

Testing the hypothesis of absence of unobserved confounding in semiparametric bivariate probit models

Giampiero Marra · Rosalba Radice ·
Silvia Missiroli

Received: 11 March 2013 / Accepted: 1 October 2013 / Published online: 15 October 2013
© Springer-Verlag Berlin Heidelberg 2013

Abstract Lagrange multiplier and Wald tests for the hypothesis of absence of unobserved confounding are extended to the context of semiparametric recursive and sample selection bivariate probit models. The finite sample size properties of the tests are examined through a Monte Carlo study using several scenarios: correct model specification, distributional and functional misspecification, with and without an exclusion restriction. The simulation results provide some guidelines which may be important for empirical analysis. The tests are illustrated using two datasets in which the issue of unobserved confounding arises.

Keywords Endogeneity · Lagrange multiplier test · Non-random sample selection · Penalized regression spline · Wald test

1 Introduction

We are concerned with testing the hypothesis of absence of unobserved confounding in the recursive and sample selection bivariate probit models (Heckman 1978, 1979;

G. Marra (✉)

Department of Statistical Science, University College London, Gower Street, London WC1E 6BT, UK
e-mail: giampiero.marra@ucl.ac.uk

R. Radice

Department of Economics, Mathematics and Statistics, Birkbeck, University of London, Malet Street, London WC1E 7HX, UK

S. Missiroli

Dipartimento di Scienze Statistiche, Università di Bologna, via Belle Arti 41, 40126 Bologna, Italy

S. Missiroli

Department of Decision Science, Bocconi University, Via Roentgen 1, 20136 Milano, Italy

Maddala 1983; de Ven and Praag 1981; Greene 2012). Recursive bivariate probit models deal with a problem which arises in observational studies when confounders (i.e., variables that are associated with treatment and response) are unobserved; this issue is known in the econometric literature as *endogeneity*. These models control for unobserved confounders by using a two-equation structural latent variable framework, where one equation models a binary response as a function of a binary treatment and some covariates whereas the other determines whether the treatment is received based on some regressors. Recent economic and biostatistical applications of such models include the study of the effect of private health insurance on medical care utilization (Buchmueller et al. 2005), impact of diabetes on employment (Latif 2009), effect of physical activity on obesity (Kawatkar and Nichol 2009), and impact of insurance on mortality among HIV-infected individuals (Goldman et al. 2001).

Sample selection models address an issue which arises when observations are not from a random sample of the population. Instead, individuals may have selected themselves (or have been selected by others) into (or out of) the sample based on a combination of observed and unobserved characteristics. If the sample of selected individuals differs in important observed characteristics from the sample of unselected individuals, then selection bias can be avoided by controlling for these features. However, if the two samples differ in important unobserved characteristics then *non-random sample selection* arises. Sample selection bivariate probit models control for unobserved confounders by simultaneously estimating two regressions: a selection equation and a response equation. The former determines whether an individual is selected into the sample whereas the latter is used to examine the substantive question of interest. Applications of these models include the study of prevention program for high school dropouts (Montmarquette et al. 2001), quantification of the effect of family-related factors on foster approval (Cuddeback et al. 2004), estimation of HIV prevalence (Bärnighausen et al. 2011) and reject-inference to predict credit quality (Banasik and Crook 2007).

The classic recursive and sample selection bivariate probit models used for the applications mentioned above do not allow for flexible functional dependence of the responses on continuous covariates. To this end, Marra and Radice (2011, 2013a) introduced a penalized likelihood estimation framework to estimate simultaneously the parameters of a system of two binary equations that include smooth functions of continuous covariates. Alternative (Bayesian) approaches are available in the literature (e.g., Chib and Greenberg 2007; Chib et al. 2009). However the lack of user-friendly software makes them unfeasible for practitioners. The need for methods flexibly modeling covariate effects arises from the observation that all parameter estimates are inconsistent when the relationship between observed confounders and outcome is misspecified (Marra and Radice 2011; Chib et al. 2009).

An important aspect of these binary bivariate models is that if the hypothesis of absence of unobserved confounding (i.e., exogeneity and random selection) can not be rejected then joint estimation of the model equations can be avoided. This is appealing because inference in simultaneous models with smooth components may become computationally demanding as the sample size and number of smooth terms increase. Therefore, before employing a complex simultaneous estimation approach, testing the hypothesis of absence of unobserved confounding is an important step that should be

undertaken at the beginning of the empirical analysis. To this end, we extend Lagrange multiplier (LM) and, for comparison, Wald (\mathcal{W}) tests to the context of semiparametric bivariate probit models. LM is particularly attractive as it is based on estimating the model equations separately. The finite sample size performance of the tests is investigated through a Monte Carlo simulation study that considers the scenarios of correct model specification, distributional and functional misspecification, with and without an exclusion restriction (ER). The simulation results allow us to infer some guidelines which may be crucial for applications. For instance, the good performance of LM should be particularly attractive to practitioners wishing to test the null hypothesis of absence of unobserved confounding while avoiding simultaneous estimation. The tests are implemented in the R package `SemiParBIVProbit` (Marra and Radice 2013b).

The article is organized as follows. Section 2 provides a brief overview of the models of interest and their estimation; this is useful to define the notation and make some remarks that are relevant to the implementation of the tests. Section 3 discusses the construction of the LM and \mathcal{W} type tests in the context of semi-parametric models, whereas Sect. 4 assesses their finite sample size performance through a Monte Carlo simulation study. Section 5 illustrates the tests using two datasets, and Sect. 6 provides a discussion. ‘‘Appendix 1’’ shows that, under the null hypothesis, the expected information matrix for the recursive bivariate probit model case is block diagonal, whereas ‘‘Appendix B’’ reports further simulation results.

2 Preliminaries

2.1 The models

The semiparametric recursive and sample selection bivariate probit models introduced by Marra and Radice (2011, 2013a) are a generalization of the parametric model versions introduced by Heckman (1978, 1979) in that continuous covariate effects are modeled flexibly.

The model structure consists of two equations. The first can be written as

$$y_{1i}^* = \mathbf{m}_{1i}^\top \boldsymbol{\theta}_1 + \sum_{k_1=1}^{K_1} f_{1k_1}(z_{1k_1i}) + \varepsilon_{1i}, \quad i = 1, \dots, n, \tag{1}$$

where n is the sample size, y_{1i}^* is a latent continuous variable and y_{1i} is determined via the rule $1(y_{1i}^* > 0)$. The second can be defined as

$$y_{2i}^* = \vartheta y_{1i} + \mathbf{m}_{2i}^\top \boldsymbol{\theta}_2 + \sum_{k_2=1}^{K_2} f_{2k_2}(z_{2k_2i}) + \varepsilon_{2i}, \tag{2}$$

where, in the non-random sample selection case, ϑ is set to zero and y_{2i} is determined as

$$y_{2i} = \begin{cases} 1 & \text{if } (y_{2i}^* > 0 \ \& \ y_{1i} = 1) \\ 0 & \text{if } (y_{2i}^* < 0 \ \& \ y_{1i} = 1) \\ - & \text{if } y_{1i} = 0 \end{cases},$$

whereas, in the endogeneity case, ϑ is allowed to be different from zero and y_{2i} is determined via the rule $1(y_{2i}^* > 0)$. Vector \mathbf{m}_{1i} contains P_1 parametric model components (such as the intercept, dummy and categorical variables), with corresponding parameter vector $\boldsymbol{\theta}_1$, and the f_{1k_1} are unknown smooth functions of the K_1 continuous covariates z_{1k_1i} . Each smooth term may also be multiplied by some predictor(s) (Hastie and Tibshirani 1993). Furthermore, smooth functions of two covariates such as $f_{11,12}(z_{11i}, z_{12i})$ can be considered (e.g., Wood 2006). Similarly, \mathbf{m}_{2i} is a vector containing P_2 parametric components, with coefficient vector $\boldsymbol{\theta}_2$, and the other terms have the obvious definitions. Smooth functions are subject to constraints, i.e. $\sum_i f_{vk_v}(z_{vk_vi}) = 0, v = 1, 2$, for all smooth components in the model. The error terms are assumed to follow the distribution $\mathcal{N}([0, 0], [1, \rho, \rho, 1])$, where ρ is the correlation coefficient and the error variances are normalized to unity (e.g., Greene 2012, p. 686). To identify the parameters of Eq. (2) it is typically assumed that the ER on the exogenous variables holds. That is, the covariates in the first equation should contain at least one or more regressors (typically referred to as instruments) not included in the second equation. These regressors have to induce variation in y_{1i} , not to directly affect y_{2i} , and be independent of $(\varepsilon_{1i}, \varepsilon_{2i})$ given covariates. However, under correct model specification, this restriction may not be necessary in estimation (Marra and Radice 2011; Wilde 2000).

The smooth functions are represented using regression splines (e.g., Eilers and Marx 1996). The basic idea is to approximate a generic function $f_k(z_{ki})$ by a linear combination of known spline basis functions, $b_{kj}(z_{ki})$, and regression parameters, β_{kj} , i.e. $\sum_{j=1}^{J_k} \beta_{kj} b_{kj}(z_{ki}) = \mathbf{B}_k(z_{ki})^\top \boldsymbol{\beta}_k$, where J_k is the number of spline bases, $\mathbf{B}_k(z_{ki}) = \{b_{k1}(z_{ki}), \dots, b_{kJ_k}(z_{ki})\}^\top$ is a vector of the basis functions evaluated at z_{ki} and $\boldsymbol{\beta}_k$ is the corresponding parameter vector. Note that subscript v has been suppressed to avoid clutter. Basis functions have to be chosen to have convenient mathematical and numerical properties. Possible choices include B-splines, cubic regression and thin plate regression splines (see, e.g., Marra and Radice (2010) for an overview). Based on the result above, Eqs. (1) and (2) can be written as $y_{1i}^* = \mathbf{m}_{1i}^\top \boldsymbol{\theta}_1 + \mathbf{B}_{1i}^\top \boldsymbol{\beta}_1 + \varepsilon_{1i} = \eta_{1i} + \varepsilon_{1i}$, and $y_{2i}^* = \vartheta y_{1i} + \mathbf{m}_{2i}^\top \boldsymbol{\theta}_2 + \mathbf{B}_{2i}^\top \boldsymbol{\beta}_2 + \varepsilon_{2i} = \eta_{2i} + \varepsilon_{2i}$, where $\mathbf{B}_{vi}^\top = \{\mathbf{B}_{v1}(z_{v1i})^\top, \dots, \mathbf{B}_{vK_v}(z_{vK_vi})^\top\}$, $\boldsymbol{\beta}_v^\top = (\beta_{v1}^\top, \dots, \beta_{vK_v}^\top)$ and η_{1i} and η_{2i} have the obvious definition.

