A Flexible Copula Regression Model with Bernoulli and Tweedie Margins for Estimating the Effect of Spending on Mental Health

Giampiero Marra
Department of Statistical Science
University College London

Matteo Fasiolo
School of Mathematics
University of Bristol

Rosalba Radice
Bayes Business School
City, University of London

Rainer Winkelmann
Department of Economics
University of Zurich

2022-05-10

Abstract

Previous evidence shows that better insurance coverage increases medical expenditure. However, formal studies on the effect of spending on health outcomes, and especially mental health, are lacking. To fill this gap, we reanalyze data from the Rand Health Insurance Experiment and estimate a joint non-linear model of spending and mental health. We address the endogeneity of spending in a flexible copula regression model with Bernoulli and Tweedie margins and discuss its implementation in the freely available `GJRM R` package. Results confirm the importance of accounting for endogeneity: in the joint model, a $1000 spending in mental care is estimated to reduce the probability of low mental health by 1.3 percentage points, but this effect is not statistically significant. Ignoring endogeneity leads to a spurious (upwardly biased) estimate.

Key Words: Binary response; Co-payment; Copula; Health expenditures; Penalized regression spline; Rand experiment; Simultaneous estimation; Tweedie distribution.

1 Introduction

It is quite common in empirical health economics to specify models where two or more outcome variables are potentially associated conditional on a set of covariates. Modeling such association matters for efficiency, but also for addressing endogeneity in a recursive system of equations. The choice of model is dictated by the nature of the dependent variables. For example, for two jointly determined binary responses, the bivariate probit is frequently used (see Waters (1999), Farbmacher et al. (2017), Humphreys et al. (2014), among others). Often, however, there will be a mix of outcomes. As typical example, consider a joint model for insurance status and health care utilization, as in Deb & Trivedi (2006) and Marra et al. (2020). In such cases, one can use copulae to introduce dependence between outcomes with known arbitrary margins. In fact, the copula approach allows the marginal distributions to be chosen to best fit the data.

In this paper, we are interested in expenditures for medical services (per person and quarter or year), and their effect on a health outcome (a binary indicator), based on a joint model with associated outcomes. Modeling medical expenditure data is a challenge because they display a substantial fraction of zeros, often more than 50%, in combination with a continuous distribution of positive amounts that is highly skewed. In single equation models, which neglect the presence of associated outcomes, estimation of regression parameters does not rely on a correctly specified distribution, and determinants of expenditure have often been estimated using the Gamma pseudo maximum likelihood estimator (Manning et al., 1987), linear regression for logarithmic non-zero expenditures (Manning & Mullahy, 2001), or a variety of two-part models (Mullahy, 1998). For the copula approach, however, it is important to find a suitable marginal distribution for expenditures. Kurz (2017) recently argued that a compound Gamma distribution can provide a good characterization of health care expenditures. Specifically, assume that expenditures Y is a sum of N independent and identically distributed Gamma random variables, where N is distributed as Poisson(ω). Then the probability of a zero expenditure is given by $\exp(-\omega)$ and the distribution of positive amounts is skewed to the right. The compound gamma distribution is a special case of the more general Tweedie family of distributions. A plot based on quantile residuals confirms the suitability of the Tweedie for modeling the marginal distribution of health care expenditures in our application.

Our main technical contribution is then to build a flexible copula regression model for the joint determination of a Tweedie variable and a Bernoulli response (in our application an indicator of

good health) and describe the steps for an efficient estimation algorithm. The proposed model is flexible in the sense that it is possible to: choose several link functions for the Bernoulli margin, incorporate (linear and non-linear) covariate effects into any distributional parameter, and allow for the exploration of a wide set of copulae. Moreover, it can be made fully interdependent, recursive, or "seemingly unrelated", by imposing zero-constraints on corresponding coefficients. A similar model has been proposed by Marra et al. (2020) who focused on binary and count margins (see, also, the references therein for alternative simpler versions, with different types of margins). The model is implemented in the `R` package `GJRM` (Marra & Radice, 2022), written for the programming language `R` (R Core Team, 2022). To the best of our knowledge, this is the first freely available implementation of a flexible copula regression model involving the Tweedie distribution.

In terms of substantive application, we use the proposed framework for estimating the causal effect of expenditure on health. We revisit data from the Rand Health Insurance Experiment (RHIE, e.g., Manning et al., 1987; Aron-Dine et al., 2013). The experiment generated exogenous variation in expenditure, by providing more or less insurance coverage to participant households. The experiment lasted for a period of up to 5 years, and health evaluations were conducted both before and after. Unfortunately, the public use files only contain information on mental rather than physical health, and that is one reason why our analysis is focused on the effect of mental health expenditures on mental health.

The other reason is that empirical studies on mental health expenditures are relatively scarce, despite their policy relevance. The OECD assesses that more spending on mental health (often outpatient psychotherapy) could massively increase productivity and well-being (OECD, 2014). The United Nations include improved mental care provision among the sustainable development goals (United Nations, 2015). Nevertheless, among the stream of papers that came out of the RHIE, only one dealt explicitly with mental health (Manning et al., 1986). A possible explanation is that mental spending in the RHIE accounted only for around 4% of overall spending. Regarding overall expenditures, earlier analyzes of the data suggested a strong "first-stage" effect (i.e., more generous insurance coverage increased expenditures) but no systematic "reduced-form" effect (of more generous insurance cover on actual health, or mental health, as in Manning et al. (1989)). A formal analysis of the direct effect of increased mental spending on mental health was, to the best of our knowledge, never conducted.

The bivariate model we implement is recursive: it postulates a direct effect of the Tweedie dis-

tributed health expenditure variable on the Bernoulli outcome "low mental health", but not vice-versa. Joint estimation using the copula approach allows for common unobserved components. Ignoring this would bias the effect size. Direct reverse causation is not an issue here because of the ordering of events: spending precedes the health measurement. Hence, the recursive model is appropriate. To better achieve empirical identification, we instrument spending using RHIE's random assignment of individuals to insurance plans with different levels of cost-sharing. Our approach is superior to ad-hoc endogeneity corrections, such as plug-in or control function approaches, as they are not designed for non-normal and non-linear reduced forms as evidently required for the mixed discrete-continuous variable "mental spending" (e.g., Rivers & Vuong, 1988; Wooldridge, 2010). Using the copula approach with appropriate margins for the binary mental health indicator and the semi-continuous mental spending variable, we address the endogeneity of spending in a model-consistent way, based on a specification of the joint probability function that could have generated the observed data.

Our main findings are as follows. The Tweedie margin fits the expenditure data well, as evidenced by approximately normally distributed quantile residuals. Among the eleven copulae and three binary link functions considered, the Gaussian copula with probit link has the smallest Akaike information criterion (AIC) value. The instrument for mental spending (made up of several categories) is highly significant (p-value of 0.00027 using the Wald test). Substantively, we find that the probability of low mental health is strongly predicted by the initial mental health score. The probability is higher for women and for the less educated, although the latter effect is not statistically significant (p-value of 0.14). The estimated Gaussian copula parameter indicates positive dependence. Ignoring this leads to a spurious estimate in the univariate model, where a $1000 increase in spending predicts an increase of 3.3 percentage points in the probability of low mental health. Once the endogeneity of spending is accounted for, the average partial effect switches sign. The point estimate of -1.3 percentage points is of modest size and is not statistically significant.

The paper is organized as follows. Section 2 introduces the Tweedie-Bernoulli copula model by discussing its components. Section 3 provides some details on parameter estimation and the related implementation in `R`. Section 4 presents the application whereas Section 5 concludes the paper.

2 Copula model

The copula regression model introduced in this paper aims at modeling in a flexible and versatile way the joint distribution of a Bernoulli outcome variable and a semi-continuous (endogenous) variable. In the application, $Y_1 \in \{0, 1\}$ is the binary outcome "low mental health" and $Y_2 \in \mathbb{R}_0^+$ is mental spending. Assume that

$$F_{12}(y_1, y_2|\boldsymbol{\vartheta}) = C\left(F_1(y_1|\pi), F_2(y_2|\mu, \sigma, \nu); \theta\right), \tag{1}$$

where $\boldsymbol{\vartheta} = (\pi, \mu, \sigma, \nu, \theta)'$, $F_1(y_1|\pi)$ and $F_2(y_2|\mu, \sigma, \nu)$ represent the marginal cumulative distribution functions (cdfs) of Y_1 and Y_2 taking values in $(0, 1)$, $C : (0, 1)^2 \to (0, 1)$ is a two-place copula function whose specification does not depend on the marginals, and θ is a copula dependence parameter that quantifies the association between the two random variables (see, e.g., Nelsen, 2006, for further details). Variable Y_1 is modeled via a Bernoulli distribution with parameter $\pi \in [0, 1]$ (representing the probability that the outcome is equal to 1), and Y_2 using a Tweedie distribution with parameters μ, σ and ν (see Section 2.1). Note that π, μ, σ, ν and θ can be specified as functions of covariate effects as detailed in Section 2.2. Hence, the marginal model for Y_1 can be regarded as a generalized additive model and that for Y_2 as a generalized additive model for location scale and shape (e.g., Rigby & Stasinopoulos, 2005; Wood, 2017). Function C can be specified as in Table 1 of Marra et al. (2020) which reports several copulae.

As opposed to classical copula regression settings, the variable Y_2 appears as an explanatory variable in π, hence giving the model a recursive structure which in turn implies that Y_2 is endogenous with respect to Y_1 if the dependence between the two marginals (captured by θ) is statistically significant; see Han & Vytlacil (2017), Marra et al. (2020) and references therein for some works which have adopted the same logic for copula regression models. Note that, for the model adopted in this paper, Sklar (1973)'s result can only guarantee that the copula is unique over the range of the outcomes. However, for applied purposes, in a regression contest, this is not problematic as noted by several authors including Joe (2014), Nikoloulopoulos & Karlis (2010) and Trivedi & Zimmer (2017).

