Econometric Reviews

Editor: Esfandiar Maasoumi

Included in this print edition:
Numbers 9–10

Volume 42 2023

Econometric Reviews

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/lecr20

A unifying switching regime regression framework with applications in health economics

Giampiero Marra, Rosalba Radice & David Zimmer

To cite this article: Giampiero Marra, Rosalba Radice & David Zimmer (11 Oct 2023): A unifying switching regime regression framework with applications in health economics, Econometric Reviews, DOI: 10.1080/07474938.2023.2255438

To link to this article: https://doi.org/10.1080/07474938.2023.2255438

ECONOMETRIC REVIEWS
https://doi.org/10.1080/07474938.2023.2255438

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS

A unifying switching regime regression framework with applications in health economics

Giampiero Marra[a], Rosalba Radice[b], and David Zimmer[c]

[a]University College London, UK; [b]Bayes Business School, City, University of London, UK; [c]Western Kentucky University, Bowling Green, Kentucky, USA

ABSTRACT

Motivated by three health economics-related case studies, we propose a unifying and flexible regression modeling 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 penalized maximum likelihood estimation approach. The proposed modeling 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.

ARTICLE HISTORY
Received 25 April 2023
Accepted 21 August 2023

KEYWORDS
Copula; penalized regression spline; simultaneous estimation; structural equation model; switching regime.

1. Introduction

Inspired by health economics-related studies, we introduce a unifying and flexible regression modeling 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 (Cox, 1958; Neyman, 1923). The main difference between the two frameworks is in the model's formalization; in economics, models are thought of in terms of

CONTACT Giampiero Marra giampiero.marra@ucl.ac.uk University College London, London, UK.
Supplemental data for this article can be accessed online at https://doi.org/10.1080/07474938.2023.2255438.

realized, 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 labor economics where it has been abundantly applied, discussed, and extended in several ways (e.g., Chen et al., 2014; Cornelissen et al., 2016; D'Haultfoeuille and Maurel, 2013; Eisenhauer et al., 2015; French and Taber, 2011; Heckman and Honore, 1990; Heckman et al., 2003; Smith, 2005). Applications and extensions have also appeared in other fields (e.g., Bayer et al., 2011; Fitawek and Hendriks, 2021; Kim, 2021; Moscelli et al., 2018), hence highlighting the relevance of the modeling framework. When the approach was first developed and formalized, multivariate normality was assumed for theoretical and computational tractability. Since this assumption is clearly questionable in applications, extensions were proposed. A notable example is Smith (2005) who introduced a class of alternatives to the multivariate Gaussian based on copulae, a modeling 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 and Sajaia (2004) and Toomet and Henningsen (2020). Hasebe (2013) provided an implementation where the normal, logistic and Student's *t* univariate distributions as well as several bivariate copulae 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 parametric but flexible 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 traditional ones. Also, since all distributional parameters can be modeled as a function of observed confounders, the information contained in the data can be exploited more fully. 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 the functional 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.

The potential drawback of the proposed framework is that treatment effect identification is essentially based on parametric functional form assumptions. However, in spite of this, the flexibility offered by the methodology enables the data to point to meaningful model structures, hence capturing in a sense the spirit of distribution-free methods. Moreover, as stressed, by Chen et al. (2014, see also references therein), the use of non parametric techniques in this context may be problematic because, among other things, 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 approach allows for many layers of complexity; however, there is no price to pay in terms of usability and interpretability. In fact, the modeling approach has been incorporated in the software package GJRM (Marra and Radice, 2023), written for the programming language R (R Core Team, 2023), which significantly eases the use of our proposal. Parameter estimation relies on a carefully structured algorithm, whereas inference exploits a Bayesian result often employed for penalized likelihood-based models. The proposed methodological developments, together with fast and reliable software implementation, represent a significant advance in switching regression modeling. 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 organized 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 (which is made up of Appendices A, B, and C) provides details on the algorithm, discusses the findings of various simulation studies, and illustrates the use of GJRM in the three empirical applications.

