Health Care Utilization and Self-Assessed Health: Specification of Bivariate Models Using Copulas[(*)]

José M. R. Murteira
Faculdade de Economia, Universidade de Coimbra, and CEMAPRE

Óscar D. Lourenço
Faculdade de Economia, Universidade de Coimbra, and CEISUC

This version: November, 2009

Abstract

The discernment of relevant factors driving health care utilization constitutes one important research topic in Health Economics. This issue is frequently addressed through specification of regression models for health care use (*y* – often measured by number of doctor visits) including, among other covariates, a measure of self-assessed health (*sah*). However, the exogeneity of *sah* within those models has been questioned, due to the possible presence of unobservables influencing both *y* and *sah*, and because individuals' health assessments may depend on the quantity of medical care received.

This paper addresses the possible simultaneity of (*sah*,*y*) by adopting a full information approach, through specification of the bivariate probability function (p.f.) of these discrete variables, conditional on a set of exogenous covariates (*x*). The approach is implemented with copula functions, which afford separate consideration of each variable margin and their dependence structure. The specification of the joint p.f. of (*sah*,*y*) enables estimation of several quantities of potential economic interest, namely features of the conditional p.f. of *y* given sah and *x*. The adopted models are estimated through maximum likelihood, with cross-section data from the Portuguese National Health Survey of 1998/99. Estimates of the margins parameters do not vary much among different copula models, while, in accordance with theoretical expectations, the dependence parameter is estimated to be negative across the various joint models.

JEL classification code: I10, C16, C51.
Key Words: health care utilization; self-assessed health; endogeneity; discrete data; copulas.

[(*)] We thank John Mullahy, Maarten Lindeboom, João Santos Silva, and an anonymous referee for helpful comments. The usual disclaimer applies. Support from Fundação para a Ciência e Tecnologia, Program FEDER/POCI 2010 is gratefully acknowledged. Address for correspondence: José Murteira, Faculdade de Economia, Universidade de Coimbra, Av. Dias da Silva, 165, 3004-512 Coimbra, Portugal. E-mail: jmurt@fe.uc.pt.

1. Introduction

The area of health economics has witnessed a steady increase in research activity over the last decades. To some extent, this growing interest can be seen as a consequence of the volume and continuous rise of health care expenditures in most industrialized countries, Portugal included.[1] Not surprisingly, the modelling and estimation of medical care demand functions has constituted an important research topic in this area, namely with a view to discern the impact of such factors as price, income or health insurance status on health care utilization. Seminal work on these issues can be found in Newhouse, Phelps and Marquis (1980), and Wagstaff (1989), who were among the first to survey the econometric analysis of medical care demand. The estimation of demand functions has been frequently addressed through specification of regression models for health care use, often represented by a count, y, measuring the number of doctor visits.[2] Most of the regression models discussed in the literature include, besides other covariates (here denoted as x), an indicator of self-assessed health (*sah*), which is usually found to be a relevant regressor for the dependent variable of interest.

Frequently, these models are estimated by use of methods that rely on the assumption of regressors' exogeneity, *sah* included. Some authors, however, have cast doubt on the exogeneity of *sah* within such models, due to the accepted fact that individuals' health assessments are, to a significant extent, both subjective and influenced by the quantity of medical care previously received. Concern about the use of *sah*-type covariates in demand equations was detailed by Manning, Newhouse and Ware (1982), who advocate treating *sah* as endogenous in regressions for cross-section data on health care utilization, with consequential use of instrumental variables (see Manning, *et al*., 1982, p. 166). The issue has also been raised more recently, namely in, Windmeijer and Santos Silva (1997), Lourenço (2007b), and Van Ourti (2004).

As is well known, recognizing *sah* as endogenous within count data regressions, calls for estimation strategies, namely nonlinear instrumental variables (NLIV)

[1] In Portugal, according to the OECD Health Data (2006), the total expenditures on health, as share of GDP, increased from 7.3% in 1994 to 10.1% in 2004.

[2] Among many other examples, use of count data models in health economics can be found in Jones and O'Donnell (2002), Bago d'Uva (2006), Lourenço, Quintal, Ferreira and Barros (2007a), Deb and Trivedi (1997, 2002), and Winkelmann (2004).

or generalized method of moments (GMM), that require valid instruments (see, *e.g.* Cameron and Trivedi 1998, ch. 11). When no such variables are available – or when the endogeneity of *sah* is neglected – researchers often take to one of two avenues: either to exclude *sah* from the regression or to adopt a nonrobust estimation method – usually nonlinear least squares (NLS) or conditional maximum likelihood (ML). Clearly, either choice involves a considerable risk of producing inconsistent estimates – not only of the parameter associated with *sah*, but of all regression parameters associated with covariates not orthogonal to *sah*.

This paper addresses the likely simultaneity of (y, sah) and its associated consequences by specifying the joint probability function (p.f.) of (y, sah), conditional on a set of exogenous regressors (*x*). This full information approach can be implemented using copula functions (Sklar 1959). One advantage of copulas is that they enable separate consideration of the marginal distribution for each dependent variable, as well as their dependence structure. This flexibility makes it possible for researchers to capture the dependence structure of the data without knowing the exact form of the joint p.f., while, at the same time, preserving desirable characteristics of the chosen marginals for the response variables.

Modelling the joint (conditional) distribution of (y, sah), given *x*, constitutes a full information alternative to the more conventional approach of specifying a structural equation, or a set of structural simultaneous equations, linking the endogenous variables (y, sah) to a set of exogenous covariates, *x* (as, *e.g.*, in Windmeijer and Santos Silva 1997). Instead of specifying these structural equations, one can conceivably think of the joint probability law of $(y, sah)\,|\,x$, and try to model it by making use of the copula theory insight.

