Computational Statistics and Data Analysis 61 (2013) 158–173

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Estimation of a regression spline sample selection model

Giampiero Marra [a,*], Rosalba Radice [b]

[a] Department of Statistical Science, University College London, London WC1E 6BT, UK

[b] Department of Economics, Mathematics and Statistics, Birkbeck, University of London, London WC1E 7HX, UK

ARTICLE INFO

Article history:
Received 23 July 2012
Received in revised form 15 December 2012
Accepted 15 December 2012
Available online 23 December 2012

Keywords:
Non-random sample selection
Penalized regression spline
Selection bias
Simultaneous equation system

ABSTRACT

It is often the case that an outcome of interest is observed for a restricted non-randomly selected sample of the population. In such a situation, standard statistical analysis yields biased results. This issue can be addressed using sample selection models which are based on the estimation of two regressions: a binary selection equation determining whether a particular statistical unit will be available in the outcome equation. Classic sample selection models assume a priori that continuous regressors have a pre-specified linear or non-linear relationship to the outcome, which can lead to erroneous conclusions. In the case of continuous response, methods in which covariate effects are modeled flexibly have been previously proposed, the most recent being based on a Bayesian Markov chain Monte Carlo approach. A frequentist counterpart which has the advantage of being computationally fast is introduced. The proposed algorithm is based on the penalized likelihood estimation framework. The construction of confidence intervals is also discussed. The empirical properties of the existing and proposed methods are studied through a simulation study. The approaches are finally illustrated by analyzing data from the RAND Health Insurance Experiment on annual health expenditures.

1. Introduction

Sample selection models are used when the observations available for statistical analysis are not from a random sample of the population. Instead, individuals may have selected themselves into (or out of) the sample based on a combination of observed and unobserved characteristics. The use of statistical models ignoring such a non-random selection can have severe detrimental effects on parameter estimation.

As a motivating example, consider the RAND Health Insurance Experiment (RHIE), a study conducted in the United States between 1974 and 1982 (Newhouse, 1999). The aim was to quantify the relationship between various demographic and socio-economic characteristics (see Table 6 in Section 4) and annual health expenditures in the population as a whole. Non-random selection arises if the sample consisting of individuals who used health care services differ in important characteristics from the sample of individuals who did not use them. When the relationship between the decision to use the services and health expenditure is through observables, selection bias can be avoided by accounting for these variables. However, if some individuals are part of the selected subsample because of some observables as well as unobservables, then regardless of whether such variables are correlated in the population they will be in the selected sample (e.g., Dubin and Rivers, 1990). Hence, the neglect of this potential correlation can lead to inconsistent estimates of the covariate effects in the equation for annual expenditure.

Statistical methods correcting for the bias induced by non-random sample selection involve the estimation of two regression models: the selection equation (e.g., decision to use health services), and outcome equation (e.g., amount of health

* Corresponding author. Tel.: +44 0 20 7679 1864; fax: +44 0 20 3108 3105.
E-mail addresses: giampiero@stats.ucl.ac.uk (G. Marra), r.radice@bbk.ac.uk (R. Radice).

0167-9473
doi:10.1016/j.csda.2012.12.010

care expenditure). The latter is used to examine the substantive question of interest, whereas the former is used to detect selection bias and obtain consistent estimates of the covariate effects in the outcome equation. Since their introduction by Heckman (1979), sample selection models have been used in various fields (e.g., Bärnighausen et al., 2011; Cuddeback et al., 2004; Montmarquette et al., 2001; Sigelman and Zeng, 1999; Winship and Mare, 1992). Most of the case studies consider parametric sample selection models where continuous predictors have a pre-specified linear or non-linear relationship to the response variable. The need for techniques modeling flexibly regressor effects, without making a priori assumptions, arises from the observation that all parameter estimates are inconsistent when the relationship between covariates and outcome is mismodeled (e.g., Chib et al., 2009; Marra and Radice, 2011). This may prevent the researcher from recognizing, for instance, strong covariate effects or revealing interesting relationships. Going back to the health expenditure example, covariates such as age and education are likely to have a non-linear relationship to both decision to use health services and amount to spend on them. Imposing a priori a linear relationship (or non-linear by simply using quadratic polynomials, for example) could mean failing to capture possibly important complex relationships.

In a parametric context, sample selection models are typically estimated using the two-step framework first introduced by Heckman (1979): using the parameter estimates of the selection equation, a component (called inverse Mills ratio) is calculated and then included in the outcome equation to correct for non-random sample selection. Such an approach was proposed to deal with violations of the assumption of normality. However, it has been found to be sensitive to correlation among covariates in the outcome and selection equations, which can be really problematic in applications (Puhani, 2000). This problem can be alleviated by imposing an exclusion restriction, which requires at least one extra covariate to be a valid predictor in the selection equation but the outcome equation. A number of estimation methods which do not impose parametric forms on the error distribution have been introduced. These are termed 'semiparametric' since only part of the model of interest (the linear predictor) is parametrically pre-specified (e.g., Ahn and Powell, 1993; Lee, 1994; Martins, 2001; Newey et al., 1990; Powell, 1994; Vella, 1998). In this direction, recent developments include Marchenko and Genton (2012) and van Hasselt (2011). The sample selection literature has also been focusing on models with non-normal responses (e.g., Boyes et al., 1989; Terza, 1998; Smith, 2003; Greene, 2012). There are other variants of the sample selection model; these include Li (2011) who considered the case in which there is more than one selection mechanism, and Omori and Miyawaki (2010) who extended selection models to allow threshold values to depend on individuals' characteristics. These models have also been compared to principal stratification in the context of causal inference with nonignorable missingness (Mealli and Pacini, 2008).

We are interested in modeling flexibly covariate effects when the response is Gaussian. Das et al. (2003) considered the estimation of non-linear effects by extending the Heckman (1979) two-step estimation procedure. Recently, Chib et al. (2009) and Wiesenfarth and Kneib (2010) have introduced two more general estimation methods. Specifically, the approach of the former authors is based on Markov chain Monte Carlo simulation techniques and uses a simultaneous equation system that incorporates Bayesian versions of penalized smoothing splines. The latter further extended this approach by introducing a Bayesian algorithm based on low rank penalized B-splines for non-linear effects, varying-coefficient terms and Markov random-field priors for spatial effects. Using a model specification that is very similar to that of Wiesenfarth and Kneib (2010), we introduce a frequentist counterpart which has the advantage of being computationally fast. Our proposal can especially appeal to practitioners already familiar with traditional frequentist techniques. The proposed algorithm is based on the penalized maximum likelihood (ML) estimation framework, and is implemented in the R package `SemiParSampleSel` (Marra and Radice, 2012). As in a Bayesian framework, the proposal supports the choice of any class of smoothers albeit without requiring extra computational effort, an advantage which is not shared by a Bayesian implementation. The construction of confidence intervals is also discussed. The performance of the proposed and available methods are examined through a simulation study. Finally, the methods are illustrated analyzing data from the RAND Health Insurance Experiment on annual health expenditures.

2. Regression spline sample selection model

2.1. Model structure

The model consists of a system of two equations. Using the latent variable representation, the selection equation is

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

where n is the sample size, and y_{1i}^{*} is a latent continuous variable which determines its observable counterpart y_{1i} through the rule $1(y_{1i}^{*} > 0)$. The outcome equation determining the response variable of interest is

$$y_{2i} = \begin{cases} \mathbf{u}_{2i}^{\mathsf{T}}\boldsymbol{\theta}_2 + \sum_{k_2=1}^{K_2} s_{2k_2}(z_{2k_2 i}) + \varepsilon_{2i} & \text{if } y_{1i}^{*} > 0 \\ \text{not observed} & \text{if } y_{1i}^{*} \leq 0. \end{cases} \tag{2}$$

Vector $\mathbf{u}_{1i}^{\mathsf{T}} = \left(1, u_{12i}, \ldots, u_{1P_1 i}\right)$ is the ith row of $\mathbf{U}_1 = \left(\mathbf{u}_{11}^{\mathsf{T}}, \ldots, \mathbf{u}_{1n}^{\mathsf{T}}\right)^{\mathsf{T}}$, the $n \times P_1$ model matrix containing P_1 parametric model components (such as the intercept, dummy and categorical variables), with corresponding parameter vector $\boldsymbol{\theta}_1$, and the s_{1k_1} are unknown smooth functions of the K_1 continuous covariates $z_{1k_1 i}$. In line with Wiesenfarth and Kneib (2010), our implementation also supports varying coefficients models, obtained by multiplying one or more smooth terms by some predictor(s) (Hastie and Tibshirani, 1993), and smooth functions of two or more (e.g., spatial) covariates as described in Wood (2006, pp. 154–167). Similarly, $\mathbf{u}_{2i}^{\mathsf{T}} = \left(1, u_{22i}, \ldots, u_{2P_2 i}\right)$ is the ith row vector of the $n_s \times P_2$ model matrix $\mathbf{U}_2 = \left(\mathbf{u}_{21}^{\mathsf{T}}, \ldots, \mathbf{u}_{2n_s}^{\mathsf{T}}\right)^{\mathsf{T}}$, with coefficient vector $\boldsymbol{\theta}_2$, and the s_{2k_2} are unknown smooth terms of the K_2 continuous regressors $z_{2k_2 i}$. n_s denotes the size of the selected sample. For identification purposes, the smooth functions are subject to the centering constraint $\sum_i s_k(z_{ki}) = 0$ (Wood, 2006). As in Chib et al. (2009) and Wiesenfarth and Kneib (2010), we make the assumption that unobserved confounders have a linear impact on the responses. That is, the errors $(\varepsilon_{1i}, \varepsilon_{2i})$ are assumed to follow the bivariate distribution

$$\begin{pmatrix} \varepsilon_{1i} \\ \varepsilon_{2i} \end{pmatrix} \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho\sigma_2 \\ \rho\sigma_2 & \sigma_2^2 \end{bmatrix} \right), \tag{3}$$

where ρ is the correlation coefficient, σ_2 the standard deviation of ε_{2i} and σ_1, the standard deviation of ε_{1i}, is set to 1 because the parameters in the selection equation can only be identified up to a scale coefficient (e.g., Greene, 2012, p. 686). The assumption of normality may be, perhaps, too restrictive for applied work; however it is typically made to obtain more tractable expressions.

The smooth functions are represented using the regression spline approach. To fix ideas and in the one-dimensional case, a generic $s_k(z_{ki})$ is approximated by a linear combination of known spline basis functions, $b_{kj}(z_{ki})$, and regression parameters, β_{kj},

$$s_k(z_{ki}) = \sum_{j=1}^{J_k} \beta_{kj} b_{kj}(z_{ki}) = \mathbf{B}_k(z_{ki})^{\mathsf{T}} \boldsymbol{\beta}_k,$$

where J_k is the number of spline bases (hence regression coefficients) used to represent s_k, $\mathbf{B}_k(z_{ki})^{\mathsf{T}}$ is the ith vector of dimension J_k containing the basis functions evaluated at the observation z_{ki}, i.e. $\mathbf{B}_k(z_{ki}) = \left\{b_{k1}(z_{ki}), b_{k2}(z_{ki}), \ldots, b_{kJ_k}(z_{ki})\right\}^{\mathsf{T}}$, and $\boldsymbol{\beta}_k$ 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 will give a (linear or non-linear) estimate for $s_k(z_k)$. The number of basis functions, J_k, determines the flexibility allowed for $s_k(z_k)$; $J_k = 10$ will lead to a "wigglier" curve estimate than when such a parameter is set to 5, for instance. Because, in practice, it is not easy to determine the optimal number of basis functions for the $s_k(z_{ki})$, a reasonably large number for the J_k is typically chosen (to allow enough flexibility in the model) and then the corresponding $\boldsymbol{\beta}_k$ penalized in order to suppress that part of non-linearity which is not supported from the data. As it will be shown in the next section, this can be achieved using a penalized estimation approach which is the mainstream in the regression spline literature (see Ruppert et al., 2003 and Wood, 2006 for more details). Basis functions should be chosen to have convenient mathematical properties and good numerical stability. Many choices are possible and supported in our implementation including B-splines, cubic regression and low rank thin plate regression splines (e.g., Ruppert et al., 2003). The case of smooths of more than one variable follows a similar construction. Based on the result above, Eqs. (1) and (2) can be written as

$$y_{1i}^* = \mathbf{u}_{1i}^{\mathsf{T}} \boldsymbol{\theta}_1 + \mathbf{B}_{1i}^{\mathsf{T}} \boldsymbol{\beta}_1 + \varepsilon_{1i}, \quad i = 1, \ldots, n, \tag{4}$$

and

$$y_{2i} = \mathbf{u}_{2i}^{\mathsf{T}} \boldsymbol{\theta}_2 + \mathbf{B}_{2i}^{\mathsf{T}} \boldsymbol{\beta}_2 + \varepsilon_{2i}, \quad i = 1, \ldots, n_s, \tag{5}$$

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

In principle, the parameters of the sample selection model are identified even if $\left(\mathbf{u}_{1i}^{\mathsf{T}}, \mathbf{B}_{1i}^{\mathsf{T}}\right) = \left(\mathbf{u}_{2i}^{\mathsf{T}}, \mathbf{B}_{2i}^{\mathsf{T}}\right)$ (Wiesenfarth and Kneib, 2010). In practice, however, the absence of equation-specific regressors may lead to a likelihood function which does not vary significantly over a wide region around the mode (e.g., Marra and Radice, 2011). Moreover, in applications, both functional form and model errors are likely to be misspecified to some degree. This suggests that empirical identification is better achieved if an exclusion restriction (ER) on the covariates in the two equations holds (e.g., Chib et al., 2009; Vella, 1998). That is, the regressors in the selection equation should contain at least one or more regressors not included in the outcome equation. See the simulation results in Section 3 for more discussion of this issue.

