A Unifying Switching Regime Regression Framework with Applications in Health Economics

Giampiero Marra, University College London, UK

Rosalba Radice, Bayes Business School, City, University of London, UK

David Zimmer, Western Kentucky University, Bowling Green, USA

2022-09-06

Abstract

Motivated by three health economics-related case studies, we propose a unifying and flexible regression modelling framework that involves regime switching. The proposal can handle the peculiar distributional shapes of the considered outcomes via a vast range of marginal distributions, allows for a wide variety of copula dependence structures and permits to specify all model parameters (including the dependence parameters) as flexible functions of covariate effects. The algorithm is based on a computationally efficient and stable penalised maximum likelihood estimation approach. The proposed modelling framework is employed in three applications in health economics, that use data from the Medical Expenditure Panel Survey, where novel patterns are uncovered. The framework has been incorporated in the `R` package `GJRM`, hence allowing users to fit the desired model(s) and produce easy-to-interpret numerical and visual summaries.

Key Words: copula; penalised regression spline; simultaneous estimation; structural equation model, switching regime.

1 Introduction

Inspired by health economics-related studies, we introduce a unifying and flexible regression modelling framework with endogenous regime switching. In particular, we consider three case studies which investigate the effect of a binary treatment on different types of outcomes. The first study involves a continuous outcome and investigates the effect of holding insurance through the employer on female wage earnings, and tests whether the theory of "compensating differentials" holds. The second study, where the outcome is discrete, addresses the question of whether visiting a doctor to obtain curative health care services affects children's school attendance. The third study, which involves a binary outcome, investigates the effect of private health insurance on health care consumption, and assesses the theory of "favorable selection". These studies use data from the Medical Expenditure Panel Survey (MEPS), collected and published by the Agency for Healthcare Research and Quality, an agency within the U.S. Department of Health and Human Services. Commencing in 1996 and still ongoing to this day, the MEPS enjoys a reputation for having the most complete individual-level information on health insurance, health care usage, and health conditions among large-scale household surveys in the U.S. The MEPS database files are freely available at `https://www.meps.ahrq.gov`.

Endogenous switching regression was originally envisioned in economics by Roy (1951) and then later exploited by Borjas (1987). A parallel approach was independently developed in statistics under the name of potential outcome framework (Neyman, 1923; Cox, 1958). The main difference between the two frameworks is in the model's formalisation; in economics, models are thought of in terms of realised, not potential, outcomes, because the counterfactual information is already enclosed in the related structural equations and hence there is no need to construct "non-realised" variables to carry this information (Pearl, 2015). Given our background knowledge and acquired experience, we adopt the structural equation approach since it naturally fits within the context of our case studies. In fact, this framework allows us to assess economic theories, check model assumptions, account for the continuous or non-continuous nature of the data, quantify the presence of omitted confounders, use the structural coefficients to obtain treatment effects, and flexibly adjust for the effect of many and different types of

observed confounders (e.g., Bollen, 2013).

Switching regression has proved to be a valuable tool in labour economics where it has been abundantly applied, discussed and extended in several directions (e.g., Chen et al., 2014; Cornelissen et al., 2016; Eisenhauer et al., 2015; Heckman, 1990; Heckman & Honore, 1990; Heckman & Hotz, 1989; D'Haultfoeuille & Maurel, 2013; Murtazashvili & Wooldridge, 2016; Smith, 2005). Various applications and extensions have also appeared in other fields (e.g., Bayer et al., 2011; Choi & Insik, 2009; Fitawek & Hendriks, 2021; Kim, 2021; Moscelli et al., 2018), hence highlighting the relevance of the modelling framework. When the approach was first developed and formalised, multivariate normality was assumed for theoretical and computational tractability. Since this assumption is clearly questionable in applications, extensions of it were proposed. A notable example is Smith (2005) who introduced a class of alternatives to the multivariate Gaussian based on copulae, a modelling strategy whereby a joint distribution is formed by specifying marginal distributions and a copula function that binds them together. In terms of software implementation, the traditional model, based on the assumption of normality, can be found in Lokshin & Sajaia (2004) and Toomet & Henningsen (2020). Hasebe (2013) provided an implementation where the normal, logistic and Student's t univariate distributions as well as several bivariate copula distributions can be employed to specify the model. Hasebe (2020) presented a switching regression model for count-data that exploits multivariate normality.

The aim of this work is to introduce a unifying copula-based switching regression framework that: a) is capable of handling binary, discrete and continuous outcomes via a vast range of marginal distributions; b) permits to model each parameter of the assumed multivariate distribution as a function of regression effects; c) can accommodate flexible regression structures; d) allows for a wide variety of copula dependence structures. These points are prompted by the three case studies. For example, in the first study, less conventional distributions such as the inverse Gaussian and Fisk distributions fit wage earnings better than more traditional ones. Also, since all distributional parameters can be modelled as a function of observed confounders, the information contained in the data can be better exploited. In the second study, the negative

binomial type II fits well the missed school days outcome, and the observed confounders enter both mean and dispersion parameters. The ability to accommodate flexibly outcome-observed confounder relationships means that new patterns and trends in the data may be uncovered and the impact of misspecification mitigated. As an example, in the first study, instead of imposing a quadratic shape on the effect of age, we let the data determine such shape which was found to be an increasing, but slightly concave, curve among females who do not hold insurance through their employers, and a non-monotonic pattern among females who do hold insurance. Finally, the copula parameters provide information on the presence and role of unobservables. In the third study, the estimated dependencies are negative for both treatment regimes, corroborating the assumption that unobservables that increase the likelihood of insurance also tend to reduce the probability of having a doctor visit. Note that the proposal requires an instrumental variable for the switching mechanism (e.g., French & Taber, 2011), as illustrated in the case studies.

The potential drawback of the proposed framework is that treatment effect identification is based on functional form assumptions. However, in spite of its parametric flavour, the flexibility offered by the methodology enables the data to point to meaningful model structures, hence capturing in a sense the spirit of semi-/non-parametric methods. Importantly, as stressed, for instance, by Chen et al. (2014, see also references therein), the use of semi-/non-parametric techniques in this context is problematic because they do not easily allow for treatment effect calculations and rely on asymptotic arguments that make such approaches less viable for empirical research.

The introduced framework allows for many layers of complexity, however there is no price to pay in terms of usability and interpretability. In fact, the modelling approach has been incorporated in the software package `GJRM` (Marra & Radice, 2022), written for the programming language `R` (R Core Team, 2022), which significantly eases the use of our switching regime framework. Parameter estimation relies on a carefully structured algorithm, whereas inference exploits a Bayesian result often employed for penalised likelihood-based models. The proposed methodological developments, together with fast and reliable software implementation, repre-

sent a significant advance in switching regression modelling. To the best of our knowledge, this is the first freely available implementation of a unifying and flexible copula-based switching regression framework.

The article is organised as follows. Section 2 introduces the general model and then discusses the related log-likelihoods (and the components that make them up) as well as the incorporation of flexible covariate effects. Section 3 describes parameter estimation, whereas Section 4 gives details on inference, some properties of the estimator, information criteria and the definition of residuals. Section 5 discusses the calculation of the average treatment effect and the procedure to obtain an interval for it. Section 6 presents the findings from our three case studies, and Section 7 concludes the paper with a discussion. The On-Line Supplementary Material provides details on the algorithm, discusses the findings of a simulation study, and illustrates the use of `GJRM` in the three case studies.

2 The Model

In switching regressions models, a random variable of interest is explained in different ways across alternate regimes. When there are two regimes, the model has a trio of underlying random variables $(Y_{1i}^*, Y_{2i}^*, Y_{3i}^*)$ which connect with observable random variables (S_i, Y_{2i}, Y_{3i}) via the rules

$$S_i = \mathbb{1}(Y_{1i}^* > 0), \quad Y_{2i} = (1 - S_i)Y_{2i}^*, \quad Y_{3i} = S_i Y_{3i}^*,$$

where $\mathbb{1}(\cdot)$ is an indicator function equaling 1 if the condition inside the braces holds and 0 otherwise. These rules imply a binary switching mechanism: if $S_i = 0$ then Y_{2i} holds the observed value of Y_{2i}^* and Y_{3i} equals 0 (which here means that Y_{3i} is missing or unobserved), and if $S_i = 1$ then Y_{3i} holds the observed value of Y_{3i}^* and Y_{2i} equals 0. Note that the first rule implies $P(S_i = 0) = P(Y_{1i}^* \leq 0)$, i.e. the cumulative distribution functions (cdfs) of S_i and Y_{1i}^* coincide at $s_i = y_{1i}^* = 0$.

Each member of the trio $(Y_{1i}^*, Y_{2i}^*, Y_{3i}^*)$ has associated marginal cdfs and probability (density or mass) functions (pdfs/pmfs) which can be denoted as $F_j(y_{ji}^*|\boldsymbol{\varphi}_j)$ and $f_j(y_{ji}^*|\boldsymbol{\varphi}_j)$, for $j = 1, 2, 3$,

where $\boldsymbol{\varphi}_j$ represents a vector of distributional parameters of dimension $w_j \in \mathbb{N}^+$ that can be specified as flexible functions of regression effects (as explained in Section 2.3). Recall that the equation related to the switching mechanism requires an instrument. The model also requires bivariate cdfs that relate the first variable in the trio to the other two variables, that is $F_{12}(y_{1i}^*, y_{2i}^*|\boldsymbol{\varphi}_1, \boldsymbol{\varphi}_2, \theta_{12})$ and $F_{13}(y_{1i}^*, y_{3i}^*|\boldsymbol{\varphi}_1, \boldsymbol{\varphi}_3, \theta_{13})$, where θ_{12} and θ_{13} are parameters capturing the dependence between the respective margins.

The next section gives details on the form of the log-likelihood for different types of outcome variables. Section 2.2 discusses the range of options for the specification of the copula and marginal distributions. Section 2.3 explains how flexible regression structures can be accommodated in the modelling framework.

2.1 Log-Likelihoods

Let (s_i, y_{2i}, y_{3i}) denote the i^{th} observation on (S_i, Y_{2i}, Y_{3i}), for $i = 1, \ldots, n$, where $n \in \mathbb{N}^+$ denotes the sample size. For a given observed random sample, the log-likelihood function can be expressed in three different ways depending on whether the outcome variable of interest is continuous, discrete or binary. In the continuous case, as in Smith (2005), we have

$$\begin{aligned}\ell(\boldsymbol{\delta}) = \sum_{i=1}^{n} (1 - s_i) \log \left\{ \frac{\partial F_{12}(0, y_{2i}^*|\boldsymbol{\varphi}_1, \boldsymbol{\varphi}_2, \theta_{12})}{\partial y_{2i}^*} \right\} \Bigg|_{y_{2i}^* = y_{2i}} + \\ s_i \log \left\{ f_3(y_{3i}^*|\boldsymbol{\varphi}_3) - \frac{\partial F_{13}(0, y_{3i}^*|\boldsymbol{\varphi}_1, \boldsymbol{\varphi}_3, \theta_{13})}{\partial y_{3i}^*} \right\} \Bigg|_{y_{3i}^* = y_{3i}}\end{aligned}, \tag{1}$$

where $\boldsymbol{\delta} = (\boldsymbol{\varphi}_1^\mathsf{T}, \boldsymbol{\varphi}_2^\mathsf{T}, \boldsymbol{\varphi}_3^\mathsf{T}, \theta_{12}, \theta_{13})^\mathsf{T}$. In the discrete case, the log-likelihood function is instead built using finite differences, i.e.

$$\begin{aligned}\ell(\boldsymbol{\delta}) = \sum_{i=1}^{n} (1 - s_i) \log \left\{ F_{12}(0, y_{2i}^*|\boldsymbol{\varphi}_1, \boldsymbol{\varphi}_2, \theta_{12}) - F_{12}(0, y_{2i}^* - 1|\boldsymbol{\varphi}_1, \boldsymbol{\varphi}_2, \theta_{12}) \right\} \Bigg|_{y_{2i}^* = y_{2i}} + \\ s_i \log \left\{ f_3(y_{3i}^*|\boldsymbol{\varphi}_3) - F_{13}(0, y_{3i}^*|\boldsymbol{\varphi}_1, \boldsymbol{\varphi}_3, \theta_{13}) + F_{13}(0, y_{3i}^* - 1|\boldsymbol{\varphi}_1, \boldsymbol{\varphi}_3, \theta_{13}) \right\} \Bigg|_{y_{3i}^* = y_{3i}}\end{aligned}. \tag{2}$$

In the binary scenario, we have four possible outcomes and hence

$$\begin{aligned}\ell(\boldsymbol{\delta}) = \sum_{i=1}^{n} & (1-s_i)(1-y_{2i}) \log \{F_{12}(0,0|\boldsymbol{\varphi}_1, \boldsymbol{\varphi}_2, \theta_{12})\} + \\ & (1-s_i) y_{2i} \log \{F_1(0|\boldsymbol{\varphi}_1) - F_{12}(0,0|\boldsymbol{\varphi}_1, \boldsymbol{\varphi}_2, \theta_{12})\} + \\ & s_i (1-y_{3i}) \log \{F_3(0|\boldsymbol{\varphi}_3) - F_{13}(0,0|\boldsymbol{\varphi}_1, \boldsymbol{\varphi}_3, \theta_{13})\} + \\ & s_i y_{3i} \log \left[(1 - F_1(0|\boldsymbol{\varphi}_1)) - \{F_3(0|\boldsymbol{\varphi}_3) - F_{13}(0,0|\boldsymbol{\varphi}_1, \boldsymbol{\varphi}_3, \theta_{13})\}\right]\end{aligned} \quad . \tag{3}$$

Note that, since the log-likelihood functions above do not depend on $F_{23}(y_{2i}^*, y_{3i}^*|\boldsymbol{\varphi}_2, \boldsymbol{\varphi}_3, \theta_{23})$, where θ_{23} would represent a further dependence parameter, potential links between Y_{2i}^* and Y_{3i}^* cannot be (directly) recovered (e.g., Smith, 2005), hence it is superfluous to specify a trivariate distribution for $(Y_{1i}^*, Y_{2i}^*, Y_{3i}^*)$.

2.2 Copulae and Marginal Distributions

This section provides a very succinct description of the copula approach; we refer the reader to, e.g., Nelsen (2006), Nikoloulopoulos & Karlis (2010), Trivedi and Zimmer (2007) and Joe (2014) for technical details in various contexts. Using the copula method, the joint cdf of the random variables of interest can be expressed as

$$F_{1j}(y_1^*, y_j^*|\boldsymbol{\varphi}_1, \boldsymbol{\varphi}_j, \theta_{1j}) = C_{1j}\left(F_1(y_1^*|\boldsymbol{\varphi}_1), F_2(y_j^*|\boldsymbol{\varphi}_j); \theta_{1j}\right), \ j = 2, 3, \tag{4}$$

where $C_{1j} : (0,1)^2 \to (0,1)$ is a two-place copula function. The main practical advantage of copulae is that, with knowledge of arbitrary F_1 and F_j and a copula function C_{1j} that glues them together, one can assemble a distribution of the otherwise difficult-to-know F_{1j}. The copulae implemented in `GJRM` are reported in Table 1. For those copulae that can only account for positive dependence (e.g., Clayton and Joe), counter-clockwise rotated versions of them can be obtained (Brechmann & Schepsmeier, 2013). For a pictorial representation of some of the copulae considered here see, e.g., Hasebe (2013).

Using (4), log-likelihood function (1) is made operational by replacing $F_{12}(y_{1i}^*, y_{2i}^*|\boldsymbol{\varphi}_1, \boldsymbol{\varphi}_2, \theta_{12})$

Copula	$C(u_1, u_2; \theta)$	Range of θ	Transformation of θ
AMH (`"AMH"`)	$\frac{u_1 u_2}{1-\theta(1-u_1)(1-u_2)}$	$[-1, 1]$	$\tanh^{-1}(\theta)$
Clayton (`"C0"`)	$\left(u_1^{-\theta} + u_2^{-\theta} - 1\right)^{-1/\theta}$	$(0, \infty)$	$\log(\theta)$
FGM (`"FGM"`)	$u_1 u_2 \{1 + \theta(1-u_1)(1-u_2)\}$	$[-1, 1]$	$\tanh^{-1}(\theta)$
Frank (`"F"`)	$-\theta^{-1} \log\{1 + (\exp\{-\theta u_1\} - 1)$ $(\exp\{-\theta u_2\} - 1)/(\exp\{-\theta\} - 1)\}$	$\mathbb{R} \setminus \{0\}$	–
Galambos (`"GAL"`)	$u_1 u_2 \exp\left[\{(-\log u_1)^{-\theta}$ $+(-\log u_2)^{-\theta}\}^{-1/\theta}\right]$	$(0, \infty)$	$\log(\theta)$
Gaussian (`"N"`)	$\Phi_2\left(\Phi^{-1}(u_1), \Phi^{-1}(u_2); \theta\right)$	$[-1, 1]$	$\tanh^{-1}(\theta)$
Gumbel (`"G0"`)	$\exp\left[-\{(-\log u_1)^{\theta}$ $+(-\log u_2)^{\theta}\}^{1/\theta}\right]$	$[1, \infty)$	$\log(\theta - 1)$
Joe (`"J0"`)	$1 - \{(1-u_1)^{\theta} + (1-u_2)^{\theta}$ $-(1-u_1)^{\theta}(1-u_2)^{\theta}\}^{1/\theta}$	$(1, \infty)$	$\log(\theta - 1)$
Plackett (`"PL"`)	$\left(Q - \sqrt{R}\right) / \{2(\theta - 1)\}$	$(0, \infty)$	$\log(\theta)$
Student's t (`"T"`)	$t_{2,\zeta}\left(t_\zeta^{-1}(u_1), t_\zeta^{-1}(u_2); \zeta, \theta\right)$	$[-1, 1]$	$\tanh^{-1}(\theta)$

Table 1: Definition of the copulae implemented in the `R` package `GJRM`, with corresponding parameter range of association parameter θ, and one-to-one transformation function of θ. $\Phi_2(\cdot, \cdot; \theta)$ denotes the cumulative distribution function (cdf) of the standard bivariate normal distribution with correlation coefficient θ, and $\Phi(\cdot)$ the cdf of the univariate standard normal distribution. $t_{2,\zeta}(\cdot, \cdot; \zeta, \theta)$ indicates the cdf of the standard bivariate Student-t distribution with correlation θ and fixed $\zeta \in (2, \infty)$ degrees of freedom, and $t_\zeta(\cdot)$ denotes the cdf of the univariate Student-t distribution with ζ degrees of freedom. Quantities Q and R are given by $1 + (\theta - 1)(u_1 + u_2)$ and $Q^2 - 4\theta(\theta - 1)u_1 u_2$, respectively. Arguments `BivD` and `BivD2` of `gjrm()` in `GJRM` allows the user to employ the desired copulae and can be set to any of the values within brackets next to the copula names in the first column; for example, `BivD = "C0"` and `BivD2 = "FGM"`. For Clayton, Galambos, Gumbel and Joe, the number after the capital letter indicates the degree of rotation required: the possible values are `0`, `90`, `180` and `270`. The rotations are defined as $C_{90}(u_1, u_2; \theta) = u_2 - C(1 - u_1, u_2)$, $C_{180}(u_1, u_2; \theta) = u_1 + u_2 - 1 + C(1 - u_1, 1 - u_2)$ and $C_{270}(u_1, u_2; \theta) = u_1 - C(u_1, 1 - u_2)$.

	$f(y\|\mu,\sigma)$	$\mathbb{E}(Y)$	$\mathbb{V}(Y)$
Poisson (`"PO"`)	$\frac{\exp(-\mu)\mu^y}{y!}$	μ	μ
Negative binomial type I (`"NBI"`)	$\frac{\Gamma(y+1/\sigma)}{\Gamma(1/\sigma)\Gamma(y+1)}\left(\frac{\sigma\mu}{1+\sigma\mu}\right)^y\left(\frac{1}{1+\sigma\mu}\right)^{1/\sigma}$	μ	$\mu+\sigma\mu^2$
Negative binomial type II (`"NBII"`)	$\frac{\Gamma(y+\mu/\sigma)\sigma^y}{\Gamma(\mu/\sigma)\Gamma(y+1)(1+\sigma)^{y+\mu/\sigma}}$	μ	$(1+\sigma)\mu$
Poisson inverse Gaussian (`"PIG"`)	$\left(\frac{2\alpha}{\pi}\right)^{0.5}\frac{\mu^y\exp(1/\sigma)K_{y-0.5}(\alpha)}{(\alpha\sigma)^y y!}$	μ	$\mu+\sigma\mu^2$

Table 2: Definition and some properties of the main discrete distributions implemented in `GJRM`. These have been parametrised according to Rigby & Stasinopoulos (2005) and are defined in terms of parameters μ and σ. In all cases, $y \in \mathbb{N}_0$ and $\mu, \sigma \in (0, \infty)$. Since the distributional parameters can only take positive values, the transformation function $\log(\cdot)$ is employed in all cases. $\alpha = \sqrt{\frac{1}{\sigma^2} + \frac{2\mu}{\sigma}}$ and $K_\lambda(t) = \frac{1}{2}\int_0^\infty x^{\lambda-1}\exp\left\{-0.5t(x+x^{-1})\right\}dx$ is the modified Bessel function of the third kind. Argument `margins` of `gjrm()` in `GJRM` allows the user to employ the desired discrete marginals This is achieved using the characters within brackets next to the names of the distributions; for instance, `margins = c("logit", "NBI", "PIG")`. Note that for the first margin other choices are possible: `probit` and `cloglog`.

and $F_{13}(y_{1i}^*, y_{3i}^*|\boldsymbol{\varphi}_1, \boldsymbol{\varphi}_3, \theta_{13})$ with $C_{12}\left(F_1(0|\boldsymbol{\varphi}_1), F_2(y_{2i}|\boldsymbol{\varphi}_2); \theta_{12}\right)$ and $C_{13}\left(F_1(0|\boldsymbol{\varphi}_1), F_3(y_{3i}|\boldsymbol{\varphi}_3); \theta_{13}\right)$, respectively. A similar reasoning applies to log-likelihood functions (2) and (3). Note that in (2), $F_{12}(y_{1i}^*, y_{2i}^* - 1|\boldsymbol{\varphi}_1, \boldsymbol{\varphi}_2, \theta_{12})$ is replaced with $C_{12}\left(F_1(0|\boldsymbol{\varphi}_1), F_2(y_{2i}|\boldsymbol{\varphi}_2) - f_2(y_{2i}|\boldsymbol{\varphi}_2); \theta_{12}\right)$, where, for the second margin, the relation $f_2(y_{2i}|\boldsymbol{\varphi}_2) = F_2(y_{2i}|\boldsymbol{\varphi}_2) - F_2(y_{2i} - 1|\boldsymbol{\varphi}_2)$ is exploited to avoid the evaluation of F_2 for negative arguments; similarly for F_{13}.

