ARTICLE IN PRESS

Journal of Statistical Planning and Inference xx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

A trivariate additive regression model with arbitrary link functions and varying correlation matrix

Panagiota Filippou [a,*], Thomas Kneib [b], Giampiero Marra [a], Rosalba Radice [c]

[a] Department of Statistical Science, University College London, Gower Street, London WC1E 6BT, UK
[b] Chairs of Statistics and Econometrics, Georg-August-Universität Göttingen, Humboldtallee 3, 37073 Göttingen, Germany
[c] Department of Economics, Mathematics and Statistics, Birkbeck, University of London, Malet Street, London WC1E 7HX, UK

ARTICLE INFO

Article history:
Received 17 October 2017
Received in revised form 13 March 2018
Accepted 5 July 2018
Available online xxxx

Keywords:
Additive predictor
Binary response
Cholesky decomposition
Penalized regression spline
Simultaneous parameter estimation
Trivariate distribution

ABSTRACT

In many empirical situations, modelling simultaneously three or more outcomes as well as their dependence structure can be of considerable relevance. Copulae provide a powerful framework to build multivariate distributions and allow one to view the specification of the marginal responses' equations and their dependence as separate but related issues. We propose a generalization of the trivariate additive probit model where the link functions can in principle be derived from any parametric distribution and the parameters describing the residual association between the responses can be made dependent on several types of covariate effects (such as linear, nonlinear, random, and spatial effects). All the coefficients of the model are estimated simultaneously within a penalized likelihood framework that uses a trust region algorithm with integrated automatic multiple smoothing parameter selection. The effectiveness of the model is assessed in simulation as well as empirically by modelling jointly three adverse birth binary outcomes in North Carolina. The approach can be easily employed via the `gjrm()` function in the `R` package `GJRM`.

1. Introduction

When the researcher is interested in modelling more than one response, univariate regression will not yield valid inferences if there is residual dependence between the outcomes conditional on covariates. The case of trivariate models has been discussed in the literature in various contexts. For example, Loureiro et al. (2010) assessed the effect of parental smoking habits on their children's smoking habits by estimating a three-equation probit regression model, whereas Zhong et al. (2012) evaluated the safety of a treatment and identified an optimal dose by jointly modelling the probabilities of toxicity, efficacy, and surrogate efficacy given a specific dose. Król et al. (2016) examined the response to a treatment on patients with metastatic colorectal cancer by analysing simultaneously three outcomes: a longitudinal marker, a set of recurrent events, and a terminal event. A mixture of powers copula-based approach to model jointly three binary and discrete outcomes was employed by Zimmer and Trivedi (2006), whereas Zhang et al. (2015) developed a Bayesian algorithm to estimate trivariate probit-ordered models affected by double sample selection.

This paper contributes to the literature by introducing a generalization of the trivariate additive probit model. Specifically, we extend and therefore enhance the model proposed by Filippou et al. (2017) by allowing (i) the link functions to be virtually derived from any parametric distribution and (ii) the model's association parameters to depend on several types of covariate effects (such as linear, nonlinear, random, and spatial effects). The first extension allows for the use of link functions

* Corresponding author.
E-mail address: panagiota.filippou@hotmail.com (P. Filippou).

https://doi.org/10.1016/j.jspi.2018.07.002
0378-3758/

other than probit. In particular, the additional link functions implemented for this work are the logit and complementary log–log which are used extensively in numerous disciplines, including the medical and social sciences. In clinical research logit models are widely employed as they provide direct information about which treatment has the best odds of benefiting a patient, for instance. Complementary log–log models have important applications in survival analysis where they can, for example, provide a clear insight into the relative reduction of risk for death or progression. Extension (ii) is of some relevance since it can help to gain insights into the way the residual association between the responses is modified by the presence of covariates. As will be further elaborated in the paper, the practical success of such extensions depends on the use of a computationally efficient and theoretically tractable parametrization for the model's correlation matrix as well as the availability of the analytical score and Hessian of the proposed model's log-likelihood which are not trivial to derive.

To the best of our knowledge, the two proposed developments have not been considered in the context of trivariate (or more generally, multivariate) binary response regression models. Note also that, despite we have focused on the trivariate case, the model's formulation in Section 2 could be easily extended to the multivariate context as would be in principle the proposed estimation framework. Finally, it is worth pointing out that our proposal may be regarded as an extension of the bivariate regression approaches introduced by Marra and Radice (2017a), Klein and Kneib (2016) and Radice et al. (2016) as well as of the popular generalized additive models (GAMs) and GAMs for location, scale and shape of Wood (2017) and Rigby and Stasinopoulos (2005). In summary, the two main contributions of this paper are to extend the model introduced by Filippou et al. (2017) as detailed above and to make the new developments available via the `gjrm()` function from the `R` package `GJRM` (Marra and Radice, 2017b).

Section 2 introduces the proposed model, Section 3 describes the log-likelihood and Section 4 provides the key details on parameter estimation. The proposal is empirically evaluated in a simulation study, presented in Section 5, and then applied to a case study in Section 6, where the interest is in modelling jointly three adverse birth binary outcomes in North Carolina. Section 7 concludes the paper.

