Two-step Parametric Estimation of Binary Treatment Effects in the Presence of Misclassification and Endogeneity for Cross-Sectional and Panel Data

Georgios Marios Chrysanthou

No. 87 /22

BATH ECONOMICS RESEARCH PAPERS

Department of Economics

Department of Economics | UNIVERSITY OF BATH

Two-step Parametric Estimation of Binary Treatment Effects in the Presence of Misclassification and Endogeneity for Cross-Sectional and Panel Data*

Georgios Marios Chrysanthou†

January, 2022

Abstract

I propose a parametric two-step estimator that corrects for the bias arising from measurement error in a binary endogenous regressor, in cross-sectional and panel data settings. The model incorporates asymmetric (unequal) misclassification rates for false negatives and false positives and is directly generalised to the symmetric misclassification case. It is demonstrated that consistent estimation of the binary treatment effect and the remaining structural form parameters is achieved via modified MLE (MMLE) estimation of the reduced form binary discrete choice model and, via modified least squares (MSL) estimation of the structural form augmented by a misclassification-corrected control function. The model is identified by the nonlinearity of the endogeneity correction terms.

JEL classification: C25, C31, C33, C35, C36

Keywords: measurement error, misclassification, binary endogenous variable, treatment effect

*I am grateful to Jason Abrevaya, Nikolas Mittag, Pedro Albarran, Raquel Carrasco and participants at the 2017 International Panel Data Conference and, the Trobada Economia: Universitat Rovira i Virgili & Universitat d'Alacant for helpful comments.

†Department of Economics, University of Bath, BA2 7AY, UK. E-mail: gc868@bath.ac.uk.

1 Introduction

In dealing with an endogenous binary (right-hand-side, RHS) variable within a two-stage parametric endogenous treatment framework, we are effectively facing measurement error problems in both estimation steps. In the second estimation stage, the structural equation includes a binary explanatory variable that is both mismeasured and endogenously determined. Regarding the first stage, the reduced form used to generate the appropriate endogeneity correction terms is a binary choice model where the dependent variable is subject to misclassification. In the case of a misclassified binary endogenous treatment if extraneous information regarding misclassification and alternative measurements or instruments are unavailable, ignoring measurement error produces inconsistent parameter estimates for the binary indicator variable treatment effect. Consistency requires modelling misclassification in the reduced form in order to modify the control function to incorporate misclassification probabilities and, adding the modified control function to the structural form to achieve endogeneity correction. To the best of my knowledge this study is the first to fill this important gap in the literature by offering tractable cross-sectional and panel data two-step methods providing consistent binary treatment effects in the presence of binary endogenous treatment misclassification.

Measurement error in an exogenous binary RHS variable produces biased and inconsistent parameter estimates and has therefore received early attention by authors such as Cochran (1968) and Aigner (1973) in the past. Aigner (1973) provides a parametric modified least squares (MLS) framework purging the problematic correlation between the misclassified binary explanatory variable and the measurement error while Bollinger (1996) studies parametric and non-parametric identification of parameter bounds.

Measurement error from the left-hand-side (LHS) i.e. in a binary dependent variable is equally problematic and also produces biased and inconsistent parameter estimates. In the case of a misclasssified binary dependent variable, using a cross-sectional setting Hausman *et al.* (1998) demonstrate that consistent estimation can be achieved via a modification of the likelihood function to correct for misclassification using both parametric and semi-parametric techniques. Meyer and Mittag (2017) extend the work of Hausman *et al.* (1998) by permitting a correlation of binary LHS variable misclassification with both observables and unobservables within a cross-sectional setting.

Chen, Hong and Nekipelov (2011) provide a survey of cross-sectional nonlinear models of measurement error with a focus on istrumentation and the availability of an additional auxiliary sample to resolve the resulting inconsistency. Mahajan (2006) studies nonparametric cross-sectional estimation with a misclassified binary regressor assuming the availability of an additional nondegenerate regressor taking at least two values to be used as an instrument. Carroll, Chen and Hu (2006) study the case of mismeasured discrete regressors using two samples both containing measurement errors within a cross-sectional setting. Heider and Stephens (2020) correct for exogenous (and weakly exogenous) misreporting by allowing binary regressor misclassification rates to vary with a discrete-valued instrumental variable in a cross-sectional setting.[1]

Exogenous misreporting of a binary treatment RHS variable leads to attenuation bias (see Aigner, 1973; Lewbel, 2007) but, the impact of endogenous binary RHS variable misreporting on average treatment effects is even more severe (see Nguimkeu *et al.*, 2019). Lewbel (2007), considers a cross-sectional setting and proposes an instrument for the mismeasured binary regressor that is conditionally independent of the treatment effect. Nguimkeu *et al.*, (2019) consider an incomplete data scenario of endogenous binary RHS variable misreporting and potentially endogenous participation where the true participation indicator is unobserved and, only a misclassified surrogate contaminated by misreporting is observed. Nguimkeu *et al.*, (2019) only consider the cross-sectional case and one-sided endogenous misreporting with endogenous participation, and do not treat bidirectional misreporting accounting for false negatives and false positives nor the panel data case. Kasahara and Shimotsu (2021) study nonparametric identification of cross-sectional models with an endogenous misclassified binary regressor assuming the instrument is correlated with the misclassification error.

Regarding panel data variants, the case of a misclassified binary LHS dependent variable has been considered by Keane and Sauer (2009, 2010) that generalise the panel probit model by nesting it in a model of dependent variable missclassification error. Unlike Hausman *et al.* (1998), Meyer and Mittag (2017) and the present study, Keane and Sauer (2009, 2010) do not directly estimate misclassification parameters by modified maximum likelihood (MMLE) but instead, compute misclassification rates from the parameters estimated from a dynamic binary model.

This study is unique and contributes to the literature by introducing a tractable two-step parametric procedure to consistently estimate models with an endogenous mismeasured binary explanatory variable within a cross-sectional, as well as, a panel data framework accounting for both symmetric and

[1]Concerning linear models Hu, Shiu and Woutersen (2015) study single-index model identification and estimation when continuous variables are endogenous and measured with error in a cross-sectional setting.

asymmetric (bidirectional) misreporting. Identification does not require an instrument nor additional alternative measurements for the misclassified binary variable. The model parameters are identified by the non-linearity of the (modified) misclassification-adjusted endogeneity correction terms.

The paper is organised as follows. Section 2 analyses the issue of misclassification in a binary endogenous explanatory variable within an endogenous treatment framework consisting of two equations. Section 3 provides the misclassification-amended endogeneity correction terms in the cross-sectional and the panel data cases, respectively. Conclusions are given in Section 4. The Appendix presents the proofs regarding the consistency of the second stage modified least squares (MLS) estimators. Longitudinal binary misclassified discrete choice modelling via modified maximum likelihood (MMLE) is treated in additional Supplementary material.

2 Misclassification Error in a Binary Endogenous Explanatory Variable

The model of interest is a binary endogenous variable model *i.e.* an endogenous treatment model which does not essentially correspond to a sample selection framework *per se* as for example in Nguimkeu *et al*. (2019) and Vella (1998). [2] We partition the structural equation population regression function as

$$\begin{aligned} y_{j,it} &= \mathbf{z}_{it}\beta_{j,t} + \delta x^*_{it} + e_{j,it} & (1) \\ e_{j,it} &= \alpha_{j,i} + \varepsilon_{j,it}, (t = 1,..,T; i = 1,...,N)\,,\; j = \{0,1\} & (2) \\ j &= (0,1)\; if\; x^*_{it} = (0,1), y_{it} = y_{j,it}\;\; if\;\; x^*_{it} = j, (t = 1,..,T; i = 1,...,N) \end{aligned}$$

where, β is a $[(k-1)\,\text{x}1]$ vector of unknown parameters, $\mathbf{z}_{it}$ is an $[1\text{x}\,(k-1)]$ vector of explanatory variables other than x^*_{it}. The random components satisfy $\alpha_{j,i} \sim iid\; N(0, \sigma^2_\alpha)$, $\varepsilon_{j,it} \sim iid\; N(0, \sigma^2_\varepsilon)$, $E(\mathbf{z}'_{it}\alpha_{j,i})$=0 and $E(\mathbf{z}'_{it}\varepsilon_{j,it})$=0. The true value of the endogenous binary explanatory variable is given by x^*_{it} and the respective treatment effect is denoted by scalar δ and is the key parameter of interest. The reduced form for x^*_{it} is

[2] The cross-sectional model variants are obtained by setting $T = 1$ and omitting the time-variant error component where applicable.

$$x_{it}^{*} = \mathbf{1}\left\{\mathbf{w}_{it}\gamma + \nu_{it} \geq 0\right\}, \nu_{it} = \theta_i + \eta_{it}, (t = 1, .., T; i = 1, ..., N) \tag{3}$$

where, $\mathbf{1}\{c\}$ takes the value of one if condition c is satisfied and zero otherwise, γ is a $[k\text{x}1]$ vector of unknown parameters, $\mathbf{w}_{it}$ is a $[1\text{x}k]$ vector of explanatory variables, $\theta_i \sim iid\ N(0, \sigma_\theta^2)$ and, $\eta_{it} \sim iid\ N(0, \sigma_\eta^2)$. Structural identification of the two-part model requires that at least one element in $\mathbf{w}_{it}$ is not included in $\mathbf{z}_{it}$.