2.2. A penalized maximum likelihood estimation approach

Recall that the error terms $(\varepsilon_{1i}, \varepsilon_{2i})$ are assumed to follow a bivariate normal distribution and define the linear predictors $\eta_{vi} = \mathbf{u}_{vi}^{\mathsf{T}} \boldsymbol{\theta}_v + \mathbf{B}_{vi}^{\mathsf{T}} \boldsymbol{\beta}_v$, $v = 1, 2$. The observations can be divided into two groups according to the type of data observed. Each

group of observations has a different form for the likelihood. When y_{2i} is observed, the likelihood function is the probability of the joint event y_{2i} and $y_{1i}^* > 0$, i.e.

$$\begin{aligned}\mathbb{P}(y_{2i}, y_{1i}^* > 0) &= f(y_{2i})\,\mathbb{P}\left(y_{1i}^* > 0 | y_{2i}\right) = f(\varepsilon_{2i})\,\mathbb{P}(\varepsilon_{1i} > -\eta_{1i}|\varepsilon_{2i}) \\ &= \frac{1}{\sigma_2}\phi\left(\frac{y_{2i}-\eta_{2i}}{\sigma_2}\right)\int_{-\eta_{1i}}^{\infty} f(\varepsilon_{1i}|\varepsilon_{2i})\,d\varepsilon_{1i} \\ &= \frac{1}{\sigma_2}\phi\left(\frac{y_{2i}-\eta_{2i}}{\sigma_2}\right)\int_{-\eta_{1i}}^{\infty}\frac{1}{\sqrt{1-\rho^2}}\phi\left\{\frac{\varepsilon_{1i}-\frac{\rho}{\sigma_2}(y_{2i}-\eta_{2i})}{\sqrt{1-\rho^2}}\right\}d\varepsilon_{1i} \\ &= \frac{1}{\sigma_2}\phi\left(\frac{y_{2i}-\eta_{2i}}{\sigma_2}\right)\Phi\left\{\frac{\eta_{1i}+\frac{\rho}{\sigma_2}(y_{2i}-\eta_{2i})}{\sqrt{1-\rho^2}}\right\}.\end{aligned}$$

When y_{2i} is not observed, the likelihood function just corresponds to the marginal probability that $y_{1i}^* \leq 0$, i.e.

$$\mathbb{P}\left(y_{1i}^* \leq 0\right) = \mathbb{P}(\varepsilon_{1i} \leq -\eta_{1i}) = \Phi(-\eta_{1i}) = 1 - \Phi(\eta_{1i}).$$

Therefore, the log-likelihood function for the complete sample of observations is

$$\ell(\boldsymbol{\delta}) = \sum_{i=1}^{n}(1-y_{1i})\log\{1-\Phi(\eta_{1i})\} + y_{1i}\left[-\log\sigma_2 + \log\phi\left(\frac{y_{2i}-\eta_{2i}}{\sigma_2}\right) + \log\Phi\left\{\frac{\eta_{1i}+\frac{\rho}{\sigma_2}(y_{2i}-\eta_{2i})}{\sqrt{1-\rho^2}}\right\}\right],$$

where $\boldsymbol{\delta}^{\mathsf{T}} = (\boldsymbol{\delta}_1^{\mathsf{T}}, \boldsymbol{\delta}_2^{\mathsf{T}}, \sigma_2, \rho)$ and $\boldsymbol{\delta}_v^{\mathsf{T}} = (\boldsymbol{\theta}_v^{\mathsf{T}}, \boldsymbol{\beta}_v^{\mathsf{T}})$, for $v = 1, 2$.

As explained in the previous section, because of the flexible linear predictor structure considered in this article, unpenalized ML estimation is likely to result in smooth term estimates which are too rough to produce practically useful results. This issue can be dealt with by augmenting the objective function with a penalty term, such as $\sum_{v=1}^{2}\sum_{k_v=1}^{K_v}\lambda_{vk_v}\int s_k''(z_{vk_v})^2 dz_{vk_v}$, measuring the (second-order, in this case) roughness of the smooth terms in the model. The λ_{vk_v} 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 $\boldsymbol{\beta}^{\mathsf{T}} = (\boldsymbol{\beta}_1^{\mathsf{T}}, \boldsymbol{\beta}_2^{\mathsf{T}})$, i.e. $\boldsymbol{\beta}^{\mathsf{T}}\mathbf{S}_{\boldsymbol{\lambda}}\boldsymbol{\beta}$ where $\mathbf{S}_{\boldsymbol{\lambda}} = \sum_{v=1}^{2}\sum_{k_v=1}^{K_v}\lambda_{vk_v}\mathbf{S}_{vk_v}$ and the $\mathbf{S}_{vk_v}$ are positive semi-definite known square matrices. The penalized log-likelihood is therefore given as

$$\ell_p(\boldsymbol{\delta}) = \ell(\boldsymbol{\delta}) - \frac{1}{2}\boldsymbol{\beta}^{\mathsf{T}}\mathbf{S}_{\boldsymbol{\lambda}}\boldsymbol{\beta}. \tag{6}$$

Because ρ is bounded in $[-1, 1]$ and σ_2 can only take positive real values, we use $\rho^* = \tanh^{-1}(\rho) = (1/2)\log\{(1+\rho)/(1-\rho)\}$ and $\sigma_2^* = \log(\sigma_2)$ in optimization. Given values for the λ_{vk_v}, we seek to maximize (6). In practice, this can be achieved by Newton–Raphson's methods iterating

$$\hat{\boldsymbol{\delta}}^{[a+1]} = \hat{\boldsymbol{\delta}}^{[a]} + (\boldsymbol{\mathcal{H}}^{[a]} - \mathbf{S}_{\boldsymbol{\lambda}}^*)^{-1}(\mathbf{S}_{\boldsymbol{\lambda}}^*\hat{\boldsymbol{\delta}}^{[a]} - \mathbf{g}^{[a]}) \tag{7}$$

until convergence, where a is an iteration index and $\mathbf{S}_{\boldsymbol{\lambda}}^*$ an overall block-diagonal penalty matrix, i.e.

$$\mathbf{S}_{\boldsymbol{\lambda}}^* = \mathrm{diag}(0_{11}, \ldots, 0_{1P_1}, \lambda_{1k_1}\mathbf{S}_{1k_1}, \ldots, \lambda_{1K_1}\mathbf{S}_{1K_1}, 0_{21}, \ldots, 0_{2P_2}, \lambda_{2k_2}\mathbf{S}_{2k_2}, \ldots, \lambda_{2K_2}\mathbf{S}_{2K_2}, 0, 0).$$

The score 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 two scalars $g_3 = \partial\ell(\boldsymbol{\delta})/\partial\sigma_2^*$ and $g_4 = \partial\ell(\boldsymbol{\delta})/\partial\rho^*$, while the Hessian matrix has a 4×4 matrix block structure with (r, h)th element $\boldsymbol{\mathcal{H}}_{r,h} = \partial^2\ell(\boldsymbol{\delta})/\partial\boldsymbol{\delta}_r\partial\boldsymbol{\delta}_h^{\mathsf{T}}$, $r, h = 1, \ldots, 4$, where $\boldsymbol{\delta}_3 = \sigma_2^*$ and $\boldsymbol{\delta}_4 = \rho^*$. The derivations of $\mathbf{g}$ and $\boldsymbol{\mathcal{H}}$ are tedious; these are given in Appendix A.

The main issues with the maximization problem (6) are that $\ell(\boldsymbol{\delta})$ is not globally concave (e.g., Toomet and Henningsen, 2008), and that the Hessian may become non-positive definite on some occasions. Preliminary work confirmed these concerns as well as that the use of classic optimization schemes, implemented using R functions `nlm()` and `optim()`, do not perform satisfactorily on this problem. To tackle such issues, (7) is implemented using a trust region algorithm with eigen-decomposition of $\boldsymbol{\mathcal{H}}$ at each iteration (e.g., Nocedal and Wright, 1999, Section 4.2), and initial values are supplied using an adaptation of the Heckman procedure (1979) which is detailed in Appendix B. This approach proved to be fast and reliable in most cases, with occasional convergence failure for small values of n and n_s.

Joint estimation of $\boldsymbol{\delta}$ and $\boldsymbol{\lambda}$ (containing the λ_{vk_v}) via maximization of (6) would result in overfitting since the highest value for $\ell_p(\boldsymbol{\delta})$ would be obtained when $\boldsymbol{\lambda} = \mathbf{0}$. This is why in (7) the λ_{vk_v} are fixed at some values. The next section illustrates how $\boldsymbol{\lambda}$ can be estimated.

2.2.1. Smoothness selection

Smoothing parameter selection is important for practical modeling. In principle, it can be achieved by direct grid search optimization of, for instance, the Akaike information criterion (AIC; Akaike, 1973). However, if the model has more than two or three smooth terms, this typically becomes computationally burdensome, hence making the model building process difficult in most applied contexts. There are a number of techniques for automatic multiple smoothing parameter estimation for univariate regression spline models. Without claim of exhaustiveness, we briefly describe some of them. Gu (1992) introduced the performance-oriented iteration method which applies generalized cross validation (GCV) or the unbiased risk estimator (UBRE; Craven and Wahba, 1979) to each working linear model of the penalized iteratively re-weighted least squares (P-IRLS) scheme used to fit the model. Wood (2004) extended this approach by providing an optimally stable computational procedure. Smoothing parameter selection can also be achieved by exploiting the mixed model representation of penalized regression spline models. Here, smoothing parameters become variance components and, as such, can be estimated by either ML or restricted maximum likelihood (REML) for the Gaussian case, and by penalized quasi-likelihood for the generalized case (e.g., Breslow and Clayton, 1993; Ruppert et al., 2003). Wahba (1985) showed that asymptotically prediction error criteria are better in a mean square error sense, even though Härdle et al. (1988) pointed out that these criteria give very slow convergence to the optimal smoothing parameters. Recent work by Reiss and Ogden (2009) shows that at finite sample sizes both GCV and AIC are more prone to undersmoothing and more likely to develop multiple minima than REML. However, as pointed out by Ruppert et al. (2003, p. 177), automatic smoothing parameter selectors might be somewhat erratic; they provide an empirical example where REML leads to severe oversmoothing. We adapt Gu's approach to the current context which, in our experience, proved to be very efficient and stable in most cases.

