trimming device (not explored here) would enable consistent estimation of the marginal effect averaged over the entire support of z.

out sieve estimation, thereby generalizing the approach in Section 5.1 below (indeed this approach could also be applied to the case where the index function is not linear as for instance in Newey and Stoker (1993)[15]). An alternative approach, pursued here, is to apply a single index estimation scheme along the lines of Powell, Stock, and Stoker (1989).

The motivation for the estimation scheme follows directly from the identification argument and enables us to identify θ up to scale for those elements of z that are continuously distributed. Suppose for simplicity that all elements of the random vector z are jointly distributed continuously over a compact set $S_z \subseteq \mathbb{R}^{d_z}$, then the subvector θ_{z0} corresponding to the covariate vector z is identified up to scale by the estimation procedure outlined below.

As previously, the estimator is motivated by the identification strategy by examining the observed moments $\mathbb{E}[w|z,v]$ and relating them to the parameter vector θ_{z0}. In particular we show in Appendix B that the directly identified function $q(\cdot)$ (defined in (46)) satisfies

$$q(z) = G(z'\theta_{z0}) \tag{22}$$

so that the model (22) is of the single index form with (unknown) differentiable link function $G : \mathbb{R} \to \mathbb{R}$.

The parameter θ_{z0} can now be estimated up to scale using average weighted derivative estimation techniques. In particular, if the k^{th} variable z_k is continuously distributed and the function $G(\cdot)$[16]

$$\nabla_{z_k} q(z) = \nabla G(z'\theta_{z0})\, \theta_{z0_{[k]}}$$

where $\theta_{z0_{[k]}}$ denotes the coefficient on z_k. Further, for any (z measurable) weight function $l(z)$ we can compute the average weighted derivative

$$\mathbb{E}[l(z)\nabla_{z_k} q(z)] = \mathbb{E}[l(z)\nabla\widetilde{g}(z'\theta_{z0})]\,\theta_{z0_{[k]}} = K_l\theta_{z0_{[k]}}$$

where K_l is a constant that depends upon the weighting function.

Using the notation introduced in Appendix A we can write $q(\cdot)$ as a function of the functions $\gamma(\cdot)$

$$q(z,\gamma) \equiv \frac{q_u(z,\gamma)}{q_l(z,\gamma)} = \frac{[\gamma_2^y(z)\gamma_1^x(z) - \gamma_2^x(z)\gamma_1^y(z) + \gamma_1^{xy}(z)\gamma_2^1(z) - \gamma_1^{xy}(z)\gamma_2^1(z)]}{(\gamma_1^x(z)\gamma_2^1(z) - \gamma_1^1(z)\gamma_2^x(z))}$$

With the choice $l(z) = (q_l(z,\gamma_0))^2$ we eliminate the 'random denominator' problem from the estimation which obviates the use of a trimming device. The object of interest is then

[15] I thank a referee for pointing this out.

[16] For a function $f(z)$, let $\nabla_{z_k} f(\tilde{z}) = \left.\frac{\partial f}{\partial z_k}\right|_{z=\tilde{z}}$ and $\nabla_z f(\tilde{z})$ denotes the entire vector of partial derivatives.

the average weighted derivative corresponding to this choice of weight function and is given by

$$\beta_0 \equiv \mathbb{E}\left[q_l(z,\gamma)^2 \nabla G(z'\theta_{z0})\right]\theta_{z0}$$

which can be rewritten as

$$\beta_0 = \mathbb{E}\left[q_l(z,\gamma)^2 \nabla_z q(z,\gamma)\right] = \mathbb{E}\left[q_l(z,\gamma)\nabla_z q_u(z,\gamma) - q_u(z,\gamma)\nabla_z q_l(z,\gamma)\right]$$

A natural estimator for β_0 then is the plug-in estimator

$$\hat{\beta} = n^{-1}\sum_{i=1}^{n}\left(q_l(z_i,\hat{\gamma})\nabla_z q_u(z_i,\hat{\gamma}) - q_u(z_i,\hat{\gamma})\nabla_z q_l(z_i,\hat{\gamma})\right)$$

where $\hat{\gamma}$ denotes a nonparametric kernel based estimator of γ. The asymptotic properties of the estimator are discussed in greater detail in Mahajan (2004) where it is shown that under suitable conditions the proposed estimator $\hat{\beta}$ is $\sqrt{n}$ consistent and asymptotically normal.

5.1 Parametric Specifications of $g(x^*, z)$

An important special case of the model occurs when the regression function is parametrically specified. We consider here the case of the binary choice model with misclassification given by (13) and assumptions 2-4, 8 and 9. There are at least two alternative methods for proceeding with estimation in this context. The first method is a minimum distance estimator along the lines of Example 2 Newey (1994a) and the second is a sieve maximum likelihood estimator which in practice reeduces to a parametric estimation procedure and is therefore relatively straightforward to implement. We first outline the sieve estimation technique.

Using (13) and Assumptions 3 the conditional probability $\mathbb{P}(y,x|z,v)$ can be written as

$$\mathbb{P}(y,x|z,v) = \sum_{x^*\in\{0,1\}}\mathbb{P}(y|x^*,z)\,\mathbb{P}(x|x^*,z)\,\mathbb{P}(x^*|z,v) \tag{23}$$

where

$$\mathbb{P}(y|x^*,z) = \mathcal{F}(\beta_1+\beta_2 x^*+\beta_3 z)^y\left(1-\mathcal{F}(\beta_1+\beta_2 x^*+\beta_3 z)\right)^{1-y}$$

As before, we place no functional form assumptions on the misclassification probabilities $\mathbb{P}(x|x^*,z)$ and the probability $\mathbb{P}(x^*|z,v)$. To ensure that the approximations to these probabilities lie between zero and one we use the log-odds ratio parameterization of the probabilities in the likelihood (23). Let $\lambda_k(\cdot)$ for $k=1,2,3$ denote the log-odds ratio for the

probabilities $\mathbb{P}(x|x^*=0,z)$, $\mathbb{P}(x|x^*=1,z)$ and $\mathbb{P}(x^*|z,v)$ respectively. Defining the logit transformation $L(x)=\frac{\exp(x)}{1+\exp(x)}$ we can write the probabilities

$$\mathbb{P}(x|x^*=0,z)=(L(\lambda_1(z)))^x(1-L(\lambda_1(z)))^{1-x}$$

$$\mathbb{P}(x|x^*=1,z)=(L(\lambda_2(z)))^{1-x}(1-L(\lambda_2(z)))^x$$

$$\mathbb{P}(x^*|z,v)=L(\lambda_3(z,v))$$

Using the notation above, the likelihood (23) is a function of the parameters $\alpha=(\beta,\lambda)=(\beta,\lambda_1(\cdot),\lambda_2(\cdot),\lambda_3(\cdot))$ and we denote it by $\mathbb{P}(y,x|z,v;\alpha)$. The parameter α belongs to the space $\mathcal{A}=\mathbf{B}\times\mathbf{\Lambda}$ where $\mathbf{B}$ is a compact subset of $\mathbb{R}^{d_z+2}$. The space $\mathbf{\Lambda}=\Pi_{k=1}^3\Lambda_k$ where Λ_1 and Λ_2 are sets of functions defined over the support of z and Λ_3 is a set of functions defined over the support of (z,v) and (under Assumption 2) satisfy $L(\lambda_1(z))+L(\lambda_2(z))<1$ for any $\lambda_1\in\Lambda_1$ and $\lambda_2\in\Lambda_2$.