Introducing the restriction of invariance of $\beta_{j,t}$ with respect to (j, t) in Eq. (1) we obtain

$$y_{it} = \mathbf{z}_{it}\beta + \delta x_{it}^{*} + e_{it}, e_{it} = x_{it}^{*}(\alpha_{1,i} + \varepsilon_{1,it}) + (1 - x_{it}^{*})(\alpha_{0,i} + \varepsilon_{0,it}) \tag{4}$$

which in vector form corresponds to

$$\mathbf{y}_i = \mathbf{Z}_i\beta + \mathbf{x}_i^{*}\delta + \mathbf{e}_i. \tag{5}$$

where $\mathbf{Z}_i$ is a $T\text{x}(K-1)$ matrix of observations of all explanatory variables except x^{*}, and $\mathbf{e}_i$ denotes a $(T\text{x}1)$ vector of residuals. Under the assumption that the error components are *iid* distributed the trivariate joint normal distribution is

$$\begin{pmatrix} \alpha_{j,i} \\ \varepsilon_{j,i} \\ \theta_i \imath + \eta_i \end{pmatrix} \sim N \begin{pmatrix} 0 & \sigma_{j,\alpha}^2 & 0 & \sigma_{j,\alpha\theta}\imath' \\ 0\ , & & \sigma_{j,\varepsilon}^2 I & \sigma_{j,\varepsilon\eta} I \\ 0 & & & \sigma_\eta^2 + \sigma_\theta^2 \imath\imath' \end{pmatrix}. \tag{6}$$

Let (x_{it}^{*}, x_{it}) be the true/observed indicators for the outcome of individual i in period t, respectively where x_{it} is subject to misclassification error. Assuming that $(x_{it}^{*},\ x_{it})$ are related via $x_{it} = x_{it}^{*} + \tau_{it}$ where τ_{it} is a random measurement error, Eq. (4) becomes

$$\begin{aligned} y_{it} &= z_{it}\beta + \delta x_{it} + u_{it}, u_{it} = (e_{it} - \delta\tau_{it}) \\ e_{it} &= (x_{it} - \tau_{it})(\alpha_{1,i} + \varepsilon_{1,it}) + (1 - x_{it} + \tau_{it})(\alpha_{0,i} + \varepsilon_{0,it}) \end{aligned} \tag{7}$$

which in vector form corresponds to

$$\mathbf{y}_i = \mathbf{Z}_i\beta + \mathbf{x}_i\delta + \mathbf{u}_i \tag{8}$$

where $\mathbf{Z}_i$ is a Tx$(K-1)$ matrix of observations of all explanatory variables other than x, and $\mathbf{u}_i$ denotes a $(T$x$1)$ vector of residuals.

Denote the probability of a false positive/negative by (λ_1,λ_2) conditional on the observed binary indicator x_{it}. It is assumed that the misclassification probabilities are constant across individuals and that $\lambda_1 \neq \lambda_2$ *i.e.* misclassification is asymmetric. The results in the remaining paper are easily extended to the case of symmetric misclassification by simply setting $\lambda_1 = \lambda_2$. For tractability in the longitudinal case (λ_1, λ_2) are specified as time-invariant so that

$$\lambda_1 = \Pr(x^*_{it} = 0|x_{it} = 1), \lambda_2 = \Pr(x^*_{it} = 1|x_{it} = 0) \tag{9}$$

noting that (λ_1,λ_2) depend on x_{it} and are assumed to be independent of $\mathbf{w}_{it}$ and ν_{it}. If misclassification is time-dependent then cross-sectional estimation should be performed for each time period separately. The expected values of the true and observed responses $(\pi, \widehat{\pi})$ are related by

$$\pi = (1-\lambda_1)\,\widehat{\pi} + \lambda_2(1-\widehat{\pi}) = \widehat{\pi}(1-\lambda_1-\lambda_2) + \lambda_2, \widehat{\pi} = (NT)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T} x_{it}. \tag{10}$$

The marginal distributions of $(x^*_{it},\ x_{it})$ are Bernoulli with parameters $(\pi, \widehat{\pi})$, respectively. By manipulation, the joint distribution of (x_{it}, τ_{it}) is:

	x_{it}		
τ_{it}	**0**	**1**	$f(\tau_{it})$
-1	$\lambda_2(1-\widehat{\pi})$	0	$\lambda_2(1-\widehat{\pi})$
0	$(1-\lambda_2)\,(1-\widehat{\pi})$	$(1-\lambda_1)\,\widehat{\pi}$	$(1-\lambda_2)\,(1-\widehat{\pi}) + (1-\lambda_1)\,\widehat{\pi}$
1	0	$\lambda_1\widehat{\pi}$	$\lambda_1\widehat{\pi}$
$f(x_{it})$	$(1-\widehat{\pi})$	$\widehat{\pi}$	**1**

There is a negative correlation between true response and the measurement error since when $x^*_{it} = 1$,

τ_{it} is either -1/0 while when $x^*_{it} = 0$, τ_{it} is either 0/1. Using the joint frequency table:

$$\begin{aligned} E(\tau_{it}) &= \lambda_1 \widehat{\pi} - \lambda_2 (1 - \widehat{\pi}) \qquad (11) \\ Var(\tau_{it}) &= [\lambda_1 \widehat{\pi} + \lambda_2 (1 - \widehat{\pi})] - [\lambda_1 \widehat{\pi} - \lambda_2 (1 - \widehat{\pi})]^2 \qquad (12) \\ E(x_{it}) &= \widehat{\pi},\ Var\,(x_{it}) = \widehat{\pi}(1 - \widehat{\pi}) \qquad (13) \\ E(x_{it}\tau_{it}) &= \lambda_1 \widehat{\pi},\ Cov(x_{it}, \tau_{it}) = (\lambda_1 + \lambda_2)\, \widehat{\pi}(1 - \widehat{\pi}) \qquad (14) \end{aligned}$$

Contrary to classical errors in variables assumptions, τ_{it} does not have a zero mean, it is negatively correlated with x^*_{it} and it is also correlated with x_{it}.

Given the MLE estimates of (λ_1, λ_2), we can compute $Cov(x_{it}, \tau_{it})$ and consistently estimate (β, δ) employing the modified least squares (MLS) procedure described in Aigner (1973) which will have to be amended by the inclusion of an appropriate control function to resolve the endogeneity of x_{it}.

Using Eq. (8), the MLS estimators for (β,δ) in the partitioned linear model, where Z is a $[T\text{x}\,(k-1)]$ matrix of observations on explanatory variables other than x, are

$$\begin{aligned} \begin{bmatrix} \widehat{\beta} \\ \widehat{\delta} \end{bmatrix} &= \begin{bmatrix} M_{ZZ} & M_{Zx} \\ M'_{Zx} & (m_{xx} - \xi) \end{bmatrix}^{-1} \begin{bmatrix} M_{Zy} \\ m_{xy} \end{bmatrix} \qquad (15) \\ M_{ZZ} &= (NT)^{-1} (Z'Z), M_{Zx} = (NT)^{-1} (Z'x), M_{Zy} = (NT)^{-1} (Z'y) \\ m_{xx} &= (NT)^{-1} (x'x), m_{xy} = (NT)^{-1} (x'y), \xi = Cov(x_{it}, \tau_{it}). \end{aligned}$$

Theorem 2.1. *The MLS estimator gives* $\underset{N \longrightarrow \infty}{plim}\, \widehat{\beta} = \beta$,

$$\underset{N \longrightarrow \infty}{plim}\, \widehat{\delta} = \delta + \underset{N \longrightarrow \infty}{plim} \left[m_{xx}^{-1} \left(\frac{1}{NT} x'e \right) \left[1 - \xi m_{xx}^{-1} \right]^{-1} \right] \qquad (16)$$

corresponding to

$$\underset{N \longrightarrow \infty}{plim}\, \widehat{\delta} = \delta + \underset{N \longrightarrow \infty}{plim} \left[(x'x)^{-1} (x'e)\, \psi \right], \psi = \left[1 - \xi m_{xx}^{-1} \right]^{-1}. \qquad (17)$$

Proof. see **Appendix, A**.

Theorem 2.1 is crucial in the ensuing analysis since unlike Aigner (1973), the endogeneity of x means that $\underset{N\longrightarrow\infty}{plim}\left(\frac{1}{NT}x'e\right)\neq 0$ and solely modifying the LS estimator to purge for the bias stemming from measurement error does not provide a consistent estimate of δ.

It is clear that the inconsistency in Eq. (17) can be resolved by adding an appropriately constructed control function, obtained from the reduced form estimates as an additional regressor in the structural form, interacted by ψ (see Fernandez-Val and Vella, 2011; Heckman, 1979; Vella, 1998; Verbeek and Nijman, 1992; Vella and Verbeek, 1999 a,b).

Consistent parameter estimation requires modifying the likelihood function in the reduced form to account for non-zero misclassification probabilities and modification of the relevant moment matrix of covariates in the structural equation- see Eq. (15). We treat this in the following Section.