Regarding the marginal distributions, for S we consider a Bernoulli distribution with parameter $\mu_1 \in (0, 1)$, representing the probability of switching. For Y_2 and Y_3 several choices are possible: a Bernoulli distribution and those listed in Tables 2 and 3. These choices imply that $\boldsymbol{\varphi}_1 = \mu_1$ and that $\boldsymbol{\varphi}_j$, for $j = 2, 3$, is equal to either μ_j, $(\mu_j, \sigma_j)^\intercal$ or $(\mu_j, \sigma_j, \nu_j)^\intercal$.

2.3 Flexible Covariate Effects

Each parameter of the model can be linked to regression effects via an unknown smooth function $m(\mathbf{z}_i) \in \mathbb{R}$, where $\mathbf{z}_i$ represents a covariate vector (containing, e.g., binary, categorical, continuous and geographic variables), and a known monotonic one-to-one transformation function ensuring that the restriction on the space of the parameter being considered is not violated. As an example, for a model with Bernoulli, Fisk and Gumbel margins and Gumbel and Clayton copulae, we would have $g_{\mu_1}(\mu_{1i}) = m_{\mu_1}(\mathbf{z}_i)$, $g_{\mu_2}(\mu_{2i}) = m_{\mu_2}(\mathbf{z}_i)$, $g_{\sigma_2}(\sigma_{2i}) = m_{\sigma_2}(\mathbf{z}_i)$,

	$F(y\|\mu,\sigma,\nu)$	$f(y\|\mu,\sigma,\nu)$	$\mathbb{E}(Y)$	$\mathbb{V}(Y)$	Support of y Parameters' ranges
beta (`"BE"`)	$I(y;\alpha_1,\alpha_2)$ $\alpha_1=\frac{\mu(1-\sigma^2)}{\sigma^2}$ $\alpha_2=\frac{(1-\mu)(1-\sigma^2)}{\sigma^2}$	$\frac{y^{\alpha_1-1}(1-y)^{\alpha_2-1}}{B(\alpha_1,\alpha_2)}$	μ	$\sigma^2\mu(1-\mu)$	$0<y<1$ $0<\mu<1, 0<\sigma<1$
Dagum (`"DAGUM"`)	$\left\{1+\left(\frac{y}{\mu}\right)^{-\sigma}\right\}^{-\nu}$	$\frac{\sigma\nu}{y}\left[\frac{\left(\frac{y}{\mu}\right)^{\sigma\nu}}{\left\{\left(\frac{y}{\mu}\right)^{\sigma}+1\right\}^{\nu+1}}\right]$	$-\frac{\mu}{\sigma}\frac{\Gamma(-\frac{1}{\sigma})\Gamma(\frac{1}{\sigma}+\nu)}{\Gamma(\nu)}$ if $\sigma>1$	$-\left(\frac{\mu}{\sigma}\right)^2\left[2\sigma\frac{\Gamma(-\frac{2}{\sigma})\Gamma(\frac{2}{\sigma}+\nu)}{\Gamma(\nu)}+\left\{\frac{\Gamma(-\frac{1}{\sigma})\Gamma(\frac{1}{\sigma}+\nu)}{\Gamma(\nu)}\right\}^2\right]$ if $\sigma>2$	$y>0$ $\mu>0,\sigma>0,\nu>0$
Fisk (`"FISK"`)	$\left\{1+\left(\frac{y}{\mu}\right)^{-\sigma}\right\}^{-1}$	$\frac{\sigma y^{\sigma-1}}{\mu^{\sigma}\left\{1+\left(\frac{y}{\mu}\right)^{\sigma}\right\}^2}$	$\frac{\mu\pi/\sigma}{\sin(\pi/\sigma)}$ if $\sigma>1$	$\mu^2\left\{\frac{2\pi/\sigma}{\sin(2\pi/\sigma)}-\frac{(\pi/\sigma)^2}{\sin(\pi/\sigma)^2}\right\}$ if $\sigma>2$	$y>0$ $\mu>0,\sigma>0$
gamma (`"GA"`)	$\frac{1}{\Gamma\left(\frac{1}{\sigma^2}\right)}\gamma\left(\frac{1}{\sigma^2},\frac{y}{\mu\sigma^2}\right)$	$\frac{1}{(\mu\sigma^2)^{\frac{1}{\sigma^2}}}\frac{y^{\frac{1}{\sigma^2}-1}\exp\left(-\frac{y}{\mu\sigma^2}\right)}{\Gamma\left(\frac{1}{\sigma^2}\right)}$	μ	$\mu^2\sigma^2$	$y>0$ $\mu>0,\sigma>0$
Gumbel (`"GU"`)	$1-\exp\left\{-\exp\left(\frac{y-\mu}{\sigma}\right)\right\}$	$\frac{1}{\sigma}\exp\left\{\left(\frac{y-\mu}{\sigma}\right)-\exp\left(\frac{y-\mu}{\sigma}\right)\right\}$	$\mu-0.57722\sigma$	$\frac{\pi^2\sigma^2}{6}$	$-\infty<y<\infty$ $-\infty<\mu<\infty,\sigma>0$
inverse Gaussian (`"iG"`)	$\Phi\left\{\frac{1}{\sqrt{y\sigma^2}}\left(\frac{y}{\mu}-1\right)\right\}+\exp\left(\frac{2}{\mu\sigma^2}\right)\Phi\left\{-\frac{1}{\sqrt{y\sigma^2}}\left(\frac{y}{\mu}+1\right)\right\}$	$\frac{1}{\sqrt{2\pi\sigma^2y^3}}\exp\left\{-\frac{1}{2\mu^2\sigma^2y}(y-\mu)^2\right\}$	μ	$\mu^3\sigma^2$	$y>0$ $\mu>0,\sigma>0$
log-normal (`"LN"`)	$\frac{1}{2}+\frac{1}{2}\text{erf}\left\{\frac{\log(y)-\mu}{\sigma\sqrt{2}}\right\}$	$\frac{1}{y\sigma\sqrt{2\pi}}\exp\left[-\frac{\{\log(y)-\mu\}^2}{2\sigma^2}\right]$	$\sqrt{\exp(\sigma^2)}\exp(\mu)$	$\exp(\sigma^2)\{\exp(\sigma^2)-1\}\exp(2\mu)$	$y>0$ $-\infty<\mu<\infty,\sigma>0$
logistic (`"LO"`)	$\frac{1}{1+\exp\left(-\frac{y-\mu}{\sigma}\right)}$	$\frac{1}{\sigma}\left\{\exp\left(-\frac{y-\mu}{\sigma}\right)\right\}\left\{1+\exp\left(-\frac{y-\mu}{\sigma}\right)\right\}^{-2}$	μ	$\frac{\pi^2\sigma^2}{3}$	$-\infty<y<\infty$ $-\infty<\mu<\infty,\sigma>0$
normal (`"N"`)	$\frac{1}{2}\left\{1+\text{erf}\left(\frac{y-\mu}{\sigma\sqrt{2}}\right)\right\}$	$\frac{1}{\sigma\sqrt{2\pi}}\exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\}$	μ	σ^2	$-\infty<y<\infty$ $-\infty<\mu<\infty,\sigma>0$
reverse Gumbel (`"rGU"`)	$\exp\left\{-\exp\left(-\frac{y-\mu}{\sigma}\right)\right\}$	$\frac{1}{\sigma}\exp\left\{\left(-\frac{y-\mu}{\sigma}\right)-\exp\left(-\frac{y-\mu}{\sigma}\right)\right\}$	$\mu+0.57722\sigma$	$\frac{\pi^2\sigma^2}{6}$	$-\infty<y<\infty$ $-\infty<\mu<\infty,\sigma>0$
Singh-Maddala (`"SM"`)	$1-\left\{1+\left(\frac{y}{\mu}\right)^{\sigma}\right\}^{-\nu}$	$\frac{\sigma\nu y^{\sigma-1}}{\mu^{\sigma}\left\{1+\left(\frac{y}{\mu}\right)^{\sigma}\right\}^{\nu+1}}$	$\mu\frac{\Gamma(1+\frac{1}{\sigma})\Gamma(-\frac{1}{\sigma}+\nu)}{\Gamma(\nu)}$ if $\sigma\nu>1$	$\mu^2\left\{\Gamma\left(1+\frac{2}{\sigma}\right)\Gamma(\nu)\Gamma\left(-\frac{2}{\sigma}+\nu\right)-\Gamma\left(1+\frac{1}{\sigma}\right)^2\Gamma\left(-\frac{1}{\sigma}+\nu\right)^2\right\}$ if $\sigma\nu>2$	$y>0$ $\mu>0,\sigma>0,\nu>0$
Weibull (`"WEI"`)	$1-\exp\left\{-\left(\frac{y}{\mu}\right)^{\sigma}\right\}$	$\frac{\sigma}{\mu}\left(\frac{y}{\mu}\right)^{\sigma-1}\exp\left\{-\left(\frac{y}{\mu}\right)^{\sigma}\right\}$	$\mu\Gamma\left(\frac{1}{\sigma}+1\right)$	$\mu^2\left[\Gamma\left(\frac{2}{\sigma}+1\right)-\left\{\Gamma\left(\frac{1}{\sigma}+1\right)\right\}^2\right]$	$y>0$ $\mu>0,\sigma>0$

Table 3: Definition and some properties of the main distributions implemented in `GJRM`. These have been conveniently parametrised according to Rigby & Stasinopoulos (2005) and are defined in terms of parameters μ, σ and ν (which sometimes represent location, scale and shape). Note that $\mathbb{E}(Y)$ and $\mathbb{V}(Y)$ of `DAGUM`, `FISK` (also known as log-logistic) and `SM` are indeterminate for certain values (or combination) of σ and ν. Also, in many cases the parameters of the distributions determine $\mathbb{E}(Y)$ and $\mathbb{V}(Y)$ through functions of them. If a parameter can only take positive values then the transformation function $\log(\cdot)$ is employed. If a parameter takes values in $(0,1)$ then the inverse of the cumulative distribution function of the standardised logistic distribution is used. $I(\cdot;\cdot,\cdot)$ is the regularized beta function, $B(\cdot,\cdot)$ is the beta function, $\Gamma(\cdot)$ is the gamma function, $\gamma(\cdot,\cdot)$ is the lower incomplete gamma function, $\Phi(\cdot)$ is the cdf of the univariate standard normal distribution, and $\text{erf}(\cdot)$ is the error function. Argument `margins` of `gjrm()` in `GJRM` allows the user to employ the desired marginals; for instance, `margins = c("probit", "iG", "FISK")`, where the first margin can also be `logit` or `cloglog`.

$g_{\mu_3}(\mu_{3i}) = m_{\mu_3}(\mathbf{z}_i)$, $g_{\sigma_3}(\sigma_{3i}) = m_{\sigma_3}(\mathbf{z}_i)$, $g_{\theta_{12}}(\theta_{12i}) = m_{\theta_{12}}(\mathbf{z}_i)$, $g_{\theta_{13}}(\theta_{13i}) = m_{\theta_{13}}(\mathbf{z}_i)$, where $g_{\mu_1}(\mu_{1i}) = \Phi^{-1}(\mu_{1i})$, with $\Phi^{-1}(\cdot)$ being the quantile function a standard normal distribution, $g_{\mu_2}(\mu_{2i}) = \log(\mu_{2i})$, $g_{\sigma_2}(\sigma_{2i}) = \log(\sigma_{2i})$, $g_{\mu_3}(\mu_{3i}) = \mu_{3i}$, $g_{\sigma_3}(\sigma_{3i}) = \log(\sigma_{3i})$, $g_{\theta_{12}}(\theta_{12i}) = \log(\theta_{12i}-1)$ and $g_{\theta_{13}}(\theta_{13i}) = \log(\theta_{13i})$. For the binary margin, we assumed a probit link, however the logit and complementary log-log functions could have been chosen instead. Furthermore, we assumed that the same covariate vector is employed for each parameter, however, if desired, different subsets of it can be adopted for different parameters.

Using $m(\mathbf{z}_i)$ makes the model very flexible. However, in practice n would have to be unfeasibly large due to the well known curse-of-dimensionality when the dimension of $\mathbf{z}_i$ is large, as in most empirical situations. To this end, we impose an additive structure on $m(\mathbf{z}_i)$ which, while it implies that not all the interaction terms among the covariates can be accounted for, still allows for a great deal of flexibility and retains good theoretical properties (e.g., Wood, 2017). Therefore, dropping for simplicity the subscript denoting which parameter the smooth function belongs to, we define

$$m(\mathbf{z}_i) = \beta_0 + \sum_{k=1}^{K} s_k(\mathbf{z}_{ki}), \tag{5}$$

where $\beta_0 \in \mathbb{R}$ is an overall intercept, $\mathbf{z}_{ki}$ denotes the k^{th} sub-vector of $\mathbf{z}_i$ and the K functions $s_k(\mathbf{z}_{ki})$ represent generic effects chosen according to the type of covariate(s) considered, as explained in the next sections. Each $s_k(\mathbf{z}_{ki})$ can be expressed as a linear combination of J_k basis functions $b_{kj_k}(\mathbf{z}_{ki})$ and regression coefficients $\beta_{kj_k} \in \mathbb{R}$,

$$\sum_{j_k=1}^{J_k} \beta_{kj_k} b_{kj_k}(\mathbf{z}_{ki}). \tag{6}$$

This means that the vector of evaluations $\{s_k(\mathbf{z}_{k1}), \ldots, s_k(\mathbf{z}_{kn})\}^\mathsf{T}$ can be written as $\mathbf{Z}_k\boldsymbol{\beta}_k$ with $\boldsymbol{\beta}_k = (\beta_{k1}, \ldots, \beta_{kJ_k})^\mathsf{T}$ and design matrix $Z_k[i, j_k] = b_{kj_k}(\mathbf{z}_{ki})$. This allows the right hand side of (5) to be written as $\beta_0\mathbf{1}_n + \mathbf{Z}_1\boldsymbol{\beta}_1 + \ldots + \mathbf{Z}_K\boldsymbol{\beta}_K$ or as $\mathbf{Z}\boldsymbol{\beta}$, where $\mathbf{Z} = (\mathbf{1}_n, \mathbf{Z}_1, \ldots, \mathbf{Z}_K)$, $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_1^\mathsf{T}, \ldots, \boldsymbol{\beta}_K^\mathsf{T})^\mathsf{T}$ and $\mathbf{1}_n$ is an n-dimensional vector made up of ones.

Each $\boldsymbol{\beta}_k$ has an associated quadratic penalty $\lambda_k\boldsymbol{\beta}_k^\mathsf{T}\mathbf{D}_k\boldsymbol{\beta}_k$ whose role is to enforce specific properties on the k^{th} function, such as smoothness. Here, $\mathbf{D}_k$ only depends on the choice of

basis functions and hence not on $\boldsymbol{\beta}_k$. Tuning or smoothing parameter $\lambda_k \in [0, \infty)$ controls the trade-off between fit and parsimony, and plays a crucial role in determining $\hat{s}_k(\mathbf{z}_{ki})$; a large value for λ_k will allow the related quadratic penalty to have a large influence on the estimation of $\boldsymbol{\beta}_k$ during model fitting, and vice versa. The overall penalty can be defined as $\boldsymbol{\beta}^\top \mathbf{D}_{\boldsymbol{\lambda}} \boldsymbol{\beta}$, where $\mathbf{D}_{\boldsymbol{\lambda}} = \text{diag}(0, \lambda_1 \mathbf{D}_1, \ldots, \lambda_K \mathbf{D}_K)$. To ensure identifiability of the model regression structure, the $s_k(\mathbf{z}_{ki})$ are subject to centering constraints which are imposed by adopting the parsimonious approach detailed in Wood (2017, Section 5.4.1). The next sections show how the above framework can be utilised to model, among others, linear and nonlinear effects; these are relevant for our case studies.

2.3.1 Parametric effects

These effects usually relate to binary and categorical variables and are represented by setting $s_k(\mathbf{z}_{ki}) = \mathbf{z}_{ki}^\top \boldsymbol{\beta}_k$. The corresponding design matrix is obtained by stacking all covariate vectors $\mathbf{z}_{ki}$ into $\mathbf{Z}_k$. No penalty is typically assigned to parametric effects, hence $\mathbf{D}_k = \mathbf{0}$. There might be, however, contexts in which it would be advisable to do so. An example is that of a factor variable with many categories and only a few observations available for some of them. The parameters of such categories may be weakly or not identified by the data in which case a ridge penalty (obtained by setting $\mathbf{D}_k = \mathbf{I}_k$, where $\mathbf{I}_k$ is an identity matrix) can be employed to circumvent such problem. Note that this is equivalent to the assumption that the coefficients are *i.i.d.* normal random effects with unknown variance (Wood, 2017, Section 5.8).

2.3.2 Nonparametric effects

Penalised regression splines are a popular and computationally efficient way for representing unknown nonlinear effects of continuous covariates. The main requirement is a global smoothness assumption on differentiability. This method makes it possible to avoid arbitrary modelling decisions, such as choosing the appropriate degree of a polynomial or specifying cut-points, which could induce misspecification. This approach has been popularised by Eilers & Marx (1996) and its theoretical properties addressed by several authors (e.g., Claeskens et al., 2009;

Kauermann et al., 2009; Wood, 2017).

For a continuous variable z_{ki}, we use representation (6), where the $b_{kj_k}(z_{ki})$ are known spline basis functions. The design matrix $\mathbf{Z}_k$ comprises the basis function evaluations for each i, and essentially contains J_k curves with varying degrees of complexity. To enforce smoothness, the penalty is based on the conventional choice $\mathbf{D}_k = \int \mathbf{d}_k(z_k)\mathbf{d}_k(z_k)^\intercal dz_k$, where the j_k^{th} element of $\mathbf{d}_k(z_k)$ is given by $\partial^2 b_{kj_k}(z_k)/\partial z_k^2$ and integration is over the range of z_k. This approach can virtually accommodate any (sensible) definition of basis function and penalty (e.g., penalised low rank thin plate regression splines, P-splines) and we refer the reader to Wood (2017, Chapter 5) for various definitions as well as a through discussion of their theoretical and computational aspects.

When setting up the basis functions, knots have to be chosen unless not required (e.g., thin plate regression splines, Wood, 2017, Section 5.5.1). A value for J_k has to be chosen too. Instead of addressing these problems, which may be computationally cumbersome, the penalised regression spline method relies on setting J_k to a large number and then using a penalty during model fitting to suppress that part of the smooth term complexity that is not supported by the data. Still, J_k has to be chosen and the theoretical analysis of the above authors suggest that J_k has to grow slowly with n to achieve statistical performance that is asymptotically indistinguishable from that of a full smoothing spline. In practice, one may perform a sensitivity analysis to assess how the estimates change for several values of J_k.

The penalised regression smoothing framework described in this section allows for several other specifications as well as a vast variety of penalised spline functions. These include interaction terms via varying coefficient smooths obtained by multiplying one or more smooth components by some covariate(s), smooth functions of two or more continuous covariates (tensor product terms), and Gaussian Markov random field, Gaussian process and adaptive smoothers, to name but a few. Such flexibility makes the scope of the introduced models and the related implementation in `GJRM` very wide and hence applicable to a large suite of empirical problems.

