Regression Spline Bivariate Probit Models: A Practical Approach to Testing for Exogeneity

Giampiero Marra[1]*, Rosalba Radice[2] and Panagiota Filippou[1]

[1]Department of Statistical Science, University College London,

Gower Street, London WC1E 6BT, U.K.

[2]Department of Economics, Mathematics and Statistics, Birkbeck, University of London,

Malet Street, London WC1E 7HX, U.K.

Abstract

Bivariate probit models can deal with a problem usually known as endogeneity. This issue is likely to arise in observational studies when confounders are unobserved. We are concerned with testing the hypothesis of exogeneity (or absence of endogeneity) when using regression spline recursive and sample selection bivariate probit models. Likelihood ratio and gradient tests are discussed in this context and their empirical properties investigated and compared with those of the Lagrange multiplier and Wald tests through a Monte Carlo study. The tests are illustrated using two datasets in which the hypothesis of exogeneity needs to be tested.

Key Words: Bivariate probit model; Endogeneity; Gradient test; Lagrange multiplier test; Likelihood ratio test; Non-random sample selection; Penalized regression spline; Wald test.

*giampiero.marra@ucl.ac.uk

1 Introduction

Recursive and sample selection bivariate probit models deal with a problem which arises in observational studies when confounders (i.e., variables that are associated with treatment, or selection, and response) are unobserved (Heckman, 1978, 1979; Maddala, 1983; de Ven & Praag, 1981; Greene, 2012). This issue is known in the econometric literature as endogeneity. Bivariate probit models control for unobserved confounders by using a two-equation structural latent variable framework, where one equation models a binary outcome as a function of observed confounders and a treatment, or selection, whereas the other equation models a binary treatment or selection process. For some economic and biostatistical applications see Banasik & Crook (2007), Bärnighausen et al. (2011), Buchmueller et al. (2005), Cuddeback et al. (2004), Goldman et al. (2001), Kawatkar & Nichol (2009), Latif (2009), Montmarquette et al. (2001) and Radice et al. (2013). Marra & Radice (2011a, 2013) introduced a penalized likelihood estimation framework to estimate recursive and sample selection bivariate probit models that include smooth functions of continuous confounders: the regression spline bivariate probit models. This extension is of some relevance as mis-modeling the relationship between observed confounders and outcomes can lead to inconsistent estimates of all model parameters (Chib & Greenberg, 2007; Marra & Radice, 2011a, and references therein).

We are concerned with testing the hypothesis of exogeneity (or absence of endogeneity) in the context of regression spline recursive and sample selection bivariate probit models. Marra et al. (2014) introduced a Lagrange multiplier (LM) test for this class of models. In this paper, we propose two more tests: the likelihood ratio (LR) and gradient (G) tests. In the classic bivariate probit context (where the functional form of covariate effects is specified a priori by the investigator), Monfardini & Radice (2008) found via an extensive simulation study that the LR test performs the best, especially in challenging scenarios. Marra et al. (2014) pointed out that the difficulty with the use of LR for regression spline bivariate probit models is that the number of degrees of freedom of the test is not guaranteed to be an integer value. We propose to address this problem by using the simple but effective idea that a penalized regression spline model can be approximated by a pure regression

spline model (e.g., Wood, 2006). We also explore the use of the G test within this class of models. The G test (Terrell, 2002) is not well known in the econometric and statistical literature, and has good practical potential (see Vargas et al., 2013, and references therein). It is important to stress that, similarly to the works of Monfardini & Radice (2008) and Marra et al. (2014), the approach taken here is practical in the sense that theoretical results are borrowed from classic asymptotic theory and some guidelines are suggested to aid the analyst in investigating whether there is an issue of endogeneity. The G test is implemented in the `R` package `SemiParBIVProbit` (Marra & Radice, 2015), whereas the implementation of the LR test is illustrated in the paper.

The article is organized as follows. Section 2 provides a brief overview of the models of interest and their estimation with the aim of defining the notation and making some remarks that are relevant to the implementation of the tests. Section 3 discusses briefly the LM test introduced by Marra et al. (2014), the classic Wald (W) test and describes the construction of the G and LR tests. Section 4 compares the finite sample size properties of the tests through a Monte Carlo simulation study. Section 5 illustrates the tests using two datasets on health care utilization and HIV, whereas Section 6 provides a discussion.

2 Preliminaries

The regression spline recursive and sample selection bivariate probit models introduced by Marra & Radice (2011a, 2013) generalize 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 treatment or selection equation can be written as

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

where n is the sample size, y_{1i}^* is a latent continuous variable and y_{1i} is determined via the rule $\mathbf{1}(y_{1i}^* > 0)$ (recall that $\mathbf{1}$ is the usual indicator function). The outcome equation in the

regression spline recursive bivariate probit model can be defined as

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

where y_{2i} is determined via the rule $\mathbf{1}(y_{2i}^{*} > 0)$ and ϑ is the effect of the treatment variable. In the sample selection case, the outcome equation can be written as

$$y_{2i}^{*} = \left\{ \mathbf{m}_{2i}^{\mathsf{T}}\boldsymbol{\theta}_2 + \sum_{k_2=1}^{K_2} s_{2k_2}(z_{2k_2 i}) + \varepsilon_{2i} \right\} \times y_{1i}, \tag{3}$$

with y_{2i} determined as

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

Vector $\mathbf{m}_{1i}$ contains P_1 parametric model components (i.e., intercept, dummy and categorical variables), $\boldsymbol{\theta}_1$ is a parameter vector, and the s_{1k_1} are unknown smooth functions of the K_1 continuous covariates $z_{1k_1 i}$. Varying coefficient terms and smooth functions of two covariates can also be considered (e.g., Hastie & Tibshirani, 1993; 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 identifiability constraints, i.e. $\sum_i s_{vk_v}(z_{vk_v i}) = 0$, $v = 1, 2$, for all smooth functions 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). The smooth functions are represented using regression splines (e.g., Eilers & Marx, 1996). In this approach, a generic function $s_k(z_{ki})$ (note that subscript v has been suppressed to avoid clutter) is approximated by a linear combination of known spline basis functions, $b_{kj}(z_{ki})$, and regression parameters, β_{kj}, i.e. $\sum_{j=1}^{J_k} \beta_{kj} b_{kj}(z_{ki}) = \mathbf{B}_k(z_{ki})^{\mathsf{T}}\boldsymbol{\beta}_k$, where J_k is the number of spline bases, $\mathbf{B}_k(z_{ki}) = \{b_{k1}(z_{ki}), \ldots, b_{kJ_k}(z_{ki})\}^{\mathsf{T}}$ is a vector of the basis functions evaluated at z_{ki} and $\boldsymbol{\beta}_k$ is the corresponding parameter vector. Basis functions with convenient mathematical and numerical properties include B-splines, cubic regression and thin plate