Given values for the λ_{vk_v}, noting that Newton–Raphson's iterative Eq. (7) can be written in the P-IRLS form, $\hat{\boldsymbol{\delta}}^{[a+1]}$ is the solution to the problem

$$\text{minimize } \|\sqrt{\mathbf{W}}^{[a]}(\mathbf{z}^{[a]} - \mathbf{X}\boldsymbol{\delta})\|^2 + \boldsymbol{\delta}^{\mathsf{T}}\mathbf{S}^*_{\boldsymbol{\lambda}}\boldsymbol{\delta} \quad \text{w.r.t. } \boldsymbol{\delta}, \tag{8}$$

where $\sqrt{\mathbf{W}}$ is any iterative weight non-diagonal matrix square root such that $\sqrt{\mathbf{W}}^{\mathsf{T}}\sqrt{\mathbf{W}} = \mathbf{W}$, and $\mathbf{z}_i$ is a 4-dimensional pseudodata vector given as $\mathbf{z}_i = \mathbf{X}_i\boldsymbol{\delta}^{[a]} + \mathbf{W}_i^{-1}\mathbf{d}_i$, where $\mathbf{d}_i = \{\partial\ell(\boldsymbol{\delta})_i/\partial\eta_{1i}, \partial\ell(\boldsymbol{\delta})_i/\partial\eta_{2i}, \partial\ell(\boldsymbol{\delta})_i/\partial\eta_{3i}, \partial\ell(\boldsymbol{\delta})_i/\partial\eta_{4i}\}^{\mathsf{T}}$, $\eta_{3i} = \sigma_2^*$, and $\eta_{4i} = \rho^*$. $\mathbf{W}_i$ is a 4×4 matrix with (r, h)th element $(\mathbf{W}_i)_{rh} = -\partial^2\ell(\boldsymbol{\delta})_i/\partial\eta_{ri}\partial\eta_{hi}$, $r, h = 1, \ldots, 4$, and, assuming without loss of generality that the spline basis dimensions for the smooth terms in the model are all equal to J, $\mathbf{X}_i$ is a $4 \times \{(P_1 + K_1 \times J) + (P_2 + K_2 \times J) + 2\}$ block diagonal matrix, i.e. $\mathbf{X}_i = \text{diag}\left\{\left(\mathbf{u}_{1i}^{\mathsf{T}}, \mathbf{B}_{1i}^{\mathsf{T}}\right), \left(\mathbf{u}_{2i}^{\mathsf{T}}, \mathbf{B}_{2i}^{\mathsf{T}}\right), 1, 1\right\}$. The superscript $[a]$ has been suppressed from $\mathbf{d}_i$, $\mathbf{z}_i$ and $\mathbf{W}_i$, and is omitted from the quantities shown in the next paragraph, to avoid clutter.

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

$$\text{minimize } \mathcal{V}_u^w(\boldsymbol{\lambda}) = \frac{1}{n_*}\|\sqrt{\mathbf{W}}(\mathbf{z} - \mathbf{X}\boldsymbol{\delta})\|^2 - 1 + \frac{2}{n_*}\text{tr}(\mathbf{A}_{\boldsymbol{\lambda}}) \quad \text{w.r.t. } \boldsymbol{\lambda}, \tag{9}$$

where the working linear model quantities are constructed for a given estimate of $\boldsymbol{\delta}$, obtained in Newton–Raphson's equation (7) or (8), $n_* = 4n$, $\mathbf{A}_{\boldsymbol{\lambda}} = \mathbf{X}(\mathbf{X}^{\mathsf{T}}\mathbf{W}\mathbf{X} + \mathbf{S}^*_{\boldsymbol{\lambda}})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{W}$ is the hat matrix and $\text{tr}(\mathbf{A}_{\boldsymbol{\lambda}})$ the estimated degrees of freedom of the penalized model. For each working linear model of the P-IRLS iteration, $\mathcal{V}_u^w(\boldsymbol{\lambda})$ is minimized by employing the approach by Wood (2004), which is based on Newton–Raphson's method and can evaluate the approximate UBRE and its derivatives in a way that is both computationally efficient and stable. Note that because $\mathbf{W}$ is a non-diagonal matrix of dimension $n_* \times n_*$, computation can be prohibitive, even for small sample sizes. To this end, $\mathbf{W}^{-1}\mathbf{d}$, $\sqrt{\mathbf{W}}\mathbf{z}$ and $\sqrt{\mathbf{W}}\mathbf{X}$ are calculated exploiting the sparse structure of $\mathbf{W}$. Hence, the working linear model in (9) can be formed in $O(n_*(m + 2))$ rather than $O(n_*^2(m + 2))$ operations, where m is the number of columns of $\mathbf{X}$.

The issue with evaluating the approximate UBRE is that the $\mathbf{W}_i$ are not guaranteed to be positive-definite, mainly because of σ_2^* and ρ^* (e.g., Marra and Radice, 2011; Yee, 2010). This is problematic in that $\sqrt{\mathbf{W}}$ and $\mathbf{W}^{-1}$ are needed in (9). As a solution, the working linear model is constructed so that its key quantities depend on all model parameters but σ_2^* and ρ^* since these are not penalized. In this way, it is possible to construct the working linear model quantities needed in (9) as well as reduce substantially the computational load and storage demand of the algorithm since in this case $n_* = 2$.

The structure of the algorithm used for estimating parameter vector $\boldsymbol{\delta}$ is given in Appendix C.

2.3. Inference

Inferential theory for penalized estimators is not standard. This is because of the presence of smoothing penalties which undermines the usefulness of classic frequentist results for practical modeling. Solutions to this problem have been introduced in the literature (see, e.g., Gu, 2002 and Wood, 2006 for an overview). Here, we show how to construct pointwise confidence intervals for the terms of a regression spline sample selection model by adapting to the current context the well known Bayesian 'confidence' intervals, originally proposed by Wahba (1983) and Silverman (1985). An appealing characteristic of these intervals is that they have close to nominal 'across-the-function' *frequentist* coverage probabilities

for the components of a generalized additive model (Marra and Wood, 2012). To see this point, consider a generic $s_k(z_{ki})$. Intervals can be constructed seeking some constants C_{ki} and A, such that

$$\text{ACP} = \frac{1}{n}\mathbb{E}\left\{\sum_i \mathbb{I}(|\hat{s}_k(z_{ki}) - s_k(z_{ki})| \leq q_{\alpha/2}A/\sqrt{C_{ki}})\right\} = 1-\alpha, \tag{10}$$

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 $b_k(z_k) = \mathbb{E}\{\hat{s}_k(z_k)\} - s_k(z_k)$ and $v_k(z_k) = \hat{s}_k(z_k) - \mathbb{E}\{\hat{s}_k(z_k)\}$, so that $\hat{s}_k - s_k = b_k + v_k$, and I to be a random variable uniformly distributed on $\{1, 2, \ldots, n\}$, we have that ACP $=$ $\Pr\left(|B_k + V_k| \leq q_{\alpha/2}A\right)$, where $B_k = \sqrt{C_{kI}}b(z_{kI})$ and $V_k = \sqrt{C_{kI}}v(z_{kI})$. It is then necessary to find the distribution of $B_k + V_k$ and values for C_{ki} and A so that the requirement (10) is met. As shown in Marra and Wood (2012), in the context of non-Gaussian response models involving several smooth components, such a requirement is approximately met when confidence intervals for the $\hat{s}_k(z_{ki})$ are constructed using

$$\boldsymbol{\delta}|\mathbf{y} \dot{\sim} \mathcal{N}(\hat{\boldsymbol{\delta}}, \mathbf{V}_{\boldsymbol{\delta}}), \tag{11}$$

where, in the current context, $\mathbf{y}$ refers to the response vectors, $\hat{\boldsymbol{\delta}}$ is an estimate of $\boldsymbol{\delta}$ and $\mathbf{V}_{\boldsymbol{\delta}} = (-\boldsymbol{\mathcal{H}} + \mathbf{S}^*_{\boldsymbol{\lambda}})^{-1}$. The structure of this variance–covariance matrix is such that it includes both a bias and a variance component in a frequentist sense (Marra and Wood, 2012). Given result (11), confidence intervals for linear and non-linear functions of the model parameters can be easily obtained. For any parametric model components, using (11) is equivalent to using classic likelihood results because such terms are not penalized. Note that there is no contradiction in fitting the sample selection model via penalized ML and then constructing intervals using a Bayesian result, and such an approach has been adopted many times in the literature (e.g., Gu, 2002; Marra and Radice, 2011; Wood, 2006). Moreover, the quantities needed to construct the intervals are obtained as a byproduct of the estimation process; hence no extra computation is really required for inferential purposes.

Frequentist approaches treat $\boldsymbol{\lambda}$ as known. It is, therefore, reasonable to expect the neglect of the variability due to smoothing parameter estimation to lead to undercoverage of the intervals. This problem should become more relevant as the number of smoothing parameters increases, which is especially the case for sample selection models as compared to single equation models. In this respect, Bayesian approach would be advantageous as a smoothing parameter uncertainty can be naturally taken into account. As shown by Marra and Wood (2012), provided that smoothing parameters are selected so that the estimation bias is not too large a proportion of the sampling variability, the empirical performance of the intervals should have little or no sensitivity to the neglect of smoothing parameter uncertainty. This suggests that a fully Bayesian approach like that of Wiesenfarth and Kneib (2010) may lead to overcoverage. The simulation study in the next section will also shed light on this issue.

3. Simulation study

In this section, we conduct a Monte Carlo simulation study to compare the proposed method with classic univariate regression and the approach of Wiesenfarth and Kneib (2010). Computations were performed in the R environment (R Development Core Team, 2012) using the package `SemiParSampleSel` (Marra and Radice, 2012), which implements the ideas discussed in the previous sections, and `bayesSampleSelection` (available at http://www.uni-goettingen.de/en/96061.html) written by Wiesenfarth. The approach by Chib et al. (2009) was not used in the comparison because of lack of R code. However, this should not be problematic as their method is closely related to that of Wiesenfarth and Kneib (2010).

The sampling experiments were based on the model

$$y^*_{1i} = \theta_{11} + \theta_{12}u_i + s_{11}(z_{1i}) + s_{12}(z_{2i}) + \varepsilon_{1i}$$
$$y_{2i} = \theta_{21} + \theta_{22}u_i + s_{21}(z_{1i}) + \varepsilon_{2i},$$

where y_{1i} and y_{2i} were determined as described in Section 2.1. The test functions used were $s_{11}(z_{1i}) = -0.7\left\{4z_{1i} + 2.5z^2_{1i} + 0.7\sin(5z_{1i}) + \cos(7.5z_{1i})\right\}$, $s_{12}(z_{2i}) = -0.4\{-0.3 - 1.6z_{2i} + \sin(5z_{2i})\}$, and $s_{21}(z_{1i}) = 0.6\{\exp(z_{1i}) + \sin(2.9z_{1i})\}$ (see Fig. 1). $(\theta_{12}, \theta_{21}, \theta_{22})$ and σ_2 were set to $(2.5, -0.68, -1.5)$ and 1. To generate binary values for y_{1i} so that approximately 25%, 50% and 75% of the total number of observations were selected to fit the outcome equation, θ_{11} was set to -0.65, 0.58 and 1.66, respectively. Regressors u_i, z_{1i} and z_{2i} were generated as three uniform covariates on (0, 1) with correlation approximately equal to 0.5. This was achieved using `rmvnorm()` in the package `mvtnorm`, generating standardized multivariate random draws with correlation 0.5 and then applying `pnorm()` (e.g., Marra and Radice, 2011). Regressor u_i was dichotomized using `round()`. Standardized bivariate normal errors with correlations $\rho = (\pm 0.1, \pm 0.5, \pm 0.9)$ were considered, and sample sizes were set to 500, 1500 and 3000. For each combination of parameter settings, the number of simulated datasets was 250. Models were fitted with and without exclusion restriction (ER and non-ER, respectively). Specifically, in the latter case z_{2i} was not included in the selection equation.

For the approach of Wiesenfarth and Kneib (2010), we used the following settings. The number of iterations for the burn-in period, number of samples used for estimation, and degree of thinning were 2000, 22 000 and 20, respectively, and smooth terms were represented using P-spline bases with 20 inner knots and penalty matrices containing second order differences.