The joint density $f_{12}\left(y_1, y_2\right)$, required for calculating the model's log-likelihood, is built by con-

sidering the four possible combinations of values that $(Y_1, Y_2)'$ can take. That is,

$$f_{12}\left(y_1, y_2\right) = \begin{cases} C\left(F_1(0|\pi), F_2(0|\mu,\sigma,\nu); \theta\right) & \text{if} \quad y_1 = 0 \quad \text{and} \quad y_2 = 0 \\ F_2(0|\mu,\sigma,\nu) - C\left(F_1(0|\pi), F_2(0|\mu,\sigma,\nu); \theta\right) & \text{if} \quad y_1 = 1 \quad \text{and} \quad y_2 = 0 \\ f_2\left(y_2|\mu,\sigma,\nu\right) \frac{\partial C(F_1(0|\pi), F_2(y_2|\mu,\sigma,\nu); \theta)}{\partial F_2(y_2|\mu,\sigma,\nu)} & \text{if} \quad y_1 = 0 \quad \text{and} \quad y_2 > 0 \\ f_2\left(y_2|\mu,\sigma,\nu\right) \left\{1 - \frac{\partial C(F_1(0|\pi), F_2(y_2|\mu,\sigma,\nu); \theta)}{\partial F_2(y_2|\mu,\sigma,\nu)}\right\} & \text{if} \quad y_1 = 1 \quad \text{and} \quad y_2 > 0 \end{cases}, \tag{2}$$

where $f_2(y_2|\mu,\sigma,\nu) = \partial F_2(y_2|\mu,\sigma,\nu)/\partial y_2$ denotes the density function of the Tweedie distribution. The first two lines of (2) show the probabilities associated with their respective events, whereas the last two lines show the densities obtained by differencing the first two lines of the equation with respect to y_2 when $y_2 > 0$. For notational convenience, we dropped observation index i from the formulae above. However, it should be clear from the context of the paper that a set of n observations is assumed to be available for practical modeling.

2.1 Tweedie distribution

The Tweedie distribution is a linear exponential dispersion model (Jørgensen, 1987) with a power mean-variance relationship, that is

$$\text{var}(Y) = \sigma\mu^{\nu},$$

where $\mu = \mathbb{E}(Y)$, $\sigma > 0$ is a scale parameter and $\nu \in \mathbb{R}$ controls the shape of the relationship. Widely used distributions such as the Gaussian, Poisson and Gamma are nested by the Tweedie family and can be recovered by setting ν to the relevant value (e.g., $\nu = 0$ for the Gaussian). In this work, we are interested in the interval $\nu \in (1, 2)$, for which a Tweedie-distributed random variable can be represented as the sum of N independent Gamma-distributed random variables, where N is Poisson distributed. The resulting density is supported on the non-negative real line and has a positive mass at $y = 0$. Prior applications of the Tweedie compound Poisson-Gamma distribution are mainly known from the actuarial sciences, where the distribution is used to model insurance claim payments data (see, e.g., Smyth & Jørgensen, 2002). Kurz (2017) provides an application to total medical expenditures.

Fitting reliably the proposed copula regression model involving a Tweedie margin, using the method that will be mentioned in Section 3, requires the ability to compute the Tweedie proba-

bility density function and cdf as well as their first and second order derivatives with respect to μ, σ and ν. As we detail in the rest of this section, such quantities are not trivial to compute.

The density of the Tweedie is

$$f(y|\mu,\sigma,\nu) = a(y,\sigma,\nu)\exp\left[\frac{1}{\sigma}\{y\theta - \kappa(\theta)\}\right],$$

where

$$\theta = \frac{\mu^{1-\nu}}{1-\nu} \text{ for } \nu \neq 1 \quad \text{and} \quad \theta = \log\mu \text{ for } \nu = 1,$$

and

$$\kappa(\theta) = \frac{\mu^{2-\nu}}{2-\nu} \text{ for } \nu \neq 2 \quad \text{and} \quad \kappa(\theta) = \log\mu \text{ for } \nu = 2.$$

As explained in Dunn & Smyth (2005), evaluating the Tweedie density requires approximating the factor $a(y,\sigma,\nu)$, which does not have a closed-form expression, using specifically designed numerical methods. Wood et al. (2017) provide methods for computing the first and second order derivatives of the log-density, while avoiding numerical problems. In the next section, we explain how to compute the Tweedie cdf and its derivatives.

2.1.1 Tweedie cdf and its derivatives

For $1 \leq \nu \leq 2$, a Tweedie variable Y can be written as the sum of N independent Gamma-distributed random variables $W_1, \ldots, W_N$ with shape $-\alpha$ and scale γ, while N follows a Poisson distribution with rate ω. In terms of the Tweedie parameters, we have

$$\omega = \frac{\mu^{2-\nu}}{\sigma(2-\nu)}, \quad \alpha = (2-\nu)/(1-\nu), \quad \gamma = \sigma(\nu-1)\mu^{\nu-1}.$$

Hence, the cdf of Y is

$$F(y) = P\left(\sum_{i=1}^{N} w_i < y\right) = \sum_{k=1}^{\infty} P_G\left(\sum_{i=1}^{k} w_i < y\right) P_P(N=k), \tag{3}$$

where the second equality holds due to the Law of Total Probability, P_G is equivalent to the cdf (F_G^k) of a Gamma distribution with parameters $-k\alpha$ and γ, while P_P is the probability mass function (f_P) of a Poisson with rate ω.

We approximate the infinite sum in (3) by

$$F(y) \approx \hat{F}(y) = \sum_{k=k_{min}}^{k_{max}} F_G^k(y) f_P(k),$$

for some $k_{\max} \geq k_{\min}$ whose values are chosen as follows. Given that f_P is maximal at $k = \lfloor \omega \rfloor$ and then monotonically decreases as k moves away from the mode, a reasonable approach is to choose $k_{\min}$ and $k_{\max}$ such that $f_P(k) < \epsilon f_P(\lfloor \omega \rfloor)$ for $k < k_{\min}$ or $k > k_{\max}$, for some small ϵ. If we do so, it is clear that a very pessimistic upper bound on the approximation error is

$$|F(y) - \hat{F}(y)| < P_P(N < k_{\min} \text{ or } N > k_{\max}) = 1 - P_P(k_{\min} < N < k_{\max}),$$

which is easy to compute. Of course, $k_{\min}$ and $k_{\max}$ are not known in advance, but we can initialize k to $\lfloor \omega \rfloor$ and then increase k until $f_P(k) < \epsilon f_P(\lfloor \omega \rfloor)$, which leads to $k_{\max}$. Then k is set to the Poisson mode and decreased until $k_{\min}$ is found or $k = 0$.

First and second derivatives of the (approximate) Tweedie cdf, $\hat{F}(y)$, w.r.t. $\boldsymbol{\theta} = \{\mu, \sigma, \nu\}$ are obtained by firstly calculating the derivatives w.r.t. $\boldsymbol{\psi} = \{\alpha, \gamma, \omega\}$ and then using the Jacobian of the transformation to convert them to the $\boldsymbol{\theta}$-based parametrization. The gradient of $\hat{F}(y)$ and the diagonal entries of the Hessian w.r.t $\boldsymbol{\psi}$ are

$$\frac{\partial^j \hat{F}}{\partial \alpha^j} \approx \sum_{k=k_{min}}^{k_{\max}} \frac{\partial^j F_G^k}{\partial \alpha^j} f_P, \quad \frac{\partial^j \hat{F}}{\partial \gamma^j} \approx \sum_{k=k_{\min}}^{k_{\max}} \frac{\partial^j F_G^k}{\partial \gamma^j} f_P, \quad \frac{\partial^j \hat{F}}{\partial \omega^j} \approx \sum_{k=k_{\min}}^{k_{\max}} F_G^k \frac{\partial^j f_P}{\partial \omega^j},$$

for $j = 1$ or 2, while the non-diagonal Hessian elements are

$$\frac{\partial^2 \hat{F}}{\partial \alpha \partial \gamma} \approx \sum_{k=k_{\min}}^{k_{\max}} \frac{\partial^2 F_G^k}{\partial \alpha \partial \gamma} f_P, \quad \frac{\partial^2 \hat{F}}{\partial \alpha \partial \omega} \approx \sum_{k=k_{\min}}^{k_{\max}} \frac{\partial F_G^k}{\partial \alpha} \frac{\partial f_P}{\partial \omega}, \quad \frac{\partial^2 \hat{F}}{\partial \gamma \partial \omega} \approx \sum_{k=k_{\min}}^{k_{\max}} \frac{\partial F_G^k}{\partial \gamma} \frac{\partial f_P}{\partial \omega}.$$

The derivatives w.r.t. γ and ν are given by

$$\frac{\partial F_G^k}{\partial \gamma} = -\frac{1}{\Gamma(-k\alpha)} \frac{y}{\gamma^2} \left(\frac{y}{\gamma}\right)^{-k\alpha-1} e^{-\frac{y}{\gamma}}, \quad \frac{\partial^2 F_G^k}{\partial \gamma \partial \alpha} = k \frac{\partial F_G^k}{\partial \gamma} \left\{ \frac{\Gamma^{(1)}(-k\alpha)}{\Gamma(-k\alpha)} - \log \frac{y}{\gamma} \right\},$$

$$\frac{\partial^2 F_G^k}{\partial \gamma^2} = \frac{y \exp\left\{\frac{-y}{\gamma} - (ka+1)\log\left(\frac{y}{\gamma}\right) - \log\Gamma(-k\alpha)\right\}\left\{\gamma(1-ka) - y\right\}}{\gamma^4},$$

$$\frac{\partial f_P}{\partial \omega} = f_P(k)\left(\frac{k}{\omega} - 1\right), \quad \frac{\partial^2 f_P}{\partial \omega^2} = \frac{\partial f_P}{\partial \omega}\left(\frac{k}{\omega} - 1\right) - \frac{k}{\omega^2} f_P(k),$$

where Γ is the gamma function and $\Gamma^{(1)}(-k\alpha)/\Gamma(-k\alpha)$ is the digamma function. Derivatives w.r.t. α require computing the derivative of the lower incomplete gamma function w.r.t. its first argument. To our best knowledge, no numerical routine is currently available to approximate this quantity efficiently and stably, hence we use finite differences to approximate (mixed) derivatives of $\hat{F}_G$ w.r.t. α. The gradient of $\hat{F}$ w.r.t. $\boldsymbol{\theta}$ is obtained by $\nabla_{\boldsymbol{\theta}} \hat{F} = \mathbf{J} \nabla_{\psi} \hat{F}$, where

$$\boldsymbol{J} = \begin{bmatrix} \frac{\partial \omega}{\partial \mu} & \frac{\partial \alpha}{\partial \mu} & \frac{\partial \gamma}{\partial \mu} \\ \frac{\partial \omega}{\partial \sigma} & \frac{\partial \alpha}{\partial \sigma} & \frac{\partial \gamma}{\partial \sigma} \\ \frac{\partial \omega}{\partial \nu} & \frac{\partial \alpha}{\partial \nu} & \frac{\partial \gamma}{\partial \nu} \end{bmatrix} = \begin{bmatrix} \frac{2-\nu}{\mu}\omega & 0 & \frac{\nu-1}{\mu}\gamma \\ -\frac{\omega}{\sigma} & 0 & \frac{\gamma}{\sigma} \\ \frac{\mu^{2-\nu}\{1+(\nu-2)\log\mu\}}{(\nu-2)^2\sigma} & \frac{1}{(\nu-1)^2} & \gamma(\log\mu + \frac{1}{\nu-1}) \end{bmatrix}$$