3 Deriving the Endogeneity Correction Terms in the Presence of Misclassification

3.1 The Cross-Sectional Case

In the cross-sectional case and absence of misclassification the model with an endogenous binary explanatory variable, assuming invariance of β_j with respect to j, corresponds to

$$
\begin{aligned}
y_i &= \mathbf{z}_i\beta + \delta x_i^* + e_i,\ \ (i=1,...,N) && (18)\\
x_i^* &= \mathbf{1}\left\{\mathbf{w}_i\gamma + \eta_i \geq 0\right\},\ (i=1,...,N) && (19)\\
j &= (0,1)\ if\ x_i^* = (0,1), y_i = y_{j,i}\ \ if\ \ x_i^* = j, (i=1,...,N)
\end{aligned}
$$

where it is assumed that e_i and η_i, $i=1,...,N$ are *iid* normally distributed

$$
\begin{pmatrix} e_i \\ \eta_i \end{pmatrix} \sim N\begin{pmatrix} 0 & & \sigma_e^2 & \sigma_{e\eta} \\ & , & & \\ 0 & & \sigma_{e\eta} & \sigma_\eta^2 \end{pmatrix}. \quad (20)
$$

It is assumed that $E(\mathbf{z}_i'e_i)$=0, while the endogeneity of x produces a non-zero conditional expectation of e_i in the structural form which in the presence of misclassification error, τ_i=$(x_i - x_i^*)$, becomes

$$
\begin{aligned}
y_i &= \mathbf{z}_i\beta + \delta x_i + u_i, u_i = (e_i - \delta\tau_i) \qquad (21)\\
e_i &= x_i^*\varepsilon_{1,i} + (1 - x_i^*)\varepsilon_{0,i}, x_i^* = (x_i - \tau_i)\\
e_i &= (x_i - \tau_i)\,\varepsilon_{1,i} + (1 - x_i + \tau_i)\varepsilon_{0,i}.
\end{aligned}
$$

Taking expectations, conditional on $(\mathbf{w}_i, x_i)$, Eq. (21) becomes

$$
E\left[y_i|\mathbf{w}_i, x_i\right] = \mathbf{z}_i\beta + \delta x_i + E\left[e_i|\mathbf{w}_i, x_i\right] - \delta E\left[\tau_i|\mathbf{w}_i, x_i\right],\; E\left[\tau_i|\mathbf{w}_i, x_i\right] = \xi
$$

$$
\therefore E\left[y_i|\mathbf{w}_i, x_i\right] = \mathbf{z}_i\beta + \delta\left(x_i - \xi\right) + E\left[e_i|\mathbf{w}_i, x_i\right]. \qquad (22)
$$

Given the result in **Theorem 2.1**, under the assumption of jointly normally distributed error terms, $E\left[e_i|\mathbf{w}_i, x_i\right]$ corresponds to

$$
E\left[e_i|\mathbf{w}_i, x_i\right] = \frac{\sigma_{e\eta}}{\sigma_\eta^2}\psi\left[\mu_i\left(\mathbf{w}_i\gamma\right)\right] \qquad (23)
$$

where

$$
\begin{aligned}
\psi &= \left[1 - \xi m_{xx}^{-1}\right]^{-1}, m_{xx} = (N)^{-1}\left(x'x\right) \qquad (24)\\
\xi &= \left(\lambda_1 + \lambda_2\right)\widehat{\pi}(1 - \widehat{\pi}), \widehat{\pi} = (N)^{-1}\sum_{i=1}^{N} x_i \qquad (25)\\
\psi &= \left[\frac{m_{xx}}{m_{xx} - \xi}\right] = \left(1 - \lambda_1 - \lambda_2\right)^{-1} \qquad (26)
\end{aligned}
$$

and $\mu_i\left(\mathbf{w}_i\gamma\right)$ denotes the modified generalised residual provided subsequently in Eq. (28).

Misclassification must be explicitly modelled since e_i in Eq. (21) is a function of τ_i and, consistent estimation of the reduced form parameters requires maximising

$$
\begin{aligned}
&\ln(L^m) = \sum_{i=1}^{N}\left\{x_i \ln\left[\lambda_2 + (1 - \lambda_1 - \lambda_2)\Phi\left(\mathbf{w}_i\gamma\right)\right]\right.\\
&\left. + \left(1 - x_i\right)\ln\left[1 - \lambda_2 - (1 - \lambda_1 - \lambda_2)\Phi\left(\mathbf{w}_i\gamma\right)\right]\right\}. \qquad (27)
\end{aligned}
$$

Setting $\frac{\partial \ln(L^m)}{\partial \gamma}$=0 we obtain the modified generalised residual, $\mu_i(\mathbf{w}_i\gamma)$, when $(\lambda_1 \neq 0, \lambda_2 \neq 0)$

$$\mu_i(\mathbf{w}_i\gamma) = \frac{\phi(\mathbf{w}_i\gamma)(1-\lambda_1-\lambda_2)[x_i - \lambda_2 - (1-\lambda_1-\lambda_2)\Phi(\mathbf{w}_i\gamma)]}{[\lambda_2 + (1-\lambda_1-\lambda_2)\Phi(\mathbf{w}_i\gamma)][1-\lambda_2-(1-\lambda_1-\lambda_2)\Phi(\mathbf{w}_i\gamma)]} \tag{28}$$

where assuming $\eta_i \sim N(0, \sigma_\eta^2)$, $\phi(.)$ is the *pdf* and $\Phi(.)$ the *cdf* of the Normal distribution, and $\mu_i(\mathbf{w}_i\gamma)$ corresponds to the conventional generalised probit residual, $\lambda_i(\mathbf{w}_i\gamma)$, only if $(\lambda_1 = \lambda_2 = 0)$.

Provided the monotonicity condition $\lambda_1 + \lambda_2 < 1$ holds, $(\gamma, \lambda_1, \lambda_2)$ can be estimated via the maximisation of the modified likelihood function in Eq. (27)- see Newey and McFadden (1994), Hausman et al. (1998). Reduced form estimation via MMLE in the cross-sectional case is treated in Section 4. Since all remaining components in Eqs. (24-26) can be computed from the underlying sample, we can construct the appropriate control function, $\psi[\mu_i(\mathbf{w}_i\gamma)]$, to be added in the structural form as an endogeneity correction term. Functional identification is achieved without any exclusion restrictions given the nonlinearity of the (modified) generalised residual. Structural identification requires that at least one element in $\mathbf{w}_i$ is not included in $\mathbf{z}_i$ (see Vella, 1998; Vella and Verbeek, 1998, 1999b; Verbeek and Nijman, 1992).

Theorem 3.1. *Ordinary Least Squares (OLS) estimation of*

$$y_i = \mathbf{z}_i\beta + \delta(x_i - \xi) + \left[\frac{\mu_i(\mathbf{w}_i\gamma)}{(1-\lambda_1-\lambda_2)}\right]\varkappa + \omega_i \tag{29}$$

produces consistent estimates of the structural form parameters $(\beta, \delta, \varkappa)$.

Summarising, the two-stage estimation procedure begins with the modified maximum likelihood estimation (MMLE) of the reduced form to obtain the misclassification probabilities (see Meyer and Mittag, 2017; Hausman *et al.*, 1998). Consistent estimation of the structural form parameters is obtained by adjusting for misclassification the structural form moment matrix of covariates and the control function in the second stage. Obtaining the analytical expression for the appropriate standard errors using the asymptotic covariance matrix given in Aigner (1973) and additionally accounting for the generated regressors along the lines of Newey (1984) is difficult. Bootstrapping standard errors over both estimation stages is an attractive alternative. Joint estimation of the two parts of the model via full-information MLE estimation (FIML) will be even more computationally intensive and, convergence problems due to collinearity are quite likely (see Puhani, 2000).

3.2 The Panel Data Case

When $T > 1$, resuming the model in Eq. (1), assuming that $E(\mathbf{z}_{it}'\alpha_{j,i})$=0, $E(\mathbf{z}_{it}'\varepsilon_{j,it})$=0 and that x_{it}^* is an endogenous binary explanatory variable the model is

$$\begin{aligned} y_{j,it} &= \mathbf{z}_{it}\beta_{j,t} + \delta x_{it}^* + e_{j,it} & (30) \\ e_{j,it} &= \alpha_{j,i} + \varepsilon_{j,it}, (t = 1,..,T; i = 1,...,N)\,,\ j = \{0,1\} \\ x_{it}^* &= \mathbf{1}\left\{\mathbf{w}_{it}\gamma + \nu_{it} \geq 0\right\}, \nu_{it} = \theta_i + \eta_{it}, (t = 1,..,T; i = 1,...,N) & (31) \\ j &= (0,1)\ if\ x_{it}^* = (0,1), y_{it} = y_{j,it}\ \ if\ \ x_{it}^* = j, (t = 1,..,T; i = 1,...,N)\,. \end{aligned}$$

Assuming invariance of $\beta_{j,t}$ with respect to (j,t) we can write

$$\begin{aligned} y_{it} &= \mathbf{z}_{it}\beta + \delta x_{it} + u_{it}, u_{it} = (e_{it} - \delta\tau_{it}) & (32) \\ e_{it} &= x_{it}^*(\alpha_{1,i} + \varepsilon_{1,it}) + (1 - x_{it}^*)(\alpha_{0,i} + \varepsilon_{0,it}), x_{it}^* = (x_{it} - \tau_{it}) \\ e_{it} &= (x_{it} - \tau_{it})\,(\alpha_{1,i} + \varepsilon_{1,it}) + (1 - x_{it} + \tau_{it})(\alpha_{0,i} + \varepsilon_{0,it}). \end{aligned}$$

