Electronic Journal of Statistics
Vol. 7 (2013) 1432–1455
ISSN: 1935-7524
DOI: 10.1214/13-EJS814

A penalized likelihood estimation approach to semiparametric sample selection binary response modeling

Giampiero Marra*

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

and

Rosalba Radice

Department of Economics, Mathematics and Statistics
Birkbeck, University of London
Malet Street, London WC1E 7HX, U.K.
e-mail: r.radice@bbk.ac.uk

Abstract: Sample selection models are employed when an outcome of interest is observed for a restricted non-randomly selected sample of the population. We consider the case in which the response is binary and continuous covariates have a nonlinear relationship to the outcome. We introduce two statistical methods for the estimation of two binary regression models involving semiparametric predictors in the presence of non-random sample selection. This is achieved using a multiple-stage procedure, and a newly developed simultaneous equation estimation scheme. Both approaches are based on the penalized likelihood estimation framework. The problems of identification and inference are also discussed. The empirical properties of the proposed approaches are studied through a simulation study. The methods are then illustrated using data from the American National Election Study where the aim is to quantify public support for school integration. If non-random sample selection is neglected then the predicted probability of giving, for instance, a supportive response may be biased, an issue that can be tackled using the proposed tools.

AMS 2000 subject classifications: Primary 46N30; secondary 97K80.
Keywords and phrases: Binary responses, bivariate probit, non-random sample selection, penalized regression spline.

Received September 2012.

Contents

*Giampiero Marra was supported by the Engineering and Physical Sciences Research Council, UK (grant EP/J006742/1).

1. Introduction

Sample selection techniques are employed when observations are not from a random sample of the population. For instance, in public surveys it may be the case that some individuals choose not to answer some specific questions because they feel that their opinion might paint them in an unfavorable light. This leaves us with a self-selected sample. So if the interest is in quantifying the relationship between various demographic and socio-economic characteristics and an outcome variable in the population as a whole, then using the responding subsample is likely to produce biased estimates [16, 9].

To fix ideas, let us consider a study of public opinion polls on school integration that uses data from an American survey. The main question was if respondents support government intervention to ensure that black and white children go to the same school. In particular, 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). Information on individual demographic and socio-economic characteristics was also recorded. If the respondents who opposed government involvement in school integration chose not to answer the question, because they felt their opinion might be perceived as socially unacceptable, then the sample of individuals who provided an opinion may have differed in systematic ways from the sample of non-respondents. To clarify this (often misunderstood) concept, let us characterize each individual by some observed and unobserved features or confounders. If the responding and nonresponding subsamples have similar characteristics, then the issue of non-random sample selection does not arise since the average (observed and unobserved) features of the responding sample are similar to those of the population. If the decision to answer is no longer random, because of differing characteristics between the responding and nonresponding individuals, then biased analyses are

expected. When the relationship between the decision to respond and outcome is only through observables, it is possible to correct for non-random sample selection by controlling for these variables in the outcome equation. However, in the presence of unobservables influencing the decision to answer and the outcome, controlling only for observables is clearly insufficient. That is, if some individuals are part of the responding subsample because of their unobserved features, then regardless of whether observables and unobservables are correlated in the overall population they will be in the selected sample [9]. This means that ignoring the potential correlation between the unobserved factors influencing the decision to answer and the outcome can lead to inconsistent estimates of the covariate impacts in the outcome equation. For other examples of non-random sample selection, see [1, 7, 30].

Statistical methods correcting for non-random selection have been developed. Many of these concern models where the response variable is Gaussian [5, 16, 11, 22, 25, 40]. There are also a number of works that go beyond Gaussian responses; these include models for skewed, count and ordinal data [35, 4, 36, 29]. We consider the case in which the response is binary. The procedures currently available to fit a sample selection binary response model are those presented in [8, 3, 11]. These involve the (separate or simultaneous) estimation of two binary regression models for the selection and outcome equations. The outcome equation is used to examine the substantive question of interest, whereas the selection equation is used to detect non-random selection and hence obtain consistent estimates of the covariate effects in the outcome equation. One potential drawback to the application of these techniques is the lack of flexibility in handling the possible presence of nonlinear covariate-response relationships. That is, because the functional shape between predictors and outcome is rarely known a priori, imposing a parametric structure may prevent the researcher from recognizing a strong covariate effect or, more generally, revealing interesting relationships [34, 42].

The contribution of this article is twofold, one methodological and the other practical. First, we extend the procedures discussed in [8, 3, 11] to incorporate semiparametric covariate effects. In particular, we present a multiple-stage estimation approach, and a penalized likelihood estimation framework for a simultaneous system of two binary equations. Both approaches allow for flexible functional dependence of the binary responses on continuous covariates. Second, we implement the methods discussed in this article in the `R` package `SemiParBIVProbit` [26]; this can be particularly attractive to practitioners who wish to fit such models. No other computational alternatives which consider a semiparametric sample selection binary response model are available in the literature. It may be argued that the model setup adopted here is fairly similar to that of [40] except that, in the current context, the outcome is binary. This suggests that the Bayesian estimation scheme introduced by [40] can be extended for fitting the model considered in this paper. We elected to follow a frequentist approach because it can especially appeal to researchers and practitioners already familiar with traditional frequentist techniques and has the advantage of being computationally fast. The empirical properties of the methods are studied

through a simulation study, and the methods then illustrated using the above-mentioned case study on public opinion polls on school integration.

2. Methods

In this section, we describe the model structure of a semiparametric binary response sample selection model, present two strategies for parameter estimation and discuss the problems of identification and inference.

2.1. The model

The model consists of a first selection equation and a second outcome equation determining the response. The selection equation, expressed using the latent variable representation, is given as

$$y_{1i}^{*} = \mathbf{x}_{1i}^{+}\boldsymbol{\theta}_1 + \sum_{k_1=1}^{K_1} f_{1k_1}(z_{1k_1 i}) + \varepsilon_{1i}, \quad i = 1, \ldots n, \tag{2.1}$$

where n denotes the sample size, and y_{1i}^{*} is a latent continuous variable which is related to its observable counterpart y_{1i} through the rule $1(y_{1i}^{*} > 0)$. The outcome equation is given as

$$y_{2i}^{*} = \mathbf{x}_{2i}^{+}\boldsymbol{\theta}_2 + \sum_{k_2=1}^{K_2} f_{2k_2}(z_{2k_2 i}) + \varepsilon_{2i}, \tag{2.2}$$

where

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

and y_{2i} is missing when $y_{1i} = 0$. In (2.1), $\mathbf{x}_{1i}^{+} = \left(1, x_{12i}^{+}, \ldots, x_{1P_1 i}^{+}\right)$ represents the i^{th} row vector of $\mathbf{x}_1^{+}$, the $n \times P_1$ model matrix for any parametric model components (i.e. intercept, binary and categorical predictors), with corresponding parameter vector $\boldsymbol{\theta}_1$. The $f_{1k_1}(z_{1k_1 i})$ are unknown smooth functions of the K_1 continuous covariates $z_{1k_1 i}$. These components are represented using the regression spline approach (see next section). Each smooth term may be multiplied by some predictor, yielding a 'varying coefficients' model [15], and smooth functions of two covariates may also be considered [42, pp. 154-167]. Similarly in (2.2), $\mathbf{x}_{2i}^{+}$ is the i^{th} row vector of the $n_{se} \times P_2$ model matrix $\mathbf{x}_2^{+}$, with coefficient vector $\boldsymbol{\theta}_2$, the $f_{2k_2}(z_{2k_2 i})$ are unknown smooth terms of the K_2 continuous regressors $z_{2k_2 i}$, and n_{se} denotes the size of the selected sample. For identification purposes, smooth terms are subject to constraints such as $\sum_i f_{vk_v}(z_{vk_v i}) = 0$, $v = 1, 2$, $k_v = 1, \ldots, K_v$. As in [40], we make the assumption that unobserved confounders have a linear impact on the responses, i.e. $(\varepsilon_{1i}, \varepsilon_{2i}) \sim \mathcal{N}([0, 0], [1, \rho, \rho, 1])$, where ρ is the correlation coefficient and the error variances are normalized to unity

since the parameters in the model can only be identified up to a scale coefficient.

It is important to stress that estimation of (2.2) alone when $\rho \neq 0$ will yield inconsistent parameter estimates. As explained in the previous section, intuitively, ignoring the correlation between the unobserved confounders influencing the decision to answer and the outcome will induce bias in the covariate impacts because of non-random sample selection on unobservables. This can be formally seen from the derivations in Section 2.2.

Going back to our example, y_{1i} and y_{2i} would correspond to the question of whether an individual had an opinion on the integration question and what that opinion was, respectively. Covariate vectors $\mathbf{x}_{1i}^{+}$ and $\mathbf{x}_{2i}^{+}$ would contain variables such as gender and region, and the $f_{1k_1}(z_{1k_1 i})$ and $f_{2k_2}(z_{2k_2 i})$ could be thought of as smooth nonlinear effects of covariates such as age and education in both the selection and outcome equations.