is the Jacobian of the transformation. The Hessian is then obtained by

$$\nabla'_{\boldsymbol{\theta}} \nabla_{\boldsymbol{\theta}} \hat{F} = \mathbf{J} \nabla'_{\boldsymbol{\psi}} \nabla_{\boldsymbol{\psi}} \hat{F} \mathbf{J}^T + \left[\frac{\partial \boldsymbol{J}}{\partial \mu} \nabla_{\boldsymbol{\psi}} \hat{F}, \frac{\partial \boldsymbol{J}}{\partial \sigma} \nabla_{\boldsymbol{\psi}} \hat{F}, \frac{\partial \boldsymbol{J}}{\partial \nu} \nabla_{\boldsymbol{\psi}} \hat{F} \right],$$

where

$$\frac{\partial \boldsymbol{J}}{\partial \mu} = \begin{bmatrix} \mathbf{J}_{11}\left(\frac{1-\nu}{\mu}\right) & 0 & \mathbf{J}_{13}\left(\frac{\nu-2}{\mu}\right) \\ -\frac{\mathbf{J}_{11}}{\sigma} & 0 & \frac{\mathbf{J}_{13}}{\sigma} \\ \frac{2-\nu}{\mu}\mathbf{J}_{31} + \frac{1}{\mu}\frac{\mu^{2-\nu}}{(\nu-2)\sigma} & 0 & \frac{\gamma}{\mu} + \mathbf{J}_{13}(\log\mu + \frac{1}{\nu-1}) \end{bmatrix}, \quad \frac{\partial \boldsymbol{J}}{\partial \sigma} = \begin{bmatrix} \frac{2-\nu}{\mu}\mathbf{J}_{21} & 0 & \frac{\nu-1}{\mu}\mathbf{J}_{23} \\ -2\frac{\mathbf{J}_{21}}{\sigma} & 0 & 0 \\ -\frac{\mathbf{J}_{31}}{\sigma} & 0 & \mathbf{J}_{23}(\log\mu + \frac{1}{\nu-1}) \end{bmatrix},$$

$$\frac{\partial \boldsymbol{J}}{\partial \nu} = \begin{bmatrix} \frac{2-\nu}{\mu}\mathbf{J}_{31} - \frac{\omega}{\mu} & 0 & \frac{\nu-1}{\mu}\mathbf{J}_{33} + \frac{\gamma}{\mu} \\ -\frac{\mathbf{J}_{31}}{\sigma} & 0 & \frac{\mathbf{J}_{33}}{\sigma} \\ -\mathbf{J}_{31}\left(\log\mu + \frac{2}{\nu-2}\right) + \frac{\mu^{2-\nu}\log\mu}{(\nu-2)^2\sigma} & -\frac{2}{(\nu-1)^3} & -\gamma\frac{1}{(\nu-1)^2} + \mathbf{J}_{33}(\log\mu + \frac{1}{\nu-1}) \end{bmatrix},$$

with $\mathbf{J}_{ij}$ indicating the element in row i and column j of the Jacobian.

2.2 Additive predictor

Each of the parameters in $\boldsymbol{\vartheta}$ is linked to covariate effects via an additive predictor $\eta \in \mathbb{R}$ and a known monotonic one-to-one transformation function g that maps the parameter space to the real line. For the proposed model, we have $g_\pi(\pi) = \eta_\pi$, $g_\mu(\mu) = \eta_\mu$, $g_\sigma(\sigma) = \eta_\sigma$, $g_\nu(\nu) = \eta_\nu$ and $g_\theta(\theta) = \eta_\theta$, where $g_\pi(\cdot)$ can be specified using a `logit`, `probit` or `cloglog` link function, $g_\mu(\mu) = \log(\mu)$,

$g_\sigma(\sigma) = \log(\sigma)$, $g_\nu(\nu) = \log((\nu - 1.001)/(1.999 - \nu))$, and g_θ depends on the chosen copula (e.g., for Gumbel $g_\theta(\theta) = \log(\theta - 1)$). Dropping for simplicity the subscript denoting the parameter the additive predictor belongs to, η_i can be written as

$$\eta_i = \beta_0 + \sum_{k=1}^{K} s_k(\mathbf{z}_{ki}),\ i = 1, \ldots, n, \tag{4}$$

where $\beta_0 \in \mathbb{R}$ is an overall intercept, $\mathbf{z}_{ki}$ denotes the k^{th} sub-vector of the complete covariate vector $\mathbf{z}_i$ (potentially containing various types of covariates) and the K functions $s_k(\mathbf{z}_{ki})$ represent generic effects which are chosen according to the type of covariate(s) considered. Each $s_k(\mathbf{z}_{ki})$ can be approximated as a linear combination of J_k basis functions $b_{kj_k}(\mathbf{z}_{ki})$ and regression coefficients $\beta_{kj_k} \in \mathbb{R}$, i.e. (e.g., Wood, 2017)

$$s_k(\mathbf{z}_{ki}) = \sum_{j_k=1}^{J_k} \beta_{kj_k} b_{kj_k}(\mathbf{z}_{ki}). \tag{5}$$

This formulation implies that the vector of evaluations $\{s_k(\mathbf{z}_{k1}), \ldots, s_k(\mathbf{z}_{kn})\}'$ can be written as $\mathbf{Z}_k\boldsymbol{\beta}_k$ with $\boldsymbol{\beta}_k = (\beta_{k1}, \ldots, \beta_{kJ_k})'$ and design matrix $Z_k[i, j_k] = b_{kj_k}(\mathbf{z}_{ki})$. This allows equation (4) to be written as

$$\boldsymbol{\eta} = \beta_0\mathbf{1}_n + \mathbf{Z}_1\boldsymbol{\beta}_1 + \ldots + \mathbf{Z}_K\boldsymbol{\beta}_K, \tag{6}$$

where $\mathbf{1}_n$ is an n-dimensional vector made up of ones. Equation (6) can also be written as $\boldsymbol{\eta} = \mathbf{Z}\boldsymbol{\beta}$, where $\mathbf{Z} = (\mathbf{1}_n, \mathbf{Z}_1, \ldots, \mathbf{Z}_K)$ and $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_1', \ldots, \boldsymbol{\beta}_K')'$.

Each $\boldsymbol{\beta}_k$ has an associated quadratic penalty $\lambda_k\boldsymbol{\beta}_k'\mathbf{D}_k\boldsymbol{\beta}_k$ whose role is to enforce specific properties on the k^{th} function, such as smoothness. Matrix $\mathbf{D}_k$ only depends on the choice of basis functions. Smoothing parameter $\lambda_k \in [0, \infty)$ controls the trade-off between fit and smoothness, and plays a crucial role in determining the shape of $\hat{s}_k(\mathbf{z}_{ki})$. The overall penalty can be defined as $\boldsymbol{\beta}'\mathbf{D}_{\boldsymbol{\lambda}}\boldsymbol{\beta}$, where $\mathbf{D}_{\boldsymbol{\lambda}} = \text{diag}(0, \lambda_1\mathbf{D}_1, \ldots, \lambda_K\mathbf{D}_K)$. Note that the first term in $\mathbf{D}_{\boldsymbol{\lambda}}$ is set to 0 since β_0 is not penalized in estimation. However, it could be replaced with $\mathbf{0}$ should some parameter vector(s) in $\boldsymbol{\beta}$ also be unpenalized. Finally, the smooth functions are subject to centering (identifiability) constraints, which impose that $\sum_{i=1}^{n} s_k(\mathbf{z}_{ki}) = 0$ for every k (see Wood (2017) for more details). The above smooth function representation allows one to specify a rich variety of covariate effects (such as linear, nonlinear and spatial Markov random field effects) and we refer the reader to Wood (2017) for full details.

Consider the compact form for the random vectors $\mathbf{Y}_1 = (Y_{11}, \ldots, Y_{1n})'$ and $\mathbf{Y}_2 = (Y_{21}, \ldots, Y_{2n})'$.

Then, by some slight abuse of notation, $\mathbf{Y}_1 \sim \mathcal{D}_1(\boldsymbol{\pi})$ and $\mathbf{Y}_2 \sim \mathcal{D}_2(\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\nu})$, where $\mathcal{D}_1$ and $\mathcal{D}_2$ denote the Bernoulli and Tweedie distributions, respectively, and $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_n)'$, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)'$, $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_n)'$ and $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_n)'$ are modeled through

$$
\begin{aligned}
\boldsymbol{\eta}_\pi &= g_\pi(\boldsymbol{\pi}) = \beta_{10}\mathbf{1}_n + \boldsymbol{\beta}_{\text{end}}\mathbf{Y}_2 + \mathbf{Z}_{11}\boldsymbol{\beta}_{11} + \ldots + \mathbf{Z}_{1K}\boldsymbol{\beta}_{1K}, \\
\boldsymbol{\eta}_\mu &= g_\mu(\boldsymbol{\mu}) = \beta_{20}\mathbf{1}_n + \mathbf{Z}_{21}\boldsymbol{\beta}_{21} + \ldots + \mathbf{Z}_{2K}\boldsymbol{\beta}_{2K}, \\
\boldsymbol{\eta}_\sigma &= g_\sigma(\boldsymbol{\sigma}) = \beta_{30}\mathbf{1}_n + \mathbf{Z}_{31}\boldsymbol{\beta}_{31} + \ldots + \mathbf{Z}_{3K}\boldsymbol{\beta}_{3K}, \\
\boldsymbol{\eta}_\nu &= g_\nu(\boldsymbol{\nu}) = \beta_{40}\mathbf{1}_n + \mathbf{Z}_{41}\boldsymbol{\beta}_{41} + \ldots + \mathbf{Z}_{4K}\boldsymbol{\beta}_{4K},
\end{aligned}
$$

where the functions g are applied element-wise. The same specification can also adopted for the dependence parameter vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)'$ where $\boldsymbol{\eta}_\theta = g_\theta(\boldsymbol{\theta}) = \beta_{50}\mathbf{1}_n + \mathbf{Z}_{51}\boldsymbol{\beta}_{51} + \ldots + \mathbf{Z}_{5K}\boldsymbol{\beta}_{5K}$. Term $\boldsymbol{\beta}_{\text{end}}\mathbf{Y}_2$, the endogenous effect of the semi-continuous variable on the binary response, here modeled using representation (5), enters equation $\boldsymbol{\eta}_\pi$, hence giving the model setup a recursive structure as explained in Section 2. In terms of specification, covariates might be common across the additive predictors, except for the equation(s) related to the endogenous variable which should include at least one variable (an instrument) that is not included in the other predictors (e.g., Han & Vytlacil, 2017). Finally, note that not all parameters have to be specified as functions of predictors.

