Two-step Parametric Estimation of Binary Treatment Effects in the Presence of Misclassification and Endogeneity

Georgios Marios Chrysanthou*

January, 2025

Abstract

I provide a two-step parametric estimator correcting for misclassification and endogeneity biases in binary treatment effects. Approximate consistency is achieved via modified maximum likelihood estimation (MMLE) of the reduced form binary discrete choice model and, modified least squares (MLS) estimation of the structural form which is augmented by a misclassification-corrected control function. The model incorporates unequal/equal misclassification probabilities for false negatives/positives, and estimates misclassification rates without reliance on extraneous information or surrogate measurements. The two-step MLS (2SMLS) estimator outperforms naive instrumental variables estimation (IV) that ignores misclassification, and OLS in terms of bias reduction. If the treatment parameter has the same sign as the error correlation, approximate OLS bias cancellation occurs rendering OLS comparable to 2MSLS if the estimable (error correlation/misclassification) ratio is equal to 1 but this may extend to ratios in $[0.5 - 1]$ as per simulations. The 2SMLS method provides estimates of the degree of endogeneity and misclassification such that practitioners can assess overall potential bias. Structural identification of the 2SMLS estimator requires a relevant exclusion restriction. The estimator is applied to study the impact of labour market inactivity on social benefit income.

JEL classification: C25, C31, C35

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

*Economic Studies, School of Business, University of Dundee, UK. E-mail: GChrysanthou001@dundee.ac.uk.

1 Introduction

This paper offers a tractable parametric estimator correcting for misclassification and endogeneity biases in binary treatment effects without reliance on external information regarding the rate of misclassification or additional measurements of the error-ridden endogenous binary treatment variable. Endogenous binary treatment is common in economic analysis (see *e.g.* Vella and Verbeek, 1999), while misclassification error in binary treatment variables is an additional econometric challenge- see Celhay *et al.* (2024) for a recent comprehensive investigation on the determinants of binary reporting error.

The related literature can be divided into three categories: **Category 1.** Exogenous binary treatment and misclassification (*e.g.* Aigner, 1973; Lewbel, 2007). **Category 2.** Endogenous binary treatment and exogenous (non-differential) misclassification (*e.g.* Battistin *et al.*, 2014; Calvi *et al.*, 2022; Tommasi and Zhang, 2024b). **Category 3.** Endogenous binary treatment and endogenous (differential) misclassification (*e.g.* Ura, 2018; Nguimkeu *et al.*, 2019; Tommasi and Zhang, 2024a).

The current study falls in **Category 3** of endogenous binary treatment and endogenous differential misclassification since it extends the MLS estimator of Aigner (1973) by incorporating endogeneity, and estimates the misclassification rate using the MMLE of Hausman *et al.* (1998). As the two sources of bias can cancel each other out (see Solon, 1985; Wooldridge, 2010, p.312; Hsiao, 2014, p.304) the range of bias cancellation is identified using simulations, and can be empirically determined using estimable parameters such that practitioners can choose between 2SMLS, naive IV accounting only for endogeneity, or uncorrected OLS. The empirical application illustrates that the proposed 2SMLS estimator provides a very similar treatment effect point estimate to the one obtained using the estimator of Tommasi and Zhang (2024a) if the latter is supplemented by the MMLE estimated mislcassification probabilities to tighten the treatment effects bounds.

Responses in survey datasets (*e.g.* UK Household Longitudinal Study, German Socio-Economic Panel, US Panel Study of Income Dynamics) are collected by means of one-time retrospective questions rendering such data susceptible to measurement error. Data collectors enquire individuals/household members at a specific time point in a given calendar year (survey wave) regarding a plethora of socioeconomic characteristics and outcomes noting that individuals are interviewed at distinct calendar dates, weeks, months or even calendar years (e.g. UKHLS waves are issued as 24 monthly samples and data collection and interviews span two years). Empirical analysts would then use these answers to construct binary indicators intended to accurately reflect individual responses/behaviour throughout the entire calendar year (or other cross-sectional unit) which may result in substantial misclassification in such binary variables.

Celhay *et al.* (2024) study the determinants of reporting error using New York State administrative microdata on government transfers linked to the American Community Survey (ACS), the Population Survey (CPS), and the Survey of Income and Program Participation (SIPP). Celhay *et al.* (2024) conclude that recall error leads to overreporting, topic importance (salience) in terms of benefit receipt duration and amount improves response quality, stigma reduces social benefit receipt reporting and interviewee cooperativeness affects response accuracy (frequent nonresponse linked to higher misreporting probabilities).

In dealing with an endogenous binary regressor within a two-stage parametric endogenous treatment framework, we are effectively facing measurement error problems in both estimation steps. The structural equation includes a binary explanatory variable that is both misclassified and endogenous. The reduced form generating the endogeneity correcting control function (generalised residual) is a misclassified binary choice model, and consistency of the second stage parameters requires adjusting the control function in order to incorporate the estimated misclassification probabilities computed via the MMLE of Hausman *et al.* (1998). [1]

Measurement error in an exogenous binary regressor (**Category 1**) produces biased and inconsistent parameter estimates (see Cochran 1968; Aigner, 1973). A parametric MLS estimator purging the correlation between the misclassified binary explanatory variable and the measurement error is offered by Aigner (1973), and Bollinger (1996) studies parametric and non-parametric identification of parameter bounds. Lewbel (2007) considers the identification and estimation of the effect of a misclassified binary explanatory variable in the context of nonparametric/semiparametric regression. Lewbel (2007) obtains the same attenuation-bias result as Aigner (1973) and introduces assumptions (an instrument for the binary regressor conditionally independent of the treatment) to identify the conditional average treatment effect of the misclassified binary regressor.

The present study falls in **Category 3** since we do not assume an exogenous treatment as in studies of **Category 2** (*e.g.* Battistin *et al.*, 2014; Calvi *et al.*, 2022; Tommasi and Zhang, 2024b). Nguimkeu

[1] Meyer and Mittag (2017) extend the Hausman *et al.* (1998) MMLE method by permitting misclassification of the binary dependent variable to be correlated with observables.

et al. (2019) achieve point identification in the presence of endogenous participation and endogenous one-sided misclassification (only false positives or false negatives) using a two-step parametric model. Nguimkeu *et al.* (2019) consider an incomplete data scenario of endogenous binary regressor misreporting and endogenous participation where the true participation indicator is unobserved and, instead a misclassified surrogate is observed. Ura (2018) studies the identifying power of an instrumental variable in the nonparametric heterogeneous treatment effect framework with a mismeasured endogenous binary treatment cocnluding that the Wald estimand is an upper bound on the local average treatment effect (LATE), but generally does not correspond to a sharp bound. Tommasi and Zhang (2024a) study LATE and the weighted average of LATEs when the binary treatment is mismeasured and using instrumental variables that are binary, discrete (or multiple-discrete).

Contrary to the present study, Ura(2018) and Tommasi and Zhang (2024a) offer interval bound estimates and do not estimate the misclassification probabilities, but rely on extraneous information regarding misclassification in order to tighten the interval bounds of the estimated treatment effect. This study contributes to the endogenous binary misclassified treatment literature by offering a point estimate of the treatment effect using a tractable two-stage parametric method. The proposed method has a bias-reduction property as it exhibits lower bias than naive IV and OLS, and accommodates for symmetric and asymmetric (bidirectional) misreporting (unlike Nguimkeu *et al.*, 2019 focusing on one-sided misreporting). Structural identification requires an exclusion restriction, but does not require additional alternative measurements or extraneous information on the misclassified binary variable.

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 term. Section 4 presents the simulations, and Section 5 presents an empirical application. Conclusions are given in Section 6. The Appendix presents the proofs regarding the consistency of the second stage modified least squares (MLS) estimator.