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 detailed in Section 2.3). Potentially useful in some empirical settings, the marginal distributions for regimes 2 and 3 need not follow identical functional forms (Smith, 2005). Our first case study demonstrates one such setting, where the presense (or lack thereof) of insurance coverage through a person's place of employment appears to alter the entire distributional shape of wage compensation. Recall that, in line with previous work (e.g., Heckman et al., 2003), the equation related to the switching mechanism usually requires an instrument or exclusion restriction. 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 dependencies across 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 modeling 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.,

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

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

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

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 and 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

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

where $C_{1j} : (0,1)^2 \rightarrow (0,1)$ 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), counterclockwise-rotated versions of them can be obtained (Brechmann and 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 $F_{12}(y_{1i}^{*}, y_{2i}^{*}|\boldsymbol{\varphi}_1, \boldsymbol{\varphi}_2, \theta_{12})$ 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 employ 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

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.

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")	$u_1 u_2 \exp\left[\left\{(-\log u_1)^{-\theta}\right.\right.$ $\left.\left.+(-\log u_2)^{-\theta}\right\}^{-1/\theta}\right]$	$(0, \infty)$	$\log(\theta)$
Gaussian ("N")	$\Phi_2\left(\Phi^{-1}(u_1), \Phi^{-1}(u_2); \theta\right)$	$[-1, 1]$	$\tanh^{-1}(\theta)$
Gumbel ("G0")	$\exp\left[-\left\{(-\log u_1)^{\theta}\right.\right.$ $\left.\left.+(-\log u_2)^{\theta}\right\}^{1/\theta}\right]$	$[1, \infty)$	$\log(\theta - 1)$
Joe ("J0")	$1 - \left\{(1 - u_1)^{\theta} + (1 - u_2)^{\theta}\right.$ $\left.-(1 - u_1)^{\theta}(1 - u_2)^{\theta}\right\}^{1/\theta}$	$(1, \infty)$	$\log(\theta - 1)$
Plackett ("PL")	$\left(Q - \sqrt{R}\right) / \{2(\theta - 1)\}$	$(0, \infty)$	$\log(\theta)$
Student's t ("T")	$t_{2,\zeta}\left(t_{\zeta}^{-1}(u_1), t_{\zeta}^{-1}(u_2); \zeta, \theta\right)$	$[-1, 1]$	$\tanh^{-1}(\theta)$

$\Phi_2(\cdot, \cdot; \theta)$ 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 `copula` and `copula2` 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; e.g., `copula = "C0"` and `copula2 = "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 $C_{270}(u_1, u_2; \theta) = u_1 - C(u_1, 1 - u_2)$.

Table 2. Definition and some properties of the main discrete distributions implemented in GJRM.

	$f(y \mid \mu, \sigma)$	$\mathbb{E}(Y)$	$\mathbb{V}(Y)$
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	$\mu + \sigma\mu^2$

These have been parametrized according to Rigby and 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 $K_{\lambda}(t) = \frac{1}{2}\int_0^{\infty} x^{\lambda-1} \exp\left\{-0.5t(x + x^{-1})\right\} dx$ are 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, which models the probability of switching, other choices are possible: `probit` and `cloglog`.

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, $(\mu_j, \sigma_j)^{\mathsf{T}}$ or $(\mu_j, \sigma_j, \nu_j)^{\mathsf{T}}$.

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

Table 3. Definition and some properties of the main distributions implemented in GJRM. These have been conveniently parametrised according to Rigby and 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 $\mathrm{erf}(\cdot)$ 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`.

