Research Article **Open Access**

Giampiero Marra* and Rosalba Radice

A joint regression modeling framework for analyzing bivariate binary data in R

https://doi.org/10.1515/demo-2017-0016
Received September 22, 2017; accepted October 18, 2017

Abstract: We discuss some of the features of the R add-on package GJRM which implements a flexible joint modeling framework for fitting a number of multivariate response regression models under various sampling schemes. In particular, we focus on the case in which the user wishes to fit bivariate binary regression models in the presence of several forms of selection bias. The framework allows for Gaussian and non-Gaussian dependencies through the use of copulae, and for the association and mean parameters to depend on flexible functions of covariates. We describe some of the methodological details underpinning the bivariate binary models implemented in the package and illustrate them by fitting interpretable models of different complexity on three data-sets.

Keywords: binary data, copula, confounding, joint model, penalized smoother, selection bias, R, simultaneous parameter estimation

MSC: 62H99, 62J02

1 Introduction

The R [43] package GJRM [Generalised Joint Regression Modelling, 34] implements a flexible joint modeling framework for fitting a number of multivariate response regression models under various sampling schemes. The package currently contains two main fitting functions: `gjrm()` which fits bivariate regression models with binary, discrete, continuous and survival margins in the presence of associated responses, endogeneity and non-random sample selection, and trivariate binary models with and without double sample selection; `gamlss()` which fits several flexible univariate regression models (this was initially designed to provide starting values for many of the joint models in the package but it was subsequently made available in the form of a proper function). This paper focuses on the case in which the user wishes to fit bivariate binary models in the presence of (i) endogeneity, (ii) non-random sample selection or (iii) partial observability. We illustrate the capabilities of such tool by fitting interpretable models of different complexity on real data. The literature on models tackling selection bias is vast and many variants have been proposed. Since our focus is on the case of binary data, all of the non-binary cases (such as those mentioned above) are not discussed here since these would deserve separate and lengthy expositions. The models considered in this article have wide applicability. Some examples are given by [9], [17], Jeliazkov and Yang [26, Chapter 13], [27], [29], [40] and [48], to name but a few. The next sections describe the three aforementioned issues and available approaches to tackle them.

***Corresponding Author: Giampiero Marra:** Department of Statistical Science, University College London, Gower Street, London WC1E 6BT, UK, E-mail: giampiero.marra@ucl.ac.uk
Rosalba Radice: Department of Economics, Mathematics and Statistics, Birkbeck, University of London, Malet Street, London WC1E 7HX, UK, E-mail: r.radice@bbk.ac.uk

1.1 Endogeneity

Quantifying the effect of a non-randomly assigned treatment on an outcome may be a challenging task in the presence of unobserved confounders (i.e., unknown or not readily quantifiable variables associated with both treatment and outcome). In this situation, the treatment is often termed endogenous and the bias resulting from neglecting unobserved confounding is typically referred to as endogenous selection bias. For the case of binary treatment and outcome, [21] introduced the bivariate probit model to address this issue (see also [18] and [30] for a gentle introduction). Alternative approaches are discussed in the excellent review by [11] which empirically shows that all methods considered (including the bivariate probit) produce very similar results. [10] and [31] extended Heckman's model by introducing Bayesian and likelihood penalized spline methods to model flexibly covariate-response relationships. To account for non-Gaussian dependence between treatment and outcome, [55] discussed the use of copulae. [44] proposed a more general approach that deals simultaneously with unobserved confounding, non-linear covariate effects and non-Gaussian dependence between treatment and outcome, and incorporated these developments in `GJRM`. Specifically, the conventional bivariate probit model (which does not model flexibly covariate effects in a data-driven manner and does not account for non-Gaussian dependencies) can be fitted in `SAS` [25] using the built-in `proc qlim` [24] and in `Stata` [52] using the built-in functions `biprobit` [51], `mvprobit` [8] and `ssm` [37]. In `R`, `VGAM` [60] may be used to estimate a bivariate binary model with endogenous treatment and non-linear covariate effects but it relies on the assumption of Gaussian errors. `mvProbit` [23] may also be employed but it is under development.

1.2 Non-random sample selection

Survey data (but not only) are often affected by systematic non-participation. This can occur through a variety of mechanisms, including directly declining to participate in the study. If individuals select themselves into (or out of) the sample based on a combination of observed and unobserved characteristics then models that ignore such a mechanism will most likely yield estimates which are not representative of the population of interest. The bias arising from neglecting this systematic non-participation is typically known as non-random sample selection bias. Selection and pattern-mixture models can deal with this issue, even when selection is based on unobserved characteristics of respondents; Fitzmaurice et al. [15, Chapter 18] provide a discussion of the features and variants of both approaches. Here the focus is on the sample selection model approach which was introduced by [19], [28] and [20], and discussed more thoroughly in [22]. When the outcome is binary, the conventional selection model is a bivariate probit [13, 54]. [35] proposed, and incorporated in `GJRM`, a selection model which allows for Gaussian and non-Gaussian dependencies, arbitrary parametric link functions, and for the association and mean parameters to depend on several types of smooth functions of covariates; this work extended the scope of the approaches introduced by [32] and [36]. More restrictive bivariate selection models, which rely on normality and on linear or pre-specified non-linear covariate-response relationships, can be fitted in `SAS` using the `proc qlim`, in `Stata` using the built-in commands `heckprob` [51], `mvprobit` and `ssm`, and in `R` using `sampleSelection` [53].

1.3 Partial observability

In some cases an observed binary outcome reflects the joint realization of the unobserved choices of two decision-makers. If this is not accounted for then partial observability bias will arise. The bivariate probit with partial observability acknowledges this by assuming that the model which determines the observed outcome is a bivariate probit in which only one of the four possible outcomes is observed. This model was first introduced by [41] and mainly consists of two equations (describing the underlying unobserved binary outcomes) which are linked through a standard bivariate Gaussian where the correlation coefficient captures the presence of unobservables influencing the two decision-makers. [42] discussed multivariate extensions and

applications of this model. The bivariate binary model with partial observability can be fitted in `Stata` using `biprobit`. In this work we have extended the model to include flexible covariate effects, and incorporated it in `GJRM`.

The paper is organized as follows. Section 2 introduces a general modeling framework for analyzing bivariate binary data. Section 3 then discusses in more detail the binary models considered in this paper. Section 4 provides an overview of `gjrm()` in `GJRM`, whereas Section 5 is devoted to three data examples that illustrate the use of the software. Section 6 concludes the paper.

2 Methodology

Let us assume that there are two binary random variables (Y_{i1}, Y_{i2}), for $i = 1, \ldots, n$, where n represents the sample size. The probability of event $(Y_{i1} = 1, Y_{i2} = 1)$ can be defined as

$$p_{11i} = P(Y_{i1} = 1, Y_{i2} = 1) = C(P(Y_{i1} = 1), P(Y_{i2} = 1); \theta_i),$$

where $P(Y_{ij} = 1) = 1 - F_j(-\eta_{ji})$ for $j = 1, 2$, $F_j(\cdot)$ is the cumulative distribution function (cdf) of a standardized univariate distribution (in this case Gaussian, logistic or Gumbel), $\eta_{ji} \in \mathbb{R}$ is an additive predictor (refer to (2) below), C is a two-place copula function [49, 50] and θ_i is an association parameter measuring the dependence between the two random variables. The notation adopted for defining $P(Y_{ij} = 1)$ is perhaps unusual. However, here we have exploited the link between the binary regression model and $Y^{\star}_{ij} = \eta_{ji} + \epsilon_i$, where $Y^{\star}_{ij}$ is a continuous latent variable, ϵ_i is an error term and Y_{ij} can be viewed as an indicator for $Y^{\star}_{ij} > 0$. Therefore, $P(Y_{ij} = 1) = P(Y^{\star}_{ij} > 0) = 1 - F_j(-\eta_{ji})$. The marginal cdfs are conditioned on covariates through η_{1i} and η_{2i}, but for notational convenience we have suppressed this when expressing them. Since the strength and direction of the association between the two marginals may, for instance, vary across groups of observations, the dependence parameter is specified as a function of an additive predictor. That is, $\theta_i = m(\eta_{ci})$, where m is a one-to-one transformation which ensures that θ_i lies in its range. This approach follows the same rationale of [45], who introduced generalized additive models for location, scale and shape, where all the parameters characterizing a chosen distribution are related to predictors via suitable link functions. The copulae implemented in `GJRM`, corresponding ranges of θ and list of transformations $m(\cdot)$ are reported in Table 1 of [33]. Rotations by 90°, 180° and 270° are also implemented; for example, rotating the Clayton, Gumbel and Joe by 90° and 270° allows these copulae to model negative dependence. Parameter θ is often not easy to interpret, in which case the well known Kendall's $\tau \in [-1, 1]$ can be employed. For full details on copulae see, for instance, [38].

The log-likelihood function of the sample can be expressed as

$$\ell = \sum_{i=1}^{n} \left\{ \sum_{a,b=0}^{1} \mathbb{1}_{abi} \log(p_{abi}) \right\}, \tag{1}$$

where $\mathbb{1}_{abi}$ is an indicator function equal to one when $(y_{i1} = a, y_{i2} = b)$ is true, $a, b \in \{0, 1\}$, and y_{i1} and y_{i2} are realizations of Y_{i1} and Y_{i2}, respectively.

2.1 Additive predictor specification

Let us consider a generic additive predictor η_{vi} and an overall covariate vector $\mathbf{z}_{vi}$. Here, subscript v can take values 1, 2 (which refer to the first and second margins) and c (which refer to the copula parameter). The predictor can be defined as

$$\eta_{vi} = \beta_{v0} + \sum_{k_v=1}^{K_v} s_{vk_v}(\mathbf{z}_{vk_vi}), \; i = 1, \ldots, n, \tag{2}$$

where $\beta_{v0} \in \mathbb{R}$ is an overall intercept, $\mathbf{z}_{vk_vi}$ denotes the k_v^{th} sub-vector of the complete vector $\mathbf{z}_{vi}$ and the K_v functions $s_{vk_v}(\mathbf{z}_{vk_vi})$ represent generic effects which are chosen according to the type of covariate(s) considered. Each $s_{vk_v}(\mathbf{z}_{vk_vi})$ can be approximated as a linear combination of Q_{vk_v} basis functions $b_{vk_vq_{vk_v}}(\mathbf{z}_{vk_vi})$ and regression coefficients $\beta_{vk_vq_{vk_v}} \in \mathbb{R}$, i.e.

$$s_{vk_v}(\mathbf{z}_{vk_vi}) = \sum_{q_{vk_v}=1}^{Q_{vk_v}} \beta_{vk_vq_{vk_v}} b_{vk_vq_{vk_v}}(\mathbf{z}_{vk_vi}). \tag{3}$$