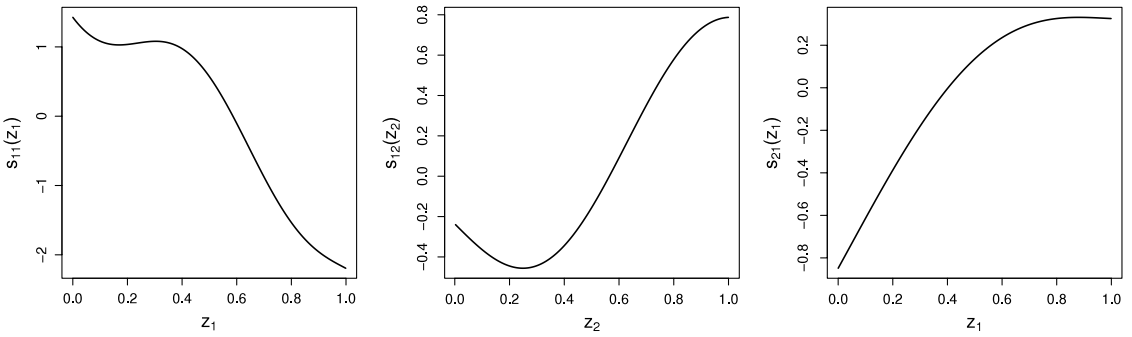

Fig. 1. The test functions used in the simulation studies.

To make a fair comparison, the smooth components of the proposed method were represented using P-splines with the same settings. Models were also fitted neglecting the sample selection issue: using the selected sample, simply fit Eq. (5) via penalized least squares, again with the same number of P-spline bases and penalty order. As this naive approach cannot correct sample selection bias, badly biased parameter estimates are clearly expected in the simulation results. However, reporting such results should be useful to highlight the negative effects that the neglect of non-random sample selection has on parameter estimates. Following a reviewer's suggestion, we also employed a standard Heckman approach where non-linear effects were modeled using second-order polynomial terms.

3.1. Results

In this section, we only show a subset of results; these are representative of all empirical findings. Since the selection equation is not affected by sample selection bias, we focus on the estimation results for the outcome equation only. Table 1 reports the percentage relative bias and the root mean squared error (RMSE) for θ_{22}, ρ and σ, when assuming that ER holds and approximately 50% of the total number of observations are available to fit the outcome equation. The approaches employed are naive, standard Heckman with second-order polynomial terms, and penalized Bayesian and ML estimation (naive, HeckP, W&K and M&R, respectively). Table 2 shows the RMSE and 95% average coverage probability (ACP) for the four approaches when estimating $s_{21}(z_1)$ under the same settings as described above. Tables 3 and 4 report the bias and RMSE for $\hat{\theta}_{22}$, $\hat{\rho}$ and $\hat{\sigma}$, and RMSE and ACP for $\hat{s}_{21}(z_1)$, respectively, for the non-ER case when approximately 75% of the total number of observations are selected for the outcome equation. This scenario should be reasonably close to the empirical illustration presented in Section 4, where there is no ER and the 77% of observations are selected. The results for the non-ER case when approximately 50% of observations are selected are reported in Appendix D (Table 8). Table 5 reports the results obtained under the ER scenario with 50% of selected observations when in the data generating process the simple non-linear function $s_{21}(z_1)$ is swapped with function $s_{11}(z_1)$. In this case, θ_{11} was set to -3.05. As in Wiesenfarth and Kneib (2010), based on the estimates for 200 fixed covariate values, RMSE($\hat{s}$) was calculated as $\sqrt{\sum_{b=1}^{200}\left\{\hat{s}(z_{1b})-s(z_{1b})\right\}^2}$. In terms of computing mean times, M&R showed lower computational cost than W&K. Specifically, mean times for M&R were between 0.02 and 0.17 min depending on n and ρ. Compared to M&R, W&K approximately took between 25 and 100 times longer to fit a sample selection model.

The main results can be summarized as follows.

- Table 1 shows that the neglect of non-random sample selection leads to seriously biased parameter estimates. When ρ is small, HeckP outperforms all other methods in terms of bias. However, as ρ increases (i.e., the sample selection issue becomes more pronounced), M&R and W&K outperform HeckP, with M&R being the best in terms of bias. For high ρ, M&R performs the best in terms of bias and precision. As n increases, the W&K, M&R and HeckP estimates show convergence to their true values. Similar conclusions are reached for $s_{21}(z_1)$ (see Table 2). The 95% ACPs for M&R are more accurate than those for W&K and HeckP. This suggests that a fully Bayesian approach, which accounts for smoothing parameter uncertainty, yields slightly conservative pointwise intervals. This finding supports the argument by Marra and Wood (2012) (see Section 2.3) and is in agreement with the results of Wiesenfarth and Kneib (2010). We did not report the ACPs for the biased naive fits because these were clearly below the nominal level.
- In the non-ER case, when 50% of observations are selected (Table 8 in Appendix D), all methods exhibit higher bias as compared to the ER case. In addition, similar trends as those in Tables 1 and 2 are detected, with W&K and M&R being the best and naive being the worst as ρ increases. When about three-fourth of the total number of observations are available for the outcome equation, the same patterns can again be observed but with lower bias and RMSE (see Tables 3 and 4). This finding complements the argument in the final paragraph of Section 2.1 by suggesting that the presence of ER is necessary to obtain good estimation results, unless the percentage of selected observations is high.
- Following a reviewer's suggestion, we also compared HeckP and M&R under the ER scenario with 50% of selected observations when in the data generating process $s_{21}(z_1)$ is swapped with $s_{11}(z_1)$ (see Table 5). As compared to the

Table 1
Percentage biases and RMSEs for $\hat{\theta}_{22}$, $\hat{\rho}$ and $\hat{\sigma}$ obtained from the ER experiments, when employing the naive, standard Heckman (with second-order polynomial terms), penalized Bayesian and ML estimation approaches (naive, HeckP, W&K and M&R). Number of simulated datasets and approximate percentage of selected observations are 250% and 50%. True values of θ_{22} and σ are -1.5 and 1. ρ and n denote the correlation between the errors of the selection and outcome equations, and the sample size. See Section 3 for further details.

ρ		$\hat{\theta}_{22}$						$\hat{\rho}$						$\hat{\sigma}$					
		Bias (%)			RMSE			Bias (%)			RMSE			Bias (%)			RMSE		
		n																	
		500	1500	3000	500	1500	3000	500	1500	3000	500	1500	3000	500	1500	3000	500	1500	3000
0.1	Naive	5.2	6.2	6.2	0.184	0.135	0.115	–	–	–	–	–	–	–	–	–	–	–	–
	HeckP	−0.5	1.7	0.1	0.390	0.220	0.163	−7.0	−25.4	−4.3	0.328	0.198	0.144	−1.0	−0.2	0.0	0.065	0.028	0.021
	W&K	2.7	3.5	1.5	0.320	0.207	0.150	−55.8	−54.0	−25.1	0.260	0.181	0.132	−1.3	−0.6	−0.8	0.108	0.059	0.045
	M&R	5.6	6.0	2.6	0.398	0.253	0.170	−102.6	−95.9	−46.1	0.365	0.239	0.157	1.2	0.3	−0.1	0.058	0.028	0.020
0.5	Naive	31.9	32.3	32.4	0.502	0.492	0.490	–	–	–	–	–	–	–	–	–	–	–	–
	HeckP	2.0	3.8	3.2	0.363	0.214	0.161	−12.3	−13.4	−11.3	0.288	0.185	0.136	−0.3	0.4	0.5	0.084	0.044	0.030
	W&K	8.0	3.7	2.1	0.320	0.173	0.126	−27.5	−12.0	−6.9	0.280	0.150	0.099	−5.0	−2.1	−1.5	0.136	0.076	0.055
	M&R	4.8	2.2	1.2	0.371	0.175	0.124	−17.2	−7.5	−4.3	0.332	0.151	0.096	−0.6	−0.7	−0.6	0.062	0.035	0.026
0.9	Naive	59.0	58.8	58.8	0.893	0.885	0.884	–	–	–	–	–	–	–	–	–	–	–	–
	HeckP	6.6	6.2	5.6	0.338	0.199	0.158	−14.8	−10.4	−8.9	0.225	0.144	0.117	1.7	1.8	1.7	0.103	0.063	0.048
	W&K	1.6	0.5	0.3	0.171	0.099	0.072	−2.8	−0.8	−0.3	0.070	0.029	0.019	−2.1	−0.8	−0.6	0.119	0.070	0.050
	M&R	−0.9	−0.5	−0.3	0.164	0.094	0.067	1.2	0.5	0.3	0.048	0.023	0.016	−0.3	−0.1	−0.2	0.056	0.032	0.023

Table 2
RMSEs and 95% average coverage probabilities for $\hat{s}_{21}(z_1)$ obtained from the ER experiments, when employing naive, HeckP, W&K and M&R. Number of simulated datasets and approximate percentage of selected observations are 250% and 50%. See the caption of Table 1 for further details.

ρ		$\hat{s}_{21}(z_1)$					
		RMSE			ACP		
		n					
		500	1500	3000	500	1500	3000
0.1	Naive	0.121	0.075	0.059	–	–	–
	HeckP	0.143	0.079	0.056	0.97	0.97	0.96
	W&K	0.137	0.085	0.066	0.98	0.97	0.97
	M&R	0.164	0.099	0.069	0.96	0.96	0.95
0.5	Naive	0.200	0.185	0.178	–	–	–
	HeckP	0.135	0.076	0.055	0.97	0.97	0.97
	W&K	0.134	0.078	0.058	0.97	0.97	0.97
	M&R	0.146	0.080	0.056	0.96	0.96	0.96
0.9	Naive	0.314	0.311	0.309	–	–	–
	HeckP	0.124	0.073	0.056	0.98	0.97	0.98
	W&K	0.099	0.065	0.049	0.98	0.98	0.97
	M&R	0.101	0.062	0.043	0.96	0.96	0.95

previous case, this scenario better highlights the advantage of penalized regression splines over polynomial models. Overall, results show that HeckP does not model adequately regressor effects. Specifically, for any value of ρ and n, the HeckP RMSE of $\hat{s}_{11}(z_1)$ is consistently higher than that of M&R. In addition, the residual confounding induced by the mismodeled non-linear effects seems to have negative consequences on the estimation of the parametric effects, where HeckP underperforms in terms of accuracy and precision.

- We also carried out additional simulation experiments where the model errors are generated according to a bivariate Student-t distribution with 3 degrees of freedom; see Tables 9 and 10 in Appendix D. As expected, results are worse than those presented in this section. However, the use of ER helps to obtain better estimates although still not as good as those produced when the assumption (3) is met. Even worse results (available upon request) are found when using asymmetric or bimodal bivariate model error distributions, case in which the presence of ER cannot really help. The reason for this result is that the likelihood of the model is wrongly specified (i.e., we assume normality but the model errors are generated from a Student-t distribution) and ML approaches are known to be sensitive to such issues.

In summary, methods which cannot control for non-random sample selection (such as naive) produce severely biased estimates. The two penalized (Bayesian and ML) regression spline approaches to sample selection modeling are effective and generally outperform standard Heckman with polynomial terms. Moreover, although W&K and M&R perform similarly, the former is more time-consuming than the latter. Finally, ER is generally required to obtain good estimation results. Of course, in the presence of severe model misspecification, ER cannot avoid obtaining considerably biased estimates.

Table 3
Percentage biases and RMSEs for $\hat{\theta}_{22}$, $\hat{\rho}$ and $\hat{\sigma}$ obtained from the non-ER experiments, when employing the naive, standard Heckman (with second-order polynomial terms), penalized Bayesian and ML estimation approaches (naive, HeckP, W&K and M&R). Number of simulated datasets and approximate percentage of selected observations are 250% and 75%. See the caption of Table 1 for further details.