2.1.1. Smooth function representation

A popular and effective way of representing smooth functions of continuous covariates is the regression spline approach [10]. The basic idea is to approximate $f_k(z_{ki})$, where subscript v has been dropped to avoid clutter, by a linear combination of known spline basis functions, $b_{kj}(z_{ki})$, and regression parameters, β_{kj}. That is, $f_k(z_{ki}) = \sum_{j=1}^{J_k} \beta_{kj} b_{kj}(z_{ki}) = \mathbf{B}_k(z_{ki})\boldsymbol{\beta}_k$, where J_k is the number of spline bases and hence regression coefficients used to represent f_k, $\mathbf{B}_k(z_{ki})$ represents the i^{th} row vector of dimension J_k consisting of the basis functions evaluated at the observation z_{ki}, i.e. $\mathbf{B}_k(z_{ki}) = \{b_{k1}(z_{ki}), b_{k2}(z_{ki}), \ldots, b_{kJ}(z_{ki})\}$, and $\boldsymbol{\beta}_k$ is the corresponding parameter vector. Calculating $\mathbf{B}_k(z_{ki})$ for each i yields J_k curves encompassing different degrees of complexity which multiplied by some real valued parameter vector $\boldsymbol{\beta}_k$ and then summed give a curve estimate for $f_k(z_k)$. Basis functions are usually chosen to have convenient mathematical properties and good numerical stability. Possible choices are B-splines, cubic regression and thin plate regression splines (see [34] for a more detailed introduction). Based on the result above, equations (2.1) and (2.2) can written as

$$y_{1i}^{*} = \mathbf{x}_{1i}^{+}\boldsymbol{\theta}_1 + \mathbf{B}_{1i}\boldsymbol{\beta}_1 + \varepsilon_{1i} = \eta_{1i} + \varepsilon_{1i}, \ i = 1, \ldots, n, \tag{2.3}$$

and

$$y_{2i}^{*} = \mathbf{x}_{2i}^{+}\boldsymbol{\theta}_2 + \mathbf{B}_{2i}\boldsymbol{\beta}_2 + \varepsilon_{2i} = \eta_{2i} + \varepsilon_{2i}, \ i \in \{j : y_{1j} = 1\}, \tag{2.4}$$

where $\mathbf{B}_{vi} = \{\mathbf{B}_{v1}(z_{v1i}), \ldots, \mathbf{B}_{vK_v}(z_{vK_v i})\}$, $\boldsymbol{\beta}_v^{\mathsf{T}} = (\boldsymbol{\beta}_{v1}^{\mathsf{T}}, \ldots, \boldsymbol{\beta}_{vK_v}^{\mathsf{T}})$ and $\eta_{vi} = \mathbf{x}_{vi}^{+}\boldsymbol{\theta}_v + \mathbf{B}_{vi}\boldsymbol{\beta}_v$, for $v = 1, 2$.

2.2. Multiple-stage estimation approach

Non-random sample selection can be dealt with by using a device which [16] introduced for an analogous problem in binary-choice regression, and that [8, 3]

employed in the sample selection context for fully parametric probit models. Here, we extend this device to incorporate semiparametric covariate effects.

The population regression for (2.4) can be written as

$$\mathbb{E}(y_{2i}^{*}|\mathbf{x}_{2i}^{+},\mathbf{B}_{2i}) = \eta_{2i},$$

while the regression for the subsample of complete observations is

$$\mathbb{E}(y_{2i}^{*}|\mathbf{x}_{2i}^{+},\mathbf{B}_{2i},y_{1i}^{*}>0) = \eta_{2i} + \mathbb{E}(\varepsilon_{2i}|\mathbf{x}_{2i}^{+},\mathbf{B}_{2i},y_{1i}^{*}>0). \tag{2.5}$$

It follows that

$$\mathbb{E}(\varepsilon_{2i}|\mathbf{x}_{2i}^{+},\mathbf{B}_{2i},y_{1i}^{*}>0) = \rho\vartheta_i, \tag{2.6}$$

where $\vartheta_i = \phi(\eta_{1i})/\Phi(\eta_{1i})$ (typically called inverse Mills ratio), and ϕ and Φ are the density and distribution functions of a standardized normal. Regression equation (2.5) can be therefore written as

$$y_{2i}^{*} = \eta_{2i} + \rho\vartheta_i + \widetilde{\varepsilon}_{2i}, \tag{2.7}$$

where $\mathbb{E}(\widetilde{\varepsilon}_{2i}|y_{1i}^{*}>0) = 0$ and $\mathbb{E}(\widetilde{\varepsilon}_{2i}^{2}|y_{1i}^{*}>0) = \tau_i^2 = 1+\rho^2\vartheta_i(-\eta_{1i}-\vartheta_i)$ [16]. It is clear that the parameter estimates for the covariates in η_{2i} that are correlated with η_{1i} are inconsistent if $\rho \neq 0$. This can be thought of as arising from an ordinary specification error with the conditional mean (2.6) deleted as a covariate in the model. Including ϑ_i as an explanatory variable, as in equation (2.7), would rectify this situation. In practice we do not know ϑ_i, but it is possible to obtain a consistent estimate of it based on the estimated coefficients of selection equation (2.3). After dividing the components in the right hand side of (2.7) by τ_i, because $\mathbb{E}(\widetilde{\varepsilon}_{2i}^{2}|y_{1i}^{*}>0) = \tau_i^2$, and using the selected sample, we can obtain parameter estimates using

$$y_{2i}^{*} = (\mathbf{x}_{2i}^{+}/\tau_i)\boldsymbol{\theta}_2 + \mathbf{B}_{2i}^{*}\boldsymbol{\beta}_2 + \rho(\vartheta_i/\tau_i) + \bar{\varepsilon}_{2i}, \tag{2.8}$$

where $\mathbf{B}_{2i}^{*}$ includes the quantities corresponding to the spline bases for the smooth functions of the covariates z_{2k_2i} rescaled by τ_i, $\mathbb{E}(\bar{\varepsilon}_{2i}|y_{1i}^{*}>0) = 0$, $\mathbb{E}(\bar{\varepsilon}_{2i}^{2}|y_{1i}^{*}>0) = 1$, and ϑ_i and τ_i can be consistently estimated as described above.

The algorithm to fit model (2.8) can be summarised as follows:

step 1 Fit a probit model for equation (2.3) to obtain consistent estimates of $\widehat{\eta}_{1i}$ and hence $\widehat{\vartheta}_i$, for all i.

step 2 Using the selected sample, obtain a consistent estimate of ρ by fitting a linear probability model for equation (2.7), where ϑ_i is replaced with $\widehat{\vartheta}_i$ for all i.

step 3 Estimate $\widehat{\tau}_i$ via $\sqrt{1+\widehat{\rho}^2\widehat{\vartheta}_i(-\widehat{\eta}_{1i}-\widehat{\vartheta}_i)}$, where $\widehat{\vartheta}_i$, $\widehat{\eta}_{1i}$ and $\widehat{\rho}$ are obtained in steps 1 and 2.

step 4 Using the selected sample and after rescaling all components in the equation of interest by the $\widehat{\tau}_i$, fit a probit model for equation (2.8).

Remark 1. Equation (2.8) is fitted using probit regression even if the normality assumption of $\bar{\varepsilon}_{2i}$, which is necessary for consistency of $\widehat{\boldsymbol{\theta}}_2$ and $\widehat{\boldsymbol{\beta}}_2$, is clearly not met. However, as shown by [8] in a fully parametric context, this approach can deliver estimates which are close, if not as good as, to those obtained using a consistent estimator such as maximum likelihood.

Remark 2. Standard errors of the parameter estimates in (2.8) are not realistic in that, for example, they do not account for that additional source of sampling variability due to the estimation of ϑ_i and τ_i. In addition, the method can produce an estimate for ρ which is not in the range $[-1, 1]$.

Remark 3. The models used in the steps outlined above may be fitted using unpenalized parameter estimation procedures. However, because of the flexible model specification considered here, this is likely to result in smooth function estimates that are too wiggly to produce sensible results. This issue can be overcome by penalized estimation, where the objective function is augmented by a penalty term, such as $\sum_k \lambda_k \int f_k''(z_k)^2 dz_k$, measuring the (second-order, in this case) roughness of the smooth terms in the model. The λ_k are smoothing parameters controlling the trade-off between fit and smoothness. Since regression splines are linear in their model parameters, such a penalty can be expressed as a quadratic form in the generic parameter vector $\boldsymbol{\beta}$ (containing the coefficients of all smooth terms in the model), i.e. $\sum_k \lambda_k \int f_k''(z_k)^2 dz_k = \boldsymbol{\beta}^\mathsf{T} \left(\sum_k \lambda_k \mathbf{S}_k\right) \boldsymbol{\beta}$, where the $\mathbf{S}_k$ are positive semi-definite known square matrices. Depending on the model employed, parameters can be estimated by either minimization of a penalized least squares criterion or maximization of a penalized log-likelihood function. The λ_k can be selected via a prediction error or likelihood criterion [34, Chapter 8].