Equation (3) implies that the vector of evaluations $\left\{s_{vk_v}(\mathbf{z}_{vk_v1}), \ldots, s_{vk_v}(\mathbf{z}_{vk_vn})\right\}^\mathsf{T}$ can be written as $\mathbf{Z}_{vk_v}\boldsymbol{\beta}_{vk_v}$ with $\boldsymbol{\beta}_{vk_v} = (\beta_{vk_v1}, \ldots, \beta_{vk_vQ_{vk_v}})^\mathsf{T}$ and design matrix $\mathbf{Z}_{vk_v}[i, q_{vk_v}] = b_{vk_vq_{vk_v}}(\mathbf{z}_{vk_vi})$. This allows the predictor in equation (2) to be written as

$$\boldsymbol{\eta}_v = \beta_{v0}\mathbf{1}_n + \mathbf{Z}_{v1}\boldsymbol{\beta}_{v1} + \ldots + \mathbf{Z}_{vK_v}\boldsymbol{\beta}_{vK_v}, \tag{4}$$

where $\mathbf{1}_n$ is an n-dimensional vector made up of ones. Equation (4) can also be written in a more compact way as $\boldsymbol{\eta}_v = \mathbf{Z}_v\boldsymbol{\beta}_v$, where $\mathbf{Z}_v = (\mathbf{1}_n, \mathbf{Z}_{v1}, \ldots, \mathbf{Z}_{vK_v})$ and $\boldsymbol{\beta}_v = (\beta_{v0}, \boldsymbol{\beta}_{v1}^\mathsf{T}, \ldots, \boldsymbol{\beta}_{vK_v}^\mathsf{T})^\mathsf{T}$. Each $\boldsymbol{\beta}_{vk}$ has an associated quadratic penalty $\lambda_{vk_v}\boldsymbol{\beta}_{vk_v}^\mathsf{T}\mathbf{D}_{vk_v}\boldsymbol{\beta}_{vk_v}$, used in fitting, whose role is to enforce specific properties on the k_v^{th} function, such as smoothness. For the case of a smooth function of a continuous regressor $\mathbf{z}_{vk_v}$, $\mathbf{D}_{vk_v}$ may be calculated as $\int \mathbf{d}_{vk_v}(z_{vk_v})\mathbf{d}_{vk_v}(z_{vk_v})^\mathsf{T} dz_{vk_v}$, where the $q_{vk_v}^{th}$ element of $\mathbf{d}_{vk_v}(z_{vk_v})$ is given by $\partial^2 b_{vk_vq_{vk_v}}(\mathbf{z}_{vk_v})/\partial z_{vk_v}^2$ and integration is over the range of z_{vk_v}. See, for instance, [33] for other examples. The smoothing parameter $\lambda_{vk_v} \in [0, \infty)$ controls the trade-off between fit and smoothness, and plays a crucial role in determining the shape of $\hat{s}_{vk_v}(\mathbf{z}_{vk_vi})$. Let us consider again the case of a smooth effect of a continuous variable. A value of $\lambda_{vk_v} = 0$ (i.e., no penalization is imposed during fitting) will result in an un-penalized regression spline estimate which will most likely over-fit the data, while $\lambda_{vk_v} \to \infty$ (i.e., the penalty has a large influence on the smooth function) will lead to a straight line estimate. The overall penalty can be defined as $\boldsymbol{\beta}_v^\mathsf{T}\mathbf{D}_v\boldsymbol{\beta}_v$, where $\mathbf{D}_v = \text{diag}(0, \lambda_{v1}\mathbf{D}_{v1}, \ldots, \lambda_{vK_v}\mathbf{D}_{vKv})$. The smoothing parameters can be collected in vector $\boldsymbol{\lambda}_v = (\lambda_{v1}, \ldots, \lambda_{vK_v})^\mathsf{T}$. Finally, smooth functions are typically subject to centering (identifiability) constraints (see [58] for more details).

The above formulation allows one to employ a rich variety of covariate effects. Specifically, `GJRM` can accommodate all terms available in `mgcv` [59], which include smooth functions of continuous covariates, smooth interactions between continuous and/or discrete variables, random effect smoothers and spatial smoothers for data sampled over a large portion of the globe or for geographic areas with complicated boundaries. These are incorporated in our modeling framework by specifying the appropriate $\mathbf{Z}_{vk_v}$ and $\mathbf{D}_{vk_v}$. Otherwise, the construction of the additive predictors and overall smoothing penalty remains essentially unchanged.

2.2 Parameter estimation

Our model specification allows for a high degree of flexibility in modeling covariate effects. If an unpenalised approach is employed to estimate the model's paramters then over-fitting is the likely consequence [e.g., 46]. To prevent this, we maximize

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

where ℓ_p is the penalized model's log-likelihood, $\boldsymbol{\delta}^\mathsf{T} = (\boldsymbol{\beta}_1^\mathsf{T}, \boldsymbol{\beta}_2^\mathsf{T}, \boldsymbol{\beta}_c^\mathsf{T})$ and $\mathbf{S} = \text{diag}(\mathbf{D}_1, \mathbf{D}_2, \mathbf{D}_c)$. The smoothing parameter vectors are collected in the overall vector $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_1^\mathsf{T}, \boldsymbol{\lambda}_2^\mathsf{T}, \boldsymbol{\lambda}_c^\mathsf{T})^\mathsf{T}$. In practice, estimation of $\boldsymbol{\delta}$ and $\boldsymbol{\lambda}$ is achieved by using a stable and efficient trust region algorithm (based on first and second order analytical derivative information) with integrated automatic multiple smoothing parameter selection [35, 44].

2.3 Further considerations

At convergence, point-wise 'confidence' intervals for linear and non-linear functions of the model's coefficients can be obtained using the Bayesian large sample approximation

$$\boldsymbol{\delta} \dot{\sim} \mathcal{N}(\hat{\boldsymbol{\delta}}, \mathbf{V}_{\boldsymbol{\delta}}), \tag{6}$$

where $\hat{\boldsymbol{\delta}}$ is a parameter vector estimate, $\mathbf{V}_{\boldsymbol{\delta}} = -\boldsymbol{H}_p(\hat{\boldsymbol{\delta}})^{-1}$ and $\boldsymbol{H}_p$ is the penalized model's Hessian. Intervals derived using (6) have good frequentist properties since they account for both sampling variability and smoothing bias; see [35] and references therein for details. Intervals for non-linear functions of the model's coefficients (e.g., τ, joint and conditional predicted probabilities) can be conveniently obtained by simulation from the posterior distribution of $\boldsymbol{\delta}$ using the following steps:

1. Draw n_{sim} random vectors from $\mathcal{N}(\hat{\boldsymbol{\delta}}, \mathbf{V}_{\boldsymbol{\delta}})$.
2. Calculate n_{sim} simulated realizations of the quantity of interest. For instance, for a Gaussian copula $\tau_i = \frac{2}{\pi} \arcsin \left[\tanh \left\{\eta_{ci}(\mathbf{Z}_{ci}; \boldsymbol{\beta}_c)\right\}\right]$ in which case $\boldsymbol{\tau}_i^{sim} = (\tau_i^{sim_1}, \tau_i^{sim_2}, \ldots, \tau_i^{sim_{n_{sim}}})^\intercal \; \forall i = 1, \ldots, n$ is obtained using $\boldsymbol{\beta}_c^{sim_j} \; \forall j = 1, \ldots, n_{sim}$.
3. For each $\boldsymbol{\tau}_i^{sim}$, calculate the lower, $\varsigma/2$, and upper, $1 - \varsigma/2$, quantiles.

A small value for n_{sim}, say 100, typically gives accurate results, whereas ς is usually set to 0.05.

Model building in our framework involves the choice of copula function, of pair of link functions and selection of relevant covariates in the model's additive predictors. To this end, we recommend using the Akaike information criterion (AIC) and/or the Bayesian information criterion (BIC), and hypothesis testing. The AIC and BIC are given by $-2\ell(\hat{\boldsymbol{\delta}}) + 2edf$ and $-2\ell(\hat{\boldsymbol{\delta}}) + \log(n)edf$, where the log-likelihood is evaluated at the penalized parameter estimates and edf represents the effective degrees of freedom (see [35] for the exact definition). Approximate p-values for testing smooth components for equality to zero are obtained using the results by Wood [56] and Wood [57].

3 The models

In the following sections, we describe in some detail the three models that can deal with the issues of (i) endogeneity, (ii) non-random sample selection and (iii) partial observability, discuss their additive predictor specifications and log-likelihood functions. We also report some typical measures of interest. For more details on the models dealing with (i) and (ii), the reader is to referred to [44] and [35].

3.1 Bivariate binary model with endogenous treatment

A bivariate binary model with endogenous treatment is mainly employed when one is interested in estimating the effect of a binary treatment on a binary outcome in the presence of unobserved confounding. In economics, this problem is commonly framed in terms of a regression model from which important covariates have been omitted and hence become a part of the model's error term. In this context, the treatment is termed exogenous if it is not associated with the error term after conditioning on the observed confounders, and endogenous otherwise. The bivariate model controls for unobserved confounding by using a two-equation structural latent variable framework where one equation essentially describes a binary outcome (say Y_{i2}) as a function of a binary treatment (Y_{i1}) whereas the other equation determines whether the treatment is received. The model is completed by including covariates and by assuming that the latent errors of the two equations follow a bivariate distribution with association parameter θ_i. Using the notation introduced in Section 2.1,

the additive predictors for this model can be expressed as

$$\begin{aligned}\boldsymbol{\eta}_1 &= \beta_{10}\mathbf{1}_n + \mathbf{Z}_{11}\boldsymbol{\beta}_{11} + \ldots + \mathbf{Z}_{1K_1}\boldsymbol{\beta}_{1K_1} = \mathbf{Z}_1\boldsymbol{\beta}_1,\\ \boldsymbol{\eta}_2 &= \beta_{20}\mathbf{1}_n + \beta_{21}\mathbf{y}_1 \quad + \ldots + \mathbf{Z}_{2K_2}\boldsymbol{\beta}_{2K_2} = \mathbf{Z}_2\boldsymbol{\beta}_2,\\ \boldsymbol{\eta}_c &= \beta_{c0}\mathbf{1}_n + \mathbf{Z}_{c1}\boldsymbol{\beta}_{c1} + \ldots + \mathbf{Z}_{cK_c}\boldsymbol{\beta}_{cK_c} = \mathbf{Z}_c\boldsymbol{\beta}_c,\end{aligned}$$

whereas log-likelihood function (1) becomes

$$\ell = \sum_{i=1}^{n} \left\{ \mathbb{1}_{11i} \log(p_{11i}) + \mathbb{1}_{10i} \log(p_{10i}) + \mathbb{1}_{01i} \log(p_{01i}) + \mathbb{1}_{00i} \log(p_{00i}) \right\},$$

where $p_{10i} = \left\{1 - F_1(-\eta_{1i})\right\} - p_{11i}$, $p_{01i} = \left\{1 - F_2(\eta_{2i})\right\} - p_{11i}$ and $p_{00i} = 1 - p_{11i} - p_{10i} - p_{01i}$.