regression splines (see, e.g., Marra & Radice (2010) for an overview). Based on the result above, equations (1), (2) and (3) can be written as $y_{1i}^* = \mathbf{m}_{1i}^\mathsf{T}\boldsymbol{\theta}_1 + \mathbf{B}_{1i}^\mathsf{T}\boldsymbol{\beta}_1 + \varepsilon_{1i} = \eta_{1i} + \varepsilon_{1i}$, $y_{2i}^* = \vartheta y_{1i} + \mathbf{m}_{2i}^\mathsf{T}\boldsymbol{\theta}_2 + \mathbf{B}_{2i}^\mathsf{T}\boldsymbol{\beta}_2 + \varepsilon_{2i} = \eta_{2i} + \varepsilon_{2i}$, and $y_{2i}^* = \left\{\mathbf{m}_{2i}^\mathsf{T}\boldsymbol{\theta}_2 + \mathbf{B}_{2i}^\mathsf{T}\boldsymbol{\beta}_2 + \varepsilon_{2i}\right\} \times y_{1i} = \left\{\eta_{2i} + \varepsilon_{2i}\right\} \times y_{1i}$, respectively, where $\mathbf{B}_{vi}^\mathsf{T} = \left\{\mathbf{B}_{v1}(z_{v1i})^\mathsf{T}, \ldots, \mathbf{B}_{vK_v}(z_{vK_vi})^\mathsf{T}\right\}$, $\boldsymbol{\beta}_v^\mathsf{T} = (\boldsymbol{\beta}_{v1}^\mathsf{T}, \ldots, \boldsymbol{\beta}_{vK_v}^\mathsf{T})$ and η_{1i} and η_{2i} have the obvious definitions. The overall parameter vector can be defined as $\boldsymbol{\delta}^\mathsf{T} = (\boldsymbol{\delta}_1^\mathsf{T}, \boldsymbol{\delta}_2^\mathsf{T}, \rho)$ where $\boldsymbol{\delta}_1^\mathsf{T} = (\boldsymbol{\theta}_1^\mathsf{T}, \boldsymbol{\beta}_1^\mathsf{T})$, and $\boldsymbol{\delta}_2^\mathsf{T} = (\vartheta, \boldsymbol{\theta}_2^\mathsf{T}, \boldsymbol{\beta}_2^\mathsf{T})$ or $\boldsymbol{\delta}_2^\mathsf{T} = (\boldsymbol{\theta}_2^\mathsf{T}, \boldsymbol{\beta}_2^\mathsf{T})$ depending on whether a recursive or sample selection bivariate probit model is employed.

To identify the parameters of equation (2) or (3) it is typically assumed that an exclusion restriction (ER) on the exogenous variables holds. That is, the covariates in equation (1) should contain at least one or more regressors (typically referred to as instruments) not included in equation (2) or (3). 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. For recursive bivariate probit models, this restriction may not be necessary in estimation (e.g., Han & Vytlacil, 2014; Marra & Radice, 2011a; Wilde, 2000). However, ER is crucial to identify the parameters of a sample selection model (e.g., Marra & Radice, 2013, and references therein).

2.1 Parameter estimation

In the sample selection case 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

$$\ell(\boldsymbol{\delta}) = \sum_{i=1}^{n} \left\{ y_{1i} y_{2i} \log(p_{11i}) + y_{1i}(1 - y_{2i}) \log(p_{10i}) + (1 - y_{1i}) \log(p_{0i}) \right\}, \tag{4}$$

where $\boldsymbol{\delta}^\mathsf{T} = (\boldsymbol{\delta}_1^\mathsf{T}, \boldsymbol{\delta}_2^\mathsf{T}, \rho)$. 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 (4), $(1 - y_{1i}) \log(p_{0i})$ is replaced by $(1 - y_{1i}) y_{2i} \log(p_{01i}) + (1 - y_{1i})(1 - y_{2i}) \log(p_{00i})$. Because of the presence of smooth functions in the equations, to avoid overfitting the models are estimated using a penalized log-likelihood (Marra & Radice, 2011a, 2013; Wood, 2006).

In particular, the model parameters are estimated by maximization of

$$\ell_p(\boldsymbol{\delta}) = \ell(\boldsymbol{\delta}) - \frac{1}{2}\boldsymbol{\beta}^{\mathsf{T}}\tilde{\mathbf{S}}_{\boldsymbol{\lambda}}\boldsymbol{\beta}, \tag{5}$$

where $\boldsymbol{\beta}^{\mathsf{T}} = (\boldsymbol{\beta}_1^{\mathsf{T}}, \boldsymbol{\beta}_2^{\mathsf{T}})$, $\tilde{\mathbf{S}}_{\boldsymbol{\lambda}} = \sum_{v=1}^{2}\sum_{k_v=1}^{K_v}\lambda_{vk_v}\mathbf{S}_{vk_v}$ and the $\mathbf{S}_{vk_v}$ are positive semi-definite symmetric known square matrices expanded with zeros everywhere except for the elements which correspond to the coefficients of the vk_v^{th} smooth term. The penalty typically measures the second-order roughness of the smooth terms in the model (e.g., Ruppert et al., 2003; Wood, 2006), i.e. $\boldsymbol{\beta}^{\mathsf{T}}\left(\sum_{v=1}^{2}\sum_{k_v=1}^{K_v}\lambda_{vk_v}\mathbf{S}_{vk_v}\right)\boldsymbol{\beta} = \sum_{v=1}^{2}\sum_{k_v=1}^{K_v}\lambda_{vk_v}\int f''_{vk_v}(z_{vk_v})^2 dz_{vk_v}$. The λ_{vk_v} are smoothing parameters controlling the trade-off between fit and smoothness. For instance, $\lambda_{vk_v} = 0$ results in an un-penalized regression spline estimate, while $\lambda_{vk_v} \to \infty$ leads to a straight line estimate for $s_{vk_v}(z_{vk_vi})$. The overall penalty can be also be defined as $1/2\boldsymbol{\delta}^{\mathsf{T}}\mathbf{S}_{\boldsymbol{\lambda}}\boldsymbol{\delta}$, where $\mathbf{S}_{\boldsymbol{\lambda}} = \text{diag}(0_{11}, \ldots, 0_{1P_1}, \lambda_{1k_1}\mathbf{S}_{1k_1}, \ldots, \lambda_{1K_1}\mathbf{S}_{1K_1}, 0_{21}, \ldots, 0_{2P_2}, \lambda_{2k_2}\mathbf{S}_{2k_2}, \ldots, \lambda_{2K_2}\mathbf{S}_{2K_2}, 0)$. The smoothing parameters are estimated by minimising a mean squared error criterion which can be shown to be equivalent to the Un-Biased Risk Estimator score, which in turn is equivalent to an approximate Akaike information criterion. We refer the reader to Marra & Radice (2011a, 2013) and Wood (2006) for more details.

3 Testing the hypothesis of exogeneity

The hypothesis of exogeneity is 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 we can say that there is exogeneity. On the contrary, $\rho \neq 0$ implies that there is a problem of endogeneity. The null hypothesis is $H_0 : \rho = 0$ and the alternative is $H_1 : \rho \neq 0$. Under H_0 the two equations are independent and the log-likelihood of the bivariate model becomes the sum of the log-likelihood functions of two univariate probit models. This implies that $\boldsymbol{\delta}_{H_0}^{\mathsf{T}} = \left(\boldsymbol{\delta}_{H_0,1}^{\mathsf{T}}, \boldsymbol{\delta}_{H_0,2}^{\mathsf{T}}, 0\right)$, where $\boldsymbol{\delta}_{H_0,1}$ and $\boldsymbol{\delta}_{H_0,2}$ are estimated by fitting equations (2) or (3) separately. Therefore, $\hat{\boldsymbol{\delta}}_{H_0,2}$ is a consistent estimator for $\boldsymbol{\delta}_2$. Under H_1 simultaneous modeling of equations (1) and (2) or equations (1) and (3) is required to obtain a consistent estimator for $\boldsymbol{\delta}_2$.