2. Model specification

This section introduces an extension of the trivariate probit that is based on copulae, arbitrary parametric link functions, additive predictors and a modified Cholesky decomposition of the model's correlation matrix.

In general, a multivariate distribution can be constructed using a copula function that joins together marginal distributions which may come from different families (Joe, 1997). Suppose that $\mathcal{C}$ denotes a joint cumulative distribution function (cdf) with support in $[0, 1]^3$ and whose one-dimensional margins are uniform. Let also $\mathcal{U}_m^{-1} : (0, 1) \to \mathbb{R}$ be a quantile function, $\forall m = 1, 2, 3$, $F_m(\eta_{mi}) : \mathbb{R} \to [0, 1]$ a univariate cdf, $F\left(\mathcal{U}_1^{-1}\{F_1(\eta_{1i})\}, \mathcal{U}_2^{-1}\{F_2(\eta_{2i})\}, \mathcal{U}_3^{-1}\{F_3(\eta_{3i})\}\right)$ a joint cdf, and η_{mi} an additive predictor (made up of regression coefficients and covariates as described in Section 2.2) for $i = 1, \ldots, n$, where n denotes the sample size. Then there exists a three-dimensional copula function $\mathcal{C} : [0, 1]^3 \to [0, 1]$ defined as

$$\mathcal{C}(F_1(\eta_{1i}), F_2(\eta_{2i}), F_3(\eta_{3i})) = F\left(\mathcal{U}_1^{-1}\{F_1(\eta_{1i})\}, \mathcal{U}_2^{-1}\{F_2(\eta_{2i})\}, \mathcal{U}_3^{-1}\{F_3(\eta_{3i})\}\right), \tag{1}$$

which satisfies: (C.1) $\mathcal{C}(F_1(\eta_{1i}), 1, 1) = F_1(\eta_{1i})$, $\mathcal{C}(1, F_2(\eta_{2i}), 1) = F_2(\eta_{2i})$, $\mathcal{C}(1, 1, F_3(\eta_{3i})) = F_3(\eta_{3i})$, $\forall F_m(\eta_{mi}) \in [0, 1]$ and $m \leq 3$; (C.2) $\mathcal{C}(F_1(\eta_{1i}), F_2(\eta_{2i}), F_3(\eta_{3i})) = 0$ if $F_m(\eta_{mi}) = 0$ for any $m \leq 3$; and (C.3) $\mathcal{C}$ is 3-increasing (Sklar, 1959). Condition (C.1) states that if the realizations of two variables are known each with marginal probability of one, then the joint probability of the three outcomes is the same as the probability of the remaining uncertain outcome. Condition (C.2) is sometimes referred to as the grounded property of a copula and states that the joint probability of all outcomes is zero if the marginal probability of any outcome is zero. Condition (C.3) means that the copula volume of any 3-dimensional interval is non-negative. A copula $\mathcal{C}$ is unique on the cartesian product of the ranges of the marginal cdfs $\mathbf{Ran}(F_1(\eta_{1i})) \times \mathbf{Ran}(F_2(\eta_{2i})) \times \mathbf{Ran}(F_3(\eta_{3i}))$. The copula is unique if the margins are continuous. Any copula lies always in the interval

$$\max\left\{\sum_{m=1}^{3} F_m(\eta_{mi}) - 2, 0\right\} \leq \mathcal{C}(F_1(\eta_{1i}), F_2(\eta_{2i}), F_3(\eta_{3i})) \leq \min\{F_1(\eta_{1i}), F_2(\eta_{2i}), F_3(\eta_{3i})\},$$

the so-called *Fréchet–Hoeffding* bounds. A desirable feature of a copula is that it should cover the sample space between the lower and upper bounds, and that as the association parameters approach the lower (upper) bound of their permissible ranges, the copula approaches the Fréchet–Hoeffding lower (upper) bound. Knowledge of the Fréchet–Hoeffding bounds is therefore important in selecting an appropriate copula. For more details see, for instance, Trivedi and Zimmer (2007) and references therein.

In this paper, we employ the trivariate Gaussian copula with dependence structure characterized by coefficients $\vartheta_{12,i}$, $\vartheta_{13,i}$ and $\vartheta_{23,i}$ forming the model's correlation matrix $\boldsymbol{\Sigma}_i$. Based on (1), we express the trivariate Gaussian copula as $\Phi_3\left(\Phi^{-1}\{F_1(\eta_{1i})\}, \Phi^{-1}\{F_2(\eta_{2i})\}, \Phi^{-1}\{F_3(\eta_{3i})\}; \mathbf{0}, \boldsymbol{\Sigma}_i\right)$, where Φ^{-1} is the quantile function of a standard normal, $F_m(\eta_{mi})$ is derived in this case from either the standardized normal, logistic or Gumbel univariate cdf which is respectively defined as

$$F_m(\eta_{mi}) = \Phi(\eta_{mi}), \quad F_m(\eta_{mi}) = \frac{\exp(\eta_{mi})}{1 + \exp(\eta_{mi})} \quad \text{and} \quad F_m(\eta_{mi}) = 1 - \exp\{-\exp(\eta_{mi})\},$$