Taking expectations, the linear projection of the structural equation is

$$\begin{aligned} E\left[y_{it}|\mathbf{w}_i, \mathbf{x}_i\right] &= \mathbf{z}_{it}\beta + \delta x_{it} + E(\alpha_{j,i}|\ \mathbf{w}_i, \mathbf{x}_i) + E(\varepsilon_{j,it}|\ \mathbf{w}_i, \mathbf{x}_i) - \delta E(\tau_{it}|\ \mathbf{w}_i, \mathbf{x}_i) \\ \mathbf{x}_i &= (\mathbf{x}_{i1}, ..., \mathbf{x}_{iT}), \mathbf{w}_i = (\mathbf{w}_{i1}, ..., \mathbf{w}_{iT}),\ E(\tau_{it}|\mathbf{w}_i, \mathbf{x}_i) = \xi \end{aligned}$$

$$\therefore E\left[y_{it}|\mathbf{w}_i, \mathbf{x}_i\right] = \mathbf{z}_{it}\beta + \delta\left(x_{it} - \xi\right) + E(\alpha_{j,i}|\mathbf{w}_i, \mathbf{x}_i) + E(\varepsilon_{j,it}|\ \mathbf{w}_i, \mathbf{x}_i). \quad (33)$$

Given the result in **Theorem 2.1** and the joint normality assumption, the conditional expectations of the error components are

$$E(\alpha_{j,i}|\ \mathbf{w}_i, \mathbf{x}_i) = \sigma_{j,\alpha\theta}\psi\left[\frac{T}{\sigma_\eta^2 + T\sigma_\theta^2}E(\bar{\nu_i}|\ \mathbf{w}_i, \mathbf{x}_i)\right] = \sigma_{j,\alpha\theta}\left[\psi B_i\right] \quad (34)$$

$$
\begin{aligned}
E(\varepsilon_{j,it}|\,\mathbf{w}_i,\mathbf{x}_i) &= \sigma_{j,\varepsilon\eta}\psi\left[\frac{E(\nu_{it}|\,\mathbf{w}_i,\mathbf{x}_i)}{\sigma_\eta^2}-\frac{T\sigma_\theta^2}{\sigma_\eta^2(\sigma_\eta^2+T\sigma_\theta^2)}E(\bar{\nu_i}|\,\mathbf{w}_i,\mathbf{x}_i)\right] \\
&= \sigma_{j,\varepsilon\eta}\left[\psi B_{it}\right] \qquad (35)
\end{aligned}
$$

where,

$$
\begin{aligned}
\psi &= \left[1-\xi m_{xx}^{-1}\right]^{-1}, m_{xx}=(NT)^{-1}\left(x'x\right), \bar{\nu}_i=T^{-1}\sum_{t=1}^{T}\nu_{it} \qquad (36) \\
\xi &= (\lambda_1+\lambda_2)\,\widehat{\pi}(1-\widehat{\pi}), \widehat{\pi}=(NT)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}x_{it} \qquad (37) \\
\psi &= \left[\frac{m_{xx}}{m_{xx}-\xi}\right]=(1-\lambda_1-\lambda_2)^{-1} \qquad (38)
\end{aligned}
$$

The conditional expectations in Eqs. (34,35) are obtained using the standard formulae for the conditional expectation of normally distributed vectors (see Vella and Verbeek, 1998; Verbeek and Nijman, 1992). Computing the conditional expectations, in Eqs. (34,35) requires estimation of the reduced form parameters and $E(\nu_{it}|\mathbf{w}_i,\mathbf{x}_i)$ which under the normality assumption corresponds to the generalised residual of the random effects probit model and has to be appropriately modified to account for misclassification error.[3]

Consistent estimates of the structural model parameters can be obtained from conditional moment restrictions such as least squares based on the estimable version of Eq. (33) by including the endogeneity correction terms defined in Eqs. (34,35) as additional regressors. Provided the monotonicity condition $\lambda_1+\lambda_2<1$ holds, using the appropriate modified likelihood function we can estimate $\left(\gamma,\lambda_1,\lambda_2,\sigma_\theta^2\right)$ and compute (B_i, B_{it}), in Eqs. (34,35), which have to be interacted by ψ - see Newey and McFadden (1994), Hausman et al. (1998). Reduced form estimation via MMLE is treated in Section 4.

Theorem 3.2. *Conditional moment restrictions such as Least Squares on*

$$
y_{it}=\mathbf{z}_{it}\beta+\delta\left(x_{it}-\xi\right)+\sigma_{j,\alpha\theta}\left[\frac{B_i}{(1-\lambda_1-\lambda_2)}\right]+\sigma_{j,\varepsilon\eta}\left[\frac{B_{it}}{(1-\lambda_1-\lambda_2)}\right]+\omega_{it} \qquad (39)
$$

[3]For a detailed discussion regarding parametric estimation in the presence of selectivity bias and endogeneity in panel data models refer to Vella (1998), Vella and Verbeek (1998, 1999b), Verbeek and Nijman (1992). Vella (1998) also treats cross-sectional estimation.

gives consistent estimates of the structural form parameters $(\beta, \delta, \sigma_{j,\alpha\theta}, \sigma_{j,\varepsilon\eta})$.

As in the cross-sectional case, functional identification is achieved without any exclusion restrictions given the nonlinearity of the (modified) generalised residual. Structural identification requires that at least one element in $\mathbf{w}_{it}$ is not included in $\mathbf{z}_{it}$ (see Vella, 1998; Vella and Verbeek, 1998, 1999b; Verbeek and Nijman, 1992).

Implementing the two-stage procedure involves MMLE estimation of the reduced form likelihood function, as well as, calculation of $E(\nu_{it}|\mathbf{w}_i, \mathbf{x}_i)$ which requires numerical integration. Note that, the first-stage panel MMLE also requires numerical integration to integrate out the time-invariant unobserved error component. The appropriate standard errors accounting for the generated endogeneity correction terms from the first-stage are given in Vella and Verbeek (1999b, pp. 259-61). Note that these should also be adjusted using the asymptotic covariance matrix in Aigner (1973) to account for the measurement error issue. Alternatively, one could bootstrap standard errors over both estimation stages.

To obtain the expression of $E(\nu_{it}|\mathbf{w}_i, \mathbf{x}_i)$ start by using the first law of iterated expectations

$$\begin{aligned} E(\nu_{it}|\mathbf{w}_i, \mathbf{x}_i) &= E(\theta_i + \eta_{it}|\mathbf{w}_i, \mathbf{x}_i) = E_{\theta_i}[\theta_i + E(\eta_{it}|\mathbf{w}_i, \mathbf{x}_i, \theta_i)] \\ &= \int_{-\infty}^{+\infty} [\theta_i + E(\eta_{it}|\mathbf{w}_i, \mathbf{x}_i, \theta_i)] f(\theta_i|\mathbf{w}_i, \mathbf{x}_i) d\theta_i \end{aligned} \tag{40}$$

where $f(\theta_i|\mathbf{w}_i, \mathbf{x}_i)$ is the conditional density of θ_i which, assuming exogeneity of $\mathbf{w}_i$,corresponds to

$$f(\theta_i|\mathbf{w}_i, \mathbf{x}_i) = \frac{f(\mathbf{x}_i|\mathbf{w}_i, \theta_i) f(\theta_i)}{f(\mathbf{x}_i|\mathbf{w}_i)}. \tag{41}$$

Substituting Eq. (41) into Eq. (40) we get

$$E(\nu_{it}|\mathbf{w}_i, \mathbf{x}_i) = \frac{1}{f(\mathbf{x}_i|\mathbf{w}_i)} \int [\theta_i + E(\eta_{it}|\mathbf{w}_i, \mathbf{x}_i, \theta_i)] f(\mathbf{x}_i|\mathbf{w}_i, \theta_i) f(\theta_i) d\theta_i$$

and integrating out θ_i in the denominator *i.e.* $f(\mathbf{x}_i|\mathbf{w}_i) = f(\mathbf{x}_i|\mathbf{w}_i, \theta_i) f(\theta_i, \mathbf{w}_i) d\theta_i$, we obtain

$$E(\nu_{it}|\mathbf{w}_i, \mathbf{x}_i) = \frac{\int [\theta_i + E(\eta_{it}|x_{it}, U_i, \theta_i)] f(\mathbf{x}_i|\mathbf{w}_i, \theta_i) f(\theta_i) d\theta_i}{f(\mathbf{x}_i|\mathbf{w}_i, \theta_i) f(\theta_i, \mathbf{w}_i) d\theta_i} \tag{42}$$

The expression in eq. (42), can be approximated by either quadrature methods or simulation noting that

$$E(x_{it}^*|\mathbf{w}_{it},\theta_i) = \Pr(x_{it}^* = 1|\ \mathbf{w}_{it},\theta_i) = \lambda_2 + (1-\lambda_1-\lambda_2)\left[\Phi\left(\frac{\mathbf{w}_{it}\gamma+\theta_i}{\sigma_\eta}\right)\right]$$

giving

$$\begin{aligned} f(\mathbf{x}_i|\mathbf{w}_i,\theta_i) &= \prod_{t=1}^{T}\left\{x_{it}\left[\lambda_2 + (1-\lambda_1-\lambda_2)\Phi\left(\frac{\mathbf{w}_{it}\gamma+\theta_i}{\sigma_\eta}\right)\right]\right. \\ &\quad \left. +(1-x_{it})\left[1-\lambda_2-(1-\lambda_1-\lambda_2)\Phi\left(\frac{\mathbf{w}_{it}\gamma+\theta_i}{\sigma_\eta}\right)\right]\right\}. \end{aligned} \tag{43}$$