3.1 LM test

The LM test is a convenient tool for testing H_0 as it does not require parameter estimates under H_1. This means that simultaneous estimation is employed only if the hypothesis of exogeneity is rejected. As shown in Marra et al. (2014), the LM statistic for regression spline bivariate models is

$$LM = \mathbf{u}^{\mathsf{T}}_{\hat{\boldsymbol{\delta}}_{H_0}} \boldsymbol{\mathcal{I}}^{-1}_{\hat{\boldsymbol{\delta}}_{H_0}} \mathbf{u}_{\hat{\boldsymbol{\delta}}_{H_0}} \xrightarrow[H_0]{d} \chi^2_1,$$

where $\mathbf{u}_{\hat{\boldsymbol{\delta}}_{H_0}}$ is a penalized gradient vector given by $\mathbf{g}_{\hat{\boldsymbol{\delta}}_{H_0}} - \mathbf{S}_{\hat{\boldsymbol{\lambda}}_{H_0}} \hat{\boldsymbol{\delta}}_{H_0}$, with $\mathbf{g}_{\hat{\boldsymbol{\delta}}_{H_0}}$ being the gradient vector of (5) with respect to $\hat{\boldsymbol{\delta}}_{H_0}$, and $\boldsymbol{\mathcal{I}}_{\hat{\boldsymbol{\delta}}_{H_0}}$ is a penalized information matrix defined as $-\boldsymbol{\mathcal{H}}_{\hat{\boldsymbol{\delta}}_{H_0}} + \mathbf{S}_{\hat{\boldsymbol{\lambda}}_{H_0}}$, where $\boldsymbol{\mathcal{H}}_{\hat{\boldsymbol{\delta}}_{H_0}}$ is the Hessian of (5) with respect to $\hat{\boldsymbol{\delta}}_{H_0}$. Estimates for the λ_{H_0,vk_v} are obtained by estimating the two model equations separately.

Under the null, $\left\{\mathbf{g}_{\hat{\boldsymbol{\delta}}_{H_0}} - \mathbf{S}_{\hat{\boldsymbol{\lambda}}_{H_0}} \hat{\boldsymbol{\delta}}_{H_0}\right\}^{\mathsf{T}} = \left\{\mathbf{0}^{\mathsf{T}}, \mathbf{0}^{\mathsf{T}}, \partial \ell(\boldsymbol{\delta})/\partial \rho|_{\boldsymbol{\delta}=\hat{\boldsymbol{\delta}}_{H_0}}\right\}$. Note that the third component in $\mathbf{u}_{\hat{\boldsymbol{\delta}}_{H_0}}$ is unpenalized; this is because ρ is not penalized in estimation (see last element in $\mathbf{S}_{\boldsymbol{\lambda}}$ in Section 2.1). Furthermore, we use the results for penalized spline models employed, for instance, by Kauermann (2005). In particular, we consider the situation in which the spline bases approximating the smooth components are of a fixed high dimension. Since the unknown smooth functions may not have an exact representation as linear combinations of given basis functions, the unknown functions and parameters may not be asymptotically identified by their estimators as the sample size grows. However, in practice basis dimensions have to be fixed and assuming that these are of a high dimension, it is possible to argue heuristically that the approximation bias should be negligible compared to estimation variability (e.g., Kauermann, 2005). Other assumptions are $\mathbf{g}^0_{\boldsymbol{\delta}_{H_0}} = O_P(n^{1/2})$, $\mathbb{E}\boldsymbol{\mathcal{H}}^0_{\boldsymbol{\delta}_{H_0}} = O(n)$, $\boldsymbol{\mathcal{H}}^0_{\boldsymbol{\delta}_{H_0}} - \mathbb{E}\boldsymbol{\mathcal{H}}^0_{\boldsymbol{\delta}_{H_0}} = O_P(n^{1/2})$ and $\mathbf{S}_{\boldsymbol{\lambda}_{H_0}} = o\left(n^{1/2}\right)$. The first three conditions are the classic assumptions of $n^{1/2}$ asymptotics. The last condition can be formulated equivalently as $\lambda_{H_0,vk_v} = o\left(n^{1/2}\right)$ for $k_v = 1, \ldots, K_v$, $v = 1, 2$, assuming that the matrices $\mathbf{S}_{vk_v}$ are asymptotically bounded; this assumption is weak and in fact smoothing parameter estimates based on a mean squared error criterion are of order $O(1)$ (Kauermann, 2005). All this suggests that it is still possible to use the classic asymptotic result that, under the null, LM has a χ^2_1 limiting distribution as in Monfardini & Radice (2008).

3.2 W test

The W test is based on the simultaneous estimation of the two model equations. Because ρ is bounded in $[-1, 1]$, for convenience $\rho_* = \tanh^{-1}(\rho) = (1/2)\log\{(1+\rho)/(1-\rho)\}$ is used in optimization. Since the original null hypothesis can also be stated as $H_0 : \rho_* = 0$ and the alternative as $H_1 : \rho_* \neq 0$, the W test can be performed directly on ρ_*. That is,

$$W = \frac{\hat{\rho}_*^2}{\text{Var}\,(\hat{\rho}_*)} \xrightarrow[H_0]{d} \chi_1^2,$$

where $\hat{\rho}_*$ is the estimator of ρ_*, $\text{Var}\,(\hat{\rho}_*)$ is estimated using the diagonal element of $\boldsymbol{\mathcal{I}}^{-1}_{\hat{\boldsymbol{\delta}}_{H_1}}$ (the inverse of the penalized information matrix at $\hat{\boldsymbol{\delta}}_{H_1}$) corresponding to $\hat{\rho}_*$. The χ_1^2 limiting distribution of W follows from the same arguments discussed in the previous section.

3.3 G test

The G statistic has been proposed by Terrell (2002) and is based on the estimation of the parameters under H_0 and H_1. For simplicity, let $\boldsymbol{\delta}_0$ denote the true parameter vector and consider a model that does not involve the use of penalties in fitting. Then, under classic regularity conditions, $\partial\ell(\boldsymbol{\delta})/\partial\boldsymbol{\delta}|_{\boldsymbol{\delta}=\boldsymbol{\delta}_0} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\mathcal{I}})$ and $\hat{\boldsymbol{\delta}} \sim \mathcal{N}(\boldsymbol{\delta}_0, \boldsymbol{\mathcal{I}}^{-1})$ in the large sample limit, where $\boldsymbol{\mathcal{I}}$ is the information matrix (often estimated using $-\boldsymbol{\mathcal{H}}$). Let now $\mathbf{A}$ denote any square root of the information matrix, i.e. $\mathbf{A}^\mathsf{T}\mathbf{A} = \boldsymbol{\mathcal{I}}$, then $(\mathbf{A}^{-1})^\mathsf{T}\,\partial\ell(\boldsymbol{\delta})/\partial\boldsymbol{\delta}|_{\boldsymbol{\delta}=\boldsymbol{\delta}_0}$ and $\mathbf{A}(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}_0)$ have the same asymptotic distribution which is a multivariate normal with mean zero and variance given by the identity matrix. The inner product of these standardized vectors gives $\left[(\mathbf{A}^{-1})^\mathsf{T}\,\partial\ell(\boldsymbol{\delta})/\partial\boldsymbol{\delta}|_{\boldsymbol{\delta}=\boldsymbol{\delta}_0}\right]^\mathsf{T}\mathbf{A}(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}_0) = \left[\partial\ell(\boldsymbol{\delta})/\partial\boldsymbol{\delta}|_{\boldsymbol{\delta}=\boldsymbol{\delta}_0}\right]^\mathsf{T}(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}_0)$ which has a χ_p^2 limiting distribution, where p is the total number of model parameters (Terrell, 2002).