3 Parameter Estimation

For a given observed random sample $\{(s_i, y_{2i}, y_{3i}, \mathbf{z}_i)\}_{i=1}^n$, because of the highly flexible regression structures allowed for by the proposed modelling framework, parameter estimation is based on an objective function augmented by an overall quadratic penalty term which is set up using the approach discussed in Section 2.3. That is,

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

where $\ell(\boldsymbol{\delta})$ is equal to either (1), (2) or (3), $\boldsymbol{\delta}$ is defined as $(\boldsymbol{\beta}_{\mu_1}^\mathsf{T}, \ldots, \boldsymbol{\beta}_{\theta_{12}}^\mathsf{T}, \boldsymbol{\beta}_{\theta_{13}}^\mathsf{T})^\mathsf{T}$, which is made up of the coefficient vectors related to $m_{\mu_1}(\mathbf{z}_i)$, ..., $m_{\theta_{12}}(\mathbf{z}_i)$, $m_{\theta_{13}}(\mathbf{z}_i)$ and whose specific set up and dimension depend on the model specification, and $\mathbf{S}_{\boldsymbol{\lambda}} = \text{diag}(\mathbf{D}_{\boldsymbol{\lambda},\mu_1}, \ldots, \mathbf{D}_{\boldsymbol{\lambda},\theta_{12}}, \mathbf{D}_{\boldsymbol{\lambda},\theta_{13}})$, which contains the overall smoothing parameter vector $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_{\mu_1}^\mathsf{T}, \ldots, \boldsymbol{\lambda}_{\theta_{12}}^\mathsf{T}, \boldsymbol{\lambda}_{\theta_{13}}^\mathsf{T})^\mathsf{T}$.

The construction of an algorithm that can estimate $\boldsymbol{\delta}$ and $\boldsymbol{\lambda}$ in a stable and efficient manner requires careful considerations and attention to certain details. For instance, experimentation based on various optimisation schemes that rely on derivative free and quasi-Newton techniques revealed a series of convergence and speed issues; in the cases explored, throughout the iterations, the score and Hessian of $\ell(\boldsymbol{\delta})$ were poorly approximated by numerical differentiation methods. We finally settled on a simultaneous estimation approach, based on analytical first and second order derivatives, implemented by adapting to this context the algorithm of Marra & Radice (2020) (see Appendix A of the On-Line Supplementary Material for details).

4 Further considerations

At convergence, instead of basing inference on the classically derived frequentist covariance matrix $-\boldsymbol{H}_p^{-1}\boldsymbol{H}\boldsymbol{H}_p^{-1}$, where $\boldsymbol{H}_p$ and $\boldsymbol{H}$ are the Hessians of $\ell_p(\boldsymbol{\delta})$ and $\ell(\boldsymbol{\delta})$, intervals for any linear function of $\boldsymbol{\delta}$, e.g. $s_k(z_{ki})$, are obtained via the Bayesian large sample approximation

$$\boldsymbol{\delta} \overset{a}{\sim} \mathcal{N}(\hat{\boldsymbol{\delta}}, -\boldsymbol{H}_p^{-1}). \tag{7}$$

Adopting the Bayesian framework in the context of penalised models implicitly assumes that overly complex models are less likely than simpler or smoother ones which indeed translates into the prior specification $f_{\boldsymbol{\delta}} \propto \exp\left(-1/2\boldsymbol{\delta}^{\mathsf{T}}\mathbf{S}_{\boldsymbol{\lambda}}\boldsymbol{\delta}\right)$. As elaborated by Wood (2017, Section 6.10, see also references therein), the Bayesian covariance matrix gives close to across-the-function frequentist coverage probabilities since it includes the bias and variance components in a frequentist sense, which is not the case for the frequentist covariance matrix. Intervals for nonlinear functions of $\boldsymbol{\delta}$ (see the next section for an example) can be conveniently obtained via posterior simulation, whereas p-values for the terms in the model can be reliably obtained by using the results summarised in Wood (2017, Section 6.12) which are based on $-\boldsymbol{H}_p^{-1}$.

Model building can be aided using tools such as the Akaike information criterion (AIC, Akaike, 1973) and the Bayesian information criterion (BIC, Schwarz, 1978), and (randomised) normalised quantile residuals (Dunn & Smyth, 1996). The AIC and BIC are defined as $-2\ell(\hat{\boldsymbol{\delta}})+2edf$ and $-2\ell(\hat{\boldsymbol{\delta}})+\log(n)edf$, respectively, where the log-likelihood is evaluated at the penalised parameter estimates and $edf = \text{tr}(\hat{\mathbf{A}})$ with $\mathbf{A} = \sqrt{-\boldsymbol{H}}\left(-\boldsymbol{H}_p\right)^{-1}\sqrt{-\boldsymbol{H}}$ (see Appendix A for details on the derivation of this quantity).

As for the residuals, for a continuous Y_2, these are defined as $q_{2i_2} = \Phi^{-1}\left\{\widehat{F}_{2|1}(y_{2i_2}|y_{1i_2}=0)\right\}$, where $i_2 = 1,\ldots,n_2$, n_2 is the size of the sub-sample related to Y_2 and $\widehat{F}_{2|1}(\cdot)$ is the estimated conditional cdf of Y_2 given $Y_1 = 0$. Similarly for Y_3, $q_{3i_3} = \Phi^{-1}\left\{\widehat{F}_{3|1}(y_{3i_3}|y_{1i_3}=1)\right\}$, where $i_3 = 1,\ldots,n_3$, n_3 is the size of the sub-sample for Y_3 and $\widehat{F}_{3|1}(\cdot)$ is the estimated conditional cdf of Y_3 given $Y_1 = 1$. The conditional cdfs are given by the ratios of the respective joint cdfs and marginal probabilities, and account for the fact that Y_2 and Y_3 are observed only for sub-samples. If the cdfs are close to their respective true distributions then the quantile residuals follow the standard normal distribution, which can be easily assessed by, e.g., inspecting the corresponding QQ-plots. For discrete margins, randomised normalised quantile residuals are used instead. For Y_2, these are based on $q_{2i_2} = \Phi^{-1}(u_{2i_2})$, where u_{2i_2} is a random value from the uniform distribution on $\left[\widehat{F}_{2|1}(y_{2i_2}-1|y_{1i_2}=0), \widehat{F}_{2|1}(y_{2i_2}|y_{1i_2}=0)\right]$. Similarly for Y_3, $q_{3i_3} = \Phi^{-1}(u_{3i_3})$, where u_{3i_3} is a random value from the uniform distribution on $\left[\widehat{F}_{3|1}(y_{3i_3}-1|y_{1i_3}=1), \widehat{F}_{3|1}(y_{3i_3}|y_{1i_3}=1)\right]$. Randomisation allows one to view the discrete

distributions as if there were continuous. With regard to S, because of its binary nature, residual analysis is not informative (e.g., Collett, 2002). In this case, a sensitivity analysis based on different link functions can be carried out; experience suggests that the model fit will not be significantly affected by this choice.

Studying the asymptotic properties of the proposed estimator is beyond the scope of the present paper. However, this could be approached by considering a fixed number of knots for the basis functions (e.g., Kauermann, 2005), in which case we would obtain, for instance, that $\hat{\boldsymbol{\delta}} \xrightarrow{P} \boldsymbol{\delta}^0 \quad and \quad \|\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}^0\| = O_P(1/\sqrt{n})$ and $\sqrt{n}(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}^0) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{i}^{-1}(\boldsymbol{\delta}^0))$, where $\mathbf{i}(\boldsymbol{\delta}^0) = \text{cov}\left[\partial \ell(\boldsymbol{\delta})/\partial \boldsymbol{\delta}|_{\boldsymbol{\delta}_0}\right]$.

Appendix B of the On-Line Supplementary Material discusses the results of a simulation study, whereas Appendix C illustrates the use of `GJRM`.

5 Average Treatment Effect

The binary switching indicator variable S_i can be referred to as treatment and takes value 1 or 0 which has implications on whether Y_2 or Y_3 can be observed. A well known measure of treatment evaluation is the average treatment effect (ATE), which quantifies the expected impact of a treatment for a randomly chosen individual from the population of interest. For a fitted model, the ATE can be estimated as follows

$$\text{ATE}(\hat{\boldsymbol{\delta}}) = \frac{1}{n} \sum_{i=1}^{n} \left\{ \widehat{\mathbb{E}(Y_{3i})} - \widehat{\mathbb{E}(Y_{2i})} \right\}, \tag{8}$$

where, for binary outcomes, the expectations can be expressed in terms of marginal probabilities (based on the probit, logit or complementary log-log link function), whereas for discrete and continuous outcomes the required formulae are reported in Tables 2 and 3. Note that in the continuous case, some expectations are given by linear or nonlinear combinations of more than one distributional parameter. Moreover, the ATE depends on $\boldsymbol{\delta}$ (strictly speaking on that subset of it that relates to the distributions of Y_2 and Y_3) through the $m(\mathbf{z}_i)$ as detailed in Section 2.3. Note that equation (8) can be easily modified to yield a percentage; this is

achieved by dividing by $\widehat{\mathbb{E}(Y_{2i})}$.

Since Y_{2i}^* and Y_{3i}^* are not available for the whole sample, imputation or prediction is required to compute (8). This is simply going to be based on the regressors and the estimated parameters of the related equations. However, some caution is needed when selecting the set of covariates to consider in the empirical analysis, especially when there are factor or categorical variables. Here, one has to check that the levels of the variables that appear in the whole sample also appear in the selected sub-samples (those related to Y_2 and Y_3). Consider, for instance, the situation in which the data contain a factor variable with five levels and that only three of them appear in the selected sample for Y_2, and four of them in the sample associated with Y_3. In such a case, a model can still be fitted but it will not be possible to carry out the prediction exercise required to compute (8).

Intervals for the ATE are obtained by employing the following procedure:

1. Draw V random vectors $\tilde{\boldsymbol{\delta}}_v, v = 1 \ldots, V$, using result (7).

2. Obtain V realisations of the function of interest, that is $\text{ATE}(\tilde{\boldsymbol{\delta}}_v)$.

3. Calculate the $(\vartheta/2)$-th and $(1-\vartheta/2)$-th quantiles of the V realisations. The interval is then given by $CI_{1-\vartheta} = [\text{ATE}(\tilde{\boldsymbol{\delta}}_v)_{\vartheta/2}, \text{ATE}(\tilde{\boldsymbol{\delta}}_v)_{1-\vartheta/2}]$.

Parameter ϑ is typically set to 0.05. Furthermore, a value of V equal to 100 usually produces representative results although it can be increased (for a little extra computational effort) if more precision is required.

6 Case Studies

This section applies the proposed copula-based switching regression framework to three case studies which investigate the effect of a binary treatment on various outcomes. To maintain a consistent theme, all studies come from the field of health economics, a discipline for which binary treatments and peculiar distributional shapes for outcomes are pervasive. As explained in the introduction, the studies use data from the MEPS.

6.1 Case study 1: continuous outcome

The majority of privately-insured Americans receive coverage through their employers as part of their total compensation. In turn, the theory of "compensating differentials" holds that an employee who receives health insurance from his employer should, all else equal, receive a lower wage than an employee who does not receive coverage (Rosen, 1986). However, compensating differentials are very difficult to observe in practice because good jobs tend to pay well and offer insurance coverage (Currie & Madrian, 1999).

Olsen (2002) finds evidence of compensating differentials using instrumental variables regression. Focusing on a sample of full-time employed females from the Current Population Survey, and using information on their husbands' employment and insurance status as instruments, he calculates that insurance reduces wages by approximately 20 percent. Our study attempts to mimic Olsen's approach by applying the proposed framework to data from the MEPS.

The dataset focuses on females in the age range $25 - 64$ from the 2012 wave of the MEPS. (We focus on 2012 to avoid complications associated with the Affordable Care Act, which, starting in 2014, required that most employers in the U.S. offer health benefits.) Similar to Olsen, the dataset focuses on females employed in the private sector full-time (at least 35 hours per week) who are not self employed. Furthermore, all females in the dataset are married to husbands who, themselves, also are employed. Table 1 from Appendix C if the On-Line Supplementary Material presents sample means partitioned according to whether the female holds insurance through her employer. The top row shows the difficulty of finding evidence of compensating differentials, with insured females having higher wage earnings. (Those wage numbers are highly statistically different, according to a conventional two-sample t-test.) The bottom of the table shows two instruments similar to the ones used in Olsen's study. The first, whether the female's husband holds insurance through his employer, suggests that wives tend to decline coverage when they have options through their husbands' plans. The second suggests that wives of husbands who work at larger firms, and thus presumably have better insurance options, also tend to decline coverage.

For the copula-based switching regression approach, the main modelling decisions involve the distribution forms for the three marginals and the two copulae. To arrive at those decisions, we explore many permutations of marginals and copulae available in the `GJRM` package, and we settle on the forms that yield the best overall fit according to the AIC and BIC, defined in Section 4, and by inspecting residual plots (see Figure 1 for the plots of the chosen model). Using such data-driven approach, the treatment variable, an indicator for whether the wife holds insurance through her employer, follows a probit specification. For the wife's wage income, the choices are the inverse Gaussian distribution among wives who do not hold insurance through their employers, and the Fisk distribution for those who do. As for the copulae, we have the Gumbel to link insurance to wages among wives who do not hold insurance through their employers, and the survival Clayton for those who do.

The object of main interest is the average treatment effect of health benefits on wage earnings. Before analysing the results obtained, however, the next three paragraphs will comment on the estimated model parameters which are reported in Appendix C of the On-Line Supplementary Material. Regarding the switching mechanism equation (`EQUATION 1` of the `R` output), the coefficients corroborate a priori expectations. For example, for a married woman, being a union member and working at a larger firm increase the likelihood of accepting her employer coverage. Regarding the regime regressions, the interpretation might or might not be that intuitive depending on the chosen distribution. For instance, since the selected distribution for regime 0 (wives who do not hold insurance through their employers) is an inverse Gaussian, the coefficients in `EQUATION 2` can be used to obtain the usual percentage effects on the mean. For example, for wives who do not hold insurance through their employers, the wage of a Hispanic is, on average, 25 percent lower as compared to that of a white woman. This interpretation does not hold for regime 1 (`EQUATION 4`) since the chosen distribution is the Fisk and, as reported in Table 3, the mean is a function of both μ and σ. However, interpretable effects can be easily obtained as illustrated in Appendix C. As for `EQUATION 3` and `EQUATION 5`, the coefficients can be used to obtain effects on the σ parameters which, however, do not correspond to standard deviations; to calculate these, the formulae reported in Table 3 have

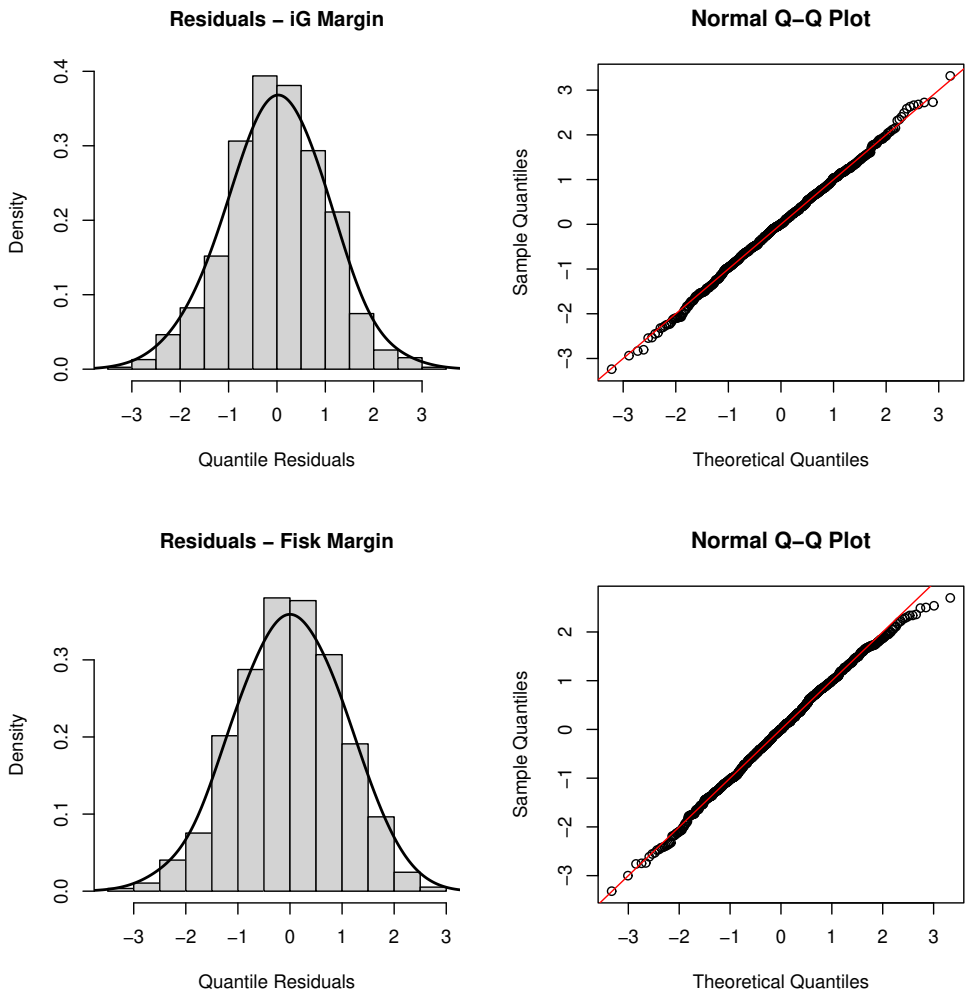


Figure 1: Histograms of normalised quantile residuals and normal Q-Q plots of residuals for the wage outcome variable under regime 0 (top) and regime 1 (bottom).

to be employed in a similar way as mentioned for the mean of the Fisk distribution.

Regarding the estimated smooth functions, reported in Figure 2, we note that age does not have an effect on the switching mechanism. For parameter μ of the inverse Gaussian, the effect of age shows an increasing, but slightly concave, shape among females who do not hold insurance through their employers. For μ of the Fisk, the effect of age follows a non-monotonic pattern, reaching its peak positive effect around age 45, among females who do hold insurance. For the σ parameters of the two distributions, age does not show any significant effect.

The dependence terms, shown in Appendix C, indicates overall that unobserved factors that increase a female's likelihood of holding insurance through her employer also correlate with

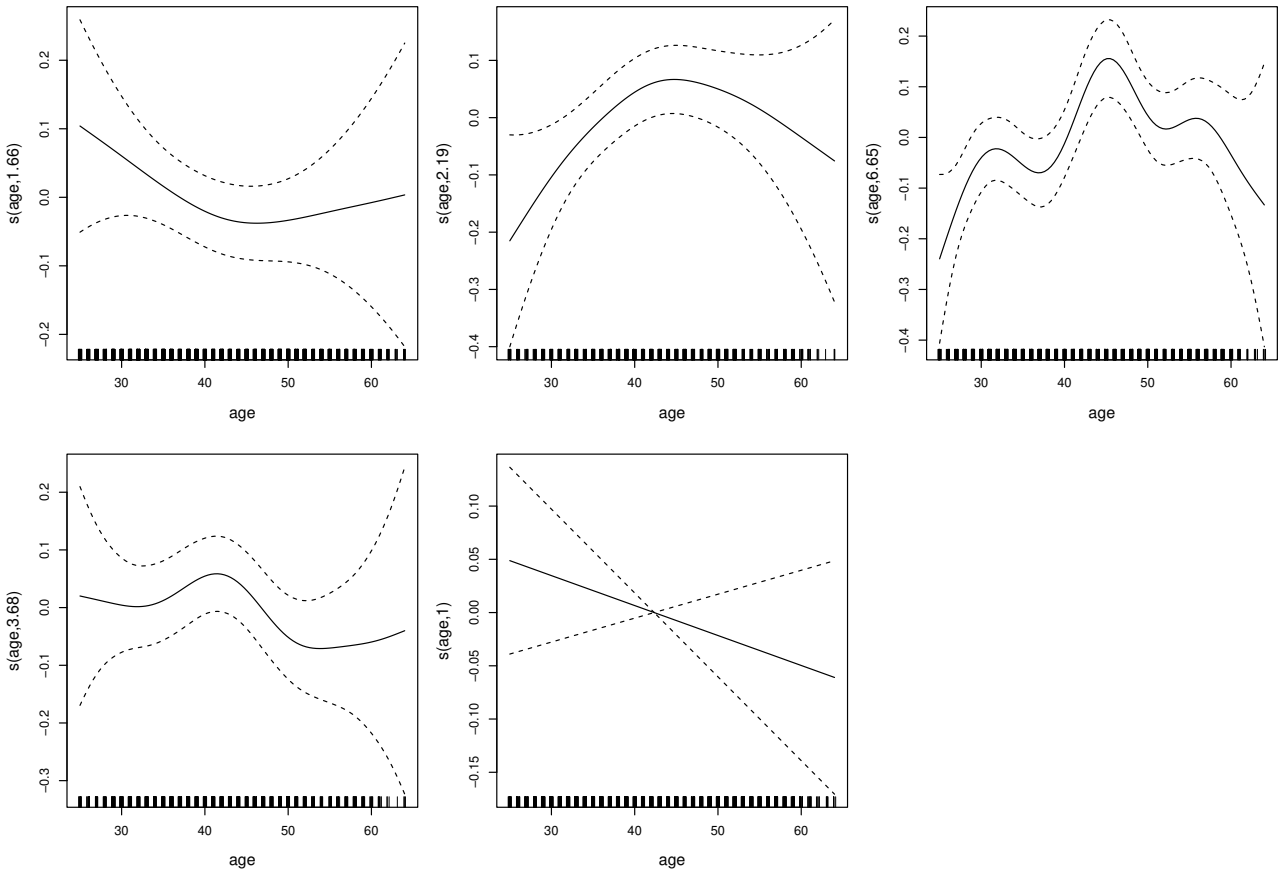


Figure 2: Estimated smooth effects of age for the probit equation, for the μ parameters of the inverse Gaussian and Fisk, and for the σ parameters of the same distributions. 95% point-wise intervals are also reported. The jittered rug plot, at the bottom of each graph, shows the covariate values. The numbers in the brackets of the y-axis captions are the *edf* of the smooth curves. Note that the estimated smooth functions are centered around zero because of the centering identifiability constraints. When $edf = 1$, the intervals correctly exhibit the behaviour displayed in the last plot because of such constraints.

higher wages. In particular, that relationship appears to be present and precisely estimated, among females who do not hold their own insurance, implying that certain types of females likely sort into jobs based on desires for health benefits. The same is not true for females who do hold their own insurance since the dependence parameter is virtually zero. Such subtle details of endogeneity bias are impossible to detect in more conventional regression setups.

