MOX–Report No. 39/2013

A semiparametric bivariate probit model for joint modeling of outcomes in STEMI patients

Ieva, F.; Marra, G.; Paganoni, A.M.; Radice, R.

MOX, Dipartimento di Matematica "F. Brioschi"
Politecnico di Milano, Via Bonardi 9 - 20133 Milano (Italy)

mox@mate.polimi.it http://mox.polimi.it

A semiparametric bivariate probit model for joint modeling of outcomes in STEMI patients

Francesca Ieva[1], Giampiero Marra[2], Anna Maria Paganoni[1], Rosalba Radice[3]

[1] MOX - Modeling and Scientific Computing, Mathematical Department, Politecnico di Milano, via Bonardi 9, 20133 Milano (Italy)

[1]Department of Statistical Science, University College London, Gower Street, London WC1E 6BT, U.K.

[3]Department of Economics, Mathematics and Statistics, Birkbeck, University of London, Malet Street, London WC1E 7HX, U.K.

Keywords: Recursive Bivariate Probit Models, Myocardial Infarction, Multiple Outcomes, Clinical Registry

Abstract

In this work we analyse the relationship among in-hospital mortality and a treatment effectiveness outcome in patients affected by ST-Elevation Myocardial Infarction. The main idea is to carry out a joint modelling of the two outcomes applying a Semiparametric Bivariate Probit Model to data arising from a clinical registry called STEMI Archive. A realistic quantification of the relationship between outcomes can be problematic for various reasons. First, latent factors associated with hospitals organization can affect the treatment efficacy and/or interact with patient's condition at admission time, then they can influence the mortality outcome. Such factors can be hardly measurable. Thus, the use of classical estimation methods will clearly result in inconsistent and biased parameter estimates. Secondly, covariate-outcomes relationships can exhibit non-linear patterns. Provided that proper statistical methods for model fitting in such framework are available, it is possible to employ a simultaneous estimation approach to account for unobservable confounders. Such a framework can also provide flexible covariate structures and model the whole conditional distribution of the response.

1 Introduction

Multiple outcomes are often used to properly characterize an effect of interest. Nevertheless, it often happens that the outcome of main interest is difficult or even impossible to measure. In general, realistic

quantification of the effect of a predictor of interest on a particular response variable can be a difficult task in statistical analysis based on observational data. A solution is to control for confounders, i.e. variables that are associated with both covariates and response. However, important confounders may be either unknown or too expensive to measure or not easily quantifiable (*unobservable confounders*). As pointed out in Sobotka et al. (2013), this problem, which is known as *endogeneity* of the explanatory variable of interest, poses serious limitations to covariate adjustment since the use of classical techniques will yield biased and inconsistent estimates. Further issues which deserve attention are the possible presence of non-linear covariate response relationships, and how these change when considering the whole response variable distribution.

Instrumental variable techniques are widely used for isolating the effect of a given predictor in the presence of unobserved confounding (e.g Marra & Radice, 2011b; Wooldridge, 2010, and references therein), and are increasingly used in epidemiological and medical studies (e.g. Goldman et al., 2001). In the context of binary responses, it is well known, from both theoretical and empirical results, that bivariate likelihood estimation methods are superior to conventional two-stage instrumental variable procedures (e.g. Bhattacharya et al., 2006; Wooldridge, 2010). First introduced by Heckman (1978), the recursive bivariate probit model represents an effective way to estimate the effect a binary regressor has on a binary outcome in the presence of unobservables. The semiparametric version of Heckman's model is an important extension since undetected nonlinearity can have severe consequences on the estimation of covariate effects (e.g. Marra & Radice, 2011a). Marra & Radice (2011a) proposed a penalized maximum likelihood fitting procedure to estimate the recursive bivariate probit model with non-linear confounder-response relationships.

The motivating problem of this work arises from the clinical context, that is a context where multiple outcomes often are used in order to characterize the patient's status or the performances of healthcare service with respect to patients' management. This is a framework where unobservable confounders are very popular as well.

In clinical context, during the last decade, the increased capability of data collection has made available a huge amount of information about procedures and outcomes. More and more often multiple outcomes are measured in order to characterize treatment effectiveness or to evaluate the impact of large policy initiatives. The case study considered in the following concerns patients affected by ST-Elevation Myocardial Infarction (STEMI) and admitted to any hospital of Lombardia, the Italian regional district whose capital is Milan. Data come from a clinical registry named STEMI Archive (Lombardia, 2009;

Ieva, 2013), which is a result of a wider comprehensive project (The Strategic Program "Exploitation, integration and study of current and future health databases in Lombardia for Acute Myocardial Infarction"). This project is funded by Italian Ministry of Health. Its main goal is to enhance the integration of different sources of health information in order to automate and streamline clinicians' workflow, so that data collected once can be used multiple times for different aims. Specifically, they can serve for measuring performances of healthcare systems, for understanding how hospitals work and for increasing efficacy of healthcare offer in terms of costs and patterns of care.