It should be noted, however, that this proposed alternative approach may not afford a mapping between the conventional structural parameters – namely the causal effect of *sah* on utilization – and the parameters of the joint p.f. of (y, sah), given *x*. This suspicion is fuelled in the present context by the fact that *y* is usually a discrete random variable and, as detailed below, *sah* is often treated as a rank variable. Clearly, their joint p.f. is far from bivariate normal, a situation in which, with suitable assumptions, one is usually able to trace back structural parameters from the parameters of the joint p.f. of endogenous variables. Indeed, an enquiry into the identification possibilities of the proposed methodology in the present context, as it relates to limited information approaches, is beyond the scope of the present text and would deserve a separate paper on its own.

In any case, the suggested approach enables identification of several entities of interest, such as features of the conditional p.f.'s (of either endogenous variable, given the other and *x*), including, *e.g.*, the conditional expectation $E(y|sah,x)$, for different *sah* values. This constitutes a prominent example of quantities of potential economic interest, representative of the influence of the individual's health status on medical care utilization. Naturally, one can think of other relevant quantities identified by the present approach, such as income-elasticities of (average) utilization, or the assessment of the effect of supplementary health insurance on average health care use. The suggested methodology is rich enough to permit the measurement of different quantities, according to specific research interests.

The paper is organized as follows. Section two details the main problem and surveys alternative econometric methodologies to deal with it. Section three presents the specification of models for the joint conditional p.f. of (y,sah), suggesting its estimation through ML. This section also includes a very brief account of copula theory, setting the general framework for the proposed specifications. Section four introduces the empirical application of the present methodology, describing the data used for estimation, a cross-section sample taken from the Portuguese National Health Survey (NHSur) of 1998/99. Section five presents and comments on estimation results. Finally, section six concludes the paper.

2. The Problem

2.1 Endogeneity

As previously mentioned, the possible endogeneity of *sah* variables in regression models for health care utilization poses relevant research issues. Formally, endogeneity (of *sah*) is referred to here according to the following, well established, definition (see Cameron and Trivedi 1998, ch. 11): denote the joint conditional p.f. of (y,sah) given *x*, as $f(y,sah\,|\,x;\theta)$. The usual factorization has

$$f(y, sah \mid x; \theta) = g(y \mid sah, x; \theta_1) f_2(sah \mid x; \theta_2),$$

where $\theta \equiv (\theta_1, \theta_2)$ denotes a parameter vector. If the marginal p.f. of *sah* depends on θ_1, estimating the parameters θ_1 by conditioning *y* on *sah* does not yield consistent estimates. In this case, *sah* is said to be endogenous.

Why can *sah* be endogenous? Two main arguments are usually invoked, that help explain the plausibility of this concern. The first reason is the possible existence of unobservables that condition individual self-assessments and, at the same time, influence the use of health care. Such factors as individual cultural background, personality characteristics or some dimensions of unmeasured health, like mental and social health (Jurges 2007) are difficult to measure (hence, not included in the regression model) and likely to influence both the dependent variable and self-judgements. Take, for instance, the case of a hypochondriac individual (usually a characteristic not accounted for): by definition, such a person will tend to display negative feelings towards his/her own health, probably rating it worse than it actually is. At the same time, he/she may also present a clear predisposition to visit the doctor often. In this case, the assumption of independence between *sah* and unobservables influencing *y* beyond the effect of observed covariates does not hold.

The endogeneity of *sah* may also be due to simultaneity of this variable and *y*. It is noted that, under the scheme adopted to collect the data used in this paper, individuals evaluate their own health state after visiting the doctor. Expectably, in these visits they acquire objective information that allows them to revise, thus update, their views about their own health.[(3)] Therefore, it is reasonable to suppose that individuals' health assessments are, to some extent, determined by the quantity of medical care recently received, which gives rise to the simultaneity of *sah* and *y* in the classical demand equation.

2.2 Econometric Choices

As previously mentioned, some authors have raised the concern of possible endogeneity of *sah* in models for health care use – see, *e.g.*, Windmeijer and Santos Silva (1997), Lourenço (2007b) and Van Ourti (2004). Each of these authors adopts a

[(3)] This information is considered to be objective, because it is provided by the doctor, possibly based on diagnostic tests, like lab tests, x-rays, etc..

different methodological course in face of this issue. While in Lourenço (2007b) *sah* is simply excluded from the regression model, the opposite is proposed in Van Ourti (2004), with *sah* included in the set of regressors, alongside with remaining covariates. As previously mentioned, both NLS and ML estimates are likely to be inconsistent, due to either omission of a potentially relevant regressor (Lourenço 2007b) or possible misspecification of the model for the conditional expectation, $E(y \mid sah, x)$ (Van Ourti 2004). Windmeijer and Santos Silva (1997), in turn, do take into account the possible endogeneity of *sah*, resorting to GMM techniques to estimate a regression model for the number of visits to the doctor by individuals.

Addressing endogeneity within a limited information framework usually requires the availability of instrumental variables. Windmeijer and Santos Silva (1997) suggest using, as instruments, variables that influence health in the long run, *e.g.*, variables which reflect behavioural attitudes like smoking- and drinking-related variables. Valid instruments are also required for the Hausman test of endogeneity, comparing NLIV to NLS or quasi-ML estimates (see, *e.g.*, Grogger 1990).