2.2 Estimation

In the sample selection model the data identify only the three possible events $(y_{1i} = 1, y_{2i} = 1), (y_{1i} = 1, y_{2i} = 0)$ and $(y_{1i} = 0)$ with probabilities $p_{11i} = \Phi_2(\eta_{1i}, \eta_{2i}; \rho), p_{10i} = \Phi(\eta_{1i}) - p_{11i}$ and $p_{0i} = \Phi(-\eta_{1i})$, where Φ and Φ_2 are the distribution functions of a standardized univariate normal and a standardized bivariate normal with correlation ρ , respectively. Therefore, the log-likelihood function is

$$\ell(\delta) = \sum_{i=1}^n \{y_{1i}y_{2i} \log(p_{11i}) + y_{1i}(1 - y_{2i}) \log(p_{10i}) + (1 - y_{1i}) \log(p_{0i})\}, \quad (3)$$

where $\delta^T = (\delta_1^T, \delta_2^T, \rho)$ and $\delta_v^T = (\theta_v^T, \beta_v^T)$, for $v = 1, 2$. In the recursive model ($y_{1i} = 0$) is replaced by ($y_{1i} = 0, y_{2i} = 1$) and ($y_{1i} = 0, y_{2i} = 0$) which have probabilities $p_{01i} = \Phi(\eta_{2i}) - p_{11i}$ and $p_{00i} = 1 - p_{11i} - p_{10i} - p_{01i}$. Hence, in (3), $(1 - y_{1i}) \log(p_{0i})$ is replaced by $(1 - y_{1i})y_{2i} \log(p_{01i}) + (1 - y_{1i})(1 - y_{2i}) \log(p_{00i})$. To avoid overfitting, the model parameters are estimated by maximization of

$$\ell_p(\delta) = \ell(\delta) - \frac{1}{2} \beta^T \mathbf{S}_\lambda \beta, \quad (4)$$

where $\beta^T = (\beta_1^T, \beta_2^T)$, $\mathbf{S}_\lambda = \sum_{v=1}^2 \sum_{k_v=1}^{K_v} \lambda_{vk_v} \mathbf{S}_{vk_v}$ and the \mathbf{S}_{vk_v} are positive semi-definite known square matrices measuring the (second-order, typically) roughness of the smooth terms in the model, i.e. $\beta^T \left(\sum_{v=1}^2 \sum_{k_v=1}^{K_v} \lambda_{vk_v} \mathbf{S}_{vk_v} \right) \beta = \sum_{v=1}^2 \sum_{k_v=1}^{K_v} \lambda_{vk_v} \int f''_{vk_v}(z_{vk_v})^2 dz_{vk_v}$. The λ_{vk_v} are smoothing parameters controlling the trade-off between fit and smoothness. Because ρ is bounded in $[-1, 1]$, $\rho^* = \tanh^{-1}(\rho) = (1/2) \log \{(1 + \rho)/(1 - \rho)\}$ is used in optimization.

Estimation of δ and $\lambda = (\lambda_{1k_1}, \dots, \lambda_{1K_1}, \lambda_{2k_2}, \dots, \lambda_{2K_2})$ is carried out in two steps. In particular, given a parameter vector value for λ , (4) is maximized using a trust region algorithm (Nocedal and Wright 2006) which is based on

$$\delta^{[a+1]} = \delta^{[a]} + (\mathcal{I}^{[a]} + \tilde{\mathbf{S}}_\lambda)^{-1} (\mathbf{g}^{[a]} - \tilde{\mathbf{S}}_\lambda \delta^{[a]}), \quad (5)$$

where a is the iteration index and $\tilde{\mathbf{S}}_\lambda$ is defined as $\tilde{\mathbf{S}}_\lambda = \text{diag}(0_{11}, \dots, 0_{1P_1}, \lambda_{1k_1} \mathbf{S}_{1k_1}, \dots, \lambda_{1K_1} \mathbf{S}_{1K_1}, 0_{21}, \dots, 0_{2P_2}, \lambda_{2k_2} \mathbf{S}_{2k_2}, \dots, \lambda_{2K_2} \mathbf{S}_{2K_2}, 0)$. The gradient vector \mathbf{g} is given by $\mathbf{g}_1 = \partial \ell(\delta) / \partial \delta_1$, $\mathbf{g}_2 = \partial \ell(\delta) / \partial \delta_2$ and $\mathbf{g}_3^* = \partial \ell(\delta) / \partial \rho^*$, while the expected information matrix has a 3×3 matrix block structure with (r, h) th element $\mathcal{I}_{r,h} = -\mathbb{E} \left[\partial^2 \ell(\delta) / \partial \delta_r \partial \delta_h^T \right]$, $r, h = 1, \dots, 3$, where $\delta_3 = \rho^*$. Given an estimate for δ , multiple smoothing parameter estimation for (5) is then achieved by minimization of

$$\mathcal{V}_u^w(\lambda) = \frac{1}{n_*} \|\sqrt{\mathbf{W}}(\mathbf{z} - \mathbf{X}\delta)\|^2 - 1 + \frac{2}{n_*} \text{tr}(\mathbf{A}_\lambda) \quad \text{w.r.t. } \lambda, \quad (6)$$

where $\mathbf{z}_i = \mathbf{X}_i \delta + \mathbf{W}_i^{-1} \mathbf{d}_i$, $\mathbf{X}_i = \text{diag} \left\{ \mathbf{m}_{1i}^T, \mathbf{B}_{1i}^T, \mathbf{m}_{2i}^T, \mathbf{B}_{2i}^T, 1 \right\}$, $\mathbf{d}_i = \{\partial \ell(\delta)_i / \partial \eta_{1i}, \partial \ell(\delta)_i / \partial \eta_{2i}, \partial \ell(\delta)_i / \partial \eta_{3i}\}^T$, $\eta_{3i} = \rho^*$, \mathbf{W}_i is the 3×3 matrix with (r, h) th element

$$(\mathbf{W}_i)_{rh} = -\mathbb{E} \left[\frac{\partial^2 \ell(\delta)_i}{\partial \eta_{ri} \partial \eta_{hi}} \right], \quad r, h = 1, \dots, 3, \quad (7)$$

$n_* = 3n$, $\mathbf{A}_\lambda = \mathbf{X}(\mathbf{X}^T \mathbf{W} \mathbf{X} + \mathbf{S}_\lambda^*)^{-1} \mathbf{X}^T \mathbf{W}$, and $\text{tr}(\mathbf{A}_\lambda)$ represents the estimated degrees of freedom (edf) of the penalized model. The square root and inverse of \mathbf{W} are obtained via eigen-decomposition. (To avoid clutter the superscript $[a]$ has been suppressed

from the quantities above.) The two steps, one for δ and the other for λ , are iterated until convergence. Full details can be found in [Marra and Radice \(2011, 2013a\)](#).

For semiparametric bivariate probit models, the expected information matrix is the only option because the \mathbf{W}_i , as defined in (7), are positive-definite over a larger region of the parameter space as compared to those obtained without taking expectations. This is crucial given that $\sqrt{\mathbf{W}}$ and \mathbf{W}^{-1} (via \mathbf{z}) are needed in (6).

3 Testing the hypothesis of absence of unobserved confounding

The hypothesis of absence of unobserved confounding can be stated in terms of ρ , which can be interpreted as the correlation between the unobserved variables in the two equations. If $\rho = 0$ then ε_{1i} and ε_{2i} are uncorrelated and hence there is not a problem of unobserved confounding. On the contrary, $\rho \neq 0$ implies that there is a problem of unobserved confounding. This leads us to the definition of the following hypothesis

$$\begin{cases} H_0 : \rho = 0 \\ H_1 : \rho \neq 0 \end{cases}$$

Under H_0 function (4) becomes the sum of the penalized log-likelihood functions of two semiparametric univariate probit models. This implies that $\hat{\delta}_{H_0}^\top = (\hat{\delta}_1^\top, \hat{\delta}_2^\top, 0)$, where $\hat{\delta}_1$ and $\hat{\delta}_2$ are obtained by estimating the model equations separately. Therefore, consistent estimates for δ_2 can be obtained by fitting Eq. (2) alone. Under H_1 simultaneous estimation is required to obtain consistent parameter estimates.

3.1 LM type tests

In the context of the models described in this article, *LM* (also known as score test) is an appropriate and computationally convenient tool for testing H_0 . Its main advantage is that it does not require parameter estimates under H_1 . This is appealing because the test is based on estimating the two model equations separately, hence obviating the need to fit the more computationally demanding semiparametric bivariate model. This implies that simultaneous estimation will be employed only if there is a problem of unobserved confounding.