2.3. Bivariate probit estimation approach

In the current sample selection context, the data identify the three possible events $(y_{1i} = 1, y_{2i} = 1)$, $(y_{1i} = 1, y_{2i} = 0)$ and $(y_{1i} = 0)$, with probabilities

$$\begin{aligned}
\mathbb{P}(y_{1i} = 1, y_{2i} = 1 | \mathbf{x}_{1i}^+, \mathbf{B}_{1i}, \mathbf{x}_{2i}^+, \mathbf{B}_{2i}) &= p_{11i} = \Phi_2(\eta_{1i}, \eta_{2i}; \rho), \\
\mathbb{P}(y_{1i} = 1, y_{2i} = 0 | \mathbf{x}_{1i}^+, \mathbf{B}_{1i}, \mathbf{x}_{2i}^+, \mathbf{B}_{2i}) &= p_{10i} = \Phi(\eta_{1i}) - \Phi_2(\eta_{1i}, \eta_{2i}; \rho), \\
\mathbb{P}(y_{1i} = 0 | \mathbf{x}_{1i}^+, \mathbf{B}_{1i}) &= p_{0i} \ = \Phi(-\eta_{1i}),
\end{aligned}$$

where Φ_2 is the distribution function of a standardized bivariate normal with correlation ρ. The log-likelihood function is therefore

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

where $\boldsymbol{\delta}^\mathsf{T} = (\boldsymbol{\delta}_1^\mathsf{T}, \boldsymbol{\delta}_2^\mathsf{T}, \rho)$, and $\boldsymbol{\delta}_v^\mathsf{T} = (\boldsymbol{\theta}_v^\mathsf{T}, \boldsymbol{\beta}_v^\mathsf{T})$.

As pointed out in Remark 3 of the previous section, in a smoothing context it is necessary to penalize the regression spline coefficients to avoid exceedingly

wiggly smooth function estimates. Hence, the model is fitted by maximization of the penalized log-likelihood

$$\ell_p(\boldsymbol{\delta}) = \ell(\boldsymbol{\delta}) - \frac{1}{2}\boldsymbol{\beta}^{\mathsf{T}}\mathbf{S}_{\boldsymbol{\lambda}}\boldsymbol{\beta}, \tag{2.9}$$

where $\boldsymbol{\beta}^{\mathsf{T}} = (\boldsymbol{\beta}_1^{\mathsf{T}}, \boldsymbol{\beta}_2^{\mathsf{T}})$ and $\mathbf{S}_{\boldsymbol{\lambda}} = \sum_{v=1}^{2}\sum_{k_v=1}^{K_v}\lambda_{vk_v}\mathbf{S}_{vk_v}$. Note that because ρ is bounded in $[-1, 1]$, we use the common transform $\rho^* = \tanh^{-1}(\rho) = (1/2)\log\{(1+\rho)/(1-\rho)\}$ in optimization. Given values for the λ_{vk_v}, we seek to maximize (2.9). This is achieved by using a trust region algorithm [31, Section 4.2] which is based on

$$\widehat{\boldsymbol{\delta}}^{[a+1]} = \widehat{\boldsymbol{\delta}}^{[a]} + (\boldsymbol{\mathcal{I}}^{[a]} + \mathbf{S}^*_{\boldsymbol{\lambda}})^{-1}(\mathbf{g}^{[a]} - \mathbf{S}^*_{\boldsymbol{\lambda}}\widehat{\boldsymbol{\delta}}^{[a]}), \tag{2.10}$$

where a is the iteration index and $\mathbf{S}^*_{\boldsymbol{\lambda}}$ an overall block-diagonal penalty matrix made up of $\lambda_{vk_v}\mathbf{S}_{vk_v}$ and $\mathbf{0}$ components. The gradient vector $\mathbf{g}$ is defined by two subvectors $\mathbf{g}_1 = \partial\ell(\boldsymbol{\delta})/\partial\boldsymbol{\delta}_1$ and $\mathbf{g}_2 = \partial\ell(\boldsymbol{\delta})/\partial\boldsymbol{\delta}_2$, and a scalar $g_3 = \partial\ell(\boldsymbol{\delta})/\partial\rho^*$, while the Fisher information matrix has a 3×3 matrix block structure with $(r,h)^{th}$ element $\boldsymbol{\mathcal{I}}_{r,h} = -\mathbb{E}\left[\partial^2\ell(\boldsymbol{\delta})/\partial\boldsymbol{\delta}_r\partial\boldsymbol{\delta}_h^{\mathsf{T}}\right]$, $r,h = 1,\ldots,3$, where $\boldsymbol{\delta}_3 = \rho^*$. The use of a trust region algorithm proved to be faster and more reliable than the standard approaches adopted in the literature to estimate likelihood-based models, with occasional convergence failure for small values of n and n_{se}. In (2.10), the smoothing parameters are fixed at some values. This is because joint estimation of $\boldsymbol{\delta}$ and $\boldsymbol{\lambda} = (\lambda_{1k_1},\ldots,\lambda_{1K_1},\lambda_{2k_2},\ldots,\lambda_{2K_2})$ via maximization of (2.9) would clearly lead to overfitting since the highest value for $\ell_p(\boldsymbol{\delta})$ would be obtained when $\boldsymbol{\lambda} = \mathbf{0}$. Hence the need to estimate $\boldsymbol{\lambda}$ using an appropriate criterion.

2.3.1. Smoothing parameter selection

Smoothing parameter selection can be achieved by direct grid search optimization of a prediction error criterion, for example. However, if the model has more than one smooth term per equation, then this can become computationally burdensome, hence making the model building process difficult in most applied contexts. There are a number of techniques for automatic multiple smoothing parameter selection within the penalized likelihood framework. Without claim of exhaustiveness, these include the performance-oriented iteration method originally proposed by [13] and mixed model approach to penalized regression spline estimation [34]. The former applies the generalized cross validation or unbiased risk estimator [UBRE; 6] to each working linear model of the penalized iteratively re-weighted least squares (P-IRLS) scheme used to fit the model. The latter consists of viewing the λ_{vk_v} as variance components so that they can be estimated, e.g., by restricted maximum likelihood. Here, we adapt the approach by [13] to the current context.

Given a parameter vector value for $\boldsymbol{\lambda}$, iterative equation (2.10) can be written in P-IRLS form

$$\|\sqrt{\mathbf{W}}^{[a]}(\mathbf{z}^{[a]} - \mathbf{X}\boldsymbol{\delta})\|^2 + \boldsymbol{\delta}^{\mathsf{T}}\mathbf{S}^*_{\boldsymbol{\lambda}}\boldsymbol{\delta}, \tag{2.11}$$

where $\sqrt{\mathbf{W}}$ is a weight non-diagonal matrix square root, $\mathbf{z}_i$ is the 3-dimensional vector $\mathbf{z}_i = \mathbf{X}_i\boldsymbol{\delta}^{[a]} + \mathbf{W}_i^{-1}\mathbf{d}_i$, $\mathbf{d}_i = \{\partial\ell(\boldsymbol{\delta})_i/\partial\eta_{1i}, \partial\ell(\boldsymbol{\delta})_i/\partial\eta_{2i}, \partial\ell(\boldsymbol{\delta})_i/\partial\eta_{3i}\}$, $\eta_{3i} = \rho^*$, $\mathbf{W}_i$ is the 3×3 matrix with $(r,h)^{th}$ element $(\mathbf{W}_i)_{rh} = -\mathbb{E}\left[\partial^2\ell(\boldsymbol{\delta})_i/\partial\eta_{ri}\partial\eta_{hi}\right]$, $r,h = 1,\ldots,3$ and $\mathbf{X}_i = \text{diag}\left\{(\mathbf{x}_{1i}^+, \mathbf{B}_{1i}), (\mathbf{x}_{2i}^+, \mathbf{B}_{2i}), 1\right\}$. The superscript $[a]$ has been suppressed from $\mathbf{d}_i$, $\mathbf{z}_i$, and $\mathbf{W}_i$, and is omitted from the quantities shown below, to avoid clutter.