ρ		$\hat{\theta}_{22}$						$\hat{\rho}$						$\hat{\sigma}$					
		Bias (%)			RMSE			Bias (%)			RMSE			Bias (%)			RMSE		
		n																	
		500	1500	3000	500	1500	3000	500	1500	3000	500	1500	3000	500	1500	3000	500	1500	3000
0.1	Naive	3.1	3.3	3.4	0.136	0.088	0.072	–	–	–	–	–	–	–	–	–	–	–	–
	HeckP	−0.1	0.4	−0.6	0.297	0.173	0.122	−24.8	−26.0	3.0	0.431	0.253	0.188	−1.6	−0.5	−0.2	0.065	0.029	0.019
	W&K	1.8	1.9	0.6	0.200	0.145	0.099	−60.8	−59.8	−25.7	0.278	0.218	0.150	−3.9	−1.4	−1.2	0.099	0.051	0.040
	M&R	2.4	2.2	0.7	0.253	0.159	0.105	−83.4	−70.4	−29.3	0.378	0.252	0.168	1.1	0.5	0.1	0.047	0.023	0.016
0.5	Naive	17.1	17.5	17.5	0.285	0.271	0.267	–	–	–	–	–	–	–	–	–	–	–	–
	HeckP	1.6	1.6	0.7	0.290	0.171	0.118	−28.0	−20.6	−15.2	0.408	0.253	0.187	−0.8	0.1	0.2	0.080	0.038	0.029
	W&K	6.9	3.0	1.8	0.233	0.138	0.090	−47.1	−24.0	−15.5	0.365	0.210	0.139	−7.0	−3.2	−1.8	0.130	0.071	0.051
	M&R	5.1	1.3	0.8	0.261	0.126	0.083	−36.7	−14.4	−11.1	0.415	0.190	0.119	−0.4	−0.4	−0.5	0.050	0.028	0.021
0.9	Naive	31.2	31.6	31.6	0.481	0.478	0.477	–	–	–	–	–	–	–	–	–	–	–	–
	HeckP	2.3	3.0	1.9	0.272	0.159	0.117	−25.4	−19.6	−15.7	0.367	0.249	0.199	0.8	1.5	1.2	0.096	0.056	0.045
	W&K	2.3	0.5	0.4	0.159	0.078	0.055	−14.0	−8.1	−7.8	0.205	0.090	0.080	−4.8	−1.5	−1.7	0.130	0.060	0.045
	M&R	0.4	−0.1	0.1	0.147	0.073	0.053	−7.6	−6.3	−6.6	0.174	0.068	0.065	−0.8	−0.4	−0.5	0.048	0.025	0.019

Table 4
RMSEs and 95% average coverage probabilities for $\hat{s}_{21}(z_1)$ obtained from the non-ER experiments, when employing naive, HeckP, W&K and M&R. Number of simulated datasets and approximate percentage of selected observations are 250% and 75%. See the caption of Table 1 for further details.

ρ		$\hat{s}_{21}(z_1)$					
		RMSE			ACP		
		n					
		500	1500	3000	500	1500	3000
0.1	Naive	0.093	0.057	0.041	–	–	–
	HeckP	0.112	0.065	0.046	0.97	0.96	0.96
	W&K	0.105	0.070	0.050	0.98	0.97	0.96
	M&R	0.114	0.071	0.047	0.96	0.95	0.95
0.5	Naive	0.126	0.104	0.097	–	–	–
	HeckP	0.108	0.063	0.045	0.98	0.97	0.97
	W&K	0.105	0.070	0.047	0.98	0.97	0.96
	M&R	0.113	0.063	0.042	0.96	0.95	0.96
0.9	Naive	0.181	0.170	0.166	–	–	–
	HeckP	0.100	0.060	0.043	0.98	0.98	0.98
	W&K	0.087	0.058	0.042	0.98	0.98	0.97
	M&R	0.089	0.051	0.039	0.95	0.96	0.95

Table 5
Percentage biases and RMSEs for $\hat{\theta}_{22}$, $\hat{\rho}$, $\hat{\sigma}$ and $\hat{s}_{11}(z_1)$ obtained from the ER experiments, when employing the standard Heckman (with second-order polynomial terms) and ML estimation approaches (HeckP and M&R). Number of simulated datasets and approximate percentage of selected observations are 250 and 50%. See Section 3 for further details.

		$\hat{\theta}_{22}$				$\hat{\rho}$				$\hat{\sigma}$				$\hat{s}_{11}(z_1)$	
		Bias (%)		RMSE		Bias (%)		RMSE		Bias (%)		RMSE		RMSE	
		n													
		500	3000	500	3000	500	3000	500	3000	500	3000	500	3000	500	3000
0.1	HeckP	6.0	7.1	0.644	0.272	−87.7	−87.4	0.425	0.194	−2.5	−1.5	0.071	0.030	1.564	1.564
	M&R	−0.1	−0.3	0.662	0.256	−30.3	−5.0	0.359	0.168	−0.9	0.0	0.064	0.021	1.526	1.528
0.5	HeckP	8.6	7.8	0.584	0.257	258.7	296.4	0.444	0.335	−1.4	−0.8	0.069	0.029	1.569	1.565
	M&R	−0.5	−0.6	0.573	0.241	−21.7	−5.3	0.335	0.147	−0.6	0.0	0.081	0.030	1.511	1.523
0.9	HeckP	9.3	8.0	0.561	0.233	607.7	694.7	0.664	0.705	−0.3	0.0	0.091	0.034	1.572	1.566
	M&R	2.4	−1.5	0.510	0.211	−18.6	−2.9	0.282	0.077	−0.2	−0.3	0.103	0.044	1.501	1.519

4. Empirical illustration

The method presented in this paper as well as naive, standard Heckman with polynomial terms and W&K are illustrated using data from the RAND Health Insurance Experiment (RHIE) which was a comprehensive study of health care cost,

Table 6
Description of the outcome and selection variables, and of the regressors.

Variable	Definition
`lnmeddol`	log of the medical expenses of the individual (*outcome variable*)
`binexp`	binary variable indicating whether the medical expenses are positive (*selection variable*)
`logc`	log of the coinsurance rate (`coins`) plus 1
`idp`	binary variable for individual deductible plans
`pi`	participation incentive payment
`fmde`	is 0 if `idp` $= 1$, and log [max {1, maximum expenditure offer/(0.01 $*$ `coins`)}] otherwise
`physlm`	physical limitations
`disea`	number of chronic diseases
`hlthg`	binary variable for good self-rated health (the baseline is excellent self-rated health)
`hlthf`	binary variable for fair self-rated health
`hlthp`	binary variable for poor self-rated health
`inc`	family income
`fam`	family size
`educdec`	education of household head in years
`xage`	age of the individual in years
`female`	binary variable for female individuals
`child`	binary variable for individuals younger than 18 years
`fchild`	binary variable for female individuals younger than 18 years
`black`	binary variable for black household heads

utilization and outcome conducted in the United States between 1974 and 1982 (Newhouse, 1999). As explained in the introductory section, the aim was to quantify the relationship between various covariates and annual health expenditures in the population as a whole.

In this context, non-random sample selection arises because the sample consisting of individuals who used health care services differ in important characteristics from the sample of individuals who did not use them. Because some characteristics cannot be observed, traditional regression modeling is likely to deliver biased estimates. We, therefore, need to correct parameter estimates for sample selection bias. We use the same subsample as in Cameron and Trivedi (2005, p. 553), and model annual health expenditures. The sample size and number of selected observations are 5574 and 4281. The variables are defined in Table 6. Additional information can be found in Cameron and Trivedi (2005, Table 20.4) and Newhouse (1999).

Following Cameron and Trivedi (2005) the outcome and the selection equations include the same set of regressors. In M&R and W&K, the two equations include `logc`, `idp`, `fmde`, `physlm`, `disea`, `hlthg`, `hlthf`, `hlthp`, `female`, `child`, `fchild` and `black` as parametric components, and smooth functions of `pi`, `inc`, `fam`, `educdec` and `xage`, represented using P-spline bases with 20 inner knots and penalty matrices based on second order differences. For naive, the same model specification is adopted but clearly a selection equation is not present. As for standard Heckman, we model the effects of `pi`, `inc`, `fam`, `educdec` and `xage` using second-order polynomials. For W&K, the number of iterations for burn-in, of samples used for estimation, and degree of thinning were the same as those employed in the simulation study.

The use of smooth functions for `xage`, `educdec` and `inc` is suggested by the fact that these covariates embody productivity and life-cycle effects that are likely to influence health expenditures non-linearly. Dismuke and Egede (2011) and Sullivan et al. (2007) consider parametric specifications where non-linear effects are modeled by categorizing these variables into groups based on intervals. However, categorizing a continuous variable has several disadvantages since, for example, it introduces problems of defining cut-points and assumes a priori that the relationship between response and covariate is flat within intervals (e.g., Marra and Radice, 2010). As for `fam` and `pi`, we do not have a priori knowledge of their effects and imposing linear or quadratic relationships may prevent us from revealing interesting non-linear relationships. Smooth functions of other covariates such as `idp` and `disea` are not considered as their number of unique covariate values is too small.

Table 7 and Fig. 2 report the parametric and smooth function estimates for the outcome equation (which is the one of interest) when applying the four approaches on the RAND RHIE dataset. Computational times for M&R and W&K were 1.03 and 18.26 min, respectively.

The parametric effects obtained using HeckP, M&R and W&K are very similar (except for the effect of `child`) and differ from those of naive. Specifically, the M&R and W&K results suggest that socioeconomic factors (`black`, `female`, `child`, and `fchild`) as well as health status variables (`physlm`, `disea`, `hlthg`, `hlthf`, `hlthp`) have a stronger effect on annual health expenses as compared to the naive results. The health insurance variables (`logc`, `idp`) seem not to determine the annual medical expenses when using naive. The estimated smooths for the socioeconomic variables (`inc`, `fam`, `educdec` and `xage`) obtained with M&R and W&K are reasonably close. This is not true for the health insurance variable `pi` where all four estimated functions are different. Notice how the estimated curves produced by HeckP are consistently U- or inverted U-shaped; this is an artifact of the polynomial specification. The estimates of ρ, which are important to ascertain the presence of selection bias, are high, positive, and statistically significant. This indicates that the unobservable factors which lead individuals to use health services also lead them to spend more on medical expenses. The estimates of σ_2 obtained with either HeckP or M&R and W&K are significantly different, with the latter showing a larger uncertainty. This result is consistent

Table 7
Parametric estimates of the annual medical expense equation obtained by applying the naive, standard Heckman (with second-order polynomial terms), penalized Bayesian and ML estimation approaches (naive, HeckP, W&K and M&R, respectively) on the RAND RHIE dataset described in Section 4. Within parentheses are 95% confidence and credible intervals. For M&R, intervals have been calculated using the result of Section 2.3.

Variable	Naive	HeckP	W&K	M&R
(Intercept)	3.75 (3.58, 3.92)	3.52 (3.35, 3.68)	3.31 (3.11, 3.52)	3.28 (3.08, 3.48)
logc	−0.02 (−0.08, 0.04)	−0.07 (−0.14, 0.00)	−0.07 (−0.13, −0.00)	−0.07 (−0.14, −0.00)
idp	−0.11 (−0.23, 0.02)	−0.18 (−0.32, −0.05)	−0.18 (−0.31, −0.06)	−0.17 (−0.30, −0.04)
fmde	−0.03 (−0.06, 0.01)	−0.02 (−0.06, 0.02)	−0.02 (−0.05, 0.02)	−0.02 (−0.06, 0.02)
physlm	0.23 (0.11, 0.38)	0.36 (0.21, 0.51)	0.31 (0.17, 0.46)	0.33 (0.18, 0.48)
disea	0.02 (0.01, 0.03)	0.03 (0.02, 0.03)	0.03 (0.02, 0.04)	0.03 (0.02, 0.04)
hlthg	0.17 (0.07, 0.26)	0.18 (0.08, 0.28)	0.21 (0.11, 0.31)	0.20 (0.10, 0.30)
hlthf	0.39 (0.22, 0.56)	0.44 (0.25, 0.63)	0.46 (0.29, 0.65)	0.47 (0.28, 0.65)
hlthp	0.78 (0.45, 1.11)	0.96 (0.60, 1.33)	0.99 (0.62, 1.36)	0.97 (0.61, 1.34)
female	0.34 (0.23, 0.45)	0.55 (0.43, 0.68)	0.52 (0.41, 0.67)	0.54 (0.41, 0.66)
child	0.05 (−0.28, 0.39)	−0.46 (−0.70, −0.23)	0.11 (−0.24, 0.49)	0.17 (−0.19, 0.54)
fchild	−0.34 (−0.52, −0.17)	−0.57 (−0.76, −0.38)	−0.53 (−0.75, −0.35)	−0.54 (−0.73, −0.35)
black	−0.20 (−0.33, −0.07)	−0.52 (−0.66, −0.37)	−0.47 (−0.62, −0.33)	−0.52 (−0.67, −0.37)
ρ	–	0.73 (0.65, 0.79)	0.69 (0.57, 0.76)	0.72 (0.63, 0.78)
σ_2	–	1.56 (1.51, 1.62)	2.32 (2.17, 2.54)	1.54 (1.49, 1.60)