The *LM* statistic for the semiparametric bivariate models is

$$LM = \left\{ \mathbf{g}_{\hat{\delta}_{H_0}} - \tilde{\mathbf{S}}_{\lambda_{H_0}} \hat{\delta}_{H_0} \right\}^\top \mathbf{I}^{-1} \left\{ \mathbf{g}_{\hat{\delta}_{H_0}} - \tilde{\mathbf{S}}_{\lambda_{H_0}} \hat{\delta}_{H_0} \right\} \xrightarrow{H_0} \chi_1^2, \quad (8)$$

where $\mathbf{g}_{\hat{\delta}_{H_0}}$ is the score vector evaluated at $\hat{\delta}_{H_0}$, $\tilde{\mathbf{S}}_{\lambda_{H_0}}$ is defined in Sect. 2.2 but with smoothing parameter estimates obtained by estimating the two univariate probit equations separately, and \mathbf{I}^{-1} is the inverse of the information matrix. Studying the limiting behavior of test statistics that involve penalty terms is not an easy task (e.g., [Wood 2006](#), Sect. 4.8). However, because ρ is not penalized, it is still possible to use the

classic result that LM has a χ_1^2 limiting distribution (Monfardini and Radice 2008). This can be seen by observing that $\{\mathbf{g}_{\hat{\delta}_{H_0}} - \tilde{\mathbf{S}}_{\lambda_{H_0}} \hat{\delta}_{H_0}\}^\top = \{\mathbf{0}, \mathbf{0}, \partial \ell(\hat{\delta}_{H_0})/\partial \rho\}$ which results from evaluating the penalized score in $\hat{\delta}_{H_0}$.

Different estimation methods for \mathbf{I} may lead to different finite sample size performances. In this paper, we consider:

- The observed information matrix $-\mathcal{H}_{\hat{\delta}_{H_0}} + \tilde{\mathbf{S}}_{\lambda_{H_0}}$ (this expression derives from the penalized score in (8)). The test obtained using this matrix will be referred to as $LM_{\mathcal{H}}$ for notational convenience.
- The expected information matrix $\mathcal{I}_{\hat{\delta}_{H_0}} + \tilde{\mathbf{S}}_{\lambda_{H_0}}$. This test will be referred to as $LM_{\mathcal{I}}$.

The analytical expressions of \mathbf{g} , \mathcal{I} and $-\mathcal{H}$ for the recursive and sample selection bivariate probit models are not reported here to save space and are implemented in function `LM.bpm()` of the R package `SemiParBIVProbit` (Marra and Radice 2013b).

For the recursive model, the expected information matrix becomes block diagonal under H_0 (see ‘‘Appendix 1’’). Hence, $LM_{\mathcal{I}}$ can be simplified to

$$LM_{\mathcal{I}} = - \left\{ \frac{\partial \ell(\hat{\delta}_{H_0})}{\partial \rho} \right\}^2 \mathbb{E} \left[\frac{\partial^2 \ell(\hat{\delta}_{H_0})}{\partial \rho \partial \rho} \right]^{-1}, \tag{9}$$

which is computationally convenient since it does not require the inversion of the information matrix. This result does not hold for the non-random sample selection case given the different structure of the information matrix.

Note that, to implement the LM type tests discussed above, we evaluate the score vector and information matrices at $\hat{\delta}_{H_0}$ (obtained from the univariate fits). Therefore, the arc-tangent transform for ρ (used in the context of simultaneous equation estimation) is not required.

3.2 \mathcal{W} test

The Wald test requires fitting the semiparametric bivariate model, hence it is not as advantageous as the LM test. However, such a test is widely used in the applied literature and therefore we consider it for comparison.

\mathcal{W} is based on estimating ρ and is given by

$$\mathcal{W} = \frac{\hat{\rho}^2}{\text{Var}(\hat{\rho})} \xrightarrow{H_0} \chi_1^2. \tag{10}$$

$\text{Var}(\hat{\rho})$ is estimated using the diagonal element of the inverse of $\mathcal{I}_{\hat{\delta}} + \tilde{\mathbf{S}}_{\lambda}$ corresponding to $\hat{\rho}$. This test will be referred to as $\mathcal{W}_{\mathcal{I}}$. In a semiparametric context this is the only option available. This is because model fitting can only be based on the expected information matrix, as explained in the final paragraph of Sect. 2.2. For the recursive model, $\text{Var}(\hat{\rho})$ can be estimated using the simplification described in the previous section, i.e. $\text{Var}(\hat{\rho}) = -1/\mathbb{E}[\partial^2 \ell(\hat{\delta})/\partial \rho \partial \rho]$. Once

again, this result does not hold for the non-random sample selection case. Because, in estimation, we make use of ρ^* rather than ρ (see Sect. 2.2), using the delta method $\text{Var}(\hat{\rho}) = 4\text{Var}(\hat{\rho}^*) \exp(2\hat{\rho}^*) / (\exp\{2\hat{\rho}^*\} + 1)^2$.

The expected information matrix employed for computing $\text{Var}(\hat{\rho})$ corresponds to the Bayesian covariance matrix typically used for constructing ‘confidence’ intervals for the terms of a penalized regression spline model (e.g., Wood 2006). Such intervals have good *frequentist* coverage probabilities. This is because they include both a bias and variance component, a feature which is not shared by their frequentist counterpart; see Marra and Wood (2012) for full details. It may be argued that since ρ is not penalized the frequentist covariance matrix, given by $\mathbf{I}^{-1}(\mathbf{I} - \hat{\mathbf{S}}_\lambda)\mathbf{I}^{-1}$, is also a sensible choice. Preliminary simulation evidence confirmed that $\text{Var}(\hat{\rho})$ is estimated well by both covariance matrices. However, the Bayesian version is computationally convenient since it can be obtained as a byproduct of the estimation procedure described in Sect. 2.2.

Another commonly used test which requires fitting the bivariate model is the likelihood ratio (LR) test. Here, the statistic is given by twice the difference of the model log-likelihoods under H_1 and H_0 , and has a χ_1^2 limiting distribution for parametric models (Monfardini and Radice 2008). In the current context, we are however faced with a difficulty which inhibits the use of this approach for testing H_0 . Specifically, in the semiparametric case the number of degrees of freedom for LR is not guaranteed to be equal to 1, and can in fact be a positive or negative real value. For instance, if simultaneous estimation of the two equations leads to a model with edf (defined in Sect. 2.2) equal to 18.37 and estimating the equations separately yields a model with edf equal to 20.23 then the number of degrees of freedom for LR is -1.86 ; the amount of smoothness required for the smooth functions of a model fitted via simultaneous estimation is likely to be different from that required when the equations are estimated separately. The same issue has been found when comparing generalized additive models (e.g., (Wood 2006, Sect. 4.10)). It is not clear whether LR can be rigorously extended to this context and further research is required.

4 Simulations

To compare the finite sample size properties of the LM and \mathcal{W} tests discussed in the previous section, we conducted simulation studies under correct and incorrect specification of both semiparametric bivariate probit models, with and without ER. The size and power of each test were calculated as the proportions of rejections based on simulation replications. All computations were performed in the R environment (R Development Core Team 2013) using the package `SemiParBIVProbit` (Marra and Radice 2013b).

The next sections provide details on the simulation and model fitting settings. The most salient features of the simulation results are then discussed. We present the endogeneity and non-random sample selection cases separately, starting with the former. Note that the study design detailed in Sects. 4.1.1 and 4.2.1 extends the simulation study of Monfardini and Radice (2008) to recognize the specific challenges that arise when there are nonlinear response-covariate relationships.

Table 1 Parameter value sets for the DGPs in the endogeneity and non-random sample selection cases

	Parametric components					Smooth components						
	θ_{10}	θ_{11}	θ_{20}	θ_{21}	ϑ	γ_1	γ_2	γ_3	γ_4	γ_5	γ_6	γ_7
Setting 1	0.32	1.25	0.25	1.00	0.75	0.90	3.00	0.35	0.75	2.00	0.25	1.00
Setting 2	0.01	-1.31	-0.29	-0.37	0.49	0.24	2.61	0.60	0.28	-2.62	0.43	-0.63
Setting 3	-0.55	0.17	0.10	0.93	0.03	-0.65	3.10	-0.67	-0.07	0.41	0.29	-0.18

4.1 Endogeneity

4.1.1 Design of the experiments

Our experiments were based on two different data generating processes (DGPs): DGP1 and DGP2. In DGP1, the specification of the semiparametric recursive bivariate probit model contained main effects only, whereas in DGP2 an interaction term between the endogenous binary variable and a continuous regressor was also considered. That is,

$$\begin{aligned} y_{1i}^* &= \theta_{10} + \theta_{11}m_{1i} + f_1(z_{1i}) + f_3(z_{2i}) + \varepsilon_{1i} \\ y_{2i}^* &= \theta_{20} + \vartheta y_{1i} + \theta_{21}m_{1i} + f_2(z_{1i}) + \varepsilon_{2i} \end{aligned} \tag{11}$$

and

$$\begin{aligned} y_{1i}^* &= \theta_{10} + \theta_{11}m_{1i} + f_1(z_{1i}) + f_3(z_{2i}) + \varepsilon_{1i} \\ y_{2i}^* &= \theta_{20} + \theta_{21}m_{1i} + (1 - y_{1i})f_2(z_{1i}) + y_{1i}f_4(z_{1i}) + \varepsilon_{2i} \end{aligned} \tag{12}$$

where the binary outcomes y_{1i} and y_{2i} were determined according to the rules described in Sect. 2.1. The smooth functions were $f_1(z_{1i}) = \gamma_1[z_{1i}^{2.5} + \exp\{\gamma_2(z_{1i} - \gamma_3)^2\}]$, $f_2(z_{1i}) = \gamma_4\{\gamma_5 \exp(z_{1i}) - z_{1i}^3\}$, $f_3(z_{2i}) = \gamma_6 z_{2i}$ and $f_4(z_{1i}) = \gamma_7 \sin(\pi z_{1i})$. For both DGPs, we used three different parameter vector settings (see Table 1). These values were obtained randomly from a standardized normal distribution.

We also considered a variant of DGP2, for the three settings, useful to test what happens in the situation of distributional misspecification. Specifically, we generated data assuming uncorrelated gamma errors with shape and scale parameters equal to 2. This was achieved using `rgamma()`. The sample sizes considered were 1,000 and 4,000. Each design was replicated 1,000 times.