To define the sieve approximation to the spaces Λ_j $j=1,2,3$, let $\{r_{j,m}\}_{m=1}^{\infty}$ represent a set of basis functions (such as power, Fourier series or splines) for Λ_j. Let $R_{j,k_n}(x)=(r_1(x),...,r_{k_n}(x))$ denote a $k_n\times 1$ vector of basis functions and $\Pi_{j,n}$ a conformable vector of constants. Then we define $\Lambda_{j,n}=\{R_{j,k_n}\Pi_{j,n}:|\Pi_{j,n}|\leq c_n\}$ for an arbitrary sequence $c_n\to\infty$[17] and the sieve space $\mathcal{A}_n=\mathbf{B}\times\Lambda_n=\mathbf{B}\times\Lambda_{1,n}\times\Lambda_{2,n}\times\Lambda_{3,n}$. For more details on the sequence of basis functions and the construction of the sieve see Mahajan (2004) and Ai and Chen (2001).

Estimation of α is carried out in a semi-parametric maximum likelihood setting using the method of sieves to estimate the infinite dimensional parameters. In the special case where we are willing to assume that the misclassification rates are independent of the other explanatory variables or when the support of the explanatory variables is finite, estimation reduces to standard maximum likelihood.

The log likelihood is given by

$$\begin{aligned} l(w,\alpha) &= \ln\mathbb{P}(y,x|z,v;\alpha) \\ &= \ln\{(1-\mathcal{F}(b_1+b_3z))^{1-y}\mathcal{F}(b_1+b_3z)^y\lambda_1(z)^x(1-\lambda_1(z))^{1-x}(1-\lambda_3(z,v)) \\ &\quad +(1-\mathcal{F}(b_1+b_2+b_3z))^{1-y}\mathcal{F}(b_1+b_2+b_3z)^y\lambda_2(z)^{1-x}(1-\lambda_2(z))^x\lambda_3(z,v)\} \end{aligned}$$

where $w=(y,x,z,v)$ and the parameter $\alpha\in\mathcal{A}=\mathbf{B}\times\mathbf{\Lambda}$. We observe a random sample on

[17]The bound c_n is not typically imposed in estimation but is required to ensure that the sieve space Λ_n is compact for each n as required in the consistency argument (though one could also consider adopting a penalized likelihood approach).

w and define the Sieve Maximum Likelihood estimator as

$$\hat{\alpha} \equiv \left(\hat{b}, \hat{\lambda}\right) = \arg\max_{\mathcal{A}_n} \frac{1}{n} \sum_{i=1}^{n} l\left(w_i, \alpha\right) \tag{24}$$

so $\hat{\alpha}_n$ maximizes the sample log likelihood over the finite dimensional 'sieve' space $\mathcal{A}_n$ and the optimization can be carried out using standard software packages. Mahajan (2004) discusses the consistency, rate of convergence and asymptotic normality for the proposed estimator $\hat{b}$ and shows that it attains the semiparametric efficiency bound.

An alternative estimation procedure is to construct an estimator along the lines of Example 2 in Newey (1994a) noting that the equation

$$\mathcal{F}^{-1}\left(g\left(0, z\right)\right) = \beta_1 + \beta_3 z$$

can be used to construct an estimator for (β_1, β_3) by choosing the parameters to minimize the distance between the left and the right hand side. Therefore, a simple alternative estimator for (β_1, β_2) is given by a least squares regression of $\mathcal{F}^{-1}\left(\hat{g}\left(0, z_i\right)\right)$ on z_i:

$$\left(\hat{\beta}_1, \hat{\beta}_3\right) = \arg\min_{(b_1, b_3)} \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_i \left(\mathcal{F}^{-1}\left(\hat{g}\left(0, z_i\right)\right) - b_1 - b_3 z_i\right)^2$$

where $\mathbb{I}_i$ is a fixed trimming set and $\hat{g}\left(0, z_i\right)$ is the non-parametric estimator defined in (19). β_2 can be estimated using $\mathcal{F}^{-1}\left(\hat{g}\left(1, z_i\right)\right)$ and the estimates $\left(\hat{\beta}_1, \hat{\beta}_3\right)$ obtained above.

6 Testing for Misclassification

A natural question is whether one can test for misclassification since in its absence, simpler estimation procedures can be implemented. While a comprehensive discussion on hypothesis testing in such models is not carried out in this paper, a simple exclusion restriction can be derived to test for misclassification and is the subject of this section. In particular, for the model (1) and under the identification assumptions discussed above, the expectation of the outcome conditional on (x, z, v) does not depend upon the ILV v if and only if there is no misclassification. We record this result and then use it to propose a simple test for the absence of misclassification.

Lemma 4 *Consider the model (1) under assumptions (1)-(5) and assume* $\mathbb{P}\left(x^* = 1 | z_a, v\right) \in$

$(0,1)$. *Then,*

$$\eta_0(z_a) = \eta_1(z_a) = 0 \text{ if and only if } \mathbb{E}(y|x, z_a, v) = \mathbb{E}(y|x, z_a) \text{ a.e.}$$

Proof. The proof of the result is straightforward and depends upon examining the form of the probability $\mathbb{P}(x^*|x, z, v)$ implied by assumptions (1)-(5).

$$\begin{aligned}
\mathbb{E}(y|x, z_a, v) &= \sum_{s\in\{0,1\}} g(s, z_a)\,\mathbb{P}(x^* = s|x, z_a, v) \\
&= \sum_{s\in\{0,1\}} g(s, z_a)\,\frac{\mathbb{P}(x|x^* = s, z_a)\,\mathbb{P}(x^* = s|z_a, v)}{\mathbb{P}(x|z_a, v)} \\
&= \{g(0, z_a)(\eta_0(z_a)\,x + (1-\eta_0(z_a))(1-x))\,\mathbb{P}(x^* = 0|z_a, v) + \\
&\quad g(1, z_a)(\eta_1(z_a)(1-x) + (1-\eta_1(z_a))\,x)\,\mathbb{P}(x^* = 1|z_a, v)\}\mathbb{P}(x|z_a, v)^{-1}
\end{aligned}$$

where the second equality follows from Bayes Rule and Assumption (3). We first show the $\Rightarrow$ part. Suppose that $\eta_s(z_a) = 0$ for $s \in \{0,1\}$, then $\mathbb{P}(x^* = 1|z_a, v) = \mathbb{P}(x = 1|z_a, v)$ and the conditional expectation reduces to

$$\begin{aligned}
\mathbb{E}(y|x, z_a, v) &= \{g(0, z_a)(1-x)\,\mathbb{P}(x = 1|z_a, v) + g(1, z_a)\,x\mathbb{P}(x = 1|z_a, v)\}\,\mathbb{P}(x|z_a, v)^{-1} \\
&= g(x, z_a)
\end{aligned}$$