Table 4 displays the results for the average treatment effect of interest. The first row shows the estimate and related interval from a univariate regression (including all controls, minus the two instruments) where the wage variable is assumed to follow the Fisk distribution (which appears to offer the best univariate fit). That number suggests that holding insurance

	Estimate	95% C.I.
Univariate Fisk	0.10	(0.04, 0.17)
Control function Fisk	-0.11	(-0.24, 0.04)
Proposal	-0.28	(-0.38, -0.15)

Table 4: Average treatment effects for case study 1.

correlates with a 10 percent higher wage. The next row shows the result from a control function approach which consists of two steps. The first involves a probit regression of the endogenous treatment (insurance status) on all controls plus the instruments, and serves to generate residuals. The second step involves a Fisk-based regression of the wage outcome on all variables (minus the two instruments) plus the first-stage residuals, which control for the endogeneity problem (e.g., Terza et al., 2008). The result from this approach suggests that insurance lowers wages by about 11 percent. Finally, the bottom row reports the effect obtained from the more flexible switching regression framework. The estimate suggests that insurance lowers wages by about 28 percent, more that 150 percent larger than that of the control function. The control function setup (analogous to traditional instrumental variables regression), by restricting the link between insurance and wages to operate via a simple intercept shift, seems to under-report the magnitude of compensating differentials. Probably, having insurance tends to reduce take-home pay by a larger amount than is typically recognised.

The flexible switching regression approach allows one to specify ancillary parameters, including copula dependence terms, as functions of covariates. Specifically, it seems plausible that the copula terms might depend upon a wife's level of educational attainment. The results reported in Appendix C show the coefficient estimates when the copula terms are given regression structures dependent on the three education dummies. Bear in mind that the coefficients are not statistically significant but for the sake of demonstrating the capabilities of the approach we provide an interpretation of the results. Focusing on the effect of college, having at least some college education appears to strengthen the positive association between having insurance and wage earnings among wives who do not have insurance, but college education does not appear to exert that same influence among wives who have insurance.

6.2 Case study 2: binary outcome

For people in the U.S. younger than age 65, having private coverage requires individuals to either find employment at a place that offers insurance or navigate the maze of insurance exchanges established by the Affordable Care Act. That level of effort implies that private coverage likely is endogenously linked to key outcomes, most importantly the consumption of health care services. Consequently, one of the core topics of health economics centers on estimating the effect of private health insurance on health care consumption. This case study draws inspiration from Deb et al. (2006) who explored this topic using methods similar to those presented here, albeit with a different econometric focus.

This case study considers the effect of a binary treatment (having insurance) on a binary outcome (having visited a doctor during the previous calendar year). Nearly all contacts with the health care system in the U.S. begin with an office-based visit to a doctor, so having a doctor visit is a reasonable proxy for a broad category of medical service usage. Drawn from the 2012 and 2013 waves of the MEPS, Table 2 from Appendix C of the On-Line Supplementary Material shows sample means for people with age in the range $18-64$ who work in the private sector. No one in the sample reports being self employed, and no one reports ever having any form of public insurance during the survey period. Shown near to the top of the table, 64 percent of insured subjects report having a doctor visit, compared to only 30 percent of uninsured subjects. Those numbers differ statistically according to a standard z-test. Inspired by Deb et al. (2006), the bottom of the table shows the instrument, firm size, which appears to strongly associate with the likelihood of having insurance.

With both treatment and outcome being binary, all marginal distributions follow logit specifications, which appear to offer the best fit. For the two copulae, the information criteria point to the 270 degree rotated Clayton to link insurance to doctor visits among people without insurance, and the Gaussian copula among people with insurance. Results for the copula-based switching regression approach are reported in Appendix C of the On-Line Supplementary Material. Most coefficient estimates corroborate a priori expectations. In particular, the instrument appears to influence the probability of having insurance, with subjects employed at larger firms

more likely to have coverage. Dependence appears to be negative for both treatment states. The overall interpretation is that unobserved factors that increase the likelihood of insurance also tend to reduce the probability of having a doctor visit. Health economists label that pattern "favorable selection" which means that insured subjects tend to be healthier and/or more risk averse.

	Estimate	95% C.I.
Univariate logit	0.25	(0.23, 0.27)
Control function logit	0.37	(0.26, 0.47)
Proposal	0.37	(0.27, 0.45)

Table 5: Average treatment effects for case study 2.

Table 5 presents average treatment effects of insurance on having a doctor visit. The first row shows an estimate from a univariate logit regression (including all controls, minus the instrument). That number suggests that having insurance increases the probability of having a doctor visit by 25 percentage points, a sizable effect relative to the sample mean of 55 percent of subjects having a doctor visit. The next row shows the result from the control function approach which suggests that having insurance increases the probability of a doctor visit by 37 percentage points. Finally, the bottom row reports the effect obtained from the more flexible switching regression framework. The effect is 0.37, which is the same as that obtained from the control function approach. Thus, in this case, the proposed approach seems to confirm the result of the more restrictive control function approach.

6.3 Case study 3: discrete outcome

A large swath of research explores reasons for school absenteeism, but the most important determinant appears to be health, with medical problems strongly linked to higher absenteeism (Basch, 2011; Holbert et al., 2002). Thus, consider the following narrowly-targeted question: Does visiting a doctor in order to obtain curative health care services cause children to miss school? Drawn from the 2015 wave of the MEPS, Table 3 from Appendix C of the On-Line Supplementary Material shows sample means for children between ages $6 - 13$, a range for

which schooling is compulsory in most U.S. jurisdictions.

The top row of the table indicates that students who visit a doctor to obtain curative medical services during the calendar year have, on average, 3.33 missed school days, compared to 1.50 missed days for students who do not have curative visits. Those numbers differ statistically according to a standard t-test. The bottom row of the table shows the variable that serves as an instrument: an indicator for whether the family of the child can arrive at its usual source of medical care in less than 30 minutes. (That variable likely reflects, in part, travel costs that families must incur in order to acquire medical services.) The numbers suggest that being located closer to one's usual source of care associates with larger probabilities of having a curative visit.

As for case study 1, exploring many permutations of marginals and copulae, the best overall fit comes when the distribution of the treatment variable (whether the child had a curative visit) follows a logit specification, and both marginals for the outcome variable (the number of missed school days) follow a Negative Binomial type II (see the residual plots in Appendix C of the On-Line Supplementary Material). For the two copulae, the link between doctor visit and missed days follows a 90 degree rotated Joe among children with no doctor visits, and a non-rotated Joe among children with doctor visits.

The estimated results are reported in Appendix C. Most of the coefficients corroborate a priori expectations. The age effects are linear in the curative visit equation. But for missed school days, as far as the μ parameter is concerned (which is also the mean of the distribution), among students who do not have curative visits, age shows an overall decreasing non linear effect which then starts increasing in the final years. Among students who do have curative visits, the age effect shows a decreasing linear pattern. As for σ, age shows an increasing linear effect among students who do not have curative visits and a quadratic effect among students who do have curative visits.

The dependence terms point to very different patterns by treatment status. The association is negative for children without curative visits, suggesting that, on average, unobserved traits that induce children to have office visits tend to reduce missed school days among children who

do not have office visits. On the other hand, the other dependence term is positive, indicating the opposite pattern among children who do have office visits. Taken together, those disparate associations suggest a pattern whereby parents who have trouble accessing medical care also tend to have children who miss school, a finding that could not have been detected with a more conventional approach.

	Estimate	95% C.I.
Univariate NBII	1.62	(1.40, 1.86)
Control function NBII	1.46	(-0.47, 4.45)
Proposal	0.94	(0.58, 1.40)

Table 6: Average treatment effects for case study 3.

Table 6 shows the effects of curative visits on missed school days. The first row presents an estimate from a univariate NBII regression (including all controls, minus the instrument). That number suggests that having a curative visits leads to approximately 1.62 more missed school days, a sizable effect relative to the sample mean of 2.10 missed days. The next row shows the result from the control function approach which shows that having a curative visits leads to 1.46 more missed days; this is similar in magnitude to the univariate NBII regression, but with a much wider confidence interval. Finally, the bottom row reports the effect obtained using the proposed switching regression framework. The resulting estimate suggests that having a curative visit leads to 0.94 more missed days, which is approximately less than one-third the magnitude of the control function estimate. Evidently, the control function method fails to detect the nuanced pattern of endogeneity, likely because, as indicated by the dependencies commented in the previous paragraph, endogeneity tugs in opposite directions and different magnitudes depending on whether children have doctor visits. In the control function approach, those opposite directions appear to cancel out as well as lessen precision, leaving an effect similar to that of the univariate NBII regression. The flexible switching regression setup, where endogeneity is considered separately by treatment state, finds a smaller treatment effect. Giving the dependence terms regression structures did not appear to produce interesting insights.

7 Conclusions

Motivated by three case studies in the field of health economics, we have introduced a unifying approach to switching regime regression. Various details including the model set up, parameter estimation and inference have been discussed. All developments have been integrated within the `R` package `GJRM` whose modularity allows for easy inclusion of virtually any parametric copula and marginal distribution.

The proposed approach makes a significant contribution in switching regression modelling since it can handle various empirical situations and is practically usable. Although the literature in this area is ample, to the best of our knowledge, until now there has existed no work that provided a methodological framework together with software implementation for the type of switching regime regression problem considered in this paper. Recall that the proposal can handle many types of outcomes via a vast range of marginal distributions, allows for a wide variety of copula dependence structures, and permits to specify all model parameters as flexible functions of covariate effects. The findings from our three case studies have provided new evidence on the problems tackled.

Future research will focus on extending the scope of the modelling framework by allowing for survival margins as well as by exploring alternative copula selection methods along the lines of Cai (2014), for instance.

Acknowledgments

The first two authors were supported by the Engineering and Physical Sciences Research Council [EP/T033061/1].

References

Akaike, H. (1987). Information theory and an extension of the maximum likelihood principle. *In: Petrov, B.N., Csaki, B.F. (eds.) Second International Symposium on Information*

Theory, Academiai Kiado, Budapest.

Bayer, P., Khan, S. & Timmins, C. (2011). Nonparametric Identification and Estimation in a Roy Model With Common Nonpecuniary Returns. *Journal of Business & Economic Statistics*, 29(2), 201–215.

Basch, C. (2011). Healthier students are better learners: a missing link in school reforms to close the achievement gap. Equity matters. *Journal of School Health*, 81(10), 593–598.

Bollen, K. A. & Pearl, J. (2013). Eight Myths About Causality and Structural Equation Models. *In: Morgan S. (eds) Handbook of Causal Analysis for Social Research. Handbooks of Sociology and Social Research. Springer, Dordrecht*, 301–328.

Borjas, G. (1987). Self-selection and the earnings of immigrants. *American Economic Review*, 77(4), 531–553.

Brechmann, E. C. & Schepsmeier, U. (2013). Modeling Dependence with C- and D-Vine copulae: The R Package CDVine. *Journal of Statistical Software*, 52(3), 1–27.

Cai, Z. & Wang, X. (2014). Selection of Mixed Copula Model via Penalized Likelihood. *Journal of the American Statistical Association*, 109(506), 788–801.

Chen, H., Fan, Y. & Wu, J. (2014). A flexible parametric approach for estimating switching regime models and treatment effect parameters. *Journal of Econometrics*, 181, 77–91.

Choi, P. & Min, I. (2009). Estimating endogenous switching regression model with a flexible parametric distribution function: application to Korean housing demand. *Applied Economics*, 41(23), 3045–3055.

Claeskens, G., Krivobokova, T. & Opsomer, J. (2009). Asymptotic Properties of Penalized Spline Estimators. *Biometrika*, 96(3), 529–544.

Collett, D. (2002). *Modelling Binary Data: Second Edition.* London. Chapman & Hall/CRC Texts in Statistical Science.

Cornelissen, T., Dustmann, C., Raute, A. & Schönberg, U. (2016). From LATE to MTE: Alternative methods for the evaluation of policy interventions. *Labour Economics*, 41, 47–60.

Cox, D. R. (1958). *Planning of Experiments*. New York: Wiley.

Currie, J. & Madrian, B. (1999). Health, Health Insurance and the Labor Market. *In Handbook of Labor Economics edited by O. Ashenfelter and D. Card*, Amsterdam: Elsevier, 3C, 3309–3406.

Deb, P., Munkin, M. & Trivedi, P. (2006). Private Insurance, Selection, and Health Care Use: A Bayesian Analysis of a Roy-Type Model. *Journal of Business and Economic Statistics*, 24(4), 403–415.

D'Haultfoeuille, X. & Maurel, A. (2013). Inference on an extended Roy model, with an application to schooling decisions in France. *Journal of Econometrics*, 174(2), 95–106.

Dunn, P. K. & Smyth, G. K. (1996). Randomized Quantile Residuals. *Journal of Computational and Graphical Statistics*, 5(3), 236–245.

Eilers, P. H. C. & Marx, B. D. (1996). Flexible Smoothing with B-splines and Penalties. *Statistical Science*, 11(2), 89–121.

Eisenhauer, P., Heckman, J. J. & Vytlacil, E. (2015). The Generalized Roy Model and the Cost-Benefit Analysis of Social Programs. *Journal of Political Economy*, 123(2), 413–443.

Fitawek, W. & Hendriks, S.L. (2021). Evaluating the Impact of Large-Scale Agricultural Investments on Household Food Security Using an Endogenous Switching Regression Model. *Land*, 10(3), 323.

French, E. & Taber, C. (2011). Identification of models of the labor market. *In Handbook of Labor Economics, Elsevier*, 4, 537–617.

Hasebe, T. (2013). Copula-based maximum-likelihood estimation of sample-selection models. *The Stata Journal*, 13(3), 547–573.

Hasebe, T. (2020). Endogenous switching regression model and treatment effects of count-data outcome. *The Stata Journal*, 20(3), 627–646.

Heckman, J. (1990). Varieties of selection bias. *American Economic Review Papers and Proceedings*, 80(2), 313–318.

Heckman, J. & Honoré, B. (1990). The empirical content of the Roy model. *Econometrica*, 58(5), 1121–1149.

Heckman, J. & Hotz, V. (1989). Choosing among alternative nonexperimental methods for estimating the impact of social programs: the case of manpower training. *Journal of the American Statistical Association*, 84(408), 862–874.

Holbert, T., Wu, L. & Stark, M. (2002). School attendance initiative: the first 3 years: 1998/99-2000/01. Report prepared for US Dept of Justice, National Institute of Law Enforcement and Criminal Justice, Office of Development and Testing.

Joe, H. (2014). *Dependence Modeling with copulae*. CRC Press: Boca Raton, FL, USA.

Kauermann, G. (2005). Penalized Spline Smoothing in Multivariable Survival Models with Varying Coefficients. *Computational Statistics & Data Analysis*, 49(1), 169–186

Kauermann, G., Krivobokova, T. & Fahrmeir, L. (2009). Some asymptotics results on generalized penalized spline smoothing. *Journal of Royal Statistical Society Series B*, 71(2), 487–503.

Kim, S.-H. (2021). Changes in Social Trust: Evidence from East German Migrants. *Social Indicators Research*, 155, 959–981.

Lokshin, M. & Sajaia, Z. (2004). Maximum likelihood estimation of endogenous switching regression models. *The Stata Journal*, 4(3), 282–289.

Marra, G. & Radice, R. (2020). Copula Link-Based Additive Models for Right-Censored Event Time Data. *Journal of the American Statistical Association*, 115(530), 886–895.

Marra, G. & Radice, R. (2022). *GJRM: Generalized Joint Regression Modeling.* R package version 0.2-6, URL `https://cran.r-project.org/package=GJRM`

Moscelli, G., Siciliani., L., Gutacker, N. & Cookson, R. (2018). Socioeconomic inequality of access to healthcare: Does choice explain the gradient? *Journal of Health Economics*, 57, 290–314.

Murtazashvili, I. & Wooldridge, J. M. (2016). A control function approach to estimating switching regression models with endogenous explanatory variables and endogenous switching. *Journal of Econometrics*, 190(2), 252–266.

Nelsen, R. (2006). *An Introduction to copulae: Second Edition.* New York: Springer.

Neyman, J. (1923). On the application of probability theory to agricultural experiments. *Essay on principles, Statistical Science*, 5(9), 465–472.

Nikoloulopoulos, A. K. & Karlis, D. (2010). Regression in a Copula Model for Bivariate Count Data. *Journal of Applied Statistics*, 37(9), 1555–1568.

Olsen, C. (2002). Do Workers Accept Lower Wages in Exchange for Health Benefits? *Journal of Labor Economics*, 20(S2), S91-S114.

Pearl, J. (2015). Trygve Haavelmo and the Emergence of Causal Calculus. *Econometric Theory*, 31(1), 152–179.

R Core Team (2022). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, URL `https://www.R-project.org/`.

Rigby, R. A. & Stasinopoulos, D. M. (2005). Generalized Additive Models for Location, Scale and Shape (with Discussion). *Journal of the Royal Statistical Society, Series C*, 54(3), 507–554.

Rosen, S. (1986). The Theory of Equalizing Differences. *In Handbook of Labor Economics, edited by O. Ashenfelter and R. Layard*, Amsterdam: North-Holland, 1, 641–692.

Roy, A. (1951). Some thoughts on the distribution of earnings. *Oxford Economic Papers*, 3(2), 70, 135–146.

Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6(2), 461–464.

Smith, M. (2005). Using copulae to model switching regimes with an application to child labour. *Economic Record*, 81(255), S47–S57.

Terza, J., Basu, A. & Rathouz, P. (2008). Two-Stage Residual Inclusion Estimation: Addressing Endogeneity in Health Econometric Modeling. *Journal of Health Economics*, 27(3), 531–543.

Toomet, O. & Henningsen, A. (2020). *Sample Selection Models in R: Package sampleSelection* R package version 1.2-12, URL `https://cran.r-project.org/package=sampleSelection`.

Trivedi, P. & Zimmer, D. (2007). Copula Modeling: An Introduction for Practitioners. *Foundations and Trends in Econometrics*, 1(1), 1–111.

Wood, S. N. (2017). *Generalized Additive Models: An Introduction With R: Second Edition*. Chapman & Hall/CRC, London.

On-Line Supplementary Material: A Unifying Switching Regime Regression Framework with Applications in Health Economics

Giampiero Marra, University College London, UK

Rosalba Radice, Bayes Business School, City, University of London, UK

David Zimmer, Western Kentucky University, Bowling Green, USA

2022-09-06

Appendix A: Parameter Estimation

The algorithm used for parameter estimation was implemented by adapting to this context the two-step approach of Marra & Radice (2020). The two steps are summarised below.

Estimation of $\boldsymbol{\delta}$

At iteration a, for a given parameter vector $\boldsymbol{\delta}^{[a]}$ and holding the smoothing parameters fixed at a vector of values (denoted as $\boldsymbol{\lambda}^{[a]}$ or $\hat{\boldsymbol{\lambda}}$), an update for $\boldsymbol{\delta}$ is found by solving the following problem

$$\boldsymbol{\delta}^{[a+1]} = \boldsymbol{\delta}^{[a]} + \underbrace{\underset{\boldsymbol{e}:\|\boldsymbol{e}\|\leq\Delta^{[a]}}{\operatorname{argmin}}\ \breve{\ell}_p(\boldsymbol{\delta}^{[a]})}_{:=\boldsymbol{e}^{[a+1]}}, \tag{1}$$

$$\breve{\ell}_p(\boldsymbol{\delta}^{[a]}) := -\{\ell_p(\boldsymbol{\delta}^{[a]}) + \boldsymbol{e}^\top \boldsymbol{g}_p^{[a]} + \frac{1}{2}\boldsymbol{e}^\top \boldsymbol{H}_p^{[a]} \boldsymbol{e}\},$$

where $\|\cdot\|$ is the Euclidean norm and $\Delta^{[a]}$ is the radius of the trust region which is adjusted through the iterations. The penalised score vector and penalised Hessian matrix are given by $\boldsymbol{g}_p^{[a]} = \boldsymbol{g}^{[a]} - \mathbf{S}_{\hat{\boldsymbol{\lambda}}}\boldsymbol{\delta}^{[a]}$ and $\boldsymbol{H}_p^{[a]} = \boldsymbol{H}^{[a]} - \mathbf{S}_{\hat{\boldsymbol{\lambda}}}$, where $\boldsymbol{g}^{[a]}$ is defined as

$$\left(\frac{\partial \ell(\boldsymbol{\delta})}{\partial \boldsymbol{\beta}_{\mu_1}}\bigg|^{\mathsf{T}}_{\boldsymbol{\beta}_{\mu_1}=\boldsymbol{\beta}_{\mu_1}^{[a]}}, \ldots, \frac{\partial \ell(\boldsymbol{\delta})}{\partial \boldsymbol{\beta}_{\theta_{12}}}\bigg|^{\mathsf{T}}_{\boldsymbol{\beta}_{\theta_{12}}=\boldsymbol{\beta}_{\theta_{12}}^{[a]}}, \frac{\partial \ell(\boldsymbol{\delta})}{\partial \boldsymbol{\beta}_{\theta_{13}}}\bigg|^{\mathsf{T}}_{\boldsymbol{\beta}_{\theta_{13}}=\boldsymbol{\beta}_{\theta_{13}}^{[a]}} \right)^{\mathsf{T}}$$

and $\boldsymbol{H}^{[a]}$ as

$$\frac{\partial^2 \ell(\boldsymbol{\delta})}{\partial \boldsymbol{\beta}_l \partial \boldsymbol{\beta}_m^{\mathsf{T}}}\bigg|_{\boldsymbol{\beta}_l=\boldsymbol{\beta}_l^{[a]}, \boldsymbol{\beta}_m=\boldsymbol{\beta}_m^{[a]}}, \; l, m = \mu_1, \ldots, \theta_{12}, \theta_{13}.$$

At each iteration, the minimisation of (1) uses a quadratic approximation of $\ell_p(\boldsymbol{\delta}^{[a]})$ to choose the best $\boldsymbol{e}^{[a+1]}$ within the ball centered in $\boldsymbol{\delta}^{[a]}$ of radius $\Delta^{[a]}$.