	$F(y\|\mu,\sigma,\nu)$	$f(y\|\mu,\sigma,\nu)$	$\mathbb{E}(Y)$	$\mathbb{V}(Y)$	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")	$\left\{1+\left(\frac{y}{\mu}\right)^{-\sigma}\right\}^{-\nu}$	$\frac{\sigma\nu}{y}\left[\frac{\left(\frac{y}{\mu}\right)^{\sigma\nu}}{\left\{\left(\frac{y}{\mu}\right)^{\sigma}+1\right\}^{\nu+1}}\right]$	$-\frac{\mu}{\sigma}\frac{\Gamma\left(-\frac{1}{\sigma}\right)\Gamma\left(\frac{1}{\sigma}+\nu\right)}{\Gamma(\nu)}$ if $\sigma > 1$	$-\left(\frac{\mu}{\sigma}\right)^2\left[2\sigma\frac{\Gamma\left(-\frac{2}{\sigma}\right)\Gamma\left(\frac{2}{\sigma}+\nu\right)}{\Gamma(\nu)} + \left\{\frac{\Gamma\left(-\frac{1}{\sigma}\right)\Gamma\left(\frac{1}{\sigma}+\nu\right)}{\Gamma(\nu)}\right\}^2\right]$ 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")	$\Phi\left\{\frac{1}{\sqrt{y\sigma^2}}\left(\frac{y}{\mu}-1\right)\right\} + \exp\left(\frac{2}{\mu\sigma^2}\right)$ $\Phi\left\{-\frac{1}{\sqrt{y\sigma^2}}\left(\frac{y}{\mu}+1\right)\right\}$	$\frac{1}{\sqrt{2\pi\sigma^2 y^3}}\exp\left\{-\frac{1}{2\mu^2\sigma^2 y}(y-\mu)^2\right\}$	μ	$\mu^3\sigma^2$	$y > 0$ $\mu > 0, \sigma > 0$

(*Continued*)

Table 3. Continued.

	$F(y\|\mu,\sigma,\nu)$	$f(y\|\mu,\sigma,\nu)$	$\mathbb{E}(Y)$	$\mathbb{V}(Y)$	Support of y Parameters' ranges
log-normal ("LN")	$\frac{1}{2}+\frac{1}{2}\text{erf}\left\{\frac{\log(y)-\mu}{\sigma\sqrt{2}}\right\}$	$\frac{1}{y\sigma\sqrt{2\pi}}\exp\left[-\frac{\{\log(y)-\mu\}^2}{2\sigma^2}\right]$	$\sqrt{\exp\left(\sigma^2\right)}\exp\left(\mu\right)$	$\exp\left(\sigma^2\right)\left\{\exp\left(\sigma^2\right)\right.$ $\left.-1\right\}\exp\left(2\mu\right)$	$y>0$ $-\infty<\mu<\infty,\sigma>0$
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")	$1-\left\{1+\left(\frac{y}{\mu}\right)^{\sigma}\right\}^{-\nu}$	$\frac{\sigma\nu y^{\sigma-1}}{\mu^{\sigma}\left\{1+\left(\frac{y}{\mu}\right)^{\sigma}\right\}^{\nu+1}}$	$\mu\frac{\Gamma\left(1+\frac{1}{\sigma}\right)\Gamma\left(-\frac{1}{\sigma}+\nu\right)}{\Gamma(\nu)}$ if $\sigma\nu>1$	$\mu^2\left\{\Gamma\left(1+\frac{2}{\sigma}\right)\Gamma(\nu)\Gamma\left(-\frac{2}{\sigma}+\nu\right)\right.$ $\left.-\Gamma\left(1+\frac{1}{\sigma}\right)^2\Gamma\left(-\frac{1}{\sigma}+\nu\right)^2\right\}$ if $\sigma\nu>2$	$y>0$ $\mu>0,\sigma>0,\nu>0$
Weibull ("WEI")	$1-\exp\left\{-\left(\frac{y}{\mu}\right)^{\sigma}\right\}$	$\frac{\sigma}{\mu}\left(\frac{y}{\mu}\right)^{\sigma-1}\exp\left\{-\left(\frac{y}{\mu}\right)^{\sigma}\right\}$	$\mu\Gamma\left(\frac{1}{\sigma}+1\right)$	$\mu^2\left[\Gamma\left(\frac{2}{\sigma}+1\right)\right.$ $\left.-\left\{\Gamma\left(\frac{1}{\sigma}+1\right)\right\}^2\right]$	$y>0$ $\mu>0,\sigma>0$