In general, clinical registries and administrative databases are more and more often used nowadays to answer epidemiological enquires (see Saia et al., 2009; Hasday et al., 2002; Dalby et al., 2003). The idea is to use information collected possibly with different purposes in order to analyze the efficacy and efficiency of the healthcare system on patients' outcomes. Thus, integrated healthcare systems for data collection measuring multiple outcomes play a fundamental role in complex clinical environments.

In such framework, the STEMI Archive consists of a clinical collection of data related to patients admitted in all hospitals of Regione Lombardia with STEMI diagnosis. One of the innovative contents of this survey is represented by process indicators recorded in it. They can be used to evaluate treatment times with the aim of designing a preferential therapeutic path to reperfusion in STEMI patients. In this sense, this survey represents an instrument both for epidemiological enquiries and for organizational optimization of the cardiological healthcare networks, quantifying the policies effects on multiple outcomes measured at patient's level. Within the data available from the STEMI Archive, there are two binary outcomes of interest: in-hospital mortality and reperfusion efficacy. The first one indicates if a patient is discharged alive from hospital. The second one indicates if a reduction greater than 70% of the ST-segment[1] elevation in the Electrocardiographic signal has been achieved after 1 hour from the reperfusion procedure, i.e., the primary angioplasty (Percutaneous Transluminal Coronary Angioplasty or PTCA). These outcomes are clearly correlated and the interest lies in their joint modelling in order to accomplish a manyfold goal:

1. performance evaluation (in terms of in-hospital survival) of the healthcare structure the patients are admitted to;

2. quantification of the influence of procedural variables on outcomes: how do the management of the patterns of care affects the quality of life after discharge?

[1]The ECG signal can be divide into different waves and segments, delimited by some relevant points (landmarks), listed in alphabetical order starting from letter P. The P wave represents atrial depolarization, the ventricular depolarization causes the QRS complex, that is followed by the ST segment. The ventricular depolarization is responsible of the T and the U waves. See Lindsay (2006), among others, for details.

3. evaluation of the relationship between outcomes, i.e., success in reperfusion practices and mortality, taking advantage by the joint modeling of their dependence on categorical and continuous covariates.

The article is then organized as follows: in Section 2 we present the recursive bivariate probit model proposed by Marra & Radice (2011a), highlighting the aspects that make such a model particulary suitable for carrying out the analysis of STEMI Archive data; in Section 3 the case-study is described and results of the analysis are proposed. Section 4 contains the discussion of results and conclusions. All the analyses have been carried out using `R` software, version 2.15.3, and in particular we refer to the R-package `SemiParBIVProbit`, presented in Marra & Radice (2012).

2 Recursive bivariate probit model

The bivariate probit model is a natural extension of probit regression model, where the disturbances of the two equations are assumed to be correlated in the same spirit as the seemingly unrelated regression model (Greene, 2012). The recursive version of the bivariate probit allows us to estimate the effect of interest while accounting for unobserved confounders (Maddala, 1983). The general specification is

$$\left.\begin{array}{l} y_{1i}^* = \mathbf{x}_{1i}^{\mathsf{T}}\boldsymbol{\alpha}_1 + \varepsilon_{1i} \\ y_{2i}^* = \gamma y_{1i} + \mathbf{x}_{2i}^{\mathsf{T}}\boldsymbol{\alpha}_2 + \varepsilon_{2i} \end{array}\right. , \; i = 1, \ldots, n, \tag{1}$$

where n denotes the sample size, y_{1i}^* and y_{2i}^* are continuous latent variables which determine the observed binary outcomes y_{1i} and y_{2i} through the rule $1(y_{vi}^* > 0)$, for $v = 1, 2$. Moreover, $\mathbf{x}_{1i}^{\mathsf{T}} = (1, x_{12i}, \ldots, x_{1P_1 i})$ is the ith row vector of the $n \times P_1$ model matrix $\mathbf{X}_1$, and $\boldsymbol{\alpha}_1$ is a parameter vector. Similarly, $\mathbf{x}_{2i}^{\mathsf{T}}$ is the ith row vectors of the $n \times P_2$ model matrix $\mathbf{X}_2$, $\boldsymbol{\alpha}_2$ is a coefficient vector and γ is the parameter of the endogenous binary variable y_{1i}. The error terms $(\varepsilon_{1i}, \varepsilon_{2i})$ are assumed to follow the distribution $\mathcal{N}([0,0],[1,\rho,\rho,1])$, where ρ is the correlation coefficient and the error variances are normalized to unity since the parameters in the model can only be identified up to a scale coefficient (e.g. Greene, 2012). To identify the parameters of the second equation in (1), it is typically assumed that the exclusion restriction on the exogenous variables holds. That is, the covariates in the first equation should contain at least one or more regressors (usually referred to as *instruments*) not included in the second equation. These regressors have to induce variation in y_{1i}, have not to directly affect y_{2i}, and have to be independent of $(\varepsilon_{1i}, \varepsilon_{2i})$ given covariates. However, as shown in Han & Vytlacil (2013), Marra & Radice (2011a) and Wilde (2000), the presence of this restriction may not be necessary to obtain consistent estimates of the model parameters.