Recalling from Section 3.1 that $\left\{\mathbf{g}_{\hat{\boldsymbol{\delta}}_{H_0}} - \mathbf{S}_{\hat{\boldsymbol{\lambda}}_{H_0}}\hat{\boldsymbol{\delta}}_{H_0}\right\}^\mathsf{T} = \left\{\mathbf{0}^\mathsf{T}, \mathbf{0}^\mathsf{T}, \partial\ell(\boldsymbol{\delta})/\partial\rho|_{\boldsymbol{\delta}=\hat{\boldsymbol{\delta}}_{H_0}}\right\}$, the test reduces to

$$G = \left.\frac{\partial\ell(\boldsymbol{\delta})}{\partial\rho}\right|_{\boldsymbol{\delta}=\hat{\boldsymbol{\delta}}_{H_0}} \hat{\rho} \xrightarrow[H_0]{d} \chi_1^2.$$

The G test has a simple form and does not require the use of an information matrix. Terrell (2002) correctly pointed out that the G test "is not transparently non-negative, even though it must be so asymptotically"; his Theorem 2 shows that if the log-likelihood function is

concave and differentiable at $\boldsymbol{\delta}_0$ then $G \geq 0$.

3.4 LR test

The LR statistic is given by twice the difference of the model log-likelihoods under H_1 and H_0, that is

$$LR = 2\left\{\ell(\hat{\boldsymbol{\delta}}_{H_1}) - \ell(\hat{\boldsymbol{\delta}}_{H_0})\right\}.$$

As discussed in the preamble of Section 3, $\ell(\hat{\boldsymbol{\delta}}_{H_0})$ can be written as $\ell_{M_1}(\hat{\boldsymbol{\delta}}_{H_0,1}) + \ell_{M_2}(\hat{\boldsymbol{\delta}}_{H_0,2})$, where ℓ_{M_1} and ℓ_{M_2} denote the log-likelihoods of the two univariate equations. Thus,

$$LR = 2\left\{\ell(\hat{\boldsymbol{\delta}}_{H_1}) - \left[\ell_{M_1}(\hat{\boldsymbol{\delta}}_{H_0,1}) + \ell_{M_2}(\hat{\boldsymbol{\delta}}_{H_0,2})\right]\right\}.$$

To calculate $\ell(\hat{\boldsymbol{\delta}}_{H_1})$ the two equations are estimated simultaneously, whereas for $\ell(\hat{\boldsymbol{\delta}}_{H_0})$ the equations are estimated separately.

For bivariate probit models which do not involve the use of penalties in estimation the limiting distribution of LR is χ^2_1: the model under H_1 contains only one more parameter to estimate (i.e. ρ) as compared to the model under H_0. This result cannot be used for penalized regression spline bivariate models. This is because the number of degrees of freedom of the test should be calculated using the notion of estimated degrees of freedom (*edf*). The total *edf* is defined as the trace of the hat matrix implied by the fitted penalized model (Wood, 2006). The *edf* of a smooth function can be understood as follows. If the number of basis functions and parameters used to represent a smooth function is equal to 20 then the *edf* can be any real number in the range from 1 (meaning straight line estimate) to 20 (wiggly estimate). The total *edf* of a model is given by the sum of all the *edf* for the smooth functions plus the number of parametric terms (i.e., intercepts, correlation, binary predictors etc.). Therefore, $LR \underset{H_0}{\xrightarrow{d}} \chi^2_{edf_{H_1} - edf_{H_0}}$, where edf_{H_1} and edf_{H_0} are the total *edf* obtained after fitting the model under H_1 and H_0 (Wood, 2006). However, this approach is problematic as $edf_{H_1} - edf_{H_0}$ is likely not be an integer value despite it may be close to 1. So, for instance, if edf_{H_1} turns out to be 14.45 and edf_{H_0} is equal to 13.34 (because it happened for the model parameters under H_0 to be penalized slightly more than those under H_1),

then the number of degrees of freedom for LR is 1.11. As a simple and effective fix, we can use the fact that for any penalized regression spline model, with given *edf* for each smooth term, there is a very similar model based on pure regression splines, with similar degrees of freedom (*df*). So, for instance, if the *edf* of an estimated smooth curve is 5.43 then it is possible to obtain a similar estimated curve using 5 basis functions and corresponding unpenalized coefficients. The theoretical foundation of this result is given in Wood (2006, Section 4.10); see also Liu & Tu (2012) and Nummi et al. (2011) for applications of this idea in different contexts. For a model based on pure splines we have that $LR \xrightarrow[H_0]{d} \chi^2_{df_{H_1}-df_{H_0}}$, where df_{H_1} and df_{H_0} are the total number of parameters of the model under H_1 and H_0. This will ensure that, in our case, $LR \xrightarrow[H_0]{d} \chi^2_1$. The procedure for implementing the test is illustrated in the Appendix.

4 Simulation study

To assess and compare the empirical properties of the LM, W, G and LR tests, we conducted a Monte Carlo simulation study. All computations were performed in the `R` environment (R Development Core Team, 2015) using the package `SemiParBIVProbit` (Marra & Radice, 2015) and the code snippets reported in the Appendix. The simulation settings used below are from Marra et al. (2014).

4.1 Design of the experiments

The data generating process (DGP) for the recursive bivariate probit model was based on

$$\begin{aligned} y^*_{1i} &= 0.32 + 1.25m_{1i} + s_1(z_{1i}) - 0.75z_{2i} + \varepsilon_{1i} \\ y^*_{2i} &= 0.25 + y_{1i} - 0.75m_{1i} + s_2(z_{1i}) + \varepsilon_{2i} \end{aligned},$$

where the binary outcomes y_{1i} and y_{2i} were determined according to the rules described in Section 2, $s_1(z_{1i}) = -0.9\left[x^{2.5} + \exp\left\{-3(x-0.35)^2\right\}\right]$ and $s_2(z_{1i}) = -2\left\{0.25\exp(x) - x^3\right\}$. For the sample selection model the outcome equation was generated as

$$y^*_{2i} = 0.25 - 0.75m_{1i} + s_2(z_{1i}) + \varepsilon_{2i},$$

with binary outcome determined as described in Section 2. 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 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. Then, the columns of this matrix were transformed using `round()` and `pnorm()` to generate one binary and two continuous uniform covariates. Bivariate normal errors were generated using `rmvnorm()`. We also considered the situation in which z_{2i} does not enter the first equation of the DGP. This case is of some practical relevance because it may not sometimes be possible to impose ER. Each design was replicated 1000 times while the sample sizes considered were 1000 and 4000. The H_0 and H_1 rejection probabilities of each test were calculated as the proportions of rejections based on simulation replications.

A variety of simulation settings for generating y_{1i}^* and y_{2i}^* were considered. Specifically, we used several smooth function definitions, different sample sizes ($n = 500, 2000, 3000$), more smooth functions in the equations, and different coefficients for the parametric terms. Here, we do not describe the exact details and discuss the results obtained as the substantive conclusions did not change. These results are available upon request.

4.2 Monte Carlo results

4.2.1 H_0 rejection probability

The rejection frequencies under $H_0 : \rho = 0$ for the recursive and sample selection models are given in Table 1. We first discuss the case when ER is present. In general, the finite sample null rejection probability of the tests are similar and are close to their nominal values. Therefore, all tests can be used for testing the hypothesis of exogeneity when ER is available. As already discussed by Marra et al. (2014), the good performance of the LM test makes it appealing to use as it does not require estimating the two model equations simultaneously.