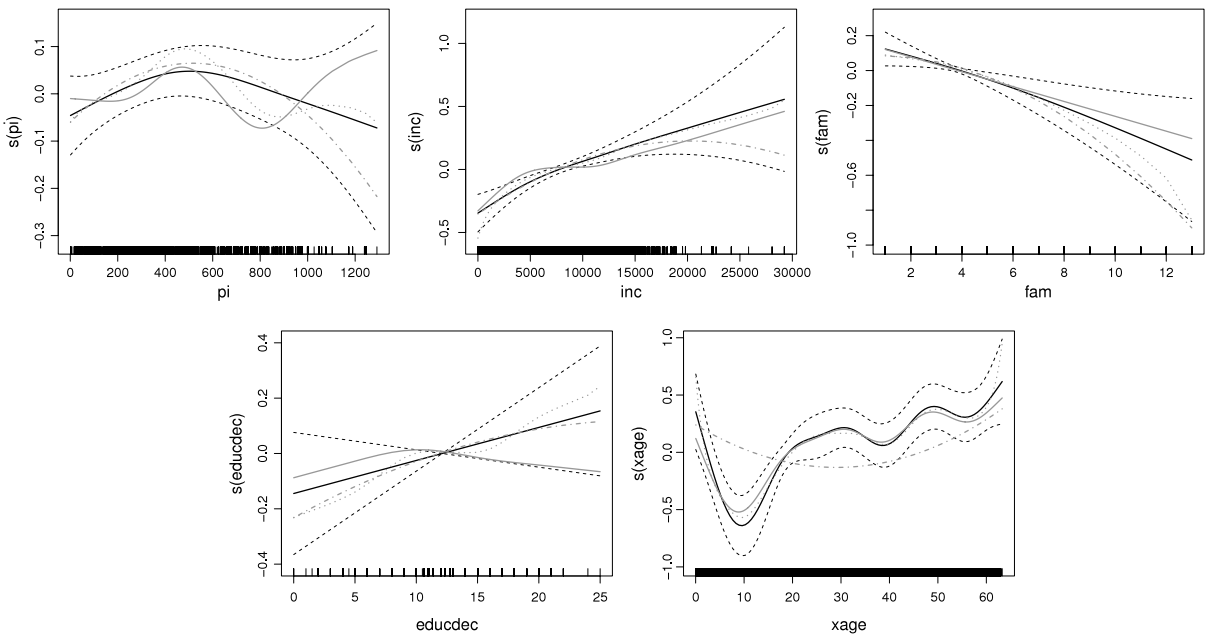

Fig. 2. Smooth function estimates obtained by applying naive (gray lines), HeckP (gray dot-dashed lines), W&K (gray dotted lines) and M&R (black lines) on the RAND RHIE dataset described in Section 4. The black dashed lines represent 95% pointwise confidence intervals calculated from the M&R estimates. The 'rug plot', at the bottom of each graph, shows the covariate values. To avoid clutter, credible intervals for W&K have not been reported. Due to the identifiability constraints, the estimated curves are centered around zero.

with that found in the simulated non-ER scenario with percentage of selected observations equal to 75%. Overall, the M&R and W&K estimates are coherent with the predictions of economic theory. For example, the results of age, education and income are consistent with the interpretation that health expenditure increases as people become older, have more years of schooling, and are wealthier. Also, individual health expenditure decreases as family size increases.

The reliability of the results presented in this section relies on whether the assumption of normality is met. As noted by Cameron and Trivedi (2005, p. 555), the underlying normality is suspect for these data because of the presence of large outliers. Testing this assumption is especially important when ER is not present, as in this case. Without ER and under violation of the assumption of normality, selection bias correction fails (e.g., Vella, 1998). It would be therefore ideal to test for normality. Cameron and Trivedi (2005) checked this assumption by applying standard tests of heteroskedasticity, skewness and kurtosis on the outcome variable `lnmeddol`. However, in a regression context, normality should be assessed more rigorously. For example, in the current case, 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 Grami–Charlier series with 9 additional parameters (e.g., Chiburis, 2010, Lee, 1984). However, it is not entirely clear how this test can be extended to the penalized likelihood framework considered in this paper. Another possibility would be to exploit the fact that a penalized regression

spline is approximately equivalent to a pure regression spline with degrees of freedom close to that of the penalized fit (e.g., Wood, 2006, pp. 210–212). This topic is beyond the scope of this paper and will be addressed in future research.

5. Conclusions

We introduced an algorithm to estimate a regression spline sample selection model for Gaussian data. The proposal is based on the penalized likelihood estimation framework. The construction of confidence intervals has also been illustrated, and the problem of identification has been discussed. The method has been tested and compared to a Bayesian counterpart and the classic Heckman sample selection model. Finally, the proposed approach and its competitors have been illustrated on data from the RAND Health Insurance Experiment on annual health expenditures. The R package `SemiParSampleSel` (Marra and Radice, 2012) implements the ideas discussed in this article.

The results of our simulation study highlighted the detrimental effects that the neglect of non-random sample selection has on parameter estimation. They also suggested that the two Bayesian and ML regression spline approaches considered in this article are effective and generally outperform standard Heckman with polynomials. The Bayesian and ML methods were found to perform similarly, with the former being more computationally expensive than the latter. We also found that ER is generally required to obtain good estimation results.

Because ML estimators are sensitive to model error misspecification, methods allowing for different bivariate distributions of the errors can be developed. For example, Marchenko and Genton (2012) introduced a sample selection model where the errors are assumed to follow a bivariate Student-t distribution. However, in their implementation the structure of the linear predictor is parametrically pre-specified. The proposed approach could be extended by adopting either a copula (e.g., Nelsen, 2006) or a nonparametric distribution function estimation framework. Future research will be conducted toward these directions.

Acknowledgments

Giampiero Marra was supported by the Engineering and Physical Sciences Research Council (grant EP/J006742/1). We are indebted to the Editor, Associate Editor and two reviewers for the detailed comments, which helped us improve the manuscript and clarify the main messages.

Appendix A. Analytical expressions for g and $\mathcal{H}$

The expressions for the gradient vector and Hessian matrix that are referred to in Section 2.2 are given below. Let us define $\mathbf{X}_{vi} = \left(\mathbf{u}_{vi}^{\top}, \mathbf{B}_{vi}^{\top}\right)$ for $v = 1, 2$, $\sigma_2 = \exp(\sigma_2^*)$, $\rho = \tanh(\rho^*)$, $e_{2i} = y_{2i} - \eta_{2i}$, $a = \sqrt{1-\rho^2}$, $A_i = (\eta_{1i} + \frac{\rho}{\sigma_2} e_{2i})/a$, $l_{1i} = \phi(-\eta_{1i})/\Phi(-\eta_{1i})$, $l_{2i} = \phi(A_i)/\Phi(A_i)$, $e_c = \exp(2\rho^*)$, $PA_i = -\left\{\Phi(A_i)\phi(A_i)A_i + \phi(A_i)^2\right\}/\Phi(A_i)^2$, $PE_i = -\left\{-\Phi(-\eta_{1i})\phi(-\eta_{1i})\eta_{1i} + \phi(-\eta_{1i})^2\right\}/\Phi(-\eta_{1i})^2$, $R = \{4\rho e_c(\rho - 1)\}/(e_c + 1)$, $M_i = (2e_c e_{2i})/\{(e_c + 1)\sigma_2 a\}$ and $C = -1 + \rho + \left\{\rho^2(\rho - 1)\right\}/a^2$. The remaining quantities are defined in Section 2.

The elements of the score vector are

$$\mathbf{g}_1 = \sum_{i=1}^{n} \left\{ -(1 - y_{1i})\, l_{1i} + \frac{y_{1i} l_{2i}}{a} \right\} \mathbf{X}_{1i},$$

$$\mathbf{g}_2 = \sum_{i=1}^{n} y_{1i} \left(\frac{e_{2i}}{\sigma_2^2} - \frac{l_{2i}\rho}{\sigma_2 a} \right) \mathbf{X}_{2i},$$

$$g_3 = \sum_{i=1}^{n} y_{1i} \left\{ -1 + \left(\frac{e_{2i}}{\sigma_2} \right)^2 - \frac{l_{2i}\rho e_{2i}}{\sigma_2 a} \right\},$$

$$g_4 = \sum_{i=1}^{n} y_{1i} \left[l_{2i} \left\{ M_i(1 - \rho) - \frac{A_i R}{2a^2} \right\} \right].$$

The elements of the Hessian are

$$\mathcal{H}_{11} = \sum_{i=1}^{n} \left\{ (1 - y_{1i})\, PE_i + \frac{y_{1i} PA_i}{a^2} \right\} \mathbf{X}_{1i}^{\top} \mathbf{X}_{1i},$$

$$\mathcal{H}_{12} = \sum_{i=1}^{n} \left(-\frac{y_{1i} PA_i \rho}{\sigma_2 a^2} \right) \mathbf{X}_{1i}^{\top} \mathbf{X}_{2i},$$

$$\mathcal{H}_{13} = \sum_{i=1}^{n} \left(-\frac{y_{1i} PA_i \rho e_{2i}}{\sigma_2 a^2} \right) \mathbf{X}_{1i},$$

$$\mathcal{H}_{14} = \sum_{i=1}^{n}\left[y_{1i}\left\{\frac{PA_i}{a}\left\{M_i\left(1-\frac{e_c-1}{e_c+1}\right)-\frac{A_iR}{2a^2}\right\}-\frac{l_{2i}R}{2a^3}\right\}\right]\mathbf{X}_{1i},$$

$$\mathcal{H}_{22} = \sum_{i=1}^{n}\left[y_{1i}\left\{PA_i\left(\frac{\rho}{\sigma_2 a}\right)^2-\frac{1}{\sigma_2}\right\}\right]\mathbf{X}_{2i}^{\mathsf{T}}\mathbf{X}_{2i},$$

$$\mathcal{H}_{23} = \sum_{i=1}^{n}\left\{y_{1i}\left(PA_ie_{2i}\left(\frac{\rho}{\sigma_2 a}\right)^2+\frac{\rho l_{2i}}{\sigma_2 a}-\frac{2e_{2i}}{\sigma_2^2}\right)\right\}\mathbf{X}_{2i},$$

$$\mathcal{H}_{24} = \sum_{i=1}^{n}\left[y_{1i}\left\{-\frac{PA_i}{(e_c+1)\,\sigma_2 a}\left\{(e_c-1)\left(M_i(1-\rho)-\frac{A_iR}{2a^2}\right)\right\}+\frac{2l_{2i}e_cC}{(e_c+1)\,\sigma_2 a}\right\}\right]\mathbf{X}_{2i},$$

$$\mathcal{H}_{33} = \sum_{i=1}^{n}\left[y_{1i}\left\{-2\left(\frac{e_{2i}}{\sigma_2}\right)^2+\frac{l_{2i}e_{2i}\rho}{a\sigma_2}+PA_i\left(\frac{\rho e_{2i}}{\sigma_2 a}\right)^2\right\}\right],$$

$$\mathcal{H}_{34} = \sum_{i=1}^{n}\left[y_{1i}\left\{-\frac{PA_i}{(e_c+1)\,\sigma_2 a}\left\{(e_c-1)\,e_{2i}\left(M_i(1-\rho)-\frac{A_iR}{2a^2}\right)\right\}-l_{2i}M_iC\right\}\right],$$

$$\mathcal{H}_{44} = \sum_{i=1}^{n}\left[y_{1i}\left(PA_i\left(M_i(1-\rho)-\frac{A_iR}{2a^2}\right)^2+l_{2i}\left\{2M_i\left(1-\frac{2e_c}{e_c+1}+\frac{2e_c\rho}{e_c+1}-\rho\right)-\frac{M_i(1-\rho)R}{a^2}+\frac{3A_iR^2}{4a^4}\right.\right.$$
$$\left.\left.-\frac{A_i}{2a^2}\frac{8e_c}{e_c+1}\left(-\frac{e_c^2}{e_c+1}+\frac{4\rho e_c}{e_c+1}-\rho-\frac{3\rho^2e_c}{e_c+1}+\rho^2\right)\right\}\right)\right].$$