Marra & Radice (2011a) proposed an extension of this model which allows for flexible functional dependence of the responses on continuous covariates: the semiparametric recursive bivariate probit model. This extension is important because the neglect of the presence of nonlinearity may have severe consequences on the estimation of covariate effects (Chib & Greenberg, 2007; Marra & Radice, 2011a). The semiparametric version of the classic bivariate probit can be written as

$$\begin{aligned} y_{1i}^* &= \tilde{\mathbf{x}}_{1i}^\mathsf{T}\boldsymbol{\delta}_1 + \sum_{k_1=1}^{K_1} s_{1k_1}(\breve{x}_{1k_1i}) + \varepsilon_{1i} \\ y_{2i}^* &= \gamma y_{1i} + \tilde{\mathbf{x}}_{2i}^\mathsf{T}\boldsymbol{\delta}_2 + \sum_{k_2=1}^{K_2} s_{2k_2}(\breve{x}_{2k_2i}) + \varepsilon_{2i} \end{aligned} \quad , \; i = 1, \ldots, n, \tag{2}$$

where $\tilde{\mathbf{x}}_{1i}^\mathsf{T} = (1, \tilde{x}_{12i}, \ldots, \tilde{x}_{1Q_1i})$ is the ith row vector of $\tilde{\mathbf{X}}_1$, the $n \times Q_1$ model matrix for the parametric model components (such as intercept, dummy and categorical variables), with corresponding parameter vector $\boldsymbol{\delta}_1$, and the s_{1k_1} are unknown smooth functions of the K_1 continuous covariates $\breve{x}_{1k_1i}$. Similarly, $\breve{\mathbf{x}}_{2i}^\mathsf{T}$ is the ith row vectors of the $n \times Q_2$ model matrix $\breve{\mathbf{X}}_2$, with coefficient vector $\boldsymbol{\delta}_2$, and the s_{2k_2} are unknown smooth terms of the K_2 continuous regressors $\breve{x}_{2k_2i}$. Smooth terms are subject to identifiability constraints such as $\sum_i s_{vk_v}(\breve{x}_{vk_vi}) = 0$, $v = 1, 2$, $k_v = 1, \ldots, K_v$. The smooth functions are approximated using regression splines (e.g. Wood, 2006). Here, function $s_k(\breve{x}_{ki})$, where subscript v has been suppressed to avoid clutter, is given by $\sum_{j=1}^{J_k} \beta_{kj} b_{kj}(\breve{x}_{ki}) = \mathbf{b}_k(\breve{x}_{ki})^\mathsf{T}\boldsymbol{\beta}_k$, where the $b_{kj}(\breve{x}_{ki})$ are known spline basis functions, with corresponding regression parameters β_{kj}, J_k is the number of spline bases, $\mathbf{b}_k(\breve{x}_{ki})$ is a vector consisting of the basis functions evaluated at $\breve{x}_{ki}$, i.e., $\mathbf{b}_k(\breve{x}_{ki})^\mathsf{T} = \{b_{k1}(\breve{x}_{ki}), \ldots, b_{kJ_k}(\breve{x}_{ki})\}$, and $\boldsymbol{\beta}_k$ is the corresponding parameter vector. Basis functions are typically chosen to have convenient mathematical properties and good numerical stability. Possible choices include B-splines, cubic regression and thin plate regression splines (see, e.g., Marra & Radice (2010) for an overview). Based on this representation, the equations in (2) can written as $y_{1i}^* = \tilde{\mathbf{x}}_{1i}^\mathsf{T}\boldsymbol{\delta}_1 + \mathbf{b}_{1i}^\mathsf{T}\boldsymbol{\beta}_1 + \varepsilon_{1i} = \eta_{1i} + \varepsilon_{1i}$, and $y_{2i}^* = \gamma y_{1i} + \tilde{\mathbf{x}}_{2i}^\mathsf{T}\boldsymbol{\delta}_2 + \mathbf{b}_{2i}^\mathsf{T}\boldsymbol{\beta}_2 + \varepsilon_{2i} = \eta_{2i} + \varepsilon_{2i}$, where, for $v = 1, 2$, $\mathbf{b}_{vi}^\mathsf{T} = \left\{\mathbf{b}_{v1}(\breve{x}_{v1i})^\mathsf{T}, \ldots, \mathbf{b}_{vK_v}(\breve{x}_{vK_vi})^\mathsf{T}\right\}$, $\boldsymbol{\beta}_v^\mathsf{T} = (\boldsymbol{\beta}_{v1}^\mathsf{T}, \ldots, \boldsymbol{\beta}_{vK_v}^\mathsf{T})$ and η_{vi} has the obvious definition.

