A Unifying and Flexible Bivariate Copula Regression Framework

Rosalba Radice
Faculty of Actuarial Science & Insurance
Bayes Business School

Insurance Data Science Conference
16^{th} June 2023

A Unifying Approach

The proposal is a *regression* technique which allows to model simultaneously two variables of interest via *bivariate copulae* and as function of *flexible* covariates effects.

It is *unifying* because it allows to fit a *large variety of models* using a general penalised maximum likelihood approach.

All models developed have been incorporated in the `GJRM R` package, to facilitate the use of such models in industry and academia in a variety of data-analytic settings, and to enhance reproducible research.

		LM `lm()`	GLM `glm()`	GAM `gam()`	GAMLSS `gamlss()`	Copula Reg `gjrm()`
Distrib	Normal	✓	✓	✓	✓	✓
	Exp. family	✗	✓	✓	✓	✓
	Any	✗	✗	✗	✓	✓
Effects	Non-flex.	✓	✓	✓	✓	✓
	Flexible	✗	✗	✓	✓	✓
Param	Mean	✓	✓	✓	✓	✓
	Any	✗	✗	✗	✓	✓
	Copulae	✗	✗	✗	✗	✓

The main building blocks of the copula regression models are: distributions, flexible covariate effects on any parameter, copulae. (Both `gam()` and `gjrm()` can also model time to event data.)

Why Modelling Bivariate Variables

- Treatment selection bias also known as endogeneity. Both treatment and outcome variable need to be modelled simultaneously. An example is estimating the effect of having private health insurance on doctor visits.

- Missing not at random also known as non-ignorable selection or sample selection. Both selection and outcome equations are modelled simultaneously. An example is modelling insurance claim payments data using a sample of accepted claims only.

- Associated outcomes; e.g., modelling multiple types of insurance claims.

Flexible Bivariate Copula Regression

Let Y_1 and Y_2 be two variables of interest, their bivariate distribution can be represented as

$$C\left(\mathcal{F}_{y_1}(y_1|\mu_1,\sigma_1,\nu_1,\ldots),\mathcal{F}_{y_2}(y_2|\mu_2,\sigma_2,\nu_2,\ldots);\theta\right).$$

- C is a two-place copula function.
- $\mathcal{F}_{y_1}(y_1|\mu_1,\sigma_1,\nu_1,\ldots)$ and $\mathcal{F}_{y_2}(y_2|\mu_2,\sigma_2,\nu_2,\ldots)$ are CDFs.
- θ measures the dependence between Y_1 and Y_2.
- $g_{\mu_1}(\mu_1) = \eta_{\mu_1}$, $g_{\sigma_1}(\sigma_1) = \eta_{\sigma_1}$, $g_{\nu_1}(\nu_1) = \eta_{\nu_1}$, ..., $g_{\mu_2}(\mu_2) = \eta_{\mu_2}$, $g_{\sigma_2}(\sigma_2) = \eta_{\sigma_2}$, $g_{\nu_2}(\nu_2) = \eta_{\nu_2}$, ..., and $g_\theta(\theta) = \eta_\theta$.
- η_{μ_1}, η_{μ_2}, ..., and η_θ are predictors (made up of covariates and parameters) which allow for spatial effects, non-linear effects, etc.
- θ is specified as a function of η_θ (allowing it to vary, e.g., according to individual and or geographic features).

Distributions

- For binary variables: Bernoulli with logit, probit and cloglog link functions.
- For discrete variables: discrete generelised Pareto, Poisson, zero truncated Poisson, negative binomial, Poisson inverse Gaussian, Tweedie.
- For continuous variables: normal, log-normal, Gumbel, reverse Gumbel, generelised Pareto, logistic, Weibull, inverse Gaussian, gamma, Dagum, Singh-Maddala, beta, Fisk, Tweedie.
- For survival models, the margins can be proportional hazards, proportional odds or probit.
- We can also model ordinal outcomes.

Flexible (Additive) Predictors

Each predictor can be written as

$$\eta_i = \beta_0 + s_1(\mathbf{x}_{1i}) + s_2(\mathbf{x}_{2i}) + \ldots + s_K(\mathbf{x}_{Ki}), \quad i = 1, \ldots, n,$$

where the $\mathbf{x}$ are covariate vectors.

For each i, the generic $s(\mathbf{x}_i)$ can be approximated as $\sum_{j=1}^{J} \beta_j b_j(\mathbf{x}_i)$.

