MONASH
BUSINESS
SCHOOL

ISSN 1440-771X

Department of Econometrics and Business Statistics

http://business.monash.edu/econometrics-and-business-statistics/research/publications

The Bivariate Probit Model, Maximum Likelihood Estimation, Pseudo True Parameters and Partial Identification

Chuhui Li, Donald S. Poskitt and Xueyan Zhao

August 2016

Working Paper 16/16

The Bivariate Probit Model, Maximum Likelihood Estimation, Pseudo True Parameters and Partial Identification

Chuhui Li, Donald S. Poskitt*, and Xueyan Zhao

Department of Econometrics and Business Statistics, Monash University

August 26, 2016

Abstract

This paper presents an examination of the finite sample performance of likelihood based estimators derived from different functional forms. We evaluate the impact of functional form miss-specification on the performance of the maximum likelihood estimator derived from the bivariate probit model. We also investigate the practical importance of available instruments in both cases of correct and incorrect distributional specifications. We analyze the finite sample properties of the endogenous dummy variable and covariate coefficient estimates, and the correlation coefficient estimates, and we examine the existence of possible "compensating effects" between the latter and estimates of parametric functions such as the predicted probabilities and the average treatment effect. Finally, we provide a bridge between the literature on the bivariate probit model and that on partial identification by demonstrating how the properties of likelihood based estimators are explicable via a link between the notion of pseudo-true parameter values and the concepts of partial identification.

Keywords: partial identification, binary outcome models, mis-specification, average treatment effect.

JEL codes: C31, C35, C36.

*Correspondence to: Donald S. Poskitt, Department of Econometrics and Business Statistics, Monash University, Victoria 3800, Australia. Tel.: +61 3 99059378, Fax: +61 3 99055474. Email: Donald.Poskitt@monash.edu

1 Introduction

In the wake of the pioneering work of Heckman (1978) on the identification and estimation of treatment effects in simultaneous equation models with endogenous dummy variables, the bivariate probit model has become the workhorse underlying much applied econometric work in various different areas of economics: labour economics, (Carrasco, 2001; Bryson et al., 2004; Morris, 2007), economics and law (Deadman and MacDonald, 2003), and in health economics (Jones and O'Donnell, 2002; Jones, 2007), for example. The bivariate probit model is typically used where a dichotomous indicator is the outcome of interest and the determinants of the probable outcome includes qualitative information in the form of a dummy variable where, even after controlling for a set of covariates, the possibility that the dummy explanatory variable is endogenous cannot be ruled out a priori. Scenarios of this type are found in diverse fields of study, such as marketing and consumer behaviour, social media and networking, as well as studies of voting behaviour, and a "Web" search reveals many research papers where the bivariate probit model finds application, far too many to list explicitly here.

Although the bivariate probit model provides a readily implemented tool for estimating the effect of an endogenous binary regressor on a binary outcome variable, the identification relies heavily on the parametric specification and distributional assumptions, including linear indexing in latent variables with a threshold crossing rule for the binary variables, and a separable error structure with a prescribed distributional form, assumptions that are unlikely to be true of real data. Moreover, the work of Manski (1988) indicates that despite the assumptions underlying the bivariate probit model being sufficient to yield statistical identification of the model parameters, they are not restrictive enough to guarantee the identification of parametric functions of interest such as the average treatment effect (ATE), the expected value of the difference between the outcome when treated and not treated. Meaningful restrictions on the values that such parametric functions may take can still be achieved, i.e. they can be partially identified, but the resulting bounds are often so wide that they are uninformative for most practical purposes unless additional constraints such as monotonicity of treatment response or monotonicity of treatment selection are imposed, see Manski (1990), Manski (1997) and Manski and Pepper

(2000). Without imposing monotone constraints, Chesher (2005) provides conditions under which features of a nonseparable structural function that depends on a discrete endogenous variable are partially identified, and in Chesher (2010) (see also Chesher, 2007) he shows that single equation instrumental variable (IV) models for discrete outcomes are in general not point but set identified for the structural functions that deliver the values of the discrete outcome.

Given the popularity of the bivariate probit model in econometric applications, and in the light of recent developments in the literature on partial identification, our aim in this paper is to provide some evidence on whether or not the bivariate probit model can still be thought of as being useful. We are motivated in this endeavour by the aphorism of the late G. E. P. Box that

> *All models are wrong but some are useful.*

We present an examination of the finite sample performance of likelihood based estimation procedures in the context of bivariate binary outcome, binary treatment models. We compare the sampling properties of likelihood based estimators derived from different functional forms and we evaluate the impact of functional form miss-specification on the performance of the maximum likelihood estimator derived from the bivariate probit model. We also investigate the practical importance of available instruments in both cases of correct and incorrect distributional specifications. We analyze the finite sample properties of the endogenous dummy variable and covariate coefficient estimates, and the correlation coefficient estimates, and we examine the existence of possible "compensating effects" between the latter and estimates of the ATE. Finally, we provide a bridge between the literature on the bivariate probit model and that on partial identification by demonstrating how the properties of likelihood based estimators are explicable via a link between the notion of pseudo-true parameter values and the concepts of partial identification.

The remainder of this paper is arranged as follows. Section 2 presents the basic bivariate model that forms the background to the paper, establishes notation, and uses this to present the recursive bivariate probit (RBVP) model frequently used in empirical studies, and to introduce the recursive bivariate skew-probit (RBVS-P) and recursive bivariate log-probit (RBVL-P) models. The latter models represent new specifications in their own right, and they are used in Section 3

to study the finite sample performance of associated maximum likelihood inference when applied to both correctly specified and miss-specified models. Section 4 reviews the properties of maximum likelihood estimates of correctly specified and miss-specified models from the perspective of partial identification. Section 5 summarises this paper.

2 The Econometric Framework

The basic specification that we will consider here is a recursive bivariate model characterized by a structural equation determining a binary outcome as a function of a binary treatment variable where the binary treatment, or dummy, variable is in turn governed by a reduced form equation:

$$\begin{aligned} Y^* &= \mathbf{X}'\boldsymbol{\beta}_Y + D\alpha + \varepsilon_1\,, \qquad Y = \mathbb{I}(Y^* > 0); \\ D^* &= \mathbf{X}'\boldsymbol{\beta}_D + \mathbf{Z}'\boldsymbol{\gamma} + \varepsilon_2\,, \qquad D = \mathbb{I}(D^* > 0), \end{aligned} \tag{1}$$

where $\mathbb{I}(\cdot)$ denotes the indicator function. In (1), $\mathbf{X}$ contains the common covariates and $\mathbf{Z}$ contains the instruments. The underlying continuous latent variables Y^* and D^* are mapped into the observed outcome Y and the observed (potentially endogenous) regressor D via threshold crossing conditions, and the joint distribution of Y and D conditional on $\mathbf{X}$ and $\mathbf{Z}$, $P(Y = \mathrm{y}, D = \mathrm{d}|\mathbf{X} = \mathbf{x}, \mathbf{Z} = \mathbf{z})$, which for notational convenience we abbreviate to P^{yd}, therefore has four elements:

$$\begin{aligned} P^{11} &= P(\varepsilon_1 > -\mathbf{x}'\boldsymbol{\beta}_Y - \alpha, \varepsilon_2 > -\mathbf{x}'\boldsymbol{\beta}_D - \mathbf{z}'\boldsymbol{\gamma})\,, \\ P^{10} &= P(\varepsilon_1 > -\mathbf{x}'\boldsymbol{\beta}_Y, \varepsilon_2 < -\mathbf{x}'\boldsymbol{\beta}_D - \mathbf{z}'\boldsymbol{\gamma})\,, \\ P^{01} &= P(\varepsilon_1 < -\mathbf{x}'\boldsymbol{\beta}_Y - \alpha, \varepsilon_2 > -\mathbf{x}'\boldsymbol{\beta}_D - \mathbf{z}'\boldsymbol{\gamma}) \quad \text{and} \\ P^{00} &= P(\varepsilon_1 < -\mathbf{x}'\boldsymbol{\beta}_Y, \varepsilon_2 < -\mathbf{x}'\boldsymbol{\beta}_D - \mathbf{z}'\boldsymbol{\gamma})\,. \end{aligned} \tag{2}$$

The probabilities in (2) are fully determined once a joint distribution for ε_1 and ε_2 has been specified, and given data consisting of N observations $(y_i, d_i, \mathbf{x}_i', \mathbf{z}_i')$ for $i = 1, \ldots, N$, the

log-likelihood function can then be calculated as

$$L(\boldsymbol{\theta}) = \sum_{i=1}^{N} \log P^{y_i d_i}(\boldsymbol{\theta}) \tag{3}$$

where $P^{y_i d_i}(\boldsymbol{\theta})$ denotes the probabilities in (2) evaluated at the point $(y_i, d_i, \mathbf{x}_i', \mathbf{z}_i')$ and emphasizes the dependence of the probabilities on the parameter $\boldsymbol{\theta}$, which contains the coefficients $\boldsymbol{\beta}_D$, $\boldsymbol{\beta}_Y$, $\boldsymbol{\gamma}$, and α, and other unknown parameters of the joint distribution of ε_1 and ε_2 that need to be estimated from the data.[1]

2.1 The Bivariate Probit Model

In the bivariate probit model it is assumed that $(\varepsilon_1, \varepsilon_2)$ is drawn from a standard bivariate normal distribution with zero means, unit variances, and correlation coefficient ρ:

$$(\varepsilon_1, \varepsilon_2) \sim \mathcal{N}_2 \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \right) . \tag{4}$$

The specification in (1) and (2) together with the assumption in (4) is commonly referred to as the recursive bivariate probit (RBVP) model. The joint distribution function of ε_1 and ε_2 in the RBVP model is therefore $\Phi_2(\varepsilon_1, \varepsilon_2; \rho)$ where $\Phi_2(\cdot, \cdot; \rho)$ denotes the cumulative distribution function of the bivariate standard normal distribution with coefficient of correlation ρ. In this case, the joint probability function of Y and D can be expressed compactly as

$$P^{yd}(\boldsymbol{\theta}) = \Phi_2(t_1, t_2; \rho^*) \tag{5}$$

where

$$t_1 = (2\mathrm{y} - 1)(\mathbf{x}'\boldsymbol{\beta}_Y + \mathrm{d}\,\alpha), \quad t_2 = (2\mathrm{d} - 1)(\mathbf{x}'\boldsymbol{\beta}_D + \mathbf{z}'\boldsymbol{\gamma}) \quad \text{and} \quad \rho^* = (2\mathrm{y} - 1)(2\mathrm{d} - 1)\rho\,,$$