2.1 Estimation and Inference

In the bivariate probit model the data identify the four possible events $(y_{1i} = 1, y_{2i} = 1)$, $(y_{1i} = 1, y_{2i} = 0)$, $(y_{1i} = 0, y_{2i} = 1)$ and $(y_{1i} = 0, y_{2i} = 0)$ with probabilities $p_{11i} = \Phi_2(\eta_{1i}, \eta_{2i}; \rho)$, $p_{10i} = \Phi(\eta_{1i}) - p_{11i}$, $p_{01i} = \Phi(\eta_{2i}) - p_{11i}$ and $p_{00i} = 1 - p_{11i} - p_{10i} - p_{01i}$, where Φ and Φ_2 are the distribution functions of a standardized univariate normal and a standardized bivariate normal with correlation ρ, respectively.

Therefore, the log-likelihood function is

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \{y_{1i}y_{2i}\log(p_{11i}) + y_{1i}(1-y_{2i})\log(p_{10i}) + (1-y_{1i})y_{2i}\log(p_{01i}) + (1-y_{1i})(1-y_{2i})\log(p_{00i})\}, \tag{3}$$

where $\boldsymbol{\theta}^{\mathsf{T}} = (\boldsymbol{\delta}_1^{\mathsf{T}}, \boldsymbol{\beta}_1^{\mathsf{T}}, \gamma, \boldsymbol{\delta}_1^{\mathsf{T}}, \boldsymbol{\beta}_2^{\mathsf{T}}, \rho)$ according to the notation introduced in the previous section. When using regression splines, to avoid overfitting, the model parameters are typically estimated by maximization of

$$\ell_p(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}) - \frac{1}{2}\boldsymbol{\beta}^{\mathsf{T}}\mathbf{S}_{\boldsymbol{\lambda}}\boldsymbol{\beta}, \tag{4}$$

where $\boldsymbol{\beta}^{\mathsf{T}} = (\boldsymbol{\beta}_1^{\mathsf{T}}, \boldsymbol{\beta}_2^{\mathsf{T}})$, $\mathbf{S}_{\boldsymbol{\lambda}} = \sum_{v=1}^{2}\sum_{k_v=1}^{K_v}\lambda_{vk_v}\mathbf{S}_{vk_v}$ and the $\mathbf{S}_{vk_v}$ are positive semi-definite known square matrices measuring the (second-order, here) roughness of the smooth terms in the model, i.e. $\boldsymbol{\beta}^{\mathsf{T}}\left(\sum_{v=1}^{2}\sum_{k_v=1}^{K_v}\lambda_{vk_v}\mathbf{S}\right.$ $\sum_{v=1}^{2}\sum_{k_v=1}^{K_v}\lambda_{vk_v}\int f''_{vk_v}(\breve{x}_{vk_v})^2 d\breve{x}_{vk_v}$. The λ_{vk_v} are smoothing parameters controlling the trade-off between fit and smoothness. Given values for λ_{vk_v}, maximization of (4) is straightforward. However, smoothing parameter estimation has to be settled in practice. This usually involves the use of specialized numerical routines minimizing, for instance, a prediction error criterion so that the estimated smooth functions are as close as possible to the true functions. In the current context, multiple smoothing parameter estimation is achieved by minimization of the approximate unbiased risk estimator (Craven & Wahba, 1979). Full computational details can be found in Marra & Radice (2011a).

The inferential theory for models involving penalized regression splines is not standard. This is because of the presence of penalties which undermines the use of classic asymptotic results for practical modelling. As explained in Marra & Radice (2011a), confidence intervals (CIs) for the components in the semiparametric bivariate probit model can be constructed using the results for the well known Bayesian 'confidence' intervals typically employed in a generalized additive model context (e.g. Gu, 2002). The resulting intervals include both a bias and variance component, a fact that makes such intervals have good observed *frequentist* coverage probabilities across the function (Marra & Wood, 2012). Interval calculations are therefore based on $\boldsymbol{\theta}|\mathbf{y} \dot{\sim} \mathcal{N}(\hat{\boldsymbol{\theta}}, \mathbf{V}_{\boldsymbol{\theta}})$, where $\mathbf{y}$ contains the response vectors, $\hat{\boldsymbol{\theta}}$ is the estimate of $\boldsymbol{\theta}$, and $\mathbf{V}_{\boldsymbol{\theta}}$ represents the inverse of the penalized information matrix obtained at convergence of the algorithm (see Marra & Radice (2011a) for further details). Given this result, CIs for linear and nonlinear functions of the model parameters can be easily obtained. Note that, for parametric model components, using the result above is equivalent to using classic likelihood results since such terms are not penalized. Also, there is no contradiction in fitting model (2) by penalized log-likelihood estimation and then constructing CIs adopting a Bayesian approach, and such a procedure has been employed many times in the literature (e.g. Gu, 2002; Wood, 2006).