For all observations, we have $\mathbf{X}\boldsymbol{\beta}$ where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_J)^\mathsf{T}$ and $X[i,j] = b_j(\mathbf{x}_i)$. Hence,

$$\boldsymbol{\eta} = \beta_0 \mathbf{1}_n + \mathbf{X}_1\boldsymbol{\beta}_1 + \ldots + \mathbf{X}_K\boldsymbol{\beta}_K.$$

Quadratic penalty $\lambda_k \boldsymbol{\beta}_k^\mathsf{T} \mathbf{S}_k \boldsymbol{\beta}_k$ imposes specific properties on the k^{th} function.

Key ingredients are $\mathbf{X}_k$ and $\mathbf{S}_k$.

Linear and Non-Linear Effects

For binary or categorical predictors, $s_k(\mathbf{x}_{ki})$ is approximated by

$$\mathbf{x}_{ki}^{\mathsf{T}}\boldsymbol{\beta}_k.$$

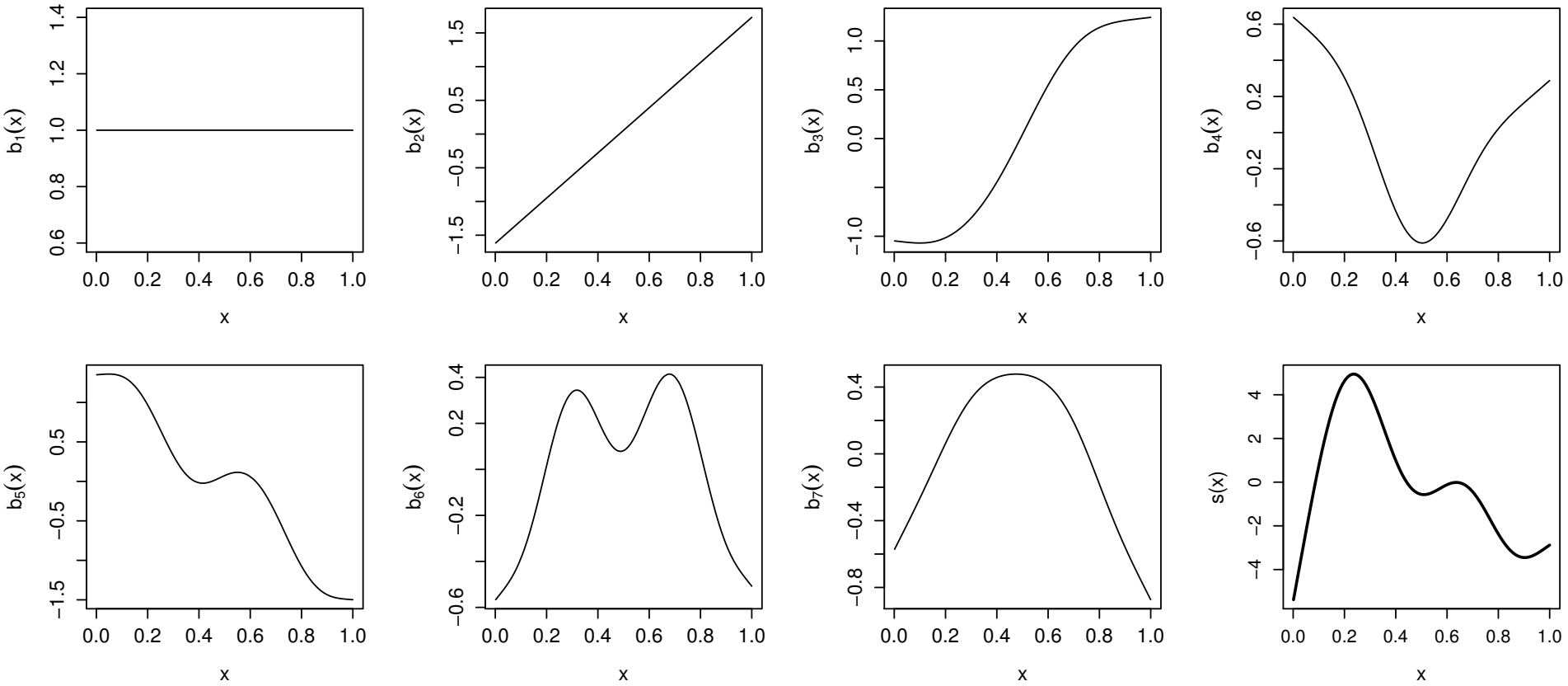
$\mathbf{X}_k$ is obtained by stacking all covariate vectors. Typically, $\mathbf{S}_k = \mathbf{0}$ but sometimes $\mathbf{S}_k = \mathbf{I}$, where $\mathbf{I}$ is an identity matrix. (This has the interpretation of a random effect.)