Finally, $E(\eta_{it}|\mathbf{w}_i,\mathbf{x}_i,\theta_i)$ denotes the generalised probit residual in the cross-sectional sense which under misclassification corresponds to

$$E(\eta_{it}|x_{it},U_i,\theta_i) =$$

$$= \frac{\phi\left(\mathbf{w}_{it}\gamma+\theta_i\right)(1-\lambda_1-\lambda_2)\sigma_\eta\left[x_{it}-\lambda_2-(1-\lambda_1-\lambda_2)\Phi\left(\frac{\mathbf{w}_{it}\gamma+\theta_i}{\sigma_\eta}\right)\right]}{\left[\lambda_2+(1-\lambda_1-\lambda_2)\Phi\left(\frac{\mathbf{w}_{it}\gamma+\theta_i}{\sigma_\eta}\right)\right]\left[1-\lambda_2-(1-\lambda_1-\lambda_2)\Phi\left(\frac{\mathbf{w}_{it}\gamma+\theta_i}{\sigma_\eta}\right)\right]}. \tag{44}$$

4 Concluding Remarks

I propose cross-sectional and longitudinal estimators of binary treatment effects in the presence of a misclassified endogenous binary treatment variable. The two-step parametric estimators do not rely on the availability of instruments (or additional measurements) and, functional identification is achieved without any exclusion restrictions given the nonlinearity of the generalised residuals.

Summarising the problem, within a parametric two-step framework: measurement error in an endogenous binary explanatory variable produces biased and inconsistent parameter estimates in both estimation stages. The structural equation includes a binary explanatory variable that is both misclassified and endogenous. The respective reduced form, used to generate the appropriate endogeneity correction terms, corresponds to a binary choice model where the dependent variable is subject to misclassification producing inconsistent parameter estimates.

Unless misclassification is appropriately modelled in the reduced form, and the correlation between the mismeasured binary variable and the measurement error in the structural equation purged, the usual two-stage endogeneity correction fails.

Given the modified MLE parameter estimates of the misclassification probabilities, we obtain the appropriate endogeneity correction terms which have to be divided by a factor of one minus the sum of the misclassification probabilities *i.e.* $(1 - \lambda_1 - \lambda_2)$. If the monotonicity condition $\lambda_1 + \lambda_2 < 1$ holds, the reduced form parameters can be consistently estimated via MLE while, the structural equation augmented by the addition of the corresponding endogeneity correction terms, can be estimated using conditional moment restrictions such as least squares.

A Appendix: Proofs - Consistency of the Modified Least Squares (MLS) Estimator

Consider the partitioned linear model where the LS normal equations become

$$\begin{bmatrix} M_{ZZ} & M_{Zx} \\ M'_{Zx} & (m_{xx}-\xi) \end{bmatrix} \begin{bmatrix} \beta \\ \delta \end{bmatrix} = \begin{bmatrix} M_{Zy} \\ m_{xy} \end{bmatrix} \tag{A.1}$$

giving

$$M_{Zy} = M_{ZZ}\beta + M_{Zx}\delta \tag{A.2}$$

and

$$m_{xy} = M'_{Zx}\beta + (m_{xx}-\xi)\,\delta. \tag{A.3}$$

Using Eq. (A.3) we obtain

$$\delta = \left[m_{xx}^{-1}m_{xy} - m_{xx}^{-1}M'_{Zx}\beta\right]\left[1-\xi m_{xx}^{-1}\right]^{-1} \tag{A.4}$$

and substituting in Eq. (A.2) we arrive at [4]

$$\begin{aligned} M_{Zy} &= (NT)^{-1} Z' \left[I - x m_{xx}^{-1} (NT)^{-1} x' \left(1-\xi m_{xx}^{-1}\right)^{-1}\right] Z\beta \\ &\quad + M_{Zx} m_{xx}^{-1} m_{xy} \left[1-\xi m_{xx}^{-1}\right]^{-1}. \end{aligned} \tag{A.5}$$

Following Aigner (1973), setting $A = \left[I - x m_{xx}^{-1} (NT)^{-1} x' \left(1-\xi m_{xx}^{-1}\right)^{-1}\right]$ in Eq. (A.5) and rearranging we get $Z'Ay = (Z'AZ)\,\beta$ which yields

$$\widehat{\beta} = (Z'AZ)^{-1}Z'Ay. \tag{A.6}$$

Substitution of Eq. (A.6) in Eq. (A.4) gives

$$\widehat{\delta} = (NT)^{-1} m_{xx}^{-1} x' \left[I - Z(Z'AZ)^{-1}Z'A\right] y \left[1-\xi m_{xx}^{-1}\right]^{-1}. \tag{A.7}$$

[4] The corresponding cross-sectional expressions are simply obtained by setting $T = 1$.

To prove consistency of the MLS estimate of the $\underset{1x(k-1)}{\beta}$ vector of parameters, we introduce Eq. (5) into Eq. (A.6) to get

$$\widehat{\beta} = (Z'AZ)^{-1}Z'A\left[Z\beta + x^{*}\delta + e\right] \tag{A.8}$$

giving

$$\widehat{\beta} = \beta + (Z'AZ)^{-1}Z'Ae + (Z'AZ)^{-1}Z'Ax^{*}\delta. \tag{A.9}$$

Taking probabity limits, noting that in both the cross-sectional case with $T = 1$ and in micropanels with short-T and large-N the asymptotics rely on N

$$\underset{N\longrightarrow\infty}{plim}\ \widehat{\beta} = \beta + \underset{N\longrightarrow\infty}{plim}\ (\frac{1}{NT}Z'AZ)^{-1}\underset{N\longrightarrow\infty}{plim}\ (\frac{1}{NT}Z'Ax^{*}\delta) \tag{A.10}$$

under the exogeneity assumptions, $E(\mathbf{z}'_{it}\alpha_{j,i})$=0 and $E(\mathbf{z}'_{it}\varepsilon_{j,it})$=0, ensuring that $\underset{N\longrightarrow\infty}{plim}\ (\frac{1}{NT}Z'Ae) = 0$.

Noting that $Ax^{*} = x^{*} - \left(1-\xi m_{xx}^{-1}\right)^{-1}x + \left(1-\xi m_{xx}^{-1}\right)^{-1}xm_{xx}^{-1}\left(\frac{1}{NT}x'\tau\right)$, where τ=$(x - x^{*})$, we get

$$Ax^{*} = x^{*} - x\left[1 - m_{xx}^{-1}\left(\frac{1}{NT}x'\tau\right)\right]\left[\left(1-\xi m_{xx}^{-1}\right)^{-1}\right].$$

Therefore,

$$\begin{aligned}\underset{N\longrightarrow\infty}{plim}\left(\frac{1}{NT}Z'Ax^{*}\right) &= \underset{N\longrightarrow\infty}{plim}\left(\frac{1}{NT}Z'x^{*}\right)\\ &\quad -\underset{N\longrightarrow\infty}{plim}\frac{1}{NT}Z'x\left[1 - m_{xx}^{-1}\left(\frac{1}{NT}x'\tau\right)\right]\left[\left(1-\xi m_{xx}^{-1}\right)^{-1}\right]\end{aligned}$$

and if $\underset{N\longrightarrow\infty}{plim}\left[1 - m_{xx}^{-1}\left(\frac{1}{NT}x'\tau\right)\right]\left[\left(1-\xi m_{xx}^{-1}\right)^{-1}\right] = 1$ since $\underset{N\longrightarrow\infty}{plim}\left(\frac{1}{NT}x'\tau\right) = \xi$, $\underset{N\longrightarrow\infty}{plim}\left(\frac{1}{NT}Z'x^{*}\right) = \underset{N\longrightarrow\infty}{plim}\left(\frac{1}{NT}Z'x\right)$ and hence, $\underset{N\longrightarrow\infty}{plim}\left(\frac{1}{NT}Z'Ax^{*}\right) = 0$ thus, proving $\underset{N\longrightarrow\infty}{plim}\ \widehat{\beta} = \beta$.