One alternative to the foregoing approaches is to adopt a full information strategy, specifying $f(y, sah \mid x)$ and estimating the resulting model through likelihood-based methods. This goal can be achieved using a particular class of cumulative distribution functions (c.d.f.'s) known as copulas. Essentially, a copula function is a joint c.d.f. whose marginals are uniform. In formal terms, the model for the joint conditional c.d.f. of $(y, sah) \mid x$ can be expressed as

$$F(y, sah \mid x) = C(F_1(y \mid x), F_2(sah \mid x) \mid x), \qquad (1)$$

where C is the copula, and F, F_1 and F_2 denote, respectively, the joint and marginal c.d.f.'s. For generality, the total set of conditioning covariates, x, is considered in the above expressions for these c.d.f.'s: formally, this poses no difficulty if the actual sets of covariates in each margin do not coincide. It simply means that exclusion restrictions (*e.g.*, advocated by economic theory) are imposed. Later in the paper (as of section 3.2), the distinction between both sets of regressors is to be made explicit, with x denoting their respective reunion.

The notion of copula has been well known for some time in statistics. It was introduced in the literature by Sklar (1959), although the main idea dates back to Hoeffding (1940). Its application to the study of economic problems is a recent but fast-growing field, namely in finance (see, *e.g.*, Bouyé, Durrleman, Nikeghbali, Riboulet and Roncalli 2000). Lee (1983), in one seminal paper, was the first to use copulas in econometrics, introducing the "normal copula" as an alternative to Heckman's (1976) two-step procedure of modelling selectivity. General surveys on copulas can be found in Joe (1997), Nelsen (2006) and Trivedi and Zimmer (2006).

The area of health economics has also witnessed a recent but fast increase in the use of copulas. Smith (2003) applies copulas to the specification of models for health care data that may suffer from selectivity bias. Zimmer and Trivedi (2006) use trivariate copulas to specify a regression joint model for three discrete response variables. These are, respectively, two counted measures of health care use by spouses, and a binary variable of insurance status. Dancer, Rammohan and Smith (2008) adopt a similar methodology to assess the degree of dependence between infant mortality and child nutrition. Quinn (2007a) addresses the simultaneity of mortality risk, health and lifestyles with a reduced-form system of equations, using a copula to define the corresponding multivariate distribution. Other examples in the area of health economics and econometrics are mentioned in the excellent survey by Quinn (2007b).

A full information methodology can also be implemented by using a bivariate mixture model for the specification of $f(y, sah \mid x)$. For instance, the joint p.f. of *sah* and *y* can be obtained upon mixing statistical independence, conditional on unobserved heterogeneity. Formally,

$$f(y, sah \mid x) = \int f_1(y \mid x, \varepsilon) f_2(sah \mid x, \varepsilon) h(\varepsilon \mid x) d\varepsilon\,, \tag{2}$$

where f_1 and f_2 represent the marginal p.f.'s and ε denotes unobserved heterogeneity, with density *h*. Except for some particular cases, one disadvantage associated with this approach is that it generally leads to criterion functions without analytical expressions, which require simulation-based or numerical approximation methods of maximization. On the other hand, such a specification enables the control of rich heterogeneity structures.

Actually, a mixture joint model can be given a copula interpretation, with the copula function implicitly defined by

$$F(y, sah \mid x) = \sum_{i \le y} \sum_{j \le sah} f(i, j \mid x),$$

and *f* as in (2). The next section presents the specification of a mixture model that is used in the present application and details its interpretation as a copula-based model.

3. Model Specification

This section presents models for the joint conditional p.f. of (y, sah), given a set of regressors. The section begins with a brief presentation of bivariate copulas, setting the general framework for the proposed copula-based models and subsequent empirical application.

3.1 Copulas

The main finding of copula theory is the fact that the joint c.d.f. of a set of real-valued random variables (r.v.'s) can be separated into its marginal c.d.f.'s and a copula, describing their dependence structure. More precisely, an *l*-variate copula (or *l*-copula) is defined as the c.d.f. of a random *l*-vector with uniform marginal c.d.f.'s. In the bivariate case, a 2-copula is a function $C:[0,1]^2 \mapsto [0,1]$ that satisfies the following properties:

i. For every $u \equiv (u_1, u_2) \in [0,1]^2$,

$C(u) = 0$, if at least one coordinate of *u* is zero;

$C(1, w) = C(w,1) = w, w \in [0,1]$.

ii. $\forall (a_1, a_2), (b_1, b_2) \in [0,1]^2$, $a_j \le b_j, j = 1,2$, $\Delta_{a_2}^{b_2} \Delta_{a_1}^{b_1} C(v) \ge 0$, where the two first-order differences of the function *C* are defined, respectively, as

$$\Delta_{a_1}^{b_1} C(v) \equiv C(b_1, v_2) - C(a_1, v_2), \quad \Delta_{a_2}^{b_2} C(v) \equiv C(v_1, b_2) - C(v_1, a_2).$$

Expression $\Delta_{a_2}^{b_2} \Delta_{a_1}^{b_1} C(u)$ is naturally interpreted as $\Pr(a_1 \le u_1 \le b_1, a_2 \le u_2 \le b_2)$.

If *F* is a bivariate c.d.f. with margins F_1, F_2, then, there exists a 2-copula *C* such that, for any random vector $z \equiv (z_1, z_2) \in R^2$,

$$F(z_1, z_2) = C(F_1(z_1), F_2(z_2)).$$

If F_1, F_2 are continuous, then C is unique; otherwise C is uniquely determined on $RanF_1 \times RanF_2$ (*Ran G* denotes the range of the function *G*). Conversely, if C is a 2-copula and F_1, F_2 are c.d.f.'s, then the function F defined above is a bivariate c.d.f. with marginal c.d.f.'s F_1, F_2.