4.1.2 Fitting details

Let the two model equations be equal to

```
eq1 <- y1 ~ m1 + s(z1) + s(z2)
eq2 <- y2 ~ y1 + m1 + s(z1)
```

for DGP1 and

```
eq1 <- y1 ~ m1 + s(z1) + s(z2)
eq2 <- y2 ~ y1 + m1 + s(z1, by = y1)
```

for DGP2. $s(\cdot)$ indicates that a smooth component is being used, and the `by` option is to allow the smooth to interact with a parametric term. Variables y_1 , y_2 , m_1 , z_1 and z_2 refer to y_{1i} , y_{2i} , m_{1i} , z_{1i} and z_{2i} . Using `rmvnorm()` in the package `mvtnorm`, regressors m_{1i} , z_{1i} and z_{2i} were obtained from a matrix of dimensions $n \times 3$ whose columns, called say `reg1`, `reg2` and `reg3`, were generated from a multivariate normal distribution with zero means and covariance matrix characterized by correlations equal to 0.5 and variances equal to 1. `reg1` and `reg2` were then transformed using `round()` and `pnorm()`, respectively, to generate binary and uniform covariates. Note that the specifications for `eq1` and `eq2` are consistent with Eqs. (11) and (12).

p values for $LM_{\mathcal{H}}$ and $LM_{\mathcal{I}}$ were calculated using `LM.bpm()` from `SemiParBIVProbit`. That is,

```
LM.bpm(eq1, eq2, data, FI = FALSE)
LM.bpm(eq1, eq2, data, FI = TRUE)
```

corresponding to $LM_{\mathcal{H}}$ and $LM_{\mathcal{I}}$, respectively. `data` is a data frame containing the variables in the two equations generated according to the DGPs described in the previous section. p values for $\mathcal{W}_{\mathcal{I}}$ were calculated using $\hat{\rho}$ and $\text{Var}(\hat{\rho})$ obtained from the output produced by `SemiParBIVProbit()`. That is,

```
outE <- SemiParBIVProbit(eq1, eq2, data, ...)
V <- outE$Vb
rho.s <- outE$fit$argument[dim(V)[2]]
rho <- tanh(rho.s)
var.rho.s <- V[dim(V)[2], dim(V)[2]]
var.rho <- 4*var.rho.s*exp(2*rho.s) / (exp(2*rho.s)+1)^2
pchisq(rho^2/var.rho, 1, lower.tail = FALSE)
```

The smooth components of the semiparametric models were represented using penalized thin plate regression splines with basis dimensions equal to 10 and penalties based on second-order derivatives (Wood 2006, pp. 154–160).

Regarding model misspecification, for DGP2, we generated data using gamma errors and employed the same tests. We also considered a scenario with functional form misspecification where the model equation for y_2 did not include the interaction term.

4.1.3 Monte Carlo results

Empirical size Table 2 provides rejection frequencies produced under $H_0 : \rho = 0$ and correct model specification using the three test statistics considered in this paper. Results are reported for the three typical critical values. The findings for the two DGPs share some important features, for the three settings. $LM_{\mathcal{H}}$ exhibits empirical sizes that are overall close to the nominal values. $LM_{\mathcal{I}}$ and $\mathcal{W}_{\mathcal{I}}$ yield unsatisfactory size results; the former under-rejects whereas the latter over-rejects. The good performance of $LM_{\mathcal{H}}$ is important for practitioners wishing to test the hypothesis of absence of unobserved confounding without coping with simultaneous estimation. In the current context, $-\mathcal{H}$ is a more adequate measure than \mathcal{I} for estimating the information matrix.

Table 2 Size results (in %) for the endogeneity case

	α (%)	n	Setting 1			Setting 2			Setting 3		
			$LM\mathcal{H}$	$LM\mathcal{I}$	$\mathcal{W}\mathcal{I}$	$LM\mathcal{H}$	$LM\mathcal{I}$	$\mathcal{W}\mathcal{I}$	$LM\mathcal{H}$	$LM\mathcal{I}$	$\mathcal{W}\mathcal{I}$
DGP1	1	1,000	1.7	0.0	29.6	1.7	0.0	16.2	1.5	0.00	22.5
	5		6.8	0.0	42.7	6.9	0.1	27.7	6.0	0.00	35.1
	10		11.4	0.2	49.1	11.9	0.7	36.6	11.1	0.00	43.6
	1	4,000	0.8	0.0	27.4	1.3	0.0	11.1	0.9	0.00	20.9
	5		4.7	0.0	40.8	4.5	0.0	23.5	5.1	0.00	33.2
	10		9.4	0.0	48.5	9.6	0.6	33.3	10.4	0.00	41.2
DGP2	1	1,000	1.8	0.0	29.2	1.3	0.0	16.0	1.8	0.00	20.9
	5		6.1	0.0	42.1	6.7	0.1	27.3	4.8	0.00	35.2
	10		10.4	0.0	50.8	12.2	0.9	35.1	9.6	0.02	42.8
	1	4,000	1.7	0.0	27.6	1.0	0.0	11.9	0.8	0.00	19.6
	5		5.4	0.0	41.8	5.2	0.0	23.4	5.1	0.00	31.4
	10		9.4	0.0	49.5	9.8	0.5	31.0	9.3	0.00	40.7

These were obtained employing the Lagrange multiplier (LM) tests based on the observed (\mathcal{H}) and expected (\mathcal{I}) information matrices, and the Wald (\mathcal{W}) test based on \mathcal{I} . DGP1 and DGP2 relate to models (11) and (12). α and n denote the critical value and sample size considered

Monfardini and Radice (2008) found the same for parametric recursive bivariate probit models. This has also been confirmed in other contexts by Maldonado and Greenland (1994) and Louis (1982). In a notable article, Efron and Hinkley (1978) argued that the observed information should be used in preference to the expected information on the grounds that the former is ‘closer to the data’ whereas the latter is an a priori expectation. They showed this by using an appropriate ancillary statistic. The main practical limitation of their theoretical argument is the reliance on an ancillary statistic, which is often hard to specify in complex models such as the ones considered in this article. However, there are other examples (see Cavanaugh and Shumway (1996), Cao and Spall (2009) and references therein) where the expected information may more accurately estimate the true information. It emerges from the literature that there is not clear theoretical evidence about the superiority of $-\mathcal{H}$ on \mathcal{I} (or vice versa) and that further research is needed towards this direction.

The performance of $LM_{\mathcal{I}}$ and $\mathcal{W}_{\mathcal{I}}$ suggests that using \mathcal{I} underestimates the variability of $\hat{\rho}$, hence causing them to produce zero and very high rejection frequencies, respectively. Note that the opposite rejection frequencies of the two tests is attributed to the way $-\mathbb{E}\left[\frac{\partial^2 \ell}{\partial \rho \partial \rho}\right]$ enters the statistics, which can be easily verified by looking at (9) and (10).

Empirical power For DGP1 and DGP2 we evaluated the empirical power of $LM_{\mathcal{H}}$ for $\rho = \{0.1, 0, 2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$. Power curves are presented in Figs. 1 and 2. $LM_{\mathcal{H}}$ gives overall satisfactory results for all settings. Obviously the power improves as both n and ρ increase.

Empirical size under misspecification We applied $LM_{\mathcal{H}}$ to data generated according to model (12) with gamma errors. We also considered the situation of data generated according to model (12) with normal errors but using the test based on a misspecified functional form (i.e., the interaction term was omitted from the model). Size results are reported in Table 3 (see ER case). The results for setting 1 are not as good as those obtained under no misspecification but still reasonable. However, under settings 2 and 3 the sizes are overestimated. To gain more evidence, we considered other parameter sets; the results lead to similar conclusions (see “Appendix 2”, Table 7). Overall, this suggests that under misspecification the performance of $LM_{\mathcal{H}}$ worsens and the good results produced for setting 1 are due to sampling variability rather than to the robustness of the test.

We also assessed the effectiveness of $LM_{\mathcal{H}}$ when the ER does not hold. Results are reported in Table 3 (see non-ER case). Overall, the test performs poorly. Comparing the ER and non-ER results, under misspecification, $LM_{\mathcal{H}}$ performs better when the ER holds, although it still has high rejection frequencies.

4.2 Non-random sample selection

4.2.1 Design of the experiments

Similarly to the endogeneity case, the sampling experiments were based on the two different DGPs: DGP3 and DGP4. In DGP3 the specification of the semiparametric sample selection bivariate probit model contained main effects only, whereas DGP4

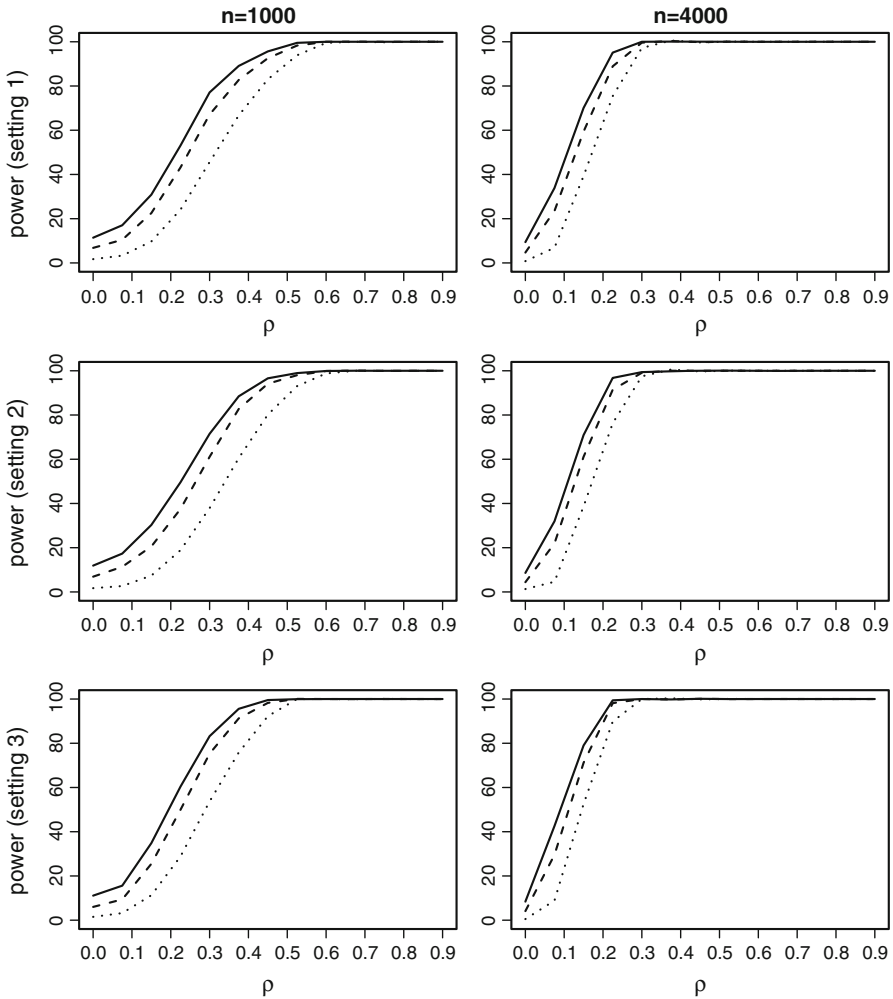


Fig. 1 Power curves for $LM_{\mathcal{H}}$ obtained under DGP1 and the three settings described in Sect. 4.1.1. The solid, dashed and dotted lines represent curves at the 10, 5 and 1% significance levels, respectively