Turning to the consistency of the MLS estimate of δ, introduce Eq.(5) in Eq.(A.7) to get

$$\widehat{\delta} = (NT)^{-1}m_{xx}^{-1}x'\left[I - Z(Z'AZ)^{-1}Z'A\right]\left[Z\beta + x^{*}\delta + e\right]\left[1-\xi m_{xx}^{-1}\right]^{-1} \tag{A.11}$$

giving

$$\widehat{\delta} = (NT)^{-1} m_{xx}^{-1} x' \delta \left[x^* - Z(Z'AZ)^{-1} Z'Ax^*\right] \left[1 - \xi m_{xx}^{-1}\right]^{-1} + (NT)^{-1} m_{xx}^{-1} x' \left[e - Z(Z'AZ)^{-1} Z'Ae\right] \left[1 - \xi m_{xx}^{-1}\right]^{-1}. \tag{A.12}$$

Taking probability limits using $\underset{N \longrightarrow \infty}{plim} \left(\frac{1}{NT} Z'Ax^*\right)$=0 and $\underset{N \longrightarrow \infty}{plim} \left(\frac{1}{NT} Z'Ae\right)$=0, given the exogeneity assumptions $E(\mathbf{z}'_{it}\alpha_{j,i}) = 0$ and $E(\mathbf{z}'_{it}\varepsilon_{j,it}) = 0$, we get

$$\underset{N \longrightarrow \infty}{plim} \widehat{\delta} = \delta \underset{N \longrightarrow \infty}{plim} \left[m_{xx}^{-1} \left(\frac{1}{NT} x'x^*\right) \left[1 - \xi m_{xx}^{-1}\right]^{-1}\right] + \underset{N \longrightarrow \infty}{plim} \left[m_{xx}^{-1} \left(\frac{1}{NT} x'e\right) \left[1 - \xi m_{xx}^{-1}\right]^{-1}\right]$$

and using τ=$(x - x^*)$ we arrive at

$$\underset{N \longrightarrow \infty}{plim} \widehat{\delta} = \delta \underset{N \longrightarrow \infty}{plim} \left[\left[1 - m_{xx}^{-1} \left(\frac{1}{NT} x'\tau\right)\right] \left[1 - \xi m_{xx}^{-1}\right]^{-1}\right] + \underset{N \longrightarrow \infty}{plim} \left[m_{xx}^{-1} \left(\frac{1}{NT} x'e\right) \left[1 - \xi m_{xx}^{-1}\right]^{-1}\right] \tag{A.13}$$

which given that $\underset{N \longrightarrow \infty}{plim} \left(\frac{1}{NT} x'\tau\right) = \xi$ simplifies to

$$\underset{N \longrightarrow \infty}{plim} \widehat{\delta} = \delta + \underset{N \longrightarrow \infty}{plim} \left[m_{xx}^{-1} \left(\frac{1}{NT} x'e\right) \left[1 - \xi m_{xx}^{-1}\right]^{-1}\right] \tag{A.14}$$

corresponding to

$$\underset{N \longrightarrow \infty}{plim} \widehat{\delta} = \delta + \underset{N \longrightarrow \infty}{plim} \left[(x'x)^{-1} (x'e) \psi\right], \psi = \left[1 - \xi m_{xx}^{-1}\right]^{-1}. \tag{A.15}$$

References

[1] Aigner, D.J., (1973). 'Regression with a Binary Independent Variable Subject to Errors of Observation', *Journal of Econometrics*, 1, pp. 49-60.

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

[3] Butler, J. and Moffitt, R. (1982). 'A Computationally Efficient Quadrature Procedure for the One Factor Multinomial Probit Model', *Econometrica*, 50, pp.761 -764.

[4] Carroll, R.J., Chen, X., Hu, Y. (2010) 'Identification and Estimation of Nonlinear Models Using Two Samples with Nonclassical Measurement Errors', *Journal of Nonparametric Statistics*, 22(4), pp.379-399.

[5] Cochran, W.G. (1968).'Errors of Measurement in Statistics', *Technometrics*, 10, pp. 637-666.

[6] Fernández-Val, I. and Vella, F. (2011). 'Bias Corrections for Two-Step Fixed Effects Panel Data Estimators', *Journal of Econometrics*, 163, pp. 144-162.

[7] Frazis, H. and Loewenstein, M. A., (2003) 'Estimating Linear Regressions with Mismeasured, Possibly Endogenous, Binary Explanatory Variables', *Journal of Econometrics*, 117, pp. 151-178.

[8] Gourieroux, C. (2000). 'Econometrics of Qualitative Dependent Variables', *Themes in Modern Econometrics,* Cambridge University Press, 2nd ed.

[9] Hausman, J.A. and Scott-Morton, F. M., (1994). 'Misclassification of a Dependent Variable in a Discrete-Response Setting', Working papers 94-19, *Massachusetts Institute of Technology* (MIT), Department of Economics.

[10] Hausman, J.A., Abrevaya, J., and Scott-Morton, F. M., (1998). 'Misclassification of the Dependent Variable in a Discrete-Response Setting', *Journal of Econometrics*, 87, pp. 239-269.

[11] Heckman, J.J. (1979). 'Sample Selection Bias as a Specification Error', *Econometrica*, 47, pp. 153-161.

[12] Heckman, J.J. (1981). 'The Incidental Parameters Problem and the Problem of Initial Conditions in Estimating a Discrete Time-Discrete Data Stochastic Process', in Manski, C. F. and McFadden, D. L., (eds.),' *Structural Analysis of Discrete Data and Econometric Applications*', Cambridge: The MIT Press.

[13] Haider, S., Stephens, M. (2020). 'Correcting for Misclassied Binary Regressors Using Instrumental Variables', IZA, Discussion Paper 13593.

[14] Hu, Y., Shiu, J. and Woutersen, T. (2015). 'Identification and Estimation of Single-Index Models with Measurement Error and Endogeneity', *The Econometrics Journal*, 18, pp. 347-362.

[15] Horrowitz, J. and Manski, C. (1995). 'Identification and Robustness with Contaminated and Corrupted Data', *Econometrica*, 63, pp. 281-302.

[16] Hsiao, C. (2003). 'Analysis of Panel Data', *Econometric Society Monographs*, Camrbidge University Press, 2nd ed.

[17] Kasahara, H., Shimotsu, K. (2021). 'Identification of Regression Models with a Missclassified and Endogenous Binary Regressor', *Econometric Theory*, pp. 1-23.

[18] Keane M.P., and Sauer, R.M., (2009). 'Classification Error in Dynamic Discrete Choice Models: Implications for Female Labour Supply Behaviour', *Econometrica*, 77, pp. 975-991.

[19] Keane M.P., and Sauer, R.M., (2010). 'A Computationally Practical Simulation Estimation Algorithm for Dynamic Panel Data Models with Unobserved Endogenous State Variables', *International Economic Review*, 51(4), pp. 925-958.

[20] Lewbel, A., (2007). 'Estimation of average treatment effects with misclassification', *Econometrica*, 75(2), 537-551.

[21] Meyer, B. and Mittag, N., (2017). 'Misclassification in Binary Choice Models. *Journal of Econometrics*, 200(2), pp. 295-311.

[22] Mahajan, A. (2006). 'Identification and Estimation of Regression Models with Misclassification', *Econometrica*, 74, pp. 631-665.

[23] Newey, W.K. (1984). 'A Method of Moments Interpretation of Sequential Estimators', *Economics Letters*, 14, pp. 201-206.

[24] Newey, W.K. and McFadden, D.L., (1994). 'Large Sample Estimation and Hypothesis Testing'. In: Engle, R. F., McFadden, D. L. (Eds.), *Handbook of Econometrics*, vol. 4. North-Holland, Amsterdam.

[25] Nguimkeu, P., Denteh, A., and Tchernis, R. (2019). 'On the Estimation of Treatment Effects with Endogenous Misreporting', *Journal of Econometrics*, 208, pp. 487-506.

[26] Puhani, P. (2000). 'The Heckman Correction for Sample Selection and Its Critique', *Journal of Economic Surveys*, 14(1), pp. 53-68.

[27] Vella, F., (1998). 'Estimating Models with Sample Selection Bias: A Survey', *Journal of Human Resources*, 33, pp. 127-169.

[28] Vella, F. and Verbeek, M. (1998). 'Whose Wages Do Unions Raise? A Dynamic Model of Unionism and Wage Rate Determination for Young Men', *Journal of Applied Econometrics*, 13, pp. 163-183.

[29] Vella, F. and Verbeek, M. (1999a). 'Estimating and Interpreting Models with Endogenous Treatment Effects', *Journal of Business & Economic Statistics*, 17, pp. 473-478.

[30] Vella, F. and Verbeek, M. (1999b). 'Two-step Estimation of Panel Data Models with Censored Endogenous Variables and Selection Bias', *Journal of Econometrics*, 90, pp. 239-263.

[31] Verbeek, M. and Nijman, T. (1992). 'Testing for Selectivity Bias in Panel Data Models', *International Economic Review*, 33, pp. 681-703.

[32] Wooldridge, J.M. (2002). '*Econometric Analysis of Cross Section and Panel Data*', Cambridge: MIT Press.

[33] Xiaohong, C., Hong, H. and Nekipelov, D. (2011). 'Nonlinear Models of Measurement Errors', *Journal of Economic Literature*, 49, pp. 901-937.