2 Misclassification 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). We partition the structural equation population regression function as

$$\begin{aligned} y_{j,i} &= \mathbf{z}_i\boldsymbol{\beta} + \delta x_i^T + \varepsilon_{j,i} \\ j &= \{0,1\}\; if\; x_i^T = \{0,1\}, y_i = y_{j,i}\;\; if\;\; x_i^T = j, (i = 1, ..., N) \end{aligned} \tag{1}$$

where, $\boldsymbol{\beta}$ is a $[(k-1)\,\mathrm{x}1]$ vector of unknown parameters assumed to be invariant with respect to j, $\mathbf{z}_i$ is an $[1\mathrm{x}\,(k-1)]$ vector of explanatory variables other than the true value of the binary treatment variable $x_i^T \in \{0,1\}$, and $y_{j,i}$ is the potential outcome of individual i if $(j = 0, 1)$.

Assumption 1. *The* $(k-1)\times(k-1)$ *matrix,* $\mathbb{E}\left(\mathbf{z_i'z_i}\right)$, *is nonsingular (and hence finite).*

Assumption 2. *The random component satisfies* $\varepsilon_{j,i} \sim iid\; N(0, \sigma_\varepsilon^2)$ *and* $\mathbb{E}(\mathbf{z}_i'\varepsilon_{j,i})$=***0***.

The key parameter of interest is the treatment effect denoted by scalar δ multiplying the true value of the endogenous binary explanatory variable (x_i^T). The reduced form for x_i^T is

$$x_i^T \;\; = \;\; \mathbf{1}\left\{\mathbf{w}_i\boldsymbol{\gamma} + \eta_i > 0\right\}, (i = 1, ..., N) \tag{2}$$

where, $\mathbf{1}\left\{c\right\}$ takes the value of one if condition c is satisfied and zero otherwise, $\boldsymbol{\gamma}$ is a $[k\mathrm{x}1]$ vector of unknown parameters, $\mathbf{w}_i$ is a $[1\mathrm{x}k]$ vector of explanatory variables.

Assumption 3. *The reduced form random component satisfies* $\eta_i \sim iid\; N(0,1)$.

Assumption 4. *The* $k \times k$ *matrix,* $\mathbb{E}\left(\mathbf{w_i'w_i}\right)$, *is nonsingular (and hence finite).*

Assumption 5. ***Structural identification*** *of the parameters in the two-part model in Equations (1,2) is achieved if* $\mathbf{w}_i \neq \mathbf{z}_i$ *i.e. at least one element in* $\mathbf{w}_i$ *is not included in* $\mathbf{z}_i$.

Equation (1) can be written as

$$y_i = \mathbf{z}_i\boldsymbol{\beta} + \delta x_i^T + e_i, e_i = x_i^T\varepsilon_{1,i} + (1 - x_i^T)\varepsilon_{0,i} \tag{3}$$

Assumption 6. *The error terms (e_i,η_i) in Equations (1,2) follow a bivariate normal distribution: (e_i, η_i), $i = 1, ..., N$ are iid normally distributed such that*

$$\begin{pmatrix} e_i \\ \eta_i \end{pmatrix} \sim N \begin{pmatrix} 0 & & \sigma_e^2 & \sigma_{e\eta} \\ & , & & \\ 0 & & \sigma_{e\eta} & 1 \end{pmatrix}. \tag{4}$$

Let (x_i^T, x_i) be the true and observed binary treatment indicators for the outcome of individual i, respectively where the binary observed indicator $x_i \in \{0, 1\}$ is subject to misclassification error. Assuming that $\left(x_i^T,\ x_i\right)$ are related via $x_i = x_i^T + \tau_i$ where τ_i is a measurement error, Eq. (3) becomes

$$\begin{aligned} y_i &= \mathbf{z}_i\boldsymbol{\beta} + \delta x_i + u_{_i}, u_{_i} = (e_i - \delta\tau_i), (i = 1, ..., N) \\ e_i &= (x_i - \tau_i)\,\varepsilon_{1,i} + (1 - x_i + \tau_i)\varepsilon_{0,i}. \end{aligned} \tag{5}$$

Equations (3) and (5) indicate that misclassification is endogenous (differential) since

$$x_j \not\perp (y_j, x_j^T), j = 0, 1.$$

Denote the probability of a false negative/positive conditional on the true binary indicator x_i^T by (λ_1,λ_2), respectively. It is assumed that the misclassification probabilities are constant across individuals and that misclassification is asymmetric *i.e.* $\lambda_1 \neq \lambda_2$. The results in the remaining paper are easily extended to the case of symmetric misclassification by simply setting $\lambda_1 = \lambda_2$. The misclassification probabilities (λ_1, λ_2) are

$$\lambda_1 = \Pr(x_i = 0|x_i^T = 1), \lambda_2 = \Pr(x_i = 1|x_i^T = 0) \tag{6}$$

noting that (λ_1,λ_2) depend on x_i and are assumed to be independent of $\mathbf{w}_i$ and η_i.

Using the joint distribution of (x_i, τ_i),

$$\begin{aligned}
\mathbb{E}(\tau_i) &= \lambda_1 \tilde{\pi} - \lambda_2 (1-\tilde{\pi}) && (7)\\
Var(\tau_i) &= [\lambda_1 \tilde{\pi} + \lambda_2(1-\tilde{\pi})] - [\lambda_1 \tilde{\pi} - \lambda_2(1-\tilde{\pi})]^2 && (8)\\
\mathbb{E}(x_i) &= \tilde{\pi},\ Var\,(x_i) = \tilde{\pi}(1-\tilde{\pi}) && (9)\\
\mathbb{E}(x_i \tau_i) &= \lambda_1 \tilde{\pi},\ Cov(x_i, \tau_i) = (\lambda_1 + \lambda_2)\,\tilde{\pi}(1-\tilde{\pi}) && (10)
\end{aligned}$$

where, $\tilde{\pi} = (N)^{-1} \sum_{i=1}^{N} x_i$ denotes the expected value of the observed responses. There is a negative correlation between true response and the measurement error since when $x_i^T = 1$, τ_i is either -1/0 while when $x_i^T = 0$, τ_i is either $0/1$. Contrary to classical errors in variables assumptions, τ_i does not have a zero mean, it is negatively correlated with x_i^T and it is also correlated with x_i.

The non-zero $Cov(x_i, \tau_i)$ can bias all least squares estimates of regression coefficients other than δ unless all covariates are orthogonal to x_i. Using the MMLE estimates of (λ_1, λ_2) we can compute $Cov(x_i, \tau_i)$ and consistently estimate ($\boldsymbol{\beta}, \delta$) employing the modified least squares (MLS) procedure of Aigner (1973), and extending it by the inclusion of a control function to account for the endogeneity of x_i. The corresponding MMLE function is specified in Eq.19, Section 3.

Using Eq. (5), the MLS estimators for ($\boldsymbol{\beta}, \delta$) in the partitioned linear model correspond to

$$\begin{aligned}
\begin{bmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\delta} \end{bmatrix} &= \begin{bmatrix} M_{ZZ} & M_{Zx} \\ M'_{Zx} & (m_{xx} - \xi) \end{bmatrix}^{-1} \begin{bmatrix} M_{Zy} \\ m_{xy} \end{bmatrix} && (11)\\
M_{ZZ} &= (N)^{-1}(Z'Z), M_{Zx} = (N)^{-1}(Z'x), M_{Zy} = (N)^{-1}(Z'y)\\
m_{xx} &= (N)^{-1}(x'x), m_{xy} = (N)^{-1}(x'y), \xi = Cov(x_i, \tau_i).
\end{aligned}$$

where Z is a Nx$(K-1)$ matrix of observations of all explanatory variables other than x.

Theorem 2.1. *The MLS estimator in Eq.(11) gives* $\underset{N\longrightarrow\infty}{plim}\ \widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}$,

and

$$\underset{N\longrightarrow\infty}{plim}\ \widehat{\delta} = \delta + \underset{N\longrightarrow\infty}{plim}\left[m_{xx}^{-1}\left(\frac{1}{N}x'e\right)\left[1-\xi m_{xx}^{-1}\right]^{-1}\right] \tag{12}$$

corresponding to

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

Proofs. See **Appendix, A1-A3**. □

Theorem 2.1 is crucial in the ensuing analysis since unlike Aigner (1973), the endogeneity of x implicates that $\underset{N\longrightarrow\infty}{plim}\ \left(\frac{1}{N}x'e\right) \neq 0$ 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. (13) 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 Heckman, 1979; Vella, 1998).