The above statement is the bivariate version of what is known as Sklar's theorem. It demonstrates the role of copulas as the link between multivariate distributions and their univariate margins. The result essentially follows from the probability integral transformation, under which, for a continuous random variable w with c.d.f. F, $F(w)$ is uniformly distributed over the range $(0,1)$. The theorem enables the construction of a joint c.d.f., once the marginal c.d.f.'s and copula are available.

The copula is not unique if any of the marginal c.d.f.'s exhibits discontinuities – as is the case for discrete r.v.'s (see Joe 1997, p. 14, for details). Nevertheless, as Zimmer and Trivedi (2006, p. 64) point out, the non-uniqueness of copula in such cases is a theoretical issue that does not hinder its use in empirical applications. Finding a unique copula representation rests on full knowledge of the joint c.d.f.. Now, one of the reasons why researchers use copulas is precisely the fact that they ignore the true form of the joint c.d.f.. Thus, once the researcher decides which marginals to adopt, the issue, for him, is one of finding a copula that is able to reflect the dependence structure of the data while preserving desirable features of those marginals.

Given the purpose of the present paper, conditional c.d.f.'s and copulas must be considered. A bivariate conditional copula is a function $C:[0,1]^2 \mapsto [0,1]$, such that, conditional on some set (name it *H*), *C* corresponds to the above definition of copula. Sklar's theorem for conditional distributions leads to (see, *e.g.*, Patton 2005)

$$F(z \mid H) = C(F_1(z_1 \mid H), F_2(z_2 \mid H) \mid H).$$

As previously mentioned, the copula describes the dependence structure of r.v.'s with a given joint c.d.f.. One trivial but important case is the bivariate product copula, $\Pi(u) \equiv u_1 u_2$, that results in case of independence. The close relationship between copulas and dependence is also reflected by the Fréchet-Hoeffding bounds inequality: for every copula C and every $u \in [0,1]^2$, it can be shown that (see, *e.g.*, Nelsen 2006)

$$W_2(u) \equiv \max\{u_1 + u_2 - 1\,, 0\} \leq C(u) \leq \min\{u\} \equiv M_2(u).$$

In the bivariate case, both bounds are themselves copulas; the upper (lower) bound arises if and only if one r.v. is almost surely a strictly increasing (decreasing) transformation of the other. Between the extremes of independence and monotone functional dependence many forms of dependence can be considered, that are described by the properties of copulas. Besides the familiar notion of linear correlation, several dependence concepts and measures have been proposed in the literature (see Joe 1997, for an extended survey). For present purposes it suffices to distinguish "positive" from "negative" bivariate dependence – with positive dependence expressing the idea that "large" (or "small") values of both r.v.'s tend to occur together, and negative dependence expressing the notion that "large" values of one r.v. tend to be associated with "small" values of the other.

In practice, marginal c.d.f.'s can be specified conditional on a set of regressors, leading to a conditional copula representation for the joint (conditional) c.d.f. of the dependent r.v.'s of interest. In addition, the copula can include one or more parameters intended to capture the dependence between the univariate margins – usually, in the bivariate case, a single dependence parameter is used.

Interpreting the dependence parameter of a copula in the discrete case is not as straightforward as for continuous r.v.'s. In the latter case, the dependence parameter is frequently converted into a concordance measure, such as Kendall's tau or Spearman's rho, both defined on the interval $[-1\,,1]$ and independent of the functional form of the margins. However, as shown by several authors (*e.g.*, Marshall 1996, Denuit and Lambert 2005), this is not so with discrete r.v.'s, for which these measures are no longer bounded on the above interval, and are sensitive to the choice of margins. Still, every copula defines a range of permissible values for its dependence parameter, thereby allowing for varying degrees of positive and/or negative dependence. Thus, a researcher should choose those families of copulas that best fit his intended application, being able to capture the dependence pattern in the available data.

3.2 Model Specification

This section presents several alternative specifications for the conditional c.d.f. $F(y, sah \mid x)$. Starting with copula-based models, the bivariate probabilistic model can be generally expressed as in (1),

$$F(y, sah \mid x; \theta, \delta) = C(F_1(y \mid x_1; \theta_1), F_2(sah \mid x_2, \theta_2); \delta),$$

where *x* represents the vector of all conditioning variables, x_1 and x_2 denote the vectors of covariates in the margin of, respectively, *y* and *sah* (including intercept terms in both x_1 and x_2), $\theta \equiv (\theta_1', \theta_2')'$ denotes the vector of the margins parameters, and δ represents a dependence parameter.

In the present application, *y* is a count variable with unbounded support. Following common practice (see Cameron and Trivedi 1998), the function $F_1(y \mid x_1; \theta_1)$ is specified as the c.d.f. of a negative binomial p.f. with conditional mean $E(y|x_1) \equiv \mu_y = \exp(x_1'\beta_1)$ and variance $V(y \mid x_1) = \mu_y + \alpha\mu_y^2$, $\alpha > 0$. Formally, the marginal p.f. of *y* can be expressed as

$$f_1(y \mid x_1; \theta_1) = \frac{\Gamma(y+\alpha)}{\Gamma(\alpha)\Gamma(y+1)} \left(\frac{\alpha}{\mu_y + \alpha} \right)^{\alpha} \left(\frac{\mu_y}{\mu_y + \alpha} \right)^{y}, \qquad (3)$$

with $\theta_1 \equiv (\beta_1', \alpha)'$. As is well known, this functional form allows for overdispersion in the data, with reference to the Poisson p.f. (which results for $\alpha = 0$), thereby providing considerable modelling flexibility.