To prove in the other direction, note that the equality of conditional expectations implies

$$\mathbb{E}(y|x, z_a, v_1) = \mathbb{E}(y|x, z_a, v_2)$$

which in turn implies after some algebra that

$$(g(1, z_a) - g(0, z_a))(\mathbb{P}(x^* = 1|x, z_a, v_1) - \mathbb{P}(x^* = 1|x, z_a, v_2)) = 0$$

Because of Assumption (5) the above implies that

$$\mathbb{P}(x^* = 1|x, z_a, v_1) - \mathbb{P}(x^* = 1|x, z_a, v_2) = 0$$

which can be rewriten as

$$(\eta_1(z_a)(1-x) + (1-\eta_1(z_a))\,x)\left\{\frac{\mathbb{P}(x^* = 1|z_a, v_1)}{\mathbb{P}(x|z_a, v_1)} - \frac{\mathbb{P}(x^* = 1|z_a, v_2)}{\mathbb{P}(x|z_a, v_2)}\right\} = 0$$

When $x = 1$ the expression above reduces to

$$(1 - \eta_1(z_a)) \left\{ \frac{\mathbb{P}(x^* = 1|z_a, v_1)}{\mathbb{P}(x = 1|z_a, v_1)} - \frac{\mathbb{P}(x^* = 1|z_a, v_2)}{\mathbb{P}(x = 1|z_a, v_2)} \right\} = 0$$

and by Assumption (2) the above implies that

$$\frac{\mathbb{P}(x^* = 1|z_a, v_1)}{\mathbb{P}(x = 1|z_a, v_1)} - \frac{\mathbb{P}(x^* = 1|z_a, v_2)}{\mathbb{P}(x = 1|z_a, v_2)} = 0$$

and using the fact that

$$\mathbb{P}(x = 1|z_a, v_2) = \eta_0(z_a) + (1 - \eta_0(z_a) - \eta_1(z_a))$$

we obtain

$$\eta_0(z_a) \left(\frac{\mathbb{P}(x^* = 1|z_a, v_1) - \mathbb{P}(x^* = 1|z_a, v_2)}{\mathbb{P}(x^* = 1|z_a, v_1)\mathbb{P}(x^* = 1|z_a, v_2)} \right) = 0$$

and by Assumption (4) this must imply that $\eta_0(z_a) = 0$. A similar argument for the case where $x = 1$ leads to the conclusion that $\eta_1(z_a) = 0$ so that the equality of conditional expectations is also sufficient for the absence of measurement error. ■

Therefore, comparing the conditional expectations can serve as the basis for a test of misclassification under the maintained assumptions 1-5 for (1). Note that the result depends upon the maintained assumptions (1)-(5) so that the a rejection of the null hypothesis may also be interpreted as evidence against the identifying assumptions (1)-(5).

The result implies that the statistic

$$T(x, z_a, v) = \widehat{\mathbb{E}}(y|x, z_a, v) - \widehat{\mathbb{E}}(y|x, z_a)$$

can be used to form a test of misclassification. This has the appealing feature that estimation of the misclassification rates themselves is not required. A discussion of the power properties of such tests in general is, however, not carried out here. We first record a standard set of regularity conditions to study the limiting distribution of the statistic T. Let r denote a discrete random variable (it will be either the variable x or the vector (v, x) in what follows). Let $\sigma^s(z_a, r_a) = \mathbb{E}(y - \mathbb{E}(y|z_a, r_a))^s$ and $f(z|r_a)$ denote the density of z conditional on $r = r_a$. For each possible value of r the following conditions hold

ASSUMPTION 15 *There exists a $\delta > 0$ such that $\sigma^{2+\delta}(z, r_a) f(z|r_a)$ is continuous and uniformly bounded in z*

ASSUMPTION 16 *The functions $(\mathbb{E}(y|z, r_a))^2 f(z|r_a)$,$\sigma^2(z, r_a) f(z|r_a)$ are continuous and uniformly bounded in z.*

ASSUMPTION 17 *The functions* $f(z|r_a), (\mathbb{E}(y|z,r_a) f(z|r_a))$ *have first and* $m \geq 2$ *partial derivatives that are continuous and uniformly bounded in* z

ASSUMPTION 18 *The Kernel function* $K : \mathbb{R}^{d_z} \to \mathbb{R}$ *satisfies for some* $m \geq 2$

$$\int (z_1)^{i_1} ... (z_{d_z})^{i_{d_z}} K(z) dz = \begin{cases} 1 & if \quad i_1 = .. = i_{d_z} = 0 \\ 0 & if \quad 0 < i_1 + ... + i_{d_z} < m \end{cases}$$

for non-negative integers $(i_1, ..., i_{d_z})$ *and* $\int \|z\|^i K(z) dz < \infty$ *for* $i = m$.

ASSUMPTION 19 $h_n \to 0$, $nh_n^{d_z} \to \infty$ *and* $\sqrt{nh_n^{d_z}} h_n^m \to 0$

Lemma 5 *Consider the model (1) under assumptions (1)-(5) and suppose* $\mathbb{P}(x^* = 1|z_a, v) \in (0,1)$. *Suppose there is no misclassification so that* $\eta_0(z_a) = \eta_1(z_a) = 0$. *For* $x_a \in \{0,1\}$ *and* $v_a \in \{v_1, v_2\}$ *Define the statistic*

$$T(x_a, z_a, v_a) = \frac{\sum_{i=1}^n y_i K\left(\frac{z_i - z_a}{h_n}\right) \mathbb{I}(x_i = x_a, v_i = v_a)}{\sum_{i=1}^n K\left(\frac{z_i - z_a}{h_n}\right) \mathbb{I}(x_i = x_a, v_i = v_a)} - \frac{\sum_{i=1}^n y_i K\left(\frac{z_i - z_a}{h_n}\right) \mathbb{I}(x_i = x_a)}{\sum_{i=1}^n K\left(\frac{z_i - z_a}{h_n}\right) \mathbb{I}(x_i = x_a)}$$

and suppose that assumptions 15-19 hold with $r = x$ *and* $r = (x, v)$. *Then,*

$$\sqrt{nh^d} T(x_a, z_a, v_a) \Rightarrow \mathcal{N}(0, V_T^a)$$

where

$$\begin{aligned} V_T^a &= (\mathbb{P}(x = x_a) \mathbb{P}(x = x_a, v = v_a) f(z_a|x_a, v_a) f(z_a|x_a))^{-1} \int K(u)^2 du \\ &\quad \left\{ \begin{array}{c} [f(z_a|x_a) \mathbb{P}(x = x_a) - 2\mathbb{P}(x = x_a, v = v_a)] f(z_a|x_a, v_a) Var(y|z_a, x_a, v_a) \\ + [\mathbb{P}(x = x_a, v = v_a) f(z_a|x_a, v_a)] Var(y|z_a, x_a) \end{array} \right\} \end{aligned}$$