[1] The model structure considered here corresponds to Heckman's case 4, see Maddala (1983, Model 6, page 122) and also Greene (2012, specification 21-41, page 710).

and the log-likelihood function for the RBVP model can be written as

$$\log L(\boldsymbol{\theta}) = \sum_{i=1}^{N} \log \Phi_2(t_{1i}, t_{2i}; \rho_i^*), \tag{6}$$

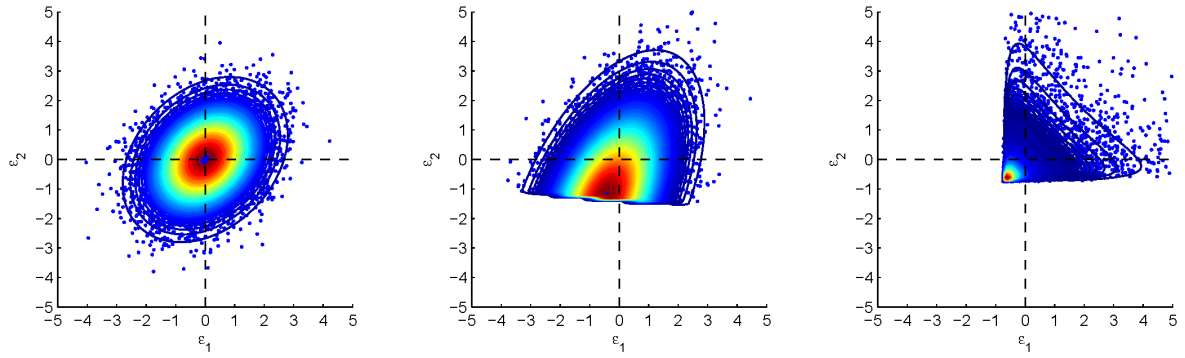
where for $i = 1, \ldots, N$, $t_{1i} = (2\mathrm{y}_i - 1)(\mathbf{x}_i'\boldsymbol{\beta}_Y + \mathrm{d}_i\,\alpha)$, $t_{2i} = (2\mathrm{d}_i - 1)(\mathbf{x}_i'\boldsymbol{\beta}_D + \mathbf{z}_i'\boldsymbol{\gamma})$ and $\rho_i^* = (2\mathrm{y}_i - 1)(2\mathrm{d}_i - 1)\rho$, with the subscript i indicating the ith unit observed in the sample.

Heckman (1978) noted that full rank of the regressor matrix is a sufficient condition for the identification of the model parameters in simultaneous equation models with endogenous dummy variables, but this has often not been recognized in the applied literature following a misleading statement in Maddala (1983, page 122) suggesting that the parameters of the structural equation are not identified in the absence of exclusion restrictions. Wilde (2000), however, notes that Maddala's statement is only valid when $\mathbf{X}$ and $\mathbf{Z}$ are both constants, and shows that as long as both equations of the model contain a varying exogenous regressor then full rank of the matrix of regressors is a sufficient condition for identification. Wilde's arguments make it clear that identification in the RBVP model does not require the presence of additional IVs in the reduced form equation, but in the absence of additional instruments identification strongly relies on functional form, i.e. normality of the stochastic disturbances.

2.2 Alternative Functional Forms

Given that in the absence of instruments the RBVP model relies on identification via functional form it is natural to ask: (i) What are the effects of using different functional forms, i.e. alternative specifications for the joint distribution of ε_1 and ε_2? and (ii) How does the presence of exclusion restrictions in the model influence the estimation of the model parameters? To address these issues we will examine the consequences of supplementing the basic model in (1) with two completely different specifications for the joint distribution of ε_1 and ε_2 that determines the probabilities in (2), namely the standardized bivariate skew-normal distribution (Azzalini and Dalla Valle, 1996) and the standardized bivariate log-normal distribution (Johnson et al., 1995).

Figure 1 presents a contour plot of the joint probability density function for each of the three different distributions – the standardized bivariate normal, standardized bivariate skew-normal and the standardized bivariate log-normal – overlaid with scatter plots of $N = 1000$ independent and identically distributed (i.i.d.) observations $(\varepsilon_{1i}, \varepsilon_{2i})$ $i = 1, \ldots, N$. For each distribution the marginal distributions of ε_1 and ε_2 have zero means and unit variances, and the correlation coefficient between ε_1 and ε_2 is $\rho = 0.3$. It is apparent that the skew-normal and log-normal distributions do not share the spherical symmetry of the normal distribution and that they produce probability structures that are very different from that generated by the normal distribution.



(a) Standard bivariate normal distribution

(b) Standardised bivariate skew-normal distribution

(c) Standardised bivariate log-normal distribution

Figure 1. Contour and scatter plots of $(\varepsilon_1, \varepsilon_2)$ from bivariate normal, skew-normal and log-normal distributions with zero means, unit variances and correlation coefficient $\rho = 0.3$.

2.2.1 The Bivariate Skewed-Probit Model

A k-dimensional random vector $\mathbf{U}$ is said to have a multivariate skew-normal distribution, denoted by $\mathbf{U} \sim SN_k(\boldsymbol{\mu}, \boldsymbol{\Omega}, \boldsymbol{\alpha})$, if it is continuous with density function

$$2\, \phi_k(\mathbf{u}; \boldsymbol{\mu}, \boldsymbol{\Omega})\, \Phi\left(\boldsymbol{\alpha}' \boldsymbol{\Lambda}^{-1} (\mathbf{u} - \boldsymbol{\mu})\right), \qquad \mathbf{u} \in \mathbb{R}^k, \tag{7}$$

where $\phi_k(\cdot; \boldsymbol{\mu}, \boldsymbol{\Omega})$ is the k-dimensional normal density with mean vector $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Omega}$, $\Phi(\cdot)$ is the standard normal cumulative distribution function, $\boldsymbol{\Lambda}$ is the diagonal matrix formed from the standard deviations of $\boldsymbol{\Omega}$, $\boldsymbol{\Lambda} = \text{diag}(\boldsymbol{\Omega})^{\frac{1}{2}}$, and $\boldsymbol{\alpha}$ is a k-

dimensional skewness parameter. Azzalini and Capitanio (1999) showed that if

$$\begin{bmatrix} V_0 \\ \mathbf{V} \end{bmatrix} \sim \mathcal{N}_{k+1}(\mathbf{0}, \boldsymbol{\Omega}^*) \,,$$

where V_0 is a scalar component,

$$\boldsymbol{\Omega}^* = \begin{bmatrix} 1 & \boldsymbol{\delta}^{\mathrm{T}} \\ \boldsymbol{\delta} & \boldsymbol{R} \end{bmatrix}$$

is a positive definite correlation matrix and

$$\boldsymbol{\delta} = \frac{\boldsymbol{R}\boldsymbol{\alpha}}{(1 + \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{R}\boldsymbol{\alpha})^{1/2}} \,, \tag{8}$$

then

$$\mathbf{U} = \begin{cases} \mathbf{V}, & \text{if } V_0 > 0, \\ -\mathbf{V}, & \text{otherwise.} \end{cases}$$

has a skew-normal distribution $SN_k(\mathbf{0}, \boldsymbol{R}, \boldsymbol{\alpha})$ where

$$\boldsymbol{\alpha} = \frac{\boldsymbol{R}^{-1}\boldsymbol{\delta}}{(1 - \boldsymbol{\delta}^{\mathrm{T}}\boldsymbol{R}^{-1}\boldsymbol{\delta})^{1/2}} \,.$$

The random vector $\boldsymbol{\mu} + \boldsymbol{\Lambda}\mathbf{U}$ then has the density function in (7) with location parameter $\boldsymbol{\mu}$ and dispersion parameter $\boldsymbol{\Omega} = \boldsymbol{\Lambda}\boldsymbol{R}\boldsymbol{\Lambda}$. The mean vector and variance-covariance matrix of $\mathbf{U}$ are

$$\boldsymbol{\mu}_{\mathbf{u}} = \mathbb{E}(\mathbf{U}) = (\frac{2}{\pi})^{1/2}\boldsymbol{\delta}, \quad \text{and} \quad \mathbb{V}ar(\mathbf{U}) = \boldsymbol{R} - \boldsymbol{\mu}_{\mathbf{u}}\boldsymbol{\mu}_{\mathbf{u}}^{\mathrm{T}},$$

and the correlation coefficient between U_i and U_j is

$$\rho_{ij} = \frac{\varrho_{ij} - 2\pi^{-1}\delta_i\delta_j}{\sqrt{(1 - 2\pi^{-1}\delta_i^2)(1 - 2\pi^{-1}\delta_j^2)}}, \tag{9}$$

where δ_i, δ_j and ϱ_{ij} are the elements in $\boldsymbol{\delta}$ and $\boldsymbol{R}$ respectively.

The above characterization provides a straightforward way to both generate skew-normal random variables and to evaluate the probabilities in (2) needed to calculate the likelihood function if the disturbances in the basic model (1) follow a normalized and standardized bivariate skew-

normal distribution. Define the random vector $\mathbf{W} = \boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbf{U} - \boldsymbol{\mu}_{\mathbf{u}})$ where $\boldsymbol{\Sigma}$ is the diagonal matrix with leading diagonal $\mathrm{diag}(\boldsymbol{R} - \boldsymbol{\mu}_{\mathbf{u}}\boldsymbol{\mu}_{\mathbf{u}}^{\mathrm{T}})$, then the elements of $\mathbf{W}$ have zero mean and unit variance and the distribution function of $\mathbf{W}$ is given by

$$
\begin{aligned}
P(\mathbf{W} \leq \mathbf{u}) &= P\left\{\boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbf{U} - \boldsymbol{\mu}_{\mathbf{u}}) \leq \mathbf{u}\right\} \\
&= P\left\{\mathbf{U} \leq (\boldsymbol{\mu}_{\mathbf{u}} + \boldsymbol{\Sigma}^{\frac{1}{2}}\,\mathbf{u})\right\} \\
&= P\left\{\mathbf{V} \leq (\boldsymbol{\mu}_{\mathbf{u}} + \boldsymbol{\Sigma}^{\frac{1}{2}}\,\mathbf{u}) | V_0 > 0\right\} \\
&= \frac{P\left\{\mathbf{V} \leq (\boldsymbol{\mu}_{\mathbf{u}} + \boldsymbol{\Sigma}^{\frac{1}{2}}\,\mathbf{u}),\ V_0 > 0\right\}}{P(V_0 > 0)} \\
&= 2P\left\{\begin{bmatrix} -V_0 \\ \mathbf{V} \end{bmatrix} \leq \begin{bmatrix} 0 \\ \boldsymbol{\mu}_u + \boldsymbol{\Sigma}^{\frac{1}{2}}\,\mathbf{u} \end{bmatrix}\right\} \\
&= 2\Phi_{k+1}\left(\begin{bmatrix} 0 \\ \boldsymbol{\mu}_{\mathbf{u}} + \boldsymbol{\Sigma}^{\frac{1}{2}}\,\mathbf{u} \end{bmatrix}; \begin{bmatrix} 1 & -\boldsymbol{\delta}^{\mathrm{T}} \\ -\boldsymbol{\delta} & \boldsymbol{R} \end{bmatrix}\right).
\end{aligned}
\tag{10}
$$