The second dependent r.v., *sah*, is a rank variable ranging from 1 to 5. Again following established literature, its marginal p.f. is specified as ordered probit, conditional on x_2 (see *e.g.*, Maddala 1983). Under this specification,

$$\Pr(sah = j \mid x_2; \theta_2) = \Phi(\lambda_{j+1} - x_2'\beta_2) - \Phi(\lambda_j - x_2'\beta_2), \; j = 1, \ldots, 5, \qquad (4)$$

with Φ denoting the standard normal c.d.f., $\theta_2 \equiv (\beta_2', \lambda')'$, $\lambda \equiv (\lambda_2, \ldots, \lambda_5)'$, $\lambda_1 = -\infty$ and $\lambda_6 = \infty$. From this it follows

$$F_2(j \mid x_2;\theta_2)=$$
$$\Pr(sah \le j \mid x_2;\theta_2)=\sum_{k=1}^{j}\Pr(sah = k \mid x_2;\theta_2)=$$
$$\Phi(\lambda_{j+1}-x_2'\beta_2),\ j=1,\ldots,5.$$

As usual, identification requires a normalization, such as 0, for the intercept in β_2 or one of the λ's.

The next step towards full specification of the c.d.f. of $(y,sah)\big| x$ consists on the choice of copula. In the present context, *y* and *sah* may well tend to move in opposite directions, thereby producing negative dependence in the data. This suggests the convenience of choosing a copula that allows for both positive and negative dependence. Several choices are possible, that satisfy this requirement.

Already referred to, the normal copula can be written as

$$C(u_1,u_2;\delta)=\Phi_2\left(\Phi^{-1}(u_1),\Phi^{-1}(u_2);\delta\right)=$$
$$\int_{-\infty}^{\Phi^{-1}(u_1)}\int_{-\infty}^{\Phi^{-1}(u_2)}\left(2\pi\sqrt{1-\delta^2}\right)^{-1}\exp\left(-\frac{v^2-2\delta vw+w^2}{2\left(1-\delta^2\right)}\right)dvdw,\quad |\delta|<1,$$

where Φ_2 and Φ^{-1} denote, respectively, the bivariate normal c.d.f. with zero mean vector, unit marginal variances and correlation coefficient δ, and the quantile function of the standard normal c.d.f.. Another example, involving two dependence parameters, is provided by the Student's *t* copula, formally expressed as

$$C(u_1,u_2;\gamma,\delta)=$$
$$\int_{-\infty}^{t_\gamma^{-1}(u_1)}\int_{-\infty}^{t_\gamma^{-1}(u_2)}\left(2\pi\sqrt{1-\delta^2}\right)^{-1}\left(1+\frac{v^2-2\delta vw+w^2}{2\left(1-\delta^2\right)}\right)^{-1-\gamma/2}dvdw,\quad \gamma\in\boldsymbol{N},\ |\delta|<1,$$

where t_γ^{-1} denotes the quantile function of the Student's *t* c.d.f. with γ degrees of freedom, and δ denotes the correlation coefficient. The γ parameter determines the thickness of the tails (the smaller the value of γ, the heavier the tails); as $\gamma\to+\infty$, the *t* copula tends to the normal copula. Both functions nest the independence copula (for $\delta=0$) and allow for positive ($\delta>0$) as well as negative ($\delta<0$) dependence. In addition, both copulas attain the upper and lower Fréchet-Hoeffding bounds, as, respectively, $\delta\to 1$ and $\delta\to -1$. Another shared feature of both copulas is the fact that they involve quantile functions of absolutely continuous c.d.f.'s (standard normal and *t*). Consequently, if, as in the present context where *y* and *sah* are discrete, the marginal c.d.f.'s are not strictly monotonous (therefore, not injective) functions of their respective arguments, ML estimation of the joint c.d.f. parameters can actually become arduous, even under theoretically identified models and with considerable sample sizes. Thus, particularly with discrete marginals and heavily parameterized models, the above two copulas may not be the more tractable functions to work with.[4]

Among few other cases allowing for negative dependence (see, *e.g*., Joe 1997, ch. 5.1), two of the most frequently encountered in the literature are the Frank copula (Frank 1979) and the Farlie-Gumbel-Morgenstern (FGM) copula, first proposed by Morgenstern (1956). Formally, these copulas can be written, respectively, as

Frank Copula

$$C(u_1,u_2;\delta)=\begin{cases}\dfrac{-1}{\delta}\log\left(1+\dfrac{(\exp(-\delta u_1)-1)(\exp(-\delta u_2)-1)}{\exp(-\delta)-1}\right), & \delta\neq 0,\\ u_1u_2, & \delta=0,\end{cases} \tag{5}$$

FGM copula

$$C(u_1,u_2;\delta)=u_1u_2(1+\delta(1-u_1)(1-u_2)),\quad |\delta|\leq 1. \tag{6}$$

Again, both functions nest the independence copula, which results for $\delta=0$. Positive and negative dependence occur with, respectively, $\delta>0$ and $\delta<0$. The Frank copula attains the Fréchet-Hoeffding upper and lower bounds, under, respectively, $\delta\to\infty$ and $\delta\to-\infty$. Despite its simplicity, the FGM copula is more restrictive, in that the dependence parameter is bounded on $[-1,1]$ and does not lead to either Fréchet-Hoeffding bound.

Let $(u_1,u_2)=(F_1(y\,|\,x_1),F_2(sah\,|\,x_2))$. Then, $F(y,sah\,|\,x)$ immediately results by plugging $F_1(y\,|\,x_1)$ and $F_2(sah\,|\,x_2)$ into (5) or (6).