The proof follows directly from Theorem 3.2.1 in Bierens (1987). Implementing the test is straightforward and only requires the use of standard kernel regression (with perhaps the use of the bootstrap to calculate standard errors). In the context of other, parametric, models, the proposed test usually reduces to testing for an exclusion restriction. For instance, testing for the absence of misclassification in the binary choice model (13) the result above implies that a direct test for the exclusion of v in the binary choice model can serve as a test for misclassification. This test can be carried out using any of the standard test statistics for testing such hypotheses in a maximum likelihood context and is therefore readily

implementable.[18]

There are, however, other possible tests although they may require placing further structure on the form of the misclassification rates and the test statistics may have more complicated limiting distributions. To take a simple example, consider the case where the misclassification rates are independent of the other covariates and are parameterized as $\eta_s(z) = (\eta_s/(1+\eta_s))$ for $s = 0, 1$. In this case, the hypothesis of no misclassification can be stated as $H : \eta = 0$ against the alternative $K : \eta \geq 0, \eta \neq 0$ for the bivariate vector (η_0, η_1). Further, suppose that the model is parametric (for instance, given by (13)) so that the object of interest is some finite dimensional parameter θ. In this situation, tests based on the Lagrange Multiplier are attractive since estimation under the null hypothesis can be carried out using standard statistical software. However, the limiting distribution of the test statistics under the null hypothesis will be not be asymptotically normal since the true parameter lies on the boundary of the parameter space. In the case of the binary choice model, the Lagrange Multiplier statistic can be calculated as the solution to a quadratic optimization problem

$$\xi = \arg\min_{\lambda \geq 0} \left(\lambda - \hat{\lambda}_H\right)' \hat{Q} \left(\lambda - \hat{\lambda}_H\right)$$

where $\hat{\lambda}_H$ is the Lagrange Multiplier associated with the optimization problem under the null hypothesis and $\hat{Q}$ is an estimated weighting matrix. The limiting distribution of the LM statistic is non-normal, in fact it will be a mixture of chi-squared distributions with weights depending upon the true (unknown) parameter values.

7 Conclusion

The evidence on measurement error in typical data sets suggests that it does not satisfy the independence assumptions usually required by error correction methods for non-linear models. This paper examined the effect of relaxing these assumptions in a general class of models. We studied the problem of parameter identification and estimation in a non-parametric regression model when the data is measured with error and the error is related to the regressors in the model. The paper showed that the regression functions and marginal effect are identified in the presence of an additional random variable that is correlated with the true unobserved underlying variable but unrelated to the measurement error (by analogy with the linear IV case we called this a "Instrument Like" variable although the estimation strategy is *not* a version of 2SLS). Next, the paper proposed various estimation strategies for

[18]Note that such a test will be a test of the null hypothesis that $\eta_0(z) = \eta_1(z) = 0$ almost everywhere rather than just the pointwise assertion of the absence of misclassification.

the parameters of interest in such models. Beginning with a fully non-parametric estimation procedure, the paper further proposed a weighted partial derivative estimation scheme for the case where the regression function satisfies a (linear) single index restriction. Next, a sieve based estimation procedure for the special case of a parametric binary choice model that is straightforward to implement and potentially useful in applications is discussed. [19] Finally, the paper proposed a simple test for the absence of misclassification based on an exclusion restriction.

Mailing Address: Dept. of Economics, Stanford University, 579 Serra Mall, Stanford, CA 94305-6072, U.S.A.; e-mail address: amahajan@stanford.edu

A Appendix

A.1 Proof of Theorem *1:* Identification

Proof of Theorem 1

We first introduce some simplifications and some notation. We suppose that v is a binary variable taking values (v_1, v_2) with $v_1 \neq v_2$. This simplifies the exposition considerably and is sufficient for identification. Increasing the points of support strengthens identification but makes the notation and discussion somewhat more complicated and is therefore not pursued here. To simplify notation further, we suppress the dependence on (z, v) unless needed and equalities involving random variables are assumed to hold almost everywhere.[20] Finally, let the vector w denote the random variables (x, y, xy).

The identification result proceeds in two steps: we first show that if the misclassification rates $\eta(z_a) \equiv (\eta_0(z_a), \eta_1(z_a))$ are identified, then the regression function $g(x^*, z_a)$ is identified. In the second step we show that the misclassification rates are indeed identified.

First Step

The step shows that under Assumptions 1-5, if $\eta(z_a)$ is identified, knowledge of the moments $\mathbb{E}(xy|z_a, v)\,(\equiv r(z_a, v))$, $\mathbb{E}(x|z_a, v)\,(\equiv \eta_2(z_a, v))$ and $\mathbb{E}(y|z_a, v)\ (\equiv t(z_a, v))$ is enough to identify $g(x^*, z_a)$. The proof is done by contradiction. Suppose that $\eta(z_a)$ is identified but the model (1) is not identified, so that $g(x^*, z_a)$ is not uniquely determined by the population distribution of the observed variables.

First, note that since $(\eta_0(\cdot), \eta_1(\cdot))$ are identified by assumption then (by Assumption 3)

[19] See Zinman (2004) for a demonstration of the proposed method in an empirical application.

[20] Similarly for two random variables $X \neq Y$ is taken to mean that $\mathbb{P}(\{\omega : X(\omega) \neq Y(\omega)\}) > 0$

$\eta_2^*(z_a, v) \equiv \mathbb{P}(x^* = 1|z_a, v)$ is also identified since

$$\eta_2^*(z_a, v) = \frac{\eta_2(z_a, v) - \eta_0(z_a)}{1 - \eta_0(z_a) - \eta_1(z_a)} \tag{25}$$

Next, some algebra yields

$$t(z_a, v) \equiv g(0, z_a)(1 - \eta_2^*(z_a, v)) + g(1, z_a)\eta_2^*(z_a, v) \tag{26}$$

Evaluating the equation above at the two values of the ILV v we obtain

$$t(z_a, v_2) \equiv g(0, z_a)(1 - \eta_2^*(z_a, v_2)) + g(1, z_a)\eta_2^*(z_a, v_2)$$

$$t(z_a, v_1) \equiv g(0, z_a)(1 - \eta_2^*(z_a, v_1)) + g(1, z_a)\eta_2^*(z_a, v_1)$$

which forms a linear system of equations in two unknowns $(g(1, z_a), g(0, z_a))$. By assumption (4) this system has a unique solution given by

$$\begin{bmatrix} g(1, z_a) \\ g(0, z_a) \end{bmatrix} = \frac{1}{\eta_2^*(z_a, v_1) - \eta_2^*(z_a, v_2)} \begin{bmatrix} 1 - \eta_2^*(z_a, v_1) & -(1 - \eta_2^*(z_a, v_2)) \\ -\eta_2^*(z_a, v_1) & \eta_2^*(z, v_2) \end{bmatrix} \begin{bmatrix} t(z_a, v_2) \\ t(z_a, v_1) \end{bmatrix}$$

Therefore, we conclude that if the misclassification rates $\eta_0(z_a), \eta_1(z_a)$ are identified, then the regression function $g(x^*, z_a)$ is identified.