Vector $\boldsymbol{\lambda}$ should be selected so that the estimated smooth functions are as close as possible to the true functions. In the current context, this is achieved using the approximate UBRE. Specifically, $\widehat{\boldsymbol{\lambda}}$ is the solution to the problem

$$\text{minimize} \quad \mathcal{V}_u^w(\boldsymbol{\lambda}) = \frac{1}{n_*}\|\sqrt{\mathbf{W}}(\mathbf{z} - \mathbf{X}\boldsymbol{\delta})\|^2 - 1 + \frac{2}{n_*}\text{tr}(\mathbf{A}_{\boldsymbol{\lambda}}) \quad \text{w.r.t.} \quad \boldsymbol{\lambda}, \quad (2.12)$$

where the working linear model quantities are constructed for a given estimate of $\boldsymbol{\delta}$, $n_* = 3n$, $\mathbf{A}_{\boldsymbol{\lambda}} = \mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{X} + \mathbf{S}_{\boldsymbol{\lambda}}^*)^{-1}\mathbf{X}^\mathsf{T}\mathbf{W}$ is the hat matrix, and $\text{tr}(\mathbf{A}_{\boldsymbol{\lambda}})$ represents the estimated degrees of freedom of the penalized model. For each working linear model of a trust region iteration, $\mathcal{V}_u^w(\boldsymbol{\lambda})$ is minimized with respect to $\boldsymbol{\lambda}$. The two steps, one for $\boldsymbol{\delta}$ the other for $\boldsymbol{\lambda}$, are iterated until convergence. This approach is implemented employing the approach by [41], which is based on Newton's method and can evaluate score (2.12) and their derivatives in a way that is both computationally efficient and stable. Generally speaking, this is achieved using $\sqrt{\mathbf{W}}\mathbf{X} = \mathbf{QR}$, obtained by pivoted QR decomposition, where $\mathbf{Q}$ and $\mathbf{R}$ are defined in the usual manner, and the singular value decomposition $\begin{bmatrix} \mathbf{R} \\ \mathbf{L} \end{bmatrix} = \mathbf{UDV}^\mathsf{T}$, where $\mathbf{L}$ is any matrix square root of $\mathbf{S}_{\boldsymbol{\lambda}}^*$ such that $\mathbf{L}^\mathsf{T}\mathbf{L} = \mathbf{S}_{\boldsymbol{\lambda}}^*$, the columns of $\mathbf{U}$ are those of an orthogonal matrix, $\mathbf{V}$ is an orthogonal matrix, and $\mathbf{D}$ is a diagonal matrix of singular values which are useful to detect numerical rank deficiency of the fitting problem. Based on this, evaluation of $\text{tr}(\mathbf{A}_{\boldsymbol{\lambda}})$ for new trial values of the smoothing parameters can be made relatively cheap, and the derivatives of $\mathcal{V}_u^w(\boldsymbol{\lambda})$ with respect to $\boldsymbol{\lambda}$ can be stably and efficiently evaluated. Note that minimization of the score is with respect to $\boldsymbol{\lambda}^* = \log(\boldsymbol{\lambda})$ since the smoothing parameters must be positive. See [41] for further details.

Remark 1. In the context of simultaneous equation estimation methods, the use of the Fisher information matrix is recommended because the $\mathbf{W}_i$ are positive-definite over a larger region of the parameter space as compared to those obtained by using the observed information. This is crucial given that $\sqrt{\mathbf{W}}$ and $\mathbf{W}^{-1}$ (via $\mathbf{z}$), obtained by eigen-decomposition, are needed in (2.12).

Remark 2. Because $\mathbf{W}$ is a non-diagonal matrix of dimension $n_* \times n_*$, computation can quickly become prohibitive, even for small sample sizes. To calculate $\mathbf{W}^{-1}\mathbf{d}$, $\sqrt{\mathbf{W}}\mathbf{z}$ and $\sqrt{\mathbf{W}}\mathbf{X}$ so that the computational load and storage demand of the algorithm is kept as low as possible, the band structure of $\mathbf{W}$ is exploited. Hence, the working linear model in (2.12) is formed in $O(n_*(m+2))$ rather than $O(n_*^2(m+2))$ operations, where m is the number of columns of $\mathbf{X}$.

Remark 3. As opposed to the multiple-stage approach discussed in Section 2.2, simultaneous estimation of all model parameters via the bivariate probit scheme introduced here does not rely on approximations and does not require the use

of quantities estimated in preliminary steps. Therefore, the procedure can yield consistent estimates for $\boldsymbol{\delta}$. Moreover, correction of standard errors is not in principle required since no inverse Mills ratio is used and all parameters in $\boldsymbol{\delta}$ are estimated jointly. This convenience comes at expense of computational cost and stability. However, these can be dealt with by using the approach described in last two sections, and supplying the multiple-stage estimates as starting values in the bivariate probit estimation scheme.

2.4. Identification

Under correct model specification, the parameters of the approach described in Section 2.2 are formally identified even if $\left(\mathbf{x}_{1i}^{+}, \mathbf{B}_{1i}\right) = \left(\mathbf{x}_{2i}^{+}, \mathbf{B}_{2i}\right)$. This is because of the nonlinearity of the inverse Mills ratio; see, e.g., [32]. However, in applications, this typically results in substantial collinearity between $\widehat{\vartheta}_i$ and the other covariates in the outcome equation, especially when the variation in $\widehat{\eta}_{1i}$ is such that the nonlinearity of the inverse Mills ratio does not play a major role. This collinearity can lead to large standard errors and instability in estimation. The parameters of the method introduced in Section 2.3 are also formally identified, but with the advantage of not having the limitations deriving from the use of the inverse Mills ratio. However, the likelihood functions of sample selection models may be affected by local maxima especially in the case of highly correlated error terms [20].

In practice, empirical identification is achieved if the exclusion restriction (ER) on the covariates in the two equations holds [37]. That is, the regressors in the selection equation should contain at least one or more regressors not included in the outcome equation. Such predictors can be regarded as instrumental variables, which, in this context, induce variation in the selection equation, do not directly affect the outcome, and are independent of $(\varepsilon_{1i}, \varepsilon_{2i})$ given the covariates [23]. In the data analysis reported in Section 4, ER was achieved by including in the selection equation the binary variable indicating whether the respondent was persuaded to participate in the survey.

2.5. Inference

The methods described in Section 2 rely on penalized estimation. Within this framework, inferential theory is not standard because of the presence of smoothing penalties which undermines the usefulness of classic frequentist results for practical modeling; see, e.g., [42]. Solutions to this problem have been proposed. In this section, we show how to construct pointwise confidence intervals for the components of a semiparametric sample selection model adapting some of the results available in the literature.

The well known Bayesian 'confidence' intervals originally proposed by [39] in the univariate spline model context are typically used to represent the uncertainty of smooth functions [14, 34, 42]. An interesting feature of these intervals is

that they have close to nominal 'across-the-function' *frequentist* coverage probabilities [27]. To better understand this point, let us consider a generic smooth component $f(z_i)$. Intervals can be constructed seeking some constants C_i and A, such that

$$ACP = \frac{1}{n}\mathbb{E}\left\{\sum_i \mathbb{I}(|\widehat{f}(z_i) - f(z_i)| \leq q_{\alpha/2}A/\sqrt{C_i})\right\} = 1-\alpha, \tag{2.13}$$

where ACP denotes average coverage probability, $\mathbb{I}$ is an indicator function, α is a constant between 0 and 1, and $q_{\alpha/2}$ is the $\alpha/2$ critical point from a standard normal distribution. Defining $b(z) = \mathbb{E}\{\widehat{f}(z)\} - f(z)$ and $v(z) = \widehat{f}(z) - \mathbb{E}\{\widehat{f}(z)\}$, so that $\widehat{f} - f = b + v$, and I to be a random variable uniformly distributed on $\{1, 2, \ldots, n\}$, we have that $ACP = \Pr\left(|B + V| \leq q_{\alpha/2}A\right)$, where $B = \sqrt{C_I}b(z_I)$ and $V = \sqrt{C_I}v(z_I)$. At this point, it is necessary to find the distribution of $B + V$ and values for the C_i and A so that requirement (2.13) is met. As shown in [27], in the context of non-Gaussian response models involving several smooth components, such a requirement is approximately met when confidence intervals for the smooth components are constructed using

$$\boldsymbol{\delta}|\mathbf{y} \dot{\sim} \mathcal{N}(\widehat{\boldsymbol{\delta}}, \mathbf{V}_{\boldsymbol{\delta}}), \tag{2.14}$$

where, for the approach described in Section 2.3, $\mathbf{y}$ refers to the response vectors, $\widehat{\boldsymbol{\delta}}$ is the estimate of $\boldsymbol{\delta}$ and $\mathbf{V}_{\boldsymbol{\delta}} = (\boldsymbol{\mathcal{I}} + \mathbf{S}^*_{\boldsymbol{\lambda}})^{-1}$ is the inverse of the penalized Fisher information matrix obtained at convergence of the algorithm used to fit the model. Note that the same distributional result can be used for the models described in Section 2.2. Given (2.14), confidence intervals for linear and nonlinear functions of the model parameters can be easily obtained. For any parametric model components, using (2.14) is equivalent to using classic likelihood results since such model terms are not penalized. It is important to stress that there is no contradiction in fitting the sample selection model via penalized log-likelihood estimation and then constructing confidence intervals using a Bayesian result, and such an approach has been employed many times in the literature; see, e.g., [14, 23, 42].

Result (2.14) should produce intervals with good coverage probabilities for the model components when using the method described in Section 2.3. However, this is not likely to be true for the approach detailed in Section 2.2, for the reasons given in Remark 2. As a solution, posterior simulation can be employed [42]. Specifically, we propose to adjust the intervals as follows:

- Let the parameter vector and covariance Bayesian matrix estimated in step 1 be $\widehat{\boldsymbol{\delta}}_1$ and $\widehat{\mathbf{V}}_{\boldsymbol{\delta}_1}$. Draw N_s random vectors from $\mathcal{N}(\widehat{\boldsymbol{\delta}}_1, \widehat{\mathbf{V}}_{\boldsymbol{\delta}_1})$ and then calculate the corresponding N_s values $\widehat{\eta}^*_{1i}$ and $\widehat{\vartheta}^*_i$, for all i.
- Fit N_s step 2 models to obtain $\widehat{\boldsymbol{\delta}}^{ols}_{2,1}, \ldots, \widehat{\boldsymbol{\delta}}^{ols}_{2,N_s}$ and $\widehat{\mathbf{V}}^{ols}_{\boldsymbol{\delta}_{2,1}}, \ldots, \widehat{\mathbf{V}}^{ols}_{\boldsymbol{\delta}_{2,N_s}}$. For each parameter vector and covariance matrix combination, draw N_s random vectors from the corresponding Gaussian distribution and then calculate the N_s^2 values $\widehat{\rho}^*$.