Consistency relies on knowing ξ and, in theory one could substitute a consistent estimate for ξ and proceed with the estimation. An estimate for ξ can be obtained from an extraneous source, *i.e.* a distinct population than the population from which the N observations have been selected but, in this case it would not make much sense to use $\underset{N\longrightarrow\infty}{plim}\ \left(\frac{1}{N}x'\tau\right) = \xi$ which is central to the consistency proofs (see Eq. (A.11), Eq. (A.14) in the Appendix). When using an extraneous sample to compute ξ as stated by Aigner (1973) "the strict classical statistician must be satisfied with a sort of approximate consistency when $\hat{\xi}$ is used in place of ξ" (p.55).

In the absence of extraneous information regarding ξ, consistent parameter estimation requires using the single available sample to compute the extent of misclassification. Consistency is achieved by modifying the likelihood function to estimate misclassification in the binary endogenous treatment, and subsequently adding the modified endogeneity correcting control function term to the structural equation. We treat this in the following Section.

3 Deriving the Control Function under Misclassification

Taking expectations, conditional on $(\mathbf{w}_i, x_i)$, Eq. (5) becomes

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

$$\therefore \mathbb{E}\left[y_i|\mathbf{w}_i, x_i\right] = \mathbf{z}_i\boldsymbol{\beta} + \delta\left(x_i - \xi\right) + \mathbb{E}\left[e_i|\mathbf{w}_i, x_i\right]. \tag{14}$$

Given the result in **Theorem 2.1**, under **Assumption 6** of jointly normally distributed error terms, $E\left[e_i|\mathbf{w}_i, x_i\right]$ corresponds to

$$\mathbb{E}\left[e_i|\mathbf{w}_i, x_i\right] = \frac{\sigma_{e\eta}}{\sigma_\eta^2}\psi\left[\mu_i\left(\mathbf{w}_i\boldsymbol{\gamma}\right)\right] \tag{15}$$

where

$$\psi = \left[1 - \xi m_{xx}^{-1}\right]^{-1}, m_{xx} = (N)^{-1}\left(x'x\right) \tag{16}$$

$$\xi = \left(\lambda_1 + \lambda_2\right)\tilde{\pi}(1 - \tilde{\pi}), \tilde{\pi} = (N)^{-1}\sum_{i=1}^{N} x_i \tag{17}$$

$$\psi = \left[\frac{m_{xx}}{m_{xx} - \xi}\right] = \left(1 - \lambda_1 - \lambda_2\right)^{-1} \tag{18}$$

and $\mu_i\left(\mathbf{w}_i\gamma\right)$ denotes the modified generalised probit residual defined subsequently in Eq. (20).

Misclassification must be explicitly modelled since e_i in Eq. (5) is a function of τ_i. Following Hausman *et al.* (1998), under **Assumption 3** and an additional monotonicity condition regarding the sum of the misclassification probabilities specified in **Assumption 7**, consistent estimation of the reduced form parameters requires maximising

$$\begin{aligned}\ln(L^m) = &\sum_{i=1}^{N}\left\{x_i \ln\left[\lambda_2 + (1 - \lambda_1 - \lambda_2)\Phi\left(\mathbf{w}_i\boldsymbol{\gamma}\right)\right]\right.\\ &\left.+ (1 - x_i)\ln\left[1 - \lambda_2 - (1 - \lambda_1 - \lambda_2)\Phi\left(\mathbf{w}_i\boldsymbol{\gamma}\right)\right]\right\}\end{aligned} \tag{19}$$

where, $\Phi(.)$ is the *cdf* of the Normal distribution.

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

$$\mu_i\left(\mathbf{w}_i\boldsymbol{\gamma}\right) = \frac{\phi\left(\mathbf{w}_i\boldsymbol{\gamma}\right)(1-\lambda_1-\lambda_2)\left[x_i - \lambda_2 - (1-\lambda_1-\lambda_2)\Phi\left(\mathbf{w}_i\boldsymbol{\gamma}\right)\right]}{\left[\lambda_2 + (1-\lambda_1-\lambda_2)\Phi\left(\mathbf{w}_i\boldsymbol{\gamma}\right)\right]\left[1-\lambda_2-(1-\lambda_1-\lambda_2)\Phi\left(\mathbf{w}_i\boldsymbol{\gamma}\right)\right]} \tag{20}$$

where, $\phi(.)$ is the *pdf* and $\Phi(.)$ the *cdf* of the Normal distribution and $\mu_i\left(\mathbf{w}_i\boldsymbol{\gamma}\right)$ corresponds to the conventional inverse Mills ratio *i.e*, the generalised probit residual of Gourieroux *et al.*, (1987) in the absence of misclassification when $\lambda_1 = 0, \lambda_2 = 0$.

Assumption 7. $\lambda_1 = \Pr(x_i = 0|x_i^T = 1), \lambda_2 = \Pr(x_i = 1|x_i^T = 0)$ *satisfy the* ***monotonicity condition:***

$$\lambda_1 + \lambda_2 < 1.$$

Assumption 7 implies that $\lambda_2 + (1-\lambda_1-\lambda_2)\Phi(\mathbf{w}_i\boldsymbol{\gamma})$ is strictly increasing in $(\mathbf{w}_i\boldsymbol{\gamma})$ if Φ is strictly increasing (*i.e.*, η has positive density everywhere). If the monotonicity **Assumption 7** fails such that $\lambda_1 + \lambda_2 = 1$, then writing $\widetilde{\lambda}_1 = 1-\lambda_2$, $\widetilde{\lambda}_2 = 1-\lambda_1$, $\widetilde{\boldsymbol{\gamma}} = -\boldsymbol{\gamma}$ makes it clear that we are unable to distinguish between $\left(\widetilde{\lambda}_1, \widetilde{\lambda}_2, \widetilde{\boldsymbol{\gamma}}\right)$ and $(\lambda_1, \lambda_2, \boldsymbol{\gamma})$ since the symmetry of the normal *cdf* implies that $\Phi(\mathbf{w}_i\widetilde{\boldsymbol{\gamma}})(1-\widetilde{\lambda}_1-\widetilde{\lambda_2}) + \widetilde{\lambda_2} = \Phi(\mathbf{w}_i\boldsymbol{\gamma})(1-\lambda_1-\lambda_2)+\lambda_2$. Alternatively, if $\lambda_1+\lambda_2 > 1$ then imposing **Assumption 7** will produce opposite sign estimates of $\boldsymbol{\gamma}$ (see Hausman *et al.*, 1998, pp. 242-43).

Theorem 3.1. *If* ***Assumption 3****,* ***Assumption 4*** *and* ***Assumption 7*** *hold, then the reduced form parameters* $(\lambda_1, \lambda_2, \boldsymbol{\gamma})$ *are identified by maximising the MMLE in Equation (19).*

The **Proof of Theorem 3.1** follows from Newey and McFadden (1994, pp. 2125-26). □

The Fisher information matrix associated with the maximisation of the MMLE in Equation (19) is given in Hausman *et al.*, 1998, pp. 244). Since all remaining components in Eqs. (16-18) can be computed from the underlying sample, we can construct the appropriate control function, $\psi\left[\mu_i\left(\mathbf{w}_i\boldsymbol{\gamma}\right)\right]$, to be added in the structural form as an endogeneity correction term.