If Assumptions 1-5 are strengthened to hold for almost all $z \in \mathbb{S}_z$ (by appending "$a.e.$ $\mathbb{P}_z$" to the displays 2,4 and 5) and we assume that the entire misclassification functions $\eta(\cdot)$ are identified, then an iteration of the argument above yields the conclusion that the entire regression function $g(\cdot)$ is identified.

Second Step

We now demonstrate that the misclassification rates $\eta(z_a)$ are identified. We start by noting two directly identified features of the model.

First, define the directly identified quantity

$$\bar{t}(z_a) \equiv \frac{t(z_a, v_1) - t(z_a, v_2)}{\eta_2(z_a, v_1) - \eta_2(z_a, v_2)} \tag{27}$$

which is akin to the Wald-IV estimator for the marginal effect of x on y using v as an instrument. Using (27) and (25) and letting $\kappa(z_a) \equiv (1 - \eta_0(z_a) - \eta_1(z_a))^{-1}$ it is seen that

$$\bar{t}(z_a) = \kappa(z_a)(g(1, z_a) - g(0, z_a)) \tag{28}$$

which is not equal to zero by assumption.

Second, using (1) and Assumption 3

$$r(z_a, v) = g(0, z_a)\,\eta_0(z_a)\,(1 - \eta_2^*(z_a, v)) + g(1, z_a)\,(1 - \eta_1(z_a))\,\eta_2^*(z_a, v)$$

(recall that $r(z_a, v) \equiv \mathbb{E}(yx|z_a, v)$). Evaluating $r(z_a, v)$ at two different values of the ILV v and taking differences

$$r(z_a, v_1) - r(z_a, v_2) = (\eta_2^*(z_a, v_1) - \eta_2^*(z_a, v_2))\,((1 - \eta_1(z_a))\,g(1, z_a) - \eta_0(z_a)\,g(0, z_a))$$

Next, using (25) define the directly identified quantity $\bar{r}(z_a)$ as

$$\bar{r}(z_a) \equiv \frac{r(z_a, v_1) - r(z_a, v_2)}{\eta_2(z_a, v_1) - \eta_2(z_a, v_2)}$$

which is interpretable as the Wald-IV estimate of the marginal effect of x on the random variable xy using v as an instrument. It is easy to show that

$$\bar{r}(z_a) = [(1 - \eta_1(z_a))\,g(1, z_a) - \eta_0(z_a)\,g(0, z_a)]\,\kappa(z_a) \tag{29}$$

Putting these together, we can rewrite $r(z_a, v)$ as

$$r(z_a, v) = \eta_2(z_a, v)\,\bar{r}(z_a) - \bar{t}(z_a)\,(1 - \eta_1(z_a))\,\eta_0(z_a)$$

so that the object $-\eta_0(z_a)\,(1 - \eta_1(z_a))$ is identified since the left hand side below is identified

$$r(z_a, v) - \eta_2(z_a, v)\,\bar{r}(z_a) = -\eta_0(z_a)\,(1 - \eta_1(z_a))\,\bar{t}(z_a) \tag{30}$$

and $\bar{t}(z_a)$ on the right hand side is identified. This implies that if any one of the misclassification probabilities is the same for two different points in the parameter space, the other probability must also coincide.

Finally, it is easily shown that

$$r(z_a, v) - t(z_a, v)\,\eta_2(z_a, v) = \bar{t}(z_a)\,(1 - \eta_1(z_a) - \eta_2(z_a, v))\,(\eta_2(z_a, v) - \eta_0(z_a)) \tag{31}$$

Now, suppose that the model is not identified so that there exist $(\bar{\eta}(z_a), \bar{g}(x^*, z_a))\,(\neq (\eta(z_a), g(x^*, z_a)))$ that cannot be distinguished from $(\eta(z_a), g(x^*, z_a))$ using the population distribution of the observed variables Note that we must have $\eta(z_a) \neq \bar{\eta}(z_a)$ since otherwise the model is identified as we saw in the first step. In addition, (30) implies that we must also have

$\eta_0(z_a) \neq \bar{\eta}_0(z_a)$ and $\eta_1(z_a) \neq \bar{\eta}_1(z_a)$.

Since the left hand side of (31) is directly identified and the model is not identified

$$\bar{t}(z_a)(1-\eta_1(z_a)-\eta_2(z_a,v))(\eta_2(z_a,v)-\eta_0(z_a)) = \bar{t}(z_a)(1-\bar{\eta}_1(z_a)-\eta_2(z_a,v))(\eta_2(z_a,v)-\bar{\eta}_0(z_a))$$

and since $g(1,z_a) \neq g(0,z_a)$ by assumption this can be simplified to

$$\begin{aligned}(1-\eta_1-\eta_2)(\eta_2-\eta_0) &= (1-\bar{\eta}_1-\eta_2)(\eta_2-\bar{\eta}_0)\\ \eta_2(1-\eta_1)-\eta_2^2+\eta_0\eta_2 &= \eta_2(1-\bar{\eta}_1)-\eta_2^2+\bar{\eta}_0\eta_2\\ \eta_2(1-\eta_1+\eta_0) &= \eta_2(1-\bar{\eta}_1+\bar{\eta}_0)\end{aligned}$$

suppressing the dependence of the quantities above on (z_a, v). This further simplifies to

$$\eta_2(\bar{\eta}_1-\eta_1+\eta_0-\bar{\eta}_0)=0$$

and evaluating the above at (z_a, v_1) and (z_a, v_2) and taking differences we obtain

$$(\eta_2(z_a,v_1)-\eta_2(z_a,v_2))(\bar{\eta}_1(z_a)-\eta_1(z_a)+\eta_0(z_a)-\bar{\eta}_0(z_a))=0$$

so that (under Assumptions 3 and 4) $\eta_0(z_a)-\eta_1(z_a)=\bar{\eta}_0(z_a)-\bar{\eta}_1(z_a)$ and define (again suppressing the dependence on z_a)

$$d \equiv \eta_0 + 1 - \eta_1$$

then

$$\begin{aligned}1-\bar{\eta}_1 &= d-\bar{\eta}_0\\ 1-\eta_1 &= d-\eta_0\end{aligned}$$

and substituting into (30)

$$\begin{aligned}\bar{\eta}_0(d-\bar{\eta}_0) &= \eta_0(d-\eta_0)\\ d(\bar{\eta}_0-\eta_0) &= (\bar{\eta}_0-\eta_0)(\bar{\eta}_0+\eta_0)\end{aligned}$$

so that $d=(\bar{\eta}_0+\eta_0)$ which in turn implies that

$$\begin{aligned}\bar{\eta}_1 &= 1-\eta_0\\ \bar{\eta}_0 &= 1-\eta_1\end{aligned}$$

but then $\bar{\eta}_0(z_a) + \bar{\eta}_1(z_a) > 1$ which violates Assumption 2. Therefore, $\eta(z_a)$ is identified and from the first step we can conclude therefore that $g(x^*, z_a)$ is identified.