The effect of the treatment Y_{i1} on the probability that $Y_{i2} = 1$ is typically of primary interest. That is, the aim is to investigate how the treatment changes the expected outcome. Thus, the treatment effect is given by the difference between the expected outcome with treatment and the expected outcome without treatment. Different measures of treatment effect have been proposed in the literature. Here, we employ the average treatment effect in the specific sample at hand, rather than in the population at large [SATE; 1]. This can be defined as

$$\text{SATE}(\boldsymbol{\beta}_2, \mathbf{Z}_{2i}) = \frac{1}{n} \sum_{i=1}^{n} \left\{ P(Y_{i2} = 1 | Y_{i1} = 1) - P(Y_{i2} = 1 | Y_{i1} = 0) \right\},$$

where $P(Y_{i2} = 1|Y_{i1} = 1) = 1 - F_2(-\eta_{2i}^{(y_{i1}=1)})$, $P(Y_{i2} = 1|Y_{i1} = 0) = 1 - F_1(-\eta_{2i}^{(y_{i1}=0)})$ and $\eta_{2i}^{(y_{i1}=a)}$ represents the additive predictor evaluated at $y_{i1} = a$, for a equal to 1 or 0. $\text{SATE}(\boldsymbol{\beta}_2, \mathbf{Z}_{2i})$ can be estimated using $\text{SATE}(\hat{\boldsymbol{\beta}}_2, \mathbf{Z}_{2i})$, whereas an interval for it can be obtained by employing Bayesian posterior simulation as explained in Section 2.3. Linear and non-linear effects of covariates on the propensities or probabilities that certain events occur can be also be easily obtained using the functions available in the package (e.g., `jc.probs()`).

3.2 Bivariate binary model with non-random sample selection

Non-random sample selection occurs when individuals select themselves into (or out of) the sample based on a combination of observed and unobserved characteristics. Models that ignore such a systematic selection may yield estimates which are not representative of the population of interest. One way to deal with this issue is to use a bivariate binary selection model which controls for non-random sample selection by using a two-equation structural latent variable framework where one equation describes the selection process (Y_{i1}) and the other describes the outcome Y_{i2}. Specifically, Y_{i1} indicates whether an individual is selected into the sample whereas Y_{i2} is the outcome which is observed only if the individual is selected. Similarly to the endogenous model, the errors of the two equations are assumed to follow a bivariate distribution with association parameter θ_i. In this case, the first additive predictor is the same as that defined in the previous section and the remaining ones look like

$$\begin{aligned}\boldsymbol{\eta}_2 &= \beta_{20}\mathbf{1}_{n_s} + \mathbf{Z}_{21}\boldsymbol{\beta}_{21} + \ldots + \mathbf{Z}_{2K_2}\boldsymbol{\beta}_{2K_2} = \mathbf{Z}_2\boldsymbol{\beta}_2,\\ \boldsymbol{\eta}_c &= \beta_{c0}\mathbf{1}_{n_s} + \mathbf{Z}_{c1}\boldsymbol{\beta}_{c1} + \ldots + \mathbf{Z}_{cK_c}\boldsymbol{\beta}_{cK_c} = \mathbf{Z}_c\boldsymbol{\beta}_c,\end{aligned}$$

where $\mathbf{1}_{n_s}$ is an n_s-dimensional vector made up of ones corresponding to the selected observations, and $\mathbf{Z}_2$ and $\mathbf{Z}_c$ have n_s rows. The log-likelihood function of the sample is

$$\ell = \sum_{i=1}^{n} \left\{ \mathbb{1}_{11i} \log(p_{11i}) + \mathbb{1}_{10i} \log(p_{10i}) + (1 - y_{i1}) \log(p_{0i}) \right\},$$

where $p_{0i} = F_1(-\eta_{1i})$.

The proportion of a population found to have a condition (i.e., prevalence) may be of interest. This is given as $P(Y_2 = 1)$ which can be estimated by

$$\text{PREV}(\hat{\boldsymbol{\beta}}_2, \mathbf{Z}_2) = \frac{\sum_{i=1}^{n} w_i \left\{1 - F_2(\hat{\eta}_{2i})\right\}}{\sum_{i=1}^{n} w_i},$$

where the w_i are survey weights. An interval for the prevalence can be derived using posterior simulation. Covariate impacts on $P(Y_2 = 1)$ or other probabilities of interest can also be obtained. Sample selection models typically require a valid exclusion restriction for empirical model identification (i.e., a variable which predicts selection but not the outcome).

3.3 Bivariate probit model with partial observability

This model tackles a problem in which an observed binary outcome reflects the joint realization of two unobserved binary outcomes. In other words, it is only possible to observe the product of two binary variables which means that $Y_{i1}Y_{i2} = 1$ only if $Y_{i1} = Y_{i2} = 1$ and 0 otherwise. Therefore, the joint event $(Y_{i1} = 1, Y_{i2} = 1)$ has probability p_{11i} whereas all the other events have probability $1 - p_{11i}$. In this paper, we extend Poirier's model to allow for the possibility of estimating flexibly various types of covariate effects. Additive predictors $\boldsymbol{\eta}_1$ and $\boldsymbol{\eta}_3$ are the same as those defined in Section 3.1 whereas the second predictor is defined as

$$\boldsymbol{\eta}_2 = \beta_{20}\mathbf{1}_n + \mathbf{Z}_{21}\boldsymbol{\beta}_{21} + \ldots + \mathbf{Z}_{2K_2}\boldsymbol{\beta}_{2K_2}.$$

The log-likelihood function can be written as

$$\ell = \sum_{i=1}^{n} \{\mathbb{1}_{11i} \log(p_{11i}) + (1 - \mathbb{1}_{11i}) \log(1 - p_{11i})\}. \tag{7}$$

Quantities of interest include estimates for p_{11i} and the impacts the covariates have on these probabilities. Note that this model is defined using Gaussian margins and a Gaussian copula [41].

The non-linearity of (7) provides local identification of the model parameters, except in certain cases which are problem specific and usually involve peculiar exogenous variable configurations [41]. Because interchanging $\boldsymbol{\eta}_1$ and $\boldsymbol{\eta}_2$ would give an observationally equivalent model (this was termed by Poirier the 'labelling' problem), the equations for the two underlying responses are typically distinguished by introducing at least one exclusion restriction on the covariates. If the unobservable variables influencing both outcomes are uncorrelated then the model can be simplified by assuming a priori that $\theta = 0$ [2], which would in turn imply that $p_{11i} = \{1 - F_1(-\eta_{1i})\}\{1 - F_2(-\eta_{2i})\}$.

4 The function `gjrm()` in the `R` package `GJRM`

The `GJRM` package is available at http://CRAN.R-project.org/package=GJRM and its main function is `gjrm()` which can be employed to fit the three main types of bivariate binary models described in this paper. The function can be called using

```
gjrm(formula, data = list(), ...)
```

where `formula` is a list of two compulsory equations and an optional extra formula for the dependence parameter, and `data` is a data frame, list or environment containing the variables in the model. These are `glm` like formulae except that smooth terms can be included in the equations in the same way as for `gam()` in `mgcv` (see the documentation of `mgcv`). An example of specification for the equations of a bivariate binary model with varying association parameter is

```
list(y1 ~ as.factor(x1) + s(x2, bs = "cr"),
     y2 ~ s(x3, bs = "tp"),
        ~ s(x4, bs = "mrf") )
```

where `y1` and `y2` are the two binary responses, `x1` is a categorical predictor, and the `s` terms represent smooth functions of predictors `x2`, `x3` and `x4`. Argument `bs` specifies the type of spline basis which has to be se-

lected depending on the nature of the varible considered; some of the possible choices are `cr` (cubic regression spline), `tp` (thin plate regression spline, the default), `re` (random effect) and `mrf` (Markov random field smoother). Bivariate smoothing can be achieved using `s(x2, x3, bs = "te")`, for instance. For more details and smooth term options see documentation of `mgcv`.

Important arguments of `gjrm()` are `Model` which indicates the type of model the user wishes to employ (`"B"` for the bivariate model with or without endogenous variable, `"BSS"` for the bivariate sample selection model, `"BPO"` for the partial observability model, `"BPO0"` for the partial observability model with zero correlation coefficient), `BivD` which denotes the bivariate distribution linking the two model equations (the list of possibilities include `"N"` (default), `"C0"`, `"J0"`, `"F"`, `"G0"`, `"G180"`, etc.) and `margins` which indicates the link functions (`"probit"`, `"logit"`, `"cloglog"`). Further details can be found in the help file of `gjrm`.

The package contains several post-estimation functions whose aim is to provide interpretable numerical and graphical summaries. The functions include:

- `AT(x, nm.end, type = "joint", n.sim = 100, prob.lev = 0.05, hd.plot = FALSE, ...)`. This function takes a fitted `gjrm` object `x` and calculates the SATE of a binary endogenous treatment, with corresponding interval obtained using posterior simulation. `nm.end` denotes the name of the binary endogenous predictor of interest, whereas `type` can take three possible values: `"naive"` (the effect is calculated ignoring the presence of observed and unobserved confounders), `"univariate"` (the effect is obtained from the univariate model which neglects the presence of unobserved confounders) and `"joint"` (the effect is obtained from the simultaneous model which accounts for observed and unobserved confounders). Arguments `n.sim` and `prob.lev` indicate the number of coefficient vectors simulated from the posterior distribution of the estimated model parameters, and overall probability of the left and right tails of the simulated SATE distribution to be used for interval calculations. If `hd.plot = TRUE` then a plot containing the histogram and kernel density estimate of the simulated SATE distribution is produced.
- `conv.check(x)` provides some information about the convergence of the algorithm.
- `prev(x, sw = NULL, type = "joint", n.sim = 100, prob.lev = 0.05, hd.plot = FALSE, ...)`. This function calculates the prevalence using sample selection model estimates, with corresponding interval obtained using posterior simulation. Many of the arguments of `prev()` are the same as those of `AT()`. `sw` allows for the use of survey weights.
- `plot(x, eq, ...)`. This function takes a fitted `gjrm` object and plots the estimated smooth functions of `eq` (`1`, `2` or `3`). This function is a wrapper for `plot.gam()` in `mgcv` to which we refer the reader for further details and options.
- `polys.map(lm, z, scheme = "gray", lab = "", zlim, rev.col = TRUE, ...)`. This function produces a map with geographic regions defined by polygons. `lm` is a named list of matrices; each matrix has two columns and each matrix row defines the vertex of a boundary polygon. `z` is a vector of values associated with each area of `lm`, `scheme` indicates how to fill the polygons in accordance with the values of `z` (possible options are `"heat"`, `"terrain"`, `"topo"`, `"cm"` and `"gray"`), `lab` is a label for the plot, `zlim` indicates the range to use for `z` (if missing then `zlim = range(pretty(z))`), and `rev.col` indicates whether the coloring scheme should be reversed. This function is essentially the same as `polys.plot()` in `mgcv` but with the added arguments `zlim` and `rev.col` and a wider set of choices for `scheme`.
- `predict(object, eq, ...)`. This function takes a fitted `gjrm` object and produces predictions for `eq` using a new set of values of the model covariates (`newdata`) or the original values used for the model fit. This function is a wrapper for `predict.gam()` in `mgcv`.
- `summary(object, n.sim = 100, prob.lev = 0.05, ...)`. This function produces some summaries from a fitted `gjrm` object and returns a list including, for instance, summary tables for the parametric and nonparametric components of the model equations and interval(s) for θ_i. `n.sim` and `prob.lev` have the same definitions as those for `AT()`.