The results for the case of absence of ER concern the recursive model only. Recall from Section 2 that ER is crucial to identify the parameters of a sample selection model. This was confirmed by the high number of convergence failures experienced for the sample

			Recursive				Sample selection			
	$\alpha(\%)$	n	LM	W	G	LR	LM	W	G	LR
ER	1	1000	1.2	1.0	0.6	0.8	1.2	1.0	0.8	0.9
	5		5.3	4.6	4.8	5.0	5.9	5.3	5.4	5.7
	10		9.3	9.2	9.1	9.7	11.0	10.6	10.5	10.9
	1	4000	1.0	1.0	0.7	0.6	0.9	1.0	0.8	0.8
	5		5.4	5.1	5.1	5.1	5.4	5.2	5.2	5.4
	10		9.6	9.6	9.4	9.8	10.6	10.2	10.4	10.8
non-ER	1	1000	2.1	29.7	0.0	2.0	-	-	-	-
	5		3.4	49.4	0.0	7.1	-	-	-	-
	10		5.2	58.3	0.0	12.4	-	-	-	-
	1	4000	2.2	38.5	0.0	1.7	-	-	-	-
	5		4.2	53.5	0.0	6.6	-	-	-	-
	10		5.8	63.9	0.2	11.7	-	-	-	-

Table 1: Null rejection frequencies (in %) were obtained by applying the Lagrange multiplier (LM), Wald (W), gradient (G), likelihood ratio (LR) tests. The results refer to the recursive and sample selection model cases when an exclusion restriction is present (ER) and absent (non-ER). α and n denote the significance level and sample size. Results are not available for the sample selection non-ER case as an exclusion restriction is always required to identify the model parameters.

			Recursive				Sample selection			
	$\alpha(\%)$	n	LM	W	G	LR	LM	W	G	LR
ER	1	1000	2.8	9.2	0.7	2.4	3.6	0.7	0.6	2.5
	5		8.7	14.8	6.8	7.0	10.0	6.1	6.0	9.6
	10		15.0	19.8	14.0	12.9	15.2	12.3	11.8	14.6
	1	4000	1.8	5.7	2.2	1.2	2.1	1.3	1.1	1.6
	5		6.5	9.9	7.2	5.8	6.8	5.3	5.5	6.2
	10		12.3	15.8	12.9	12.2	12.2	11.1	11.6	11.6
non-ER	1	1000	2.4	12.6	0.0	5.0	-	-	-	-
	5		3.1	16.6	0.1	11.8	-	-	-	-
	10		4.8	21.4	0.5	18.5	-	-	-	-
	1	4000	1.0	13.6	0.0	4.4	-	-	-	-
	5		1.6	22.4	0.0	10.6	-	-	-	-
	10		2.2	27.8	0.3	14.5	-	-	-	-

Table 2: Null rejection frequencies (in %) were obtained by applying the Lagrange multiplier (LM), Wald (W), gradient (G), likelihood ratio (LR) tests. The results refer to the recursive and sample selection model cases when an exclusion restriction is present (ER) and absent (non-ER), and the error terms follow a gamma distribution with shape and scale parameters equal to 2. α and n denote the significance level and sample size.

selection case when ER was absent, which prevented us from calculating meaningful rejection frequency. The performance of W and G is the worst when ER is not present. Relatively to W and G, LR performs better although it tends to over-reject the null across all values of α. LM also over-rejects the null when $\alpha = 0.01$, whereas for the other two values of α it tends to be conservative as compared to LR. The reasonable performance of LR and LM is attractive if the hypothesis of exogeneity has to be tested in the absence of ER. This is because identification of a valid instrument is not always straightforward and in certain situations it may just not be feasible to find a suitable ER.

Following Monfardini & Radice (2008), we also explored the performance of the tests in the situation of distributional misspecification. This was achieved by generating uncorrelated gamma errors with shape and scale parameters equal to 2 using `rgamma()`. This case is of some relevance as the assumption of bivariate normality is often criticized. As shown in Table 2, the finite sample null rejection frequencies are worse than those obtained when the assumption of bivariate normality is not violated. Nevertheless, the null rejection frequencies are still reasonable for $n = 4000$. It is worth noting that, when ER does not hold, LM and LR still perform best relative to the other tests.

4.2.2 H_1 rejection probability

The H_1 rejection frequency of the LM, W, G and LR tests were calculated for several values of ρ, i.e. $\rho = \{0.1, 0, 2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$, when ER was present. Rejection frequency curves are presented in Figures 1 and 2 for the recursive and sample selection model cases. All curves are nearly identical except for $n = 1000$ where the LM test seems to have a marginal advantage. In all cases, the H_1 rejection frequency improves as ρ and n increase.

5 Applications

We illustrate the tests using two case studies in which the issue of endogeneity arises. The first concerns a study, conducted in USA, on the impact of private health insurance on health care utilization. Data are from the 2008 Medical Expenditure Panel Survey (MEPS;

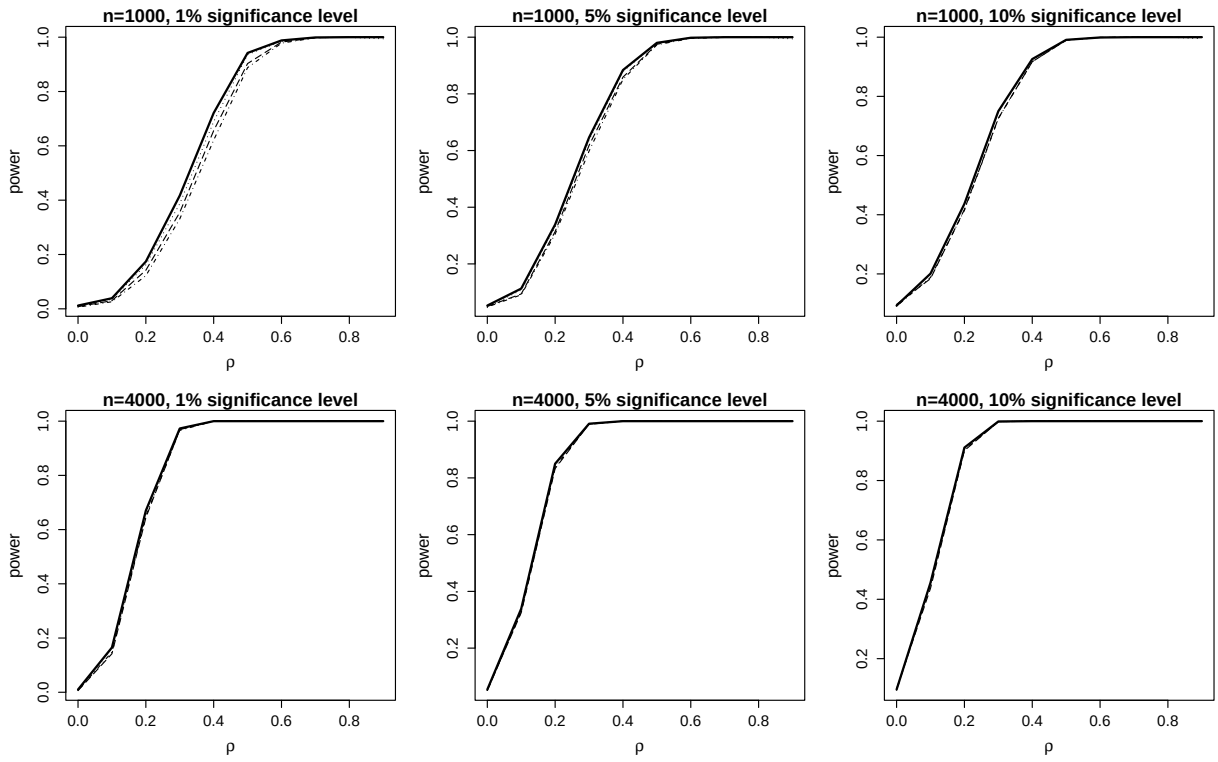

Figure 1: Rejection frequency curves for the LM (solid curve), W (dotted), G (dotdashed) and LR (longdashed) tests. The results refer to the recursive model case when ER is present. Note that in almost all cases the curves differ only minimally, hence they can hardly be distinguished.