For continuous variables,

$$\sum_{j_k=1}^{J_k} \beta_{kj_k} b_{kj_k}(x_{ki}),$$

where the $b_{kj_k}(x_{ki})$ are known spline basis functions. $\mathbf{X}_k$ comprises the basis function evaluations for each i. $\mathbf{S}_k = \int \mathbf{d}_k(x_k)\mathbf{d}_k(x_k)^{\mathsf{T}} dx_k$, where the j_k^{th} element of $\mathbf{d}_k(x_k)$ is given by $\partial^2 b_{kj_k}(x_k)/\partial x_k^2$.

Thin Plate Regression Spline Example



Example of Spatial Effects

Geographic location of statistical units is exploited using

$$\mathbf{x}_{ki}^{\mathsf{T}}\boldsymbol{\beta}_k,$$

where $\boldsymbol{\beta}_k = (\beta_{k1}, \ldots, \beta_{kR})^{\mathsf{T}}$ contains R spatial effects and

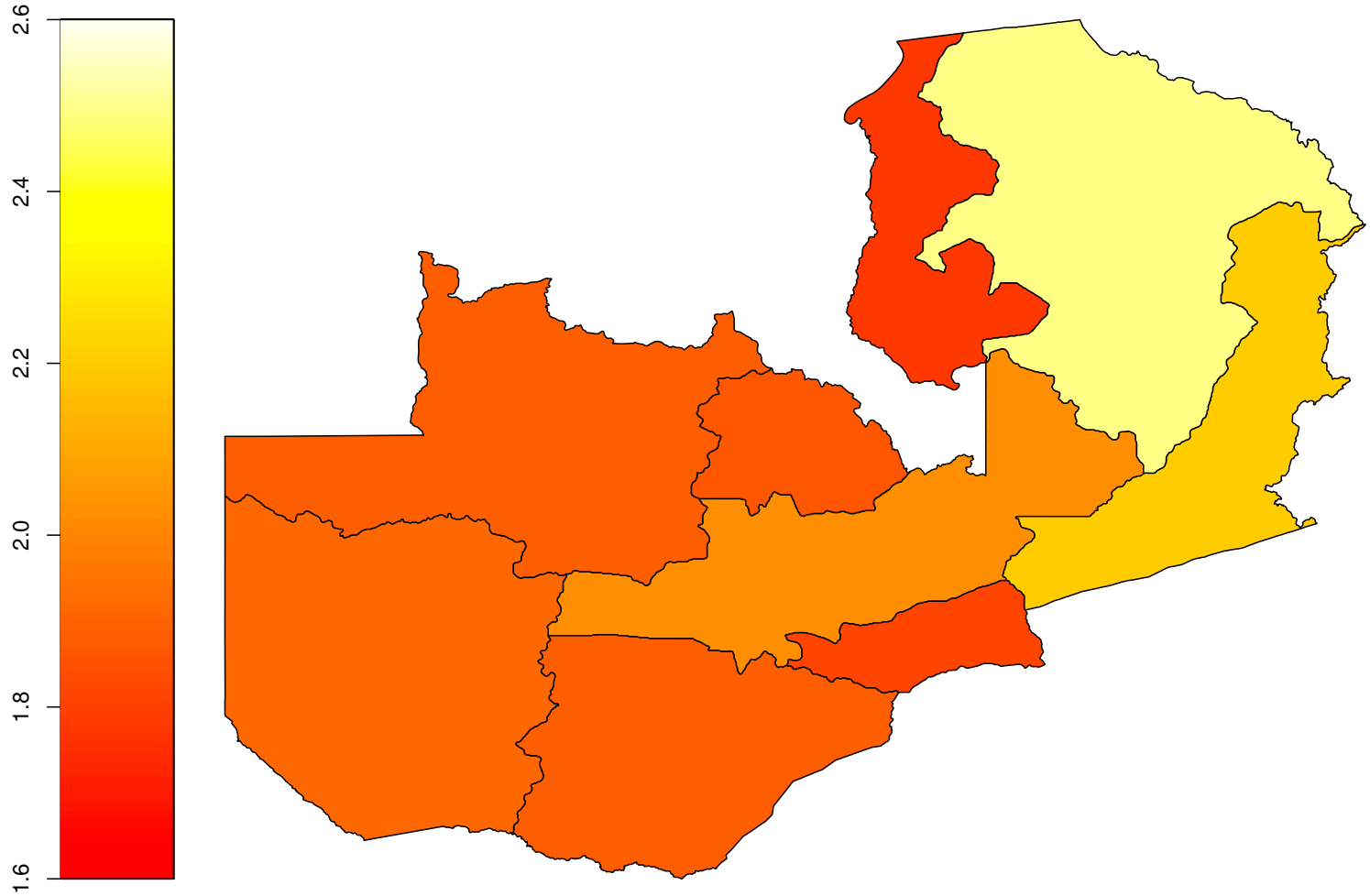
$$\mathbf{X}_k[i, r] = \begin{cases} 1 & \text{if the observation belongs to region } r \\ 0 & \text{otherwise} \end{cases}, \quad r = 1, \ldots, R.$$