If Assumptions 1-5 are modified to hold for almost all $z \in \mathbb{S}_z$ (by appending "$a.e.\ \mathbb{P}_z$" to the displays 2,4 and 5) then the argument above can be modified in a straightforward manner to show the complete misclassification functions $\eta(\cdot)$ as well as the entire regression function $g(\cdot)$ are identified.

The final result is quite intuitive. We show that identification failure in the presence of the ILV results *only* from the "flipping" of the misclassification probabilities as given by the last two equalities. Once these flips are ruled out by Assumption 2, the model is identified.

A.2 Proof for Identification of the Endogenous Misclassified Model

We discuss here the extension of the basic model to the one given by (10), (11) and $\mathbb{E}(\varepsilon|z_a, v) = 0$. An application of Theorem 1 yields identification of the regression

$$\mathbb{E}(y|x^*, z_a) = g^*(x^*, z_a) + \mathbb{E}(\varepsilon|x^*, z_a)$$

Next,

$$\begin{aligned}
\mathbb{E}(y|z_a, v) &= \sum_{s\in\{0,1\}} \mathbb{E}(y|x^* = s, z_a, v)\,\mathbb{P}(x^* = s|z_a, v) \\
&= g^*(0, z)(1 - \eta_2^*(z_a, v)) + g^*(1, z)\,\eta_2^*(z_a, v) \\
&\quad + \sum_{s\in\{0,1\}} \mathbb{E}(\varepsilon|x^* = s, z_a, v)\,\mathbb{P}(x^* = s|z_a, v) \\
&= g^*(0, z_a)(1 - \eta_2^*(z_a, v)) + g^*(1, z_a)\,\eta_2^*(z_a, v)
\end{aligned}$$

where the last equality holds since $\mathbb{E}(\varepsilon|z_a, v) = 0$. We can then use the variation of $\eta_2^*(z_a, v)$ in v (which is also identified by Theorem 1) to identify $g^*(x^*, z)$ exactly as we did in the argument preceding (7).

A.3 Proof of Lemma 1

Proof: The proof is a direct application of Theorem 1.

A.4 Estimation

We next discuss the elements required to implement the estimation strategy outlined in the text. We directly identify the quantities $g(1, z_a)$ and $g(0, z_a)$ and hence also the marginal effect $g(1, z_a) - g(0, z_a)$. To this end, define

$$q_0(z_a, v) \equiv t(z_a, v) + \bar{r}(z_a) - \eta_2(z_a, v)\,\bar{t}(z_a)$$

Using (25)-(29) it is easy to see that

$$\begin{aligned} q_0(z_a, v) &= (1 - \eta_2^*(z_a, v))\, g(0, z_a) + \eta_2^*(z_a, v)\, g(1, z_a) \\ &\quad + (1 - \eta_2^*(z_a, v))\, g(1, z_a) + \eta_2^*(z_a, v)\, g(0, z_a) \\ &= g(1, z_a) + g(0, z_a) \end{aligned}$$

and finally define

$$\begin{aligned} q(z_a) &= \frac{1}{2}\left(q_0(z_a, v_1) + q_0(z_a, v_2)\right) \\ &= g(1, z_a) + g(0, z_a) \end{aligned} \tag{32}$$

which will be useful in the sequel.

Using (30) it is straightforward to show that

$$\frac{\tilde{c}(z_a)}{\bar{t}(z_a)} = -\eta_0(z_a)\left(1 - \eta_1(z_a)\right) \tag{33}$$

where

$$\tilde{c}(z_a) = \frac{1}{2}\left\{r(z_a, v_1) + r(z_a, v_2) - \left(\eta_2(z_a, v_1) + \eta_2(z_a, v_2)\right)\bar{r}(z_a)\right\} \tag{34}$$

and similar calculations yield

$$\frac{\tilde{b}(z_a)}{\bar{t}(z_a)} = \eta_1(z_a) - \eta_0(z_a) \tag{35}$$

for

$$\tilde{b}(z_a) = \frac{1}{2}\left\{t(z_a, v_1) + t(z_a, v_2) - \left(\eta_2(z_a, v_1) + \eta_2(z_a, v_2) - 2\right)\bar{t}(z_a) - 2\bar{r}(z_a)\right\} \tag{36}$$

which we rewrite to make its depedence on $(\mathbb{E}(w|z_a, v_1), \mathbb{E}(w|z_a, v_2))$ explicit as

$$\begin{aligned}\tilde{b}(z_a) &= \frac{1}{2}\left(\mathbb{E}(y|z_a, v_1) + \mathbb{E}(y|z_a, v_2) - \left(\mathbb{E}(x|z_a, v_1) + \mathbb{E}(x|z_a, v_2)\right)\bar{t}\left(\mathbb{E}(w|z_a, v_1), \mathbb{E}(w|z_a, v_2)\right)\right) \\ &\quad -\bar{r}\left(\mathbb{E}(w|z_a, v_1), \mathbb{E}(w|z_a, v_2)\right)\end{aligned}$$

where $\bar{t}$ and $\bar{r}$ are defined in (44) and (45) respectively.

Equations (33) and (35) imply a quadratic equation for η_1 (suppressing the dependence on z) in terms of the directly identified quantities $\tilde{c}(z), \tilde{b}(z)$ and $\bar{t}(z)$. Solving this quadratic equation for η_1 (and using the root with the negative square root term in the quadratic formula [21]) and substituting into (35) yields

$$\sqrt{\left(\frac{\tilde{b}+\bar{t}}{\bar{t}}\right)^2 - 4\left(\frac{\tilde{b}-\tilde{c}}{\bar{t}}\right)} = 1 - \eta_0 - \eta_1 \tag{37}$$

where it is easily seen (using the definitions of the quantities defined in (33) and (35) and Assumption 2) that the term under the square root sign is always non-negative. Solving for the misclassification rates

$$\eta_1(z_a) = \left(1 + h_1\left(\mathbb{E}(w|z_a, v_1), \mathbb{E}(w|z_a, v_2)\right) - h_0\left(\mathbb{E}(w|z_a, v_1), \mathbb{E}(w|z_a, v_2)\right)\right)2^{-1} \tag{38}$$

$$\eta_0(z_a) = \left(1 - h_1\left(\mathbb{E}(w|z_a, v_1), \mathbb{E}(w|z_a, v_2)\right) - h_0\left(\mathbb{E}(w|z_a, v_1), \mathbb{E}(w|z_a, v_2)\right)\right)2^{-1} \tag{39}$$

where

$$h_1\left(\mathbb{E}(w|z_a, v_1), \mathbb{E}(w|z_a, v_2)\right) = \frac{\tilde{b}(z_a)}{\bar{t}(z_a)} \tag{40}$$

$$h_0(z_a) = \sqrt{\left(\frac{\tilde{b}(z_a)+\bar{t}(z_a)}{\bar{t}(z_a)}\right)^2 - 4\left(\frac{\tilde{b}(z_a)-\tilde{c}(z_a)}{\bar{t}(z_a)}\right)} \tag{41}$$