As a result, the probabilities in (2) that enter into the likelihood function of what we will label the recursive bivariate skew-probit (RBVS-P) model can be readily calculated as

$$
\begin{aligned}
P^{yd}(\boldsymbol{\theta}) &= 2\,\Phi_3\left(\begin{bmatrix} 0 \\ t_1 \\ t_2 \end{bmatrix}, \begin{bmatrix} 1 & \delta_1^* & \delta_2^* \\ \delta_1^* & 1 & \varrho_{12}^* \\ \delta_2^* & \varrho_{12}^* & 1 \end{bmatrix}\right) \\
&= 2\Phi_3(t_0, t_1, t_2, \Sigma^*)\,,
\end{aligned}
\tag{11}
$$

where $t_0 \equiv 0$,

$$
\begin{aligned}
t_1 &= (2\mathrm{y} - 1)(-\mu_{u1} + \sigma_{u1}(\mathbf{x}'\boldsymbol{\beta}_Y + \mathrm{d}\alpha))\,, \\
t_2 &= (2\mathrm{d} - 1)(-\mu_{u2} + \sigma_{u2}(\mathbf{x}'\boldsymbol{\beta}_D + \mathbf{z}'\boldsymbol{\gamma})) \quad \text{and} \\
\delta_1^* &= (2\mathrm{y} - 1)\delta_1, \quad \delta_2^* = (2\mathrm{d} - 1)\delta_2, \quad \text{and} \quad \varrho_{12}^* = (2\mathrm{y} - 1)(2\mathrm{d} - 1)\varrho_{12}\,.
\end{aligned}
$$

2.2.2 The Bivariate Log-Probit Model

The k-dimensional vector $\mathbf{X} = (X_1, \ldots, X_k)'$ is said to be log-normally distributed with parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ if $\log \mathbf{X} = (\log X_1, \ldots, \log X_k)' \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Let $(V_1, V_2)'$ denote a pair of log-normally distributed variables. The related bivariate normally distributed variables,

$(U_1, U_2)'$ say, have the bivariate distribution

$$\begin{bmatrix} U_1 \\ U_2 \end{bmatrix} = \begin{bmatrix} \ln V_1 \\ \ln V_2 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_{n_1} \\ \mu_{n_2} \end{bmatrix}, \begin{bmatrix} \sigma^2_{n_1} & \rho_n \sigma_{n_1} \sigma_{n_2} \\ \rho_n \sigma_{n_1} \sigma_{n_2} & \sigma^2_{n_2} \end{bmatrix} \right).$$

Denote the mean vector and variance-covariance matrix of $(V_1, V_2)'$ by

$$\mathbb{E}\begin{bmatrix} V_1 \\ V_2 \end{bmatrix} = \begin{bmatrix} \mu_{l_1} \\ \mu_{l_2} \end{bmatrix}, \quad \text{and} \quad \mathbb{V}ar\begin{bmatrix} V_1 \\ V_2 \end{bmatrix} = \begin{bmatrix} \sigma^2_{l_1} & \rho_l \sigma_{l_1} \sigma_{l_2} \\ \rho_l \sigma_{l_1} \sigma_{l_2} & \sigma^2_{l_2} \end{bmatrix},$$

where we have used the subscripts l and n, respectively, to distinguish the mean, variance, and correlation coefficient of the log-normal distribution from those of the normal distribution. Then the relationship between the parameters of the log-normal distribution and the moments of the normally distributed variables is as follows:

$$\begin{aligned} \mu_{n_1} &= \ln \mu_{l_1} - \frac{1}{2}\sigma^2_{n_1}, \qquad \sigma_{n_1} = \sqrt{\ln(1 + \sigma^2_{l_1}/\mu^2_{l_1})}, \\ \mu_{n_2} &= \ln \mu_{l_2} - \frac{1}{2}\sigma^2_{n_2}, \qquad \sigma_{n_2} = \sqrt{\ln(1 + \sigma^2_{l_2}/\mu^2_{l_2})}, \\ \rho_n &= \frac{\ln\left(1 + \rho_l \sqrt{(e^{\sigma^2_{n_1}} - 1)(e^{\sigma^2_{n_2}} - 1)}\right)}{\sigma_{n_1} \sigma_{n_2}}. \end{aligned} \tag{12}$$

Now, if we first generate (U_1, U_2) from a bivariate normal distribution, then $(V_1, V_2) = (e^{U_1}, e^{U_2})$ will posses a bivariate log-normal distribution. We can then standardize (V_1, V_2) to obtain $(\varepsilon_1, \varepsilon_2)$ where $\varepsilon_1 = (V_1 - \mu_{l_1})/\sigma_{l_1}$ and $\varepsilon_2 = (V_2 - \mu_{l_2})/\sigma_{l_2}$, so that $(\varepsilon_1, \varepsilon_2)$ will have a bivariate log-normal distribution with zero means and unit variances.

The above relationships can be exploited in a similar manner to construct the probabilities in (2) needed to determine the likelihood function of the base model in (1). For such a model, which we will christen the recursive bivariate log-probit (RBVL-P) model, the probabilities entering the likelihood function are given by

$$P^{yd}(\boldsymbol{\theta}) = \Phi_2(t_1, t_2, \rho^*) \tag{13}$$

where

$$t_1 = (2\mathrm{y} - 1)\left(\frac{\mu_{n_1} - \ln\left[\mu_{l_1} - (\mathbf{x}'\boldsymbol{\beta}_Y + \mathrm{d}\alpha)\sigma_{l_1}\right]}{\sigma_{n_1}}\right),$$

$$t_2 = (2\mathrm{d} - 1)\left(\frac{\mu_{n_2} - \ln\left[\mu_{l_2} - (\mathbf{x}'\boldsymbol{\beta}_D + \mathbf{z}'\boldsymbol{\gamma})\sigma_{l_2}\right]}{\sigma_{n_2}}\right) \quad \text{and}$$

$$\rho^* = (2\mathrm{y} - 1)(2\mathrm{d} - 1)\rho_n.$$

2.3 Maximum Likelihood and Quasi Maximum Likelihood Estimation

Comparing the expressions for the probabilities required to calculate the likelihood functions of the RBVS-P model and the RBVL-P model in (11) and (13) with those for the RBVP model in (5), we can see that they are all couched in terms of normal distribution functions that give the same variance-covariance structure for $(\varepsilon_1, \varepsilon_2)$ but different shifted and re-scaled conditional mean values derived from the structural and reduced form equations and latent variable threshold crossing rules of the basic model. Hence we can conclude that arguments that parallel those employed in Heckman (1978) and Wilde (2000) can be carried over to the RBVS-P and RBVL-P models to show that the RBVP, RBVS-P and RBVL-P models will all be identified in the absence of exclusion constraints if the matrix of regressors has full rank. Furthermore, following the argument in Greene (2012, pages 715-716) it can be shown that the terms that enter the likelihood functions are the same as those that appear in the corresponding bivariate probability model, the appearance of the dummy variable in the reduced form equation notwithstanding. The endogenous nature of the dummy regressor in the structural equation can therefore be ignored in formulating the likelihood and thus the models can be consistently and fully efficiently estimated using the maximum likelihood estimator (MLE).

Consistency and efficiency of the MLE is contingent, however, on the presumption that the model fitted to the data coincides with the true data generating process (DGP). If the model is miss-specified optimality properties of the MLE can no longer be guaranteed. Nevertheless, the likelihood function of a model can still be used to construct an estimator, and the resulting quasi-maximum likelihood estimator (QMLE) will be consistent for the pseudo-true parameter and asymptotically normal (White, 1982), and certain optimality features can also be ascer-

tained given suitable regularity (see Heyde, 1997, for detailed particulars). In what follows we will investigate the possible consequences that might arise when estimating of the recursive bivariate model parameters and basing the identification on an assumed functional form. In particular, we will examine properties of the QMLE constructed using the RBVP model when applied to data derived from DGPs that correspond to RBVS-P and RBVL-P processes.

3 Finite Sample Performance

In this section we present the results of Monte Carlo experiments designed to provide some evidence on the finite sample performance of the MLE for correctly specified models, and the QMLE for miss-specified models. For each set of the experiments we consider two primary designs for the exogenous regressors. The first design corresponds to the absence of exclusion restrictions in the process generating the data, and possible problems in the empirical identification of the parameters in this case are the genuine consequence of basing the identification only on the assumed functional form. In the first design there are no IVs in the model so it is only the matrix of exogenous regressors $\mathbf{X}$ that appears in the endogenous treatment equation and the outcome equation. We chose a continuous variable in $\mathbf{X}$ to mimic variables such as age and income, and also a dummy variable to represent a qualitative characteristic of the type that might be present in empirical applications. The exogenous regressors in $\mathbf{X}$ were drawn from $(X_1, X_2, X_3)' = \Big(1, \log(100 \times U), I(V > .25)\Big)'$ where U and V are independent and uniformly distributed in the unit interval. In the second design we introduce additional IVs $\mathbf{Z}$ into the endogenous dummy or treatment variable equation. The instruments were generated from the standard normal distribution, $Z_0 \sim \mathcal{N}(0, 1)$, and two Bernoulli random variables Z_1 and Z_2 with means equal to $P(Z_1 = 1) = 0.3$ and $P(Z_2 = 1) = 0.7$. The instrument Z_0 mimics a continuous variable, and Z_1 and Z_2 reflect that it has been common practice to use additional qualitative characteristics as instruments. In the simulations a set of N i.i.d. draws of $(\mathbf{X}, \mathbf{Z})$ was generated once and subsequently held fixed, then D and Y were generated via the latent variable equations and indicator functions.