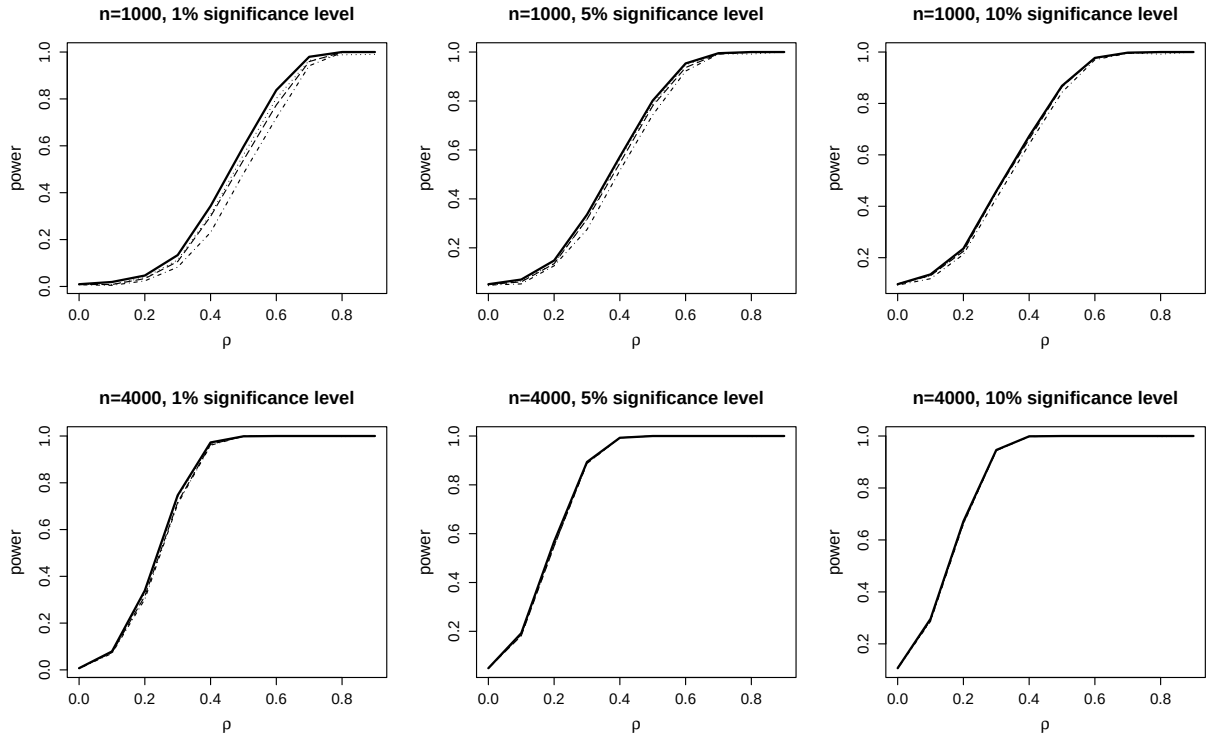

Figure 2: Rejection frequency curves for the LM (solid curve), W (dotted), G (dotdashed) and LR (longdashed) tests. The results refer to the sample selection model case when ER is present.

http://www.meps.ahrq.gov/) which provides a sample that includes information on demographics, individual health status, health care utilization and private health insurance coverage. Private health insurance status is equal to 1 if the individual had a private health insurance and 0 otherwise, and health care utilization is equal to 1 if the individual had at least one visit to hospital outpatient departments. As suggested by many scholars, estimation of such an effect can be biased by the possible endogeneity arising because unobserved confounders (e.g., allergy and risk aversiveness) are likely to influence both health service utilization and private insurance decision; full details can be found in Radice et al. (2015) and references therein. The second dataset concerns the 2007 Zambian Demographic and Health Survey (DHS; http://www.dhsprogram.com/) and the aim is to estimate the HIV prevalence for men. HIV status (1 = positive, 0 = negative) and consent to test (1 = consent, 0 = no consent) are recorded. Traditional methods assume that there are no unobserved confounders associated with both HIV status and consent to test (Perez et al., 2006). This assumption is unlikely to hold since a person's belief about his or her HIV status may be related to the actual status and hence influence the likelihood of consenting to a test through unobservables. For instance, individuals who already know or suspect that they are HIV positive may be less likely to consent. If this is the case then there will be a selection bias in population prevalence estimates based on an incorrect assumption of exogeneity (e.g., Bärnighausen et al., 2011).

5.1 Analysis of health care utilization data

Following previous work on the subject (e.g., Radice et al., 2015, and references therein), we specified a regression spline recursive bivariate probit model with main terms only. The description of the variables used in the model are reported in Table 3. Specifically, the linear predictors of the treatment (private) and outcome (visits.hosp) equations are

```
treat.eq <-   private ~ as.factor(health) + as.factor(race) + as.factor(region)
                      + limitation + gender + diabetes + hypertension
                      + hyperlipidemia + self + s(bmi) + s(income) + s(age)
                      + s(education)
```

Variable	Definition
Outcome	
`visits.hosp`	=1 at least one visit to hospital outpatient departments
Treatment	
`private`	=1 private health insurance
Observed confounders	
`age`	age in years
`gender`	=1 male
`race`	=1 white, =2 black, =3 native American, =4 others
`education`	years of education
`income`	income (000's)
`region`	=1 northeast, =2 midwest, =3 south, =4 west
`health`	=1 excellent, =2 very good, =3 good, =4 fair,=5 poor
`bmi`	body mass index
`diabetes`	=1 diabetic
`hypertension`	=1 hypertensive
`hyperlipidemia`	=1 hyperlipidemic
`limitation`	=1 health limits physical activity

Table 3: Description of outcome and treatment variables, and observed confounders for MEPS data.

```
out.eq <- visits.hosp ~ private + as.factor(health) + as.factor(race)
                      + as.factor(region) + limitation + gender + diabetes
                      + hypertension + hyperlipidemia + self + s(bmi) + s(income)
                      + s(age) + s(education)
```

where the smooth functions were represented using penalized thin plate regression splines with basis dimensions equal to 20 and penalties based on second order derivatives (Wood, 2006). The non-linear specification for `bmi`, `income`, `age` and `education` arises from the fact that these covariate embody productivity and life-cycle effects that are likely to affect $\mathbb{P}(\texttt{private} = 1)$ and $\mathbb{P}(\texttt{visits.hosp} = 1)$ non-linearly. In fact, in related studies, Holly et al. (1998) considered a model for health care utilization that contains linear and quadratic terms in `bmi`, `income`, `age` and `education`, whereas Marra & Radice (2011b) specified a model containing smooth functions of them. Note that for this case study it was not possible to identify a suitable ER.