Theorem 3.2. *Under* ***Assumptions 1-7****, Ordinary Least Squares (OLS) estimation of*

$$y_i = \mathbf{z}_i\boldsymbol{\beta} + \delta\left(x_i - \xi\right) + \left[\frac{\mu_i\left(\mathbf{w}_i\boldsymbol{\gamma}\right)}{(1-\lambda_1-\lambda_2)}\right]\varkappa + \omega_i \tag{21}$$

can provide consistent estimates of the structural form parameters $(\boldsymbol{\beta}, \delta, \varkappa)$. □

Structural identification of the full set of the two-stage model parameters, requires that at least one element in $\mathbf{w}_i$ is not included in $\mathbf{z}_i$ (see Vella, 1998; Puhani, 2000). While the inverse Mills ratio is nonlinear in the single index $(\mathbf{w}_i\boldsymbol{\gamma})$ the function mapping this index into the inverse Mills ratio is linear for certain ranges of the index, see Vella (1998). Leung and Yu (1996) suggest using the correlation between $\mu_i(\mathbf{w}_i\boldsymbol{\gamma})$ and $(\mathbf{w}_i\boldsymbol{\gamma})$. Accordingly the inclusion of additional variables in $\mathbf{w}_i$ in the reduced form can be important for structural identification of the second step estimates. Leung and Yu (1996) conclude that the Inverse Mills ratio is linear over wide range of its argument, but becomes nonlinear at extreme values of the index $(\mathbf{w}_i\boldsymbol{\gamma})$. This implicates that, if $(\mathbf{w}_i\boldsymbol{\gamma})$ spans a relatively large value range, even in the absence of exclusion restrictions functional identification can be achieved, though the practitioner is advised to plot $\mu_i(\mathbf{w}_i\boldsymbol{\gamma})$ against $(\mathbf{w}_i\boldsymbol{\gamma})$ in order to determine non-reliance on exclusion restrictions.

Summarising, the two-stage estimation procedure requires modified maximum likelihood estimation of the reduced form to obtain the misclassification probabilities (see Hausman *et al.*, 1998). Consistent estimation of the structural form parameters is subsequently achieved by adjusting the moment matrix of covariates and the control function such that they incorporate misclassification- see Theorem 3.2). Finally note that, 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. Alternatively, joint estimation of the two parts of the model via full-information maximum likelihood estimation (FIML) may frequently lead to convergence problems due to collinearity induced by the endogeneity correction function being an approximately linear function over a wide range of its argument (see Puhani, 2000). Using a two-step estimation process and bootstrapping standard errors over both estimation stages is an attractive alternative and the most robust amongst the simple-to-compute estimators (see Puhani, 2000).

4 Simulations

This section presents the results of Monte Carlo simulations (in Tables 1 and 2) comparing the proposed two-step Modified Least Squares (2SMLS) estimator to naive Instrumental Variables Estimation (IV), and OLS. The aim is consistent estimation of the binary treatment parameter δ.

4.1 Simulation Design and Results

The first-step MMLE Monte Carlo design follows the data generation process of Hausman *et al.* (1998) and has three covariates: the first variable, w_1, is drawn from a lognormal distribution; the second, w_2, is a dummy variable equal to one with probability $1/3$; the third, w_3, is distributed uniformly. The error disturbance, ε, is drawn from a standard normal distribution. The latent dependent variable is given by

$$x_i^* = -1 + 0.2w_{i1} + 1.5w_{i2} - (0.6)w_{i3} + \eta_i. \tag{22}$$

The observed dependent variable is generated using asymmetric misclassification (*i.e.*, $\lambda_1 \neq \lambda_2$) noting that (λ_1, λ_2) are on average only 1% different such that the simulations are directly generalisable to symmetric misclassification (*i.e.*, $\lambda_1 = \lambda_2$). We consider combined false negative and positive rates $(\lambda_1 + \lambda_2)$ of (2%, 5%, 10%, 20%, 30%, 40%, 50%, 60%). The MMLE model parameters and misclassification rates were consistently estimated for combined misclassification rates up to 60%. Note that, 70% combined misclassification rates were considered but, the respective simulations are not reported since the estimated MMLE parameters were often inconsistent.[2]

The structural equation design is given below, where the key binary treatment parameter δ is set to 0.5 noting that we study both positive and negative treatment and, we have explicitly used w as opposed to z to denote structural equation covariates to emphasise that the set of covariates in Eq. (22) and Eq. (23) is the same except for w_2 which is excluded from Eq. (23) for identification purposes:

$$y_i = 1.5 \pm (0.5)x_i + 2w_{i1} - (0.9)w_{i3} + e_i \tag{23}$$

Naive IV (IV in Tables 1 and 2) corresponds to standard instrumental variables estimation ignoring misclassification and corresponds to a two-stage least-squares (2SLS) IV estimator implemented using the *ivregress* Stata command (2SLS, GMM, LIML give identical results). The naive IV method, estimates Eq. (23) by 2SLS using the binary dummy variable w_2 as an instrument for the binary misclassified treatment variable x. 2SMLS outperforms naive IV estimation since the latter accounts for endogeneity bias, but ignores misclassification bias. Naive IV estimation only gives a comparable relative bias to the 2SMLS estimator at the two lowest levels of combined misclassification (2%, 5%) particularly when

[2] Hausman *et al.* (1998) only consider 2%, 5%, 20% symmetric misclassification.

the treatment and error correlation signs are equal. Of note, naive IV always displays positive bias independently of the treatment and endogeneity correlation signs, while its relative-bias performance is unaffected by the degree of endogeneity (correlation among the structural and reduced form errors), since it purges the endogeneity bias, and deteriorates substantially as the degree of misclassification increases.

The endogeneity and misclassification bias correction of the 2SMLS estimator relies on the modified generalised residual $\mu_i\left(\mathbf{w}_i\boldsymbol{\gamma}\right)$ divided by the attrition factor $(1-\lambda_1-\lambda_2)$, see Equations (15, 21). Since the impact of the modified generalised residual, $\mu_i\left(\mathbf{w}_i\right)$, is captured by $\rho = \operatorname{corr}\left(e_i, \eta_i\right)$ the bias correction term can be loosely viewed as equivalent to a function of

$$\left|\frac{\rho}{\lambda_1+\lambda_2}\right|.$$

2SMLS significantly outperforms OLS and can be approximately asymptotically consistent if the treatment effect (δ) sign is opposite to the correlation ($\rho = \operatorname{corr}\left(e_i, \eta_i\right)$) sign. If δ and $\rho = \operatorname{corr}\left(e_i, \eta_i\right)$ have the same sign, the endogeneity bias and the misclassification attrition biases cancel out when

$$\left|\frac{corr(e_i, \eta_i)}{\lambda_1+\lambda_2}\right| \simeq 1.$$

The "bias cancellation region" identified by the simulations, depends on the degree of misclassification and the error correlation and lies within the ratio range:

$$0.5 \leq \left|\frac{corr(e_i, \eta_i)}{\lambda_1+\lambda_2}\right| \leq 1$$

such that OLS bias can be comparable or lower than 2SMLS- see Tables 1 and 2. Within the [0.5-1] bias cancellation region, the 2SMLS bias can be reduced by adjusting the endogeneity correction term (denoted as $\mu_i\left(\mathbf{w}_i\boldsymbol{\gamma}\right)^I$) via usage of the inverted misclassification rates such that

$$\mu_i\left(\mathbf{w}_i\boldsymbol{\gamma}\right)^I = \frac{\phi\left(\mathbf{w}_i\boldsymbol{\gamma}\right)(1-\lambda_1-\lambda_2)^{-1}\left[x_i-(\lambda_2)^{-1}-(1-\lambda_1-\lambda_2)^{-1}\Phi\left(\mathbf{w}_i\boldsymbol{\gamma}\right)\right]}{\left[\lambda_2^{-1}+(1-\lambda_1-\lambda_2)^{-1}\Phi\left(\mathbf{w}_i\boldsymbol{\gamma}\right)\right]\left[1-(\lambda_2)^{-1}-(1-\lambda_1-\lambda_2)^{-1}\Phi\left(\mathbf{w}_i\boldsymbol{\gamma}\right)\right]}. \tag{24}$$

Using the inverted misclassification rates to adjust the endogeneity correction term within the bias cancellation region of [0.5-1] renders the 2SMLS bias lower or approximately equal to the OLS bias- see

Tables (1,2) where the first bold entry corresponds to cases where 2SMLS performs worse than OLS within the bias cancellation region and the second repeated bold entry (for a given combination of misclassification and error correlation) provides the inverted misclassification 2SMLS relative bias.