There are three sets of experiments; in the first set the error terms are drawn from a bivariate

normal distribution, in the second set the error terms are drawn from a skew-normal distribution, and the errors were drawn from a log-normal distribution in the third set of experiments. To mimic various degrees of endogeneity the correlation coefficient ρ was varied from -0.9 to 0.9 in steps of length 0.2 in each design. The parameters in the structural and reduced form equations were set at $\boldsymbol{\beta}_Y = (0.6, 0.3, -2)'$, $\alpha = 0.6$ and $\boldsymbol{\beta}_D = (-1, 1, -3)'$ for the first design. For the second design $\boldsymbol{\beta}_Y = (0.6, 0.3, -2)'$, $\alpha = 0.6$, $\boldsymbol{\beta}_D = (-1, 1, -3)'$, and the coefficients on the instruments were set at $\gamma_{Z_0} = -0.6$, $\gamma_{Z_1} = 0.6$, and $\gamma_{Z_2} = -0.6$. The parameter values were chosen in such a way that both binary outcome variables Y and D have a distribution that is roughly "balanced" between the two outcomes (approximately half 0's and half 1's) in both designs. In this way we achieve maximum variation in the data and avoid problems that might be caused by having excessive numbers of zeros or ones in the observed response variables. In order to investigate the impact of sample size on the parameter estimates we included three different sample sizes in the simulations: $N = 1000$, 10000, and 30000.[2]

For each set of experiments and each design we generated $R = 1000$ replications, and in each experiment we derived the coefficient estimates, the predicted probabilities, and the estimated ATE. The estimation was undertaken in the Gauss matrix programming language, using the CML maximum likelihood estimation add-in module.[3] We summarize the results by presenting the true values, the coefficient estimates averaged over the R replications, the root mean square error of the R estimates relative to the true values (RMSE), as well as the empirical coverage probabilities (CP), measured as the percentages of times the true value falls within estimated 95% confidence intervals.

3.1 Performance of the MLE

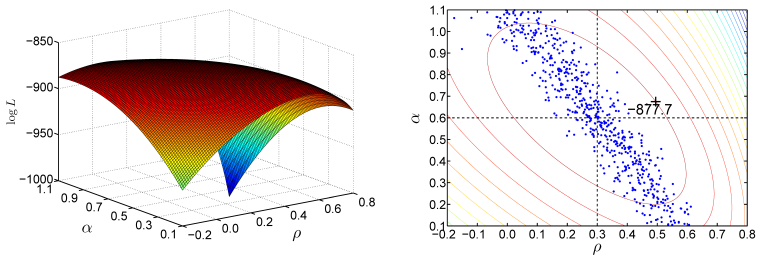
In order to gain some insight into the identification of the model parameters and the impact of including additional instruments in correctly specified models the log-likelihood function

[2]To aid in making comparisons among experiments from different designs, the samples were generated using random numbers drawn from the same random number seed.

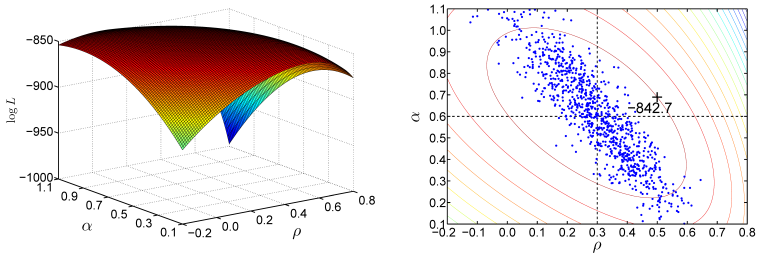
[3]The CML package provides for the estimation of statistical models by maximum likelihood while allowing for the imposition of general constraints on the parameters, see http://www.aptech.com/products/gauss-applications/constrained-maximum-likelihood-mt/

was calculated for the basic model both with and without additional IVs. Note that the first equation of the base model in (1) gives the conditional probability of Y given D, and the recursive bivariate model thereby introduces two sources of dependence between Y and D via the parameters α and ρ. The correlation coefficient between the error terms ε_1 and ε_2 acts as a measure of the endogeneity of the binary treatment variable D in the outcome equation for Y, and is of course another parameter that has to be estimated, but statistical independence of the structural and reduced form errors ($\rho = 0$) does not imply that Y and D are functionally independent. Full independence of Y and D requires that both $\rho = 0$ and $\alpha = 0$. In what follows we therefore focus on the structural equation treatment parameter α and the correlation parameter ρ and, due to space considerations, we do not present detailed results for $\boldsymbol{\beta}_Y$, $\boldsymbol{\beta}_D$, and γ. We also only provide simulation results for the case of moderate endogeneity with $\rho = 0.3$.

Figure 2 provides a graphical representation of the outcomes obtained for the RBVP model when $N = 1000$, with the log-likelihood plotted as a function of α and ρ, with $\boldsymbol{\beta}_Y$, $\boldsymbol{\beta}_D$ and γ set equal to their true values. In Figure 2 the left hand panel graphs $\bar{L}(\boldsymbol{\theta})$, the average



(a) $N = 1000$, without IV



(b) $N = 1000$, with IV

Figure 2. Comparison of log-likelihood surfaces and their contours. Correctly specified RBVP model, $\alpha = 0.6$ and $\rho = 0.3$. Numerical values of the highest contour level are displayed next to the line.

value of $L(\boldsymbol{\theta})$ observed in the R replications, plotted as a three-dimensional surface, and the right hand panel plots the contours of $\bar{L}(\boldsymbol{\theta})$ in the (ρ, α) plain with a scatter plot of the R pairs $(\hat{\rho}, \hat{\alpha})$ superimposed. Apart from a difference in overall level, the profile of the log-likelihood surface and the density of the log-likelihood contours of both designs exhibit similar features to each other. In both cases the estimates $(\hat{\rho}, \hat{\alpha})$ are spread around the maximum of the log-likelihood surface with a marked negative correlation, as would be expected since ρ measures the correlation between Y and D that remains after the influence of the regressors has been accounted for. In both designs the $(\hat{\rho}, \hat{\alpha})$ estimates are concentrated around the true value $(\rho, \alpha) = (0.3, 0.6)$, though the estimates for the second design, where additional IVs are included, are rather more densely packed around $(\rho, \alpha) = (0.3, 0.6)$ than are those in the first design. This suggests that even though the log-likelihood surfaces and their contours look similar for the two different designs, the addition of IVs can improve the estimation of the model parameters.[4]

Coefficient Estimates Table 1 presents the properties of MLE coefficient estimates for two correctly specified models, the RBVP model and the RBVL-P model. All the coefficient estimates for both the RBVP model and the RBVL-P model have very small biases, and the estimates obtained using the model with IVs generally have a smaller RMSEs than those obtained using the model without IVs (given the same sample size); the RMSEs of $\hat{\alpha}$ and $\hat{\rho}$ for the model with IVs are about one half to one third of those for the model without IVs. The coefficient estimates of the reduced form treatment equation (not presented in the table) for the model with IVs are only slightly better than those for the model without IVs, for the same sample size, whereas, estimates of the $\boldsymbol{\beta}_Y$ (also not presented in the table) in the model with IVs have much smaller RMSEs than those of the outcome equation in the model without IVs. It is perhaps worth noting that the use of additional IVs in the reduced form treatment equation can compensate for a lack of sample size; for example, the RMSE for $\hat{\alpha}$ from the RBVP model is 0.498 when $N = 1000$ and 0.176 when $N = 10000$ when no IVs are employed compared

[4]When the log-likelihood function is divided by the sample size N, changes in the log-likelihood surface and the log-likelihood contours from one sample size to the next as N is increased become virtually undetectable visually. This, of course, reflects that $N^{-1}L(\boldsymbol{\theta})$ will convergence to its expectation as $N \to \infty$.

Table 1. MLE coefficient estimates of α and ρ

True		Without IV			With IV		
		$N = 1000$	$N = 10000$	$N = 30000$	$N = 1000$	$N = 10000$	$N = 30000$
RBVP model							
$\alpha = .6$	$\bar{\alpha}$	.597	.603	.601	.607	.607	.603
	RMSE	(.498)	(.176)	(.115)	(.244)	(.078)	(.047)
	CP	.939	.947	.931	.244	.078	.047
$\rho = .3$	$\bar{\rho}$	.288	.298	.299	.295	.297	.298
	RMSE	(.279)	(.096)	(.063)	(.141)	(.046)	(.027)
	CP	.943	.948	.931	.947	.941	.938
RBVL-P model							
$\alpha = .6$	$\bar{\alpha}$	.560	.607	.598	.604	.602	.601
	RMSE	(.266)	(.103)	(.062)	(.129)	(.041)	(.022)
	CP	.904	.951	.937	.951	.941	.951
$\rho = .3$	$\bar{\rho}$	.339	.298	.302	.308	.301	.300
	RMSE	(.177)	(.067)	(.041)	(.094)	(.031)	(.017)
	CP	.920	.949	.942	.957	.941	.951

to 0.244 when $N = 1000$ and 0.078 when $N = 10000$ when IVs are used, the corresponding figures for the RBVL-P model are 0.266 when $N = 1000$ and 0.103 when $N = 10000$ when no IVs are employed compared to 0.129 when $N = 1000$ and 0.041 when $N = 10000$ when IVs are used.

Probabilities Table 2 presents the estimated predicted marginal probabilities $P(Y = 1)$ and $P(D = 1)$, and the estimated conditional probability $P(Y = 1|D = 1)$, constructed using the MLE coefficient estimates. Their RMSEs and CPs are also presented. It is apparent from the table that both correctly specified RBVP and RBVL-P models, with or without IVs, have generated accurate predicted probabilities with small RMSEs. The RMSEs and CPs for the predicted probabilities are also reasonably similar for the two models. These features presumably reflect that the MLE maximizes the log-likelihood function in Eq. (6) and thereby matches the probability of occurrence of the events via the observed relative frequencies, which will converge to the true probabilities as N increases.