also contained an interaction term between two regressors. That is,

$$\begin{aligned} y_{1i}^* &= \theta_{10} + \theta_{11}m_{1i} + f_1(z_{1i}) + f_3(z_{2i}) + \varepsilon_{1i} \\ y_{2i}^* &= \theta_{20} + \theta_{21}m_{1i} + f_2(z_{1i}) + \varepsilon_{2i} \end{aligned} \tag{13}$$

and

$$\begin{aligned} y_{1i}^* &= \theta_{10} + \theta_{11}m_{1i} + f_1(z_{1i}) + f_3(z_{2i}) + \varepsilon_{1i} \\ y_{2i}^* &= \theta_{20} + (1 - m_{1i})f_2(z_{1i}) + m_{1i}f_4(z_{1i}) + \varepsilon_{2i} \end{aligned} \tag{14}$$

where the binary outcomes y_{1i} and y_{2i} were determined according to the rules described in Sect. 2.1. For both DGPs, we considered the three different param-

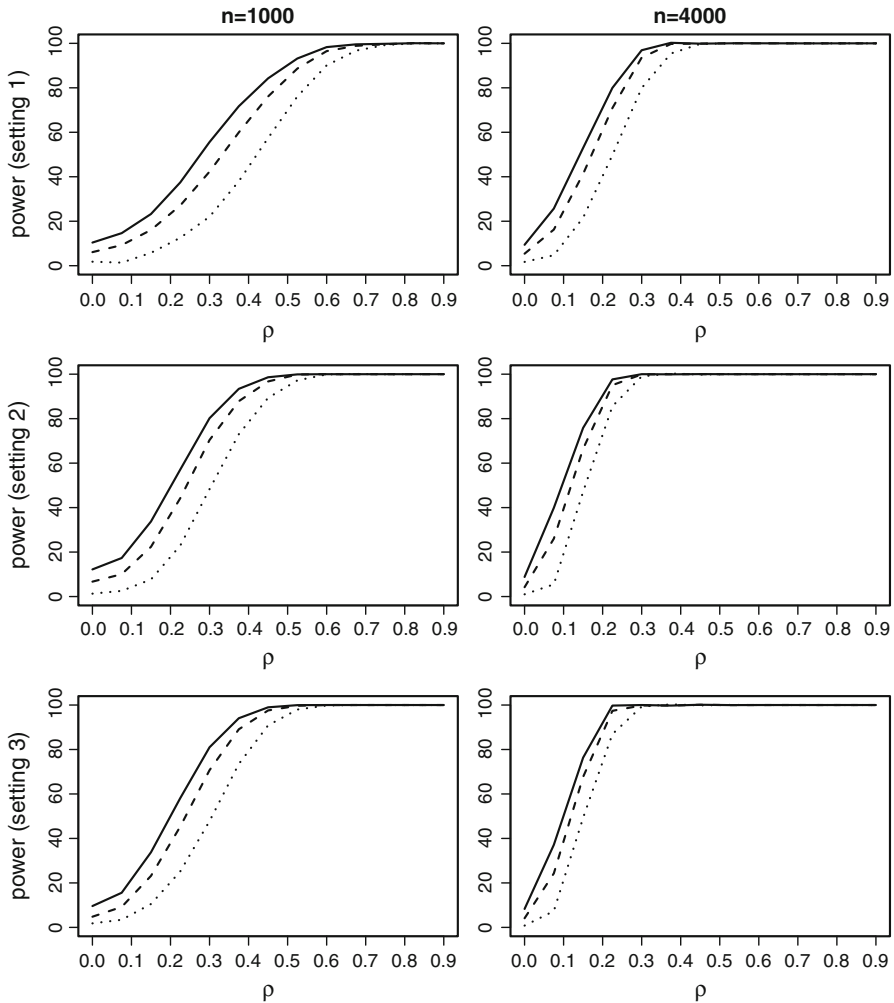


Fig. 2 Power curves for $LM_{\mathcal{H}}$ obtained under DGP2 and the three settings described in Sect. 4.1.1. The solid, dashed and dotted lines represent curves at the 10, 5 and 1% significance levels, respectively

ter settings reported in Table 1. We also considered a variant of DGP4, useful to test what happens in the situation of distributional misspecification. As before, we generated data assuming uncorrelated gamma errors with shape and scale parameters equal to 2. The sample sizes considered were 1,000 and 4,000. Each design was replicated 1,000 times.

4.2.2 Fitting details

We specified the two model equations

$$\begin{aligned} \text{eq3} &<- y1 \sim m1 + s(z1) + s(z2) \\ \text{eq4} &<- y2 \sim m1 + s(z1) \end{aligned}$$

Table 3 Size results (in %) for $LM\gamma\mathcal{H}$ in the endogeneity case under the scenarios of no misspecification and misspecified error distribution and functional form

α (%)	n	No misspecification			Error distribution			Functional form			
		Setting 1	Setting 2	Setting 3	Setting 1	Setting 2	Setting 3	Setting 1	Setting 2	Setting 3	
		ER	1	1,000	1.8	1.3	1.8	4.9	1.9	1.8	1.1
	5		6.1	6.7	4.8	10.2	7.7	7.8	8.6	7.6	7.4
	10		10.4	12.2	9.6	15.2	13.0	13.1	10.2	14.7	13.5
	1	4,000	1.7	1.0	0.8	1.2	2.2	2.0	1.5	2.4	2.9
	5		5.4	5.2	5.1	6.2	8.5	7.1	5.3	9.0	8.4
	10		9.4	9.8	9.3	10.3	12.6	13.3	8.8	15.3	14.3
Non-ER	1	1,000	56.2	7.3	13.3	23.4	23.3	8.2	3.9	12.1	4.5
	5		62.5	13.3	19.3	31.8	31.9	15.6	8.1	27.6	9.7
	10		66.8	19.2	23.6	37.5	35.8	21.3	19.5	42.9	19.8
	1	4,000	59.7	8.9	12.2	47.2	16.6	9.7	3.1	38.8	4.9
	5		65.2	16.9	17.4	60.0	23.5	18.6	7.8	66.0	9.8
	10		70.5	23.1	21.2	66.3	30.4	25.3	18.9	77.4	20.0

Data were generated using model (12) of DGP2. Non-ER refers to the case in which ϵ_{2i} (z_2) is not included in the first Eq. (eq1)

for DGP3 and

```
eq3 <- y1 ~ m1 + s(z1) + s(z2)
eq4 <- y2 ~ m1 + s(z1, by = m1)
```

for DGP4. p values for $LM_{\mathcal{H}}$ and $LM_{\mathcal{I}}$ were calculated using

```
LM.bpm(eq3, eq4, data, FI = FALSE, selection = TRUE)
LM.bpm(eq3, eq4, data, FI = TRUE, selection = TRUE)
```

where `selection` was set to `TRUE` to use the tests for the sample selection model case. p values for $\mathcal{W}_{\mathcal{I}}$ were calculated using the quantities $\hat{\rho}$ and $\text{Var}(\hat{\rho})$ obtained from the output produced by `SemiParBIVProbit(eq3, eq4, data, selection = TRUE)`.

In line with the endogeneity case, for DGP4, we generated data using gamma errors and employed the same tests. We also considered a scenario with functional form misspecification where the model equation for y_2 did not include the interaction term.

4.2.3 Monte Carlo results

Because the conclusions in this section are similar to those obtained in Sect. 4.1.3, we only report some of the results; “Appendix 2” contains the full set of results.

Regarding the empirical sizes, $LM_{\mathcal{H}}$ performs well, whereas the tests based on \mathcal{I} perform poorly (see Table 4). The good performance of $LM_{\mathcal{H}}$ is also confirmed by the power results reported in Figs. 3 and 4 in “Appendix B”.

As for model misspecification, the same conclusions as those for the endogeneity case are reached here; under functional and distributional misspecification the performance of $LM_{\mathcal{H}}$ worsens (see Table 8, ER case, “Appendix 2”). Also, the test’s performance is poor in the absence of ER and is slightly better in the presence of ER. The main findings of our simulation study can be summarized as follows. 1) $LM_{\mathcal{H}}$ yields close to nominal empirical sizes and strongly outperforms the tests based on the expected information. $LM_{\mathcal{I}}$ and $W_{\mathcal{I}}$ are characterized by zero and very high rejection frequencies, respectively. 2) $LM_{\mathcal{H}}$ produces satisfactory power results which improve as n and ρ increase. 3) Under misspecification, the performance of $LM_{\mathcal{H}}$ worsens. 4) When the ER does not hold, the test is not reliable. 5) The good performance of $LM_{\mathcal{H}}$ is important for practitioners wishing to test the hypothesis of absence of unobserved confounding without coping with simultaneous estimation.