3 Some fitting details

Let us assume that a random sample $\{(y_{1i}, y_{2i}, \mathbf{z}_i)\}_{i=1}^n$ is available, then the log-likelihood function, $\ell(\boldsymbol{\delta})$, can be obtained by taking the logarithm of $f_{12}(y_1, y_2)$ defined in (2) and creating the indicator variables corresponding to the four possible combinations of the responses. The parameters are defined as $\pi_i = g_\pi^{-1}(\eta_{\pi_i})$, $\mu_i = g_\mu^{-1}(\eta_{\mu_i})$, $\sigma_i = g_\sigma^{-1}(\eta_{\sigma_i})$, $\nu_i = g_\nu^{-1}(\eta_{\nu_i})$ and $\theta_i = g_\theta^{-1}(\eta_{\theta_i})$, and $\boldsymbol{\delta}$ is given by the coefficient vectors associated with η_{π_i}, η_{μ_i}, η_{σ_i},η_{ν_i} and η_{θ_i}, that is $\boldsymbol{\delta} = (\boldsymbol{\beta}_\pi', \boldsymbol{\beta}_\mu', \boldsymbol{\beta}_\sigma', \boldsymbol{\beta}_\nu', \boldsymbol{\beta}_\theta')'$.

Because of the modeling flexibility offered by the additive predictors in the model, parameter estimation is achieved by maximizing the penalized log-likelihood

$$
\ell_p(\boldsymbol{\delta}) = \ell(\boldsymbol{\delta}) - \frac{1}{2}\boldsymbol{\delta}'\mathbf{S}_{\boldsymbol{\lambda}}\boldsymbol{\delta}, \tag{7}
$$

where $\mathbf{S}_{\boldsymbol{\lambda}} = \text{diag}(\boldsymbol{\lambda}_\pi\mathbf{D}_\pi, \boldsymbol{\lambda}_\mu\mathbf{D}_\mu, \boldsymbol{\lambda}_\sigma\mathbf{D}_\sigma, \boldsymbol{\lambda}_\nu\mathbf{D}_\nu, \boldsymbol{\lambda}_\theta\mathbf{D}_\theta)$, each smoothing parameter vector in $\mathbf{S}_{\boldsymbol{\lambda}}$ contains all the smoothing parameters related to the corresponding $\mathbf{D}$ component, and $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_\pi', \boldsymbol{\lambda}_\mu', \boldsymbol{\lambda}_\sigma', \boldsymbol{\lambda}_\nu', \boldsymbol{\lambda}_\theta')'$.

Estimation of $\boldsymbol{\delta}$ and $\boldsymbol{\lambda}$ is based on an extension of the efficient and stable trust region algorithm with integrated automatic multiple smoothing parameter selection discussed in Marra et al. (2020), which uses analytical derivative information of $\ell_p(\boldsymbol{\delta})$. We would like to mention that, while the implementation of the proposed model exploited the infrastructure of `GJRM` as well as several of its internal functions, extending `GJRM` to accommodate a Tweedie marginal required a great deal of programming work.

Inferential results are derived using known theory for general penalized likelihood-based models. Specifically, at convergence, reliable intervals for any linear function of $\boldsymbol{\delta}$ are obtained using the Bayesian large sample approximation $\boldsymbol{\delta} \overset{a}{\sim} N(\hat{\boldsymbol{\delta}}, (\boldsymbol{H}(\hat{\boldsymbol{\delta}}) + \mathbf{S}_{\boldsymbol{\lambda}})^{-1})$, where $\boldsymbol{H}(\hat{\boldsymbol{\delta}})$ is the observed information matrix (Hessian of the negative log likelihood) at $\hat{\boldsymbol{\delta}}$. Intervals for non-linear functions of $\boldsymbol{\delta}$ can be conveniently obtained via posterior simulation (e.g., Marra et al., 2020), whereas p-values for all the terms in the model can be obtained by using the results in Wood (2017) which are based on $\boldsymbol{H}_p(\hat{\boldsymbol{\delta}})^{-1}$. Under some classical model assumptions, it can be proved that $\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}^0 = O_P(n^{-1/2})$ as $n \to \infty$, where $\boldsymbol{\delta}^0$ denotes the 'true' parameter vector. For more information, the reader can consult the on-line supplementary material of Marra et al. (2020) which provides general asymptotic arguments that can also be applied to the current context. Appendix A, in the on-line supplementary material, reports the results of a simulation study which support the empirical effectiveness of the model discussed in this paper.

3.1 Fitting the model in `R`

The proposed copula model can be employed via the `gjrm()` function in the `R` package `GJRM` (Marra & Radice, 2022). An example of the syntax is

```
fl <- list(y1 ~ y2 + z1 + s(z3),
           y2 ~ z1 + s(z2) + s(z3),
              ~ z1 + s(z3),
              ~ z1 + s(z2) + s(z3)
              ~ z1 + s(z3))
mo <- gjrm(fl, margins = c("probit", "TW"), data = md, BivD = "C0", Model = "B")
```

where `fl` is a list containing five equations: the first equation is for parameter π of the Bernoulli distribution of the binary outcome `y1` with `probit` link function (`logit` and `cloglog` are also allowed for); the second, third and fourth equations are for parameters μ, σ and ν of the Tweedie distribution used to model `y2`; the fifth equation is for the copula dependence parameter θ. Argument `BivD` specifies the copula function (Clayton in this case), `Model = "B"` implies that a bivariate model is employed

and `md` is a data frame. Symbol `s()` refers to the smooth function mentioned in Section 2. Default is `bs = "tp"` (penalized 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` (penalized cubic regression spline), `ps` (P-spline) and `mrf` (Markov random field), to name but a few. Note that, e.g., for uni-dimensional smooth functions of continuous covariates, the specific choice of spline definition will not have an impact on the estimated curves. Furthermore, the default value of `k = 10` (which can be increased if desired) is arbitrary although it generally offers enough modeling flexibility in applications. Functions such as `AIC()`, `summary()` and `predict()` can be employed in the usual manner. Function `post.check()` will produce, for the Tweedie margin, a histogram and normal Q-Q plot of modified normalized quantile residuals constructed as follows. For a continuous distribution, $F_2(y_{2i}|\mu_i, \sigma_i, \nu_i)$ is uniform. Under correct specification, therefore, the sample values $F_2(y_{2i}|\hat{\mu}_i, \hat{\sigma}_i, \hat{\nu}_i)$, $i = 1, \ldots, n$, should be approximately uniform, and the transformed values $\Phi^{-1}(F_2(y_{2i}|\hat{\mu}_i, \hat{\sigma}_i, \hat{\nu}_i))$, where $\Phi^{-1}(\cdot)$ is the quantile function of a standard normal distribution, approximately standard normal. To account for the probability mass at zero, we employ an adjustment based on the idea of sampling uniform variates with bounds given by 0 and the upper probability of those observations (Dunn & Smyth, 1996). Residual analysis for the binary margin is not informative (e.g. Collett, 2002). Here, 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.

4 Application to mental care spending and health

The main aim of the present empirical application is to estimate the effect of mental care spending on mental health, using data from the RHIE whose purpose was to determine the impact of health insurance on expenditures and health, if any. The experiment was conducted between 1974 and 1982 at six different sites in the U.S. It randomly assigned, for a period of 3 or 5 years, about 5800 participants, living in 2000 mostly middle-class households, to 14 different insurance plans. Plans can be classified into essentially four types of coverage: full insurance; individual deductible plans; coinsurance plans; and catastrophic coverage only. The data are available as public use files from `https://www.icpsr.umich.edu/icpsrweb/NACDA/studies/06439`; see Deb & Trivedi (2002) or Aron-Dine et al. (2013) for a more detailed description of data and variables, and the latter also for a discussion of enrollment refusal and attrition.

Although the RHIE data collection was completed almost 40 years ago, the study continues to attract interest by researchers and policymakers to this day (e.g., Newhouse & Normand, 2017; Lin & Sacks, 2019). A series of methodological papers in the wake of the RHIE dealt with semi-continuous outcomes when there is a substantial fraction of zeros, as is typically the case for health expenditure or utilization data (e.g., Duan et al., 1983; Deb & Trivedi, 2002). One option is to analyze the number of episodes of treatment separately from the cost per episode, using count data models for the former and linear regression models for the latter (e.g., Keeler & Rolph, 1988). A recent example for a direct analysis of total spending is Kurz (2017), who used a single-equation Tweedie model and showed that it provides a better description of the data than other distributions within the class of generalized linear models.

Our application departs from the earlier literature in a number of ways. First, we focus on spending on mental health care and its effect on mental health. The RHIE provides a mental health score both at baseline and on exit from the program. In addition, mental health expenditures are listed as a separate cost category in the data. It is true that overall spending is small, a possible reason why it has not been much analyzed (Aron-Dine et al., 2013). However, this is mostly due to the high share of non-users (90% in our sample). The average annual amount spent among users is $236. While the high share of zeros can be a problem for traditional modeling approaches, it is compatible with the compound Poisson-Gamma framework of the Tweedie distribution.

Second, we use random plan-assignment as an instrument for spending on mental care. This is in contrast to most of the previous literature that has focused on intention-to-treat effects of plan assignment on expenditures or health. We find evidence for relevance: spending is higher for plans with more generous coverage. Assuming plan-type has no direct effect on mental health then this allows us to estimate the effect of spending on actual health. Our approach addresses the obvious endogeneity problem when regressing mental health on spending. A negative association (more spending, lower health) is likely, simply because low-mental health individuals tend to have a history of low mental health which leads to higher spending on related care. But this is not the causal effect we are interested in. While we can condition on a mental health score at entry, this may be insufficient to control for the full extent of time-invariant, or highly persistent, unobserved heterogeneity in mental care needs. Hence, a joint model that accounts for endogeneity of expenditures is called for.

Third, we specify a recursive copula two-equation model for spending (a semi-continuous variable with a mass point at zero) and mental health (coded as a binary variable). Association between the

two outcomes is modeled using copulae. The benchmark specification combines Bernoulli (with probit link) and Tweedie marginal models using a Gaussian copula. Other variants are explored. The effects of continuous covariates are modeled via spline functions.

Our application also provides an analysis template, as the same methodology can be used to estimate the effect of any type of health care spending on any type of binary health outcome. Also, the order of the equations can be reversed to accommodate a situation of a Tweedie model with binary endogenous regressor.

4.1 Description of data and model

To obtain our final sample, we excluded all individuals under the age of 18, as well as those with missing information on any of the variables employed in the analysis. This leaves us with 2777 observations. Of those, 69% were enrolled in the program for 3 years and the remaining 31% for five years. The spending variable is obtained as the average yearly spending for mental care during the enrollment period (`mentalexp`). Mental health was evaluated upon exit from the program, on a 0-100 scale with mean 76; we classify all individuals with a score under 50 as having "low mental health" (`lowmhix`). This applies to 5% of the sample. We thereby focus on the more severe mental health conditions that typically would require, and benefit most from, medical attention. Setting the cut-off at 60, 70 and 75 did not change the substantive conclusions.