Trust region and classical line search methods employ a quadratic model of the objective function to generate steps from one iterate to the next. Line search methods find a search direction and then suitable step lengths along this direction. Instead, trust region approaches search the step that minimises the objective function within a previously defined region around the current iterate. If a function exhibits, e.g., long plateaus and the current iterate $\boldsymbol{\delta}^{[a]}$ is in that region, line search methods may search the next step $\boldsymbol{\delta}^{[a+1]}$ far away from the current iterate, hence reducing the efficiency of the algorithm. Also, the evaluation of the objective function may be, e.g., non-definite, therefore causing algorithmic failure. On the contrary, before evaluating the objective function, trust region methods define a maximum distance based on the trust region. This is advantageous because the new iterate will not lie too far away from the current one and also because in case of non-definite evaluation of $\breve{\ell}_p$ step $\boldsymbol{e}^{[a+1]}$ will not be accepted. When a candidate does not improve the function sufficiently or gives a non-definite evaluation of $\breve{\ell}_p$, the trust region will shrink and the algorithm will move back to the previous step. If the improvement is large enough then the trust region will expand in the next iteration. Since the proposed implementation is based on the analytical score and Hessian of $\ell(\boldsymbol{\delta})$, the algorithm will converge super-linearly to a point satisfying the second-order sufficient conditions. Moreover, near the solution, the approach will become asymptotically similar to the Newton-Raphson method, hence benefitting from the fast convergence rate of this technique.

See Nocedal & Wright (2006, Chapter 4) for more details.

Estimation of $\boldsymbol{\lambda}$

To discuss the criterion adopted for multiple smoothing parameter estimation, we first need to express the estimator for $\boldsymbol{\delta}$ in terms of score and Hessian. A first order Taylor expansion of $\boldsymbol{g}_p^{[a+1]}$ about $\boldsymbol{\delta}^{[a]}$ yields $\mathbf{0} = \boldsymbol{g}_p^{[a+1]} \approx \boldsymbol{g}_p^{[a]} + \left(\boldsymbol{\delta}^{[a+1]} - \boldsymbol{\delta}^{[a]}\right) \boldsymbol{H}_p^{[a]}$, which, after some manipulation, leads to $\boldsymbol{\delta}^{[a+1]} = \left(-\boldsymbol{H}_p^{[a]}\right)^{-1} \sqrt{-\boldsymbol{H}^{[a]}}\mathbf{M}^{[a]}$, where $\mathbf{M}^{[a]} = \boldsymbol{\mu}_{\mathbf{M}}^{[a]} + \boldsymbol{\epsilon}^{[a]}$, $\boldsymbol{\mu}_{\mathbf{M}}^{[a]} = \sqrt{-\boldsymbol{H}^{[a]}}\boldsymbol{\delta}^{[a]}$ and $\boldsymbol{\epsilon}^{[a]} = \sqrt{-\boldsymbol{H}^{[a]}}^{-1}\boldsymbol{g}^{[a]}$. From likelihood theory, $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_\psi)$, $\mathbf{M} \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{M}}, \mathbf{I}_\psi)$, where $\mathbf{I}_\psi$ is an identity matrix of dimension ψ (the length of $\boldsymbol{\delta}$), $\boldsymbol{\mu}_{\mathbf{M}} = \sqrt{-\boldsymbol{H}}\boldsymbol{\delta}^0$ and $\boldsymbol{\delta}^0$ is the true parameter vector. The predicted value vector for $\mathbf{M}$ is $\hat{\boldsymbol{\mu}}_{\mathbf{M}} = \sqrt{-\boldsymbol{H}}\hat{\boldsymbol{\delta}} = \mathbf{A}\mathbf{M}$, where $\mathbf{A} = \sqrt{-\boldsymbol{H}}\left(-\boldsymbol{H}_p\right)^{-1}\sqrt{-\boldsymbol{H}}$. To calibrate the the trade-off between fit and parsimony in a data-driven manner, $\boldsymbol{\lambda}$ is determined by minimising $\mathbb{E}\left(\|\boldsymbol{\mu}_{\mathbf{M}} - \hat{\boldsymbol{\mu}}_{\mathbf{M}}\|^2\right)$ which, after some manipulation, is equal to

$$\mathbb{E}\left(\|\mathbf{M} - \mathbf{A}\mathbf{M}\|^2\right) - Wn + 2\mathrm{tr}(\mathbf{A}), \tag{2}$$

where $W = w_1 + w_2 + w_3 + 2$ and $\mathrm{tr}(\mathbf{A})$ represents the effective degrees of freedom (*edf*) of the penalised model. Note that (2) depends on $\boldsymbol{\lambda}$ through $\mathbf{A}$. In practice, an estimate of (2) in required, i.e.

$$\widehat{\|\boldsymbol{\mu}_{\mathbf{M}} - \hat{\boldsymbol{\mu}}_{\mathbf{M}}\|^2} = \|\mathbf{M} - \mathbf{A}\mathbf{M}\|^2 - Wn + 2\mathrm{tr}(\mathbf{A}). \tag{3}$$

So, for a given $\boldsymbol{\lambda}^{[a]}$ and holding $\boldsymbol{\delta}^{[a+1]}$ fixed, the following problem is solved

$$\boldsymbol{\lambda}^{[a+1]} = \arg\min_{\boldsymbol{\lambda}} \|\mathbf{M}^{[a+1]} - \mathbf{A}^{[a+1]}\mathbf{M}^{[a+1]}\|^2 - Wn + 2\mathrm{tr}(\mathbf{A}^{[a+1]}), \tag{4}$$

using the stable and efficient computational routine of Wood (2017, Section 6.5.1). Basing smoothing parameter estimation on a parametrisation of $\mathbf{M}$ that employs $\boldsymbol{H}$ and $\boldsymbol{g}$ as a whole instead of the n components that make them up is advantageous in terms of stability, efficiency and generality of the approach (see Marra et al. (2017) for a through discussion on this). The

additional benefit is that the score and Hessian, which are needed to set up the quantities in (4), are obtained as a byproduct of the estimation step for $\boldsymbol{\delta}$, hence reducing the computational effort made for the smoothing step.

To within an additive constant, the first term on the RHS of (3) is a quadratic approximation to $-2\ell(\hat{\boldsymbol{\delta}})$, hence (3) is approximately equivalent to the Akaike information criterion (AIC, Akaike, 1973) with degrees of freedom given by $\text{tr}(\mathbf{A})$. The *edf* of a model that only has unpenalised terms is ψ, since in this case $\text{tr}(\mathbf{A}) = \text{tr}(\mathbf{I}_\psi)$. The *edf* of a penalised model is $\text{tr}(\mathbf{A})$ which can be written as $\psi - \text{tr}\left\{(-\boldsymbol{H} + \mathbf{S}_{\boldsymbol{\lambda}})^{-1}\mathbf{S}_{\boldsymbol{\lambda}}\right\}$; if $\boldsymbol{\lambda} \to \mathbf{0}$ then $\text{tr}(\mathbf{A}) \to \psi$ and if $\boldsymbol{\lambda} \to \boldsymbol{\infty}$ then $\text{tr}(\mathbf{A}) \to \psi - \zeta$, where ζ is the total number of parameters subject to penalisation. When $\mathbf{0} < \boldsymbol{\lambda} < \boldsymbol{\infty}$, the *edf* is a value in the range $[\psi - \zeta, \psi]$. The *edf* associated to each penalised term in the model is given by the sum of the related trace elements.

The steps for estimating $\boldsymbol{\delta}$ and $\boldsymbol{\lambda}$ are iterated until $\left|\ell(\boldsymbol{\delta}^{[a+1]}) - \ell(\boldsymbol{\delta}^{[a]})\right| / \left(0.1 + \left|\ell(\boldsymbol{\delta}^{[a+1]})\right|\right) < 1e-07$ is satisfied. Starting values for the parameters associated to the marginal distributions are obtained by fitting three suitable univariate regression models (related to the three responses of the joint model), whereas initial values for the copula parameters are obtained by using transformations of the empirical correlations between the residuals of the above regressions.

Appendix B: Simulation Study

This section shows the results from three sets of Monte Carlo studies, each mimicking the outcomes of the three case studies (continuous, discrete and binary). For each study, we considered sample sizes of 2000 and 4000, while the number of replicates was set to 1000.

Continuous case

The simulations rely on a data generating process (DGP) of the form

$$s \sim \text{Bernoulli with } \mu_1 = \frac{\{\exp(0.5 + s_1(z_1) + 0.5z_2 - z_3)\}}{1 + \{\exp(0.5 + s_1(z_1) + 0.5z_2 - z_3)\}},$$
$$y_2^* \sim \text{Weibull with } \mu_2 = \exp\{11 - 0.2z_1 + 0.3s_2(z_2)\} \text{ and } \sigma_2 = \exp\{1 + 0.2z_2\}$$

for regime 0, and of the form

$$s \sim \text{Bernoulli with } \mu_1 = \frac{\{\exp(0.5 + s_1(z_1) + 0.5z_2 - z_3)\}}{1 + \{\exp(0.5 + s_1(z_1) + 0.5z_2 - z_3)\}},$$
$$y_3^* \sim \text{Fisk with } \mu_2 = \exp\{10 + 0.3z_1 - 0.2z_2\} \text{ and } \sigma_2 = \exp\{1 + 0.6s_3(z_1)\}$$

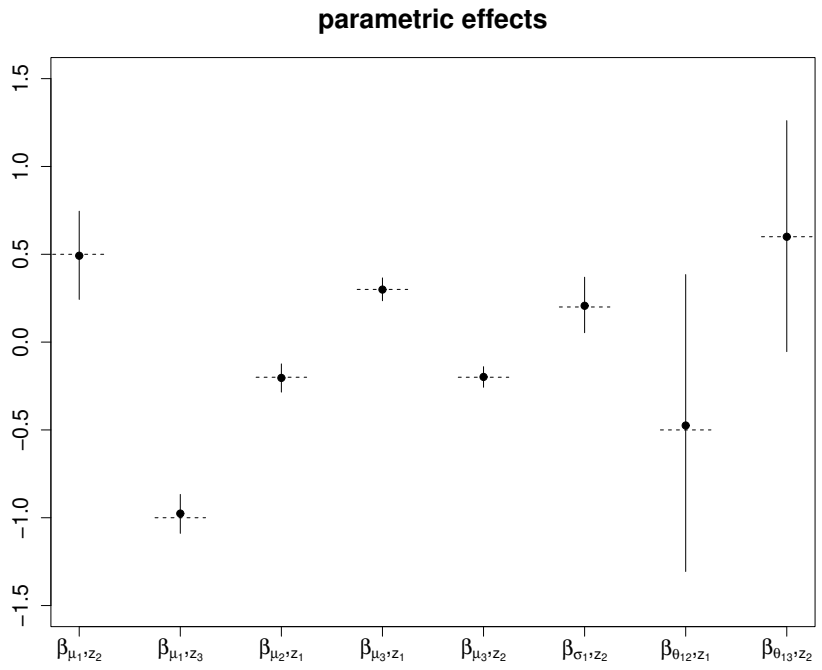
for regime 1. The smooths are defined as $s_1(z) = 0.6\sin(2\pi z)$, $s_2(z) = 0.6\{\exp(z) + \sin(2.9z)\}$ and $s_3(z) = z + \exp(-30(z - 0.5)^2)$. Variables z_1, z_2 and z_3 are generated using a multivariate standard Gaussian with correlation parameters set at 0.5, which are then transformed using the distribution function of a standard Gaussian. Associated values of s and y_2^*, and s and y_3^* are generated via function `archmCopula()`, from the `R` package `copula`, using the Joe and Clayton copulae, respectively. The dependence parameters θ_{12} and θ_{13} are specified as $\exp(2.5 - 0.5z_1) + 1$ and $\exp(1.5 + 0.6z_2)$, respectively. The data generated from the two regimes are then combined such that the data under regime 0 are consistent with $s = 0$ and data under regime 1 are consistent with $s = 1$.

Using `gjrm()`, for each simulated dataset, we fitted the endogenous switching regression model with logit and Weibull marginals and Joe copula for regime 0, and logit and Fisk

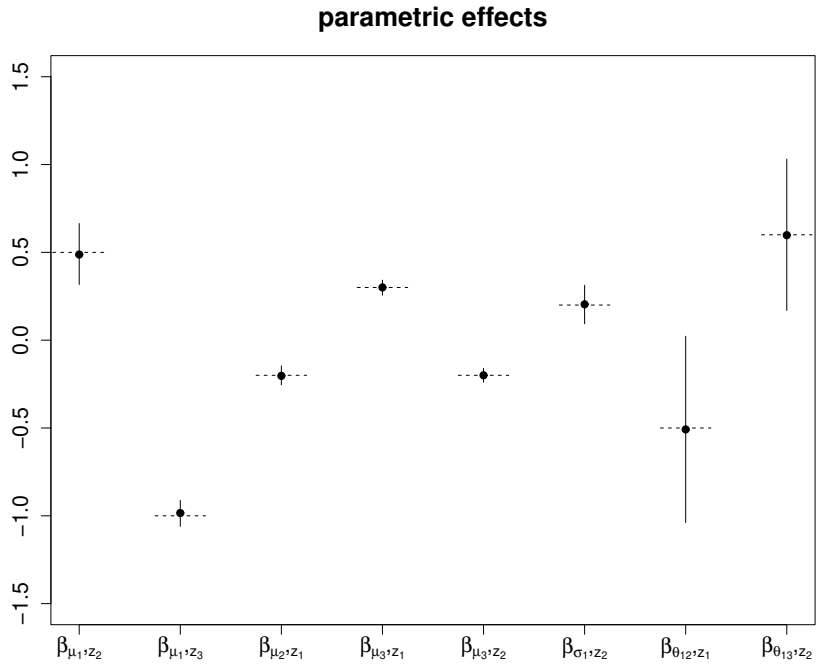
marginals and Clayton copula for regime 1. The smooth components in the models were represented using penalised low rank thin plate splines with second order penalty and 10 bases. For each replicate, curve estimates were constructed using 200 equally spaced fixed values in the $(0, 1)$ range.

Figure 1 summarises the findings of the simulation study when looking at the parametric estimates. The results show that the bias and variability of the estimates, across the two sample sizes, are low for all parameters of the two margins, and that they decrease as the sample size increases. As for the effects associated with the copula parameters, the results are more variable although the precision increases with the sample size. This finding was expected as the copula parameters are notoriously more difficult to estimate since the related profile likelihoods tend to be less sharp around the optimum.

Figure 2 shows that the true smooth functions are recovered well by the estimation method. Moreover, the results, in terms of bias and variability, improve as the sample size increases. Alternative scenarios were tried out and the findings were similar to the ones reported here.

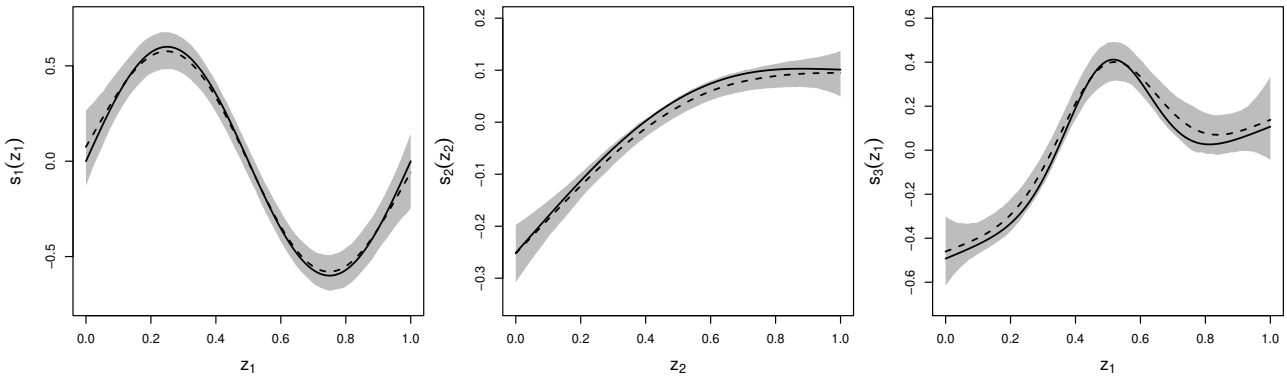


(a) $n = 2000$

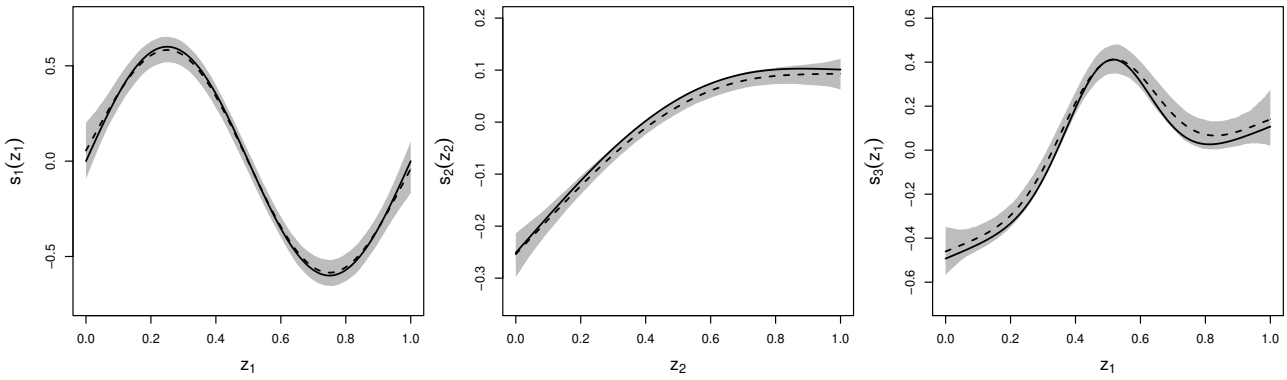


(b) $n = 4000$

Figure 1: Estimation results for parametric effects under the continuous outcome scenario.



(a) $n = 2000$



(b) $n = 4000$

Figure 2: Estimation results for smooth effects under the continuous outcome scenario.