Appendix B. Starting value procedure

Sensible starting values can be provided by adapting Heckman's approach (1979) to the regression spline context. Regression function (5) can be written as

$$\mathbb{E}(y_{2i}|y_{1i}^*>0) = \mathbf{u}_{2i}^{\mathsf{T}}\boldsymbol{\theta}_2+\mathbf{B}_{2i}^{\mathsf{T}}\boldsymbol{\beta}_2+\mathbb{E}(\varepsilon_{2i}|y_{1i}^*>0). \tag{12}$$

It then follows that

$$\mathbb{E}(\varepsilon_{2i}|y_{1i}^*>0) = \theta_\vartheta\,\vartheta_i, \tag{13}$$

where $\theta_\vartheta = \sigma_2\rho$, $\vartheta_i = \phi(\eta_{1i})/\Phi(\eta_{1i})$ (the inverse Mills ratio) and $\eta_{1i} = \mathbf{u}_{1i}^{\mathsf{T}}\boldsymbol{\theta}_1+\mathbf{B}_{1i}^{\mathsf{T}}\boldsymbol{\beta}_1$. Therefore, Eq. (12) can be written as

$$y_{2i} = \mathbf{u}_{2i}^{\mathsf{T}}\boldsymbol{\theta}_2+\mathbf{B}_{2i}^{\mathsf{T}}\boldsymbol{\beta}_2+\theta_\vartheta\,\vartheta_i+\widetilde{\varepsilon}_{2i}, \tag{14}$$

where $\widetilde{\varepsilon}_{2i}$ is a new disturbance term which, by construction, is uncorrelated with $\mathbf{u}_{2i}$, $\mathbf{B}_{2i}$ and ϑ_i. The coefficient estimates in (5) that would be obtained using a non-random selected subsample are biased if $\rho\neq 0$. This can be seen as an ordinary specification error with the conditional mean (13) deleted as a regressor in the model. Including ϑ_i as an explanatory variable, as in Eq. (14), would in principle rectify this situation. But ϑ_i is unknown; however it is possible to obtain a consistent estimate of it using the estimated coefficients of selection Eq. (4).

The two-step procedure to fit model (14) can be summarized as follows.

step 1 Fit a probit model for Eq. (4) and obtain estimates of $\widehat{\eta}_{1i}$ and $\widehat{\vartheta}_i$, for all i.
step 2 Using the selected sample only, fit model (14), where ϑ_i is replaced with $\widehat{\vartheta}_i$, for all i.

The correlation parameter ρ can be estimated by $\widehat{\rho} = \widehat{\theta}_\vartheta/\widehat{\sigma}_2$, where $\widehat{\sigma}_2 = \sqrt{\sum_{i=1}^{n_s}\widehat{\widetilde{\varepsilon}}_{2i}^2/n_s+\widehat{\theta}_\vartheta^2\sum_{i=1}^{n_s}\widehat{\gamma}_i/n_s}$, $\widehat{\widetilde{\varepsilon}}_{2i}$ is the residual resulting from estimation of (14) and $\widehat{\gamma}_i = \widehat{\vartheta}_i(\widehat{\vartheta}_i+\widehat{\eta}_{1i})$ (e.g., Toomet and Henningsen, 2008). Note that since $\widehat{\rho}$ can be outside of $[-1, 1]$, this quantity is truncated to stay within this range. Moreover, although the parameters of the models in the two steps can be estimated using an unpenalized procedure, this is not advisable in practice (see Section 2). Therefore, the models in the two-step procedure are estimated by maximization of a penalized log-likelihood function and by minimization of a penalized least squares criterion, respectively. Standard statistical software is available to achieve this (Ruppert et al., 2003; Wood, 2006).

In principle, because of the non-linearity of the inverse Mills ratio and the use of flexible covariate effects, the parameters of the two-step procedure are identified even if $(\mathbf{u}_{1i}^{\mathsf{T}}, \mathbf{B}_{1i}^{\mathsf{T}}) = (\mathbf{u}_{2i}^{\mathsf{T}}, \mathbf{B}_{2i}^{\mathsf{T}})$. However, since it is typically the case that $\widehat{\vartheta}_i$ can be approximated well by a linear function of the covariates in the model, there will be substantial collinearity between $\widehat{\vartheta}_i$ and the regressors in the outcome equation, which can affect parameter estimation. This will be especially the case when the range of values of η_{1i} is not very large. The presence of ER can alleviate this problem (see, e.g., Leung and Yu, 2000 for other remedies). We do not elaborate on this further as the presented two-step approach just serves as a starting value procedure.

Appendix C. Algorithm structure

Based on the methods presented in Section 2.2, parameter vector $\boldsymbol{\delta}$ is estimated using the following algorithm structure.

step 1 For a given $\boldsymbol{\lambda}$, find an estimate of $\boldsymbol{\delta}$:

$$\widehat{\boldsymbol{\delta}} = \underset{\boldsymbol{\delta}}{\operatorname{argmax}}\ \ell_p(\boldsymbol{\delta}).$$

step 2 Iterate the following steps until convergence:

step 2.1 For fixed $\boldsymbol{\lambda}^{[a]}$, $\sigma_2^{*[a]}$ and $\rho^{*[a]}$, find an estimate of $(\boldsymbol{\delta}_1, \boldsymbol{\delta}_2)$:

$$\left(\widehat{\boldsymbol{\delta}}_1^{[a+1]}, \widehat{\boldsymbol{\delta}}_2^{[a+1]}\right) = \underset{(\boldsymbol{\delta}_1, \boldsymbol{\delta}_2)}{\operatorname{argmax}} \ell_p(\boldsymbol{\delta}_1, \boldsymbol{\delta}_2, \sigma_2^{*[a]}, \rho^{*[a]}).$$

step 2.2 Using $\left(\widehat{\boldsymbol{\delta}}_1^{[a+1]}, \widehat{\boldsymbol{\delta}}_2^{[a+1]}\right)$, construct the working linear model quantities needed in (9) and find an estimate of $\boldsymbol{\lambda}$:

$$\widehat{\boldsymbol{\lambda}}^{[a+1]} = \arg\min_{\boldsymbol{\lambda}}\ \mathcal{V}_u^w(\boldsymbol{\lambda}).$$

step 2.3 For fixed $\boldsymbol{\lambda}^{[a+1]}$, $\sigma_2^{*[a]}$ and $\rho^{*[a]}$ find an estimate of $(\boldsymbol{\delta}_1, \boldsymbol{\delta}_2)$:

$$\left(\widehat{\boldsymbol{\delta}}_1^{[a+2]}, \widehat{\boldsymbol{\delta}}_2^{[a+2]}\right) = \underset{(\boldsymbol{\delta}_1, \boldsymbol{\delta}_2)}{\operatorname{argmax}}\ \ell_p(\boldsymbol{\delta}_1, \boldsymbol{\delta}_2, \sigma_2^{*[a]}, \rho^{*[a]}).$$

step 2.4 For fixed $\boldsymbol{\lambda}^{[a+1]}$ and $\left(\boldsymbol{\delta}_1^{[a+2]}, \boldsymbol{\delta}_2^{[a+2]}\right)$, find an estimate of $\left(\sigma_2^*, \rho^*\right)$:

$$\left(\widehat{\sigma}_2^{*[a+1]}, \widehat{\rho}^{*[a+1]}\right) = \underset{(\sigma_2^*, \rho^*)}{\operatorname{argmax}}\ \ell(\boldsymbol{\delta}_1^{[a+2]}, \boldsymbol{\delta}_2^{[a+2]}, \sigma_2^*, \rho^*).$$

step 3 Given estimates of $\boldsymbol{\lambda}$, $(\boldsymbol{\delta}_1, \boldsymbol{\delta}_2)$ and $\left(\sigma_2^*, \rho^*\right)$, obtained at convergence of step 2, repeat step 1.

Note that steps 2.1–2.4 can be seen as *leapfrog* iterations and have good convergence properties despite $\left(\widehat{\boldsymbol{\delta}}_1, \widehat{\boldsymbol{\delta}}_2\right)$ and $\left(\widehat{\sigma}_2^*, \widehat{\rho}^*\right)$ are not orthogonal (Smith, 1996).

Appendix D. Additional simulation results

See Tables 8–10.

Table 8

Percentage biases and RMSEs for $\hat{\theta}_{22}$, $\hat{\rho}$, $\hat{\sigma}$ and $\hat{s}_{21}(z_1)$ obtained from the non-ER experiments, when employing the naive, standard Heckman (with second-order polynomial terms), penalized Bayesian and ML estimation approaches (naive, HeckP, W&K and M&R). Number of simulated datasets and approximate percentage of selected observations are 250% and 50%. True values of θ_{22} and σ are -1.5 and 1. ρ and n denote the correlation between the errors of the selection and outcome equations, and the sample size. See Section 3 for further details.

ρ		$\hat{\theta}_{22}$				$\hat{\rho}$				$\hat{\sigma}$				$\hat{s}_{21}(z_1)$	
		Bias (%)		RMSE		Bias (%)		RMSE		Bias (%)		RMSE		RMSE	
		n													
		500	3000	500	3000	500	3000	500	3000	500	3000	500	3000		
0.1	Naive	5.2	6.2	0.184	0.115	–	–	–	–	–	–	–	–	0.121	0.059
	HeckP	3.8	2.0	0.827	0.342	−76.7	−39.9	0.578	0.302	−6.8	−1.1	0.201	0.040	0.255	0.100
	W&K	8.4	5.3	0.361	0.212	−135.9	−83.3	0.286	0.188	−2.9	−1.5	0.120	0.050	0.142	0.077
	M&R	14.8	8.8	0.429	0.261	−232.7	−148.8	0.386	0.251	2.7	1.0	0.067	0.026	0.176	0.085
0.5	Naive	31.9	32.4	0.502	0.490	–	–	–	–	–	–	–	–	0.200	0.178
	HeckP	18.8	18.4	0.845	0.416	−70.7	−59.9	0.664	0.408	−3.8	1.4	0.179	0.046	0.253	0.123
	W&K	24.4	11.1	0.544	0.315	−78.4	−38.7	0.505	0.308	−10.2	−4.3	0.167	0.081	0.185	0.095
	M&R	20.5	10.6	0.581	0.311	−66.1	−30.8	0.583	0.305	−3.4	−1.0	0.063	0.030	0.196	0.095
0.9	Naive	59.0	58.8	0.893	0.884	–	–	–	–	–	–	–	–	0.314	0.309
	HeckP	35.2	33.5	0.862	0.577	−67.1	−56.2	0.813	0.572	3.7	7.4	0.147	0.122	0.262	0.179
	W&K	14.6	2.5	0.512	0.196	−30.5	−10.0	0.491	0.201	−8.5	−1.2	0.200	0.056	0.148	0.056
	M&R	13.2	2.6	0.503	0.183	−28.6	−9.2	0.472	0.191	−2.5	−0.6	0.073	0.027	0.150	0.053

Table 9

Percentage biases and RMSEs for $\hat{\theta}_{22}$, $\hat{\rho}$, $\hat{\sigma}$ and $\hat{s}_{21}(z_1)$ obtained from the ER experiments, when employing the naive, standard Heckman (with second-order polynomial terms), penalized Bayesian and ML estimation approaches (naive, HeckP, W&K and M&R). Number of simulated datasets and approximate percentage of selected observations are 250% and 50%. True values of θ_{22} and σ are -1.5 and 1. The errors were generated according to a bivariate Student-t distribution with 3 degrees of freedom. See Section 3 for further details.