Note that in the absence of smooth functions in the model, as in (1), classic unpenalised maximum likelihood estimation can be reliably employed and traditional frequentist results used for inference. However, it is not possible to know whether smooth components are required before the analysis. In fact, using a more flexible model can help reducing the risk of misspecification due to undetected nonlinearity, which, as mentioned earlier, can have severe consequences on parameter estimation (Chib & Greenberg, 2007; Marra & Radice, 2011a).

2.2 Average treatment effect

Since latent variables do not typically have well-defined unit of measurements, parameter γ in model (1) may not be interpretable. For this reason the effect of the treatment y_{1i} on the response probability $P(y_{2i} = 1|y_{1i})$ is calculated. This can be done using the average treatment effect (ATE; e.g. Wooldridge, 2010). Given estimates for the model components, the ATE can be estimated as follows

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\Phi_2\left(\hat{\eta}_{2i}^{(y_{1i}=1)},\hat{\eta}_{1i};\hat{\rho}\right)}{\Phi(\hat{\eta}_{1i})}-\frac{\Phi_2\left(\hat{\eta}_{2i}^{(y_{1i}=0)},-\hat{\eta}_{1i};-\hat{\rho}\right)}{1-\Phi(\hat{\eta}_{1i})},$$

where $\hat{\eta}_{2i}^{(y_{1i}=r)}$ indicates the linear predictor evaluated at r equal to 1 or 0.

Coefficient ρ is also of interest as it is useful to ascertain the presence of unobserved confounding (endogeneity). Specifically, ρ can be interpreted as the correlation between the unobserved confounders in the two equations (e.g. Monfardini & Radice, 2008). If $\rho = 0$ then ε_{1i} and ε_{2i} are uncorrelated and hence there is not a problem of endogeneity. In this case, estimation of the second equation in either (1) or (2) will yield consistent parameter estimates. Moreover, if the model (1) or (2) were fitted with intercepts only, it turns out that ρ is precisely the tetrachoric correlation, i.e., a Pearson correlation between two bivariate normal variables that have been observed on a dichotomous scale (Pearson, 1900). Confidence intervals can be obtained using the delta method (see Chiburis et al., 2011).

3 Case study

In this section we present the analyses carried out fitting a semiparametric bivariate probit model like in (2) to the data arising from STEMI Archive, the clinical registry gathering patients admitted with ST-segment Elevation Myocardial Infarction diagnosis in any hospital of Regione Lombardia district. A complete description of this clinical registry is provided in Ieva (2013), where the Archive is presented together with the motivating clinical setting. Among the most important patient information provided by this clinical registry there are:

- *mode of admission*, i.e., if a patient reaches the hospital on her/his own or delivered by three different types of rescue units of 118 (the national toll-free number for emergencies);

- *demographic features*, like age and sex;

- *clinical appearance*, i.e., variables describing the patient's status at admission. Among others, we focus on killip class (binary variable categorizing the severity of infarction into 0 = less severe and 1 = more severe) and Ejection Fraction (EF);

- *risk factors*, like hypertension, diabetes, smoking and Chronic Kidney Disease (CKD);

- *times to treatment* (on/off hours), *times to treatment*, *times to intervention* and all the *process indicators* concerned with pre- and in-hospital phase (Symptom Onset to Door time - OD, Door to Balloon time - DB, total ischaemic time - OB, etc.);

- *clinical outcomes*, i.e., in-hospital mortality and treatment efficacy (STres), quantified by a reduction of ST segment elevation in the ECG.

We focus our readings on patients who underwent PTCA (Primary Transluminal Coronary Angioplasty), the most common reperfusion procedure for Acute Myocardial Infarctions. The population considered for the following analyses consists of 1069 statistical units.

In this application, the binary outcomes of interest are patients' in-hospital mortality and the efficacy of the reperfusion treatment they undergo. The efficacy is determined by the reduction of ST segment elevation one hour after the surgery: if the reduction is over 70% the procedure is considered effective. Thus it is clear why the joint modelling of in-hospital mortality and reperfusion efficacy makes sense. Not only they are likely to be correlated, but a strong clinical interest lies in quantifying the degree of correlation among these two. Moreover, reperfusion efficacy indicator is a binary variable whose values depend on the latent recovered ability of the coronary arteries to work properly. So the framework presented in Marra & Radice (2011a) seems to be the proper way to address the problem of interest.

The variable selection has been carried out according to both clinical knowhow and the statistical stepwise approach, similarly to what proposed in Ieva & Paganoni (2011); Grieco et al. (2012) and Guglielmi et al. (2012, 2013). Then the model (2) has been fitted to STEMI Archive data with the following specifications:

1. for the outcome $\mathbf{y_1^*}$ (the *reperfusion efficacy*, `STres`) we retained a binary variable ($\tilde{x}_{11}$) indicating if the patient reaches the hospital on her/his own (`access`), and two continuous variables ($\breve{x}_{11}$ and

$\breve{x}_{12}$), being the age (`age`) of the patient and her/his total ischaemic time (`O2B`, i.e., the time between the symptom onset and the PTCA procedure), respectively;

2. for the outcome $\mathbf{y_2^*}$ (the *in-hospital mortality*, `mortality`) we retained a categorical variable ($\tilde{x}_{21}$) indicating the patient's killip class (`killip`), and a continuous variable ($\breve{x}_{21}$) measuring her/his ejection fraction (`EF`) at the entrance.