The joint conditional p.f. of (y,sah) can also be expressed as a bivariate mixture model. Under this approach, conditional on *x* and unobserved heterogeneity, $\varepsilon\equiv(\varepsilon_1,\varepsilon_2)$, (y,sah) are assumed independent, with the above conditional margins: $y\,|\,(x_1,\varepsilon_1)$ is distributed as in (3), but now $E(y\,|\,x_1,\varepsilon_1)=\exp(x_1'\beta_1+\varepsilon_1)$, and

[4] See, however, Van Ophem (1999), who uses the normal copula to analyze dependence within a bivariate count data model.

$\Pr(sah = j \mid x_2, \varepsilon_2; \theta_2) = \Phi(\lambda_{j+1} - x_2'\beta_2 - \varepsilon_2) - \Phi(\lambda_j - x_2'\beta_2 - \varepsilon_2)$, $j = 1,\ldots,5$. Then, with $(\varepsilon_1, \varepsilon_2)$ assumed bivariate normal, independent of the regressors, with null mean vector, common variance, σ^2, and correlation coefficient δ, the model results as

$$f(y, sah \mid x; \theta, \sigma^2, \delta) = \int f_1(y \mid x_1, \varepsilon_1; \theta_1) f_2(sah \mid x_2, \varepsilon_2; \theta_2) \phi_{\sigma^2,\delta}(\varepsilon_1, \varepsilon_2) d\varepsilon_1 d\varepsilon_2, \tag{7}$$

where $\phi_{\sigma^2,\delta}$ denotes the bivariate normal density with parameters $(0, \sigma^2, \delta)$.

This formulation is equivalent to a model with random intercepts in f_1 and f_2. The assumption of Gaussian heterogeneity is common in the literature (see, *e.g.*, Train 2003). Although estimation is computationally demanding, requiring simulation-based methods or numerical approximations, the specification leads to easily interpretable parameters, namely the dependence parameter, δ. Within this framework, independence can easily be checked with the usual statistical tests. The assumptions of common variance and independence from regressors do not seem unreasonable in the present context and add to computational convenience; other schemes can be considered, such as random coefficients, varying dispersion parameters and/or dependence with respect to regressors. However, the usefulness of such sophistications in the present context is questionable, namely in view of the added estimation difficulty they are bound to represent. In any case, it is noted that two correlated heterogeneity terms are allowed for, instead of a shared term in f_1 and f_2. In the present context, these terms can naturally be seen as correlated unobserved heterogeneity influencing both *y* and *sah*. The assumption is also useful because it enables discernment of negative from positive dependence in the data (through the sign of δ), not just whether or not there is dependence (as the case would be with just one term).

As previously mentioned, the mixture model can be given a copula interpretation. In this case, the function *C* in (1) is defined as

$$F(y, sah \mid x) = \sum_{i=0}^{y} \sum_{j=1}^{sah} f(i, j \mid x) = \int F_1(y \mid x_1, \varepsilon_1) F_2(sah \mid x_2, \varepsilon_2) \phi_{\sigma^2,\delta}(\varepsilon_1, \varepsilon_2) d\varepsilon_1 d\varepsilon_2 = \int \Pi(F_1(y \mid x_1, \varepsilon_1), F_2(y \mid x_2, \varepsilon_2)) \phi_{\sigma^2,\delta}(\varepsilon_1, \varepsilon_2) d\varepsilon_1 d\varepsilon_2 ,$$

where F_k, $k=1,2$, now denote the marginal c.d.f.'s given (x_k, ε_k), and Π denotes the (conditional) independence copula.

3.3 Estimation

Maximum likelihood (ML) estimation of the above models requires the joint p.f. of (y, sah), given x, $f(y, sah \mid x)$. Under copula-based models for continuous response variables this is obtained as the second-order derivative of the copula, that is (conditioning on x is omitted),

$$f(z_1, z_2) = \frac{\partial^2 F(z_1, z_2)}{\partial z_1 \partial z_2} = \frac{\partial^2 C(u_1, u_2)}{\partial u_1 \partial u_2} f_1(z_1) f_2(z_2),$$

where $(u_1, u_2) \equiv (F_1(z_1), F_2(z_2))$. In the present case, involving discrete r.v.'s, $f(z_1, z_2 \mid x)$ is formed by taking differences. Formally,

$$\begin{aligned}&f(z_1, z_2) = \\ &F(z_1, z_2) - F(z_1 - 1, z_2) - F(z_1, z_2 - 1) + F(z_1 - 1, z_2 - 1) = \\ &C(F_1(z_1), F_2(z_2)) - C(F_1(z_1 - 1), F_2(z_2)) - C(F_1(z_1), F_2(z_2 - 1)) + C(F_1(z_1 - 1), F_2(z_2 - 1)).\end{aligned}$$

Then, upon the choice of copula, the individual contribution to the log-likelihood is formed by taking the logarithm of this last expression. After simultaneous ML estimation of all the parameters, variances of the estimates are obtained through the robust sandwich formula. It is noted that, defined as before, all the copulas referred to above are differentiable to order two at any particular value of δ, within its admissible range, so independence can be assessed with the usual likelihood-based tests.

Estimation of model (7) requires either maximum simulated likelihood (MSL) or numerical approximation. The former is used here, with (7) being approximated by direct Monte Carlo (MC) integration, that is,

$$f(y, sah \mid x; \theta, \sigma^2, \delta) \approx \frac{1}{S} \sum_{s=1}^{S} f_1(y \mid x_1, \varepsilon_1^s; \theta_1) f_2(sah \mid x_2, \varepsilon_2^s; \theta_2), \qquad (8)$$

where $\left(\varepsilon_1^s, \varepsilon_2^s\right)$, $s = 1, \ldots, S$ denote random draws from the bivariate normal, $\phi_{\sigma^2,\delta}$, and S is the number of draws. Gouriéroux and Monfort (1991) show that, under regularity conditions, the MSL estimator has the same asymptotic distribution as the ordinary ML estimator, provided that $\sqrt{n}/S \to 0$ as $n, S \to \infty$ (n denotes sample size). The number of draws used in the application of the present paper is selected *ad hoc*, mostly for reasons of computational convenience, on the basis of rough comparisons between results for various S values.