For the Tweedie model, only the location parameter μ is specified as a function of covariates; modeling σ and ν using additive predictors does not alter the substantive conclusions, hence we focus on the location parameter only. Therefore, the model equations for the Bernoulli and Tweedie margins are

$$\pi_{\texttt{lowmhix}} = g_{\pi}^{-1}(\beta_{10}+\beta_{11}s_{11}(\texttt{mentalexp})+\beta_{12}\texttt{female}+\beta_{13}\texttt{white}+\beta_{14}\texttt{poor}+s_{12}(\texttt{mhi})+s_{13}(\texttt{age})+s_{14}(\texttt{educ}))$$

and

$$\mu_{\texttt{mentalexp}} = \exp(\beta_{20} + \beta_{21}\texttt{female} + \beta_{22}\texttt{white} + \beta_{23}\texttt{poor} + \beta_{24}\mathbb{1}(\texttt{plantype} = 2) + \beta_{25}\mathbb{1}(\texttt{plantype} = 3)$$

$$+\beta_{26}\mathbb{1}(\texttt{plantype} = 4) + s_{21}(\texttt{mhi}) + s_{22}(\texttt{age}) + s_{23}(\texttt{educ}))$$

were π is the binary response probability, g_{π}^{-1} is derived according to the link function chosen for

π, and μ is the expected value of mental expenditure. Thus, both equations have six variables in common, namely the mental health score at enrollment (`mhi`), `age`, gender (`female` = 1), race (`white` = 1), the education of the household head in years (`educ`), as well as the indicator variable `poor`. Participants are classified as being poor if the per-capita household income falls into the lower tercile of the distribution.

The factor variable `plantype` is an instrument, i.e., it is included in the second equation but excluded from the first equation. In our data, 33% of observations are assigned to the first, full coverage plan; 22% to the deductible plan; 27% to the coinsurance plan, and the remaining 18% to the catastrophic coverage plan. F-tests for balancing of the pre-treatment variables by plan type do not reject the null hypothesis of no difference (p-values between 0.16 and 0.64) except for the variable `poor`, which is underrepresented in the catastrophic plan, a finding echoing Aron-Dine et al. (2013).

4.2 Results for the univariate models

Table 1 shows estimation results obtained using the univariate approach where the Bernoulli and Tweedie marginals have been estimated separately, as single equation models. The model fits have been obtained using the `R` code reported in the on-line supplementary material (Appendix B). For the Bernoulli margin, using information criteria, the probit link was selected.

The Tweedie coefficients determine the relative change in expected spending associated with an increase in the covariate value. For example, women are predicted to spend $[\exp(0.5692) - 1] \times 100 = 76.7\%$ more than men, keeping everything else constant. The omitted `plantype1` is "full insurance", so `plantype4` (catastrophic insurance) cuts spending on mental care relative to full insurance by approximately 60% ($\exp(-0.8792) - 1 = -0.5849$). The `plantype` variable as a whole has a significant impact on the response (p-value = 0.00027 obtained using the Wald test), as are all the covariates as well as the three spline functions for `mhi`, `age` and `educ`.

Significant predictors of `lowmhix` are `female`, initial `mhi` as well as `educ`. Age, race and poverty status seem not to matter. Note that we initially considered a smooth function of `mentalexp`, that is s_{11}(`mentalexp`). Since this resulted in a straight line estimate, we eventually let the variable enter the equation parametrically. Spending on mental care has a positive coefficient and a z-value of 2.762. Thus, we find indeed that, in this univariate model, higher spending is associated with an increased probability of low mental health. Specifically, the probit average partial effect (APE) predicts a 3.3 percentage points increase in low mental health per 1000 increase in spending. A 95% interval for the

Dependent variable:	`mentalexp` (Tweedie)		`lowmhix` (Bernoulli with probit link)	
	Estimate	Std. Error	Estimate	Std. Error
`(Intercept)`	0.2965	0.4213	-2.1539	0.1528
`mentalexp`$\times 10^3$			0.6444	0.2333
`female`	0.5692	0.1881	0.3157	0.0991
`white`	1.9382	0.4009	0.0717	0.1330
`poor`	0.4539	0.2114	-0.0025	0.1023
`plantype2`	0.1550	0.2279		
`plantype3`	-0.6482	0.2403		
`plantype4`	-0.8792	0.2873		
`s(mhi)`	1.611	p-val $<$ 2e-16	2.051	p-val $<$ 2e-16
`s(age)`	6.389	p-val = 0.028	1.000	p-val = 0.551
`s(educ)`	6.059	p-val $<$ 2e-16	1.002	p-val = 0.022
σ	69.9	(64.2,76.2)		
ν	1.55	(1.53,1.59)		

Table 1: Results from single equation models ($n = 2777$). The first two columns refer to the Tweedie margin and the last two to the Bernoulli (with probit link) margin. As for the smooth functions, we report effective degrees of freedoms (representing the degrees of complexity of the estimated curves) and p-values. 95% intervals for σ and ν are obtained by posterior simulation.

APE obtained by Bayesian posterior simulation (based on the result mentioned in Section 3) is given by [0.010, 0.059], confirming the statistical significance of the effect. This counterintuitive finding likely reflects associated "shocks". During the enrollment period of three or five years, individual unobserved traits related to mental problems can drive up both `mentalexp` as well as the probability of low mental health upon exit, hence generating a spurious positive association in the univariate model. This is exactly the type of problem that estimation of the joint model can address.

We tested for interactions between `poor` and `plantype`, but found no significant evidence for heterogeneous effects. For the sake of completeness, we report in Figure 1 the histogram as well as the QQ-plot for the Tweedie quantile residuals, the latter together with 95% confidence bands. Almost all points are within the intervals, hence supporting the fact that the Tweedie distribution provides an appropriate characterization for mental spending.

4.3 Results for the bivariate model

The great advantage of the copula approach in conjunction with flexible model margins is the modularity that gives the practitioner the opportunity to explore many unique combinations of elements in order to determine, on one hand, the best fitting model, and, on the other, the sensitivity of key results of interest to modeling assumptions. In our case, we take the Tweedie margin as well as the set of covariates as given. In contrast, we consider three link functions for the Bernoulli `lowmhix`

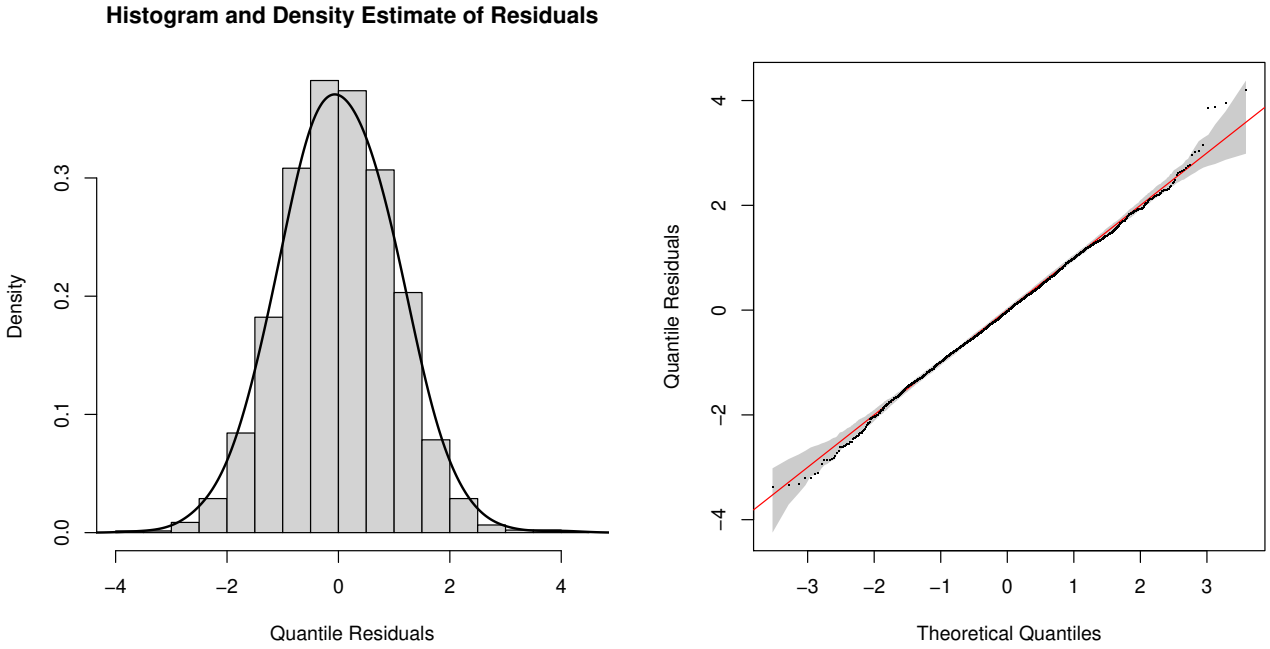


Figure 1: Histogram and normal Q–Q plot (with 95% confidence bands) of normalized quantile residuals for the mental spending variable produced after fitting a Tweedie univariate regression model.

equation (`probit`, `logit` and `cloglog`) and eleven different copulae to account for the dependence between the two equations. The copula models were based on the equations shown in the previous section and were fitted using the syntax discussed in Section 3. Specifying the dependence parameter as function of covariates did not, in this case, return any interesting results as all covariate effects were insignificant. We employed the AIC and Bayesian information criterion (BIC) in order to find the best fitting model given the covariates and the Tweedie margin for spending. Results are provided in Table 2.

The information criteria select the same copula function, namely the Gaussian, but not the same link function for the binary response (`probit` using the AIC and `cloglog` using the BIC), although the substantive conclusions are robust with respect to the selected copula and link function. The full results of the joint Gaussian copula model with Bernoulli (with probit link) and Tweedie margins are given in Table 3, whereas the three estimated smooth functions (for `mhi`, `age` and `educ`) are shown in Figure 2, panel for the mental health equation).

We find that better initial mental health is associated with lower spending as well as a reduced probability of low mental health on exit from the program. The estimated smooth functions are

Link function	Copula	AIC	BIC
probit	N	5891.504	6107.676
logit	N	5895.133	6100.153
cloglog	N	5896.859	6095.220
probit	C0	5897.271	6114.967
logit	C0	5900.432	6106.005
cloglog	C0	5901.629	6103.694
probit	C180	5895.136	6110.731
logit	C180	5898.891	6102.091
cloglog	C180	5900.447	6098.598
probit	J0	5899.430	6114.035
logit	J0	5903.010	6105.085
cloglog	J0	5904.366	6101.348
probit	J180	5897.499	6115.236
logit	J180	5900.621	6107.059
cloglog	J180	5901.842	6104.033
probit	G0	5897.583	6112.597
logit	G0	5901.302	6104.262
cloglog	G0	5902.787	6100.008
probit	G180	5891.693	6108.226
logit	G180	5895.074	6100.828
cloglog	G180	5896.893	6095.932
probit	AMH	5896.485	6106.838
logit	AMH	5899.617	6098.679
cloglog	AMH	5901.260	6096.626
probit	FGM	5895.972	6105.752
logit	FGM	5899.083	6097.999
cloglog	FGM	5900.974	6095.574
probit	T	5897.080	6113.935
logit	T	5901.212	6108.210
cloglog	T	5903.132	6099.329
probit	F	5893.310	6110.670
logit	F	5896.291	6103.940
cloglog	F	5898.574	6098.664

Table 2: AIC and BIC values by different copulae and link functions for the Bernoulli margin.

nearly linear in both cases. The positive and significant association indicates that controlling for initial mental health takes out at least some of the individual heterogeneity that drives both spending and subsequent mental health. The effect of `age` is again similar in both equations, displaying a pronounced M-shape with estimated peaks in the late twenties and mid-fifties. The effect of education is less uniform and only imprecisely determined, especially for the spending equation.