ρ		$\hat{\theta}_{22}$				$\hat{\rho}$				$\hat{\sigma}$				$\hat{s}_{21}(z_1)$	
		Bias (%)		RMSE		Bias (%)		RMSE		Bias (%)		RMSE		RMSE	
		n													
		500	3000	500	3000	500	3000	500	3000	500	3000	500	3000	500	3000
0.1	Naive	6.4	8.2	0.359	0.169	–	–	–	–	–	–	–	–	0.197	0.088
	HeckP	−2.4	−2.5	0.790	0.319	−29.0	3.8	0.398	0.186	−54.2	−48.7	1.021	0.769	0.257	0.107
	W&K	2.1	1.7	0.731	0.299	−59.2	−27.6	0.353	0.169	−53.2	−48.8	1.612	0.991	0.243	0.121
	M&R	4.0	2.4	0.813	0.326	−94.1	−44.2	0.361	0.186	−53.3	−48.5	0.992	0.766	0.279	0.124
0.5	Naive	41.9	43.0	0.706	0.654	–	–	–	–	–	–	–	–	0.283	0.235
	HeckP	−6.5	−8.2	0.772	0.337	−15.8	−14.8	0.354	0.183	−52.5	−49.6	1.009	0.768	0.249	0.109
	W&K	3.3	−7.4	0.771	0.339	−48.4	−13.8	0.366	0.168	−73.9	−61.5	1.807	1.013	0.237	0.110
	M&R	1.5	−6.0	0.774	0.337	−36.0	−7.0	0.375	0.173	−50.1	−48.7	0.965	0.754	0.273	0.118
0.9	Naive	76.8	78.2	1.188	1.178	–	–	–	–	–	–	–	–	0.423	0.405
	HeckP	−9.3	−14.9	0.815	0.620	−26.8	−7.9	0.476	0.224	−50.5	−51.2	0.997	0.792	0.260	0.136
	W&K	−2.4	−13.9	0.729	0.349	−24.5	−6.1	0.311	0.062	−53.7	−48.0	1.013	0.795	0.250	0.122
	M&R	−1.5	−13.4	0.710	0.337	−20.1	0.3	0.308	0.054	−43.3	−47.8	0.853	0.737	0.253	0.119

Table 10

Percentage biases and RMSEs for $\hat{\theta}_{22}$, $\hat{\rho}$, $\hat{\sigma}$ and $\hat{s}_{21}(z_1)$ obtained from the non-ER experiments, when employing naive, HeckP, W&K and M&R. Number of simulated datasets and approximate percentage of selected observations are 250% and 50%. The errors were generated according to a bivariate Student-t distribution with 3 degrees of freedom. See the caption of Table 9 for further details.

ρ		$\hat{\theta}_{22}$				$\hat{\rho}$				$\hat{\sigma}$				$\hat{s}_{21}(z_1)$	
		Bias (%)		RMSE		Bias (%)		RMSE		Bias (%)		RMSE		RMSE	
		n													
		500	3000	500	3000	500	3000	500	3000	500	3000	500	3000	500	3000
0.1	Naive	6.4	8.2	0.359	0.169	–	–	–	–	–	–	–	–	−0.196	0.088
	HeckP	3.6	−8.5	2.337	0.905	−88.7	39.0	0.716	0.473	−84.2	−55.8	1.658	0.884	0.621	0.269
	W&K	18.3	4.1	2.139	0.844	−180.5	−59.9	0.593	0.411	−76.3	−55.9	2.173	1.431	0.533	0.267
	M&R	23.1	7.4	2.277	0.867	−224.8	−101.7	0.609	0.426	−75.4	−54.5	1.522	0.861	0.553	0.269
0.5	Naive	41.9	43.0	0.706	0.654	–	–	–	–	–	–	–	–	0.283	0.235
	HeckP	−24.2	−22.1	2.285	0.917	−44.6	−0.5	0.706	0.391	−82.3	−58.5	1.613	0.933	0.607	0.269
	W&K	−10.4	−24.4	2.022	1.033	−63.7	−19.9	0.630	0.372	−99.7	−71.5	2.343	1.447	0.511	0.296
	M&R	−7.0	−21.3	2.025	1.025	−58.0	−11.0	0.631	0.380	−69.3	−58.7	1.365	0.948	0.521	0.300
0.9	Naive	76.8	78.2	1.188	1.178	–	–	–	–	–	–	–	–	0.423	0.405
	HeckP	−22.9	−37.3	1.791	0.929	−37.8	−8.6	0.621	0.214	−66.6	−63.7	1.316	1.024	0.471	0.267
	W&K	−14.7	−58.1	1.689	1.299	−40.7	−7.1	0.577	0.143	−57.8	−69.3	1.206	1.100	0.401	0.353
	M&R	−11.4	−57.0	1.582	1.228	−35.8	0.4	0.563	0.139	−54.0	−68.4	1.061	1.111	0.406	0.349

References

Ahn, H., Powell, J.L., 1993. Semiparametric estimation of censored selection models with a nonparametric selection mechanism. Journal of Econometrics 58, 3–29.

Akaike, H., 1973. Information theory and an extension of the maximum likelihood principle. In: Petrov, B.N., Csaki, F. (Eds.), International Symposium on Information Theory. Akademiai Kiado, Budapest, pp. 267–281.

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.

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

Breslow, N.E., Clayton, D.G., 1993. Approximate inference in generalized linear mixed models. Journal of the American Statistical Association 88, 9–25.

Cameron, A.C., Trivedi, P.K., 2005. Microeconometrics: Methods and Applications. Cambridge University Press, New York.

Chib, S., Greenberg, E., Jeliazkov, I., 2009. Estimation of semiparametric models in the presence of endogeneity and sample Selection. Journal of Computational and Graphical Statistics 18, 321–348.

Chiburis, R.C., 2010. Score tests of normality in bivariate probit models: comment. Working Paper. Available at: https://webspace.utexas.edu/rcc485/www/research.html.

Craven, P., Wahba, G., 1979. Smoothing noisy data with spline functions. Numerische Mathematik 31, 377–403.

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

Das, M., Newey, W., Vella, F., 2003. Estimation of sample selection models. The Review of Economic Studies 70, 33–58.

Dismuke, C.E., Egede, L.E., 2011. Association of serious psychological distress with health services expenditures and utilization in a national sample of US adults. General Hospital Psychiatry 33, 311–317.

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

Greene, W.H., 2012. Econometric Analysis. Prentice Hall, New York.
Gu, C., 1992. Cross validating non-Gaussian data. Journal of Computational and Graphical Statistics 1, 169–179.
Gu, C., 2002. Smoothing Spline ANOVA Models. Springer-Verlag, London.
Härdle, W., Hall, P., Marron, J.S., 1988. How far are automatically chosen regression smoothing parameters from their optimum? Journal of the American Statistical Association 83, 86–95.
Hastie, T., Tibshirani, R., 1993. Varying-coefficient models. Journal of the Royal Statistical Society: Series B 55, 757–796.
Heckman, J.J., 1979. Sample selection bias as a specification error. Econometrica 47, 153–162.
Lee, L.F., 1984. Tests for the bivariate normal distribution in econometric models with selectivity. Econometrica 52, 843–863.
Lee, L.F., 1994. Semiparametric two-stage estimation of sample selection models subject to Tobit-type selection rules. Journal of Econometrics 61, 305–344.
Leung, S.F., Yu, S., 2000. Collinearity and two-step estimation of sample selection models: problems, origins, and remedies. Computational Economics 15, 173–199.
Li, P., 2011. Estimation of sample selection models with two selection mechanisms. Computational Statistics and Data Analysis 55, 1099–1108.
Marchenko, Y.V., Genton, M.G., 2012. A Heckman selection-t model. Journal of the American Statistical Association 107, 304–317.
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., 2011. Estimation of a semiparametric recursive bivariate probit model in the presence of endogeneity. Canadian Journal of Statistics 39, 259–279.
Marra, G., Radice, R., 2012. SemiParSampleSel: semiparametric sample selection modelling. R Package Version 0.1. http://cran.r-project.org/package=SemiParSampleSel.
Marra, G., Wood, S.N., 2012. Coverage properties of confidence intervals for generalized additive model components. Scandinavian Journal of Statistics 39, 53–74.
Martins, M.F.O., 2001. Parametric and semiparametric estimation of sample selection models: an empirical application to the female labour force in Portugal. Journal of Applied Econometrics 16, 23–39.
Mealli, F., Pacini, B., 2008. Comparing principal stratification and selection models in parametric causal inference with nonignorable missingness. Computational Statistics and Data Analysis 53, 507–516.
Montmarquette, C., Mahseredjian, S., Houle, R., 2001. The determinants of university dropouts: a bivariate probability model with sample selection. Economics of Education Review 20, 475–484.
Nelsen, R.B., 2006. An Introduction to Copulas. Springer-Verlag, New York.
Newey, W., Powell, J., Walker, J., 1990. Semiparametric estimation of selection models: some empirical results. The American Economic Review 80, 324–328.
Newhouse, J.P., 1999. RAND health insurance experiment [in metropolitan and non-metropolitan areas of the United States], 1974–1982. Aggregated Claims Series, 1, Codebook for Fee-for-Service Annual Expenditures and Visit Counts ICPSR 6439, ICPSR Inter-university Consortium for Political and Social Research.
Nocedal, J., Wright, S.J., 1999. Numerical Optimization. Springer-Verlag, New York.
Omori, Y., Miyawaki, K., 2010. Tobit model with covariate dependent thresholds. Computational Statistics and Data Analysis 54, 2736–2752.
Powell, J.L., 1994. Estimation of semiparametric models. In: Engle, R.F., McFadden, D.L. (Eds.), Handbook of Econometrics, Volume 4. Elsevier, pp. 2443–2521 (Chapter 41).
Puhani, P.A., 2000. The Heckman correction for sample selection and its critique. Journal of Economic Surveys 14, 53–68.
R Development Core Team, 2012. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.
Reiss, P.T., Ogden, R.T., 2009. Smoothing parameter selection for a class of semiparametric linear models. Journal of the Royal Statistical Society: Series B 71, 505–524.
Ruppert, D., Wand, M.P., Carroll, R.J., 2003. Semiparametric Regression. Cambridge University Press, London.
Sigelman, L., Zeng, L., 1999. Analyzing censored and sample-selected data with Tobit and Heckit models. Political Analysis 8, 167–182.
Silverman, B.W., 1985. Some aspects of the spline smoothing approach to non-parametric regression curve fitting. Journal of the Royal Statistical Society: Series B 47, 1–52.
Smith, G.K., 1996. Partitioned algorithms for maximum likelihood and other non-linear estimation. Statistics and Computing 6, 201–216.
Smith, M.D., 2003. Modelling sample selection using Archimedean copulas. The Econometrics Journal 6, 99–123.
Sullivan, P.W., Ghushchyan, V., Wyatt, H.R., Wu, E.Q., Hill, J.O., 2007. Productivity costs associated with cardiometabolic risk factor clusters in the United States. Value in Health 10, 443–450.
Terza, J.V., 1998. Estimating count data models with endogenous switching: sample selection and endogenous treatment effects. Journal of Econometrics 84, 129–154.
Toomet, O., Henningsen, A., 2008. Sample selection models in R: package sampleSelection. Journal of Statistical Software 27, 1–23.
van Hasselt, M., 2011. Bayesian inference in a sample selection model. Journal of Econometrics 165, 221–232.
Vella, F., 1998. Estimating models with sample selection bias: a survey. Journal of Human Resources 33, 127–169.
Wahba, G., 1983. Bayesian 'confidence intervals' for the cross-validated smoothing spline. Journal of the Royal Statistical Society: Series B 45, 133–150.
Wahba, G., 1985. A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem. The Annals of Statistics 13, 1378–1402.
Wiesenfarth, M., Kneib, T., 2010. Bayesian geoadditive sample selection models. Journal of the Royal Statistical Society: Series C 59, 381–404.
Winship, C., Mare, R.D., 1992. Models for sample selection bias. Annual Review of Sociology 18, 327–350.
Wood, S.N., 2004. Stable and efficient multiple smoothing parameter estimation for generalized additive models. Journal of the American Statistical Association 99, 673–686.
Wood, S.N., 2006. Generalized Additive Models: An Introduction with R. Chapman & Hall, London.
Yee, T.W., 2010. The VGAM package for categorical data analysis. Journal of Statistical Software 32, 1–34.