5 Real data illustrations

We illustrate the tests using two case studies in which the issues of endogeneity and non-random sample selection arise. The first concerns a study, conducted in Botswana, on the impact of education on women’s fertility (www.measuredhs.com) and contains around 4,300 observations. Education is equal to 1 if the woman had at least 8 years of education and 0 otherwise, and fertility is equal to 1 if the woman had at least one child. The proportion of 1’s for the two variables is 28.9 and 74 %, respectively. As suggested by many scholars, estimation of such an effect can be biased by the possible endogeneity arising because unobserved confounders (e.g., ability and motivation)

Table 4 Size results (in %) for the non-random sample selection case

	α (%)	n	Setting 1			Setting 2			Setting 3		
			$LM\mathcal{H}$	$LM\mathcal{I}$	$\mathcal{W}\mathcal{I}$	$LM\mathcal{H}$	$LM\mathcal{I}$	$\mathcal{W}\mathcal{I}$	$LM\mathcal{H}$	$LM\mathcal{I}$	$\mathcal{W}\mathcal{I}$
DGP3	1	1,000	1.5	0.0	19.6	1.6	0.0	19.2	1.6	0.00	22.1
	5		6.2	0.0	32.7	6.3	0.1	37.1	6.1	0.00	36.1
	10		11.0	0.2	40.1	11.2	0.3	40.6	11.3	0.00	43.1
	1	4,000	0.9	0.0	17.4	1.5	0.0	17.1	1.5	0.00	22.1
	5		4.9	0.0	30.8	5.4	0.0	33.5	5.4	0.00	30.2
	10		9.8	0.2	40.5	9.0	0.4	40.3	11.4	0.00	42.3
DGP4	1	1,000	1.6	0.0	29.1	1.4	0.0	17.0	1.6	0.00	22.9
	5		6.0	0.0	43.0	6.8	0.0	23.3	5.8	0.00	35.3
	10		9.4	0.0	51.8	12.1	0.0	31.0	10.6	0.02	44.7
	1	4,000	1.5	0.0	26.5	1.1	0.0	3.9	0.9	0.00	19.0
	5		5.7	0.0	39.1	5.2	0.0	21.4	4.7	0.00	30.1
	10		10.4	0.0	48.5	9.8	0.0	38.2	9.3	0.02	40.2

DGP3 and DGP4 relate to models (13) and (14). α and n denote the critical value and sample size considered. These were obtained employing the Lagrange multiplier (LM) tests based on the observed ($-\mathcal{H}$) and expected (\mathcal{I}) information matrices, and the Wald (\mathcal{W}) test based on \mathcal{I}

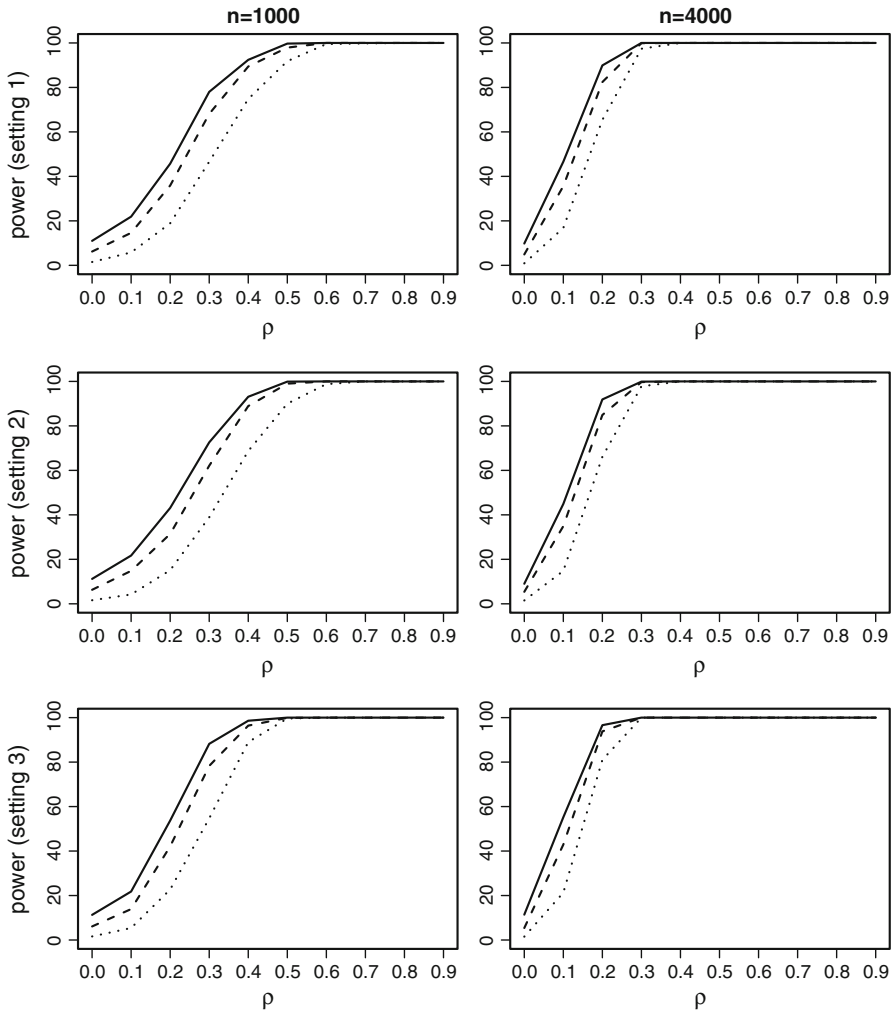


Fig. 3 Power curves for $LM_{\mathcal{H}}$ obtained under DGP3 and the three settings described in Sect. 4.2.1. The *solid*, *dashed* and *dotted lines* represent curves at the 10, 5 and 1% significance levels, respectively

are associated with both fertility and education; full details can be found in [Marra and Radice \(2011\)](#) and references therein. The second dataset considers an American survey of public opinion polls on school integration (www.electionstudies.org) where about 700 individuals were first asked if they had an opinion on the integration question (0=no, 1=yes) and then what that opinion was (0=no integration, 1=yes integration). This gave respondents an opportunity to opt out of the question answering process at an earlier stage. 64.57% of the individuals chose to answer the integration question. Among these, the proportion of yes answers was 46.43%. Because it is reasonable to assume that the decision to answer was not random, sample selection bias can occur when estimating the model parameters; full details can be found in [Marra and Radice \(2013a\)](#) and references therein.

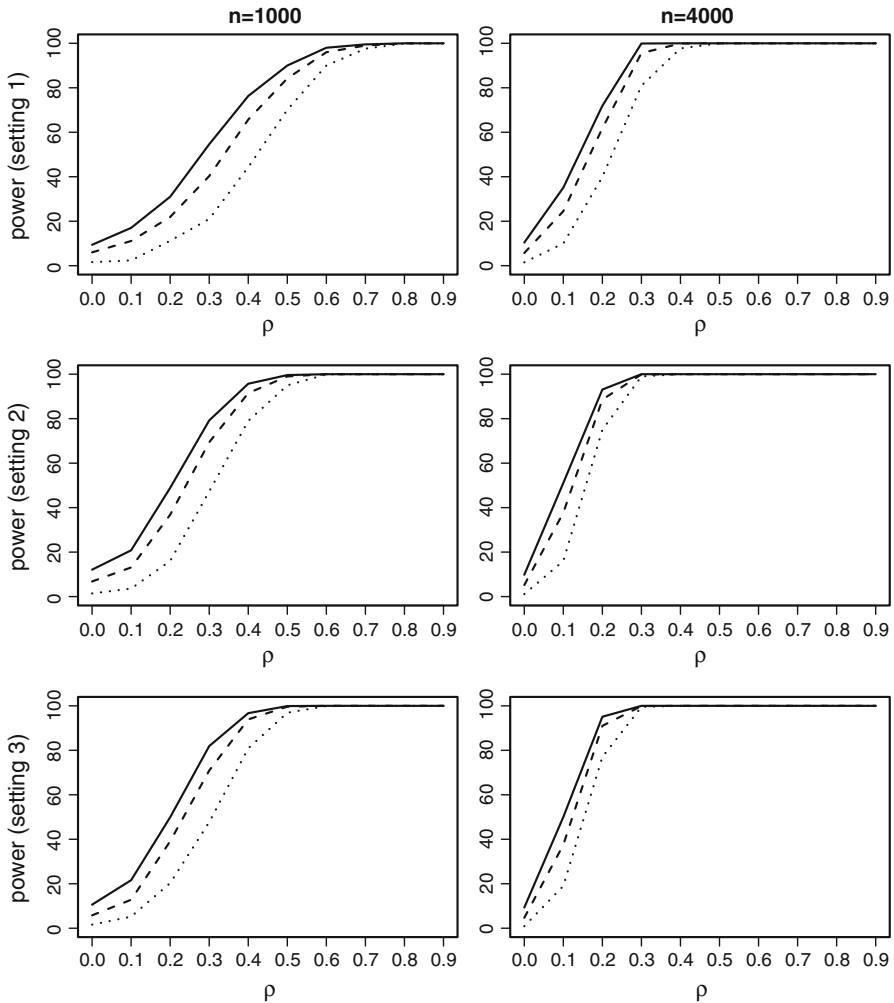


Fig. 4 Power curves for $LM_{\mathcal{H}}$ obtained under DGP4 and the three settings described in Sect. 4.2.1. The solid, dashed and dotted lines represent curves at the 10, 5 and 1% significance levels, respectively

In the fertility study, the direction of the bias due to unobserved confounding can not be determined a priori. The reason for this ambiguity is because of different substitution and income effects (Cygán-Rehm and Maeder 2012). As for the school integration study, the bias is expected to be negative because some individuals choose not to answer the integration question as they feel that their opinion may be perceived as socially unacceptable (Berinsky 1999).