Table 2. MLE Predicted probabilities

		Without IV				With IV		
	True	$N = 1000$	$N = 10000$	$N = 30000$	True	$N = 1000$	$N = 10000$	$N = 30000$
RBVP model								
$\bar{P}(Y = 1)$	.602	.611	.600	.602	.603	.612	.601	.602
RMSE		(.013)	(.004)	(.002)		(.013)	(.004)	(.002)
CP		.993	.997	1.000		.995	1.000	1.000
$\bar{P}(D = 1)$	.550	.559	.548	.550	.551	.562	.550	.549
RMSE		(.011)	(.004)	(.002)		(.011)	(.004)	(.002)
CP		.988	.993	.990		.996	.999	.993
$\bar{P}(Y = 1\|D = 1)$	.854	.860	.854	.853	.845	.850	.845	.845
RMSE		(.014)	(.004)	(.003)		(.014)	(.005)	(.003)
CP		.983	.991	.988		.987	.992	.996
RBVL-P model								
$\bar{P}(Y = 1)$	.516	.526	.513	.516	.524	.536	.522	.523
RMSE		(.012)	(.004)	(.002)		(.012)	(.004)	(.002)
CP		.987	.988	.999		.993	.996	1.000
$\bar{P}(D = 1)$	.487	.498	.484	.486	.516	.531	.515	.514
RMSE		(.011)	(.003)	(.002)		(.009)	(.003)	(.002)
CP		.960	.967	.976		.962	.994	.952
$\bar{P}(Y = 1\|D = 1)$	.902	.907	.902	.901	.862	.867	.860	.862
RMSE		(.012)	(.004)	(.002)		(.013)	(.004)	(.002)
CP		.969	.987	.982		.984	.995	.997

Table 3. MLE ATE estimates

		Without IV			With IV		
True		$N = 1000$	$N = 10000$	$N = 30000$	$N = 1000$	$N = 10000$	$N = 30000$
RBVP model							
.180	$\overline{\text{ATE}}$	.178	.181	.181	.179	.182	.181
	RMSE	(.149)	(.054)	(.036)	(.073)	(.024)	(.014)
	CP	.938	.946	.928	.943	.946	.936
RBVPL-P model							
.248	$\overline{\text{ATE}}$	.225	.251	.248	.241	.248	.249
	RMSE	(.103)	(.040)	(.024)	(.043)	(.015)	(.008)
	CP	.912	.953	.940	.961	.946	.972

Average Treatment Effect Table 3 shows the ATE of the binary endogenous treatment variable D on the binary outcome variable of interest Y for the two correctly specified models. From the results we can see that the ATE MLE estimates are very close to the true value irrespective of the presence of IVs, even for a relatively small sample size of 1000. However, the

RMSEs of the ATE estimates are quite different according to the presence or absence of IVs. In general the RMSE of the ATE estimate for the models with IVs is roughly one half of that for the models without IVs. This reflects the difference between the RMSEs of the coefficient estimates for the different models. The MLE estimates of the ATE for both models do not show any significant bias, however, implying that the variance of the ATE estimates for the model with IV is much lower than that of the estimates for the model without IVs.

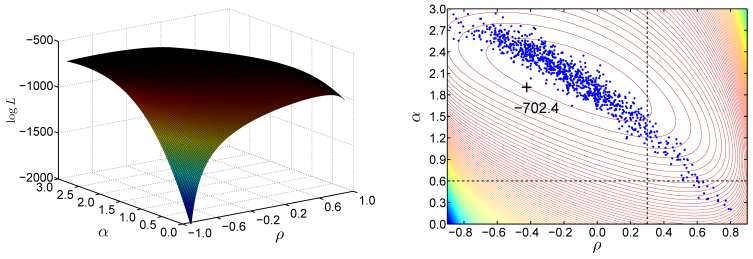
Remark: As previously observed in Fig. 1, the standardized bivariate log-normal distribution has a probability structure that is very different from that of the normal distribution, nevertheless, the qualitative characteristics of the MLE coefficient estimates, predicted probabilities, and the estimated ATEs of the RBVL-P model are not significantly differently from those of the RBVP model. Though not reported explicitly here, this invariance of the properties of the MLE estimates to the distributional specification was also observed with the RBVS-P model when the errors were generated via the standardized bivariate skew-normal distribution.

3.2 Performance of the QMLE

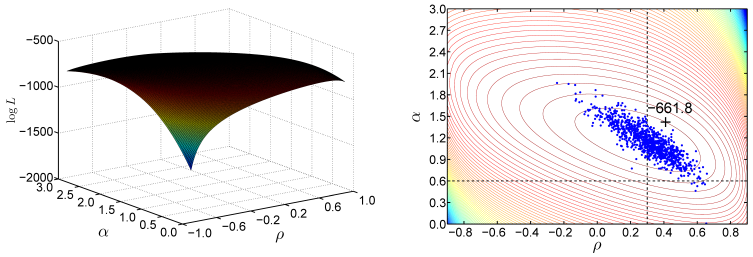
The former evidence was obtained by fitting a correctly specified model to the data via maximum likelihood, and the experimental results indicated that the finite sample properties of the MLE based on the RBVP, RBVL-P and RBVS-P models exhibited a qualitative invariance in both designs. The first design analyzed corresponds to the absence of exclusion restrictions in the DGP, i.e. Y and D depend on the same exogenous regressor $\mathbf{X}$. In the second design exclusion restrictions were imposed, i.e. in the process generating the data the dummy or treatment variable depends on the additional regressor $\mathbf{Z}$. We observed that although the presence of exclusion restrictions in the structural equation is not required for identification in these models, it is likely that the addition of instruments into the reduce form equation will improve the performance of the MLE. This suggests that the inclusion of exclusion restrictions might help in making the estimation results more robust to distributional miss-specification, an issue that we will examine here by investigating the performance of the QMLE obtained by fitting

the RBVP model, which is commonly used in applied studies, to data generated from RBVS-P and RBVL-P processes.

Figure 3 provides a counterpart to Figure 2 and depicts the log-likelihood functions of the RBVP model when calculated from data generated by a DGP corresponding to a RBVL-P model. As in Figure 2, the left hand panel graphs the average value of $\bar{L}(\boldsymbol{\theta})$ plotted as a three-



(a) $N = 1000$, without IV



(b) $N = 1000$, with IV

Figure 3. Comparison of log-likelihood surfaces and their contours. Incorrectly specified RBVP model with DGP corresponding to RBVL-P process, $\alpha = 0.6$ and $\rho = 0.3$. Numerical values of the highest contour level are displayed next to the line.

dimensional surface, and the right hand panel plots the contours of $\bar{L}(\boldsymbol{\theta})$ in the (ρ, α) plain with a scatter plot of the R pairs $(\hat{\rho}, \hat{\alpha})$ superimposed. Unlike the correctly specified case, the log-likelihood surface and the log-likelihood contours of the two designs have some distinct features. The log-likelihood from the first design is rather less peaked than that of the second design, and the estimates $(\hat{\rho}, \hat{\alpha})$ for the first design are more dispersed than those of the second. Perhaps the most striking feature of Figure 3, apart from a difference in the overall level of the log-likelihood surfaces, is that although in both cases the estimates $(\hat{\rho}, \hat{\alpha})$ are spread around the maximum of the log-likelihood surface with a marked negative correlation, the estimates $(\hat{\rho}, \hat{\alpha})$ in the first design are not concentrated around the true parameter value but deviate from $(\rho, \alpha) = (0.3, 0.6)$ by a considerable margin, whereas for the second design, where additional

IVs are included, $\hat{\alpha}$ deviates from $\alpha = 0.6$ by a much smaller margin and $\hat{\rho}$ is centered around $\rho = 0.3$. This indicates that although exclusion restrictions in the structural equation are not required for the statistical identification of the model parameters, the presence of additional instruments in the reduce form equation will help in making the QMLE more robust to the distributional miss-specification inherent in its evaluation.

Table 4. QMLE estimates of α and ρ

		Without IV			With IV		
True		$N = 1000$	$N = 10000$	$N = 30000$	$N = 1000$	$N = 10000$	$N = 30000$
$\alpha = .6$	$\bar{\alpha}$	2.020	2.029	1.895	1.151	1.176	1.186
	RMSE	(1.506)	(1.481)	(1.375)	(.598)	(.604)	(.623)
	CP	.215	.016	.010	.278	.008	.007
$\rho = .3$	$\bar{\rho}$	$-.158$	$-.153$	$-.075$	.306	.281	.278
	RMSE	(.342)	(.514)	(.459)	(.143)	(.107)	(.138)
	CP	.571	.064	.058	.912	.766	.596

Coefficient Estimates Table 4 presents a summary of the properties of the QMLE estimates of α and ρ when the RBVP model is fitted to data generated from a RBVL-P process. The most obvious feature is that the QMLE parameter estimators have an enormous bias and a large RMSE when there is no IV in the model. For example, without exclusion restrictions, the mean value of $\hat{\alpha}$ is 1.895, more than three times the true value of α, and its RMSE is 1.375, more than twenty times larger than the RMSE of the MLE of α, even when the sample size is 30000. The estimated value of $\hat{\rho}$ is actually negative when there are no IVs. The performance of the QMLE improves when additional IVs are introduced into the model: The mean value of $\hat{\alpha}$ is 1.151 when $N = 1000$, with a RMSE of 0.598, less than half of the value obtained when there are no exclusion constraints. The mean value of $\hat{\rho}$ has the correct sign and is quite close to the true value, with a relative RMSE $RMSE(\hat{\rho})/\rho = 47\%$ when $N = 1000$ compared to a value of 114% for the model without IVs.

Probabilities The properties of the QMLE estimates of the marginal probabilities $P(Y = 1)$ and $P(D = 1)$, and the conditional probability $P(Y = 1|D = 1)$, their average, RMSEs and CPs, are presented in Table 5. The most obvious feature to observe here is that despite the bias

and RMSEs of the QMLE coefficient estimates being large, the QMLE predicted probabilities are fairly close to the true probabilities, even when the sample size is small and there are no IVs in the model. This result can be attributed to the fact that, as with the MLE, in order to maximize the log-likelihood function in Eq. (6) the QMLE manipulates the parameters of the model so as to maximize the probability(likelihood) of occurrence of the data. But in order to match the predicted probability of occurrence derived from the wrong model with the observed relative frequency of events that comes from a different DGP the QMLE must "distort" the parameters of the model. By so doing the QMLE attempts to match the true probabilities as closely as possible, and the figures in Table 5 suggest that the QMLE does this reasonably precisely.