- Calculate the N_s^2 values $\widehat{\tau}_i^*$, for all i, using $\widehat{\vartheta}_i^*$, $\widehat{\eta}_{1i}^*$ and $\widehat{\rho}^*$.
- Fit N_s^2 step 4 models, using each of the $\widehat{\vartheta}_i^*$ and $\widehat{\tau}_i^*$ combination, to obtain $\widehat{\boldsymbol{\delta}}_{2,1}, \ldots, \widehat{\boldsymbol{\delta}}_{2,N_s^2}$ and $\widehat{\mathbf{V}}_{\boldsymbol{\delta}_{2,1}}, \ldots, \widehat{\mathbf{V}}_{\boldsymbol{\delta}_{2,N_s^2}}$. For each parameter vector and covariance matrix combination, draw N_d random vectors from the corresponding Gaussian distribution so to obtain $N_s^2 \times N_d$ random draws from which approximate intervals for the component functions of model (2.8) can be constructed.

This procedure can account for the extra source of variability introduced via the quantities calculated in steps 1–3 of the multiple-stage estimation approach. As in [24], simulation experience suggested that small values for N_s and N_d, say 20 and 100, will be tolerable in practice.

3. Simulation study

To gain insights into the effectiveness of the estimation approaches detailed in the previous sections, a Monte Carlo simulation study was conducted. All computations were performed in the `R` environment [33] using the package `SemiParBIVProbit` which implements the ideas discussed in this article [26].

3.1. Design and model fitting details

The sampling experiments were based on the model

$$\begin{aligned} y_{1i}^* &= \theta_{11} + \theta_{12} x_i^+ + f_{11}(z_{1i}) + f_{12}(z_{2i}) + \varepsilon_{1i} \\ y_{2i}^* &= \theta_{21} + \theta_{22} x_i^+ + f_{21}(z_{1i}) + \varepsilon_{2i} \end{aligned},$$

where the binary outcomes y_{1i} and y_{2i} were determined according to the rules described in Section 2.1. The test functions are displayed in Figure 1 and are given as $f_{11}(z_{1i}) = -0.7\left\{4z_{1i} + 2.5z_{1i}^2 + 0.7\sin(5z_{1i}) + \cos(7.5z_{1i})\right\}$, $f_{12}(z_{2i}) = -0.4\left\{-0.3 - 1.6z_{2i} + \sin(5z_{2i})\right\}$ and $f_{21}(z_{1i}) = 0.6\left\{\exp(z_{1i}) + \sin(2.9z_{1i})\right\}$. Parameter vector $(\theta_{12}, \theta_{22})$ was set to $(2.5, -1.5)$. To generate binary values for y_{1i} so that approximately 50% of the total number of observations were selected to fit the outcome equation, and values for y_{2i} which appeared in approximately identical numbers, $(\theta_{11}, \theta_{21})$ was set to $(0.58, -0.68)$. Predictors x_i^+, z_{1i} and z_{2i} were generated as three uniform correlated variables on $(0, 1)$. This was achieved using `rmvnorm()`, in the package `mvtnorm`, drawing standardized multivariate random variables with correlation 0.5 and then applying `pnorm()` [23]. Variable x_i^+ was eventually dichotomized using `round()`. Standardized bivariate normal errors with correlations $\rho = (\pm 0.1, \pm 0.5, \pm 0.9)$ were considered, and sample sizes set to 500, 1000 and 3000. In a full factorial design fashion, 1000 replications of each combination of parameter settings were obtained.

The smooth components were represented using penalized thin plate regression splines with basis dimensions equal to 10 and penalties based on second-order derivatives [42]. In cross-sectional studies, 10 bases typically suffice to

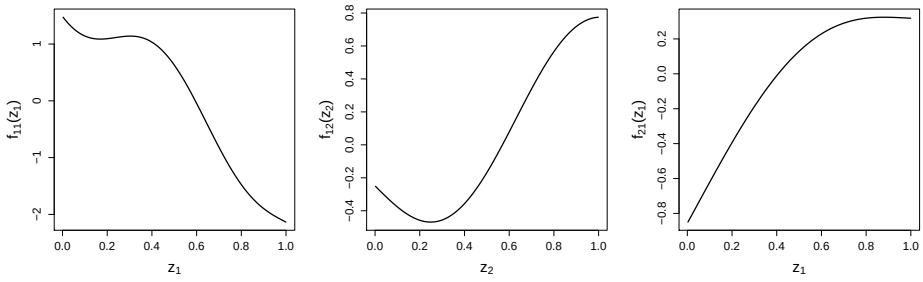

FIG 1. *The test functions used in the simulation studies.*

represent reasonably well smooth functions, although sensitivity analysis using fewer or more spline bases is advisable in applied work. Models were also fitted neglecting non-random sample selection, i.e. simply fitting equation (2.4) on the selected sample (henceforth, this will be referred to as naive approach); since this can not account for sample selection, biased parameter estimates are expected. We decided to report the naive results because they represent a benchmark for evaluating more realistic models as well as highlight the substantial detrimental effects that the neglect of non-random sample selection may have on the parameter estimates.

3.2. Results

In this section, we only show a subset of results; these are representative of all empirical findings. Since the parameters of the selection equation are not in principle affected by bias, we focus on the estimation results for the outcome equation.

Figures 2, 3 and 4 present the boxplots of the estimates for θ_{22}, ρ, and the empirical root mean squared errors (RMSE) of $\widehat{f}_{21}(z_1)$ when employing the naive, multiple-stage and bivariate probit estimation approaches, ER holds and approximately 50% of the total number of observations are available to fit the outcome equation. Figure 5 shows the estimated smooth functions for $f_{21}(z_1)$ averaged over the simulation runs. As in [40], based on the estimates for 200 fixed covariate values, $\text{RMSE}(\widehat{f}_{21})$ was calculated as $\sqrt{\sum_{b=1}^{200}\{\widehat{f}_{21}(z_{1b}) - f_{21}(z_{1b})\}^2}$.

The results can be summarized as follows:

- Figure 2 shows that at all sample sizes and $\rho = (0.5, 0.9)$ the sample selection estimators outperform the naive approach in terms of bias, and that bivariate probit is much more accurate than multiple-stage. However, for $n = (500, 1000)$ the introduced estimators are less precise than naive, especially when the sample selection issue is negligible.
- Figure 3 suggests overall that the multiple-stage estimates for ρ are systematically biased as compared to those of bivariate probit. When $\rho = 0.9$ the bias in the multiple-stage estimates worsens as n increases. This is

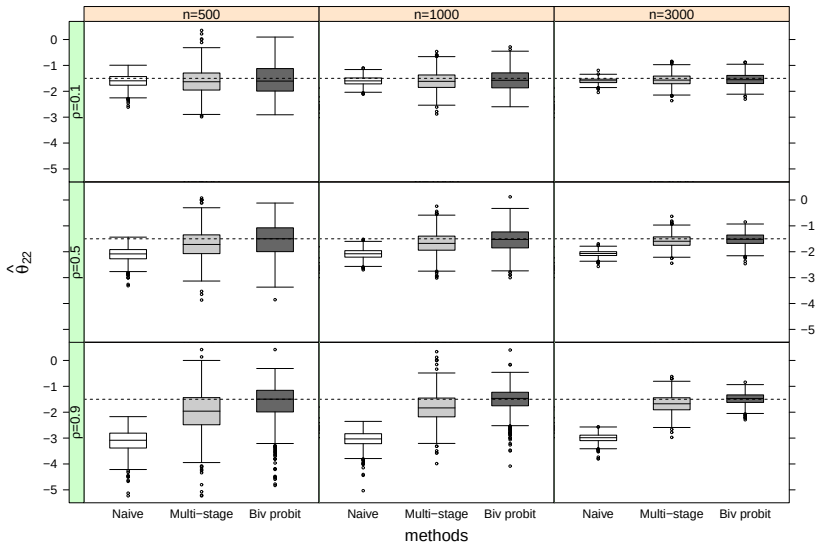

FIG 2. *Boxplots of the parameter estimates for θ_{22} based on 1000 replications when employing the naive, multiple-stage and bivariate probit estimation approaches and approximately 50% of the total number of observations are available to fit the outcome equation. The true value (dashed lines) is -1.5. ρ and n denote the correlation between the errors of the selection and outcome equations, and the sample size. See Section 3.1 for further details.*