Discrete case

The simulations are based on the following DGP:

$$s \sim \text{Bernoulli with } \mu_1 = \frac{\{\exp(0.5 + s_1(z_1) + 0.5z_2 - z_3)\}}{1 + \{\exp(0.5 + s_1(z_1) + 0.5z_2 - z_3)\}},$$

$$y_2^* \sim \text{Negative Binomial type II with } \mu_2 = \exp\{0.5 - 0.2z_1 + 0.6s_2(z_2)\} \text{ and } \sigma_2 = \exp\{0.2 + 0.4z_2\}$$

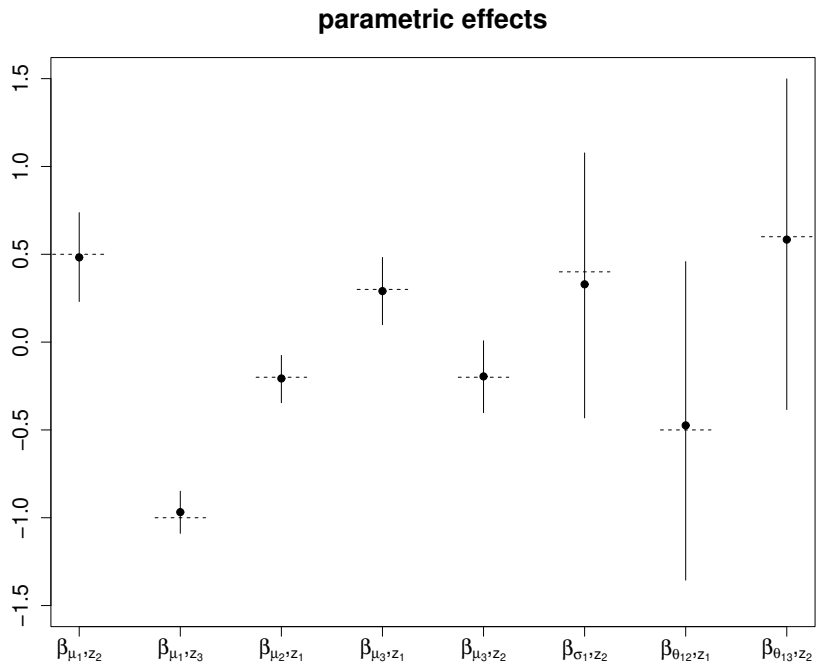
for regime 0, and

$$s \sim \text{Bernoulli with } \mu_1 = \frac{\{\exp(0.5 + s_1(z_1) + 0.5z_2 - z_3)\}}{1 + \{\exp(0.5 + s_1(z_1) + 0.5z_2 - z_3)\}},$$

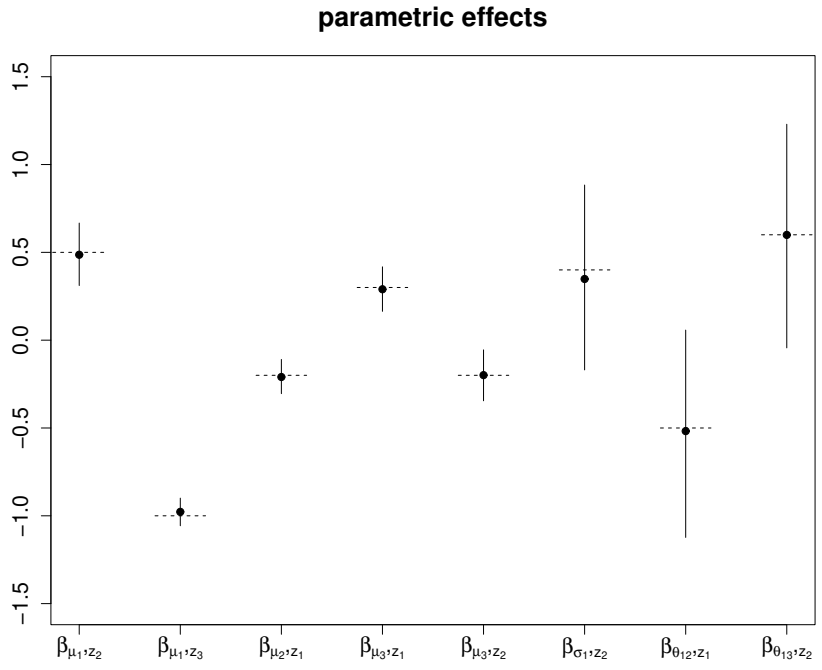
$$y_3^* \sim \text{Negative Binomial type II with } \mu_2 = \exp\{0.5 + 0.3z_1 - .2z_2\} \text{ and } \sigma_2 = \exp\{1.2s_3(z_1)\}$$

for regime 1. The smooth functions $s_1(z)$, $s_2(z)$ and $s_3(z)$ are the same as those defined for the continuous case and variables z_1, z_2 and z_3 have also been generated in the same way. Associated values of s and y_2^*, and s and y_3^* are generated using the Joe and Clayton copulae, where θ_{12} and θ_{13} are specified as $\exp(2.5 + 0.5z_1) + 1$ and $\exp(1.5 + 0.6z_2)$, respectively.

Using `gjrm()`, we fitted the endogenous switching regression model with logit and negative binomial type II marginals and Joe copula for regime 0, and logit and negative binomial type II marginals and Clayton copula for regime 1. As shown in Figures 3 and 4, the findings are similar to those previously described for the continuous case.

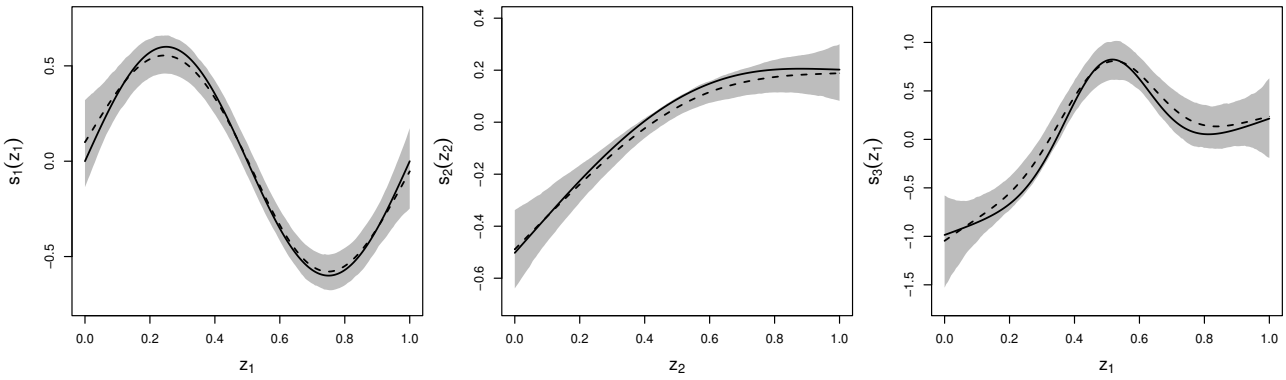


(a) $n = 2000$

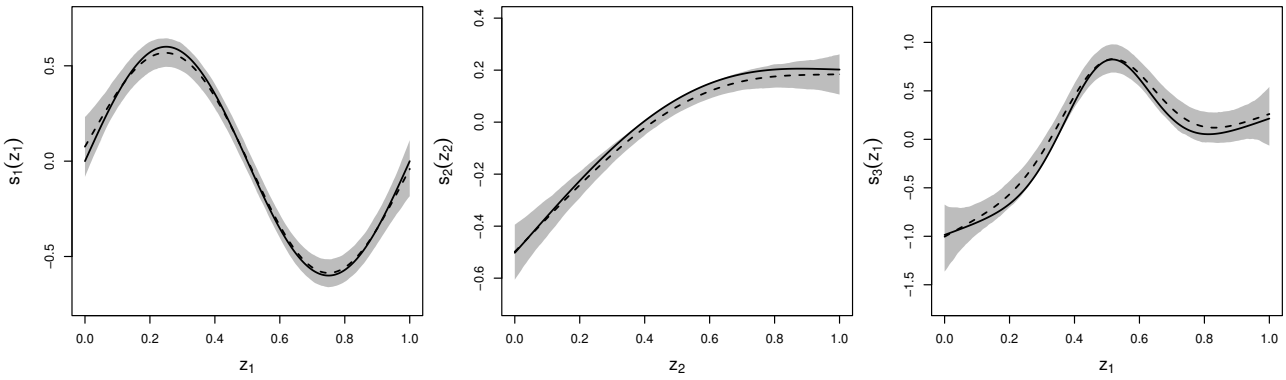


(b) $n = 4000$

Figure 3: Estimation results for parametric effects under the discrete outcome scenario.



(a) $n = 2000$



(b) $n = 4000$

Figure 4: Estimation results for smooth effects under the discrete outcome scenario.

Binary case

The simulations rely on a DGP of the form

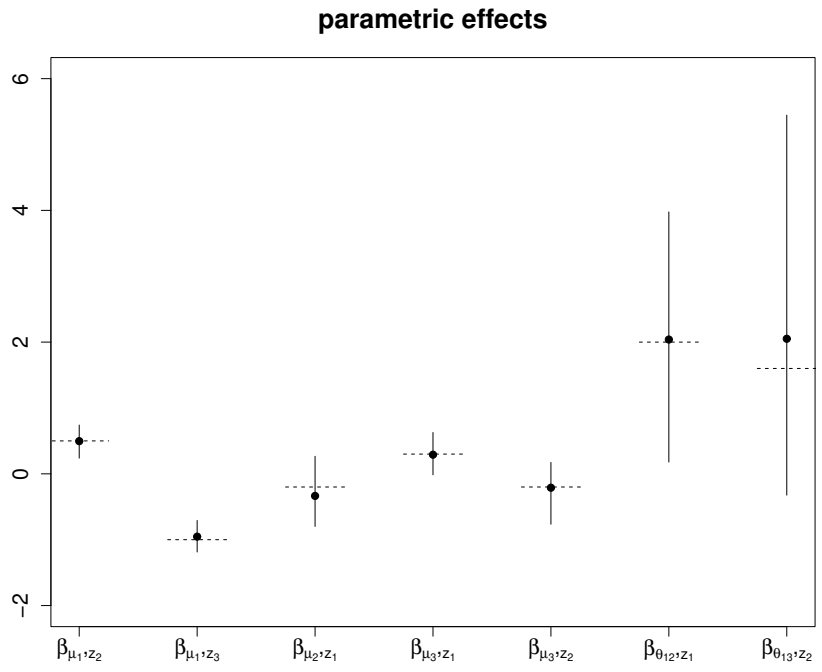
$$s \sim \text{Bernoulli with } \mu_1 = \frac{\{\exp\left(0.5 + s_1\left(z_1\right) + 0.5z_2 - z_3\right)\}}{1 + \{\exp\left(0.5 + s_1\left(z_1\right) + 0.5z_2 - z_3\right)\}},$$
$$y_2^* \sim \text{Bernoulli with } \mu_2 = \Phi\{-0.75 - 0.2z_1 + 0.6s_2\left(z_2\right)\}$$

for regime 0, and of the form

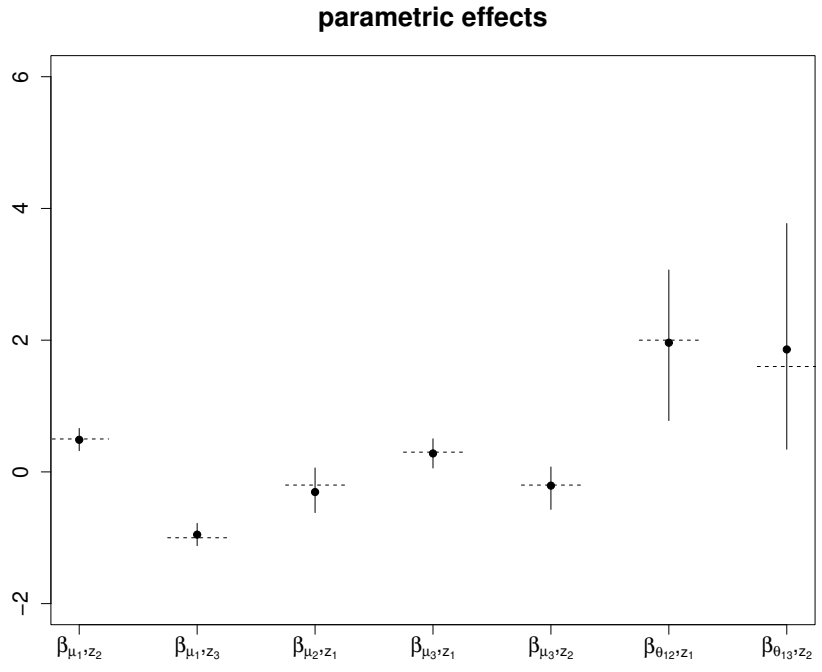
$$s \sim \text{Bernoulli with } \mu_1 = \frac{\{\exp\left(0.5 + s_1\left(z_1\right) + 0.5z_2 - z_3\right)\}}{1 + \{\exp\left(0.5 + s_1\left(z_1\right) + 0.5z_2 - z_3\right)\}},$$
$$y_3^* \sim \text{Bernoulli with } \mu_2 = \Phi\{-0.35 + 0.3z_1 - 0.2z_2\}$$

for regime 1. $\Phi(\cdot)$ denotes the cdf of a standard normal distribution. The smooth functions $s_1(z)$ and $s_2(z)$ have been previously defined, as are variables z_1, z_2 and z_3. Associated values of s and y_2^*, and s and y_3^* are generated using the Joe and Clayton copulae, where θ_{12} and θ_{13} are specified as $\exp\left(0.5 + 2z_1\right) + 1$ and $\exp\left(0.25 + 1.6z_2\right)$, respectively.

Using `gjrm()`, we fitted the endogenous switching regression model with logit and probit marginals and Joe copula for regime 0, and logit and probit marginals and Clayton copula for regime 1. The findings are similar to those previously described for the continuous and discrete cases. The bias and variability of the estimates across the two sample sizes are low for all parameters, and tend to improve with the sample size (see Figure 5). Similarly, Figure 6 shows that the true smooth functions are recovered well by the estimation method, and that the results, in terms of bias and variability, improve as the sample size increase.

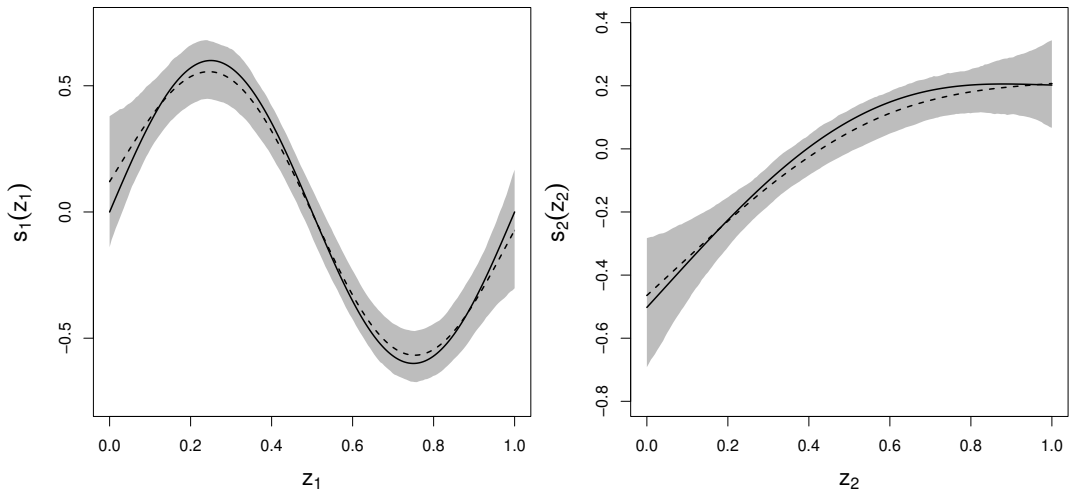


(a) $n = 2000$

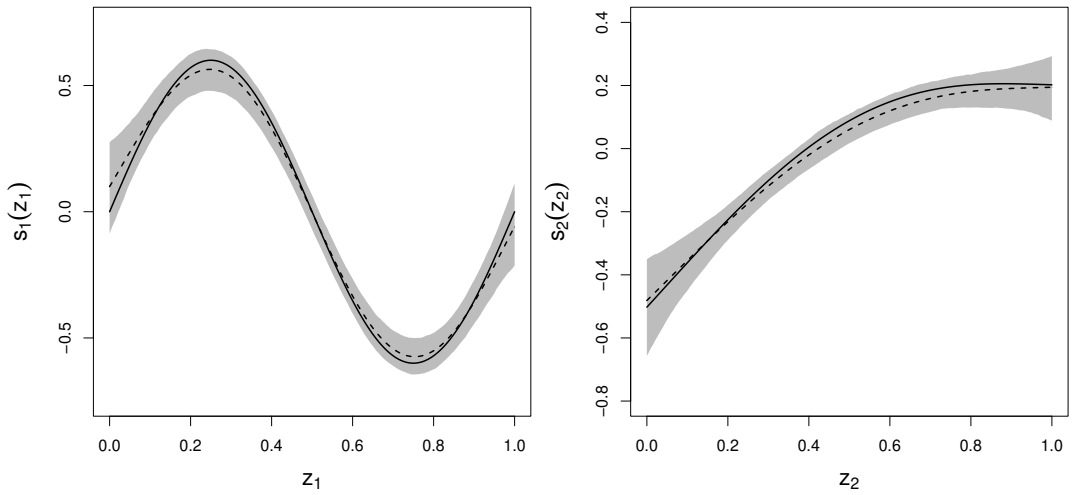


(b) $n = 4000$

Figure 5: Estimation results for parametric effects under the binary outcome scenario.



(a) $n = 2000$



(b) $n = 4000$

Figure 6: Estimation results for smooth effects under the binary outcome scenario.

Appendix C: Model fitting using GJRM

The proposed methodology has been implemented within the R package GJRM (Marra & Radice, 2022), which required extending the gjrm() function. This package has been created to enhance reproducible research and to disseminate results in a straightforward and transparent way. The function is easy to use, especially if the user is already familiar with the syntax of (generalised) linear and additive models in R. R code chunks are reported below for all three case studies: continuous, binary and discrete.

Case study 1: continuous outcome

Summary statistics for the data used in this analysis are given in Table 1.

	Insured by her employer $n = 1,141$	Not insured by her employer $n = 777$
wage	51,650	43,881
age	42.3	42.1
white	reference category	
black	0.13	0.12
hispanic	0.18	0.23
famsze2: family size of 2	reference category	
famsze3: family size of 3	0.24	0.21
famsze4: family size of 4	0.28	0.29
famsze5: family size greater than 4	0.17	0.25
edlh: less than high school	reference category	
edhs: high school degree	0.13	0.15
edsc: some college	0.22	0.23
edco: college degree	0.37	0.30
west	reference category	
neast: northest	0.15	0.17
mwest: midwest	0.20	0.20
south	0.35	0.39
union: union member	0.07	0.03
numemp: firm size	188.0	115.3
h_held: husband holds insurance	0.45	0.77
h_numemp: husband firm size	126.0	176.5

Table 1: Sample means for data from the 2012 wave of the MEPS used for case study 1.

One of the calls used for modelling the continuous data from the MEPS is

```
eq1   <- held ~ s(age) + black + hispanic + as.factor(famsze) + edhs + edsc + edco +
                neast + mwest + south + unions + numemp + h_held + h_numemp
eq23  <- wage ~ s(age) + black + hispanic + as.factor(famsze) + edhs + edsc + edco +
                neast + mwest + south + unions + numemp
eq45   <-     ~ s(age) + black + hispanic + as.factor(famsze) + edhs + edsc + edco +
                neast + mwest + south + unions + numemp
eq67   <-     ~ 1


fl    <- list(eq1, eq23, eq23, eq45, eq45, eq67, eq67)
out   <- gjrm(fl1, margins = c("probit", "iG", "FISK"), data = mydata, Model = "ROY",
              BivD = "G0", BivD2 = "C180")
```

where `fl` is a list of seven equations, one (`eq1`) related to S, two (`eq23` and `eq45`) related to Y_2, two (`eq23` and `eq45`) related to Y_3 and the remaining two (`eq67` and `eq67`) related to θ_{12} and θ_{13}, respectively. Options for the first (binary) equation are `probit`, `logit` and `cloglog`. Equation `eq1` is for μ_1, equation `eq23` is for parameters μ_2 and μ_3 of the continuous (inverse Gaussian and Fisk, in this case) distributions. Equation `eq45` is for parameters σ_2 and σ_3 of the inverse Gaussian and Fisk distributions, respectively, and `eq67` is the equation for the copula parameters θ_{12} and θ_{13}. The list of possible options for Y_2 and Y_3 are Bernoulli distributions (with `probit`, `logit` or `cloglog` links) and those listed in Tables 2 and 3 of the main paper. Arguments `BivD` and `BivD2` specify the copulae adopted for F_{12} and F_{13} (see Table 1 for the possible choices) and `Model = "ROY"` has the obvious interpretation. Symbol `s()` stands for smooth function as defined in Section 2.3 of the main paper. Default is `bs = "tp"` (penalised low rank thin plate spline) with `k = 10` (number of basis functions) and `m = 2` (order of derivatives). However, argument `bs` can also be set to, for example, `cr` (penalised cubic regression spline), `ps` (P-spline) and `mrf` (Markov random field), to name but a few. `mydata` is a data frame containing all the variables already defined in Section 7.1 of the main paper. After fitting the model, functions `conv.check()` and `post.check()` can be used to check that convergence has been achieved and that the chosen distributions adequately fit the data, respectively.

```
> conv.check(out)
```

```
Largest absolute gradient value: 1.438028e-05
Observed information matrix is positive definite
Eigenvalue range: [9.280739e-08,7.780443e+12]

Trust region iterations before smoothing parameter estimation: 21
Loops for smoothing parameter estimation: 4
Trust region iterations within smoothing loops: 23
Estimated overall probability range: 0.1314135 0.9806641
Estimated overall density range: 6.812654e-08 4.20202
```

`conv.check()` provides various information about the estimation process. Convergence is assessed by checking that the maximum of the absolute value of the score vector is virtually equal to 0 and that the observed information matrix is positive definite. `post.check()` produces the histograms and normal Q-Q plots of the normalised quantile residuals, reported in Figure 1 of the main paper, which show that the chosen distributions fit fairly well the data.