The main result of interest is the effect of spending on mental health. The point estimate in the probit equation of the bivariate recursive copula model is given by -0.2647, with a standard error of

Dependent variable:	`mentalexp` (Tweedie)		`lowmhix` (Bernoulli with probit link)	
	Estimate	Std. Error	Estimate	Std. Error
`(Intercept)`	0.3344	0.4168	-2.1177	0.1520
`mentalexp`$\times 10^3$			-0.2647	0.2821
`female`	0.5793	0.1868	0.3298	0.0979
`white`	1.8936	0.3960	0.0492	0.1308
`poor`	0.4685	0.2096	-0.0334	0.1004
`plantype2`	0.1614	0.2246		
`plantype3`	-0.6830	0.2365		
`plantype4`	-0.8356	0.2825		
`s(mhi)`	1.461	p-val $<$ 2e-16	2.219	p-val $<$ 2e-16
`s(age)`	6.270	p-val = 0.0461	4.458	p-val = 0.469
`s(educ)`	6.050	p-val $<$ 2e-16	1.000	p-val = 0.144
σ	70	(63.1,77)		
ν	1.55	(1.52,1.58)		
θ	0.386	(0.249,0.492)		

Table 3: Results from the chosen copula model ($n = 2777$). 95% intervals for σ, ν, and θ are obtained by posterior simulation.

0.2821. Allowing for endogeneity of spending overturns the earlier finding of a positive and statistically significant association from the single-equation approach. The point estimate as such corresponds to an APE of -0.013 (with a Bayesian 95% confidence set equal to $[-0.044, 0.015]$), slightly negative but statistically insignificant. Not finding an effect of mental health spending on mental health may be surprising at first. However, this in line with Brook et al. (1983), who use the same data source and show that there is no significant effect of insurance status on mental health. Using a regression discontinuity design that exploits the onset of medicare when turning 65, Rhodes (2018) also finds no effect of insurance coverage on mental health. The authors above employed reduced form regression analyzes. This is opposed to our approach which, by modeling the joint distribution of the outcomes with carefully chosen marginals and flexible covariate effects, can reduce the detrimental impact of potential biases due to mis-specification and also yield efficiency gains, aspects that could have led to different substantive conclusions.

There are a number of potential explanations for this null-finding. First, using a subjective evaluation of mental health may suffer from substantial measurement error, hence reducing the precision of estimated effect sizes. Second, in our case, the exogenous change in spending initiated by the insurance instrument may be small, on average, to make a sizeable difference. For example, from the first column of Table 3, the average effect of moving a person from catastrophic coverage to full coverage (from `plantype 4` to `plantype 1`) points to a decrease in the probability of no spending by

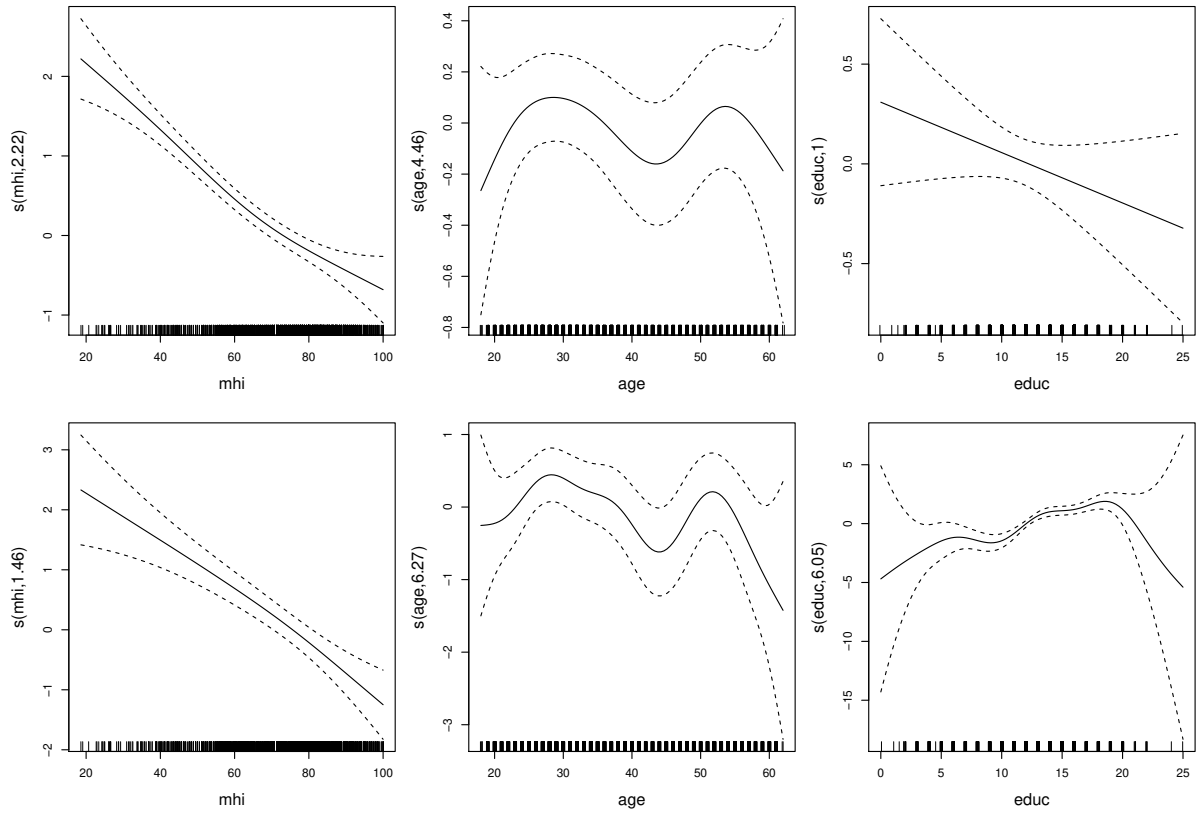


Figure 2: Estimated smooth effects of `mhi`, `age` and `educ` and associated 95% point-wise intervals obtained when fitting a Gaussian copula model with Bernoulli (with probit link) and Tweedie margins to data from the RHIE. The top plots refer to the probit equation whereas the bottom ones to the Tweedie equation. The jittered rug plot, at the bottom of each graph, shows the covariate values. The number in brackets in the y-axis caption represents the effective degrees of freedom of the respective smooth curve.

about 4 percentage points, and an increase in the average amount spent among those having at least some expenditure by $93, on average per year; these are quantitatively not large shifts in spending. And third, health benefits may accumulate only slowly over time, and thus not be visible when the outcome is measured in the short run.

In summary, there is strong evidence that the single-equation adverse "effect" of mental spending on mental health is driven by the endogeneity of spending, for example due to health "shocks" ocurring after the initial mental health score is determined. A copula model can address this endogeneity problem in the context of joint estimation of two non-linear outcome equations. Doing so leads to a null-finding that mirrors the consensus of the previous RHIE literature: while insurance generosity had a sizable effect on spending (of any type, not only mental care), its effects on health outcomes measured upon conclusion of the experimental period were negligible.

5 Conclusions

At a time of ever increasing health expenditures in many developed economies, it is a pressing question whether this is money "well-spent", i.e., whether more spending delivers improvements in health outcomes. Prior research has mostly considered country-level evidence on, for example, child mortality or life expectancy over time and countries. Here, we take a different tack on the issue, by reconsidering individual level data from the Rand Health Insurance Experiment. In either case, with individual or country level data, the main empirical challenge is how to obtain causal effects of spending on health in the presence of the endogeneity: there are potentially many factors unobserved to the analyst that affect both spending and outcome. In our case, the random insurance plan assignment provides a potential instrument for spending. A further issue is the semi-continuity of the individual expenditure variable which in our case has around 90% of zeros and we argue for using the Tweedie distribution in this context.

This paper presents an estimation approach, based on copulae, that consistently estimates the effect of an endogenous semi-continuous regressor on a binary outcome, and applies it to estimating the effect of mental care spending on mental health. The method allows for a great deal of flexibility in the specification of the model and also provides, via the copula dependence parameter, a convenient means to assess the presence of endogeneity. The proposed model can be easily employed via the `R` package `GJRM`.

The methodology presented in this paper is fundamentally parametric and as a such it may suffer from the usual potential drawbacks resulting from model mis-specifications. However, the modeling framework enables a large amount of model exploration via the various functional forms available in `GJRM` which indeed allows for a wide set of copulae as well as a large degree of flexibility in the way regressor effects are specified. Therefore, a researcher can straightforwardly explore several permutations of functional forms and regressor effects. Such exploration is in a way consistent with the philosophy of non-parametric methods in that it enables the data to point to meaningful model structures.

References

Aron-Dine, A., Einav, L. & Finkelstein, A. (2013). The RAND Health Insurance Experiment, Three Decades Later. *Journal of Economic Perspectives*, 27, 197–222.

Brook, R.H. et al. (1983). Does Free Care Improve Adults' Health? Results from a Randomized Controlled Trial. *New England Journal of Medicine*, 309, 1426–1434.

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

Deb, P. & P. Trivedi (2006) Specification and simulated likelihood estimation of a non-normal treatment-outcome model with selection: Application to health care utilization *The Econometrics Journal*, 9.

Deb P., & Trivedi P.K. (2002). The structure of demand for health care: latent class versus two-part models. *Journal of Health Economics*, 21(4), 601–625.

Duan, N., Manning, W.G., Morris, C.N., & Newhouse J.P. (1983). A comparison of alternative models for the demand for medical care. *Journal of Business & Economic Statistics*, 1(2), 115–126.

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

Dunn, P.K. & Smyth, G.K. (2005). Series evaluation of Tweedie exponential dispersion model densities. *Statistics and Computing*, 15(4), 267–280.

Farbmacher H., Ihle P., Schubert I., Winter J., & A. Wuppermann (2017) Heterogeneous Effects of a Nonlinear Price Schedule for Outpatient Care. *Health Economics*, 26(10), 1234–1248.

Han, S. & Vytlacil, E.J. (2017). Identification in a Generalization of Bivariate Probit Models with Dummy Endogenous Regressors. *Journal of Econometrics*, 199(1), 63–73.