4. Data, Variables and Summary Statistics

The described approach is now applied to cross-sectional data on (y, sah) and a set of regressors, x, taken from the NHSur. In this application, different models of the joint p.f. of (y, sah) are estimated, from which the average effects, $E(y \mid sah, x)$, are evaluated. These average effects are also compared to the corresponding estimates obtained from mainstream conditional models for the p.f. of $y \mid (sah, x)$.

The NHSur, which collected the data used in this study, took place in 1998/99.[5] The survey gathered information on a representative sample from the non-institutionalized population residing in Portugal mainland, covering 48,606 individuals belonging to, roughly, 20,000 households. The information collected in the cross-section dataset includes individuals' socio-demographic and economic characteristics (age, sex, marital status, educational attainment, activity status, income, place of residence, health insurance status), health status (self-assessed health, chronic conditions, functional status, etc.), medical care utilization (number of doctor visits in a three month period) as well as some variables reflecting individuals' lifestyles, potentially affecting health (tobacco consumption habits, physical activity, etc.).

The working sample for the present study contains 27,044 observations, obtained after deletion of 21,562 records from the initial dataset. These exclusions are mostly due to incomplete records, that is, records with missing values on any of the variables used in the study. The item on self-assessed health is responsible for most of the incomplete records, due to the following reason: according to the interview proce-

[5] Detailed expositions about the survey procedures and data can be found, among others, in Barros, Machado and Galdeano (2008), and Ministério da Saúde - Instituto Nacional de Saúde (1999).

dure, survey information could be provided by a third person within the same household, except for the question on self-assessed health when addressed at an individual over fourteen. In this case, the self-assessed health question could only be answered by the individual himself. As a consequence, whenever the individual to interview was absent from home at the time of the interview, the information relative to her/his own self-assessed health was reported missing. It is noted that the present study assumes, at the outset, that these observations are missing at random, that is, the working sample (27,044) is not considered to be the result of some form of selectivity. Allowing for this issue within a full information approach requires consideration of trivariate joint c.d.f.'s, which, in itself, suggests an autonomous extension of the present paper.

About 4% of the records in the initial sample, relating to individuals who reported benefiting from a voluntary health insurance contract, are also dropped. Voluntary health insurance introduces endogeneity concerns, so this type of information is side-stepped in the present study (inference results refer to the sub-population of individuals with no voluntary health insurance contract). It is noted that the analysis of voluntary health insurance and its effect on medical care utilization or on individuals' health status is beyond the purpose of the present work.

The variables (y, sah), meant to reflect, respectively, health care utilization and self-assessed health status, are measured as follows: y equals the number of visits to the doctor in the three months before the survey interview; the value of sah is the coded answer to the question "In general, how do you rate your health status?". Five response alternatives yield a rank variable with categories 1 ("very bad") to 5 ("very good") – see table 1.

Summary statistics for y and sah are presented in table 1. Over the three months prior to the survey, each individual consulted the doctor 1.42 times, on average (standard deviation: 2.08) and about 41% of individuals did not visit the physician at all. With regard to self-assessed health, most individuals rate her/his own health as fair (37.9%), with the empirical distribution of sah exhibiting some degree of negative skewness.

Table 1 – Dependent Variables

y	Rel. Freq.	*sah*	Rel. Freq.
0	.413		
1	.247	1 (very bad)	.045
2	.138	2 (bad)	.174
3	.104	3 (fair)	.379
4	.038	4 (good)	.363
5	.022	5 (very good)	.039
6	.018		
7	.005		
8	.004		
9	.001		
10	.004		
11	.0004		
12	.003		
> 12	.003		
sample size (n):	27044		
average	1.42	average	3.18
variance	4.33	variance	.84

The covariates are described in table 2, being grouped under five headings: socioeconomic and demographic, health status, life styles, supply of medical care services and health insurance status. As detailed in the note below table 2, the marginals of the joint model do not share all regressors $(x_1 \neq x_2)$.[6] The selection of covariates for each marginal follows well-established research, both theoretical and empirical, developed by several authors (Grossman 1972; Muurinen 1982; Wagstaff 1986; Kenkel, 1995). Besides economic and behavioural criteria, practical considerations, such as data availability and computational tractability, are also at play in the choice of covariates.

The variables *age* and *agesq* (*age* squared) are included in both marginals so as to capture the health capital depreciation rate, which increases with age (Grossman 1972). The oldest individual in the working sample is aged 95, and the youngest is less than 1 year old (classified as 0 years old in the dataset). The regressor *rural* (= 1 if the individual resides in a rural area, 0 otherwise) is also included in both marginals

[6] In any case, all regressors are included in the models for the conditional p.f. of $y|(sah,x)$ (Poisson and negative binomial), in order to allow for comparisons between average effects estimates from both approaches.

because it can directly influence both health care use and health status. It is included here in order to allow for possible behavioural differences in individuals living in rural regions, relative to those who live in urban areas, as well as for differences in the supply of medical care services, which may affect the time-price of medical care use and, consequently, the actual use of health services. The variable *education* represents the maximum educational achievement of the individual, measured in number of fully completed schooling years. If the person is a child (age < 15), this variable measures the number of schooling years of the most educated adult in the household. *Education* is included in both marginals because it may influence both the efficiency in the production of health and the propensity to seek care. The average number of years of education is 5.68, with a minimum of 0, which means that some individuals, especially the elderly, never attended formal schooling. It should be noted that Portugal is characterized by low levels of educational achievement, especially among the older cohorts. Gender (*female*), *income* and the individual's retirement status (*retired*) are also included in both marginals. *Income* measures monthly disposable household income per equivalent adult computed using the modified OECD scale (the scale assigns a weight of 1.0 to the first adult in the household, for each additional adult (aged over 13) a weight of 0.5 and, for each child, a weight of 0.3). Total household reported income includes regular wages, retirement pensions and all sorts of social security subsidies received by the household members.