5 Examples

The modeling framework is illustrated in the next sections using three data-sets: Medical Expenditure Panel Survey (MEPS) of 2008, a data-set based on the real HIV 2007 Zambian Demographic and Health Survey (DHS), and a data-set on civil war onset from [14]'s seminal study. The data and code used for the analyses below are available in the `GJRM` package (see the documentations of `meps`, `hiv` and `war`).

5.1 Impact of private health insurance on utilization of health services

We consider a case study which uses data from the 2008 MEPS (http://www.meps.ahrq.gov/) and whose goal is to estimate the effect of having private health insurance on the probability of using health care services. Private health insurance status is an important determinant of the use of health services and is a potentially endogenous variable. This is because unobserved variables, such as allergy and risk aversiveness, are likely to influence both health service utilization and private insurance decision. Sometimes the effect of private health insurance can be interpreted as adverse selection or moral hazard [e.g., 7]. Adverse selection occurs when individuals with a greater demand for medical care (because of poor health, for instance) are expected to have a greater demand for insurance. Moral hazard refers to the tendency of people to be more inclined to seek health services and doctors to be more inclined to refer them when all costs are covered. The matter is further complicated by the fact that the effects of observed confounders, such as age and education, may be complex since they embody productivity and life-cycle effects that are likely to influence private health insurance and health care utilization non-linearly. If these relationships are mismodeled then the effect of insurance on the probability of using health services may be biased. Moreover, insurance status and health care utilization may exhibit a non-Gaussian association [55].

The 2008 MEPS data-set includes information on demographics, individual health status, health care utilization and private health insurance coverage. The data-set considers individuals aged between 18 and 64 years old. Individuals that did not have a complete set of socioeconomic and demographic control variables were excluded from the sample (e.g., missing values for education or income). After exclusions, the final data-set contains 18592 observations. Table 1 in the Appendix summarizes the variables used in the analysis. The choice of these variables was motivated largely by the findings reported in previous related studies. See [47], and references therein, for further details.

We load `GJRM`, read the data-set and specify the treatment and outcome equations by including smooth functions for `bmi`, `income`, `age` and `education`.

```
R> library("GJRM")
R> data("meps", package = "GJRM")
R> treat.eq <- private ~ s(bmi) + s(income) + s(age) + s(education) +
+                        as.factor(health) + as.factor(race) +
+                        as.factor(limitation) + as.factor(region) +
+                        gender  + hypertension + hyperlipidemia + diabetes
R> out.eq <- visits.hosp ~ private + s(bmi) + s(income) + s(age) +
+                          s(education) + as.factor(health) +
+                          as.factor(race) + as.factor(limitation) +
+                          as.factor(region) + gender + hypertension +
+                          hyperlipidemia + diabetes
```

We estimate several copula models with endogenous treatment, where the bivariate distributions are chosen so that positive dependence is allowed for. This is because the models based on the Gaussian and Frank copulae suggest that the dependence between the outcomes is positive, therefore it would not make sense to employ copulae which allow for negative association when the data do not support this [e.g., 44].

```
R> f.list <- list(treat.eq, out.eq)
R> mr     <- c("probit", "probit")
R> bpN    <- gjrm(f.list, data = meps, Model = "B",
                  margins = mr)
R> bpF    <- gjrm(f.list, data = meps, BivD = "F", Model = "B",
                  margins = mr)
R> bpC0   <- gjrm(f.list, data = meps, BivD = "C0", Model = "B",
                  margins = mr)
R> bpC180 <- gjrm(f.list, data = meps, BivD = "C180", Model = "B",
                  margins = mr)
R> bpG0   <- gjrm(f.list, data = meps, BivD = "G0", Model = "B",
                  margins = mr)
R> bpG180 <- gjrm(f.list, data = meps, BivD = "G180", Model = "B",
                  margins = mr)
```

`conv.check()` can be used to check convergence. For instance,

```
R> conv.check(bpC180)

Largest absolute gradient value: 1.532329e-09
Observed information matrix is positive definite
Eigenvalue range: [0.3580183,4.533308e+13]

Trust region iterations before smoothing parameter estimation: 8
Loops for smoothing parameter estimation: 3
Trust region iterations within smoothing loops: 6
```

Based on the AIC the preferred model is the survival Clayton copula.

```
R> AIC(bpN, bpF, bpC0, bpC180, bpG0, bpG180)

             df      AIC
bpN    72.27753 30737.89
bpF    72.01256 30740.67
bpC0   72.20412 30743.14
bpC180 71.71113 30730.36
bpG0   72.22814 30731.85
bpG180 72.20412 30743.14
```

We then try several combinations of link functions but the results indicate that probit links are adequate in this case.

```
R> bpC180.1 <- gjrm(f.list, data = meps, BivD = "C180", Model = "B",
+              margins = c("logit", "logit"))
R> bpC180.2 <- gjrm(f.list, data = meps, BivD = "C180", Model = "B",
+              margins = c("logit", "cloglog"))
R> bpC180.3 <- gjrm(f.list, data = meps, BivD = "C180", Model = "B",
+              margins = c("logit", "probit"))
R> bpC180.4 <- gjrm(f.list, data = meps, BivD = "C180", Model = "B",
+              margins = c("cloglog", "probit"))

R> AIC(bpC180, bpC180.1, bpC180.2, bpC180.3, bpC180.4)
```

```
               df      AIC
bpC180   71.71113 30730.36
bpC180.1 71.47172 30746.34
bpC180.2 71.29488 30771.21
bpC180.3 71.49500 30737.17
bpC180.4 72.66715 30761.16
```

We can now look at the results.

```
R> set.seed(1)
R> summary(bpC180)

COPULA:   180 Clayton
MARGIN 1: Bernoulli
MARGIN 2: Bernoulli

EQUATION 1
Link function for mu.1: probit
Formula: private ~ s(bmi) + s(income) + s(age) + s(education) +
    as.factor(health) + as.factor(health) + as.factor(race) +
    as.factor(limitation) + as.factor(region) + gender + hypertension +
    hyperlipidemia + diabetes

Parametric coefficients:
                         Estimate Std. Error z value Pr(>|z|)
(Intercept)             0.2869055  0.0560439   5.119 3.07e-07 ***
as.factor(health)6     -0.0601848  0.0290515  -2.072  0.03830 *
as.factor(health)7     -0.1367727  0.0307420  -4.449 8.62e-06 ***
as.factor(health)8     -0.3176189  0.0426018  -7.456 8.95e-14 ***
as.factor(health)9     -0.4971039  0.0680706  -7.303 2.82e-13 ***
as.factor(race)3       -0.0349365  0.0285036  -1.226  0.22032
as.factor(race)4       -0.2296738  0.1093789  -2.100  0.03575 *
as.factor(race)5        0.0911878  0.0418697   2.178  0.02941 *
as.factor(limitation)6  0.1393116  0.0440623   3.162  0.00157 **
as.factor(region)3      0.2773288  0.0376227   7.371 1.69e-13 ***
as.factor(region)4      0.0811418  0.0326925   2.482  0.01307 *
as.factor(region)5      0.0183348  0.0349897   0.524  0.60027
gender                 -0.0004841  0.0221287  -0.022  0.98255
hypertension            0.0693851  0.0304990   2.275  0.02291 *
hyperlipidemia          0.1589956  0.0306816   5.182 2.19e-07 ***
diabetes               -0.0094771  0.0443395  -0.214  0.83075
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Smooth components’ approximate significance:
               edf Ref.df   Chi.sq p-value
s(bmi)       4.416  5.415    6.188   0.314
s(income)    8.493  8.923 2335.550  <2e-16 ***
s(age)       6.371  7.518  116.092  <2e-16 ***
s(education) 6.898  7.847 1153.674  <2e-16 ***
```

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


EQUATION 2
Link function for mu.2: probit
Formula: visits.hosp ~ private + s(bmi) + s(income) + s(age) +
    s(education) + as.factor(health) + as.factor(race) +
    as.factor(limitation) + as.factor(region) + gender +
    hypertension + hyperlipidemia + diabetes

Parametric coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)           -0.63295    0.07189  -8.804  < 2e-16 ***
private               -0.05685    0.07319  -0.777 0.437315
as.factor(health)6     0.11841    0.03448   3.435 0.000593 ***
as.factor(health)7     0.16677    0.03660   4.556 5.21e-06 ***
as.factor(health)8     0.30112    0.04799   6.274 3.51e-10 ***
as.factor(health)9     0.51663    0.06863   7.528 5.17e-14 ***
as.factor(race)3      -0.10541    0.03357  -3.140 0.001687 **
as.factor(race)4      -0.09301    0.12751  -0.729 0.465729
as.factor(race)5      -0.13672    0.04801  -2.848 0.004402 **
as.factor(limitation)6 -0.45311    0.04202 -10.784  < 2e-16 ***
as.factor(region)3     0.16104    0.03899   4.131 3.62e-05 ***
as.factor(region)4    -0.23138    0.03607  -6.415 1.41e-10 ***
as.factor(region)5    -0.35540    0.03991  -8.906  < 2e-16 ***
gender                -0.38202    0.02555 -14.953  < 2e-16 ***
hypertension           0.09921    0.03150   3.149 0.001637 **
hyperlipidemia         0.25949    0.03042   8.529  < 2e-16 ***
diabetes               0.10593    0.04362   2.429 0.015158 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Smooth components’ approximate significance:
               edf Ref.df  Chi.sq  p-value
s(bmi)       1.285  1.522   1.015   0.3556
s(income)    2.308  2.964   8.918   0.0292 *
s(age)       1.000  1.000 115.124  < 2e-16 ***
s(education) 6.940  7.847  68.572 1.58e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


n = 18592  theta = 0.313(0.177,0.476)  tau = 0.135(0.0814,0.192)
total edf = 71.7


n = 18592
theta = 0.313(0.177,0.476)  tau = 0.135(0.0814,0.192)
total edf = 71.7
```

Note that we have set a seed before `summary()`. This allows us to recover the same results for the intervals of θ and τ reported at the bottom of the summary output; recall that intervals for non-linear functions of the model parameters are calculated using posterior simulation. The small yet significant dependence parameter obtained for the Clayton copula indicates that there exists some positive association between the unstructured terms of the model equations. This suggests that individuals with private health coverage are more likely to use health care services as compared to those without coverage. The estimated effects for the binary and categorical variables have the expected signs and we refer the reader to [47] for a thorough discussion of these. Using `plot()`, we produce the smooth function estimates for the treatment and outcome equations which are reported in Figures 1 and 2.