Humphreys, B.R., L. McLeod, & J.E. Ruseski, (2014) Physical activity and health outcomes: evidence from Canada. *Journal of Health Economics*, 23(1), 33–54.

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

Jørgensen, B. (1987). Exponential dispersion models. *Journal of the Royal Statistical Society B*, 49, 127–162.

Keeler, E.B. & Rolph, J.E. (1988). The demand for episodes of treatment in the health insurance experiment. *Journal of Health Economics*, 7(4), 337–367.

Kurz, C.F. (2017). Tweedie distributions for fitting semicontinuous health care utilization cost data. *BMC Medical Research Methodology*, 17:171.

Lin, H. & Sacks, D.W. (2019) Intertemporal substitution in health care demand: Evidence from the RAND Health Insurance Experiment. *Journal of Public Economics*, 175, 29–43.

Manning, W.G., Newhouse, J.P., Duan, N., Keeler, E.B. & Leibowitz, A. (1987). Health insurance and the demand for medical care: evidence from a randomized experiment. *American economic review*, 77(3), 251-277.

Manning, W.G., Wells, K.B., Duan, N., Newhouse, J.P. & Ware, J.E. (1986). How Cost Sharing Affects the Use of Ambulatory Mental Health Services. *Journal of the American Medical Association*, 256(14), 1930–1934.

Manning, W.G., Wells, K.B., Buchanan, J.L., Keeler, E.B., Burciaga Valdez, R. & Newhouse, J.P. (1989). Effects of Mental Health Insurance: Evidence from the Health Insurance Experiment. Santa Monica, CA: RAND Corporation. https://www.rand.org/pubs/reports/R3815.html.

Manning, W.G. & J. Mullahy (2001) Estimating log models: to transform or not to transform? *Journal of Health Economics*, 20(4), 461–494.

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

Marra, G., Radice, R. & Zimmer, D. M. (2020). Estimating the Binary Endogenous Effect of Insurance on Doctor Visits by Copula-Based Regression Additive Models. *Journal of the Royal Statistical Society Series C*, 69(4), 953–971.

Mullahy, J. (1998). Much ado about two: reconsidering retransformation and the two-part model in health econometrics *Journal of Health Economics*, 17(3), 247–281.

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

Newhouse, J.P. & Normand, S.T. (2017). Health Policy Trials. *New England Journal of Medicine*, 376(22), 2160–2167.

Nikoloulopoulos, A. K. & Karlis, D. (2010). Regression in a copula model for bivariate count data. *Journal of Applied Statistics*, 37(9), 1555–1568.

OECD (2014). Making Mental Health Count: The Social and Economic Costs of Neglecting Mental Health Care.

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/`.

Rivers, D. & Vuong, Q.H. (1988). Limited information estimators and exogeneity tests for simultaneous probit models. *Journal of Econometrics*, 39(3), 347–366.

Rhodes, J.H. (2018) Changes in the Utilization of Mental Health Care Services and Mental Health at the Onset of Medicare. *Journal of Mental Health Policy and Economics*, 21(1), 29–41.

Schepsmeier, U., Stoeber, Brechmann, E.C., Graeler, B., Nagler, T., Erhardt, T., Almeida, C., Min, A., Czado, C., Hofmann, M., Killiches, M., Joe, H. & Vatter, T. (2019). *VineCopula: Statistical Inference of Vine Copulas.* R package version 2.2.0, URL `https://cran.r-project.org/package=VineCopula`.

Sklar, A. (1973). Random variables, joint distributions, and copulas. *Kybernetica*, 9(6), 449–460.

Smyth, G.K. & Jørgensen, B. (2002). Fitting Tweedie's compound Poisson model to insurance claims data: dispersion modeling. *Astin Bulletin*, 32(1), 143–157.

Rigby, R. A. and Stasinopoulos, D. M. (2005), Generalized additive models for location, scale and shape *Journal of the Royal Statistical Society: Series C*, 54(3), 507–554.

Trivedi, P. & Zimmer, D. (2017). A Note on Identification of Bivariate Copulas for Discrete Count Data. *Econometrics*, 5(1), 10.

United Nations (2015). UN General Assembly, Transforming our world : the 2030 Agenda for Sustainable Development. 21 October 2015, A/RES/70/1, available at: https://www.refworld.org/docid/57b6e3e44.html

Waters, H.R. (1999). Measuring the impact of health insurance with a correction for selection bias - a case study of Ecuador. *Health Economics*, 8(5), 473–483.

Wood, S.N., Pya, N., Säfken B. (2017). Smoothing parameter and model selection for general smooth models. *Journal of the American Statistical Association* 111(516), 1548–1563.

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

Wooldridge, J.M. (2010). *Econometric Analysis of Cross Section and Panel Data*. The MIT Press, London.

On-Line Supplementary Material

Appendix A: Simulation Study

This section provides simulation evidence on the practical performance of the proposed approach. The data generating process (DGP) has been designed to mimic some of the features of the data as well as results from the case study discussed in Section 4 of the article. For instance, the distributions of the binary and semi-continuous variables have been simulated to look very similar to their observed versions, and the values for σ, ν and θ chosen to be very close to the estimates obtained from the empirical analysis. The DGP has the form

$$
\begin{aligned}
Y_2 &\sim \text{TW with } \mu = g_\mu^{-1}(1 + s_2(z_1) + s_3(z_2) + z_3),\ \sigma = g_\sigma^{-1}(4.24),\ \nu = g_\nu^{-1}(0.22), \\
Y_1 &\sim \text{ Bernoulli with } \pi = \Phi\left\{-3 + \beta_{\text{end}} y_2 + s_1(z_1) + z_2\right\},
\end{aligned}
$$

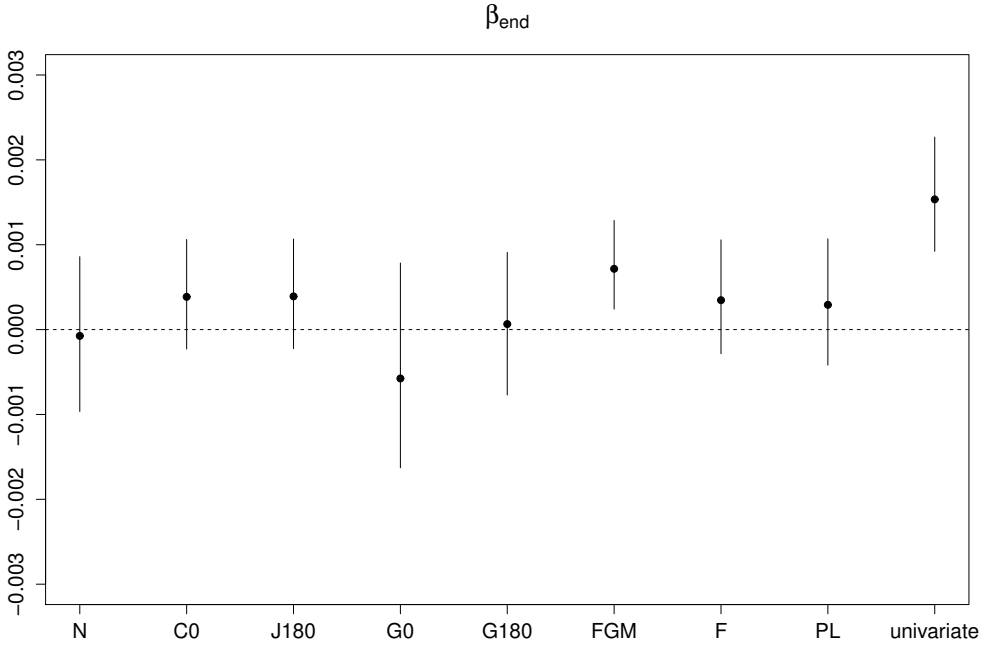
where $s_1(z) = z + \exp\left\{-30(z-0.5)^2\right\}$, $s_2(z) = 0.6\left\{\exp(z) + \sin(2.9z)\right\}$ and $s_3(z) = 0.6\sin(2\pi z)$, and the inverses of the g functions can be worked out from the expressions reported in Section 2.2 of the paper. In line with the results from the empirical analysis β_{end} is set to 0. Variables z_1, z_2 (the instrument) and z_3 are generated using a multivariate standard Gaussian with correlation parameters set at 0.5, and then transformed using the distribution function of a standard Gaussian. Regressor z_3 is dichotomised by rounding it. Associated responses are generated via function `BiCopSim()`, from the package `VineCopula` (Schepsmeier et al., 2019), using the Gaussian copula with dependence parameter specified as $\theta = \tanh(0.41)$. We considered sample sizes of 1500 and 3000, while the number of replicates was set to 500. Appendix B reports the code used to simulate the data.

Using `gjrm()` we fitted models with `probit` and `TW` marginals, and a handful of copulae available in package, specifically `N`, `C0`, `J180`, `G0`, `G180`, `FGM`, `F` and `PL`. Using `gamlss()` within `GJRM`, we also fitted the univariate model with `TW` distribution. The correct model is the one based on the Gaussian copula, whereas all the others served to assess the impact that misspecifying the copula has on the endogenous effect of interest. The smooth components in the models were represented using penalized low rank thin plate splines with second order penalty and 10 bases (Wood, 2017). For each replicate,

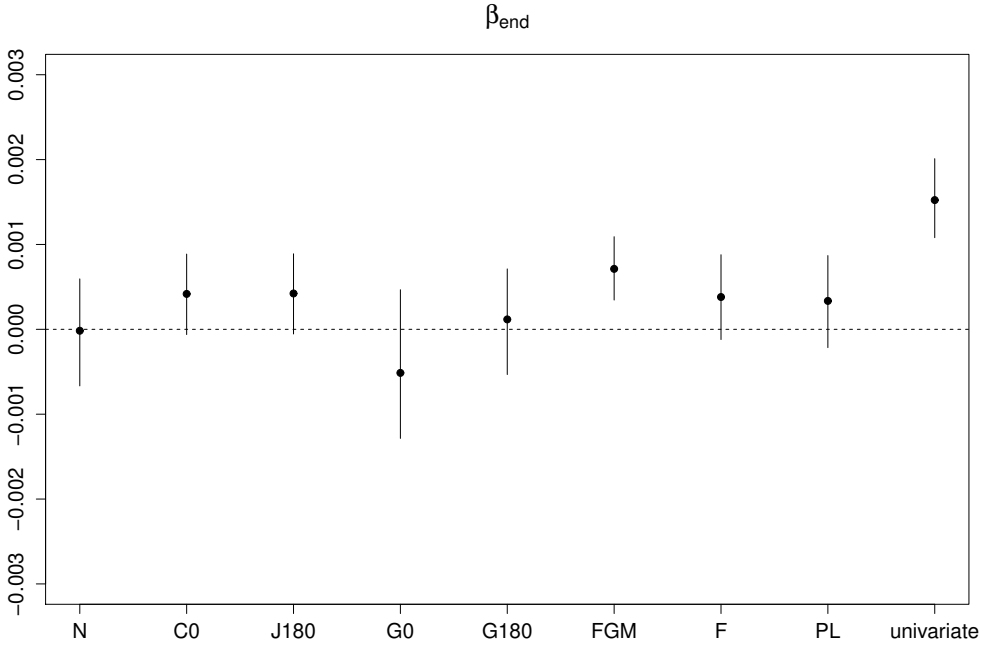
curve estimates were constructed using 200 equally spaced fixed values in the $(0, 1)$ range.