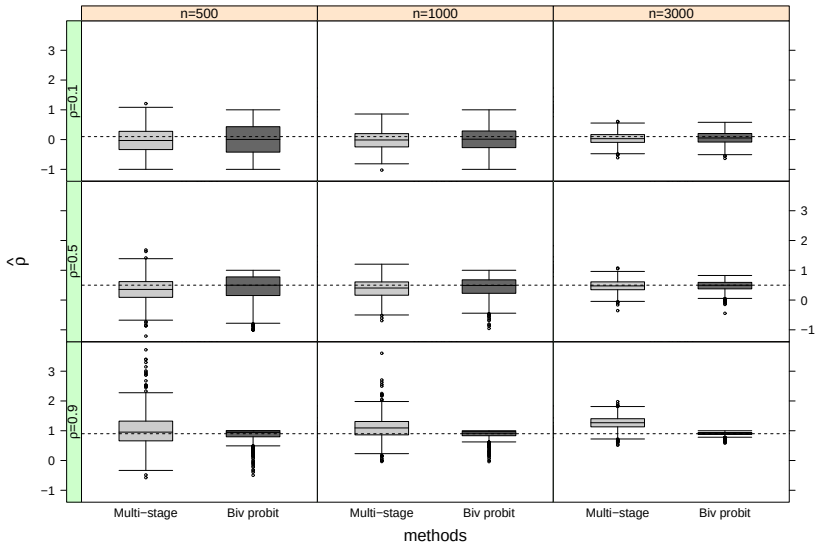

FIG 3. *Boxplots of the parameter estimates for ρ based on 1000 replications when employing the multiple-stage and bivariate probit estimation approaches. The dashed lines indicate the true values. Further details are given in the caption of Figure 2.*

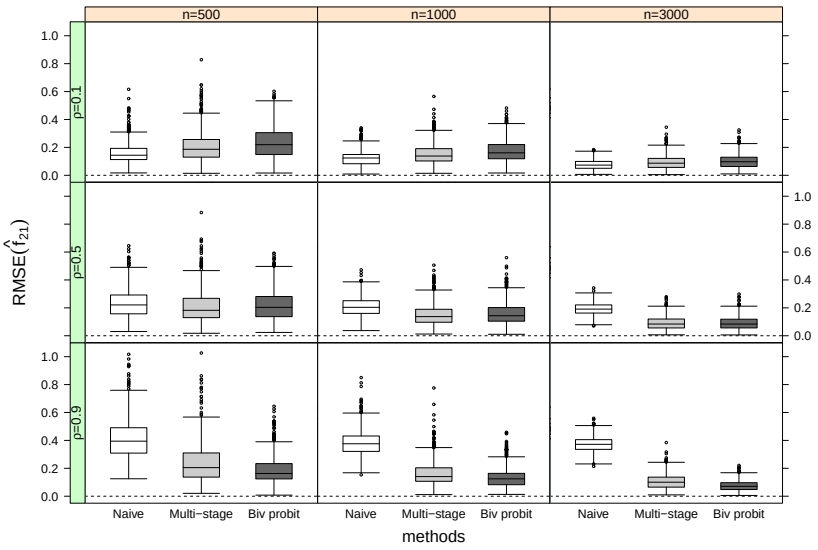

FIG 4. *Boxplots of the empirical root mean squared errors (RMSE) of* $\widehat{f}_{21}(z_1)$ *based on 1000 replications when employing the naive, multiple-stage and bivariate probit estimation approaches. Further details are given in the caption of Figure 2.*

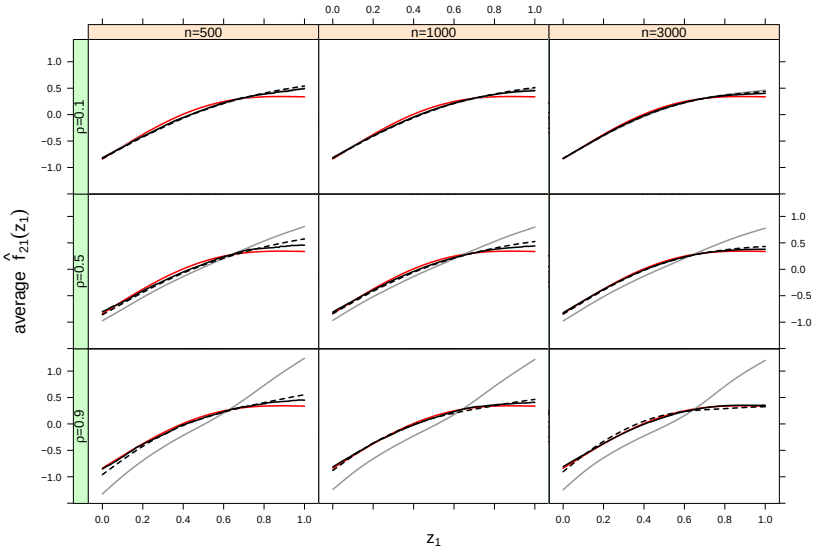

FIG 5. *Average fits calculated using the estimated smooth functions for* $f_{21}(z_1)$ *based on 1000 replications when employing the naive (grey solid lines), multiple-stage (black dashed lines) and bivariate probit (black solid lines) estimation approaches. The true function is represented by the red solid lines. Note that in some cases the curves differ only minimally, hence they can hardly be distinguished. Further details are given in the caption of Figure 2.*

most likely due to the violation of the normality assumption (see Remark 1 of Section 2.2).

- Figure 4 indicates that for $\rho = 0.1$ the performance of the naive estimator is superior at all sample sizes, and is comparable to that of the sample selection approaches for $\rho = 0.5$ and $n = 500$. In all the other cases, multiple-stage and bivariate probit outperform naive, with bivariate probit being the best for $\rho = 0.9$. The average fits, in Figure 5, show that the sample selection estimators yield better curve estimates except for $\rho = 0.1$ where naive recovers the underlying function fairly well.

In summary, in the presence of non-random sample selection and non-linear covariate effects, the sample selection estimators are less biased as compared to the naive estimator with the bivariate probit approach being the best. When the sample selection issue is negligible and the sample size small, the introduced estimators are more variable than naive. Models were also fitted using the data generating process described in the previous section but with no ER, i.e. without including z_{2i} in the selection equation. The results from this scenario (not reported here and available upon request) showed that the estimates of all methods are biased. This confirmed that empirical identification is achieved when the ER is present (see Section 2.4).

Average coverage probabilities of the 95% confidence intervals for θ_{22} and $f_{21}(z_1)$, constructed as described in Section 2.5, were also calculated. N_s and N_d were set to 20 and 100. For θ_{22}, the coverage rates of the multiple-stage approach were below the nominal level, with values going from 0.92 to 0.79 as ρ increases. For the bivariate probit method, despite its good accuracy in estimating θ_{22}, rates were low with values ranging from 0.73 to 0.88; here, nonparametric bootstrap percentile intervals based on 199 replications appeared to be effective, with rates in the interval $(0.91, 0.97)$. As for $f_{21}(z_1)$, nominal coverages were satisfactory for both methods with values in the range $(0.92, 0.97)$. For the multiple-stage approach, confidence intervals calculated employing the correction procedure described in Section 2.5 offered marginal improvements.

Coverage rates for θ_{22} were not satisfactory. For the multiple-stage estimator, this was most likely due to the bias in the estimates highlighted in Figure 2. For bivariate probit, the issue seems to lie in the information matrix which underestimates the variability of the parametric components. Limited simulation evidence suggests that using the observed (rather than Fisher) information matrix in (2.10) can yield improved coverage rates for parametric terms. However, as pointed out in Remark 1 of Section 2.3.1, the use of the Fisher information is crucial to be able to carry out the smoothing parameter selection step. Here, further research is needed to exploit the properties of both information matrices and avoid the use of computationally intensive bootstrap procedures.

4. Application

In this section, we illustrate the proposed methods using data on public opinion polls on school integration. As argued in [38], in certain situations, opinion polls

may poorly reflect collective public sentiment because some individuals choose not to respond to some specific questions as they feel that their opinion may be perceived as socially unacceptable. When some individuals are not willing to show their views, polls measuring collective opinion on sensitive topics typically provide misleading estimates of preferences in the population as a whole. A number of studies have recognized the presence of this phenomenon which can be problematic for policy information; see, e.g., [12]. A survey of public opinion polls on school integration conducted in the USA is one such study [2].

4.1. American national election study

We use data from the American National Election Survey (ANES) conducted in 1992[1]. This study is part of a time-series collection of national surveys fielded continuously since 1952. The election studies were designed to present data on Americans' social backgrounds, enduring political predispositions, social and political values, perceptions and evaluations of groups and candidates, opinions on questions of public policy, and participation in political life. The 1992 ANES study entailed both a pre-election interview and a post-election reinterview [28].

As mentioned in the introductory section, the main question was whether respondents support government intervention to ensure that black and white children go to the same school. 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%. The dataset also included information on individual demographic and socio-economic characteristics. The variables considered were `age` (in years), `educ` (number of years of education), `sex` (0 = female, 1 = male), `race` (0 = black, 1 = white), `reg` (0 = North-Central, 1 = North-East, 2 = South, 3 = West), `child` (number of children in the household), `discpol` (0 = never discuss politics, 1 = discuss politics), `moralcons` (1 = support for moral conservatism, 2 = no support for moral conservatism, 3 = neither), and `perslett` which was a binary variable indicating whether the interviewer attempted to convert a respondent who initially refused to participate in the survey.