The smoothing penalty is an adjacency matrix

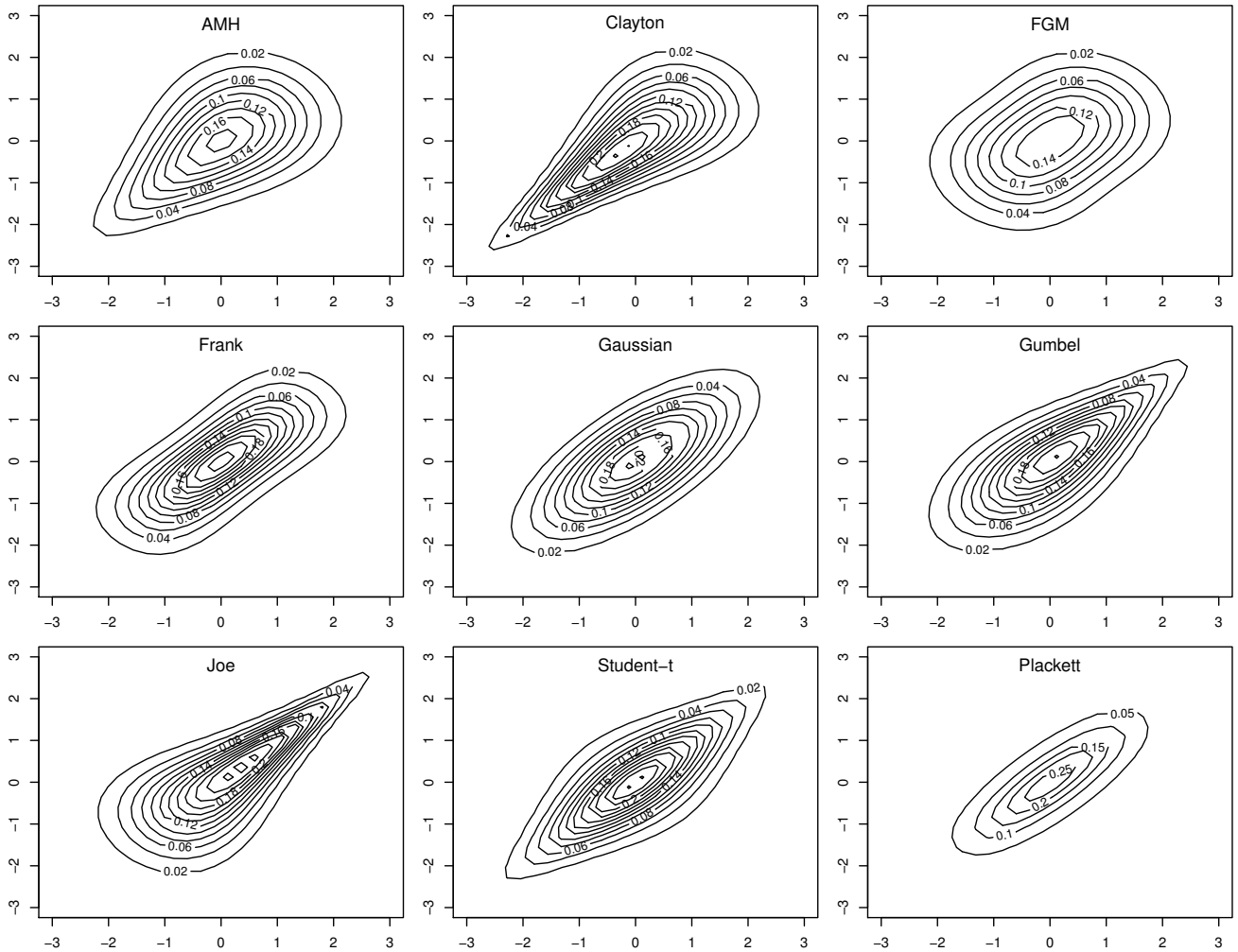
$$\mathbf{S}_k[r, q] = \begin{cases} -1 \text{ if } r \neq q \wedge r \text{ and } q \text{ are adjacent neighbors} \\ 0 \quad \text{if } r \neq q \wedge r \text{ and } q \text{ are not adjacent neighbors} \\ N_r \text{ if } r = q \end{cases},$$