The correlation between structural and reduced form errors, $\operatorname{corr}(e_i, \eta_i) = \rho$, can be computed using the structural form, Eq. (21), parameter estimates. Following Heckman (1979), a consistent estimator is given by

$$\widehat{\rho} = \frac{\hat{\varkappa}}{\widehat{\sigma}} \tag{25}$$

where, $\hat{\varkappa}$ is the estimated coefficient on the adjusted endogeneity correction term $\left[\frac{\mu_i(\mathbf{w}_i\boldsymbol{\gamma})}{(1-\lambda_1-\lambda_2)}\right]$, and the denominator $\widehat{\sigma} = \sqrt{\frac{\hat{\mathbf{e}}'\hat{\mathbf{e}}}{N}}$ denotes the standard error of the residuals in the structural form (Equation 21).

Table 1: Relative Bias $[\frac{\hat{\delta}-\delta}{\delta}]$
$\delta = \pm 0.5$, $\rho = corr\,(e_i, \eta_i) = \{\pm 0.1, \pm 0.2, \pm 0.3\}$

$M = \lambda_1 + \lambda_2, \lambda_1 \neq \lambda_2$, $\rho = corr\,(e_i, \eta_i)$, $(N = 15000, R = 3000)$

	$\delta > 0$	$\rho > 0$			$\rho < 0$		
	$\delta < 0$	$\rho < 0$			$\rho > 0$		
$\lvert\rho/M\rvert$	$\lvert ratio\rvert$	2SMLS	OLS	IV	2SMLS	OLS	IV
10/2	5.00	0.0771	0.1344	0.0858	0.0708	-0.3407	0.0881
10/5	2.00	0.0920	0.1035	0.1138	0.0829	-0.3592	0.1085
10/10	1.00	0.1067	0.0523	0.1543	0.1035	-0.3885	0.1535
10/10	1.00	0.0478	0.0523	0.1543			
10/20	0.50	0.1371	-0.0458	0.2485	0.1240	-0.4454	0.2512
10/20	0.50	-0.0492	-0.0458	0.2485			
10/30	0.33	0.1390	-0.1849	0.4320	0.1258	-0.5257	0.4253
10/40	0.25	0.1129	-0.3152	0.6650	0.0992	-0.6013	0.6685
10/50	0.20	0.0504	-0.4379	0.9993	0.0315	-0.6738	1.0034
10/60	0.17	-0.0543	-0.6026	1.7816	-0.0743	-0.7693	1.7857
20/2	10.00	0.0809	0.3676	0.0858	0.0661	-0.5738	0.0881
20/5	4.00	0.0966	0.3302	0.1136	0.0782	-0.5859	0.1087
20/10	2.00	0.1113	0.2671	0.1543	0.0989	-0.6035	0.1535
20/20	1.00	0.1414	0.1502	0.2485	0.1191	-0.6414	0.2512
20/30	0.67	0.1487	-0.0168	0.4317	0.1239	-0.6936	0.4255
20/30	0.67	-0.0107	-0.0168	0.4317			
20/40	0.50	0.1094	-0.1741	0.6680	0.0850	-0.7423	0.6680
20/50	0.40	0.0586	-0.3225	0.9992	0.0233	-0.7893	1.0034
20/60	0.33	-0.0442	-0.5209	1.7790	-0.0796	-0.8509	1.7853
30/2	15.00	0.0852	0.6207	0.0862	0.0601	-0.8270	0.0882
30/5	6.00	0.0999	0.5761	0.1121	0.0748	-0.8317	0.1103
30/10	3.00	0.1164	0.5023	0.1543	0.0938	-0.8385	0.1535
30/20	1.50	0.1476	0.3630	0.2485	0.1129	-0.8542	0.2512
30/30	1.00	0.1529	0.1658	0.4314	0.1119	-0.8762	0.4302
30/40	0.75	0.1256	-0.0213	0.6671	0.0814	-0.8951	0.6690
30/40	0.75	-0.0069	-0.0213	0.6671			
30/50	0.60	0.0676	-0.1971	0.9992	0.0143	-0.9146	1.0035
30/60	0.50	-0.0329	-0.4323	1.7790	-0.0908	-0.9395	1.7854

Table 2: Relative Bias $[\frac{\hat{\delta}-\delta}{\delta}]$
$\delta = \pm 0.5$, $\rho = corr\,(e_i, \eta_i) = \{\pm 0.4, \pm 0.5, \pm 0.6\}$

$M = \lambda_1 + \lambda_2, \lambda_1 \neq \lambda_2$, $\rho = corr\,(e_i, \eta_i)$, $(N = 15000, R = 3000)$

	$\delta > 0$	$\rho > 0$			$\rho < 0$		
	$\delta < 0$	$\rho < 0$			$\rho > 0$		
$\lvert\rho/M\rvert$	$\lvert ratio\rvert$	2SMLS	OLS	IV	2SMLS	OLS	IV
40/2	20.00	0.0900	0.8924	0.0858	0.0570	-1.0986	0.0882
40/5	8.00	0.1049	0.8400	0.1131	0.0698	-1.0956	0.1092
40/10	4.00	0.1218	0.7559	0.1542	0.0884	-1.0921	0.1535
40/20	2.00	0.1542	0.5912	0.2484	0.1063	-1.0824	0.2513
40/30	1.33	0.1639	0.3611	0.4311	0.1001	-1.0715	0.4262
40/40	1.00	0.1340	0.1425	0.6650	0.0730	-1.0590	0.6686
40/50	0.80	0.0772	-0.0628	0.9992	0.0047	-1.0490	1.0035
40/50	0.80	-0.0476	-0.0628	0.9992			
40/60	0.67	-0.0208	-0.3373	1.7789	-0.1029	-1.0345	1.7854
50/2	25.00	0.0955	1.2092	0.0883	0.0515	-1.4154	0.0855
50/5	10.00	0.1134	1.1475	0.1119	0.0641	-1.4035	0.1109
50/10	5.00	0.1282	1.0478	0.1542	0.0820	-1.3840	0.1537
50/20	2.50	0.1620	0.8574	0.2484	0.0985	-1.3486	0.2513
50/30	1.67	0.1696	0.5886	0.4269	0.0952	-1.2990	0.4310
50/40	1.25	0.1465	0.3333	0.6670	0.0656	-1.2499	0.6661
50/50	1.00	0.0884	0.0940	0.9986	-0.0065	-1.2058	1.0029
50/60	0.83	-0.0067	-0.2265	1.7724	-0.1170	-1.1454	1.7877
60/2	30.00	0.1070	1.6167	0.0857	0.0395	-1.8230	0.0882
60/5	12.00	0.1182	1.5437	0.1124	0.0543	-1.7994	0.1099
60/10	6.00	0.1438	1.4256	0.1542	0.0656	-1.7619	0.1536
60/20	3.00	0.1771	1.1996	0.2484	0.0799	-1.6908	0.2513
60/30	2.00	0.1824	0.8811	0.4264	0.0808	-1.5917	0.4316
60/40	1.50	0.1628	0.5803	0.6669	0.0460	-1.4968	0.6692
60/50	1.20	0.1029	0.2956	0.9991	-0.0209	-1.4074	1.0036
60/60	1.00	0.0114	-0.0839	1.7788	-0.1351	-1.2879	1.7856

5 Empirical Application

The 2SMLS estimator is applied to estimate the impact of labour market inactivity/unemployment ("unemployed/economically inactive") on "total household social benefit income" using two distinct cross-sections (wave 1, 2009-10) from the Understanding Society (UK Household Longitudinal Study, UKHLS) dataset.

The binary inactivity/unemployment treatment variable is constructed using the "Current economic activity" variable responses and takes the value of 0 if the individual is employed (paid ft/pt employment, self-employed) and the value of 1 otherwise (unemployed, retired, on maternity leave, family care, ft student, LT sick/disabled, governmental training scheme, unpaid family business, on apprenticeship, on furlough, temporarily laid off/short term working, doing something else).