5 Supplementary Material: Modelling Misclassification in Binary Discrete Choice Models

The longitudinal binary choice model in the absence of misclassification is

$$x_{it}^{*} = \mathbf{1}\left\{\mathbf{w}_{it}\gamma + \nu_{it} \geq 0\right\}, \nu_{it} = \theta_i + \eta_{it}, (t = 1, .., T; i = 1, ..., N) \tag{1}$$

where η_{it} is an *iid* idiosyncratic error disturbance with *cdf* F conditional on θ_i and $\mathbf{w}_i = (\mathbf{w}_{i1}, ..., \mathbf{w}_{iT})$. Assuming that θ_i, *i.e.* the unobserved individual-specific time-invariant heterogeneity, is additive inside the *cdf* and strict exogeneity of $\mathbf{w}_{it}$ (conditional on θ_i) we can write

$$Pr(x_{it} = 1|\mathbf{w}_i, \theta_i) = Pr(x_{it} = 1|\mathbf{w}_{it}, \theta_i) = F(\mathbf{w}_{it}\gamma + \theta_i),\ t = 1, .., T. \tag{2}$$

Further assuming that $(x_{i1}, ..., x_{iT})$ are independent conditional on $(\mathbf{w}_i, \theta_i)$, under the strict conditional exogeneity of $\mathbf{w}_{it}$, we can derive the density of $(x_{i1}, ..., x_{iT})$ conditional on $(\mathbf{w}_i, \theta_i)$

$$f(x_1, ..., x_T|\mathbf{w}_i, \theta_i; \gamma) = \prod_{t=1}^{T} f(x_t|\mathbf{w}_{it}, \theta_i; \gamma) \tag{3}$$

where

$$f(x_t|\mathbf{w}_t, \theta; \gamma) = F(\mathbf{w}_t\gamma + \theta)^{x_t} \left[1 - F(\mathbf{w}_t\gamma + \theta)\right]^{1-x_t}. \tag{4}$$

Let (x_{it}^{*}, x_{it}) be the true/observed indicators for the outcome of individual i in period t, respectively where x_{it} is subject to misclassification error. Denote the probability of a false positive/negative by (λ_1,λ_2) conditional on the observed binary indicator x_{it}. It is assumed that the misclassification probabilities are constant across individuals and that $\lambda_1 \neq \lambda_2$. Additionally, for tractability, (λ_1, λ_2) are assumed to be time-invariant so that

$$\lambda_1 = \Pr(x_{it}^{*} = 0|x_{it} = 1), \lambda_2 = \Pr(x_{it}^{*} = 1|x_{it} = 0) \tag{5}$$

noting that (λ_1,λ_2) depend on x_{it} and are independent of $\mathbf{w}_{it}$ and ν_{it}. The expected value of the true response is

$$\pi = E(x_{it}^{*}|\,\mathbf{w}_{it}) = \Pr(x_{it}^{*} = 1|\,\mathbf{w}_{it}) = (1-\lambda_1)F(\mathbf{w}_{it}\gamma + \theta_i) + \lambda_2\left[1 - F(\mathbf{w}_{it}\gamma + \theta_i)\right] \tag{6}$$

giving

$$\pi = E(x_{it}^{*}|\,\mathbf{w}_{it}) = \Pr(x_{it}^{*} = 1|\,\mathbf{w}_{it}) = F(\mathbf{w}_{it}\gamma + \theta_i)(1-\lambda_1-\lambda_2) + \lambda_2 \tag{7}$$

indicating that $E(x_{it}^{*}|\mathbf{w}_{it}) = F(\mathbf{w}_{it}\gamma + \theta_i)$ only if (λ_1=λ_2=0).

Alternatively, the expected value of the true response, π, is related to the expected value of the observed response, $\widehat{\pi}$, by

$$\pi = (1-\lambda_1)\,\widehat{\pi} + \lambda_2(1-\widehat{\pi}) = \widehat{\pi}(1-\lambda_1-\lambda_2) + \lambda_2, \widehat{\pi} = (NT)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T} x_{it}. \tag{8}$$

Given knowledge of F, *e.g.* normal in econometrics or logistic in (bio)statistics, and assuming that θ_i had been observed, we could parametrically estimate $(\gamma, \lambda_1, \lambda_2)$ by maximising the following (modified) log-likelihood function

$$\begin{aligned} \ln(L^m) \;\; = \;\; & \sum_{i=1}^{N}\sum_{t=1}^{T}\{x_{it}\ln\left[\lambda_2 + (1-\lambda_1-\lambda_2)F(\mathbf{w}_{it}\gamma + \theta_i)\right] \\ & + (1-x_{it})\ln\left[1-\lambda_2-(1-\lambda_1-\lambda_2)F(\mathbf{w}_{it}\gamma + \theta_i)\right]\} \end{aligned} \tag{9}$$

over $(\gamma, \lambda_1, \lambda_2)$ provided that the monotonicity condition $\lambda_1 + \lambda_2 < 1$ holds noting that parameter identification stems from the nonlinearity of F.[5]

If $\lambda_1 + \lambda_2 = 1$, denoting $\widetilde{\lambda}_1 = 1 - \lambda_2$, $\widetilde{\lambda}_2 = 1 - \lambda_1$, $\widetilde{\gamma} = -\gamma$ we are unable to distinguish between $\left(\widetilde{\lambda}_1, \widetilde{\lambda}_2, \widetilde{\gamma}\right)$ and $(\lambda_1, \lambda_2, \gamma)$ since $F(\mathbf{w}_{it}\widetilde{\gamma} + \theta_i)(1-\widetilde{\lambda}_1-\widetilde{\lambda_2}) + \widetilde{\lambda_2} = F(\mathbf{w}_{it}\gamma + \theta_i)(1-\lambda_1-\lambda_2) + \lambda_2$, (see Hausman *et al.*, 1998, pp. 242).

When $\lambda_1 + \lambda_2 > 1$ in practice on average one could not tell which outcome occurred and the appli-

[5]In the linear probability model the parameters cannot be separately identified as

$$\begin{aligned} E(x_{it}^{*}|\mathbf{w}_{it}) \;\; &= \;\; \mathbf{w}_{it}\gamma(1-\lambda_1-\lambda_2) + \lambda_2,\; \mathbf{w}_{it} = (1, \mathbf{r}_{it}),\; \gamma = (\gamma_0, \gamma_1) \\ &= \;\; \mathbf{r}_{it}\gamma_1(1-\lambda_1-\lambda_2) + (\gamma_0 + \lambda_2)\,. \end{aligned}$$

cation should not be pursued while, imposition of the monotonicity condition will produce estimates of γ and consequently marginal effects (of continuous covariates) of the opposite sign.

Regarding the continuous elements of $\mathbf{w}$ denoted by $\mathbf{w_c}$ the marginal effect of $w_{cj}, \ j = 1, ..., k_c$, on the true and observed responses are correspondingly

$$\frac{\partial \Pr(x^* = 1|\mathbf{w})}{\partial w_{cj}} = f\left(\mathbf{w}_{it}\gamma + \theta_i\right)\gamma_j \neq \frac{\partial \Pr(x = 1|\mathbf{w})}{\partial w_{cj}} = (1 - \lambda_1 - \lambda_2) f\left(\mathbf{w}_{it}\gamma + \theta_i\right)\gamma_j \tag{10}$$

while for the discrete elements of $\mathbf{w}$ denoted by $\mathbf{w_d}$ the marginal effect of w_{dl}, $l = 1, ..., k_l$, expressed using counter-factual probabilities is

$$\begin{aligned} &\Pr(x^* = 1|\mathbf{w}, w_{dl} = 1) - \Pr(x^* = 1|\mathbf{w}, w_{dl} = 0) = F(\mathbf{w}_{it}\gamma + \gamma_l + \theta_i) - F(\mathbf{w}_{it}\gamma + \theta_i) \neq \\ &\Pr\left(x = 1|\mathbf{w}, w_{dl} = 1\right) - \Pr\left(x = 1|\mathbf{w}, w_{dl} = 0\right) = \frac{\left[F\left(\mathbf{w}_{it}\gamma + \gamma_l + \theta_i\right) - F\left(\mathbf{w}_{it}\gamma + \theta_i\right)\right]}{\left(1 - \lambda_1 - \lambda_2\right)^{-1}} \end{aligned} \tag{11}$$

where f denotes the pdf of F. Clearly, the corresponding marginal effect on the observed response will be attenuated by a factor of $(1 - \lambda_1 - \lambda_2)$ while, if $\lambda_1 + \lambda_2 > 1$ imposing the monotonicity condition produces wrongly signed marginal effects.

The monotonicity condition guarantees $\lambda_2 + F(\mathbf{w}_{it}\gamma + \theta_i)(1 - \lambda_1 - \lambda_2)$ is strictly increasing in $(\mathbf{w}_{it}\gamma + \theta_i)$ if η_{it} has a positive density everywhere so that F is strictly increasing. The results in Newey and McFadden (1994, pp. 2125) can be extended to the panel data case since in the case of a normal *cdf* if $E(\mathbf{w}'\mathbf{w})$ exists and is nonsingular, provided $\lambda_1 + \lambda_2 < 1$ then $(\lambda_1, \lambda_2, \gamma)$ can be identified using MLE which is also efficient if F is known.

Ignoring misclassification and maximising the likelihood setting $(\lambda_1 = \lambda_2 = 0)$ produces biased and inconsistent parameter estimates if the misclassification rates are non-zero and statistically significant. However, the Fisher information matrix for the modified likelihood in Eq. (9) is not block diagonal in the presence of misclassification error and remains non-block diagonal even when $(\lambda_1 = \lambda_2 = 0)$. Therefore, the modified likelihood produces less efficient estimates of γ and one should use the uncorrected likelihood upon obtaining statistically insignificant estimates of (λ_1, λ_2)- for the corresponding information matrices refer to Hausman *et al.*, (1998, p. 244).