where N_r is the total number of neighbors for region r. (This has the interpretation of a Gaussian Markov random field.)

Cont'd



Some of the Copulae in GJRM



Rotations of Clayton, Joe and Gumbel can also be employed.

Penalised MLE

Estimation is based on direct optimisation of

$$\ell_p(\boldsymbol{\psi}) = \ell(\boldsymbol{\psi}) - \frac{1}{2}\boldsymbol{\psi}^{\mathsf{T}}\left(\sum_k \lambda_k \mathbf{S}_k\right)\boldsymbol{\psi},$$

where $\ell(\boldsymbol{\psi})$ is the model log.-lik. with overall parameter vector $\boldsymbol{\psi}$. Main challenges were to: (i) develop a stable algorithm for problems which may not be concave and/or exhibit regions that are close to flat; (ii) carry out smoothing using a fast and automatic approach; (iii) allow for modularity and generality. Solution: carefully structured trust region algorithm with general automatic smoothing parameter selection, based on analytical score and Hessian.

Intervals for (linear and non-linear functions) of model coefficients obtained using $\boldsymbol{\psi} \dot{\sim} \mathcal{N}(\hat{\boldsymbol{\psi}}, -\widehat{\boldsymbol{\mathcal{H}}}_p^{-1})$. Intervals for non-linear functions of $\boldsymbol{\psi}$ conveniently obtained by simulation.

Case Study

Interested in the potential endogenous effect of insurance status (binary variable) on number of doctor visits (count variable) in the presence of unobserved confouders (e.g., health related traits).

2010 Medical Expenditure Panel Survey, $n = 13137$, respondents between aged $25 - 64$.

Instruments: firm size (`bigfirm`) and indicator of whether the firm has multiple locations (`multloc`). They should influence insurance status (due to economies of scale that make it cheaper for larger firms to offer health coverage) and are are unlikely to (directly) affect usage (especially after controlling for employment status).

Tweedie Distribution

- The Tweedie distribution is a linear exponential dispersion model with the power mean-variance relationship $\text{var}(Y) = \sigma\mu^{\nu}$, where $\mu = \mathbb{E}(Y) > 0$, $\sigma > 0$ is a scale parameter, and $\nu \in \mathbb{R}$ the shape parameter.
- 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.
- When $\nu \in (1, 2)$, a Tweedie becomes a compound Poisson-gamma distribution.
- The resulting density is supported on the non-negative real line and has a positive mass at 0.
- Within our framework, μ, σ and ν can be specified as functions of predictors.

Cont'd

The density of the Tweedie is

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

where

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

and

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

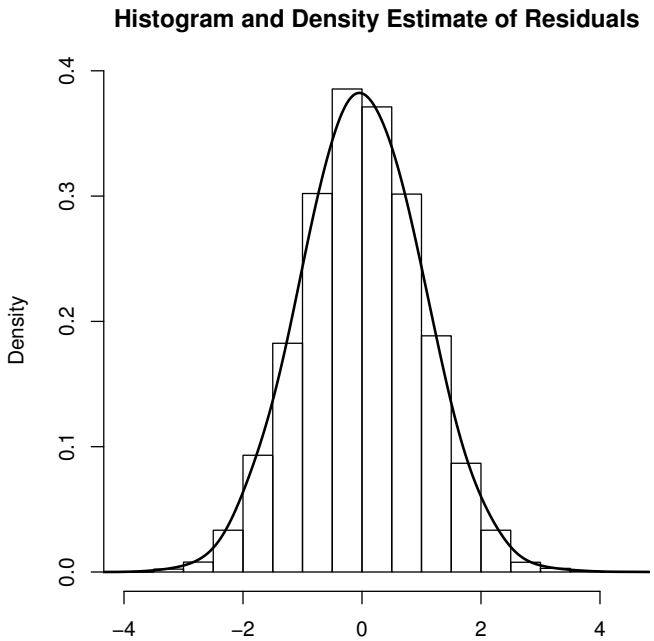
Methods for computing the density, in particular factor $a(y,\sigma,\nu)$, first and second order derivatives of the log-density, while avoiding numerical problems, are available.