```
R> par(mfrow = c(2, 2), mar = c(4.5, 4.5, 2, 2),
+      cex.axis = 1.6, cex.lab = 1.6)
R> plot(bpC180, eq = 1, seWithMean = TRUE, scale = 0, shade = TRUE,
+       pages = 1, jit = TRUE)
R> par(mfrow = c(2, 2), mar = c(4.5, 4.5, 2, 2),
+      cex.axis = 1.6, cex.lab = 1.6)
R> plot(bpC180, eq = 2, seWithMean = TRUE, scale = 0, shade = TRUE,
+       pages = 1, jit = TRUE)
```

The effects of `bmi`, `income`, `age` and `education` in the treatment and outcome equations show different degrees of non-linearity. The point-wise confidence intervals of the smooth functions for `bmi` in the treatment and outcome equations contain the zero line for the whole range of the covariate values. The intervals of the smooth for `income` in the outcome equation contain the zero line for most of the covariate value range. This suggests that `bmi` is a weak predictor of private health insurance and health care utilization, and that `income` might not be a very important determinant of hospital utilization. Similar conclusions can be drawn by looking at the p-values reported in the summary output. As for the remaining variables, the estimated effects have the expected patterns. For example, `age` is a significant determinant in both equations. The probability of purchasing a private health insurance is found to increase with `age`. The likelihood of using health care services also increases with `age`. Insurance decision as well as health care utilization appear to be highly associated with `education`. Education is likely to increase individuals' awareness of health care services and the benefits of purchasing a private health insurance. Higher household income is associated with an increased propensity of purchasing a private health insurance. See for example [7] for further details.

The estimated SATE (in %) and corresponding interval are given below.

```
R> set.seed(1)
R> AT(bpC180, nm.end = "private", hd.plot = TRUE, cex.axis = 1.5,
+     cex.lab = 1.5, cex.main = 1.6)

Average treatment effect (%) with 95% interval:

-1.11 (-4.14,1.59)
```

Figures 3 displays a plot of the histogram and kernel density estimate of the simulated SATE distribution for the fitted Clayton180 copula model. For completeness, we also calculate SATE for the case in which unobserved confounding is not accounted for and that in which both observed and unobserved confounding is not account for.

```
R> AT(bpC180, nm.end = "private", type = "univariate")

Average treatment effect (%) with 95% interval:

3.94 (2.97,5.05)
```

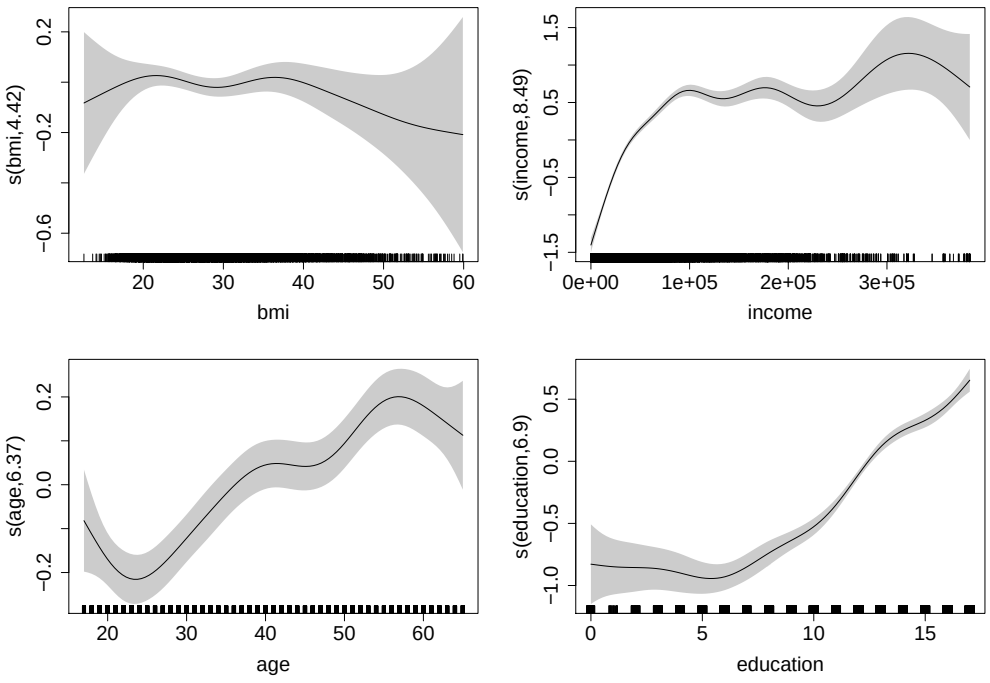

Figure 1: Treatment equation: smooth function estimates and associated 95% point-wise confidence intervals obtained by fitting the Clayton180 copula model on the 2008 MEPS data. Results are plotted on the scale of the additive predictor. The jittered rug plot, at the bottom of each graph, shows the covariate values. The numbers in brackets in the y-axis captions are the edf of the smooth curves.

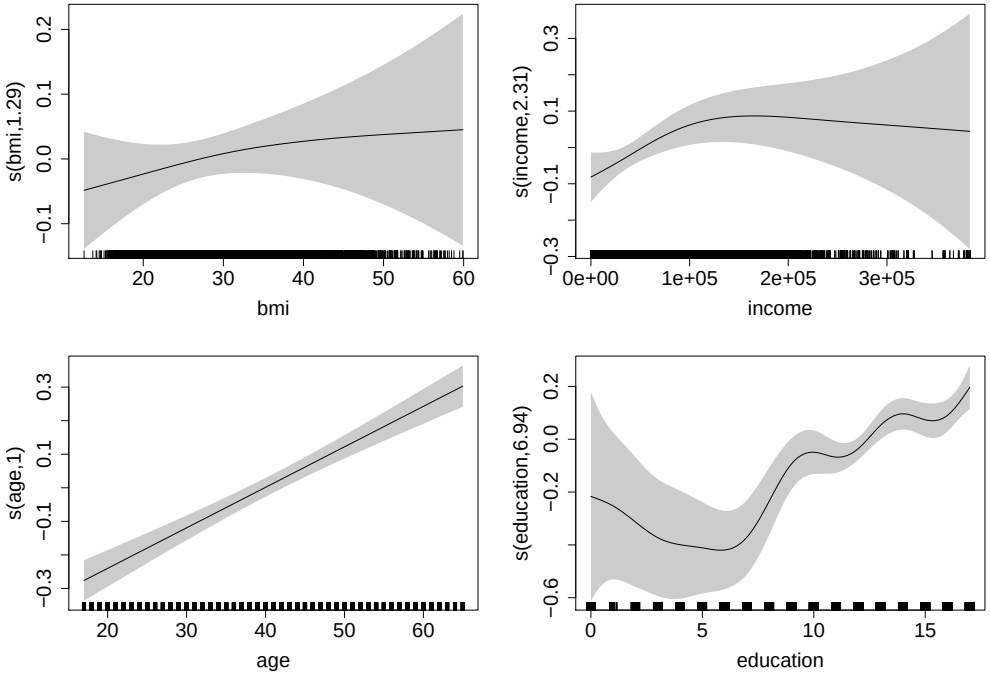

Figure 2: Outcome equation: smooth function estimates and associated 95% point-wise confidence intervals obtained by fitting the Clayton180 copula model on the 2008 MEPS data. Results are plotted on the scale of the additive predictor.

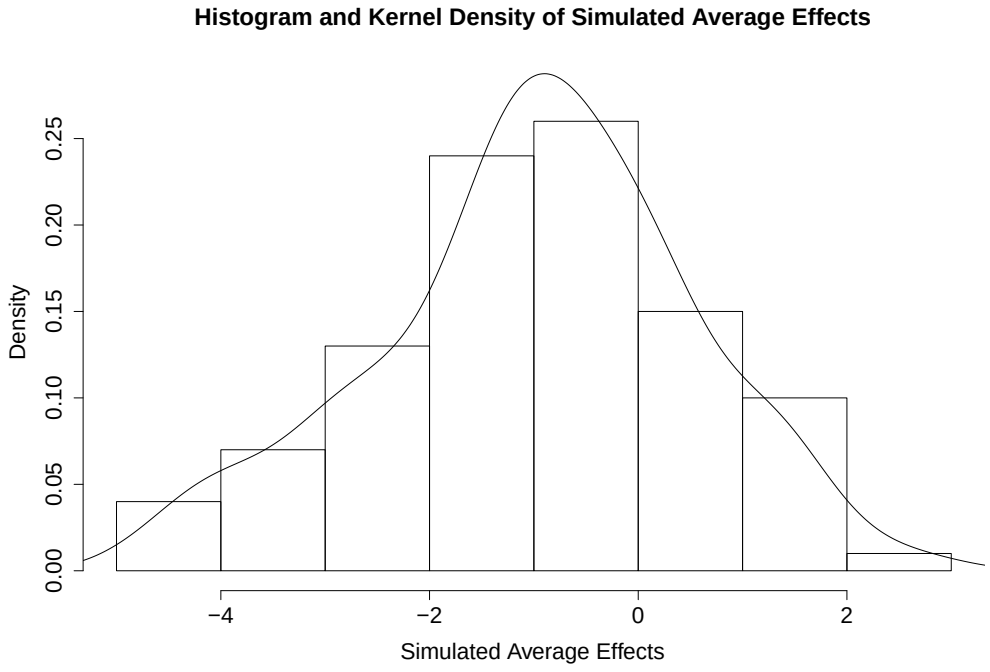

Figure 3: Histogram and kernel density estimate of the simulated SATE distribution for the fitted Clayton180 copula model.

```
R> AT(bpC180, nm.end = "private", type = "naive")

Average treatment effect (%) with 95% interval:

4.63 (3.95,5.31)
```

The naive estimate is larger than those obtained when observed and/or unobserved confounders are accounted for. Focusing on the univariate and bivariate estimates, we can see that the bivariate model indicates that private insurance does not influence significantly the outcome of interest, whereas the univariate model suggests that the impact is positive and significant. The results are not in agreement and the researcher should be careful when adopting a particular estimate for policy planning, for instance. Functions `OR()` and `RR()` calculate the odds ratio and risk ratio, respectively, and can also be used to assess the impact of the endogenous variable.

5.2 HIV prevalence

The sample selection bivariate binary model is illustrated on a data-set generated using the real 2007 Zambian DHS on HIV; details can be found in [36]. Estimates of HIV prevalence are important for policy in order to establish the health status of a country's population, to evaluate the effectiveness of population-based interventions and campaigns, to identify the most at risk members of the population, and to target those most in need of treatment. However, data in low and middle income countries are often derived from HIV testing conducted as part of household surveys, where participation rates in testing can be low. Low participation rates may be attributed to HIV positive individuals being less likely to participate because they fear disclosure, in which case, estimates obtained using conventional approaches to deal with non-participation, such as imputation-based methods, will be biased. In addition, establishing which population sub-groups are most in need of intervention requires modeling of both spatial dependence and the predictors of HIV status, which

is complicated by data censoring due to non-participation. See [5] and [36] for full details. All these issues are taken into account in the analysis below.

In the relevant survey, respondents are asked, at the end of their individual interview, if they would consent to test for HIV. If they consent then a blood sample is drawn by finger prick by the interviewer, and subsequently the dried blood spot is sent to be laboratory tested for HIV. The model includes the variables described in Table 2 in the Appendix.

We specify smooth functions of all the continuous variables, and employ Markov random field smoothers to model spatial variation. All these components enter the additive predictors for the selection and HIV status equations. The selection variable (exclusion restriction) is `interviewerID` and enters the first equation only. We apply a ridge penalty to the coefficients of this variable in order to account for the difficulties associated with its use (e.g., `interviewerID` can be collinear with other independent variables since interviewers are often matched to participants on the basis of some group-level characteristics such as language and ethnicity). The additive predictor for the copula parameter only depends on the Markov random field term and allows the association parameter to vary by region. See [35] for a detailed discussion of this model specification.

We first read the data-set and region shape list (`hiv.polys`). Then, to account for geographic clustering of HIV we store the neighborhood structure information in an object `xt` which is then used in specification of the Gaussian Markov random filed smoother. The model is defined below. Note that the employed specification is fairly complex and it has been adopted to illustrate the flexibility of the modeling approach.