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 $g_{\mu_1}(\mu_{1i}) = m_{\mu_1}(\mathbf{z}_i)$, $g_{\mu_2}(\mu_{2i}) = m_{\mu_2}(\mathbf{z}_i)$, $g_{\sigma_2}(\sigma_{2i}) = m_{\sigma_2}(\mathbf{z}_i)$, $g_{\mu_3}(\mu_{3i}) = m_{\mu_3}(\mathbf{z}_i)$, $g_{\sigma_3}(\sigma_{3i}) = m_{\sigma_3}(\mathbf{z}_i)$, $g_{\theta_{12}}(\theta_{12i}) = m_{\theta_{12}}(\mathbf{z}_i)$, $g_{\theta_{13}}(\theta_{13i}) = m_{\theta_{13}}(\mathbf{z}_i)$, 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, although 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 $\boldsymbol{\beta}^{\mathsf{T}}\mathbf{D}_{\boldsymbol{\lambda}}\boldsymbol{\beta}$, where $\mathbf{D}_{\boldsymbol{\lambda}} = \mathrm{diag}(0, \lambda_1\mathbf{D}_1, \ldots, \lambda_K\mathbf{D}_K)$. 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 utilized to model, among others, linear and non linear effects; these are relevant to our case studies.

2.3.1. Parametric effects

These effects usually relate to binary and categorical variables and are represented by setting $s_k(\mathbf{z}_{ki}) = \mathbf{z}_{ki}^{\mathsf{T}}\boldsymbol{\beta}_k$. 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. Non parametric effects

Penalized regression splines are a popular and computationally efficient way for representing unknown non linear effects of continuous covariates. The main requirement is a global smoothness assumption on differentiability. This method makes it possible to avoid arbitrary modeling decisions, such as choosing the degree of a polynomial or specifying cut-points, which could induce misspecification. This approach has been popularized by Eilers and 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 $\mathbf{D}_k = \int \mathbf{d}_k(z_k)\mathbf{d}_k(z_k)^{\mathsf{T}} dz_k$, 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., penalized 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 penalized 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 penalized regression smoothing framework described in this section allows for several other specifications as well as a vast variety of penalized 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), 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 $\left\{(s_i, y_{2i}, y_{3i}, \mathbf{z}_i)\right\}_{i=1}^n$, because of the highly flexible regression structures allowed for by the proposed modeling 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 $(\boldsymbol{\beta}_{\mu_1}^{\mathsf{T}}, \dots, \boldsymbol{\beta}_{\theta_{12}}^{\mathsf{T}}, \boldsymbol{\beta}_{\theta_{13}}^{\mathsf{T}})^{\mathsf{T}}$, which is made up of the coefficient vectors related to $m_{\mu_1}(\mathbf{z}_i), \dots, m_{\theta_{12}}(\mathbf{z}_i), m_{\theta_{13}}(\mathbf{z}_i)$ and whose specific set up and dimension depend on the model specification, and $\mathbf{S}_{\boldsymbol{\lambda}} = \mathrm{diag}(\mathbf{D}_{\boldsymbol{\lambda},\mu_1}, \dots, \mathbf{D}_{\boldsymbol{\lambda},\theta_{12}}, \mathbf{D}_{\boldsymbol{\lambda},\theta_{13}})$, which contains the overall smoothing parameter vector $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_{\mu_1}^{\mathsf{T}}, \dots, \boldsymbol{\lambda}_{\theta_{12}}^{\mathsf{T}}, \boldsymbol{\lambda}_{\theta_{13}}^{\mathsf{T}})^{\mathsf{T}}$.

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 optimization 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 and Radice (2020) (see Appendix A 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 penalized models implicitly assumes that overly complex models are less likely than simpler or smoother ones which 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 non linear 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 summarized 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), the Bayesian information criterion (BIC, Schwarz, 1978), and (randomized) normalized quantile residuals (Dunn and 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 penalized 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, randomized normalized 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 $\left[\widehat{F}_{3|1}(y_{3i_3} - 1|y_{1i_3} = 1), \widehat{F}_{3|1}(y_{3i_3}|y_{1i_3} = 1)\right]$. Randomization allows one to view the discrete distributions as if they 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 is not typically affected by this choice.