To obtain summary statistics, we can use `summary()` which works in a similar fashion as that of (generalised) linear and additive models.

```
summary(out)

COPULA 1-2: Gumbel
COPULA 1-3: 180  Clayton
MARGIN 1: Switching Mechanism - Bernoulli
MARGIN 2: Regime 0 - inverse Gaussian
MARGIN 3: Regime 1 - Fisk

EQUATION 1 - Switching Mechanism
Link function for mu.1: probit
Formula: held ~ s(age) + black + hispanic + as.factor(famsze) + edhs + edsc + edco +
                neast + mwest + south + unions + numemp + h_held + h_numemp
```

```
Parametric coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)          0.933655   0.101384   9.209  < 2e-16 ***
black                0.101623   0.097433   1.043 0.296945
hispanic            -0.184453   0.080160  -2.301 0.021388 *
as.factor(famsze)3  -0.026368   0.087565  -0.301 0.763322
as.factor(famsze)4  -0.204154   0.083774  -2.437 0.014811 *
as.factor(famsze)5  -0.448442   0.092390  -4.854 1.21e-06 ***
edhs                -0.091903   0.097280  -0.945 0.344799
edsc                 0.030079   0.078901   0.381 0.703034
edco                 0.153875   0.072721   2.116 0.034349 *
neast               -0.399121   0.100419  -3.975 7.05e-05 ***
mwest               -0.212922   0.092810  -2.294 0.021781 *
south               -0.205110   0.080223  -2.557 0.010565 *
unions               0.484815   0.150086   3.230 0.001237 **
numemp               0.014222   0.001815   7.836 4.64e-15 ***
h_held              -0.841109   0.066233 -12.699  < 2e-16 ***
h_numemp            -0.006644   0.001775  -3.743 0.000182 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Smooth components’ approximate significance:
         edf Ref.df Chi.sq p-value
s(age) 1.657  2.078   2.34   0.327


EQUATION 2 - Regime 0
Link function for mu.2: log
Formula: wage ~ s(age) + black + hispanic + as.factor(famsze) + edhs + edsc + edco +
```

```
         neast + mwest + south + unions + numemp


Parametric coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)         11.351653   0.149213  76.077  < 2e-16 ***
black               -0.112143   0.099426  -1.128 0.259360
hispanic            -0.285558   0.073942  -3.862 0.000112 ***
as.factor(famsze)3   0.033817   0.081295   0.416 0.677424
as.factor(famsze)4  -0.108853   0.076053  -1.431 0.152353
as.factor(famsze)5  -0.042998   0.090385  -0.476 0.634273
edhs                -0.484049   0.087366  -5.540 3.02e-08 ***
edsc                -0.140298   0.078980  -1.776 0.075674 .
edco                 0.277421   0.074680   3.715 0.000203 ***
neast               -0.180956   0.096452  -1.876 0.060638 .
mwest               -0.247084   0.095008  -2.601 0.009304 **
south               -0.219210   0.085300  -2.570 0.010173 *
unions               0.221818   0.170199   1.303 0.192476
numemp               0.004229   0.001632   2.591 0.009561 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Smooth components’ approximate significance:
         edf Ref.df Chi.sq p-value
s(age) 2.192  2.778  6.769  0.0694 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


EQUATION 3 - Regime 1
Link function for mu.3: log
```

```
Formula: wage ~ s(age) + black + hispanic + as.factor(famsze) + edhs + edsc + edco +
                neast + mwest + south + unions + numemp


Parametric coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)         10.6565864  0.0484040 220.159  < 2e-16 ***
black               -0.1123528  0.0442188  -2.541 0.011059 *
hispanic            -0.2053208  0.0430297  -4.772 1.83e-06 ***
as.factor(famsze)3   0.0270304  0.0420589   0.643 0.520432
as.factor(famsze)4   0.0310057  0.0440658   0.704 0.481667
as.factor(famsze)5  -0.1062012  0.0490476  -2.165 0.030367 *
edhs                -0.1751615  0.0466851  -3.752 0.000175 ***
edsc                -0.0914366  0.0401495  -2.277 0.022762 *
edco                 0.2606299  0.0362928   7.181 6.90e-13 ***
neast                0.0655319  0.0504348   1.299 0.193827
mwest               -0.0993563  0.0466613  -2.129 0.033229 *
south               -0.0644598  0.0420978  -1.531 0.125722
unions               0.0007900  0.0542413   0.015 0.988379
numemp               0.0035178  0.0007799   4.511 6.46e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Smooth components’ approximate significance:
         edf Ref.df Chi.sq  p-value
s(age) 6.651  7.769  28.44 0.000377 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


EQUATION 4 - Regime 0
```

```
Link function for sigma.2: log
Formula: ~ s(age) + black + hispanic + as.factor(famsze) + edhs + edsc + edco +
           neast + mwest + south + unions + numemp

Parametric coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)         -5.801558   0.074761 -77.602  < 2e-16 ***
black                0.228957   0.075935   3.015  0.00257 **
hispanic             0.154055   0.059355   2.595  0.00945 **
as.factor(famsze)3  -0.013479   0.069503  -0.194  0.84622
as.factor(famsze)4   0.049197   0.066638   0.738  0.46035
as.factor(famsze)5   0.208282   0.072791   2.861  0.00422 **
edhs                -0.163184   0.074302  -2.196  0.02808 *
edsc                -0.108400   0.060173  -1.801  0.07163 .
edco                -0.131152   0.055761  -2.352  0.01867 *
neast               -0.015360   0.077093  -0.199  0.84207
mwest                0.010062   0.072596   0.139  0.88976
south                0.001511   0.061292   0.025  0.98033
unions              -0.097420   0.138126  -0.705  0.48062
numemp              -0.009414   0.001437  -6.551  5.7e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Smooth components’ approximate significance:
         edf Ref.df Chi.sq p-value
s(age) 3.683  4.592  4.063   0.433


EQUATION 5 - Regime 1
Link function for sigma.3: log
```

```
Formula: ~ s(age) + black + hispanic + as.factor(famsze) + edhs + edsc + edco +
           neast + mwest + south + unions + numemp


Parametric coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)          1.0374722  0.0718558  14.438   <2e-16 ***
black                0.1221001  0.0787434   1.551   0.1210
hispanic             0.0292991  0.0705065   0.416   0.6777
as.factor(famsze)3   0.1096391  0.0689862   1.589   0.1120
as.factor(famsze)4   0.0484750  0.0664561   0.729   0.4657
as.factor(famsze)5   0.0726353  0.0772118   0.941   0.3468
edhs                 0.1292537  0.0851834   1.517   0.1292
edsc                 0.0222331  0.0662316   0.336   0.7371
edco                -0.0179561  0.0583704  -0.308   0.7584
neast                0.1269937  0.0819533   1.550   0.1212
mwest                0.1298293  0.0752446   1.725   0.0844 .
south                0.1043364  0.0654631   1.594   0.1110
unions               0.1501032  0.1015155   1.479   0.1392
numemp               0.0001106  0.0013518   0.082   0.9348
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Smooth components’ approximate significance:
       edf Ref.df Chi.sq p-value
s(age)   1      1  1.236   0.266



EQUATION 6 - Regime 0
Link function for theta.12: log(  - 1)
Formula: ~1
```

```
Parametric coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.03992    0.25177   0.159    0.874


EQUATION 7 - Regime 1
Link function for theta.13: log
Formula: ~1

Parametric coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   -16.47    3282.53  -0.005    0.996


n = 1918  n.sel0 = 777  n.sel1 = 1141
sigma.2 = 0.00271(0.00227,0.00325)  sigma.3 = 3.42(2.86,4.13)
theta.12 = 2.04(1.63,2.55)  theta.13 = 7.07e-08(4.14e-08,28)
tau.12 = 0.51(0.387,0.607)  tau.13 = 3.53e-08(2.07e-08,0.933)
total edf = 88.2
```

These summaries have been commented in the main manuscript. Function `plot()` can be used to visualise results.

```
par(mfrow = c(2, 3), mar = c(4, 5, 2, 0) + 0.1)
plot(out, eq = 1, scale = 0, select = 1, rug = TRUE, jit = TRUE)
plot(out, eq = 2, scale = 0, select = 1, rug = TRUE, jit = TRUE)
plot(out, eq = 3, scale = 0, select = 1, rug = TRUE, jit = TRUE)
plot(out, eq = 4, scale = 0, select = 1, rug = TRUE, jit = TRUE)
plot(out, eq = 5, scale = 0, select = 1, rug = TRUE, jit = TRUE)
```

They correspond to the five estimated smooth functions reported in Figure 2 of the manuscript.

To calculate the average treatment effect, with corresponding interval, function `AT()` can be used. Argument `percent = TRUE` allows the user to produce results in terms of percentage and `n.sim = 1000` is the number of simulated coefficient vectors from the posterior distribution of the estimated model parameters used to obtain intervals.

```
AT(out, percent = TRUE, n.sim = 1000)
```

```
Average treatment effect with 95% interval:
```

```
-0.282 (-0.383,-0.146)
```

Interpretable effects for the Fisk distribution can be obtained using prediction. For example, to calculate the average effect that a one-unit increase in the age variable has on the outcome, one can use for example

```
d0 <- data.frame(held = 1, age = mydata$age, black = mydata$black,
                 hispanic = mydata$hispanic, famsze5 = mydata$famsze5,
                 edhs = mydata$edhs, edsc = mydata$edsc,
                 edco = mydata$edco, neast = mydata$neast,
                 mwest = mydata$mwest, south = mydata$south,
                 unions = mydata$unions, numemp = mydata$numemp)
d0 <- data.frame(held = 1, age = mydata$age + 1, black = mydata$black,
                 hispanic = mydata$hispanic, famsze5 = mydata$famsze5,
                 edhs = mydata$edhs, edsc = mydata$edsc,
                 edco = mydata$edco, neast = mydata$neast,
                 mwest = mydata$mwest, south = mydata$south,
                 unions = mydata$unions, numemp = mydata$numemp)
eta0.1 <- predict(res.lBIC, eq = 2, newdata = d0)
eta0.2 <- predict(res.lBIC, eq = 4, newdata = d0)
eta1.1 <- predict(res.lBIC, eq = 3, newdata = d1)
eta1.2 <- predict(res.lBIC, eq = 5, newdata = d1)
```

```
mu0 <- ( exp(eta0.1)*pi/exp(eta0.2) )/( sin(pi/exp(eta0.2)) )
mu1 <- ( exp(eta1.1)*pi/exp(eta1.2) )/( sin(pi/exp(eta1.2)) )
mean((mu1 - mu0)/mu0)
```

The model with dependence parameters specified as functions of covariates can also be estimated. That is,

```
eq1   <- held ~ s(age) + black + hispanic + as.factor(famsze) + edhs + edsc + edco +
                neast + mwest + south + unions + numemp + h_held + h_numemp
eq23  <- wage ~ s(age) + black + hispanic + as.factor(famsze) + edhs + edsc + edco +
                neast + mwest + south + unions + numemp
eq45   <-     ~ s(age) + black + hispanic + as.factor(famsze) + edhs + edsc + edco +
                neast + mwest + south + unions + numemp
eq67   <-      ~ edhs + edsc + edco

fl    <- list(eq1, eq23, eq23, eq45, eq45, eq67, eq67)

out   <- gjrm(fl, margins = c("probit", "iG", "FISK"), data = mydata, Model = "ROY",
              BivD = "G0", BivD2 = "C180")
```

The following (edited) R output shows the output for the coefficients of the dependence parameters.

```
summary(out)
...
EQUATION 6 - Regime 0
Link function for theta.12: log(  - 1)
Formula: ~edhs + edsc + edco

Parametric coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.08378    0.35897  -0.233    0.815
edhs         0.30036    0.61054   0.492    0.623
```

```
edsc         -0.32284    1.01659  -0.318    0.751
edco          0.44819    0.47250   0.949    0.343


EQUATION 7 - Regime 1
Link function for theta.13: log
Formula: ~edhs + edsc + edco

Parametric coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -19.68    1595.59  -0.012    0.990
edhs           18.70    1595.59   0.012    0.991
edsc           18.85    1595.59   0.012    0.991
edco          -18.63    2136.31  -0.009    0.993
```

Finally, other familiar functions such as `AIC()`, `BIC()`, `predict()` can be used in the usual manner. Further details can be found in the documentation of the `GJRM` package in `R`.

Case study 2: binary outcome

Summary statistics for the data used in this analysis are given in Table 2.

	Insured n = 11,262	Not insured n = 4,022
`obdrv`: doctor visit?	0.64	0.30
`age`	42.7	39.7
`female`	0.49	0.41
`white`:	reference category	
`black`	0.17	0.16
`hispanic`	0.21	0.54
`married`	0.61	0.42
`famsze1`: family size of 1	reference category	
`famsze2`: family size of 2	0.27	0.19
`famsze3`: family size of 3	0.20	0.18
`famsze4`: family size of 4	0.22	0.20
`famsze5`: family size of 5	0.11	0.14
`famsze6`: family size of 6	0.04	0.07
`famsze7`: family size greater than 6	0.02	0.06
`edlh`: less than high school	reference category	
`edhs`: high school degree	0.16	0.19
`edsc`: some college	0.23	0.16
`edco`: college degree	0.29	0.08
`ttlpx`: income/1000	51.9	23.7
`west`	reference category	
`neast`: northest	0.16	0.12
`mwest`: midwest	0.22	0.14
`south`	0.350	0.46
`fairpoor`: fair or poor health	0.08	0.14
`numemp`: firm size	162.3	56.7

Table 2: Sample means for data from the 2012 and 2013 waves of the MEPS used for case study 2.

Similarly to the continuous case, the `R` syntax to fit a flexible endogenous switching regression model using data from the MEPS is

```
eq1  <- prvev ~ age + female + black + hispanic + married + as.factor(famsze6) + edhs +
                edsc + edco + ttlpx + neast + mwest + south + fairpoor + numemp
eq23 <- obdrv ~ age + female + black + hispanic + married + as.factor(famsze6) + edhs +
                edsc + edco + ttlpx + neast + mwest + south + fairpoor
eq45  <-      ~ 1
```

```
fl <- list(eq1, eq23, eq23, eq45, eq45)
out <- gjrm(fl, margins = c("logit", "logit", "logit"), data = mydata, Model = "ROY",
            BivD = "C270", BivD2 = "N")
```

The model parameter estimates are

```
summary(out)
COPULA 1-2: 270  Clayton
COPULA 1-3: Gaussian
MARGIN 1: Switching Mechanism - Bernoulli
MARGIN 2: Regime 0 - Bernoulli
MARGIN 3: Regime 1 - Bernoulli

EQUATION 1 - Switching Mechanism
Link function for mu.1: logit
Formula: prvev ~ age + female + black + hispanic + married + as.factor(famsze6) +
                 edhs + edsc + edco + ttlpx + neast + mwest + south + fairpoor +
                 numemp

Parametric coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)         -1.390675   0.128451 -10.826  < 2e-16 ***
age                  0.007424   0.002210   3.360 0.000781 ***
female               0.583377   0.046232  12.619  < 2e-16 ***
black               -0.153000   0.064563  -2.370 0.017800 *
hispanic            -1.002183   0.052129 -19.225  < 2e-16 ***
married              0.715041   0.051885  13.781  < 2e-16 ***
as.factor(famsze6)2  0.012691   0.076297   0.166 0.867887
as.factor(famsze6)3 -0.104904   0.080418  -1.304 0.192067
as.factor(famsze6)4 -0.172606   0.082278  -2.098 0.035920 *
as.factor(famsze6)5 -0.327042   0.090846  -3.600 0.000318 ***
as.factor(famsze6)6 -0.293134   0.114201  -2.567 0.010264 *
```

```
as.factor(famsze6)7 -0.854401   0.132905  -6.429 1.29e-10 ***
edhs                 0.127725   0.058959   2.166 0.030285 *
edsc                 0.300245   0.059165   5.075 3.88e-07 ***
edco                 0.594349   0.073103   8.130 4.28e-16 ***
ttlpx                0.048546   0.001564  31.033  < 2e-16 ***
neast               -0.011402   0.076354  -0.149 0.881295
mwest                0.020847   0.070886   0.294 0.768685
south               -0.280942   0.057068  -4.923 8.53e-07 ***
fairpoor            -0.169658   0.069468  -2.442 0.014596 *
numemp               0.038496   0.001897  20.297  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


EQUATION 2 - Regime 0
Link function for mu.2: logit
Formula: obdrv ~ age + female + black + hispanic + married + as.factor(famsze6) +
                 edhs + edsc + edco + ttlpx + neast + mwest + south + fairpoor


Parametric coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)         -2.526329   0.226887 -11.135  < 2e-16 ***
age                  0.022212   0.003923   5.661 1.50e-08 ***
female               0.924082   0.082390  11.216  < 2e-16 ***
black               -0.069081   0.119726  -0.577 0.563948
hispanic             0.131477   0.114602   1.147 0.251277
married              0.395872   0.093352   4.241 2.23e-05 ***
as.factor(famsze6)2 -0.048398   0.131243  -0.369 0.712300
as.factor(famsze6)3 -0.251465   0.136761  -1.839 0.065956 .
as.factor(famsze6)4 -0.386101   0.140067  -2.757 0.005842 **
```

```
as.factor(famsze6)5 -0.564675   0.155062  -3.642 0.000271 ***
as.factor(famsze6)6 -0.669130   0.193743  -3.454 0.000553 ***
as.factor(famsze6)7 -0.780401   0.208262  -3.747 0.000179 ***
edhs                -0.073853   0.103878  -0.711 0.477110
edsc                 0.101509   0.108853   0.933 0.351060
edco                 0.009480   0.151926   0.062 0.950243
ttlpx                0.001359   0.005007   0.271 0.786031
neast               -0.011043   0.137838  -0.080 0.936148
mwest                0.289097   0.124775   2.317 0.020507 *
south                0.018084   0.095836   0.189 0.850327
fairpoor             0.756052   0.102962   7.343 2.09e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
EQUATION 3 - Regime 1
Link function for mu.3: logit
Formula: obdrv ~ age + female + black + hispanic + married + as.factor(famsze6) +
                 edhs + edsc + edco + ttlpx + neast + mwest + south + fairpoor

Parametric coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)         -0.6863875  0.1320077  -5.200 2.00e-07 ***
age                  0.0233058  0.0020249  11.510  < 2e-16 ***
female               0.7861674  0.0443886  17.711  < 2e-16 ***
black               -0.2494801  0.0584396  -4.269 1.96e-05 ***
hispanic             0.0419265  0.0603593   0.695 0.487296
married              0.2886930  0.0538823   5.358 8.42e-08 ***
as.factor(famsze6)2 -0.2379138  0.0729861  -3.260 0.001115 **
as.factor(famsze6)3 -0.4359142  0.0780003  -5.589 2.29e-08 ***
```

```
as.factor(famsze6)4 -0.5281958  0.0799416  -6.607 3.91e-11 ***
as.factor(famsze6)5 -0.4960410  0.0913915  -5.428 5.71e-08 ***
as.factor(famsze6)6 -1.0579122  0.1208198  -8.756  < 2e-16 ***
as.factor(famsze6)7 -0.6797689  0.1603574  -4.239 2.24e-05 ***
edhs                -0.0546203  0.0604281  -0.904 0.366054
edsc                 0.0778587  0.0538152   1.447 0.147959
edco                 0.1077880  0.0534270   2.017 0.043645 *
ttlpx                0.0017990  0.0007355   2.446 0.014444 *
neast                0.0851615  0.0645468   1.319 0.187043
mwest                0.2258218  0.0604011   3.739 0.000185 ***
south                0.1730747  0.0543029   3.187 0.001437 **
fairpoor             0.6564940  0.0832818   7.883 3.20e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


EQUATION 4 - Regime 0
Link function for theta.12: log(-  )
Formula: ~1

Parametric coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.9401     0.5271  -3.681 0.000232 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


EQUATION 5 - Regime 1
Link function for theta.13: atanh
Formula: ~1
```

```
Parametric coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.32610    0.07157  -4.556 5.21e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


n = 15284  n.sel0 = 4022  n.sel1 = 11262
theta.12 = -0.144(-0.423,-0.054)  theta.13 = -0.315(-0.456,-0.174)
tau.12 = -0.067(-0.175,-0.0263)  tau.13 = -0.204(-0.301,-0.112)
total edf = 63
```

The average treatment effect, with corresponding interval, is

```
AT(out, n.sim = 1000)


Average treatment effect with 95% interval:


0.366 (0.259,0.443)
```

Discrete case

Summary statistics for the data used in this analysis are given in Table 3.

	Curative visit $n = 1,412$	No curative visit $n = 2,908$
`days`: Missed school days	3.33	1.50
`age`	9.29	9.55
`female`	0.48	0.47
`black`	0.15	0.25
`hispanic`	0.38	0.44
`west`	reference category	
`neast`: northest	0.13	0.13
`mwest`: midwest	0.21	0.17
`south`	0.38	0.42
`famsze2`: family size of 2	reference category	
`famsze3`: family size of 3	0.17	0.14
`famsze4`: family size of 4	0.34	0.29
`famsze5`: family size of 5	0.25	0.27
`famsze6`: family size of 6	0.11	0.15
`famsze7`: family size of 7	0.04	0.06
`famsze8`: family size greater than 7	0.02	0.05
`povcat1`: poverty category 1	reference category	
`povcat2`: poverty category 2	0.08	0.08
`povcat3`: poverty category 3	0.16	0.20
`povcat4`: poverty category 4	0.24	0.24
`povcat5`: poverty category 5	0.22	0.12
`notenglish`: language other than English spoken at home	0.42	0.49
`uscquick`: can travel to doctor in < 30 minutes	0.90	0.80

Table 3: Sample means for data from the 2015 wave of the MEPS used for case study 3.

The `R` code for fitting a flexible endogenous switching regression model using discrete data is

```
eq1  <- cure ~ s(age, k = 5) + female + black + hispanic + r_noreast + r_midwest +
               r_south + as.factor(famsze8) + as.factor(povcat) + notenglish +
               uscquick
eq23 <- days ~ s(age, k = 5) + female + black + hispanic + r_noreast + r_midwest +
               r_south + as.factor(famsze8) + as.factor(povcat) + notenglish
eq45  <-     ~ s(age, k = 5) + female + black + hispanic + r_noreast + r_midwest +
                r_south + as.factor(famsze8) + as.factor(povcat) + notenglish
```

```
eq67  <-     ~ 1

fl  <- list(eq1, eq23, eq23, eq45, eq45, eq67, eq67)

out <- gjrm(fl, margins = c("logit", "NBII", "NBII"), data = mydata, Model = "ROY",
            BivD = "J90", BivD2 = "J0")

post.check(out)
```

The randomised normalised quantile residuals (see Figure 7) show that the model seems to fit the data adequately.