The effect of `private` on `visits.hosp` (expressed in terms of average treatment effect (ATE)) may be biased by the possible presence of unobserved confounding. We employed the *LM*, *W*, *G* and *LR* tests for the null hypothesis of exogeneity; the p-values obtained

were 0.005, 0.004 ,0.243 and 0.001 respectively. Because in simulation, LR and LM showed reasonable empirical properties in the absence of ER, based on these tests' results we can assume that there is an issue of endogeneity. For completeness, we report the ATE (and confidence interval) obtained when fitting the regression spline recursive bivariate probit and the regression spline univariate probit for the outcome equation (i.e., the model not accounting for endogeneity): 5.34% (3.85%, 6.83%) and 4.29% (2.98%, 5.60%), respectively. The magnitude of the ATE found when accounting for endogeneity is higher, although the confidence intervals overlap. In this case study the direction of the bias appears to be downward. This result is counter-intuitive at first. If we assume that possible confounders are allergy and risk aversiveness, then an upward bias should be expected (individuals with a greater demand for medical care, say for reasons of poor health, are expected to have a greater demand for insurance). The explanation behind this apparent contradiction is that employer-provided insurance is generally limited to full-time workers and is positively related to worker income. The empirical evidence indicates that workers who are in poorer health are less likely to obtain employer-sponsored coverage (Buchmueller et al., 2005).

5.2 Analysis of HIV data

For the HIV dataset, using the variables in Table 4 and in line with the work by Bärnighausen et al. (2011), we specified the equations of the regression spline sample selection probit model as

```
sel.eq <- hivconsent ~ education + as.factor(wealth) + as.factor(region)
                     + as.factor(marital) + std + as.factor(agesex) + highhiv
                     + as.factor(partner) + condom + aidscare + knowsdiedofaids
                     + evertestedHIV + smoke + as.factor(religion)
                     + as.factor(ethnicity) + as.factor(language) + s(age)
                     + as.factor(interviewerID)

out.eq <-        hiv ~ education + as.factor(wealth) + as.factor(region)
                     + as.factor(marital) + std + as.factor(agesex) + highhiv
                     + as.factor(partner) + condom + aidscare + knowsdiedofaids
```

```
+ evertestedHIV + smoke + as.factor(religion)
+ as.factor(ethnicity) + as.factor(language) + s(age)
```

Variable `age` is expected to have a non-linear impact on `hiv` as well as `hivconsent`, whereas `education` 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 (`hivconsent`) also included `interviewerID` which served as ER. This was because, according to Bärnighausen et al. (2011) some interviewers are better than others at eliciting consent to HIV testing. For example, respect for the elderly is high in Zambia and people may find it more difficult to refuse testing from older interviewers than from younger ones. In addition, interviewers' personality traits, such as agreeableness or extraversion, may affect the respondents' likelihood of consenting to test.

The p-values obtained using the LM, W, G and LR tests were < 0.000, < 0.000, 0.025 and 0.003. In the presence of ER, the simulation results showed that all tests perform well. In this case study, the hypothesis of absence of endogenous sampling is rejected by all tests, hence indicating that non-random selection should be accounted for when estimating the HIV prevalence for men in Zambia. The HIV prevalence (and confidence interval) obtained using the regression spline sample selection bivariate probit was 21% (20%, 22%), which is significantly higher as compared to 12% (11%, 13%) calculated using the individuals who consented to test.

6 Discussion

Observational studies are likely to be affected by the presence of unobserved confounders, which may lead to inconsistent parameter estimates if not accounted for. For the case of binary responses, the regression spline recursive and sample selection bivariate probit models can be employed to deal with this problem. We discussed the likelihood ratio and gradient tests within this class of models and compared their empirical performance with that of the Lagrange multiplier and Wald tests through a Monte Carlo simulation study. The results obtained allowed us to derive some guidelines which may be important for empirical applications:

Variable	Definition
Outcome	
`hiv`	=1 positive HIV status
Selection	
`hivconsent`	=1 individual participated in HIV survey
Observed confounders	
`wealth`	=1 poorest, =2 poorer,=3 middle, =4 richer, =5 richest
`region`	=1 central, =2 copperbelt, =3 eastern, =4 luapula =5 lusaka, =6 northwestern, southern=7, =8 western
`marital`	=1 never married, =2 currently married =3 formerly married
`std`	=1 had sexually transmitted disease
`agesex`	=1 never had sex,=2 had sex < 16, =3 first sex > 15
`highhiv`	=1 not at high risk of HIV
`partner`	=1 none number of partners, =2 one partner =3 more than 2
`condom`	=1 no condom during last intercourse
`aidscare`	=1 would not care for HIV relative
`knowsdiedofaids`	=1 know someone died of AIDS
`evertestedHIV`	=1 previously HIV tested
`smoke`	=1 smoker
`religion`	=1 catholic, =2 protestant, =3 muslim, =4 other
`ethnicity`	defined by 99 categories
`language`	defined by 9 categories
`age`	age in years
Instrument	
`interviewerID`	interviewer identity

Table 4: Description of outcome and selection variables, instrument, and observed confounders for DHS data.

- when an exclusion restriction is present and under correct distributional specification, all tests perform well;

- when it is not possible to impose an exclusion restriction and under correct distributional specification, the tests which showed a reasonable performance are the likelihood ratio and Lagrange multiplier tests. This finding is of some interest as, in applied studies, identification of a valid instrument is not completely obvious and sometimes not possible;

- when the assumption of correct distributional specification does not hold and an exclusion restriction is present, the performance of all tests is worse than that observed under correct specification. Nevertheless, the likelihood ratio and Lagrange multiplier tests perform better relatively to the others;

- under the most challenging scenario (distributional misspecification and absence of exclusion restriction), all tests do not exhibit a satisfactory performance although the likelihood ratio and Lagrange multiplier tests seem to have a slight advantage over the others.

Because the performance of the tests worsens when the assumption of normality does not hold, tests based on different distributional assumptions may be considered. In the HIV example, respondents with a strong negative score on the test variable have a high risk of being HIV positive. This means that we would expect the presence of a non-Gaussian dependence that a linear measure of association, such as the correlation coefficient, would not be able to fully capture. In this direction, Radice et al. (2015) proposed regression spline bivariate probit models which allow for non-linear dependencies between the outcome equation and treatment or selection equation. This is achieved using Archimedean copula functions as well as their rotated versions. Future research will look into the feasibility of extending the tests discussed in this paper to copula bivariate models.

Acknowledgements

We thank the Editor and reviewer for their constructive criticism which helped to improve the presentation of the article considerably.

Appendix: Further simulation details

In `R`, the equations for the recursive bivariate probit model were

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

where the `s()` indicate the unknown smooth functions described in Section 2, while `y1`, `y2`, `m1`, `z1` and `z2` refer to $y_{1i}, y_{2i}, m_{1i}, z_{1i}$ and z_{2i} in Section 4. In the sample selection case, the outcome equation was given as

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

For the case in which z_{2i} did not enter the first equation of the DGP the equation was

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

P-values for the LM test were obtained using `LM.bpm()` from `SemiParBIVProbit`, that is

```
LM.bpm(list(eq1, eq2), data = dataSim, Model = ES),
```

where `dataSim` represents the data frame containing the variables associated with the two equations simulated as explained in the previous section, and `ES` was set to `"B"` or `"BSS"` depending on whether a recursive or sample selection bivariate probit model was considered. The simultaneous bivariate model which was needed in order to calculate the p-values for the other tests was fitted using

```
out <- SemiParBIVProbit(list(eq1, eq2), data = dataSim, Model = ES)
```