Then using (28), it is easily seen that

$$g(1, z_a) - g(0, z_a) = \bar{t}(z_a)\sqrt{\left(\frac{\tilde{b}(z_a)+\bar{t}(z_a)}{\bar{t}(z_a)}\right)^2 - 4\left(\frac{\tilde{b}(z_a)-\tilde{c}(z_a)}{\bar{t}(z_a)}\right)}$$

[21] This is because in (37) this is the root that satisfies Assumption 2.

and using (32) above,

$$g(1,z_a)=\frac{1}{2}q(z_a)+\frac{\bar{t}(z_a)}{2}\sqrt{\left(\frac{\tilde{b}(z_a)+\bar{t}(z_a)}{\bar{t}(z_a)}\right)^2-4\left(\frac{\tilde{b}(z_a)-\tilde{c}(z_a)}{\bar{t}(z_a)}\right)} \tag{42}$$

where each term on the right hand side is directly identified. Similarly,

$$g(0,z_a)=\frac{1}{2}q(z_a)-\frac{\bar{t}(z_a)}{2}\sqrt{\left(\frac{\tilde{b}(z_a)+\bar{t}(z_a)}{\bar{t}(z_a)}\right)^2-4\left(\frac{\tilde{b}(z_a)-\tilde{c}(z_a)}{\bar{t}(z_a)}\right)} \tag{43}$$

so that we have directly identified the regression function $g(x^*,z_a)$.

We next discuss in greater detail the components of the directly identified terms in the displays above as a function of the observed moments $\mathbb{E}(w|z,v_1)$ and $\mathbb{E}(w|z,v_2)$ (where recall that $w=(x,y,xy)$). The term $\bar{t}$ is what we refer to as the Wald-IV estimator of the effect of x on y

$$\bar{t}(\mathbb{E}(w|z_a,v_1),\mathbb{E}(w|z_a,v_2))=\frac{\mathbb{E}(y|z_a,v_1)-\mathbb{E}(y|z_a,v_2)}{\mathbb{E}(x|z_a,v_1)-\mathbb{E}(x|z_a,v_2)} \tag{44}$$

which we abbreviate as $\bar{t}(z_a)\,(\equiv\frac{\bar{t}_n(z_a)}{\bar{t}_d(z_a)})$. Similarly, we abbreviate as $\bar{r}(z)$ the Wald-IV expression

$$\bar{r}(\mathbb{E}(w|z_a,v_1),\mathbb{E}(w|z_a,v_2))=\frac{\mathbb{E}(xy|z_a,v_1)-\mathbb{E}(xy|z_a,v_2)}{\mathbb{E}(x|z_a,v_1)-\mathbb{E}(x|z_a,v_2)} \tag{45}$$

We next define

$$q_0(z_a,v)\equiv q_0(\mathbb{E}(w|z_a,v))=\mathbb{E}(y|z_a,v)+\bar{r}(z_a)-\mathbb{E}(x|z_a,v)\,\bar{t}(z_a)$$

and the function we abbreviate as $q(z_a)$ the function $q(\mathbb{E}(w|z_a,v_1),\mathbb{E}(w|z_a,v_2))$

$$q(\mathbb{E}(w|z_a,v_1),\mathbb{E}(w|z_a,v_2))=\frac{1}{2}(q_0(z_a,v_1)+q_0(z_a,v_2)) \tag{46}$$

The expressions $\tilde{b}(\cdot)$ and $\tilde{c}(\cdot)$ are defined previously.

Estimation proceeds, as outlined in the text, by first non-parametrically estimating the conditional expectations $\mathbb{E}(w|z,v)$ using kernel based methods and plugging in the estimates into the expressions above. Consistency and asymptotic normality follow using standard results from the kernel literature which we outline below.

A.4.1 Proof of Lemma 2 and Lemma 3

The first proof is a direct application of Theorem 3.2.1 (and the Cramer-Wold device) in Bierens (1987) and the conditions 10-14 in the text are precisely his conditions 3.2.1 on p.118.

The second proof follows from the "delta" method as detailed for instance in Theorem 3.1 of van der Vaart (1998). The form of the variance matrices follows from considering the derivatives of the functions $q(\cdot)$, $\bar{t}(\cdot), \tilde{b}(\cdot)$ and $\tilde{c}(\cdot)$ with respect to the objects $(\mathbb{E}(w|z, v_1), \mathbb{E}(w|z, v_2))$. More compactly, consider

$$\begin{aligned} g(1, z_a) &= \frac{1}{2} q(z_a) + \frac{\bar{t}(z_a)}{2} \sqrt{\left(\frac{\tilde{b}(z_a) + \bar{t}(z_a)}{\bar{t}(z_a)} \right)^2 - 4 \left(\frac{\tilde{b}(z_a) - \tilde{c}(z_a)}{\bar{t}(z_a)} \right)} \\ &= g_1(\mathbb{E}(w|z_a, v_1), \mathbb{E}(w|z_a, v_2)) \\ &= g_1(\mathbb{E}(x|z_a, v_1), \mathbb{E}(y|z_a, v_1), \mathbb{E}(xy|z_a, v_1), \mathbb{E}(x|z_a, v_2), \mathbb{E}(y|z_a, v_2), \mathbb{E}(xy|z_a, v_2)) \end{aligned}$$

where we emphasize in the last display that $g(1, z_a)$ is a function of six arguments and we denote by $\nabla g_1(\cdot)$ the derivative of $g(1, z_a)$ with respect to these six arguments.

Then, by the "delta" method the asymptotic variance matrix for the estimator $\hat{g}(1, z_a)$ will be given by

$$\begin{aligned} \Omega_1 &= \left(\nabla g_1(\mathbb{E}(w|z_a, v_1), \mathbb{E}(w|z_a, v_2))' \right) \begin{bmatrix} V_1 & 0 \\ 0 & V_2 \end{bmatrix} (\nabla g_1(\mathbb{E}(w|z_a, v_1), \mathbb{E}(w|z_a, v_2))) \\ &= \frac{f_1(\mathbb{E}(w|z_a, v_1), \mathbb{E}(w|z_a, v_2), V(w|z_a, v_1), V(w|z_a, v_2))}{2[(\mathbb{E}(y|z_a, v_1) - \mathbb{E}(y|z_a, v_2))(\mathbb{E}(x|z_a, v_1) - \mathbb{E}(x|z_a, v_2))]^2 (1 - \eta_0(z_a) - \eta_1(z_a))} \end{aligned}$$

where $V(w|z_a, v_k)$ denotes the conditional variance-covariance matrix of the vector (x, y, xy). The precise form of the function $f_1(\cdot)$ is somewhat complicated and does not add insight and is therefore omitted.