Figure 3 shows the estimates for the endogenous parameter of interest β_{end} when using several copulae. As expected, the simple univariate model (`univariate`), which does not correct for endogeneity, is clearly not able to capture the true parameter value; the estimates are upwardly biased, reflecting the positive copula dependence. All in all, when the copula is misspecified the results show some non-vanishing bias, hence suggesting that mis-modeled dependence structure plays a role in the estimation of the effect of interest. However, the estimates are still reasonable if we consider how biased the estimates from the univariate model are. The results obtained under the correct copula (`N`) exhibit the best bias-variance tradeoff.

We also report the simulation results for the other model's terms, namely β_{12}, β_{21}, σ, ν, θ, $s_1(\cdot)$, $s_2(\cdot)$ and $s_3(\cdot)$; these are displayed in Figures 4 and 5. The findings further support the empirical effectiveness of the proposed approach since the true values are overall well recovered.

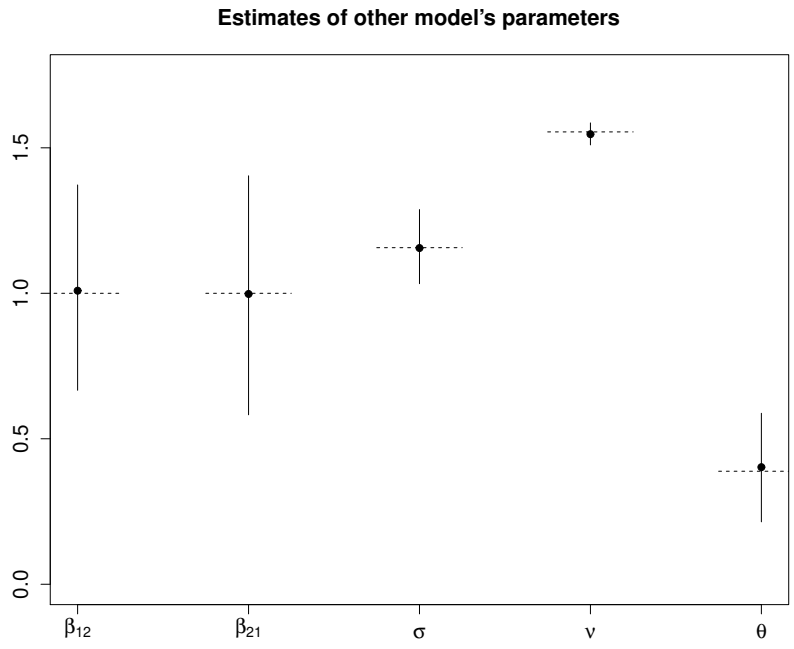


(a) $n = 1500$

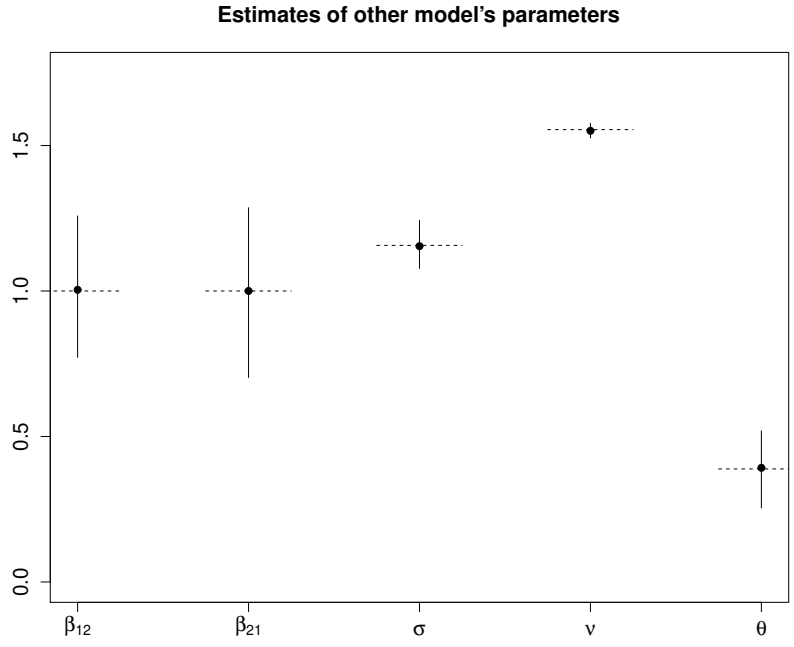


(b) $n = 3000$

Figure 3: Estimates for β_{end} (the endogenous parameter of interest) obtained by applying `gjrm()` and `gamlss()`, both from `GJRM`, to simulated data based on the Gaussian copula with probit and Tweedie margins. The correct model is denoted by `N`, while the other models are misspecified. Circles indicate mean estimates while bars represent the estimates' ranges resulting from 5% and 95% quantiles. The true value is 0 and is denoted by the dashed horizontal line.

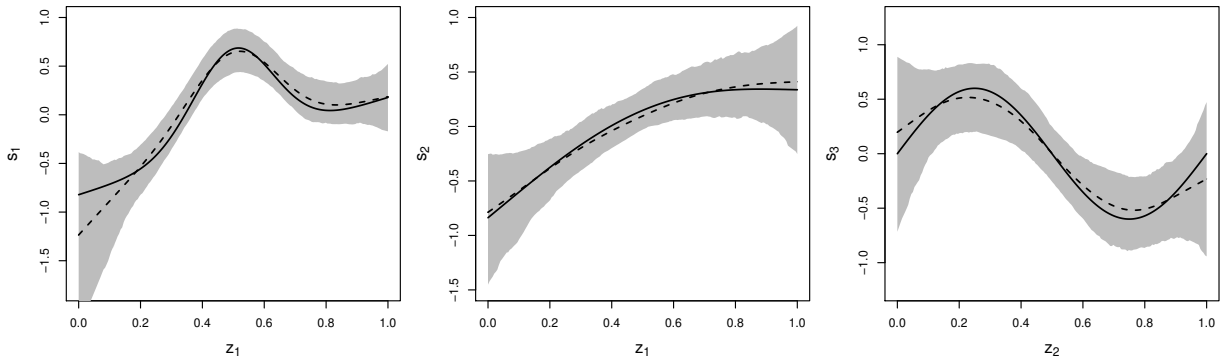


(a) $n = 1500$

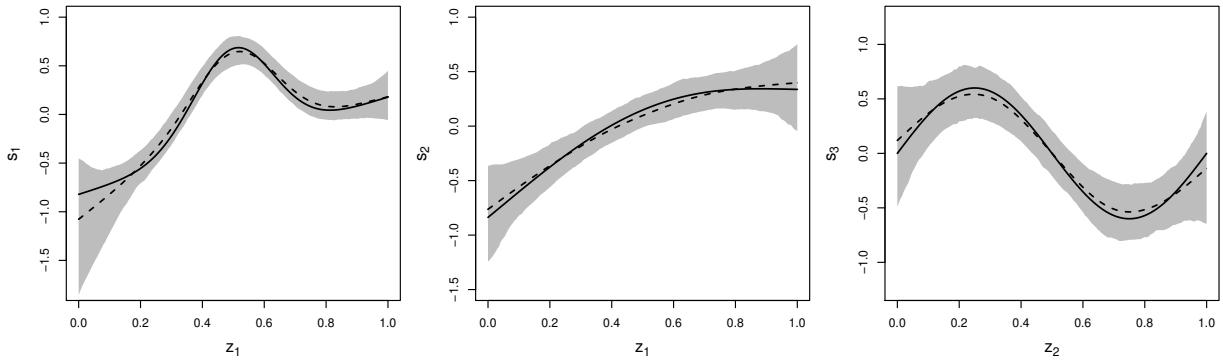


(b) $n = 3000$

Figure 4: Estimates of other model's parameters obtained by applying `gjrm()` to simulated data based on a Gaussian copula with probit and Tweedie margins when fitting correctly specified models. Circles indicate mean estimates while bars represent the estimates' ranges resulting from 5% and 95% quantiles. True values are indicated by dashed horizontal lines.



(a) $n = 1500$



(b) $n = 3000$

Figure 5: Estimates of smooth effects obtained by applying `gjrm()` to simulated data based on a Gaussian copula with probit and Tweedie margins when fitting correctly specified models. True functions are represented by black solid lines, mean estimates by dashed lines and point-wise ranges resulting from 5% and 95% quantiles by shaded areas.

Appendix B: R code

Univariate models

```
eq1 <- lowmhix   ~ mentalexp + s(mhi) + female + white + s(age) + s(educ) + poor
eq2 <- mentalexp ~ s(mhi) + female + white + s(age) + s(educ) + plantype + poor

# gam() from mgcv
out1 <- gam(eq1, data = mydata, family = binomial(link = "probit"))

# gamlss() from GJRM
out2 <- gamlss(list(eq2, ~ 1, ~ 1), data = mydata, margin = "TW")
```

Simulated data

```
library(GJRM)
library(VineCopula)
library(tweedie)

n       <- 1500 # 3000
n.rep   <- 500
cor.cov <- 0.5

s1 <- function(x) x + exp(-30*(x-0.5)^2)
s2 <- function(x) 0.6*(exp(x) + sin(2.9*x))
s3 <- function(x) 0.6*sin(2*pi*x)

cor.cov <- matrix(0.5, 3, 3); diag(cor.cov) <- 1

cov <- rMVN(n, rep(0,3), cor.cov)
cov <- pnorm(cov)

z1 <- cov[, 1]
z2 <- cov[, 2]
z3 <- round(cov[, 3])

theta <- tanh(0.41)

u1u2 <- BiCopSim(n, family = 1, par = theta)
u1   <- u1u2[,1]
u2   <- u1u2[,2]

mu.st    <- 1 + s2(z1) + s3(z2) + z3
sigma.st <- 4.24
nu.st    <- 0.22

mu    <- exp(mu.st)
sigma <- esp.tr(sigma.st, "TW")$vrb
nu    <- enu.tr(nu.st, "TW")$vrb

y2 <- NA
for(j in 1:n) y2[j] <- qtweedie(u2[j], xi = sigma, mu = mu[j], phi = nu)
```

```
y1 <- as.numeric( ( -3 + s1(z1) + z2 + qnorm(u1) ) > 0 )

md <- data.frame(y1 = y1, y2 = y2, z1 = z1, z2 = z2, z3 = z3)
```