```
R> library(GJRM)
R> data("hiv", package = "GJRM")
R> data("hiv.polys", package = "GJRM")
R> xt <- list(polys = hiv.polys)
R> sel.eq <- hivconsent ~ s(age) + s(education) + s(wealth) +
+                         s(region, bs = "mrf", xt = xt, k = 7) +
+                         marital + std + age1sex_cat + highhiv +
+                         partner + condom + aidscare +
+                         knowsdiedofaids + evertestedHIV +
+                         smoke + religion + ethnicity +
+                         language + s(interviewerID, bs = "re")
R> out.eq <- hiv        ~ s(age) + s(education) + s(wealth) +
+                         s(region, bs = "mrf", xt = xt, k = 7) +
+                         marital + std + age1sex_cat + highhiv +
+                         partner + condom + aidscare +
+                         knowsdiedofaids + evertestedHIV +
+                         smoke + religion + ethnicity +
+                         language
R> theta.eq <-          ~ s(region, bs = "mrf", xt = xt, k = 7)
R> fl <- list(sel.eq, out.eq, theta.eq)
R> bss <- gjrm(fl, data = hiv, BivD = "J90", Model = "BSS",
+              margins = c("probit", "probit"))
R> mean(bss$theta)

-8.459184
```

The estimated dependence parameter is −8.45. Note, however, that this is an average of the copula coefficients corresponding to the nine Zambian regions considered in the analysis. This result supports the hypothesis that those who are most likely to be HIV positive are those who are also most likely to decline to participate in testing. The estimated smooth functions of `age`, `education`, `wealth` and `region`, and the effects of the binary and categorical variables can be extracted as in the previous example. See [35] for a full analysis of these. The estimated prevalences from the naive, univariate and selection models are given below.

```
R> prev(bss, sw = hiv$sw, type = "naive")

Estimated prevalence (%) with 95% interval:

12.1 (11.2,13.0)

R> set.seed(1)
R> prev(bss, sw = hiv$sw, type = "univariate")

Estimated prevalence (%) with 95% interval:

12.1 (11.6,13.2)

R> prev(bss, sw = hiv$sw)

Estimated prevalence (%) with 95% interval:

22.9 (19.9,26.3)
```

These estimates show that the selection model HIV prevalence is significantly higher than that of the imputation-based and naive models. At regional level the selection model HIV prevalences range from 13% to 28%. Note that prevalence estimates, and more generally model estimates, can be adjusted for clustering using `adjCov()` or `adjCovSD()`. Figure 4 shows maps for the selection model and single imputation estimates as well as the dependence parameter estimates.

```
R> lr <- length(hiv.polys)
R> prevBYreg  <- matrix(NA, lr, 2)
R> thetaBYreg <- NA
R> for(i in 1:lr) {
+  prevBYreg[i,1] <- prev(bss, sw = hiv$sw, ind = hiv$region==i,
                          type = "univariate")$res[2]
+  prevBYreg[i,2] <- prev(bss, sw = hiv$sw, ind = hiv$region==i)$res[2]
+  thetaBYreg[i]  <- bss$theta[hiv$region==i][1]
+  }
R> zlim <- range(prevBYreg*100)  # to establish a common prevalence range
R> par(mfrow = c(1, 3), cex.axis = 1.3)
R> polys.map(hiv.polys, prevBYreg[,1]*100, zlim = zlim, lab = "",
+            cex.lab = 1.5, cex.main = 1.5,
+            main = "HIV (%) - Imputation Model")
R> polys.map(hiv.polys, prevBYreg[,2]*100, zlim = zlim, cex.main = 1.5,
+            main = "HIV (%) - Selection Model")
R> polys.map(hiv.polys, thetaBYreg, rev.col = FALSE, cex.main = 1.7,
+            main = expression(paste("Copula parameter (",hat(theta),")")))
```

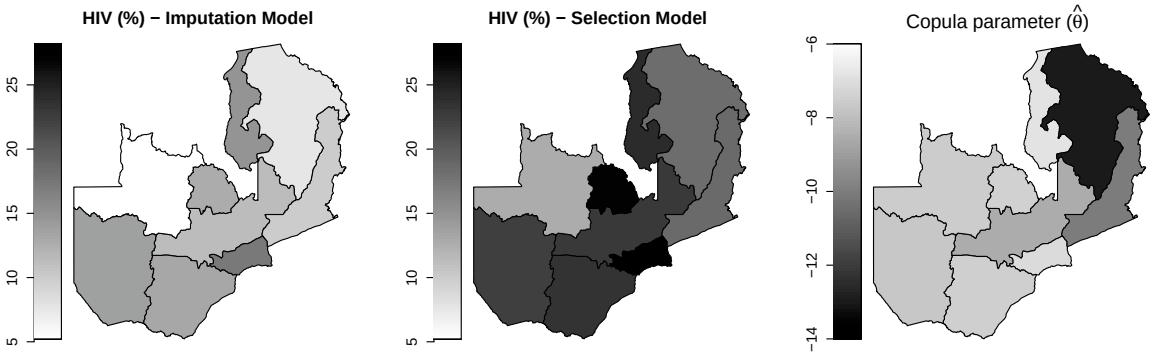

Figure 4: HIV prevalence estimates by region obtained by applying the imputation and sample selection models. The copula parameter plot reports the values of the estimated associations with range $(-\infty, -1)$ in the 90° Joe copula. The higher the absolute value, the stronger the association between the selection and outcome equations.

Intervals for the θ_i can extracted using the summary function. For instance,

```
R> set.seed(1)
R> CItheta <- summary(bss)$CItheta
R> CItheta[1,]

      2.5%      97.5%
-17.916304  -4.264049
```

5.3 Determinants of civil war onset

To highlight the benefits of using a bivariate probit model with partial observability, we re-estimate the model proposed in the [14]'s seminal study on civil war onset which has also been analyzed more recently by [39].

Civil wars are often theorized as the outcome of an interaction between an opposition group and the government [12, 14]. This means that we can only observe their joint decision (war onset) rather than the decisions of the single decision-makers (the opposition 'challenges' and the state 'fights'). The study by [14] aims at identifying the variables that increase the likelihood of civil war onset, however cannot distinguish between variables that drive local populations to rebel against the government and variables that influence government's fight. As in [39], the model includes the variables described in Table 3 in the Appendix.

We specify two equations, fit the model and check that convergence has been achieved.

```
R> library(GJRM)
R> data("war", package = "GJRM")
R> reb.eq <- onset ~ instab + oil + warl + lpopl + lmtnest +
+                    ethfrac + polity2l + s(gdpenl) + s(relfrac)
R> gov.eq <- onset ~ instab + oil + warl + ncontig + nwstate +
                     s(gdpenl)
R> bpo <- gjrm(list(reb.eq, gov.eq), data = war, Model = "BPO",
               margins = c("probit","probit") )
R> conv.check(bpo)

Largest absolute gradient value: 0.1752897
Observed information matrix is positive definite
Eigenvalue range: [0.1111491,5.648263e+13]
```

```
Trust region iterations before smoothing parameter estimation: 20
Loops for smoothing parameter estimation: 2
Trust region iterations within smoothing loops: 3
```

The convergence diagnostics suggest that the model is perhaps too complex (the gradient is close but not equal to 0 and the condition number of the information matrix relatively large). We check the estimate obtained for θ and interval for it.

```
R> set.seed(1)
R> sbpo <- summary(bpo)
R> sbpo$theta; sbpo$CItheta

     theta
0.05459771
           2.5%     97.5%
[1,] -0.8903422 0.9390838
```

This result suggests that the unobservable variables influencing local populations to rebel against the government and government's decision to fight back are uncorrelated. Following [2], we can therefore simplify the model by assuming a priori that $\theta = 0$. This implies that $p_{11i} = \Phi(\eta_{1i})\Phi(\eta_{2i})$.

```
R> bpo0 <- gjrm(list(reb.eq, gov.eq), data = war,
                Model = "BPO0", margins = c("probit","probit"))
R> conv.check(bpo0)

Largest absolute gradient value: 0.0725329
Observed information matrix is positive definite
Eigenvalue range: [0.1355123,4.740461e+13]

Trust region iterations before smoothing parameter estimation: 20
Loops for smoothing parameter estimation: 2
Trust region iterations within smoothing loops: 3
```

The gradient is now closer to zero. However, looking at the summary results (below) one notes that the estimated smooth functions have $edf = 1$. Hence, `gdpenl` and `relfrac` can in principle enter the model parametrically; this is what makes the condition number large in this case.

```
R> summary(bpo0)

COPULA:   Gaussian
MARGIN 1: Bernoulli
MARGIN 2: Bernoulli

EQUATION 1
Link function for mu.1: probit
Formula: onset ~ instab + oil + warl + lpopl + lmtnest + ethfrac +
    polity2l + s(gdpenl) + s(relfrac)

Parametric coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.316782   0.374343  -8.860  < 2e-16 ***
instab      -0.120379   0.259779  -0.463  0.64308
```

```
oil          1.074245   0.470581   2.283  0.02244 *
warl        -0.598469   0.320963  -1.865  0.06224 .
lpopl        0.116009   0.045151   2.569  0.01019 *
lmtnest      0.111640   0.043090   2.591  0.00957 **
ethfrac      0.085478   0.196891   0.434  0.66419
polity2l     0.010101   0.008683   1.163  0.24471
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Smooth components’ approximate significance:
           edf Ref.df Chi.sq  p-value
s(gdpenl)    1      1 13.763 0.000207 ***
s(relfrac)   1      1  0.565 0.452208
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


EQUATION 2
Link function for mu.2: probit
Formula: onset ~ instab + oil + warl + ncontig + nwstate + s(gdpenl)

Parametric coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.4317     0.6604  -0.654    0.513
instab        0.8721     0.6107   1.428    0.153
oil          -0.9638     0.6070  -1.588    0.112
warl          0.2113     0.6875   0.307    0.759
ncontig       0.5862     0.4968   1.180    0.238
nwstate       2.6507     2.6739   0.991    0.322

Smooth components’ approximate significance:
          edf Ref.df Chi.sq p-value
s(gdpenl)   1      1  0.434    0.51


n = 6326  total edf = 17
```

For comparison, using `mgcv`, we also fit a probit model where the joint decision of the opposition group and of the government is modeled without distinguishing between the opposition's challenge and the government's decision to fight back.