Studying the asymptotic properties of the proposed estimator is beyond the scope of this article. 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 $\hat{\boldsymbol{\delta}} \xrightarrow{P} \boldsymbol{\delta}^0$ *and* $\|\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}^0\| = O_P(1/\sqrt{n})$ 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 discusses the results of a series of simulation studies. In summary, the proposed approach works well under correct model specification, yields less precise estimates in the absence of an exclusion restriction, produces biased estimates when choosing incorrectly the marginal distributions, and it is not too sensitive to the choice of copulae as long as the magnitudes of the dependencies are captured. 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 functions), 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 non linear 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 Eq. (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 based on the regressors and 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 realizations 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 realizations. 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.

Since our approach links all the parameters of the assumed multivariate distribution to covariate effects, treatment effects on several aspects of such distribution (e.g., variance, skewness, kurtosis) can be obtained. Although in the three case studies considered in this article, treatment effects beyond the mean were not deemed relevant, they can easily be computed by modifying Eq. (8) accordingly.

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

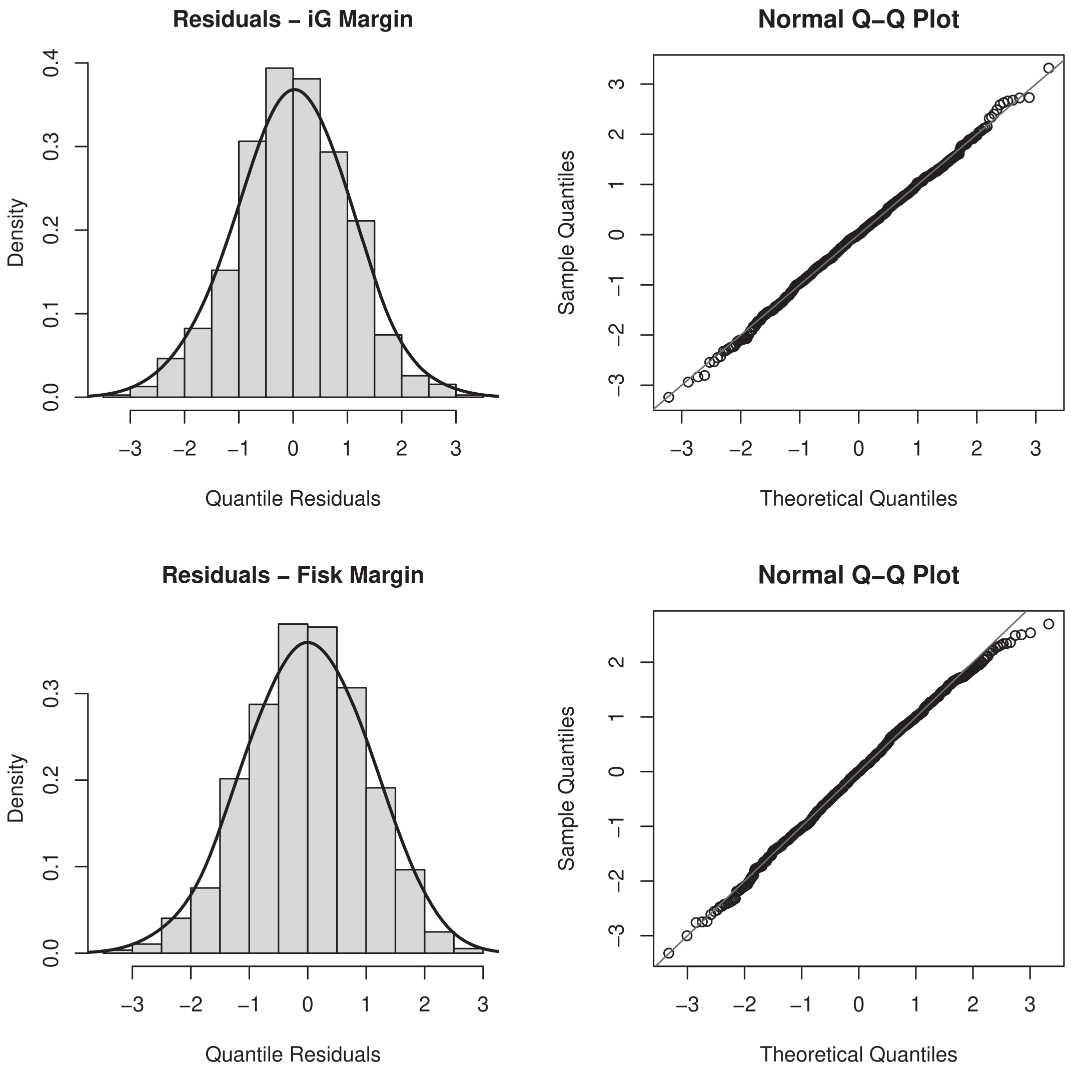
The majority of privately-insured Americans receive coverage through their employers as part of their total compensation. In turn, the theory of "compensating differentials" holds that an employee who receives health insurance from his employer should, all else equal, receive a lower wage than an employee 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 and Madrian, 1999).