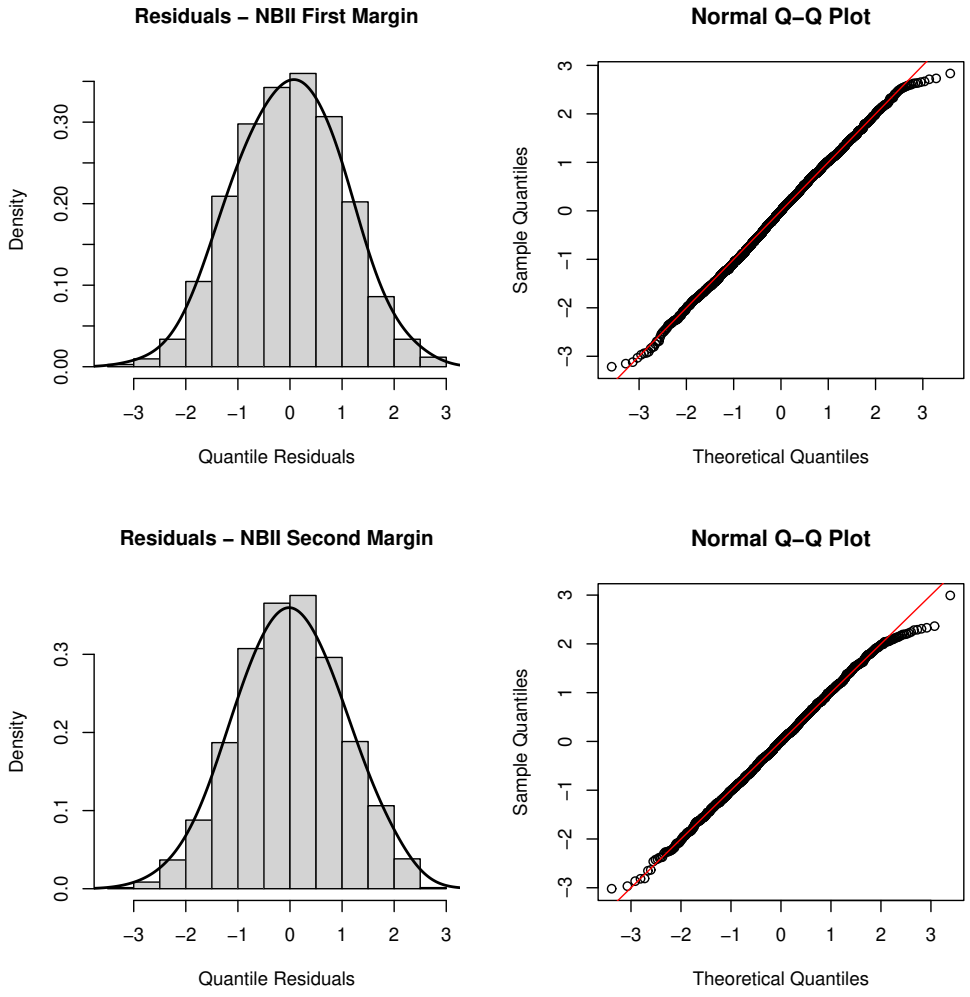


Figure 7: Histograms of randomised normalised quantile residuals and normal Q-Q plots of residuals for the outcome variable under regime 0 (top) and regime 1 (bottom).

The estimated smooth functions (see Figure 8), parametric coefficients and average treatment effect can easily be obtained as illustrated below.

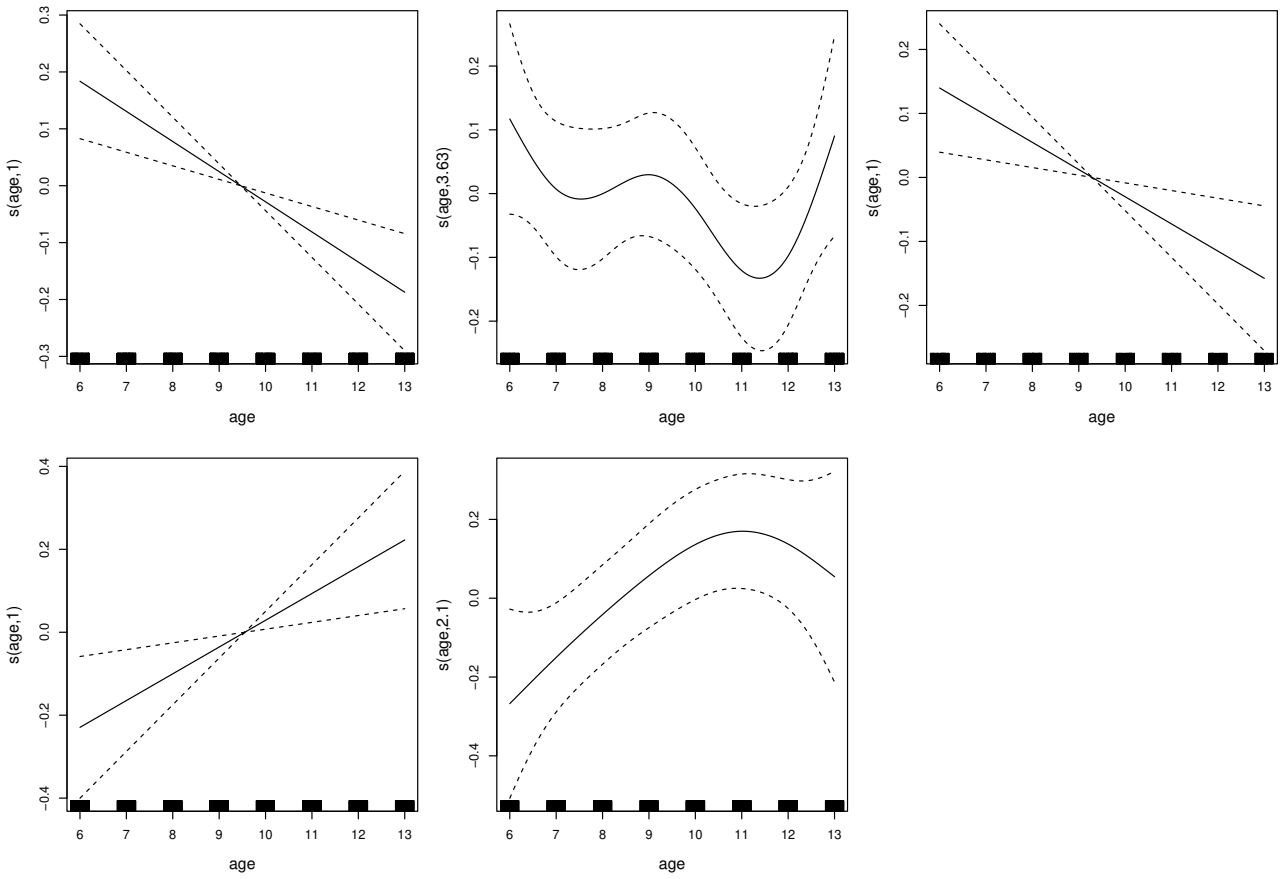


Figure 8: Estimated smooth effects of age for the logit equation, for the μ parameters of the negative binomial type II distributions, and for the σ parameters of the same distributions. 95% point-wise intervals are also reported. The jittered rug plot, at the bottom of each graph, shows the covariate values. The numbers in the brackets of the y-axis captions are the *edf* of the smooth curves. Note that the estimated smooth functions are centered around zero because of the centering identifiability constraints. When $edf = 1$, the intervals correctly exhibit the behaviour displayed in the related plots.

```
par(mfrow = c(2, 3), mar = c(4, 5, 2, 0) + 0.1  )
```

```
plot(out, eq = 1, scale = 0, select = 1, rug = TRUE, jit = TRUE)
plot(out, eq = 2, scale = 0, select = 1, rug = TRUE, jit = TRUE)
plot(out, eq = 3, scale = 0, select = 1, rug = TRUE, jit = TRUE)
plot(out, eq = 4, scale = 0, select = 1, rug = TRUE, jit = TRUE)
plot(out, eq = 5, scale = 0, select = 1, rug = TRUE, jit = TRUE)
```

```
summary(out)
```

```
COPULA 1-2: 90  Joe

COPULA 1-3: Joe

MARGIN 1: Switching Mechanism - Bernoulli

MARGIN 2: Regime 0 - Negative Binomial - Type II

MARGIN 3: Regime 1 - Negative Binomial - Type II


EQUATION 1 - Switching Mechanism

Link function for mu.1: logit

Formula: cure ~ s(age, k = 5) + female + black + hispanic + r_noreast + r_midwest +

                r_south + as.factor(famsze8) + as.factor(povcat) + notenglish +

    uscquick


Parametric coefficients:

                    Estimate Std. Error z value Pr(>|z|)

(Intercept)         -0.64852    0.19604  -3.308 0.000939 ***

female               0.03691    0.06657   0.554 0.579301

black               -0.83719    0.10182  -8.223  < 2e-16 ***

hispanic            -0.17196    0.09863  -1.743 0.081267 .

r_noreast           -0.06304    0.11312  -0.557 0.577361

r_midwest           -0.02118    0.10298  -0.206 0.837019

r_south             -0.01471    0.08577  -0.171 0.863862

as.factor(famsze8)3 -0.08582    0.16760  -0.512 0.608622

as.factor(famsze8)4 -0.23589    0.15812  -1.492 0.135734

as.factor(famsze8)5 -0.37090    0.16063  -2.309 0.020943 *

as.factor(famsze8)6 -0.57074    0.17569  -3.248 0.001160 **

as.factor(famsze8)7 -0.71895    0.21802  -3.298 0.000975 ***

as.factor(famsze8)8 -1.03579    0.24237  -4.274 1.92e-05 ***

as.factor(povcat)2   0.06563    0.13099   0.501 0.616377

as.factor(povcat)3  -0.09616    0.09980  -0.964 0.335292
```

```
as.factor(povcat)4  -0.07487    0.09393  -0.797 0.425404
as.factor(povcat)5   0.36425    0.11033   3.301 0.000962 ***
notenglish          -0.29781    0.09284  -3.208 0.001337 **
uscquick             0.70362    0.09510   7.398 1.38e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Smooth components’ approximate significance:
       edf Ref.df Chi.sq  p-value
s(age)   1      1  13.23 0.000275 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


EQUATION 2 - Regime 0
Link function for mu.2: log
Formula: days ~ s(age, k = 5) + female + black + hispanic + r_noreast + r_midwest +
                r_south + as.factor(famsze8) + as.factor(povcat) + notenglish

Parametric coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)          0.878924   0.199342   4.409 1.04e-05 ***
female              -0.117858   0.070407  -1.674 0.094137 .
black               -0.230860   0.103011  -2.241 0.025019 *
hispanic            -0.033035   0.103519  -0.319 0.749637
r_noreast            0.163015   0.122739   1.328 0.184131
r_midwest           -0.062925   0.103254  -0.609 0.542248
r_south             -0.364064   0.092632  -3.930 8.49e-05 ***
as.factor(famsze8)3 -0.168890   0.178784  -0.945 0.344834
as.factor(famsze8)4 -0.215881   0.168542  -1.281 0.200237
```

```
as.factor(famsze8)5 -0.408128   0.171584  -2.379 0.017379 *
as.factor(famsze8)6 -0.546609   0.190494  -2.869 0.004112 **
as.factor(famsze8)7 -0.346811   0.231838  -1.496 0.134675
as.factor(famsze8)8 -0.009186   0.218985  -0.042 0.966540
as.factor(povcat)2  -0.050693   0.131185  -0.386 0.699180
as.factor(povcat)3   0.007062   0.101505   0.070 0.944532
as.factor(povcat)4  -0.088240   0.095661  -0.922 0.356308
as.factor(povcat)5  -0.384332   0.128276  -2.996 0.002734 **
notenglish          -0.348412   0.098191  -3.548 0.000388 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Smooth components’ approximate significance:
        edf Ref.df Chi.sq p-value
s(age) 3.63  3.919  8.027   0.117



EQUATION 3 - Regime 1
Link function for mu.3: log
Formula: days ~ s(age, k = 5) + female + black + hispanic + r_noreast + r_midwest +
                r_south + as.factor(famsze8) + as.factor(povcat) + notenglish

Parametric coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)          1.564042   0.154438  10.127  < 2e-16 ***
female               0.067505   0.067215   1.004 0.315222
black               -0.679874   0.137693  -4.938 7.91e-07 ***
hispanic            -0.031854   0.097389  -0.327 0.743608
r_noreast            0.203741   0.112285   1.814 0.069603 .
r_midwest           -0.056513   0.096579  -0.585 0.558445
```

```
r_south             -0.171323   0.091225  -1.878 0.060379 .
as.factor(famsze8)3 -0.251047   0.133555  -1.880 0.060145 .
as.factor(famsze8)4 -0.474404   0.126485  -3.751 0.000176 ***
as.factor(famsze8)5 -0.407027   0.131573  -3.094 0.001978 **
as.factor(famsze8)6 -1.008446   0.188318  -5.355 8.56e-08 ***
as.factor(famsze8)7 -0.642992   0.226565  -2.838 0.004540 **
as.factor(famsze8)8 -0.569796   0.279354  -2.040 0.041381 *
as.factor(povcat)2   0.015392   0.131000   0.117 0.906466
as.factor(povcat)3  -0.001824   0.115683  -0.016 0.987420
as.factor(povcat)4  -0.076960   0.098379  -0.782 0.434049
as.factor(povcat)5  -0.120577   0.104563  -1.153 0.248847
notenglish          -0.428705   0.102566  -4.180 2.92e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Smooth components’ approximate significance:
       edf Ref.df Chi.sq p-value
s(age)   1      1  7.764 0.00533 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1



EQUATION 4 - Regime 0
Link function for sigma.2: log
Formula: ~ s(age, k = 5) + female + black + hispanic + r_noreast + r_midwest +
           r_south + as.factor(famsze8) + as.factor(povcat) + notenglish

Parametric coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)          1.8993685  0.3052245   6.223 4.88e-10 ***
```

```
female               -0.2029161  0.1125355  -1.803  0.07137 .
black                -0.0139303  0.1558955  -0.089  0.92880
hispanic              0.0002043  0.1559999   0.001  0.99895
r_noreast             0.4783852  0.1851500   2.584  0.00977 **
r_midwest            -0.1382934  0.1693023  -0.817  0.41402
r_south              -0.0393854  0.1482455  -0.266  0.79049
as.factor(famsze8)3 -0.4323473  0.2695046  -1.604  0.10866
as.factor(famsze8)4 -0.3983620  0.2509564  -1.587  0.11243
as.factor(famsze8)5 -0.4063417  0.2553452  -1.591  0.11153
as.factor(famsze8)6 -0.3714422  0.2844513  -1.306  0.19161
as.factor(famsze8)7 -0.0609952  0.3423269  -0.178  0.85858
as.factor(famsze8)8 -0.1899950  0.3386801  -0.561  0.57481
as.factor(povcat)2  -0.2912458  0.2138583  -1.362  0.17324
as.factor(povcat)3  -0.1255781  0.1600029  -0.785  0.43254
as.factor(povcat)4  -0.2100805  0.1485085  -1.415  0.15719
as.factor(povcat)5  -0.3348258  0.1969505  -1.700  0.08912 .
notenglish           -0.0258460  0.1464059  -0.177  0.85987
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Smooth components’ approximate significance:
       edf Ref.df Chi.sq p-value
s(age)   1      1  7.208 0.00726 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1



EQUATION 5 - Regime 1
Link function for sigma.3: log
Formula: ~ s(age, k = 5) + female + black + hispanic + r_noreast + r_midwest +
```

```
          r_south + as.factor(famsze8) + as.factor(povcat) + notenglish


Parametric coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)          1.098824   0.303125   3.625 0.000289 ***
female               0.190474   0.121641   1.566 0.117379
black               -0.137023   0.184193  -0.744 0.456932
hispanic            -0.408446   0.167109  -2.444 0.014518 *
r_noreast            0.001641   0.204763   0.008 0.993605
r_midwest           -0.521681   0.184385  -2.829 0.004665 **
r_south             -0.124930   0.154306  -0.810 0.418156
as.factor(famsze8)3  0.325554   0.289224   1.126 0.260330
as.factor(famsze8)4  0.237064   0.277049   0.856 0.392177
as.factor(famsze8)5  0.243794   0.281351   0.867 0.386209
as.factor(famsze8)6  0.530289   0.323299   1.640 0.100954
as.factor(famsze8)7  0.037756   0.395628   0.095 0.923970
as.factor(famsze8)8  0.161187   0.467116   0.345 0.730043
as.factor(povcat)2  -0.181393   0.237303  -0.764 0.444631
as.factor(povcat)3   0.067904   0.183664   0.370 0.711594
as.factor(povcat)4  -0.280581   0.170674  -1.644 0.100186
as.factor(povcat)5  -0.451696   0.196651  -2.297 0.021622 *
notenglish           0.215595   0.156001   1.382 0.166968
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


Smooth components’ approximate significance:
         edf Ref.df Chi.sq p-value
s(age) 2.099  2.565  5.752  0.0692 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
EQUATION 6 - Regime 0
Link function for theta.12: log(-   - 1)
Formula: ~1

Parametric coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.1767     0.2943    -0.6    0.548



EQUATION 7 - Regime 1
Link function for theta.13: log(  - 1)
Formula: ~1

Parametric coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.9547     0.3036  -3.145  0.00166 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


n = 4320  n.sel0 = 2908  n.sel1 = 1412
sigma.2 = 3.84(2.47,6.07)  sigma.3 = 3.12(1.86,5.21)
theta.12 = -1.84(-2.31,-1.49)  theta.13 = 1.38(1.19,1.72)
tau.12 = -0.317(-0.417,-0.217)  tau.13 = 0.179(0.0984,0.285)
total edf = 102

AT(res.lAIC, n.sim = 1000)
```

```
Average treatment effect with 95% interval:


0.938 (0.551,1.377)
```

An alternative model with dependence parameters specified as functions of covariate(s) can also be estimated.

```
eq1  <- cure ~ s(age, k = 5) + female + black + hispanic + r_noreast + r_midwest +
               r_south + as.factor(famsze8) + as.factor(povcat) + notenglish +
               uscquick
eq23 <- days ~ s(age, k = 5) + female + black + hispanic + r_noreast + r_midwest +
               r_south + as.factor(famsze8) + as.factor(povcat) + notenglish
eq45  <-     ~ s(age, k = 5) + female + black + hispanic + r_noreast + r_midwest +
               r_south + as.factor(famsze8) + as.factor(povcat) + notenglish
eq67  <-     ~ notenglish
fl  <- list(eq1, eq23, eq23, eq45, eq45, eq67, eq67)
out <- gjrm(fl, margins = c("logit", "NBII", "NBII"), data = mydata, Model = "ROY",
            BivD = "J90", BivD2 = "J0")
```

The following (edited) R output shows the coefficients for the dependence parameters.

```
summary(out)
...
EQUATION 6 - Regime 0
Link function for theta.12: log(-   - 1)
Formula: ~notenglish


Parametric coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.0555     0.2698   0.206    0.837
notenglish   -0.9873     0.7387  -1.337    0.181
```

```
EQUATION 7 - Regime 1
Link function for theta.13: log(  - 1)
Formula: ~ notenglish

Parametric coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.8663     0.3617  -2.395   0.0166 *
notenglish   -0.1766     0.4678  -0.378   0.7058
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

n = 4320  n.sel0 = 2908  n.sel1 = 1412
sigma.2 = 3.83(2.42,6.11)  sigma.3 = 3.12(1.87,5.19)
theta.12 = -1.75(-2.53,-1.37)  theta.13 = 1.39(1.2,1.82)
tau.12 = -0.281(-0.451,-0.162)  tau.13 = 0.18(0.102,0.311)
total edf = 104
```

The results show that the dependence parameters do not have any association with `notenglish` and that a more parsimonious model, as the one previously fitted, is more appropriate.

References

Akaike, H. (1987). Information theory and an extension of the maximum likelihood principle. *In: Petrov, B.N., Csaki, B.F. (eds.) Second International Symposium on Information Theory*, Academiai Kiado, Budapest.

Marra, G., Radice, R., Bärnighausen, T., Wood, S. N., & McGovern, M. E. (2017). A Simultaneous Equation Approach to Estimating HIV Prevalence with Non-Ignorable Missing Responses. *Journal of the American Statistical Association*, 112(518), 484–496.

Marra, G. & Radice, R. (2020). Copula Link-Based Additive Models for Right-Censored Event Time Data. *Journal of the American Statistical Association*, 115(530), 886–895.

Marra, G. & Radice, R. (2022). *GJRM: Generalized Joint Regression Modeling.* R package version 0.2-6, URL `https://cran.r-project.org/package=GJRM`

Nocedal, J. & Wright, S. J. (2006). *Numerical Optimization: Second Edition.* Springer, New York, NY, USA.

Wood, S. N. (2017). *Generalized Additive Models: An Introduction With R: Second Edition.* Chapman & Hall/CRC, London.