```
R> war.eq <- onset ~ instab + oil + warl +  ncontig + nwstate + lpopl +
+                    lmtnest + ethfrac + polity2l + s(gdpenl) + s(relfrac)
R> Probit <- gam(war.eq, family = binomial(link = "probit"), data = war)
R> summary(Probit)

Family: binomial
Link function: probit

Formula:
onset ~ instab + oil + warl + ncontig + nwstate + lpopl + lmtnest +
```

```
    ethfrac + polity2l + s(gdpenl) + s(relfrac)

Parametric coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.641277   0.294255 -12.375  < 2e-16 ***
instab       0.261447   0.100910   2.591 0.009573 **
oil          0.363108   0.122069   2.975 0.002934 **
warl        -0.378155   0.129964  -2.910 0.003618 **
ncontig      0.155754   0.121656   1.280 0.200446
nwstate      0.759497   0.163264   4.652 3.29e-06 ***
lpopl        0.104802   0.031235   3.355 0.000793 ***
lmtnest      0.091518   0.034332   2.666 0.007684 **
ethfrac      0.078613   0.157390   0.499 0.617443
polity2l     0.009303   0.007004   1.328 0.184115
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
             edf Ref.df Chi.sq p-value
s(gdpenl)  1.002  1.004 22.845 1.8e-06 ***
s(relfrac) 1.001  1.002  0.366   0.546
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.0314   Deviance explained = 10.5%
UBRE = -0.845  Scale est. = 1         n = 6326
```

Although both the probit and bivariate probit models recover coefficients with the same signs, there are several differences in the statistical significance of these parameters (`nwstate`, `instab`, for example). What is of greater consequence, however, is that, unlike probit, the partial observability model allows for a more nuanced separation of alternative theoretical mechanisms. For instance, `instab`, `oil` and `warl` are all statistically significant in the probit model; each of these variables may affect the onset of civil war through the two theoretical mechanisms, associated with opposition and government. The partial observability model permits for evaluating each of the player-specific and outcome-specific theoretical components. To demonstrate this point, let us focus, for instance, on how the two models separate the competing mechanisms linking civil war and GDP per capita. The probit model shows that `gdpenl` has a negative linear and statistically significant effect. The effect is linear because $edf = 1$ and has a negative impact as illustrated below.

```
R> coef(Probit)[(which(names(coef(Probit)) == "s(gdpenl).9"))]

s(gdpenl).9
   -0.58988
```

(When using thin plate regression splines with basis dimensions equal to 10 and second-order penalties, if $edf = 1$ then the coefficient of the ninth spline basis corresponds to the parametric linear effect.) While this suggests that GDP per capita reduces the propensity of civil war onset, we cannot determine which of the two alternative mechanisms are supported by this result. In other words, a negative coefficient for `gdpenl` in the probit model may indicate that (i) states with greater capacities are more efficient at deterring insurgents or (ii) prospective rebels are less likely to challenge the state in the presence of higher opportunity costs or (iii) both (i) and (ii). In contrast, the partial observability model provides some insights in regard to these processes. `gdpenl` is negative, linear and statistically significant in the rebels' challenge equation. This indicates that

as GDP per capita increases, potential rebel groups are less likely to challenge the government. In contrast, `gdpenl` is not significant in government's fight back equation.

```
R> coef(bpo0)[(which(names(coef(bpo)) == "s(gdpenl).9"))]

s(gdpenl).9 s(gdpenl).9
 -0.9214988   0.4603390
```

Figure 5 displays the predicted probabilities of several outcomes (war onset, rebels challenging the state, and government fighting back) across varying values in GDP per capita.

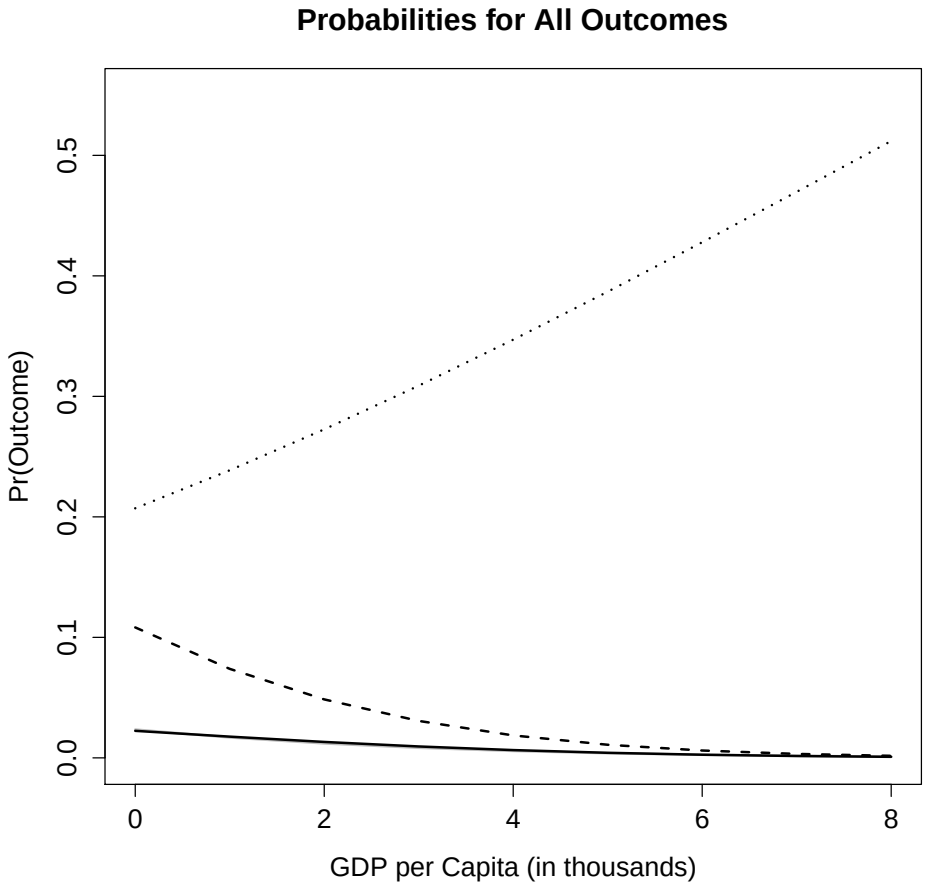

Figure 5: The probabilities of civil war predicted from the probit model are depicted as a continuous gray line. The probabilities of rebels challenging the state, of government fighting back, and of civil war from the partial observability model are depicted as dashed, dotted and continuous black lines, respectively. Note that the probabilities of civil war for both models can not be distinguished as in this case they are nearly identical.

```
R> probitW <- bpoW <- bpoReb <- bpoGov <- NA
R> gdp.grid <- seq(0, 8)
R> median.values <- data.frame(t(apply(war, 2, FUN = median)))
R> for (i in 1:length(gdp.grid)){
+    newd <- median.values; newd$gdpenl <- gdp.grid[i]
+    eta1 <- predict(bpo0, eq = 1, newd)
```