Olsen (2002) finds evidence of compensating differentials using instrumental variable 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%. Our study attempts to mimic Olsen's approach by applying the proposed framework to MEPS data.

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 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 modeling decisions involve the distributional 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 Fig. 1 for the plots of the chosen model). Using such a data-driven approach, the treatment variable, an indicator for whether the wife holds insurance through her employer, follows a probit specification. However, owing to the binary nature of such variable, using logit and cloglog links produced virtually identical results for the ATE of health benefits on wage earnings, the measure of interest. For the wife's wage income, the choices are the inverse Gaussian distribution among wives who do not hold insurance through their employers, and Fisk for those who do. Using marginal distributions that do not yield a good fit (e.g., Weibull, Gamma, log-normal; for an example, see Figure 17 in Appendix C), the ATE was found to be about 40% lower or higher than that from the chosen model, hence highlighting the importance of finding marginal distributions that fit the data well. 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 Gaussian for those who do. Different choices produced similar estimated effects but with larger intervals, hence suggesting that employing the copulae that are most supported by the data leads to efficiency gains. Note that the set of alternative copulae should be selected so that the magnitudes of the dependencies present in the data can still be captured, otherwise parameter estimates will be adversely affected. For instance, using AMH and FGM copulae, the ATE was found to be about 80% lower than that from the chosen model.

Before analyzing the ATE results, the next three paragraphs comment on the estimated model parameters which are reported in Appendix C. 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



Residuals – iG Margin

Normal Q–Q Plot

Residuals – Fisk Margin

Normal Q–Q Plot

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

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

Regarding the estimated smooth functions, reported in Fig. 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