P-values for the W test were obtained as follows

```
athr2  <- coef(out)["theta.star"]^2
v      <- out$Vb
v.athr <- v[dim(v)[2], dim(v)[2]]
W      <- athr2/v.athr
pchisq(W, 1, lower.tail = FALSE)
```

where `athr2` and `v.athr` denote the estimates of ρ_* and $\text{Var}(\rho_*)$. P-values for the G test were obtained using `gt.bpm()` from `SemiParBIVProbit`, specifically `gt.bpm(out)`. Finally, p-values for LR in the recursive model case were obtained as follows

```
df11 <- round(out$edf1[1]) + 1
df12 <- round(out$edf1[2]) + 1
df21 <- round(out$edf2[1]) + 1
eq1FP <- y1 ~      m1 + s(z1, fx = TRUE, k = df11) + s(z2, fx = TRUE, k = df12)
eq2FP <- y2 ~ y1 + m1 + s(z1, fx = TRUE, k = df21)
outFP <- SemiParBIVProbit(list(eq1FP, eq2FP), fp = TRUE, data = dataSim, Model = ES)
log.sep <- logLik(outFP$gam1) + logLik(outFP$gam2)
log.sim <- logLik(outFP)
LR <- as.numeric(2*(log.sim-log.sep))
pchisq(LR, 1, lower.tail = FALSE)
```

where the `edf` are the estimated degrees of freedom of the smooth functions, and the `df` refer to the number of basis functions used for each smooth term when fitting the unpenalized bivariate probit model with splines (here achieved setting `fx = TRUE` and `fp = TRUE`). `gam1` and `gam2` refer to the fitted regression spline univariate probit models for the two model equations; these were obtained using `mgcv`. Because of identifiability constraints, one basis function (specifically, the constant flat basis function) is always dropped from each smooth term (Wood, 2006). Thus, each `df` needs to be increased by 1 to use the desired number of degrees of freedom. If `round(edf)` is equal to 1 then there is no need to use a smooth function and `s()` can be dropped from the variable involved. Of course, the expressions for `eq1FP` and `eq2FP` varied based on the DGP and model (recursive or sample selection) considered. We also considered constructing the LR test based on `df11`, `df12` and `df21`

chosen using the univariate fits from `gam1` and `gam2`. This did not lead to different results as, in the majority of the replicates, the numbers of basis functions selected for the smooth terms in the models were the same as those determined using the approach described above.

The smooth functions were represented using penalized thin plate regression splines with basis dimensions equal to 20 and penalties based on second-order derivatives (Wood, 2006, pp. 154–160). For the LR test, the basis dimension for each smooth function was equal to $df + 1$ and unpenalized thin plate regression splines were used.

References

Banasik, J. & Crook, J. (2007). Reject inference, augmentation, and sample selection. *European Journal of Operational Research*, 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.

Buchmueller, T. C., Grumbach, K., Kronick, R., & Kahn, J. G. (2005). Book review: The effect of health insurance on medical care utilization and implications for insurance expansion: A review of the literature. *Medical Care Research and Review*, 62, 3–30.

Chib, S. & Greenberg, E. (2007). Semiparametric modeling and estimation of instrumental variable models. *Journal of Computational and Graphical Statistics*, 16, 86–114.

Cuddeback, G., Wilson, E., Orme, J., & Combs-Orme, T. (2004). Detecting and statistically correcting sample selection bias. *Journal of Social Service Research*, 30, 19–33.

de Ven, W. V. & Praag, B. V. (1981). The demand for deductibles in private health insurance: a probit model with sample selection. *Journal of Econometrics*, 17, 229–252.

Eilers, P. H. C. & Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, 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. *Journal of the American Statistical Association*, 96, 883–894.

Greene, W. H. (2012). *Econometric Analysis*. Prentice Hall, New York.

Han, S. & Vytlacil, E. J. (2014). Identification in a generalization of bivariate probit models with endogenous regressors. `http://econ.sites.olt.ubc.ca/files/2013/12/pdf_paper_seminar_sukjin_han.pdf`.

Hastie, T. & Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society Series 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.

Holly, A., Gardiol, L., Domenighetti, G., & Brigitte, B. (1998). An econometric model of health care utilization and health insurance in switzerland. *European Economic Review*, 42(3-5), 513–522.

Kauermann, G. (2005). Penalized spline smoothing in multivariable survival models with varying coefficients. *Computational Statistics and Data Analysis*, 49, 169–186.

Kawatkar, A. A. & Nichol, M. B. (2009). Estimation of causal effects of physical activity on obesity by a recursive bivariate probit model. *Value in Health*, 12, A131–A132.

Latif, E. (2009). The impact of diabetes on employment in Canada. *Health Economics*, 18, 577–589.

Liu, H. & Tu, W. (2012). A semiparametric regression model for paired longitudinal outcomes with application in childhood blood pressure development. *Annals of Applied Statistics*, 6, 1861–1882.

Maddala, G. S. (1983). *Limited Dependent and Qualitative Variables in Econometrics*. Cambridge University Press, Cambridge.

Marra, G. & Radice, R. (2010). Penalised regression splines: theory and application to medical research. *Statistical Methods in Medical Research*, 19, 107–125.

Marra, G. & Radice, R. (2011a). Estimation of a semiparametric recursive bivariate probit model in the presence of endogeneity. *Canadian Journal of Statistics*, 39, 259–279.

Marra, G. & Radice, R. (2011b). A flexible instrumental variable approach. *Statistical Modelling*, 11, 581–603.

Marra, G. & Radice, R. (2013). A penalized likelihood estimation approach to semiparametric sample selection binary response modeling. *Electronic Journal of Statistics*, 7, 1432–1455.

Marra, G. & Radice, R. (2015). *SemiParBIVProbit: Semiparametric Bivariate Probit Modelling*. R package version 3.2-13.2.

Marra, G., Radice, R., & Missiroli, S. (2014). Testing the hypothesis of absence of unobserved confounding in semiparametric bivariate probit models. *Computational Statistics*, 29, 715–741.

Monfardini, C. & Radice, R. (2008). Testing exogeneity in the bivariate probit model: A monte carlo study. *Oxford Bulletin of Economics and Statistics*, 70, 271–282.

Montmarquette, C., Mahseredjiana, S., & Houle, R. (2001). The determinants of university dropouts: a bivariate probability model with sample selection. *Economics of Education Review*, 20, 475–484.

Nummi, T., Pan, J., Siren, T., & Liu, K. (2011). Testing for cubic smoothing splines under dependent data. *Biometrics*, 67, 871–875.

Perez, F., Zvandaziva, C., Engelsmann, B., & Dabis, F. (2006). Acceptability of routine hiv testing (opt-out) in antenatal services in two rural districts of zimbabwe. *Journal of Acquired Immune Deficiency Syndromes*, 41, 514–520.

R Development Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Radice, R., Marra, G., & Wojtys, M. (2015). Copula regression spline models for binary outcomes.

Radice, R., Marra, G., & Zanin, L. (2013). On the effect of obesity on employment in the presence of observed and unobserved confounding. *Statistica Neerlandica*, 67, 436–455.

Ruppert, D., Wand, M., & Carroll, R. (2003). *Semiparametric Regression*. Cambridge University Press, New York.

Terrell, G. (2002). The gradient statistic. *Computing Science and Statistics*, 34, 206–215.

Vargas, T., Ferrari, S., & Lemonte, A. (2013). Gradient statistic: Higher-order asymptotics and bartlett-type correction. *World Bank Economic Review*, 7, 43–61.

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

Wood, S. N. (2006). *Generalized Additive Models: An Introduction With R*. Chapman & Hall/CRC, London.