UKHLS data collection stretches across 24 months noting that, while the survey fieldwork period is 24 months, every individual is interviewed at approximately 12 month intervals and individual responses are only collected once per UKHLS wave such that individual responses for a given survey wave correspond to single cross-sectional observations. The interview dates are roughly equally split between 2009 and 2010 regarding the cross-sectional data from wave 1 (only 3.35% of the interviews were conducted in 2011). The binary treatment variable is likely to be misclassified in that, the stated current economic activity on the date of interview may not be an accurate reflection of individual economic activity throughout the duration of the corresponding wave *i.e.* when modelling the reduced form for inactivity/unemployment, individual observations with large linear index ($\mathbf{w}_i\gamma$) values will predict 1 with probability $\Phi\left(\mathbf{w}_i\gamma\right)$ close to 1 (and 0 for small linear index values) independently of the observed stated individual current economic activity responses.

The dependent variable has been subjected to the inverse hyperbolic sine transformation $\operatorname{arsinh}(y) = \ln\left(y + \sqrt{y^2+1}\right)$ such that we are able to include in the analysis observations with zeros. "Total household social benefit income" corresponds to the total income aid received by the UK government as part of the "universal credit" and contains up to 39 components including pension income, incapacity benefit, income support, job seeker's allowance, child benefit, maternity allowance, housing benefit, and council tax benefit.

The binary inactivity/unemployment treatment indicator is likely to be both misclassified and endogenous as the unobserved individual determinants of economic activity may be correlated with the unobserved factors determining household benefit income.

Descriptive statistics are given in Table 3. The estimation samples include individuals of prime working age and those approaching retirement such that the age range is 25-65. All explanatory variable controls included in Tables 4 and 5 are binary variables taking the value of 1 for the stated category except, age (in years) and the "Number of children in household" taking the value 0 if none, 1 if one and 2 if more. The base regional control is "North East". "Total household social benefit income" is given as "arsinh(hh social benefit income)" and, the total number of observations is 26896.

For structural identification of the 2SMLS model, current financial state is excluded from the structural form estimation (column 1, Table 5) and is included in the reduced form in Table 4 since self-reported individual current financial evaluations can affect labour market participation decisions, but are not direct determinants of social benefit income received from the government. The five-point scale ordered current subjective financial state variable has been used to create the binary indicator variable "Doing Alright Financially" which is the highest frequency of self-reported current financial state in the dataset. "Doing Alright Financially" takes the value of 1 if an individual indicated "Doing Alright", and 0 if they reported any of "Finding it very difficult", "Finding it quite difficult", "Just about getting by/don't know", "Living comfortably".

The first-stage reduced form MMLE estimates are given in Table 4. The estimated misclassification rates provided at the bottom of Table 4, are unilaterally statistically significant and misclassification is clearly asymmetric since the χ^2 test-statistic for the equality of $\lambda_1 = \lambda_2$ is 66.03 with a zero p-value. The sum of the estimated misclassification probabilities $\widehat{\lambda_1} + \widehat{\lambda_2}$ is 0.2600632. The estimated probability to observe an individual at the individual-specific interview date to be employed when their true predicted status is inactive/unemployed, $\widehat{\lambda_1}$, is correspondingly 0.1964 which is substantially higher than the estimated probability to be observed as inactive/unemployed on the date of interview when their predicted true status is employed, $\widehat{\lambda_2}$, corresponding to 0.0637. Thus, the estimated misclassification rates indicate that exiting employment is more likely than entering employment during 2009-10. This aligns with Celhay *et al.* (2024) estimating notably higher false negative probabilities, $\widehat{\lambda_1}$, in binary social benefit receipt responses using the ACS, ACP, SIPP datasets between [0.18, 0.59] and lower false positives, $\widehat{\lambda_2}$, within [0.03, 0.013], respectively.

Note that, Unemployed/Economically Inactive (x_i) in the case of the 2SMLS estimator (first column, Table 5) is replaced by $(x_i - \hat{\xi})$ where, $\hat{\xi} = \left(\widehat{\lambda_1} + \widehat{\lambda_2}\right) \tilde{\pi}(1 - \tilde{\pi}), \tilde{\pi} = (N)^{-1} \sum_{i=1}^{N} x_i$, see Eq. (17). The impact of "Unemployed/Inactive" on "Total Household Social Benefit Income" (henceforth referred to as treatment effect) using 2SMLS, OLS and IV is respectively given in columns 1 and 2 of Table 5. The IV estimates reported in (column 3, Table 5) correspond to naive IV estimation which is single-equation two-stage instrumental-variables regression using "Doing Alright Financially" to instrument "Unemployed/Economically Inactive" obtained using the Stata *ivregress* command. The estimated error correlation between the unobservables in the reduced form for the probability of being "Unemployed/Inactive" and the structural form for "Total Household Social Benefit Income", $\widehat{rho} = \widehat{corr\,(e_i}, \eta_i)$, is approximately -0.2879 giving an absolute value ratio of error correlation to misclassification of around 1.1072 (see Table 5). Being unemployed/inactive has a positive effect on social benefit income in the case of all three (2SMLS, OLS, IV) noting that, the OLS and naive IV estimated effects (columns 2 and 3, Table 5) are biased downwards and upwards, respectively. The bias direction of the estimated treatment effects is in line with the simulation bias predictions in Table 1. Note that, the modified generalised residual (defined in Eq. (20)) statistical significance in Table 5, is reduced (p-value=0.07) due to the sampling variation induced by bootstrapping both 2SMLS estimation stages to compute the standard errors increasing the respective standard error to 0.3988 from 0.1116 in the case of the unadjusted standard error (with corresponding p-value=0.00).

To compare the estimated treatment effect to the simulations observe that, $\widehat{\lambda_1} + \widehat{\lambda_2} = 0.2600631$, $\widehat{rho} = \widehat{corr\,(e_i}, \eta_i) = -0.28793126$, and $\left| \dfrac{\widehat{\rho}}{\widehat{\lambda_1} + \widehat{\lambda_2}} \right| = 1.1072$ such that using the closest simulated correlation/misclassification ratio $|\rho/M| = 30/30$ with $(\delta > 0, \rho < 0)$ in Table 1, the relative OLS bias is -0.8762 (i.e. downward bias of 87.62%). The empirical estimated 2SMLS treatment effect is 3.1210 (column one, Table 5) and the OLS estimated treatment effect (column two, Table 5) is 2.0526 such that the estimated relative OLS bias (compared to 2SMLS) of 0.5206 (52.06%). Adjusting the 2SMLS treatment effect estimate for the respective upward 0.1119 relative bias the true estimate should be approximately 3.4703 giving a relative downward bias of the OLS estimate (compared to the true estimate) of -0.6907 which is close to the simulated bias difference prediction of -0.7643 (*i.e.* 0.1119-0.8762). The upwardly biased naive IV treatment effect is 3.8808 (column 3, Table 5) and is (0.2434, 0.1183) higher than the corresponding estimated 2SMLS and true treatment effects of (3.121, 3.4703) aligning with the simulations in Table 1.

To see the policy relevance, consider the semi-elasticity of social benefit income with respect to changes in unemployment or labour market inactivity incidence. The 2SMLS and OLS estimated semi-elasticities are (0.9842, 0.6549). Compared to OLS, the 2SMLS estimates indicate that the demand for social benefits is far more responsive to rises in unemployment and labour market inactivity incidence by 0.5029 percent (half as elastic).

Employing the *ivbounds* Stata command (Lin et al., 2021, 2024) we implement Tommasi and Zhang's (2024a) IV estimator (using "Doing Alright Financially" to instrument "Unemployed/Economically Inactive"). Tommasi and Zhang's (2024a) IV estimator gives bounds of the treatment effect and delivers a point estimate when the misclassification probabilities are known. Using the MMLE estimated false negative and false positive probabilities ($\widehat{\lambda_1}$, $\widehat{\lambda_2}$), the mismeasured IV treatment estimand using Tommasi and Zhang's (2024a) estimator is 4.116 with corresponding 95% confidence interval bounds $[2.5313, 5.7006]$, using 100 bootstraps, such that the corrected treatment effect point estimate corresponds to $4.116(1-(\widehat{\lambda_1}+\widehat{\lambda_2})) = 4.116(0.7399) = 3.0454$ (see Tommasi and Zhang, 2024a). The 2SMLS 95% confidence bounds of the treatment effect correspond to $[2.7898, 3.4523]$ and the corresponding bootstrapped Normal-based 95% confidence interval bounds for the 2SMLS estimated treatment effect in Table 5 are $[1.8561, 4.386]$ which is similar to the Tommasi and Zhang (2024a) interval of $[1.873, 4.2181]$ computed using 100 bootstraps.