Therefore, for $i = 1, \ldots, 1069$, model (2) becomes:

$$\begin{aligned} &\text{Equation 1:} \quad \texttt{STres}^*_{1i} = \delta_{10} + \delta_{11} \times \texttt{access}_i + \delta_{12} \times \texttt{age}_i + s_{11}(\texttt{O2B}_{1i}) + \varepsilon_{1i} \\ &\text{Equation 2:} \quad \texttt{Mortality}^*_{1i} = \delta_{20} + \delta_{21} \times \texttt{STres} + \delta_{22} \times \texttt{killip} + \delta_{23} \times \texttt{EF} + \varepsilon_{2i} \end{aligned} \tag{5}$$

Table 1 shows the estimates provided by the bivariate probit model for the mortality outcome (second equation of (5), lower panel in the Table) and the indicator of successful reperfusion therapy (first equation of (5), upper panel in the Table).

		Coefficient	estimate	std. err.	p-val
equation 1	`intercept`	δ_{10}	1.3811	0.2488	< 0.0000
	`access`	δ_{11}	0.2129	0.0933	0.0225
	`age`	δ_{12}	-0.0088	0.0036	0.0140
		Smooth term	edf	est. rank	p-val
	s(`O2B`)	s_{11}	1.356	2	0.0041

		Coefficient	estimate	std. err.	p-val
equation 2	`intercept`	δ_{20}	1.7708	0.4656	0.0001
	`STres`	δ_{21}	-1.2480	0.2109	< 0.0000
	`killip`	δ_{22}	0.7223	0.2544	0.0045
	`EF`	δ_{23}	-0.0748	0.0119	< 0.0000

Table 1: Parametric and smoothed coefficients' estimates obtained fitting the semiparametric bivariate PROBIT model in (2) to STEMI Archive data.

It can be noticed that all the selected covariates are significant. In particular, the treatment efficacy decrases as the age increases, as expected. The way of admission of patients delivered by 118 eases the good prognosis, too. It is worth noting that the total ischaemic time effect is nonlinear, being the smoother degrees of freedom significantly greater than one. This confirms the clinical knowhow according to which the way the delay affects the treatment efficacy is definitely nonlinear Gersh et al. (2005).

Concerning the mortality outcome, as expected the more severe the infarction (quantified by killip class), the higher the mortality. Also a reduced ejection fraction plays the role of increasing the mortality, as it is known by clinial practice.

The estimated correlation coefficient of the recursive bivariate probit model is equal to 0.394 and it is significantly different from zero at the 5% level ($\text{IC}(\rho) = (0.0637, 0.644)$), hence supporting the presence

of unobserved confounders. In fact, it is usual that a lot of unexplained variability exists in complex healthcare processes where patterns of care consist of multiple phases. It derives from the variability existing at patient's level plus a variability induced by the complex process of patient's management. Then it is crucial to find correlations and to identify which procedures clinicians can act upon in order to improve the process of care.

Table 2 shows ATE estimates obtained using the bivariate probit model (upper line) and the naive model (lower line). Since the naive model does not account for unobserved confounding, in this case the ATE has been estimated by fitting the equation of interest alone using univariate probit regression.

ATE of	estimate	CI
SBP model	-2.05	(-4.88,0.77)
naive model	-1.92	(-20.63,16.78)

Table 2: ATE estimates obtained by fitting the semiparametric bivariate probit and the naive approach, respectively

It can be observed that the confidence intervals of ATE strongly modify if we take into account the unobservables. It is reasonable to expect ATE to be negative, since the better the efficacy, the lower the mortality probability. Then, even if there is no evidence for considering the point estimates different from 0, in both cases the estimates are concentrated on negatives values. Moreover, the modification of the CIs seems to suggest that with a greater number of observation, our tenet on ATE could be confirmed. Anyway, in this case results suggest that the presence of unobserved confounders detected by the bivariate model may be regarded as variables which do not interfere with the relationship of interest but whose presence inflates the variance of the estimates.

In this study we were concerned with the possible detrimental effect of unobserved confounders on the effect of interest (reperfusion efficacy on mortality). Based on our analysis, this does not seem to be an issue as the SBP and naive models produced similar point estimates. However, the use of a bivariate probit model may still be preferred as it may allow for more reliable inferences.