The argument for the asymptotic variance Ω_0 of $\hat{g}(0, z)$ and the variance Ω_M of the marginal effect $\hat{g}(1, z) - \hat{g}(0, z)$ follows similarly. The nonparametric bootstrap can be used to consistently estimate the standard errors (see for instance Jones and Wand (1995) or Shao and Tu (1996)) and we omit details here.

B Semiparametric Estimation

Recall from Appendix A.4 that

$$q(z) = g(1, z) + g(0, z)$$

which for the single (linear) index model reduces to

$$q(z) = g(\theta_{z1} + z'\theta_{z0}) + g(z'\theta_{z0})) \equiv G(z'\theta_{z0})$$

as stated in the text.

References

Abrevaya, J., and J. Hausman (1999): "Semiparametric Estimation with Mismeasured Dependent Variables: An Application to Duration Models with Unemployment Spells," *Annales d'Economie et de Statistique*, 55/56, 243–275.

Abrevaya, J., J. Hausman, and F. Scott-Morton (1998): "Misclassification of the dependent variable in a discrete response Setting," *Journal of Econometrics*, 87, 239–269.

Ai, C., and X. Chen (2001): "Efficient Sieve Minimum Distance Estimation of Semiparametric Conditional Moment Models," Working Paper.

Aigner, D. (1973): "Regression with a Binary Independent Variable Subject to Errors of Observations," *Journal of Econometrics*, 1, 49–60.

Amemiya, Y. (1985): "Two-Stage Instrumental Variable Estimators for the Non-linear Errors-in-Variables Model," *Journal of Econometrics*, 44, 311–332.

Bierens, H. (1987): "Kernel Estimators of Regression Functions," in *Fifth World Congress, Volume I*, ed. by T. F. Bewley, vol. I of *Advances in Econometrics*, chap. 4, pp. 99–144. Cambridge University Press.

Black, D., M. C. Berger, and F. A. Scott (2000): "Bounding Parameter Estimates with Nonclassical Measurement Error," *Journal of the American Statistical Association*, 95(451), 739–748.

Bollinger, C. (1996): "Bounding Mean Regressions When A Binary Regressor is Mismeasured," *Journal of Econometrics*, 73(2), pp 387–399.

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

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

Bound, J., C. Brown, and N. Mathiowetz (2000): "Measurement Error in Survey Data," Institute for Social Research, University of Michigan.

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

Buzas, J., and L. Stefanski (1996): "Instrumental Variable Estimation in Generalized Linear Measurement Error Models," *Journal of the American Statistical Association*, 91(435), 999–1006.

Card, D. (1996): "The Effect of Unions on the Structure of Wages: A Longitudnal Analysis," *Econometrica*, 64(4), 957–979.

Carroll, R., D. Ruppert, and L. Stefanski (1995): *Measurement Error in Non-Linear Models*. Chapman and Hall.

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

Chen, X., H. Hong, and E. Tamer (2002): "Measurement Error Models with Auxiliary Data," Princeton University.

Chesher, A. (1991): "The Effect of Measurement Error," *Biometrika*, 78, 451–462.

Das, M. (2004): "Instrumental Variables Estimation of Nonparametric Models with Discrete Endogenous Regressors," *Journal of Econometrics*, 124, 335–361.

Hausman, J., W. Newey, H. Ichimura, and J. Powell (1991): "Identification and Estimation of Polynomial Errors-in-Variables Models," *Journal of Econometrics*, 50(3), 273–296.

Hong, H., and E. Tamer (2001): "A Simple Estimator for Non Linear Errors in Variables Models," Princeton University.

Horowitz, J. (1998): *Semiparametric Methods in Econometrics*. Springer-Verlag.

Horowitz, J., and C. Manski (1995): "Identification and Robustness with Contaminated and Corrupt Data," *Econometrica*, 63(2), 281–302.

Ichimura, H. (1993): "Semiparametric Least Squares and Weighted SLS Estimation of Single Index Models," *Journal of Econometrics*, 58(1-2), 71–120.

Imbens, G. W., and D. Hyslop (2000): "Bias from Classical and other forms of Measurement Error," Discussion Paper 257, National Bureau of Economic Research.

Jones, M., and M. Wand (1995): *Kernel Smoothing*. Chapman and Hall.

Kane, T., C. E. Rouse, and D. Staiger (1999): "Estimating Returns to Schooling when Schooling is Misreported," Discussion Paper 7235, National Bureau of Economic Research.

Lee, L., and J. Sepanski (1995): "Estimation of Linear and Nonlinear Errors-in-Variables Models Using Validation Data," *Journal of the American Statistical Association*, 90, 130–140.

Lewbel, A. (1996): "Demand Estimation with Expenditure Measurement Errors on the Left and Right Hand Side," *The Review of Economics and Statistics*, 78, 718–725.

——— (2000): "Identification of the Binary Choice Model with Misclassification," *Econometric Theory*, 16, 603–60.

——— (2004): "Estimation of Average Treatment Effects Under Misclassification," Working Paper, Boston College.

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

Mahajan, A. (2003): "Misclassified Regressors in Binary Choice Models," Stanford University.

——— (2004): "Identification and Estimation of Single-Index Models with Misclassified Regressors," Stanford University.

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

Molinari, F. (2004): "Partial Identification of Probability Distributions with Misclassified Data," Cornell University.

Newey, W. (1994a): "The Asymptotic Variance of Semiparametric Estimators," *Econometrica*, 62(6), 1349–1382.

——— (1994b): "Kernel Estimation of Partial Means and a General Variance Estimator," *Econometric Theory*, 10, 233–253.

——— (2001): "Flexible Simulated Moment Estimation of Nonlinear Errors-in-Variables Models," *Review of Economics and Statistics*, 83, 616–627.

Newey, W., and D. McFadden (1994): "Large Sample Estimation and Hypothesis Testing," in *Handbook of Econometrics*, ed. by R. Engle, and D. McFadden, vol. IV, chap. 36, pp. 2111–2245. Elsevier Science.

Newey, W., and T. Stoker (1993): "Efficiency of Weighted Average Derivative Estimators and Index Models," *Econometrica*, 61, 1199–1223.

Powell, J., J. Stock, and T. Stoker (1989): "Semiparametric Estimation of Index Coefficients," *Econometrica*, 57(6), 1403–1430.

Schennach, S. (2004a): "Estimation of Non Linear Models with Measurement Error," *Econometrica*, 72(1), 33–75.

——— (2004b): "Instrumental Variable Estimation of Nonlinear Errors-in-Variables Models," University of Chicago.

Shao, J., and D. Tu (1996): *The Jackknife and Bootstrap*. Springer Verlag.

Stefanski, L. A., and J. Buzas (1995): "Instrumental Variable Estimation in Binary Regression Measurement Error Models," *Journal of the American Statistical Association*, 90(430), 541–550.

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

van der Vaart, A. (1998): *Asymptotic Statistics*. Cambridge University Press.

Zinman, J. (2004): "Why Use Debit Instead of Credit? Consumer Choice in a Trillion Dollar Market," New York Federal Reserve.