5.1 Fertility dataset

Following previous work on the subject (Wooldridge 2010; Marra and Radice 2011; Sobotka et al. 2013), we specified a semiparametric recursive bivariate probit model

Table 5 Variables in the education-fertility data set

Variable	Explanation
child	Number of children—binary: 1 (at least one child), 0 (otherwise)
ed	Number of years of schooling—binary: 1 (at least 8 years), 0 (otherwise)
elz	Household has electricity—binary: yes, no
urb	Household lives in urban area—binary: yes, no
em	Ever married—binary: yes, no
age	Age in years
fhalf	Born during the first 6 months of the year—binary: yes, no

with main terms only. The description of the variables used in the model are reported in Table 5. Specifically, the linear predictors of the treatment (ed) and outcome (child) equations included urb, em and el, whereas fhalf and ed entered the former and latter predictors, respectively. These variables entered the model as parametric components. Thin plate regression splines of age, with the same settings as those employed for the simulation study, were employed. The two equations are

$$\begin{aligned} \text{ed}_i^* &= \theta_{10} + \theta_{11}\text{el}_i + \theta_{12}\text{urb}_i + \theta_{13}\text{em}_i + \theta_{14}\text{fhalf}_i + f_{1\text{age}}(\text{age}_i) + \varepsilon_{1i}, \\ \text{child}_i^* &= \theta_{20} + \vartheta \text{ed}_i + \theta_{21}\text{el}_i + \theta_{22}\text{urb}_i + \theta_{23}\text{em}_i + f_{2\text{age}}(\text{age}_i) + \varepsilon_{2i}. \end{aligned}$$

As in Wooldridge (2010), the binary variable ‘born during the first 6 months of the year’ (fhalf) was used as an instrument on the grounds that it does not have a direct effect on fertility given covariates, influences education, and is unlikely to be associated with unobservable confounders such as ability and motivation. The nonparametric specification for age arises from the fact that this covariate embodies productivity and life-cycle effects that are likely to affect child non-linearly. Wooldridge (2010) considered a model for fertility that contains linear and quadratic terms in age, whereas Marra and Radice (2011) and Sobotka et al. (2013) specified a model containing a smooth function of age. Here, we found non-linear effects of age that are almost identical to those reported in Marra and Radice (2011); results are available upon request. The quantity of interest is the average treatment effect (ATE) of ed on child, which measures the effect of ed on the probability of having at least one child, i.e. $\mathbb{P}(\text{child} = 1)$.

Because the effect of education on fertility may be biased by the possible presence of endogeneity, as a first step of the analysis we tested the null hypothesis of absence of unobserved confounding. Specifically, we employed $LM_{\mathcal{H}}$ as, in simulation, it was shown to have good size and power properties. The p value obtained using $LM_{\mathcal{H}}$ was 0.00, suggesting that there is an issue of endogeneity. For completeness, we also report the results of the tests based on the expected information; the p values obtained using $LM_{\mathcal{I}}$ and $\mathcal{W}_{\mathcal{I}}$ were 0.94 and 0.00, respectively. The p value of $\mathcal{W}_{\mathcal{I}}$ suggests that there is an issue of endogeneity, whereas that of $LM_{\mathcal{I}}$ suggests that endogeneity is not present. The conflicting conclusions can be attributed to the fact that, as shown in our simulations, $LM_{\mathcal{I}}$ and $\mathcal{W}_{\mathcal{I}}$ are not reliable as under the null they yield zero and

high rejection frequencies, respectively. Based on the result produced by $LM_{\mathcal{H}}$, we estimated a recursive bivariate probit model and then calculated the ATE. The estimate in % (with 95% confidence interval) was -3.01 ($-5.90, -0.08$). This is statistically different from zero and suggests that having at least 8 years of schooling decreases the probability of having at least one child by 3%.

5.2 School integration dataset

For the school integration dataset, using the variables in Table 6 and in line with the works by Berinsky (1999) and Marra and Radice (2013a), we specified a semiparametric sample selection probit model based on the two equations

$$\begin{aligned} \text{opinion}_i^* &= \theta_{11} + \theta_{12}\text{sex}_i + \theta_{13}\text{race}_i + \theta_{14}\text{reg.northeast}_i \\ &\quad + \theta_{15}\text{reg.south}_i + \theta_{16}\text{reg.west}_i + \theta_{17}\text{chil} \\ &\quad + \theta_{18}\text{discpol} + \theta_{19}\text{moralcons} + \theta_{110}\text{perslett} \\ &\quad + f_{1\text{age}}(\text{age}_i) + f_{1\text{educ}}(\text{educ}_i) + \varepsilon_{1i}, \\ \text{integration}_i^* &= \theta_{21} + \theta_{22}\text{sex}_i + \theta_{23}\text{race}_i + \theta_{24}\text{reg.northeast}_i \\ &\quad + \theta_{25}\text{reg.south}_i + \theta_{26}\text{reg.west}_i + \theta_{27}\text{chil} + \theta_{28}\text{discpol} \\ &\quad + \theta_{29}\text{moralcons} + f_{2\text{age}}(\text{age}_i) + f_{2\text{educ}}(\text{educ}_i) + \varepsilon_{2i}, \end{aligned}$$

where, the equations included `sex`, `race`, `reg.northeast`, `reg.south`, `reg.west`, `discpol`, `moralcons` and `chil` as parametric components, and smooth functions of `age` and `educ`. These two continuous covariates are expected to have non-linear impacts on `integration` as well as `opinion`. `chil` was included as a parametric component because it did not have enough unique covariate values to justify the use of a smooth function. The selection equation (`opinion`) also included

Table 6 Variables in the school integration dataset

Variable	Explanation
<code>opinion</code>	Individual had opinion on the integration question—binary: yes, no
<code>integration</code>	Individual supports integration—binary: yes, no
<code>chil</code>	Number of children
<code>age</code>	Age in years
<code>educ</code>	Number of years of education
<code>sex</code>	Respondent is man—binary: yes, no
<code>race</code>	Respondent is white—binary: yes, no
<code>reg.northeast</code>	Respondent lives in north-east region—binary: yes, no
<code>reg.south</code>	Respondent lives in south region—binary: yes, no
<code>reg.west</code>	Respondent lives in west region—binary: yes, no
<code>discpol</code>	Respondent discusses politics—binary: yes, no
<code>moralcons</code>	Moral conservatism – 1 = support, 2 = no support, 3 = neither
<code>perslett</code>	Respondent was persuaded to participate in the survey—binary: yes, no

perslett. This was because, according to Berinsky (1999), those individuals who are difficult to reach are also reluctant to answer specific survey questions. The inclusion of this variable in the selection equation served as ER. As in the fertility study, we found estimated linear and non-linear effects of the continuous variables that are almost identical to those reported in Marra and Radice (2013a).

The p value obtained using $LM_{\mathcal{H}}$ was 0.00, whereas those obtained using $LM_{\mathcal{I}}$ and $\mathcal{W}_{\mathcal{I}}$ were 0.76 and 0.00. These lead to the same conclusions reached for the fertility study. In summary, $LM_{\mathcal{H}}$, which is the most reliable test, supports the presence of non-random sample selection. Therefore, we estimated the model parameters using the sample selection bivariate probit model and calculated the mean predicted probability (and associated confidence interval) of giving a supportive response. This was 0.34 (0.22, 0.45).

In both examples (fertility and school integration) $LM_{\mathcal{H}}$ suggests rejecting the null hypothesis of absence of unobserved confounding. In such cases, estimates of the parameters of interest (ATE and mean predicted probability) are obtained by using bivariate probit models. If we were not to reject the null hypothesis then we would estimate the quantities of interest by using univariate models, hence avoiding the use of a simultaneous estimation approach.

Following a reviewer's suggestion, for both case studies, we obtained p values using $LM_{\mathcal{H}}$ based on a parametric specification of the model equations (i.e., the continuous covariates were assumed to have a linear impact). The values were 0.12 and 0.03 for the fertility and school integration studies, respectively. In the first case, the conclusion would be that there is not an issue of endogeneity, which is not consistent with the results reported in this section as well as in the literature on this topic. In the second case, conclusions would not be altered. In general, we recommend using a model specification that reduces the risk of functional form misspecification which may have a detrimental impact on the reliability of the test.

6 Discussion

We extended LM and \mathcal{W} type tests for the hypothesis of absence of unobserved confounding to the context of semiparametric recursive and sample selection bivariate probit models. The finite sample size performance of the tests was examined via a Monte Carlo study under several scenarios: correct model specification, distributional and functional misspecification, with and without an ER. The results allowed us to derive some guidelines which may be important for empirical applications. First, under correct model specification, LM based on $-\mathcal{H}$ (the observed information) performs well, whereas the statistics based on \mathcal{I} (the expected information) are characterized by zero and very high rejection frequencies, suggesting these tests should not to be used for empirical analysis. Second, LM performs satisfactorily only when the ER holds. Third, the availability of a valid instrument can alleviate but not eliminate the detrimental effect that model misspecification has on the empirical performance of the test.

The good performance of $LM_{\mathcal{H}}$ should be particularly attractive to practitioners wishing to test the null hypothesis of absence of unobserved confounding while

avoiding simultaneous estimation. This implies that simultaneous estimation will be employed only if unobserved confounding is detected.

The *LM* statistic presented here can in principle be generalized to any other model which controls for endogeneity or non-random sample selection. Examples are copula or count data endogenous and non-random sample selection models (Bratti and Miranda 2011; Winkelmann 2011; Zimmer and Trivedi 2006; Smith 2003). More generally, this test could be extended to any other context where there is a system of equations and testing their independence is important (Kiefer 1982; Yee and Wild 1996).

Acknowledgments Giampiero Marra was supported by the Engineering and Physical Sciences Research Council, UK (Grant EP/J006742/1). We wish to thank Ioannis Kosmidis for his extremely helpful comments after reading a revised version of the work, and the Editor, Associate Editor and two reviewers for their constructive criticism which helped to improve the presentation of the article.