We analyzed the ANES dataset using the naive, multiple-stage and bivariate probit estimation approaches with the same model fitting settings as those described in the simulation study section. The outcome equation included the variables `sex`, `race`, `reg`, `child`, `moralcons` and `discpol` as parametric components, and smooth functions of `age` and `educ`. The selection equation included the same variables plus `perslett`. The inclusion of `perslett` in the selection equation was used as ER on the ground that it may be regarded as a good predictor of the propensity to answer and is independent of the outcome. `child`

[1] http://www.electionstudies.org/studypages/1992prepost/1992prepost.htm.

was included as a parametric component because it did not have enough unique covariate values to justify the use of a smooth function.

4.2. Results and interpretation

Table 1 and Figure 6 report the parametric and smooth function estimates for the outcome equation and, for completeness, also those for the selection equation, when applying the three approaches on the ANES dataset.

In the selection equation, the parameter of `perslett`, obtained using the sample selection estimators, is statistically significant at the 5% level, hence supporting its use as ER. Although there are some differences in the significance of the parametric terms of the sample selection models, the magnitude and sign of the coefficients are similar. For example, the negative parameter estimate for `race.white` is consistent with the interpretation that the propensity

TABLE 1

Parametric estimates obtained applying the naive, multiple-stage and bivariate probit estimation approaches on the ANES dataset described in Section 4.1. Within parentheses are 95% confidence intervals calculated as described in Section 2.5 with $N_s = 20$ and $N_d = 100$ for multiple-stage, and nonparametric bootstrap based on 199 replications for bivariate probit. Note that results from the naive approach concern only the outcome equation

Variable	Naive	Multiple-stage	Bivariate probit
Selection Eq.			
`(Intercept)`	-	0.10 (−0.33, 0.54)	0.12 (−0.17, 0.41)
`sex.male`	-	0.05 (−0.16, 0.26)	0.04 (−0.10, 0.18)
`race.white`	-	−0.27 (−0.60, 0.06)	−0.27 (−0.53, −0.01)
`reg.northeast`	-	0.31 (0.00, 0.63)	0.34 (0.12, 0.56)
`reg.south`	-	0.18 (−0.10, 0.46)	0.19 (0.03, 0.35)
`reg.west`	-	0.36 (0.05, 0.66)	0.37 (0.17, 0.56)
`child`	-	0.08 (−0.03, 0.20)	0.09 (−0.01, 0.18)
`discpol`	-	0.24 (−0.01, 0.48)	0.21 (0.07, 0.34)
`moralcons.disagree`	-	0.26 (0.03, 0.49)	0.23 (0.06, 0.40)
`moralcons.neither`	-	−0.30 (−0.62, 0.02)	−0.33 (−0.52, −0.13)
`perslett`	-	0.25 (0.01, 0.50)	0.29 (0.07, 0.52)
Outcome Eq.			
`(Intercept)`	0.62 (0.07, 1.17)	0.28 (−0.96, 1.51)	−0.22 (−0.67, 0.24)
`sex.male`	−0.13 (−0.39, 0.13)	−0.12 (−0.37, 0.14)	−0.06 (−0.29, 0.18)
`race.white`	−0.71 (−1.12, −0.31)	−0.76 (−1.20, −0.31)	−0.61 (−0.94, −0.29)
`reg.northeast`	0.41 (0.02, 0.80)	0.48 (0.01, 0.95)	0.45 (0.08, 0.82)
`reg.south`	0.40 (0.03, 0.76)	0.43 (0.04, 0.82)	0.38 (0.05, 0.72)
`reg.west`	0.37 (−0.01, 0.75)	0.44 (−0.02, 0.91)	0.45 (0.09, 0.81)
`child`	0.11 (−0.03, 0.25)	0.13 (−0.02, 0.28)	0.12 (0.00, 0.25)
`discpol`	−0.33 (−0.66, 0.00)	−0.26 (−0.62, 0.10)	−0.10 (−0.36, 0.16)
`moralcons.disagree`	−0.33 (−0.61, −0.05)	−0.26 (−0.59, 0.08)	−0.11 (−0.37, 0.15)
`moralcons.neither`	0.01 (−0.44, 0.46)	−0.08 (−0.60, 0.44)	−0.16 (−0.49, 0.17)
ρ	-	0.77 (−0.36, 1.91)	0.86 (0.46, 1.00)

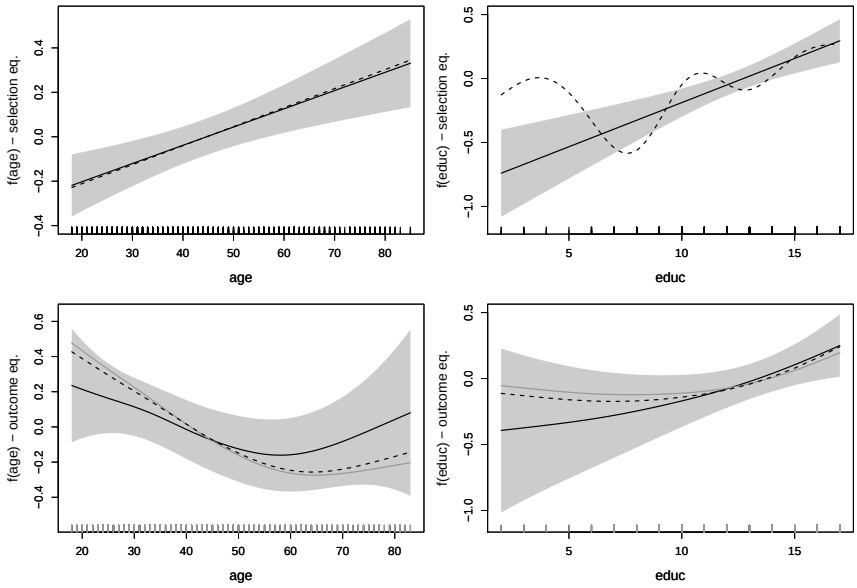

FIG 6. *Smooth function estimates obtained applying the naive (grey solid lines), multiple-stage (black dashed lines) and bivariate probit (black solid lines) estimation approaches on the ANES dataset described in Section 4.1. The results are reported on the scale of the linear predictors of the selection and outcome equations. The shaded regions represent 95% Bayesian 'confidence' intervals calculated from the bivariate probit estimates. The 'rug plot', at the bottom of each graph, shows the covariate values. Note that to avoid clutter corrected confidence intervals for the multiple-stage approach have not been reported, results from the naive approach concern only the outcome equation, and due to the identifiability constraints the estimated curves are centered around zero.*

that a white individual answers an integration question is lower than that of a non-white individual. Also, the fact that the interviewer attempted to convert a respondent who initially refused to participate in the survey has a positive effect on the probability of answering the integration question. As for the smooth components, the effect of `age` is linear for both sample selection models. The function estimates of `educ` are linear and nonlinear for bivariate probit and multiple-stage, respectively, and the bivariate probit intervals do not contain the multiple-stage curve for a part of the covariate value range. These findings support the presence of linear parametric effects for all terms in the selection equation of bivariate probit, and of linear and nonlinear covariate effects for multiple-stage. The reason for the difference in some of the estimates between the two sample selection estimators can perhaps be ascribed to sampling variability; at small sample sizes, the two methods are likely to be affected differently by some bias.

In the outcome equation, the probability that a white respondent supports integration is significantly lower as compared to that of a non-white; this conclusion is common to all approaches. The effects of `age` and `educ` show different

degrees of nonlinearity across the three methods. The pointwise confidence intervals of bivariate probit contain the zero line, suggesting that neither `age` nor `educ` have (non-linear or linear) effects. These results suggest that information in the data is too weak to clearly support the need for inclusion of smooth terms.

The estimates of ρ, which are important to ascertain the presence of bias induced by non-random sample selection, are high and, for bivariate probit, statistically significant. This means that the process by which individuals decide to answer the integration question is connected to the process by which they decide what the answer is. This could not have been detected using the naive approach. The positive sign of $\hat{\rho}$ indicates that the unobserved factors which lead individuals to take part in the survey also lead them to take a more supportive stance on the integration issue.

A comparison among the parametric estimates of the outcome equation obtained from the three approaches indicates that, while none of them change sign, the magnitude of some parameters is altered by the correction for non-random sample selection. Specifically, the movement of the estimates of `(Intercept)` is of interest. The naive estimate is more than double that of multiple-stage and four times that of bivariate probit. This means that once selection effects are accounted for, respondents are more likely to oppose school integration than the naive estimate suggests. This result is consistent with that of [2] who also found that correcting for non-random sample selection decreases the probability of supporting government efforts to integrate schools.