Table 5. QMLE predicted probabilities

		Without IV				With IV		
	True	$N = 1000$	$N = 10000$	$N = 30000$	True	$N = 1000$	$N = 10000$	$N = 30000$
$\bar{P}(Y = 1)$	.516	.528	.515	.519	.524	.540	.528	.529
RMSE		(.012)	(.018)	(.031)		(.012)	(.036)	(.018)
CP		.976	.960	.925		.982	.994	.884
$\bar{P}(D = 1)$	.487	.501	.486	.488	.516	.538	.523	.521
RMSE		(.011)	(.012)	(.012)		(.010)	(.022)	(.014)
CP		.950	.957	.940		.937	.930	.834
$\bar{P}(Y = 1\|D = 1)$	.902	.907	.902	.899	.862	.867	.859	.862
RMSE		(.012)	(.009)	(.037)		(.014)	(.033)	(.015)
CP		.966	.937	.865		.972	.939	.862

Table 6. QMLE ATE estimates

		$N = 1000$		$N = 10000$		$N = 30000$	
	True	No IV	IV	No IV	IV	No IV	IV
$\overline{\text{ATE}}$	.248	.519	.300	.534	.314	.499	.295
RMSE		(.124)	(.061)	(.099)	(.050)	(.116)	(.059)
CP		.371	.885	.023	.125	.018	.041

Average Treatment Effect The ATE estimates for the QMLE are presented in Table 6. From the table we can see that the QMLE estimate of the ATE is about twice the magnitude of the true value for the model without exclusion restrictions. When additional IVs are included in the model the QMLE ATE estimates are much closer to the true value, with a RMSE that is

about one half that achieved for the model without instruments. The most significant feature, however, is the collapse of the coverage probabilities relative to those seen for the MLE in Table 3.

Remark: When the errors in DGP are generated from the standardized bivariate skew-normal distribution the performance of the QMLE based upon the RBVP model is similar to that presented here for RBVP model QMLE applied to the DGP where the errors are generated from standardized bivariate log-normal distribution. The difference between the qualitative features of the QMLE and the MLE are somewhat smaller in the former case, especially when IVs are present, and this presumably reflects that the standardized bivariate skew-normal distribution does not deviate from the standard normal distribution by as much as does the standardized bivariate log-normal distribution, as can be seen from visual inspection of Figure 1.

3.3 Summary of Simulation Results

When the model is correctly specified the estimates of the model parameters, predicted probabilities, and ATE exhibit features that are consistent with the known properties of the MLE. The MLE estimates show little bias, even at the smallest sample size, and the relative RMSE can drop from approximately $40\,\%$ when $N = 1000$ to $20\,\%$ when $N = 10000$, and can be as low as 5.6% by the time $N = 30000$. Without IVs, the model can still be estimated with reasonable precision provided the sample size is sufficiently large, but in order to have precise coefficient estimates it seems desirable to have a sample size as large as 10000 even when there are exclusion restrictions in the model. Generally the use of IVs improves the identification and estimation of the model parameters *ceteris paribus*. For example, the RMSE of the estimates of α in Table 1 for models with IVs are a third to one half of those obtained for the model without IVs.

When the model is misspecified it is obvious that detailed particulars of the behaviour of the QMLE based upon the RBVP model will depend on the true DGP, nevertheless some general comments can be made: Although the performance of QMLE estimates need not be too dis-

similar from that of the MLE, the QMLE coefficient estimates can be heavily biased in finite samples and consequently the QMLE coefficient estimates have a significant RMSE. Moreover, whereas increases in sample size clearly improve the RMSE performance of the MLE, in accord with its known consistency and efficiency properties, such increases do not necessarily improve the RMSE performance of the QMLE. This is due to the fact that in order to maximize the likelihood of a miss-specified model the QMLE must "distort" the parameters of the model so as to match the probabilities of occurrence from the true DGP, and this results in the QMLE parameter estimates having a non-trivial asymptotic bias. The use of IVs in the model can dramatically improve the identification and estimation of the model parameters for the QMLE. For example, the estimates of ρ in Table 4 for models without IVs have an incorrect sign and magnitude, irrespective of sample size, but the model with IVs captures the sign and level of endogeneity correctly.

Although the QMLE produces asymptotically biased estimates of the parameters of the true DGP, the QMLE estimates of parametric functions such as the predicted probabilities and ATE perform surprisingly well. That the predicted probabilities and ATE estimates are not sensitive to the miss-specification of the error distribution, and that the RBVP QMLE is able to reproduce the true probabilities and ATE with reasonable accuracy despite the model being misspecified can be explained by linking the notion of pseudo-true parameters with the concept of partial identification, as we will show in the following section.[5]

4 Pseudo True Parameters and Partial Identification

Let $\mathrm{P}^{YD}(\boldsymbol{\theta}_0)$ denote the true probability distribution function of (Y, D) for given values of $\mathbf{X}$ and $\mathbf{Z}$, i.e. the probability distribution that characterizes the DGP, and set

$$K\left(\boldsymbol{\theta} : \boldsymbol{\theta}_0\right) = \mathbb{E}\left[\log\left\{\frac{\mathrm{P}^{YD}(\boldsymbol{\theta}_0)}{P^{YD}(\boldsymbol{\theta})}\right\}\right] = \sum_{y=0}^{1}\sum_{d=0}^{1}\log\left\{\frac{\mathrm{P}^{yd}(\boldsymbol{\theta}_0)}{P^{yd}(\boldsymbol{\theta})}\right\}\mathrm{P}^{yd}(\boldsymbol{\theta}_0)$$

[5] As with any Monte Carlo experiments the above results are conditional on the specific experimental design employed. More extensive simulation studies that encompass the experimental results presented here and add further experimental evidence supporting the conclusions reached above can be found in Li (2015).

where $P^{YD}(\boldsymbol{\theta})$ is the probability distribution function specified by the model to be fitted to the data. Then $K(\boldsymbol{\theta} : \boldsymbol{\theta}_0)$ equals the Kullback-Leibler divergence of the two distributions, and via an application of Jensen's inequality to $-\log(x)$ it can be shown that

(i) $K(\boldsymbol{\theta} : \boldsymbol{\theta}_0) \geq 0$ and

(ii) $K(\boldsymbol{\theta} : \boldsymbol{\theta}_0) = 0$ if and only if $P^{yd}(\boldsymbol{\theta}) = \mathrm{P}^{yd}(\boldsymbol{\theta}_0)$ for all $(y, d) \in \{0, 1\} \times \{0, 1\}$.

Now let $L(\boldsymbol{\theta})$ be defined as in Equation (6) and set $\mathrm{L}(\boldsymbol{\theta}) = \sum_{i=1}^{N} \log \mathrm{P}^{y_i d_i}(\boldsymbol{\theta})$, the log-likelihood function for the correctly specified model. Treating each log-likelihood as a random variable, a function of the random variables (Y_i, D_i), $i = 1, \ldots, N$, given $\mathbf{X} = \mathbf{x}_i$ and $\mathbf{Z} = \mathbf{z}_i$, $i = 1, \ldots, N$, and given values of theta, set

$$K_N(\boldsymbol{\theta} : \boldsymbol{\theta}_0) = \mathbb{E}\left[\mathrm{L}(\boldsymbol{\theta}_0) - L(\boldsymbol{\theta})\right] = \mathbb{E}\left[\log\left\{\frac{\prod_{i=1}^{N} \mathrm{P}^{Y_i D_i}(\boldsymbol{\theta}_0)}{\prod_{i=1}^{N} P^{Y_i D_i}(\boldsymbol{\theta})}\right\}\right].$$

Rearranging the products on the right hand side gives

$$K_N(\boldsymbol{\theta} : \boldsymbol{\theta}_0) = \sum_{i=1}^{N} \mathbb{E}\left[\log\left\{\frac{\mathrm{P}^{Y_i D_i}(\boldsymbol{\theta}_0)}{P^{Y_i D_i}(\boldsymbol{\theta})}\right\}\right] = \sum_{i=1}^{N} K_i(\boldsymbol{\theta} : \boldsymbol{\theta}_0) \geq 0\,.$$

It is a trivial exercise to verify that $K_N(\boldsymbol{\theta} : \boldsymbol{\theta}_0) = \sum_{i=1}^{N} \mathbb{E}\left[\log \mathrm{P}^{Y_i D_i}(\boldsymbol{\theta}_0)\right] - \mathbb{E}\left[\log P^{Y_i D_i}(\boldsymbol{\theta})\right]$ and we can therefore conclude that if $\boldsymbol{\theta}^* = \arg\min K_N(\boldsymbol{\theta} : \boldsymbol{\theta}_0)$ then $\mathbb{E}[L(\boldsymbol{\theta})]$ must be maximized at $\boldsymbol{\theta} = \boldsymbol{\theta}^*$.

Since, by the law of large numbers, $N^{-1}L(\boldsymbol{\theta})$ converges to $N^{-1}\mathbb{E}[L(\boldsymbol{\theta})]$ it follows that the MLE, which is constructed using the true probability distribution function, will converge to $\boldsymbol{\theta}^* = \boldsymbol{\theta}_0$, the true parameter value, as is well known. That the MLE is able to precisely reproduce the probability of occurrence of events determined by the DGP (as seen in Table 2) is then a consequence of the fact that asymptotically the MLE achieves the global minimum of the Kullback-Leibler divergence, namely zero. The QMLE, on the other hand, is based upon a misspecified model and it will converge to a pseudo-true value $\boldsymbol{\theta}^* \neq \boldsymbol{\theta}_0$. That the QMLE predicted probabilities can match the probabilities of the true DGP reasonably well, though not exactly and with various levels of accuracy (as seen in Table 5), reflects that the QMLE

minimizes the Kullback-Leibler divergence but $K_N(\boldsymbol{\theta}^* : \boldsymbol{\theta}_0) > 0$.

In order to link the Kullback-Leibler divergence and pseudo-true parameter to the constructs of partial identification recall the basic recursive bivariate model in (1). Define a new random variable $U = F_{\epsilon_1}(\varepsilon_1)$, where $F_{\epsilon_1}(\cdot)$ is the marginal distribution function of the stochastic error on the structural equation. Then U is uniformly distributed in the unit interval ($U \sim \text{Unif}(0,1)$) and from the assumed exogenuity of $\mathbf{X}$ and $\mathbf{Z}$ we have $U \perp\!\!\!\perp \mathbf{Z}|\mathbf{X}$. We can now define a structural function

$$h(D, \mathbf{X}, U) = \mathbb{I}[U \geq F_{\epsilon_1}(-\mathbf{X}\boldsymbol{\beta}_Y - D\alpha)]$$

in which h is weakly monotonic in U, $P(U \leq \tau|\mathbf{Z} = \mathbf{z}) = \tau$ for all $\tau \in (0,1)$ and all $\mathbf{z}$ in the support of $\mathbf{Z}$, $\Omega_{\mathbf{Z}}$ say, and

$$Y = h(D, \mathbf{X}, U) = \begin{cases} 0, & \text{if } 0 < U \leq p(D, \mathbf{X}) \\ 1, & \text{if } p(D, \mathbf{X}) < U \leq 1 \end{cases} \tag{14}$$

where the probability function of the structural equation is given by

$$p(D, \mathbf{X}) = 1 - F_{\epsilon_1}(-\mathbf{X}\boldsymbol{\beta}_Y - D\alpha)\,.$$

The specification in (14) satisfies the assumptions of the structural equation model with a binary outcome and binary endogenous variable as defined and discussed in Chesher (2010). Thus the linear index threshold crossing model in (1) is equivalent to a single equation structural model augmented with parametric assumptions and a specification for the endogenous dummy or treatment variable.