4 Conclusions

It is more and more frequent in clinical practice that multiple outcomes are measured for properly characterizing an effect of interest in terms of diseases or for assessing healthcare policies and performances. Nevertheless, it often happens that (some) outcomes are difficult (even impossible) to be measured, or that confounders are difficult to be accounted for when modelling such outcomes by means of suitable covariates. Instrumental variables are nowadays an established method for isolating the effect of a given

predictor in the presence of unobserved confounding.

In this work we showed an application of a semiparametric bivariate probit model to a couple of binary outcomes representing the in-hospital mortality and an indicator of the reperfusion efficacy in patients affected by Acute Myocardial Infarction. The efficacy is determined by the reduction of ST segment elevation one hour after the surgery: if the reduction is over 70% the procedure is considered effective. Data come from a clinical registry called STEMI Archive Lombardia (2009). This case study claims for the joint modelling of the in-hospital mortality and reperfusion efficacy outcomes. It makes sense not only because they are likely to be correlated, but a strong clinical interest lies in quantifying the degree of correlation among these two. Moreover, reperfusion efficacy indicator is a binary variable whose values depend on the latent recovered ability of the coronary arteries to work properly. In this sense, this modelling strategy represents a step forward with respect to the results pointed out in Ieva & Paganoni (2011) and then in (Ieva & Paganoni, 2010; Grieco et al., 2012), since a joint modelling of correlated outcomes is possible, as well as parametric and nonparametric definition of the relationship between outcomes and covariates.

Results are strongly coherent with clinical practice. This enables a better comprehension of the disease-recovery dynamics, and enables better predictions for new patients entering the study. In general, accounting and adjusting for confounders is extremely important in complex processes such clinical ones are, since so many source of latent interaction may arise. This appears clearly looking at results reported in Table 2.

In general, diagnosis and management of AMI patients are difficult and may strongly benefit of the aid of statistical models that provide effective risk stratification of patients. In fact, flexible models that are able to properly profile patients adjusting for case mix and confounders are extremely of interest in the context of modern clinical practice, since the more accurate predictions and more reliable prognoses they provide enable to gain insights of the economic burden of AMI, supporting an effective clinical decision making.

References

Bhattacharya, J., Goldman, D., & McCaffrey, D. (2006). Estimating probit models with self-selected treatments. *Statistics in Medicine*, 25, 389–413.

Chib, S. & Greenberg, E. (2007). Semiparametric modeling and estimation of instrumental variable models. *Journal of Computational and Graphical Statistics*, 16, 86–114.

Chiburis, R. C., Das, J., & Lokshin, M. (2011). A practical comparison of the bivariate probit and linear iv estimators. *World Bank Policy Research Working Paper 5601*.

Craven, P. & Wahba, G. (1979). Smoothing noisy data with spline functions. *Numerische Mathematik*, 31, 377–403.

Dalby, M., Bouzamondo, A., Lechat, P., & Montalescot, G. (2003). Transfer for primary angioplasty versus immediate thrombolysis in acute myocardial infarction: a meta-analysis. *Circulation*, 108(15), 1809–1814.

Gersh, B. J., Stone, G. W., White, H. D., & Holmes, D. R. (2005). Pharmacological facilitation of primary percutaneous coronary intervention for acute myocardial infarction: Is the slope of the curve the shape of the future? *Journal of the American Medical Association*, 293, 979–986.

Goldman, D., Bhattacharya, J., McCaffrey, D., Duan, N., Leibowitz, A., Joyce, G., & Morton, S. (2001). Effect of insurance on mortality in an hiv-positive population in care. *Journal of the American Statistical Association*, 96, 883–894.

Greene, W. H. (2012). *Econometric Analysis*. Prentice Hall, New York.

Grieco, N., Ieva, F., & Paganoni, A. M. (2012). Performance assessment using mixed effects models: a case study on coronary patient care. *IMA Journal of Management Mathematics*, 23, 117–131.

Gu, C. (2002). *Smoothing Spline ANOVA Models*. London: Springer-Verlag.

Guglielmi, A., Ieva, F., Paganoni, A. M., & Ruggeri, F. (2012). A bayesian random effects model for survival probabilities after acute myocardial infarction. *Chilean Journal of Statistics*, 3, 1–15.

Guglielmi, A., Ieva, F., Paganoni, A. M., Ruggeri, F., & Soriano, J. (2013). Semiparametric bayesian modeling for the classification of patients with high observed survival probabilities. *Journal of the Royal Statistical Society - Series C*, Forthcoming.

Han, S. & Vytlacil, E. J. (2013). Identification in a generalization of bivariate probit models with endogenous regressors. *Working paper*.

Hasday, D., Behari, S., & Wallentini, L. (2002). A prospective survey of the characteristics, treatments and outcomes of patients with acute coronary syndromes in europe and the mediterranean basin. the euro heart survey of acute coronary syndromes (euro heart survey acs). *European Heart Journal*, 23(15), 1190–1210.

Heckman, J. (1978). Dummy endogenous variables in a simultaneous equation system. *Econometrica*, 46, 931–959.