Differentiating $\ln(L^m)$ with respect to the parameter vector γ, yields the (score) vector of first derivatives

$$\frac{\partial \ln(L^m)}{\partial \gamma} = \sum_{i=1}^{N}\sum_{t=1}^{T}\left\{\left[\frac{x_{it}\mathbf{w}_{it}f(\mathbf{w}_{it}\gamma+\theta_i)(1-\lambda_1-\lambda_2)}{\lambda_2+(1-\lambda_1-\lambda_2)F(\mathbf{w}_{it}\gamma+\theta_i)}\right] - \left[\frac{(1-x_{it})\mathbf{w}_{it}f(\mathbf{w}_{it}\gamma+\theta_i)(1-\lambda_1-\lambda_2)}{1-\lambda_2-(1-\lambda_1-\lambda_2)F(\mathbf{w}_{it}\gamma+\theta_i)}\right]\right\} \tag{12}$$

where f is the *pdf* corresponding to F and by rearrangement the score vector, providing the orthogonality conditions under misclassification, becomes

$$\frac{\partial \ln(L^m)}{\partial \gamma} = \sum_{i=1}^{N}\sum_{t=1}^{T}\left[\frac{(x_{it}-\zeta_{it})}{\zeta_{it}(1-\zeta_{it})}f(\mathbf{w}_{it}\gamma+\theta_i)(1-\lambda_1-\lambda_2)\mathbf{w}_{it}\right] \tag{13}$$

where ζ_{it}=$\lambda_2+(1-\lambda_1-\lambda_2)F(\mathbf{w}_{it}\gamma+\theta_i)$. Setting $\frac{\partial \ln(L^m)}{\partial \gamma}$=0, we obtain the generalised residual in the presence of misclassification

$$\frac{(x_{it}-\zeta_{it})}{\zeta_{it}(1-\zeta_{it})}f(\mathbf{w}_{it}\gamma+\theta_i)(1-\lambda_1-\lambda_2) \tag{14}$$

that corresponds to the conventional generalised residual only if (λ_1=λ_2=0).

The inconsistency produced by misclassification can be clearly seen by noting that the MLE first-order condition for a maximum sets the score equal to zero under the assumption of no misclassification. Maximising the likelihood under the assumption that $E(x_{it}^*|\mathbf{w}_{it},\lambda_1=\lambda_2=0)=F(\mathbf{w_{it}}\gamma+\theta_i)$, and setting $\frac{\partial \ln(L)}{\partial \gamma}$=0 produces a zero expectation for the score since

$$E\left[\frac{\partial \ln(L)}{\partial \gamma}\right] = \sum_{i=1}^{N}\sum_{t=1}^{T}\left\{\frac{\mathbf{w}_{it}f(\mathbf{w}_{it}\gamma+\theta_i)}{F(\mathbf{w}_{it}\gamma+\theta_i)}\left[F(\mathbf{w}_{it}\gamma+\theta_i)\right]\right\} - \sum_{i=1}^{N}\sum_{t=1}^{T}\left\{\frac{\mathbf{w}_{it}f(\mathbf{w}_{it}\gamma+\theta_i)}{1-F(\mathbf{w}_{it}\gamma+\theta_i)}\left[1-F(\mathbf{w}_{it}\gamma+\theta_i)\right]\right\}=0 \tag{15}$$

while if ($\lambda_1\neq 0,\lambda_2\neq 0$) then $E(x_{it}^*|\mathbf{w}_{it})=F(\mathbf{w_{it}}\gamma+\theta_i)(1-\lambda_1-\lambda_2)+\lambda_2$ and the expected value of the uncorrected score, $E\left[\frac{\partial \ln(L)}{\partial \gamma}\right]$, will generally not be equal to zero. Appropriate modification of the likelihood function yields

$$E\left[\frac{\partial \ln(L^m)}{\partial \gamma}\right] = \sum_{i=1}^{N}\sum_{t=1}^{T}\left\{\frac{\mathbf{w}_{it}f(\mathbf{w}_{it}\gamma+\theta_i)(1-\lambda_1-\lambda_2)}{\lambda_2+(1-\lambda_1-\lambda_2)F(\mathbf{w}_{it}\gamma+\theta_i)}\left[\lambda_2+(1-\lambda_1-\lambda_2)F(\mathbf{w}_{it}\gamma+\theta_i)\right]\right\}$$

$$\sum_{i=1}^{N}\sum_{t=1}^{T}\left\{\frac{\mathbf{w}_{it}f(\mathbf{w}_{it}\gamma+\theta_i)(1-\lambda_1-\lambda_2)}{1-\lambda_2-(1-\lambda_1-\lambda_2)F(\mathbf{w}_i\gamma+\theta_i)}\left[1-\lambda_2-(1-\lambda_1-\lambda_2)F(\mathbf{w}_{it}\gamma+\theta_i)\right]\right\}=0. \quad (16)$$

Note that, observations with a high index value such that $F(\mathbf{w}_{it}\gamma+\theta_i)$ is local to unity that are misclassified will be assigned to the second term of $\frac{\partial\ln(L)}{\partial\gamma}$, *i.e.* $\frac{\mathbf{w}_{it}f(\mathbf{w}_{it}\gamma+\theta_i)}{1-F(\mathbf{w}_{it}\gamma+\theta_i)}$ in the expectation operator of Eq. (15), making the denominator zero and driving the whole term towards infinity. Similarly, observations with $F(\mathbf{w}_{it}\gamma)$ local to zero that are misclassified are assigned to the first term of $\frac{\partial\ln(L)}{\partial\gamma}$, *i.e.* $\frac{\mathbf{w}_{it}f(\mathbf{w}_{it}\gamma+\theta_i)}{F(\mathbf{w}_{it}\gamma+\theta_i)}$ in the expectation operator of Eq. (15), causing the same problem. Paradoxically then, the inconsistency increases with model fit- see Hausman and Scott-Morton, (1994).

In the absence of an individual-specific effect ($\theta_i=0$) or when $\sigma_\theta^2=0$, then one can obtain a $\sqrt{N}$-consistent estimator of γ via the maximisation of the partial log-likelihood referred to as pooled estimation (see Wooldridge, 2002, p.482). The cross-sectional case is equivalent to the pooled estimator with $T=1$ and is treated in detail in Hausman and Scott-Morton, (1994), Hausman *et al.* (1998) and Meyer and Mittag (2017). Otherwise, if $\sigma_\theta^2\neq 0$, we should introduce assumptions regarding the relationship between $\mathbf{w}_{it}$ and θ_i. Unlike the linear case, treating the θ_i as parameters to be jointly estimated with γ, introduces the incidental parameters problem (see Heckman, 1981). The (modified) log-likelihood function given $\sigma_\theta^2=0$ would then be

$$\begin{aligned}\ln(L^m) \quad = \quad & \sum_{i=1}^{N}\sum_{t=1}^{T}\{x_{it}\ln\left[\lambda_2+(1-\lambda_1-\lambda_2)F(\mathbf{w}_{it}\gamma)\right] \\ & +(1-x_{it})\ln\left[1-\lambda_2-(1-\lambda_1-\lambda_2)F(\mathbf{w}_{it}\gamma)\right]\}\end{aligned} \quad (17)$$

Alternatively, treating θ_i as a random (unobserved) variable drawn with $(x_i,\mathbf{w}_i)$ while additionally assuming $\theta_i|\mathbf{w}_i\sim N(0,\sigma_\theta^2)$ provides consistent parameter estimates for (γ,σ_θ^2) via conditional maximum likelihood. Finding the conditional joint distribution of $(x_{i1},...,x_{iT})$ requires integrating out θ_i and under the assumption that $\theta_i\sim N(0,\sigma_\theta^2)$

$$f(x_1,...,x_T|\mathbf{w}_i;\mathbf{c})=\int_{-\infty}^{+\infty}\left[\prod_{t=1}^{T}f(x_t|\mathbf{w}_{it},\theta;\gamma)\right]\left(\frac{1}{\sigma_\theta}\right)\phi\left(\frac{\theta}{\sigma_\theta}\right)d\theta \quad (18)$$

where assuming no misclassification

$$f(x_t|\mathbf{w}_t, \theta; \gamma) = \Phi\left(\frac{\mathbf{w}_t\gamma + \theta}{\sigma_\eta}\right)^{x_t}\left[1 - \Phi\left(\frac{\mathbf{w}_t\gamma + \theta}{\sigma_\eta}\right)\right]^{1-x_t} \tag{19}$$

and $\mathbf{c}$ contains γ and σ_θ^2 while (ϕ, Φ) denote the Normal *pdf* and *cdf*, respectively. The respective (modified) log-likelihood function becomes

$$\begin{aligned} \ln(L^m) &= \sum_{i=1}^{N}\left\{\ln\int\left[\prod_{t=1}^{T}\left[\lambda_2 + (1-\lambda_1-\lambda_2)\Phi\left(\frac{\mathbf{w}_{it}\gamma+\theta}{\sigma_\eta}\right)\right]^{x_{it}}\right.\right. \\ &\left.\left.\left[1-\lambda_2-(1-\lambda_1-\lambda_2)\Phi\left(\frac{\mathbf{w}_{it}\gamma+\theta}{\sigma_\eta}\right)\right]^{1-x_{it}}\right]\left(\frac{1}{\sigma_\theta}\right)\phi\left(\frac{\theta}{\sigma_\theta}\right)d\theta\right\} \end{aligned} \tag{20}$$

noting that the integral has no analytical solution and can be either approximated using simulation or, the Gauss-Hermite quadrature- see Butler and Moffitt (1982). Since x_{it} is binary a normalisation is required and it is assumed that $\sigma_\eta^2 = 1$.