From Eq. (14) it is clear that the distribution of Y, given D and $\mathbf{X}$, is determined by the probability function $p(D, \mathbf{X})$. Suppose, for the sake of argument, that $p(D, \mathbf{X})$ is unknown.

Let

$$
\begin{aligned}
f_0(\mathbf{x},\mathbf{z}) &\equiv P[Y=0|\mathbf{X}=\mathbf{x},D=0,\mathbf{Z}=\mathbf{z}]\,,\\
f_1(\mathbf{x},\mathbf{z}) &\equiv P[Y=0|\mathbf{X}=\mathbf{x},D=1,\mathbf{Z}=\mathbf{z}]\,,\\
g_0(\mathbf{x},\mathbf{z}) &\equiv P[D=0|\mathbf{X}=\mathbf{x},\mathbf{Z}=\mathbf{z}] \quad \text{and}\\
g_1(\mathbf{x},\mathbf{z}) &\equiv P[D=1|\mathbf{X}=\mathbf{x},\mathbf{Z}=\mathbf{z}]\,,
\end{aligned}
\tag{15}
$$

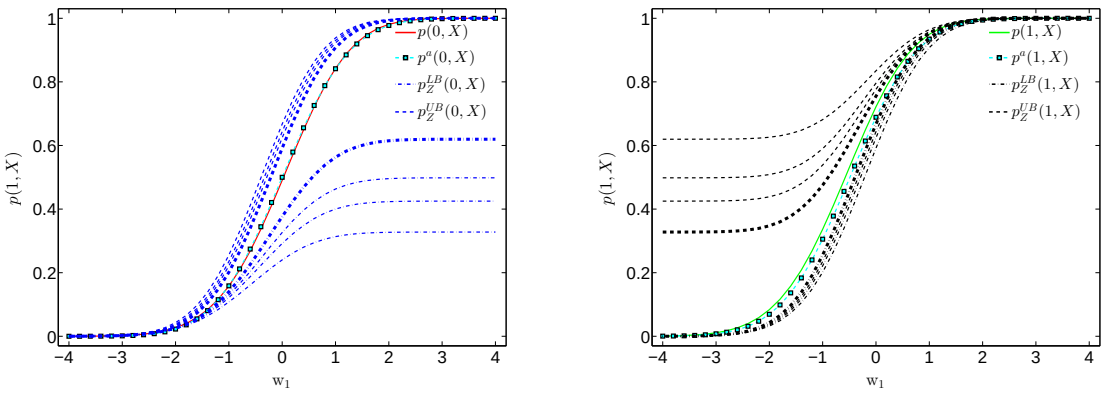
denote the stated conditional probabilities and, to state a standard convention, set $p(d,\mathbf{x}) = p(D,\mathbf{X})|_{(D=d,\mathbf{X}=\mathbf{x})}$. From the developments in Chesher (2010, Section 2) the following inequalities for $p(0,\mathbf{x})$ and $p(1,\mathbf{x})$ can be derived:

$$
\begin{aligned}
p(0,\mathbf{x}) < p(1,\mathbf{x}): \quad & f_0(\mathbf{x},\mathbf{z})g_0(\mathbf{x},\mathbf{z}) \le p(0,\mathbf{x}) \le f_0(\mathbf{x},\mathbf{z})g_0(\mathbf{x},\mathbf{z}) + f_1(\mathbf{x},\mathbf{z})g_1(\mathbf{x},\mathbf{z})\\
& \le p(1,\mathbf{x}) \le g_0(\mathbf{x},\mathbf{z}) + f_1(\mathbf{x},\mathbf{z})g_1(\mathbf{x},\mathbf{z}),\\
p(0,\mathbf{x}) \ge p(1,\mathbf{x}): \quad & f_1(\mathbf{x},\mathbf{z})g_1(\mathbf{x},\mathbf{z}) \le p(1,\mathbf{x}) \le f_0(\mathbf{x},\mathbf{z})g_0(\mathbf{x},\mathbf{z}) + f_1(\mathbf{x},\mathbf{z})g_1(\mathbf{x},\mathbf{z})\\
& \le p(0,\mathbf{x}) \le g_1(\mathbf{x},\mathbf{z}) + f_0(\mathbf{x},\mathbf{z})g_0(\mathbf{x},\mathbf{z}).
\end{aligned}
\tag{16}
$$

By taking the intersection of the intervals in (16) for different values of $\mathbf{z} \in \Omega_{\mathbf{Z}}$ the bounds on $p(1,\mathbf{x})$ and $p(0,\mathbf{x})$ can be tightened to give a least upper bound (l.u.b.) and a greatest lower bound (g.l.b.). Hence, if we presume that the DGP is characterized by a process that satisfies the model in (1) we can derive upper and lower bounds for the structural functions $p(0,\mathbf{x})$ and $p(1,\mathbf{x})$ via (16) by calculating the conditional probabilities in (15). Consequently, any alternative specification that generates a probability function lying between the intersection bounds of $p(0,\mathbf{x})$ and $p(1,\mathbf{x})$ across the support of the IVs will be observationally equivalent to that of the presumed model.

Figure 4 illustrates the partial identification of the probability function $p(D,\mathbf{X})$ when the true DGP corresponds to the RBVS-P model with parameters $\boldsymbol{\beta}_Y = (0.6, 0.3, -2)'$ and $\alpha = 0.6$ in the structural equation and $\boldsymbol{\beta}_D = (-1, 1, -3)'$, and $\gamma_{Z_0} = 0$, $\gamma_{Z_1} = 0.6$, and $\gamma_{Z_2} = -0.6$ in the reduced form. In this case $F_{\epsilon_1}(\varepsilon_1)$ is given by the marginal distribution function of ε_1 where $(\varepsilon_1, \varepsilon_2)'$ is generated from the bivariate skew-normal distribution. Figure 4 plots the probability function $p(D,\mathbf{X})$ constructed from the true RBVS-P DGP, together with its upper and lower

bounds. Figure 4(a) presents $p(0, \mathbf{x})$ and its upper and lower bounds plotted as functions of



(a) Upper and lower bounds of $p(0, \mathbf{X})$ (b) Upper and lower bounds of $p(1, \mathbf{X})$

Figure 4. Partial identification of probability function $p(D, \mathbf{X})$, RBVS-P process.

the linear index $\mathrm{w}_1 = \mathbf{x}'\boldsymbol{\beta}_Y$ when, without loss of generality, the index $\mathrm{w}_2 = \mathbf{x}'\boldsymbol{\beta}_D = 0$. The solid red curve plots $p(0, \mathbf{x})$ and the four upper bounds of $p(0, \mathbf{x})$ for the four different values of the pair (Z_1, Z_2) are plotted as dashed blue lines, and the corresponding lower bounds are plotted as dash-doted blue lines. The intersection of the upper and lower bounds is given by the region between the l.u.b., the bold blue dashed line, and the g.l.b., the bold blue dash-doted line. Similarly, Fig. 4(b) plots the true structural function $p(1, \mathbf{x})$ as a solid green curve, and the upper and lower bounds of $p(1, \mathbf{x})$ are plotted as dashed and dash-doted black lines respectively. The bold black lines mark the l.u.b. and the g.l.b., the borders of the bound intersection.

From the partial identification perspective, any properly defined probability function, $p'(D, \mathbf{X})$ say, such that $p'(0, \mathbf{x})$ lies between the two bold blue lines and $p'(1, \mathbf{x})$ lies between the two bold black lines is observationally equivalent to the true $p(D, \mathbf{X})$. The curves labeled $p^a(0, \mathbf{x})$ and $p^a(1, \mathbf{x})$, shown in cyan with square markers in Figure 4(a) and Figure 4(b), denote the probability functions derived from the RBVP model using QMLE parameter estimates. The parameter values used to construct $p^a(0, \mathbf{x})$ and $p^a(1, \mathbf{x})$ were taken as the average value of the QMLE estimates observed over $R = 1000$ replications when $N = 30,000$. It seems reasonable to conjecture that the latter parameter estimates should be close to $\boldsymbol{\theta}^*$. Both $p^a(0, \mathbf{x})$ and $p^a(1, \mathbf{x})$ fall into the regions enveloped by their respective l.u.b. and g.l.b., and the proximity of the estimated probability functions to the true $p(d, \mathbf{x})$ functions from the bivariate skew-

normal distribution indicate that $K_N(\boldsymbol{\theta}^* : \boldsymbol{\theta}_0)$ is close to zero. Thus we find that the QMLE generates a pseudo-true parameter value that minimizes the Kullback-Leibler divergence and thereby produces an observationally equivalent characterization that maximizes the proximity of the probability function constructed from the assumed model to the $p(D, \mathbf{X})$ function of the true process.