Summarising, the 3.0454 point estimate of the treatment effect obtained using Tommasi and Zhang's (2024a) estimator with known misclassification probabilities is quite close to the 2SMLS estimated treatment effect of 3.1210 indicating that the two estimators are comparable. The clear advantage of the 2SMLS estimator is that the misclassification probabilities are estimated by MMLE in the first stage of the model such that no extraneous information regarding misclassification is necessary as in the case of Tommasi and Zhang (2024a) that rely on external knowledge of the misclassification probabilities in order to improve the estimated treatment effect bounds.

Table 3: Descriptive Statistics

	mean	standard deviation
Unemployed/Economically inactive	0.3058	(0.0363)
Doing Alright financially	0.3150	(0.0333)
Age	44.8965	(0.0017)
Female	0.5683	(0.0313)
Married/Civil Partnership	0.5734	(0.0341)
Number of children in household	0.5856	(0.0216)
LT illness/disability	0.3531	(0.0339)
University degree	0.2532	(0.0365)
House owned outright/mortgage	0.6918	(0.0370)
North West	0.1207	(0.0837)
Yorkshire and the Humber	0.0880	(0.0879)
East Midlands	0.0812	(0.0891)
West Midlands	0.0910	(0.0873)
East of England	0.0985	(0.0863)
London	0.1044	(0.0859)
South East	0.1398	(0.0821)
South West	0.0895	(0.0876)
Wales	0.0521	(0.0978)
Scotland	0.0883	(0.0878)
arsinh(household social benefit income)	4.3720	(3.2670)
Number of Observations	26896	

Table 4: Probability of Unemployment/Labour Market Inactivity, MMLE

Doing Alright Financially	-0.3165***
	(0.0343)
Age	0.0482***
	(0.0036)
Female	0.5381***
	(0.0396)
Married/Civil Partnership	-0.1475***
	(0.0296)
Number of children in household	0.2727***
	(0.0260)
LT illness/disability	0.6261***
	(0.0418)
University degree	-0.3907***
	(0.0437)
House owned outright/mortgage	-1.0952***
	(0.0797)
North West	0.0389
	(0.0689)
Yorkshire and the Humber	0.0576
	(0.0727)
East Midlands	-0.1152
	(0.0738)
West Midlands	0.0556
	(0.0722)
East of England	-0.1861***
	(0.0719)
London	0.1130
	(0.0720)
South East	-0.1710**
	(0.0683)
South West	-0.1285*
	(0.0720)
Wales	0.0353
	(0.0798)
Scotland	-0.1842**
	(0.0729)
Constant	-2.3526***
	(0.1543)
$\widehat{\lambda_1}$	0.1964***
	(0.0290)
$\widehat{\lambda_2}$	0.0637***
	(0.0083)
Number of Observations	26896

1. Standard errors in parentheses; 2. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 5: Total Household Social Benefit Income

	2SMLS	OLS	IV
Unemployed/Economically Inactive	3.1210***	2.0526***	3.8808***
	(0.6454)	(0.0362)	(0.5523)
Modified Generalised Residual	-0.7222*		
	(0.3988)		
Age	0.0411***	0.0509***	0.0341***
	(0.0062)	(0.0017)	(0.0054)
Female	0.1286	0.2457***	0.0454
	(0.0789)	(0.0313)	(0.0686)
Married/Civil Partnership	-0.0275	-0.0685**	0.0016
	(0.0420)	(0.0341)	(0.0414)
Number of children in household	2.1456***	2.1972***	2.1088***
	(0.0374)	(0.0216)	(0.0350)
LT illness/disability	0.3184***	0.4772***	0.2056**
	(0.0941)	(0.0339)	(0.0892)
University degree	-0.6569***	-0.7333***	-0.6026***
	(0.0565)	(0.0365)	(0.0549)
House owned outright/mortgage	-0.4399***	-0.6934***	-0.2597*
	(0.1499)	(0.0370)	(0.1363)
North West	-0.1101	-0.1088	-0.1110
	(0.0934)	(0.0837)	(0.0875)
Yorkshire and the Humber	-0.0692	-0.0630	-0.0736
	(0.1003)	(0.0879)	(0.0920)
East Midlands	-0.2997***	-0.3326***	-0.2764***
	(0.1089)	(0.0892)	(0.0948)
West Midlands	-0.1634	-0.1616*	-0.1648*
	(0.1067)	(0.0874)	(0.0914)
East of England	-0.3590***	-0.4094***	-0.3231***
	(0.1028)	(0.0863)	(0.0939)
London	-0.4818***	-0.4647***	-0.4939***
	(0.0992)	(0.0859)	(0.0903)
South East	-0.3897***	-0.4341***	-0.3581***
	(0.0937)	(0.0822)	(0.0889)
South West	-0.2887**	-0.3272***	-0.2613***
	(0.1127)	(0.0876)	(0.0938)
Wales	0.0532	0.0567	0.0508
	(0.1282)	(0.0978)	(0.1023)
Scotland	-0.2675***	-0.3111***	-0.2365**
	(0.0962)	(0.0878)	(0.0946)
Constant	1.0352***	0.8309***	0.8856***
	(0.1367)	(0.1039)	(0.1099)
$\widehat{rho} = \widehat{corr\,(e_i, \eta_i)}$	-0.2879		
$\left\lvert \frac{\widehat{\rho}}{\widehat{\lambda_1} + \widehat{\lambda_2}} \right\rvert$	1.1072		
Number of Observations	26896	26896	26896

1. Standard errors in parentheses; 2. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

2. 2SMLS standard errors are Bootstrapped (100 replications).

6 Concluding Remarks

I propose a tractable procedure to estimate binary treatment effects in the presence of a misclassified endogenous binary treatment variable. The two-step parametric estimator does not rely on additional measurements or exogenous information on the mismeasured binary treatment but, structural identification requires an exclusion restriction. The proposed 2SMLS estimator outperforms OLS and naive IV estimators in terms of bias reduction.

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 term, corresponds to a binary choice model where the dependent variable is subject to misclassification producing inconsistent parameter estimates.

Given the modified MLE (MMLE) 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.

Even if the misclassification rates are unknown an approximately consistent 2SMLS procedure can be implemented using the estimated misclassification rates provided an exclusion restriction is imposed in the structural from (*i.e.* the MMLE contains at least one covariate excluded from the second stage structural MLS estimator). The simulations indicate that if the correlation sign between the error terms in the structural and reduced forms is opposite to the treatment effect sign, the proposed 2SMLS procedure always outperforms OLS in terms of relative bias. If the treatment effect and error correlation signs are identical then, the endogeneity bias may approximately cancel out with the misclassification attrition bias such that the OLS relative bias is lower which, can occur if the absolute value of the correlation to misclassification ratio lies between [0.5-1]. However, using the estimable misclassification rates and the estimated error correlation, the modified endogeneity correction term can be adjusted accordingly within the [0.5-1] potential bias cancellation region such that, the 2SMLS bias is reduced to a lower (or approximately equal) bias compared to uncorrected OLS.

The empirical application compares the 2SMLS treatment effect to the markedly biased OLS and naive IV estimators. Further, the application indicates the advantage of 2SMLS compared to the IV bounding treatment-effects estimator of Tommasi and Zhang (2024a) that gives similar estimates, but relies on the validity and availability of extraneous misclassification information to improve, tighten the treatment effect bounds, and effectively give the appropriate point estimate.