To gauge the aggregate effects of the sample selection issue in the question-answering process, we estimated the mean predicted probabilities of giving a supportive response under the three methods. 95% confidence intervals were conveniently obtained via posterior simulation using the results in Section 2.5. The results were 0.46 (0.43, 0.50), 0.41 (0.35, 0.46) and 0.31 (0.28, 0.35) for the naive, multiple-stage and bivariate probit approaches. So, predicted support for school integration is significantly lower when sample selection is accounted for, hence indicating that expressed opinion on the school integration question is a poor barometer of underlying support for integrationist policy. Based on our simulation evidence and the remarks given in Sections 2.2 and 2.3.1, the estimation results obtained using the bivariate probit approach may perhaps be regarded as the most accurate. The findings of this section may offer further empirical insights into how the school integration issue could be handled. For instance, support for school integration may be increased by investing in campaigns for social sensibilization.

The goodness of the results presented in this section relies especially on whether the assumption of normality is met. For the simultaneous equation estimation approach, a possibility would be to employ a score test of bivariate normality whose density of the errors under the alternative hypothesis is based on a type AA bivariate Gram Charlier series with 9 additional parameters [19]. However, it is not clear whether this test can be extended to the penalized framework proposed in this article.

5. Discussion

We introduced two statistical methods for the (separate or simultaneous) estimation of two binary regression models involving semiparametric predictors in the presence of non-random sample selection. The problems of identification and inference have also been discussed. The approaches have been illustrated using data on public opinion polls on school integration; predicted support for school integration calculated using the proposed tools is lower as compared to what a naive estimate would suggest.

The results of our simulation study showed that the sample selection estimators are less biased as compared to the naive estimator and that the bivariate probit approach can produce consistent parameter estimates. However, when the sample selection issue is negligible and the sample size small, the introduced approaches are more variable than naive; here, stability of sample selection estimators may be tenuous.

Because maximum likelihood estimation schemes are typically sensitive to model error misspecification, extensions of our proposals allowing for different joint distributions of the model errors seem feasible adopting copula functions. This approach has already been adopted in the context of non-random sample selection [18, 21, 35, 43, and references therein]. An alternative solution could be based on a nonparametric distribution function framework; see, e.g., [17]. To accommodate more complex data structures arising, for instance, in longitudinal studies, future research will also focus on extending the methods presented in this article to allow for random effects in the linear predictors. Finally, it would be interesting to determine whether a generalized least squares approach can be exploited to improve the efficiency of the sample selection estimators.

Acknowledgements

We are indebted to the Associate Editor and two reviewers for the detailed comments which helped us improve the quality of the manuscript.

References

[1] T. Bärnighausen, J. Bor, S. Wandira-Kazibw, and D. Canning. Correcting hiv prevalence estimates for survey nonparticipation using heckman-type selection models. *Epidemiology*, 22:27–35, 2011.

[2] A. J. Berinsky. The two faces of public opinion. *American Journal of Political Science*, 43:1209–1230, 1999.

[3] W. J. Boyes, D. L. Hoffman, and S. A. Low. An econometric analysis of the bank credit scoring problem. *Journal of Econometrics*, 40:3–14, 1989.

[4] M. Bratti and A. Miranda. Endogenous treatment effects for count data models with endogenous participation or sample selection. *Health Economics*, 20:90–1109, 2011.

[5] S. Chib and E. Greenberg. Semiparametric modeling and estimation of instrumental variable models. *Journal of Computational and Graphical Statistics*, 16:86–114, 2007. MR2345749

[6] P. Craven and G. Wahba. Smoothing noisy data with spline functions. *Numerische Mathematik*, 31:377–403, 1979. MR0516581

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

[8] W. P. M. M. Van de Ven and B. M. S. Van Praag. The demand for deductibles in private health insurance: a probit model with sample selection. *Journal of Econometrics*, 17:229–252, 1981.

[9] J. A. Dubin and D. Rivers. Selection bias in linear regression, logit and probit models. *Sociological Methods and Research*, 18:360–390, 1990.

[10] Paul H. C. Eilers and Brian D. Marx. Flexible smoothing with B-splines and penalties. *Statistical Science*, 11(2):89–121, 1996. MR1435485

[11] W. H. Greene. *Econometric Analysis*. Prentice Hall, New York, 2012.

[12] R. M. Groves, D. A. Dillman, J. L. Eltinge, and R. J. A. Little. *Survey Nonresponse*. Wiley, New York, 2001.

[13] C. Gu. Cross validating non-gaussian data. *Journal of Computational and Graphical Statistics*, 1:169–179, 1992.

[14] C. Gu. *Smoothing Spline ANOVA Models*. London: Springer-Verlag, 2002. MR1876599

[15] T. Hastie and R. Tibshirani. Varying-coefficient models. *Journal of the Royal Statistical Society Series B*, 55:757–796, 1993. MR1229881

[16] J. J. Heckman. Sample selection bias as a specification error. *Econometrica*, 47:153–162, 1979. MR0518832

[17] R. Klein and R. Spady. An efficient semiparametric estimator of the binary choice model. *Econometrica*, 61:387–421, 1993. MR1209737

[18] L. F. Lee. Generalized econometric models with selectivity. *Econometrica*, 51:507–512, 1983. MR0688735

[19] L. F. Lee. Tests for the bivariate normal distribution in econometric models with selectivity. *Econometrica*, 52:843–863, 1984. MR0750363

[20] S. F. Leung and S. Yu. Collinearity and two-step estimation of sample selection models: problems, origins, and remedies. *Computational Economics*, 15:173–199, 2000.

[21] P. Li and M. A. Rahman. Bayesian analysis of multivariate sample selection models using gaussian copulas. In D. M. Drukker, editor, *Missing Data Methods: Cross-sectional Methods and Applications. Volume 27 of Advances in Econometrics*, pages 269–288. Emerald Group Publishing Limited, 2011.

[22] G. S. Maddala. *Limited Dependent and Qualitative Variables in Econometrics*. Cambridge University Press, Cambridge, 1983. MR0799154

[23] G. Marra and R. Radice. Estimation of a semiparametric recursive bivariate probit model in the presence of endogeneity. *Canadian Journal of Statistics*, 39:259–279, 2011. MR2839480

[24] G. Marra and R. Radice. A flexible instrumental variable approach. *Statistical Modelling*, 11:581–279, 2011. MR2961695

[25] G. Marra and R. Radice. Estimation of a regression spline sample selection model. *Computational Statistics and Data Analysis*, 2013.

[26] G. Marra and R. Radice. *SemiParBIVProbit: Semiparametric Bivariate Probit Modelling*, 2013. R package version 3.2-6.

[27] G. Marra and S. N. Wood. Coverage properties of confidence intervals for generalized additive model components. *Scandinavian Journal of Statistics*, 39:53–74, 2012. MR2896791

[28] W. E. Miller, R. D. R. Kinder, S. J. Rosenstone, and National Election Studies. American national election study, 1992: Pre- and post-election survey [enhanced with 1990 and 1991 data]. Technical report, Inter-university Consortium for Political and Social Research [distributor], 1999.

[29] A. Miranda and S. Rabe-Hesketh. Maximum likelihood estimation of endogenous switching and sample selection models for binary, ordinal, and count variables. *Stata Journal*, 6:285–308, 2006.

[30] C. Montmarquettea, S. Mahseredjiana, and R. Houle. The determinants of university dropouts: a bivariate probability model with sample selection. *Economics of Education Review*, 20:475–484, 2001.

[31] J. Nocedal and S. J. Wright. *Numerical Optimization*. New York: Springer-Verlag, 2006. MR2244940

[32] P. A. Puhani. The heckman correction for sample selection and its critique. *Journal of Economic Surveys*, 14:53–68, 2000.

[33] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013. ISBN 3-900051-07-0.

[34] D. Ruppert, M. P. Wand, and R. J. Carroll. *Semiparametric Regression*. Cambridge University Press, New York, 2003. MR1998720

[35] M. D. Smith. Modelling sample selection using archimedean copulas. *Econometrics Journal*, 6:99–123, 2003. MR1992394

[36] J. V. Terza. Estimating count data models with endogenous switching: Sample selection and endogenous treatment effects. *Journal of Econometrics*, 84:129–154, 1998. MR1621944

[37] F. Vella. Estimating models with sample selection bias: a survey. *Journal of Human Resources*, 33:127–169, 1998.

[38] S. Verba, K. L. Schlozman, and H. E. Brady. *Voice and Equality: Civic Voluntarism in American Politics*. Cambridge: Harvard University Press, 1995.

[39] G. Wahba. Bayesian 'confidence intervals' for the cross-validated smoothing spline. *Journal of the Royal Statistical Society Series B*, 45:133–150, 1983. MR0701084

[40] M. Wiesenfarth and T. Kneib. Bayesian geoadditive sample selection models. *Journal of the Royal Statistical Society Series C*, 59:381–404, 2011. MR2756541

[41] S. N. Wood. Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association*, 99:673–686, 2004. MR2090902

[42] S. N. Wood. *Generalized Additive Models: An Introduction With R*. London: Chapman & Hall/CRC, 2006. MR2206355

[43] D. M. Zimmer and P. K. Trivedi. 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, 1983. MR2234712