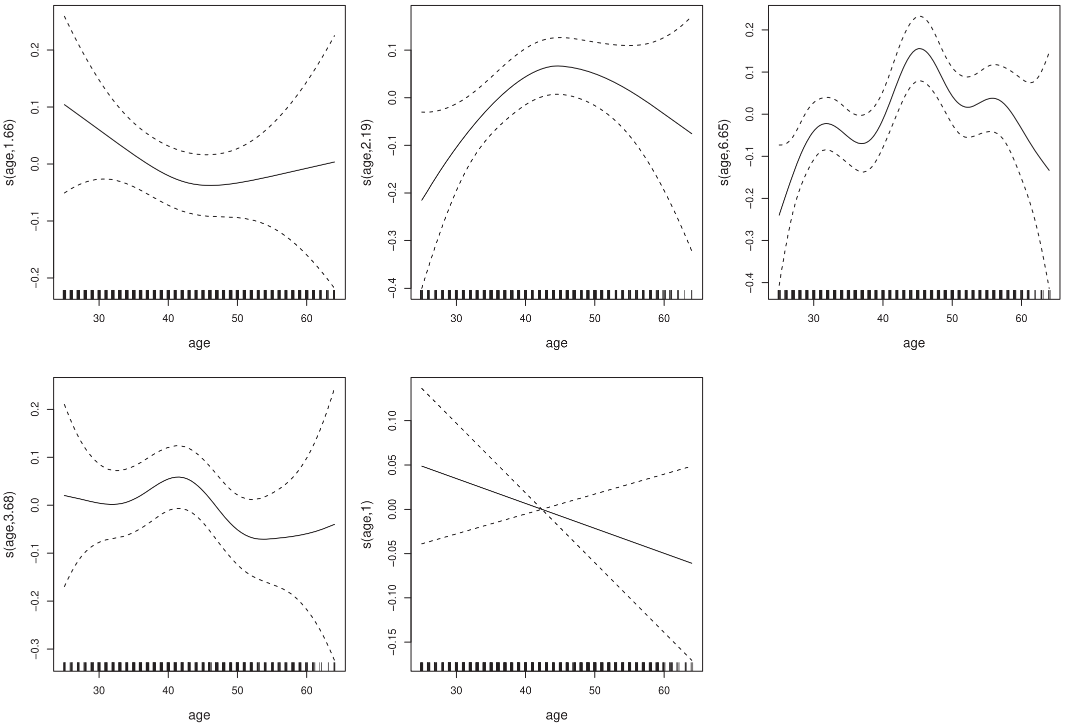


Figure 2. Estimated smooth effects of age for the probit equation, for the μ parameters of the inverse Gaussian and Fisk, and for the σ parameters of the same distributions. 95% point-wise intervals are also reported. The jittered rug plot, at the bottom of each graph, shows the covariate values. The 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 behavior displayed in the last plot because of such constraints.

Table 4. Average treatment effects for case study 1.

	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, −0.13)

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 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 ATE. 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 is associated with a 10% 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 (Terza et al., 2008). The ATE calculation for this approach is based on Eq. (13) of these authors' paper; the result suggests that insurance lowers wages by about 11%. 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%, more that 150% 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 recognized.

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 have insurance, but college education exerts a more ambiguous influence among wives who do not 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 shows sample means for people with ages 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% of insured subjects report having a doctor visit, compared to only 30% of uninsured subjects. Those numbers differ statistically according to a standard t-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. Since the responses are binary, using different combinations of link function specifications for the marginals did not affect the results. 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. As for alternative copulae, findings and considerations were similar to those of the previous case study. As an example, using AMH copulae, the ATE was found to be 24% lower than that from the chosen model.

Results for the copula-based switching regression approach are reported in Appendix C. 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

Table 5. Average treatment effects for case study 2.

	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)

doctor visit. Health economists label that pattern "favorable selection" which means that insured subjects tend to be healthier and/or more risk averse.

Table 5 presents ATEs 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% points, a sizable effect relative to the sample mean of 55% 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% 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 proposal confirms 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 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 chosen 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.

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 Figure 18, Appendix C). 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. Using Poisson marginals (clearly not appropriate here as shown in Figure 20, Appendix C), the ATE was found to be 58% higher than that from the chosen model. As for alternative copulae, findings and considerations were similar to those of the previous case studies.

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.

Table 6. Average treatment effects for case study 3.

	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)

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.

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 ATE. 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 and flexible 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 modeling 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 article. 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 modeling framework by allowing for survival margins as well as by exploring alternative copula selection methods along the lines of Cai (2014), for instance. We will also look into developing a testing approach for weak instruments.

Acknowledgments

We would like to thank the Editor and reviewer for their insightful comments that have helped to improve and clarify several aspects of the paper.

Funding

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*, Budapest: Academiai Kiado.

Basch, C. E. (2011). Healthier students are better learners: A missing link in school reforms to close the achievement gap. *The Journal of School Health* 81(10):593–598. doi:10.1111/j.1746-1561.2011.00632.x 21923870

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. doi:10.1198/jbes.2010.08083

Bollen, K. A., Pearl, J. (2013). Eight myths about causality and structural equation models. In: Morgan S., ed., *Handbook of Causal Analysis for Social Research*, Dordrecht: Springer, pp. 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. doi:10.1080/01621459.2013.873366

Chen, H., Fan, Y., Wu, J. (2014). A flexible parametric approach for estimating switching regime models and treatment effect parameters. *Journal of Econometrics* 181(2):77–91. doi:10.1016/j.jeconom.2013.06.006

Claeskens, G., Krivobokova, T., Opsomer, J. D. (2009). Asymptotic properties of penalized spline estimators. *Biometrika* 96(3):529–544. doi:10.1093/biomet/asp035

Collett, D. (2002). *Modelling Binary Data*, 2nd ed. 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. doi:10.1016/j.labeco.2016.06.004