Appendix

Appendix A

For the recursive bivariate probit model case, matrix \mathcal{I} becomes block diagonal under $H_0 : \rho = 0$. This is because \mathcal{I}_{12} , \mathcal{I}_{13} and \mathcal{I}_{23} , reported in Marra and Radice (2011), are equal to $\mathbf{0}$ under the null. In what follows, we ignore the role of the penalty because it is block diagonal and hence does not affect the aforementioned quantities. Consider, for instance, \mathcal{I}_{13} . Under the null

$$\begin{aligned} \mathcal{I}_{13} &= \sum_{i=1}^N \phi_{1i}^2 \phi_{2i} \left[\Phi_{2i} \left\{ \frac{1}{\Phi_{1i} \Phi_{2i}} + \frac{1}{(1 - \Phi_{1i}) \Phi_{2i}} \right\} \right. \\ &\quad \left. - (1 - \Phi_{2i}) \left\{ \frac{1}{\Phi_{1i} (1 - \Phi_{2i})} + \frac{1}{(1 - \Phi_{1i}) (1 - \Phi_{2i})} \right\} \right] \mathbf{X}_{1i} \\ &= \sum_{i=1}^N \phi_{1i}^2 \phi_{2i} \left[\frac{1}{\Phi_{1i}} + \frac{1}{1 - \Phi_{1i}} - \frac{1}{\Phi_{1i}} - \frac{1}{1 - \Phi_{1i}} \right] \mathbf{X}_{1i} = \mathbf{0}, \end{aligned}$$

where, using a shorthand notation, $\Phi_{1i} \Phi_{2i}$, $\Phi_{1i} (1 - \Phi_{2i})$, $(1 - \Phi_{1i}) \Phi_{2i}$ and $(1 - \Phi_{1i}) (1 - \Phi_{2i})$ are the quantities for p_{11i} , p_{10i} , p_{01i} and p_{00i} obtained under H_0 , and Φ_{1i} , Φ_{2i} , ϕ_{1i} and ϕ_{2i} denote the probability and density functions of a standardized normal evaluated at their corresponding linear predictors η_{1i} and η_{2i} .

Appendix B

Table 7 Further size results (in %) for $LM_{\mathcal{H}}$ in the endogeneity case under the scenarios of misspecified error distribution and functional form

	α (%)	n	Error distribution			Functional form		
			Setting 4	Setting 5	Setting 6	Setting 4	Setting 5	Setting 6
			ER	1	1,000	2.3	1.8	2.3
	5		8.2	7.9	8.8	8.6	8.6	7.3
	10		14.2	13.2	13.6	12.1	14.9	12.5
	1	4,000	2.5	2.3	2.2	2.5	2.5	2.7
	5		8.2	8.7	8.1	8.3	8.5	8.1
	10		13.3	12.1	13.6	12.8	14.3	14.8

Data were generated using model (12) of DGP2. ER refers to the cases in which z_2 is included in the first equation (eq1). Parameter vector values for the three settings were obtained randomly from a standardized normal distribution

Table 8 Size results (in %) for $LM\mathcal{H}$ in the non-random sample selection case under the scenarios of no misspecification and misspecified error distribution and functional form

α (%)	n	No misspecification			Error distribution			Functional form			
		Setting 1	Setting 2	Setting 3	Setting 1	Setting 2	Setting 3	Setting 1	Setting 2	Setting 3	
		ER	1	1,000	1.6	1.4	1.6	2.6	2.0	2.0	1.5
	5		6.0	6.8	5.8	8.5	7.8	8.8	8.9	7.9	8.1
	10		9.4	12.1	10.6	15.2	12.7	15.1	10.0	13.3	12.9
	1	4,000	1.5	1.1	0.9	2.5	2.1	2.3	1.7	2.1	2.6
	5		5.7	5.2	4.7	6.4	9.9	7.9	6.3	10.0	8.8
	10		10.4	9.8	9.3	12.1	13.1	14.3	9.8	15.1	13.3
Non-ER	1	1,000	50.1	8.1	14.1	22.4	21.9	8.5	4.0	10.1	4.9
	5		61.1	14.1	19.9	30.1	30.8	14.6	9.1	26.5	10.1
	10		64.9	20.2	24.5	36.2	39.5	20.1	20.1	44.8	20.2
	1	4,000	53.7	9.9	13.1	46.1	17.1	10.0	3.7	39.1	4.8
	5		66.0	16.6	18.2	58.9	25.1	20.0	8.8	69.0	10.8
	10		71.5	24.0	22.1	67.1	29.9	24.1	19.6	75.3	21.1

Data were generated using model (15) of DGP4. Non-ER refers to the case in which z_{2j} (z_2) is not included in the first equation (eq3)

References

- Banasik J, Crook J (2007) Reject inference, augmentation, and sample selection. *Eur J Oper Res* 183: 1582–1594
- Bärnighausen T, Bor J, Wandira-Kazibwe S, Canning D (2011) Correcting HIV prevalence estimates for survey nonparticipation using heckman-type selection models. *Epidemiology* 22:27–35
- Berinsky A (1999) The two faces of public opinion. *Am J Polit Sci* 43:1209–1230
- Bratti M, Miranda A (2011) Endogenous treatment effects for count data models with endogenous participation or sample selection. *Health Econ* 20:90–1109
- Buchmueller TC, Grumbach K, Kronick R, Kahn JG (2005) Book review: the effect of health insurance on medical care utilization and implications for insurance expansion: a review of the literature. *Med Care Res Rev* 62:3–30
- Cao X, Spall JC (2009) Preliminary results on relative performance of expected and observed fisher information. In: Proceedings of the 48th IEEE conference on decision and control, CDC 2009, combined with the 28th Chinese control conference, Dec 16–18 2009. IEEE, Shanghai, China, pp 1538–1543
- Cavanaugh JE, Shumway RH (1996) On computing the expected fisher information matrix for state-space model parameters. *Stat Prob Lett* 26:347–355
- Chib S, Greenberg E (2007) Semiparametric modeling and estimation of instrumental variable models. *J Comput Graph Stat* 16:86–114
- Chib S, Greenberg E, Jeliaskov I (2009) Estimation of semiparametric models in the presence of endogeneity and sample selection. *J Comput Graph Stat* 18:321–348
- Cuddeback G, Wilson E, Orme J, Combs-Orme T (2004) Detecting and statistically correcting sample selection bias. *J Soc Serv Res* 30:19–33
- Cygan-Rehm K, Maeder M (2012) The effect of education on fertility: evidence from a compulsory schooling reform. Working Papers 121, Bavarian Graduate Program in Economics (BGPE). http://ideas.repec.org/p/bav/wpaper/121_cyganrehmmaeder.html
- de Ven WV, Praag BV (1981) The demand for deductibles in private health insurance: a probit model with sample selection. *J Econom* 17:229–252
- Efron B, Hinkley DV (1978) Assessing the accuracy of the maximum likelihood estimator: observed versus expected fisher information. *Biometrika* 65:457–487
- Eilers PHC, Marx BD (1996) Flexible smoothing with B-splines and penalties. *Stat Sci* 11(2):89–121
- Goldman D, Bhattacharya J, McCaffrey D, Duan N, Leibowitz A, Joyce G, Morton S (2001) Effect of insurance on mortality in an HIV-positive population in care. *J Am Stat Assoc* 96:883–894
- Greene WH (2012) *Econometric analysis*. Prentice Hall, New York
- Hastie T, Tibshirani R (1993) Varying-coefficient models. *J Roy Stat Soc Ser B* 55:757–796
- Heckman J (1978) Dummy endogenous variables in a simultaneous equation system. *Econometrica* 46: 931–959
- Heckman J (1979) Sample selection bias as a specification error. *Econometrica* 47:153–161
- Kawatkar AA, Nichol MB (2009) Estimation of causal effects of physical activity on obesity by a recursive bivariate probit model. *Value Health* 12:A131–A132
- Kiefer NM (1982) Testing for dependence in multivariate probit models. *Biometrika* 69:161–166
- Latif E (2009) The impact of diabetes on employment in Canada. *Health Econ* 18:577–589
- Louis TA (1982) Finding the observed information matrix when using the em algorithm. *J Roy Stat Soc Ser B* 44:226–233
- Maddala GS (1983) *Limited dependent and qualitative variables in econometrics*. Cambridge University Press, Cambridge
- Maldonado G, Greenland S (1994) A comparison of the performance of model-based confidence intervals when the correct model form is unknown: coverage of asymptotic means. *Epidemiology* 5:171–182
- Marra G, Radice R (2010) Penalised regression splines: theory and application to medical research. *Stat Methods Med Res* 19:107–125
- Marra G, Radice R (2011) Estimation of a semiparametric recursive bivariate probit model in the presence of endogeneity. *Canad J Stat* 39:259–279
- Marra G, Wood S (2012) Coverage properties of confidence intervals for generalized additive model components. *Scand J Stat* 39:53–74
- Marra G, Radice R (2013a) A penalized likelihood estimation approach to semiparametric sample selection binary response modeling. *Electron J Stat* 7:1432–1455

- Marra G, Radice R (2013b) SemiParBIVProbit: semiparametric bivariate probit modelling. R package version 3.2-6
- Monfardini C, Radice R (2008) Testing exogeneity in the bivariate probit model: a Monte Carlo study. *Oxf Bull Econ Stat* 70:271–282
- Montmarquette C, Mahseredjiana S, Houle R (2001) The determinants of university dropouts: a bivariate probability model with sample selection. *Econ Educ Rev* 20:475–484
- Nocedal J, Wright SJ (2006) Numerical optimization. Springer, New York
- R Development Core Team (2013) R: a Language and environment for statistical computing. R foundation for statistical computing. Vienna, Austria. ISBN 3-900051-07-0
- Smith MD (2003) Modelling sample selection using archimedean copulas. *Econom J* 6:99–123
- Sobotka F, Radice R, Marra G, Kneib T (2013) Estimating the relationship of women's education and fertility in botswana using an instrumental variable approach to semiparametric expectile regression. *J Roy Stat Soc Ser C* 62:1–21
- Wilde J (2000) Identification of multiple equation probit models with endogenous dummy regressors. *Econ Lett* 69:309–312
- Winkelmann R (2011) Copula bivariate probit models: with an application to medical expenditures. *Health Econ* 21:1444–1455
- Wood SN (2006) Generalized additive models: an introduction with R. Chapman and Hall, London
- Wooldridge JM (2010) Econometric analysis of cross section and panel data. MIT Press, Cambridge
- Yee TW, Wild CJ (1996) Vector generalized additive models. *J Roy Stat Soc Ser B* 58:481–493
- Zimmer DM, Trivedi PK (2006) Using trivariate copulas to model sample selection and treatment effects: application to family health care demand. *Journal of Business and Economic Statistics* 24:63–76