From the definition of the structural model it follows that the probability that Y is unity, given D and $\mathbf{X}$, equals $1 - p(D, \mathbf{X})$. Thus we have that $\mathbb{E}[Y|D = 1, \mathbf{X} = \mathbf{x}] = 1 - p(1, \mathbf{x})$ and $\mathbb{E}[Y|D = 0, \mathbf{X} = \mathbf{x}] = 1 - p(0, \mathbf{x})$, and the ATE for an individual with features characterized by $\mathbf{X} = \mathbf{x}$ is therefore

$$ATE(\mathbf{x}) = p(0, \mathbf{x}) - p(1, \mathbf{x})\,. \tag{17}$$

When $p(D, X)$ is unknown the inequalities in (16) can be used to bounded $ATE(\mathbf{x})$ by the interval

$$\Big[\sup_{\mathbf{z}\in\Omega_Z} f_0(\mathbf{x},\mathbf{z})g_0(\mathbf{x},\mathbf{z}) - \inf_{\mathbf{z}\in\Omega_Z}\Big\{g_0(\mathbf{x},\mathbf{z}) + f_1(\mathbf{x},\mathbf{z})g_1(\mathbf{x},\mathbf{z})\Big\},$$
$$\inf_{\mathbf{z}\in\Omega_Z}\Big\{f_0(\mathbf{x},\mathbf{z})g_0(\mathbf{x},\mathbf{z}) + f_1(\mathbf{x},\mathbf{z})g_1(\mathbf{x},\mathbf{z})\Big\} - \sup_{\mathbf{z}\in\Omega_Z}\Big\{f_0(\mathbf{x},\mathbf{z})g_0(\mathbf{x},\mathbf{z}) + f_1(\mathbf{x},\mathbf{z})g_1(\mathbf{x},\mathbf{z})\Big\}\Big] \tag{18}$$

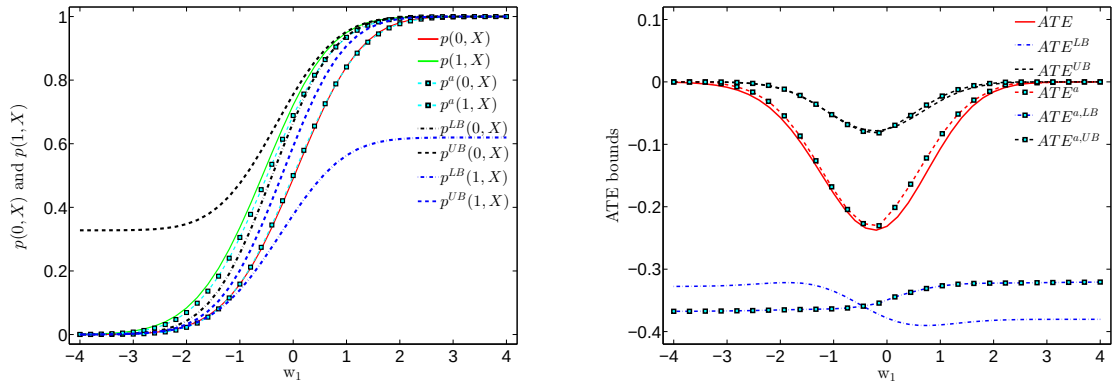
when $p(0, \mathbf{x}) \geq p(1, \mathbf{x})$, and

$$\Big[\sup_{\mathbf{z}\in\Omega_Z}\Big\{f_0(\mathbf{x},\mathbf{z})g_0(\mathbf{x},\mathbf{z}) + f_1(\mathbf{x},\mathbf{z})g_1(\mathbf{x},\mathbf{z})\Big\} - \inf_{\mathbf{z}\in\Omega_Z}\Big\{f_0(\mathbf{x},\mathbf{z})g_0(\mathbf{x},\mathbf{z}) + f_1(\mathbf{x},\mathbf{z})g_1(\mathbf{x},\mathbf{z})\Big\},$$
$$\inf_{\mathbf{z}\in\Omega_Z}\Big\{g_1(\mathbf{x},\mathbf{z}) + f_0(\mathbf{x},\mathbf{z})g_0(\mathbf{x},\mathbf{z})\Big\} - \sup_{\mathbf{z}\in\Omega_Z} f_1(\mathbf{x},\mathbf{z})g_1(\mathbf{x},\mathbf{z})\Big]\,. \tag{19}$$

when $p(0, \mathbf{x}) < p(1, \mathbf{x})$.

Figure 5 illustrates the evaluation of the ATE and the ATE bounds using the example considered above to construct Figure 4, that is, a DGP corresponding to a RBVS-P model and the QMLE based upon a RBVP model. The true $p(d, \mathbf{x})$ probability functions and their intersection bounds as presented in Figure 4(a) and Figure 4(b) are reproduced superimposed on each other in

Figure 5(a). The l.u.b. and g.l.b. of $p(0,\mathbf{x})$ and $p(1,\mathbf{x})$ are denoted by $p^{\mathrm{UB}}(0,\mathbf{x})$ and $p^{\mathrm{UB}}(1,\mathbf{x})$, and $p^{\mathrm{LB}}(0,\mathbf{x})$ and $p^{\mathrm{LB}}(1,\mathbf{x})$, respectively. The resulting ATE and its upper and lower bounds are plotted in Figure 5(b). The red solid curve is the ATE value calculated from Equation (17) using the probability functions derived from the DGP, namely the RBVS-P process, while the black dashed curve and the blue dash-doted curve graph the corresponding upper and lower bounds of the ATE.



(a) Intersected upper and lower bounds of $p(0,\mathbf{x})$ and $p(1,\mathbf{x})$

(b) Upper and lower bounds of ATE

Figure 5. Intersected bounds of $p(0,\mathbf{x})$ and $p(1,\mathbf{x})$ and ATE bounds

In this example the ATE is negative since $p(0,\mathbf{x}) < p(1,\mathbf{x})$ and the ATE upper and lower bounds are calculated using Eq. (19). From an inspection of the inequalities in (16) and the formulation in Equation (19) it can be deduced that (in the notation of Figure 5) the mapping of the bounds in Figure 5(a) to those in Figure 5(b) is given by

$$\begin{aligned} \mathrm{ATE}^{\mathrm{LB}}(\mathbf{x}) &= p^{\mathrm{LB}}(0,\mathbf{x}) - p^{\mathrm{UB}}(1,\mathbf{x}) \quad \text{and} \\ \mathrm{ATE}^{\mathrm{UB}}(\mathbf{x}) &= p^{\mathrm{UB}}(0,\mathbf{x}) - p^{\mathrm{LB}}(1,\mathbf{x})\,. \end{aligned} \tag{20}$$

The red doted curve, and the black dashed curve and the blue dash-doted curves shown with square markers graph the ATE and its upper and lower bounds calculated from the RBVP model using the QMLE parameter estimates. In this example, the QMLE probability function estimates are actually very close to the true $p(D,X)$ functions coming from the DGP bivariate skew-normal distribution, and as a result the QMLE ATE estimates are very close to the true

ATE values. For example, the ATE is -0.222 when $\mathrm{w}_1 = -0.6$ with a partially identified interval of $[-0.352, -0.072]$. The QMLE estimate of the ATE when $\mathrm{w}_1 = -0.6$ is -0.218 with a partially identified interval of $[-0.361, -0.075]$. When $\mathrm{w}_1 = 1$ the ATE is -0.107 with a partially identified interval of $[-0.389, -0.030]$ and the QMLE estimate of the ATE is -0.083 with a partially identified interval of $[-0.327, -0.023]$.

5 Conclusion

The RBVP model is commonly employed by applied researchers in situations where the outcome of interest is a dichotomous indicator and the determinants of the probable outcome includes qualitative information in the form of an endogenous dummy or treatment variable. The identification of the RBVP model relies heavily on the parametric specification and distributional assumptions, however, so called "identification by functional form" and the literature in this area contains conflicting statements regarding "identification by functional form", particularly in empirical studies. In this paper we have clarified the notion of "identification by functional form" and presented Monte-Carlo results that highlight the fact that when a practitioner presumes that a particular model generates the data, the availability of suitable IVs is not an issue for the statistical identification of the model parameters, but is a matter of concern for the finite sample performance of the estimates. In general, RMSE performance can be significantly improved (*ceteris paribus*) by the availability of IVs, particularly for the QMLE based upon an incorrectly specified model. As observed above, the assumptions underlying the RBVP model are unlikely to be true of real data and the latter observation may go some way in explaining the perceived improvement brought about in practice by the use of suitable instruments.

Finally, the RBVP model is frequently used by empirical researchers in policy evaluation because this model allows for the estimation of the ATE. An important message from the results presented here is that if a RBVP model is used to estimate the ATE, then the resulting QMLE can produce reasonably accurate estimates even in the presence of gross distributional misspecification. When we analyze the identification of the ATE within the partial identification

framework, we find that the QMLE generates pseudo-true parameter values that yield estimates of the ATE and the ATE partially identified set that are close to those generated by the true DGP.

In summary, our results suggest that a response to Box's aphorism is that not only is the RBVP model a readily implementable tool for estimating the effect of an endogenous binary regressor on a binary outcome variable, but it is also a useful tool whose results can be readily interpreted from a partial identification perspective.

References

Azzalini A, Capitanio A. 1999. Statistical applications of the multivariate skew normal distribution. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **61**: 579–602.

Azzalini A, Dalla Valle A. 1996. The multivariate skew-normal distribution. *Biometrika* **83**: 715–726.

Bryson A, Cappellari L, Lucifora C. 2004. Does union membership really reduce job satisfaction? *British Journal of Industrial Relations* **42**: 439–459.

Carrasco R. 2001. Binary choice with binary endogenous regressors in panel data: Estimating the effect of fertility on female labor participation. *Journal of Business & Economic Statistics* **19**: 385–394.

Chesher A. 2005. Nonparametric identification under discrete variation. *Econometrica* **73**: 1525–1550.

Chesher A. 2007. Endogeneity and discrete outcomes. *CeMMAP Working Papers CWP 05/07* .

Chesher A. 2010. Instrumental variable models for discrete outcomes. *Econometrica* **78**: 575–601.

Deadman D, MacDonald Z. 2003. Offenders as victims of crime?: an investigation into the relationship between criminal behaviour and victimization. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **167**: 53–67.

Greene WH. 2012. *Econometric Analysis*. Prentice Hall, Upper Saddle River, NJ, 7th edition.

Heckman JJ. 1978. Dummy endogenous variables in a simultaneous equation system. *Econometrica* **46**: 931–959.

Heyde CC. 1997. *Quasi-likelihood and Its Application: A General Approach to Optimal Parameter Estimation*. Springer-Verlag: New York.

Johnson NL, Kotz S, Balakrishnan N. 1995. *Continuous Univariate Distributions*, volume 2. Wiley Series in Probability and Statistics.

Jones A. 2007. Identification of treatment effects in Health Economics. *Health Economics* **16**: 1127–1131.

Jones AM, O'Donnell O (eds.) . 2002. *Econometric Analysis of Health Data*. Wiley: Chichester.

Maddala GS. 1983. *Limited-dependent and Qualitative Variables in Econometrics*. Cambridge University Press.

Manski CF. 1988. Identification of binary response models. *Journal of the American Statistical Association* **83**: 729–738.

Manski CF. 1990. Nonparametric bounds on treatment effects. *The American Economic Review* **80**: 319–323.

Manski CF. 1997. Monotone treatment response. *Econometrica* **65**: 1311–1334.

Manski CF, Pepper JV. 2000. Monotone instrumental variables: with an application to the returns to schooling. *Econometrica* **68**: 997–1010.

Morris S. 2007. The impact of obesity on employment. *Labour Economics* **14**: 413–433.

White H. 1982. Maximum likelihood estimation of misspecified models. *Econometrica* **50**: 1–25.

Wilde J. 2000. Identification of multiple equation probit models with endogenous dummy regressors. *Economics Letters* **69**: 309–312.