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

Currie, J., Madrian, B. (1999). Health, health insurance and the labor market. In: Ashenfelter, O., Card, D., eds., *Handbook of Labor Economics*, Vol. 3, Amsterdam: Elsevier, pp. 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'Haultføeuille, 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. doi:10.1016/j.jeconom.2013.01.005

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. (2021). Evaluating the impact of large-scale agricultural investments on household food security using an endogenous switching regression model. *Land* 10(3):323. doi:10.3390/land10030323

French, E., Taber, C. (2011). Identification of models of the labor market. In Orley Ashenfelter, David Card, eds., *Handbook of Labor Economics,* Vol. 4, Part A, Elsevier, pp. 537–617.

Hasebe, T. (2013). Copula-based maximum-likelihood estimation of sample-selection models. *The Stata Journal: Promoting Communications on Statistics and Stata* 13(3):547–573. doi:10.1177/1536867X1301300307

Hasebe, T. (2020). Endogenous switching regression model and treatment effects of count-data outcome. *The Stata Journal: Promoting Communications on Statistics and Stata* 20(3):627–646. doi:10.1177/1536867X20953573

Heckman, J. J., Honore, B. E. (1990). The empirical content of the roy model. *Econometrica* 58(5):1121. doi:10.2307/2938303

Heckman, J., Tobias, J. L., Vytlacil, E. (2003). Simple estimators for treatment parameters in a latent-variable framework. *Review of Economics and Statistics* 85(3):748–755. doi:10.1162/003465303322369867

Holbert, T., Wu, L., Stark, M. (2002). *School attendance initiative: The first 3 years*, Report prepared for US Dept of Justice: Office of Development and Testing. National Institute of Law Enforcement and Criminal Justice. https://www.ojp.gov/ncjrs/virtual-library/abstracts/school-attendance-initiative-first-3-years-199899-200001.

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

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 asymptotic results on generalized penalized spline smoothing. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 71(2):487–503. doi:10.1111/j.1467-9868.2008.00691.x

Kim, S. H. (2021). Changes in social trust: Evidence from east german migrants. *Social Indicators Research* 155(3):959–981. doi:10.1007/s11205-021-02611-z

Lokshin, M., Sajaia, Z. (2004). Maximum likelihood estimation of endogenous switching regression models. *The Stata Journal: Promoting Communications on Statistics and Stata* 4(3):282–289. doi:10.1177/1536867X0400400306

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. (2023). *GJRM: Generalized Joint Regression Modeling.* R package version 0.2-6.4. 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. doi:10.1016/j.jhealeco.2017.06.005 28935158

Nelsen, R. (2006). *An Introduction to Copulae*. 2nd ed. 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. doi:10.1080/02664760903093591

Olsen, C. (2002). Do workers accept lower wages in exchange for health benefits?. *Journal of Labor Economics*, 20(S2):91–114.

Pearl, J. (2015). Trygve haavelmo and the emergence of causal calculus. *Econometric Theory* 31(1):152–179. doi:10.1017/S0266466614000231

R Core Team. (2023). *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Rigby, R. A., Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society Series C: Applied Statistics* 54(3):507–554. doi:10.1111/j.1467-9876.2005.00510.x

Rosen, S. (1986). The theory of equalizing differences. In: Ashenfelter, O., Layard, R., eds., *Handbook of Labor Economics*, Vol. 1 Amsterdam: North-Holland, pp. 641–692.

Roy, A. D. (1951). Some thoughts on the distribution OF earnings. *Oxford Economic Papers* 3(2):135–146. doi:10.1093/oxfordjournals.oep.a041827

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):47–57.

Terza, J. V., Basu, A., Rathouz, P. J. (2008). Two-stage residual inclusion estimation: addressing endogeneity in health econometric modeling. *Journal of Health Economics* 27(3):531–543. doi:10.1016/j.jhealeco.2007.09.009 18192044

Toomet, O., Henningsen, A. (2020). *Sample Selection Models in R: Package sampleSelection R package version 1.2-12*, 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*. 2nd ed. London: Chapman & Hall/CRC.