```
+    eta2 <- predict(bpo0, eq = 2, newd)
+    probitW[i] <- predict(Probit, newd, type = "response")
+    bpoW[i]    <- pnorm(eta1)*pnorm(eta2)
+    bpoReb[i]  <- pnorm(eta1)
+    bpoGov[i]  <- pnorm(eta2)
+  }
R> plot(gdp.grid, probitW, type = "l", ylim = c(0, 0.55), lwd = 2,
+       col = "grey", xlab = "GDP per Capita (in thousands)",
+       ylab = "Pr(Outcome)", main = "Probabilities for All Outcomes",
+       cex.main = 1.5, cex.lab = 1.3, cex.axis = 1.3)
R> lines(gdp.grid, bpoW,   lwd = 2)
R> lines(gdp.grid, bpoReb, lwd = 2, lty = 2)
R> lines(gdp.grid, bpoGov, lwd = 2, lty = 3)
```

The probit and partial observability models yield identical results as far as the probability of civil war is concerned. However, the partial observability model reveals additional information about the effect of GDP per capita on the rebel-government interaction, by also allowing to estimate the probabilities of rebels challenging the state and government fighting back. We see, for example, that while the former decreases as GDP per capita increases, the latter increases.

6 Discussion

We described the bivariate binary models implemented in the R add-on package GJRM and illustrated them using three case studies in which the issues of endogeneity, non-random sample selection and partial observability were prevalent. The framework allows the user to specify flexibly covariate effects and the dependence structure between the margins. Given the modular structure of the estimation algorithm, other copulae and link functions can be incorporated in the package with little programming work.

Since link functions other than the ones implemented in the package may be plausible in applications, we explored the empirical performance of skew probit links, derived from the standard skew-normal distribution by [3], and power probit and reciprocal power probit links [6]. We opted for these links as they include the probit as special case and have desirable mathematical properties. We found that the use of these approaches causes numerical difficulties, which is in line with the arguments of [4]. Moreover, even when numerical convergence is achieved, the empirical results are virtually identical to those obtained when assuming probit links. We also considered non-exchangeable copulae and, following the approach detailed in [16], assessed the feasibility of using $C_{\kappa_1,\kappa_2}(u,v) = u^{1-\kappa_1}v^{1-\kappa_2}C(u^{\kappa_1}, v^{\kappa_2})$, $0 < \kappa_1, \kappa_2 < 1$ in the context of bivariate binary data. We encountered the same issues mentioned above, even when employing models with a small number of covariates and without nonlinear effects.

As mentioned in the introduction, the package allows for the modeling of several types of multivariate responses in a flexible regression context. We are currently working on several extensions of the models in GJRM and incorporating new ones.

Appendix - Variable definitions

Table 1: MEPS data: description of the outcome and treatment variables, and observed confounders.

Variable	Definition
Outcome	
visits.hosp	= 1 at least one visit to hospital outpatient departments
Treatment	
private	= 1 private health insurance
Demographic-socioeconomic	
age	age in years
gender	= 1 male
race	= 2 white, = 3 black, = 4 native American, = 5 others
education	years of education
income	income (000's)
region	= 2 northeast, = 3 mid-west, = 4 south, = 5 west
Health-related	
health	= 5 excellent, = 6 very good, = 7 good, = 8 fair, = 9 poor
bmi	body mass index
diabetes	= 1 diabetic
hypertension	= 1 hypertensive
hyperlipidemia	= 1 hyperlipidemic
limitation	= 1 health limits physical activity

Table 2: HIV data: description of the outcome and selection variables, and observed confounders.

Variable	Definition
Selection	
hivconsent	consent to test for HIV
Outcome	
hiv	HIV positive
Demographic-socioeconomic	
age	age in years
education	years of education
region	= 1 central, = 2 copperbelt, = 3 eastern, = 4 luapula, = 5 lusaka, = 6 northwestern, = 7 northern, = 8 southern, = 9 western
et	bemba, lunda (luapula), lala, ushi, lamba, tonga, luvale, lunda (northwestern), mbunda, kaonde, lozi, chewa, nsenga, ngoni, mambwe, namwanga, tumbuka, other
language	English, Bemba, Lozi, Nyanja, Tonga, other
marital	never married, currently married, formerly married
interviewerID	interviewer identifier
Health sex-related	
std	had a sexually transmitted disease
highhiv	had high risk sex
condom	used condom during last intercourse
aidscare	= 1 if would care for an HIV-infected relative
knowsdiedofaids	= 1 if know someone who died of HIV
evertestedHIV	= 1 if previously tested for HIV
smoke	smoker

Table 3: Civil war data: description of the outcome variable and covariates.

Variable	Definition
Outcome	
`onset`	= 1 for all country-years in which a civil war started
Covariates	
`instab`	= 1 unstable government
`oil`	= 1 for oil exporter country
`warl`	= 1 if the country had a distinct civil war ongoing in the previous year
`lpopl`	log(population size)
`lmtnest`	log(%mountainous)
`ethfrac`	measure of ethnic fractionalization (calculated as the probability that two randomly drawn individuals from a country are not from the same ethnicity)
`polity2l`	measure of political democracy (ranges from −10 to 10) lagged one year
`gdpenl`	GDP per capita (measured as thousands of 1985 U.S. dollars) lagged one year
`relfrac`	measure of religious fractionalization
`ncontig`	= 1 for non-contiguous state
`nwstate`	= 1 for new state

Acknowledgement: We would like to thank Bear F. Braumoeller for suggesting the implementation of the bivariate probit model with partial observability, and a reviewer for pointing out various corrections which have improved the readability and message of the article.

References

[1] Abadie, A., D. Drukker, J. L. Herr, and G. W. Imbens (2004). Implementing matching estimators for average treatment effects in Stata. *Stata J. 4*(3), 290–311.
[2] Abowd, J. M. and H. S. Farber (1982). Job queues and the union status of workers. *Ind. Labor. Relat. Rev. 35*(3), 354–367.
[3] Azzalini, A. (1985). A class of distributions which includes the normal one. *Scand. J. Stat. 12*(2), 171–178.
[4] Azzalini, A. and R. B. Arellano-Valle (2013). Maximum penalized likelihood estimation for skew-normal and skew-t distributions. *J. Stat. Plan. Infer. 143*(2), 419–433.
[5] Bärnighausen, T., J. Bor, S. Wandira-Kazibwe, and D. Canning (2011). Correcting HIV prevalence estimates for survey nonparticipation using Heckman-type selection models. *Epidemiology 22*(1), 27–35.
[6] Bazan, J. L., H. Bolfarinez, and M. B. Branco (2010). A framework for skew-probit links in binary regression. *Commun. Stat. Simulat. 39*(4), 678–697.
[7] Buchmueller, T. C., K. Grumbach, R. Kronick, and J. G. Kahn (2005). The effect of health insurance on medical care utilization and implications for insurance expansion: a review of the literature. *Med. Care Res. Rev. 62*(1), 3–30.
[8] Cappellari, L. and S. P. Jenkins (2003). Multivariate probit regression using simulated maximum likelihood. *Stata J. 3*(3), 278–294.
[9] Chen, G. G. and T. Åstebro (2012). Bound and collapse bayesian reject inference for credit scoring. *J. Oper. Res. Soc. 63*(10), 1374–1387.
[10] Chib, S. and E. Greenberg (2007). Semiparametric modeling and estimation of instrumental variable models. *J. Comput. Graph. Stat. 16*(1), 86–114.
[11] Clarke, P. S. and F. Windmeijer (2012). Instrumental variable estimators for binary outcomes. *J. Amer. Statist. Assoc. 107*, 1638–1652.
[12] Collier, P. and A. Hoeffler (2004). Greed and grievance in civil war. *Oxford Econ. Pap. 56*, 563–595.
[13] Dubin, J. A. and D. Rivers (1989). Selection bias in linear regression, logit and probit models. *Sociol. Method Res. 18*(2–3), 360–390.
[14] Fearon, J. D. and D. D. Laitin (2003). Ethnicity, insurgency, and civil war. *Am. Polit. Sci. Rev. 97*(1), 75–90.
[15] Fitzmaurice, G., M. Davidian, G. Verbeke, and G. Molenberghs (2008). *Longitudinal Data Analysis*. Chapman & Hall/CRC, London.
[16] Frees, E. W. and E. A. Valdez (1998). Understanding relationships using copulas. *N. Am. Actuar. J. 2*(1), 1–25.

[17] Goldman, D. P., J. Bhattacharya, D. F. McCaffrey, N. Duan, A. A. Leibowitz, G. F. Joyce, and S. C. Morton (2001). Effect of insurance on mortality in an HIV-positive population in care. *J. Amer. Statist. Assoc. 96*, 883–894.
[18] Greene, W. H. (2012). *Econometric Analysis*. Prentice Hall, New York.
[19] Gronau, R. (1974). Wage comparisons: A selectivity bias. *J. Polit. Econ. 82*(6), 1119–1143.
[20] Heckman, J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Ann. Econ. Soc. Meas. 5*(4), 475–492.
[21] Heckman, J. J. (1978). Dummy endogenous variables in a simultaneous equation system. *Econometrica 46*(4), 931–959.
[22] Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica 47*(1), 153–161.
[23] Henningsen, A. (2015). *mvProbit: Multivariate Probit Models*. R package version 0.1-8. Available on CRAN.
[24] Inc., S. I. (2017a). *SAS/ETS(R) 14.2 User's Guide*. Cary, NC.
[25] Inc., S. I. (2017b). *SAS/STAT Software, Version 9.4*. Cary, NC.
[26] Jeliazkov, I. and X. S. Yang (2014). *Bayesian Inference in the Social Sciences*. John Wiley & Sons, Hoboken NJ.
[27] Latif, E. (2009). The impact of diabetes on employment in Canada. *Health Econ. 18*(5), 577–589.
[28] Lewis, H. G. (1974). Comments on selectivity biases in wage comparisons. *J. Polit. Econ. 82*(6), 1145–1155.
[29] Li, Y. and G. A. Jensen (2011). The impact of private long-term care insurance on the use of long-term care. *Inquiry 48*(1), 34–50.
[30] Maddala, G. S. (1983). *Limited Dependent and Qualitative Variables in Econometrics*. Cambridge University Press.
[31] Marra, G. and R. Radice (2011). Estimation of a semiparametric recursive bivariate probit model in the presence of endogeneity. *Can. J. Stat. 39*(2), 259–279.
[32] Marra, G. and R. Radice (2013). A penalized likelihood estimation approach to semiparametric sample selection binary response modeling. *Electron. J. Stat. 7*, 1432–1455.
[33] Marra, G. and R. Radice (2017a). Bivariate copula additive models for location, scale and shape. *Comput. Stat. Data An. 112*, 99–113.
[34] Marra, G. and R. Radice (2017b). *GJRM: Generalised Joint Regression Modelling*. R package version 0.1-2. Available on CRAN.
[35] Marra, G., R. Radice, T. Bärnighausen, S. N. Wood, and M. E. McGovern (2017). A simultaneous equation approach to estimating HIV prevalence with non-ignorable missing responses. *J. Amer. Statist. Assoc. 112*(518), 484–496.
[36] McGovern, M. E., T. Bärnighausen, G. Marra, and R. Radice (2015). On the assumption of bivariate normality in selection models: a copula approach applied to estimating HIV prevalence. *Epidemiology 26*(2), 229–237.
[37] Miranda, A. and S. Rabe-Hesketh (2006). Maximum likelihood estimation of endogenous switching and sample selection models for binary, ordinal, and count variables. *Stata J. 6*(3), 285–308.
[38] Nelsen, R. (2006). *An Introduction to Copulas*. Second edition. Springer, New York.
[39] Nieman, M. D. (2015). Statistical analysis of strategic interaction with unobserved player actions: Introducing a strategic probit with partial observability. *Polit. Anal. 23*(3), 429–448.
[40] Pianzola, J. (2014). Selection biases in voting advice application research. *Elect. Stud. 36*, 272–280.
[41] Poirier, D. J. (1980). Partial observability in bivariate probit models. *J. Econometrics 12*(2), 209–217.
[42] Poirier, D. J. (2014). Identification in multivariate partial observability probit. *Int. J. Math. Model. Num. Optim. 5*(1–2), 45–63.
[43] R Development Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna.
[44] Radice, R., G. Marra, and M. Wojtys (2016). Copula regression spline models for binary outcomes. *Stat. Comput. 26*(5), 981–995.
[45] Rigby, R. A. and D. M. Stasinopoulos (2005). Generalized additive models for location, scale and shape. *J. Roy. Statist. Soc. Ser. C 54*(3), 507–554.
[46] Ruppert, D., M. P. Wand, and R. J. Carroll (2003). *Semiparametric Regression*. Cambridge University Press, New York.
[47] Shane, D. and P. K. Trivedi (2012). What drives differences in health care demand? The role of health insurance and selection bias. HEDG Working Papers 12/09. Available at https://www.york.ac.uk/media/economics/documents/herc/wp/12_09.pdf.
[48] Shideler, G. S., D. W. Carter, C. Liese, and J. E. Serafy (2015). Lifting the goliath grouper harvest ban: Angler perspectives and willingness to pay. *Fish. Res. 161*, 156–165.
[49] Sklar, A. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris 8*, 229–231.
[50] Sklar, A. (1973). Random variables, joint distributions, and copulas. *Kybernetica 9*, 449–460.
[51] StataCorp (2015a). *Stata 14 Base Reference Manual*. StataCorp LP, College Station TX.
[52] StataCorp (2015b). *Stata Statistical Software: Release 14*. StataCorp LP, College Station TX.
[53] Toomet, O. and A. Henningsen (2008). Sample selection models in R: Package sampleselection. *J. Stat. Softw. 27*(7), 1–23.
[54] Van de Ven, W. P. and B. Van Praag (1981). The demand for deductibles in private health insurance: A probit model with sample selection. *J. Econometrics 17*(2), 229–252.
[55] Winkelmann, R. (2011). Copula bivariate probit models: with an application to medical expenditures. *Health Econ. 21*, 1444–1455.
[56] Wood, S. N. (2013a). On p-values for smooth components of an extended generalized additive model. *Biometrika 100*(1), 221–228.

[57] Wood, S. N. (2013b). A simple test for random effects in regression models. *Biometrika 100*(4), 1005–1010.
[58] Wood, S. N. (2017a). *Generalized Additive Models: An Introduction With R. Second edition*. Chapman & Hall/CRC, London.
[59] Wood, S. N. (2017b). *mgcv: Mixed GAM Computation Vehicle with GCV/AIC/REML Smoothness Estimation*. R package version 1.8-19. Available on CRAN.
[60] Yee, T. W. (2017). *VGAM: Vector Generalized Linear and Additive Models*. R package version 1.0-4. Available on CRAN.