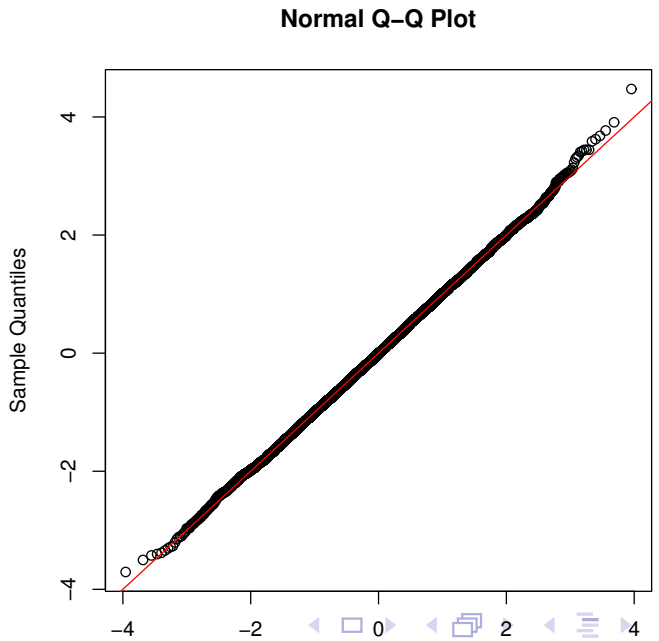
However, specific numerical methods were required to compute the Tweedie CDF and its derivatives, needed for copula modelling.

library(GJRM)

```
eq.mu1   <- priv   ~         s(age) + female + fairpoor + employed +
                             bigfirm + multloc
eq.mu2   <- docvis ~ priv + s(age) + female + fairpoor + employed
eq.sig2  <-        ~ 1
eq.nu2   <-        ~ 1
eq.theta <-        ~ 1
f.l <- list(eq.mu1, eq.mu2, eq.sig2, eq.nu2, eq.theta)

out <- gjrm(f.l, data = mydata, margins = c("probit", "TW"),
                 Model = "B", BivD = "PL")
```





Interpretation of Endogenous Effect and Copula Parameter

- Univariate Tweedie: $1.85\ (1.70, 2.02)$.
- Copula (Plackett) model with Tweedie marginal: $5.23\ (4.39, 6.24)$.
- The average number of doctor visits for an insured individual is approx. 5 times that for an individual with no insurance.
- The copula parameter is $0.27\ (0.19, 0.39)$.
- For the Plackett copula when $0 < \theta < 1$ the dependence is negative, hence the estimate suggests the presence of favorable selection.
- Once favourable selection has been taken into account, the effect of insurance is larger than when it is ignored.

Some key references

- Marra, G., Fasiolo, M., Radice, R. and Winkelmann, R. (2023). A flexible copula regression model with Bernoulli and Tweedie margins for estimating the effect of spending on mental health. *Health Economics*, 32(6), pp. 1305–1322.
- Ranjbar, S., Cantoni, E., Chavez-Demoulin, V., Marra, G., Radice, R. and Jaton, K. (2022). Modelling the Extremes of Seasonal Viruses and Hospital Congestion: The Example of Flu in a Swiss Hospital. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 71(4), pp. 884–905.
- Marra, G., Radice, R. and 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: Applied Statistics*, 69(4), pp. 953–971.
- Marra, G. and Radice, R. (2020). Copula Link-Based Additive Models for Right-Censored Event Time Data. *Journal of the American Statistical Association*, 115(530), pp. 886–895.
- Marra G, Radice R, Barnighausen T, Wood SN, McGovern ME (2017), A Simultaneous Equation Approach to Estimating HIV Prevalence with Non-Ignorable Missing Responses, *Journal of the American Statistical Association*, 112(518), pp. 484-496.