A Appendix: Proofs

A.1 Consistency of the Modified Least Squares (MLS) Estimators

Consider the partitioned linear model where the LS normal equations become

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

giving

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

and

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

Using Eq. (A.3) we obtain

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

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

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

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

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

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

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

A.2 Consistency of the Modified Least Squares (MLS) Estimator of β

To prove MLS consistency of the $\underset{1x(k-1)}{\boldsymbol{\beta}}$ parameter vector, insert Eq. (3) in Eq. (A.6) to get

$$\widehat{\boldsymbol{\beta}} = (Z'AZ)^{-1}Z'A\left[Z\boldsymbol{\beta} + x^T\delta + e\right] \tag{A.8}$$

giving

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (Z'AZ)^{-1}Z'Ae + (Z'AZ)^{-1}Z'Ax^T\delta. \tag{A.9}$$

Taking probability limits,

$$\underset{N\longrightarrow\infty}{plim}\ \widehat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \underset{N\longrightarrow\infty}{plim}\ (\frac{1}{N}Z'AZ)^{-1}\ \underset{N\longrightarrow\infty}{plim}\ (\frac{1}{N}Z'Ax^T\delta) \tag{A.10}$$

since under the exogeneity assumption $E(\mathbf{z}_i'\varepsilon_{j,i})$=0, ensuring that $\underset{N\longrightarrow\infty}{plim}\ (\frac{1}{N}Z'Ae) = 0$.

Noting that $Ax^T = x^T - \left(1 - \xi m_{xx}^{-1}\right)^{-1} x + \left(1 - \xi m_{xx}^{-1}\right)^{-1} xm_{xx}^{-1}\left(\frac{1}{N}x'\tau\right)$, where τ=$(x - x^T)$, we get

$$Ax^T = x^T - x\left[1 - m_{xx}^{-1}\left(\frac{1}{N}x'\tau\right)\right]\left[\left(1 - \xi m_{xx}^{-1}\right)^{-1}\right].$$

Therefore,

$$\begin{aligned}\underset{N\longrightarrow\infty}{plim}\left(\frac{1}{N}Z'Ax^T\right) &= \underset{N\longrightarrow\infty}{plim}\left(\frac{1}{N}Z'x^T\right)\\ &\quad - \underset{N\longrightarrow\infty}{plim}\frac{1}{N}Z'x\left[1 - m_{xx}^{-1}\left(\frac{1}{N}x'\tau\right)\right]\left[\left(1 - \xi m_{xx}^{-1}\right)^{-1}\right]\end{aligned} \tag{A.11}$$

and if $\underset{N\longrightarrow\infty}{plim}\ \left[1 - m_{xx}^{-1}\left(\frac{1}{N}x'\tau\right)\right]\left[\left(1 - \xi m_{xx}^{-1}\right)^{-1}\right] = 1$ since

$$\underset{N\longrightarrow\infty}{plim}\left(\frac{1}{N}x'\tau\right) = \xi$$

along with, $\underset{N\longrightarrow\infty}{plim}\ \left(\frac{1}{N}Z'x^T\right) = \underset{N\longrightarrow\infty}{plim}\ \left(\frac{1}{N}Z'x\right)$, implicate that $\underset{N\longrightarrow\infty}{plim}\ \left(\frac{1}{N}Z'Ax^T\right) = 0$ such that introducing this last expression in Eq. (A.10) proves

$$\underset{N\longrightarrow\infty}{plim}\ \widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}$$

□

A.3 Consistency of the Modified Least Squares (MLS) Estimator of δ

Turning to the consistency of the MLS estimate of δ, introduce Eq. (3) in Eq. (A.7) to get

$$\widehat{\delta} = (N)^{-1}\, m_{xx}^{-1} x' \left[I - Z(Z'AZ)^{-1}Z'A\right] \left[Z\boldsymbol{\beta} + x^T\delta + e\right] \left[1 - \xi m_{xx}^{-1}\right]^{-1} \tag{A.12}$$

giving

$$\begin{aligned} \widehat{\delta} \quad &= \quad (N)^{-1}\, m_{xx}^{-1} x'\delta \left[x^T - Z(Z'AZ)^{-1}Z'Ax^T\right] \left[1 - \xi m_{xx}^{-1}\right]^{-1} \\ &\quad + (N)^{-1}\, m_{xx}^{-1} x' \left[e - Z(Z'AZ)^{-1}Z'Ae\right] \left[1 - \xi m_{xx}^{-1}\right]^{-1}. \end{aligned} \tag{A.13}$$

Taking probability limits using $\underset{N\longrightarrow\infty}{plim}\ \left(\frac{1}{N}Z'Ax^T\right)$=0 and $\underset{N\longrightarrow\infty}{plim}\ (\frac{1}{N}Z'Ae)$=0, given the exogeneity assumption $E(\mathbf{z}_i'\varepsilon_{j,i}) = 0$ we get

$$\begin{aligned} \underset{N\longrightarrow\infty}{plim}\ \widehat{\delta} \quad &= \quad \delta\ \underset{N\longrightarrow\infty}{plim} \left[m_{xx}^{-1}\left(\frac{1}{N}x'x^T\right)\left[1 - \xi m_{xx}^{-1}\right]^{-1}\right] \\ &\quad + \underset{N\longrightarrow\infty}{plim} \left[m_{xx}^{-1}\left(\frac{1}{N}x'e\right)\left[1 - \xi m_{xx}^{-1}\right]^{-1}\right] \end{aligned}$$

and using τ=$\left(x - x^T\right)$ we arrive at

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

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

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

corresponding to

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

□

The variances of $(\widehat{\boldsymbol{\beta}}, \widehat{\delta})$ are given in Aigner (1973), p.58.

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] Battistin, E., De Nadai, M., Barbara Sianesi, B., (2014). ‘Misreported Schooling, Multiple Measures and Returns to Educational Qualifications’, *Journal of Econometrics*, 181, pp. 136-150.

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

[4] Calvi, R., Lewbel, A., and Tommasi, D., (2022). ‘LATE With Missing or Mismeasured Treatment’, *Journal of Business Economic Statistics*, 40(4), pp. 1701-1717.

[5] Celhay, P., Meyer, B. D., Mittag, N., (2024). ‘What Leads to Measurement Errors? Evidence from Reports of Program Participation in Three Surveys’, *Journal of Econometrics*, Volume 238, 2, pp. 1-19.

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

[7] Gourieroux, C., Monfort, A., Renault, E., Trognon, A., (1987). ‘Generalised residuals’, *Journal of Econometrics*, 34, pp. 5-32.

[8] 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.

[9] 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.

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

[11] Hsiao, C., (2003). ‘Analysis of Panel Data’, (2nd ed.), Econometric Society Monographs, Cambridge: Cambridge University Press.

[12] Leung, S.F., Yu, S., (1996). ‘On the choice between sample selection and two-part models’, *Journal of Econometrics*, 72, pp. 197-229.

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

[14] Lin, A., Tommasi, D., and Zhang, L., (2021). 'IVBOUNDS: Stata module providing instrumental variable method to bound treatment-effects estimates with potentially misreported and endogenous program participation', Statistical Software Components S458967, Boston College Department of Economics.

[15] Lin, A., Tommasi, D., Zhang, L., (2024). 'Bounding Program Benefits when Participation is Misreported: Estimation and Inference with Stata', *The Stata Journal*, 24(2), pp. 185-212.

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

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

[18] 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.

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

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

[21] Solon, G., (1985), 'Comment on Benefits and Limitations of Panel Data' by C. Hsiao, *Econometric Reviews*, 4, 183-186.

[22] Tommasi, D., and Zhang ,L., (2024a). L. 'Bounding Program Benefits When Participation Is Misreported, *Journal of Econometrics*, 238, pp. 1-14.

[23] Tommasi, D., and Zhang ,L., (2024b). 'Identifying Program Benefits when Participation is Misreported, *Journal of Applied Econometrics*, 39(6), 1123-48.

[24] Ura, T., (2018), 'Heterogeneous Treatment Effects with Mismeasured Endogenous Treatment', *Quantitative Economics*, 9(3), pp. 1335-1370.

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

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

[27] Wooldridge, J. M., (2010). ‘Econometric Analysis of Cross Section and Panel Data’ (2nd ed.). MIT Press.