Ieva, F. (2013). Designing and mining a multicenter observational clinical registry concerning patients with acute coronary syndromes. In N. Grieco, M. Marzegalli, & A. M. Paganoni (Eds.), *New diagnostic,*

therapeutic and organizational strategies for patients with Acute Coronary Syndromes (pp. 47–60).: Springer.

Ieva, F. & Paganoni, A. M. (2010). Multilevel models for clinical registers concerning stemi patients in a complex urban reality: a statistical analysis of momi^2 survey. *Communications in Applied and Industrial Mathematics*, 1, 128–147.

Ieva, F. & Paganoni, A. M. (2011). Process indicators for assessing quality of hospital care: a case study on stemi patients. *JP Journal of Biostatistics*, 6, 53–75.

Lindsay, A. (2006). Ecg learning centre.

Lombardia (2009). Determinazioni in merito alla rete per il trattamento dei pazienti con infarto miocardico con tratto st elevato (stemi).

Maddala, G. S. (1983). *Limited Dependent and Qualitative Variables in Econometrics*. Cambridge University Press, Cambridge.

Marra, G. & Radice, R. (2010). Penalised regression splines: theory and application to medical research. *Statistical Methods in Medical Research*, 19, 107–125.

Marra, G. & Radice, R. (2011a). Estimation of a semiparametric recursive bivariate probit model in the presence of endogeneity. *Canadian Journal of Statistics*, 39, 259–279.

Marra, G. & Radice, R. (2011b). A flexible instrumental variable approach. *Statistical Modelling*, 11, 581–279.

Marra, G. & Radice, R. (2012). *SemiParBIVProbit: Semiparametric Bivariate Probit Modelling*. R package version 3.2-1.

Marra, G. & Wood, S. (2012). Coverage properties of confidence intervals for generalized additive model components. *Scandinavian Journal of Statistics*, 39, 53–74.

Monfardini, C. & Radice, R. (2008). Testing exogeneity in the bivariate probit model: A monte carlo study. *Oxford Bulletin of Economics and Statistics*, 70, 271–282.

Pearson, K. (1900). Mathematical contributions to the theory of evolution. VII. on the correlation of characters not quantitatively measurable. *Philosophical Transactions. Royal Society of London. Series A. Mathematical and Physical Sciences*, 195, 1–47.

Saia, F., Marzocchi, A., Manari, G., Guastaroba, P., Vignali, L., & Varani, E. (2009). Patient selection to enhance the long-term benefit of first generation drug-eluting stents for coronary revascularization procedures: insights from a large multicenter registry. *Eurointervention*, 5(1), 57–66.

Sobotka, F., Radice, R., Marrai, G., & Kneib, T. (2013). Estimating the relationship between womens education and fertility in botswana by using an instrumental variable approach to semiparametric

expectile regression. *Journal of the Royal Statistical Society - Series C*, 62, 25–45.

Wilde, J. (2000). Identification of multiple equation probit models with endogenous dummy regressors. *Economics Letters*, 69, 309–312.

Wood, S. N. (2006). *Generalized Additive Models: An Introduction With R*. Chapman & Hall/CRC.

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

MOX Technical Reports, last issues

Dipartimento di Matematica "F. Brioschi",
Politecnico di Milano, Via Bonardi 9 - 20133 Milano (Italy)

39/2013 Ieva, F.; Marra, G.; Paganoni, A.M.; Radice, R.
A semiparametric bivariate probit model for joint modeling of outcomes in STEMI patients

38/2013 Bonizzoni, F.; Nobile, F.
Perturbation analysis for the Darcy problem with log-normal permeability

34/2013 Tavakoli, A.; Antonietti, P.F.; Verani, M.
Automatic computation of the impermeability of woven fabrics through image processing

35/2013 Cattaneo, L.; Formaggia, L.; Iori G. F.; Scotti, A,; Zunino, P.
Stabilized extended finite elements for the approximation of saddle point problems with unfitted interfaces

36/2013 Ferran Garcia, Luca Bonaventura, Marta Net, Juan Sanchez
Exponential versus IMEX high-order time integrators for thermal convection in rotating spherical shells

37/2013 Lever, V.; Porta, G.; Tamellini, L.; Riva, M.
Characterization of basin-scale systems under mechanical and geochemical compaction

33/2013 Menafoglio, A; Guadagnini, A; Secchi, P
A Kriging Approach based on Aitchison Geometry for the Characterization of Particle-Size Curves in Heterogeneous Aquifers

32/2013 Taddei, T.; Perotto, S.; Quarteroni, A.
Reduced basis techniques for nonlinear conservation laws

31/2013 Dassi, F.; Ettinger, B.; Perotto, S.; Sangalli, L.M.
A mesh simplification strategy for a spatial regression analysis over the cortical surface of the brain

30/2013 Cagnoni, D.; Agostini, F.; Christen, T.; de Falco, C.; Parolini, N.; Stefanovic, I.
Multiphysics simulation of corona discharge induced ionic wind