In the socio-economic and demographic group of covariates, *married* and *not_work* are included only in the marginal that explains medical care use. The activity status of the individual (*not_work*) is likely to affect the time price of medical care use but not health status, at least not directly. The variable *not_work* equals 0, if the individual is employed or self-employed and worked in the two weeks before the interview, and 1, if the individual did not work. This last group includes children, students, individuals on sick leave, unemployed, retired, those doing housework and other individuals economically inactive. 67% of the individuals in the sample reported not to work during the two weeks before the interview.

Table 2
Regressors – Definition and Summary Statistics

Variable name	Variable Definition	average	st.dev.	min.	max.
	Socioeconomic and Demographic				
age	Age, in years, divided by 10	4.26	2.49	0	9.50
agesq	*age* squared	24.29	20.97	0	90.25
rural	= 1 if the individual lives in a rural area	.18	.39	0	1
education	Years of schooling. If child, years of schooling of the most educated adult in the household.	5.68	4.27	0	24
female	= 1 if the individual is female	.60	.49	0	1
income	Monthly disposable household income per equivalent adult (unit: 100 euros)	3.63	2.77	.23	24.94
married [(1)]	= 1 if the individual is married	.54	.50	0	1
not_work [(1)]	= 1 if the individual did not work in the two weeks prior to taking the survey	.67	.47	0	1
retired	= 1 if the individual is retired	.23	.42	0	1
	Health Status				
limitation	= 1 if the individual has some physical handicap that prevents him from executing certain physical daily activities	.04	.19	0	1
n_chronic	Number of chronic conditions reported	.96	1.03	0	6
no_dental [(2)]	= 1 if the individual reports no dental hygiene habits	.06	.23	0	1
ill	= 1 if the individual reports being ill in the previous two weeks	.37	.48	0	1
stress	= 1 if the individual took sleeping pills in the last two weeks	.12	.33	0	1
	Life Styles				
smoke [(2)]	= 1 if the individual smokes on a daily basis	.11	.31	0	1
sedentary [(2)]	= 1 if the individual's daily activities require no physical activity	.58	.49	0	1
	Supply Side				
med_supply[(1)]	Total number of licensed physicians per 1000 inhabitants	2.75	2.22	.58	9.15
	Health Insurance Status				
nhs_only	= 1 if the individual is covered only through the National Health Service	.84	.36	0	1

[(1)] Regressor in $f_1(y|x_1)$ but not in $f_2(sah|x_2)$. [(2)] Regressor in $f_2(sah|x_2)$ but not in $f_1(y|x_1)$.

Apart from *no_dental* (=1 if the individual reports no dental hygiene habits), all remaining health status covariates are assumed to influence both the propensity to seek medical care and self-assessed health. Limitations on functional status (*limitation*) and a count of chronic diseases (*n_chronic*) are included in order to reflect the physical component of health. In addition, the variable *ill* (=1 if the individual has been feeling sick in the two weeks before the interview) is also used, so as to control the effect of the occurrence of an acute illness. The variable *stress* is used as a proxy to reflect the individual's level of anxiety, thus capturing the mental component of health status (Manning *et al*. 1982).

The variables under the heading Life Styles are *smoke* (= 1 if the individual smokes daily) and *sedentary* (= 1 if the individual's daily activities require no physical activity). Both covariates enter only the second margin, directly influencing the individual's health status. The data show that 11% of the population smokes on a daily basis, and 58% have daily activities requiring no physical activity.

The regressor *med_supply* (total number of licensed physicians per 1000 residents in the individual's area of residency) is included only in the marginal for *y*, as it has no direct bearing on the conditional p.f. of *sah*. Its inclusion is intended to capture the effect, upon the conditional distribution of health care demand, of the availability of medical care services in the individual's area of residence.

Turning to the insurance status covariate (*nhs_only*), a previous short note about the Portuguese health care system helps to understand why, in the Portuguese context, health insurance covariates should be considered exogenous. The Portuguese health care system provides and finances medical care to Portuguese citizens through a mix of public and private providers and financing bodies. In what concerns the structure providing health insurance, two main coexisting overlapping coverage systems can be identified: 1) the National Health Service (NHS), mainly financed through taxation and covering 100% of the population; 2) special public and private health insurance schemes, membership to which is based on profession/occupation (usually known as health subsystems). Altogether, these schemes, which provide health insurance to about 16% of the population, are financed through a mix of employer and employee contributions, and membership is mandatory, thus not depending on the will of the individual (Barros and Almeida Simões 2007). Given this context, covariates reflecting health insurance status can therefore be considered exogenous in the application. For an extended discussion of this issue, see Barros, *et al*. (2008). The covariate *nhs_only* (=1 if the individual is covered only through the NHS, 0 if the individual has a supplementary health insurance through a health subsystem scheme) is included in both marginals.

5. Estimation Results

Estimation results are presented in tables 3 and 4. Table 3 contains estimates of the parameters in the conditional p.f. of y given (sah, x), $g(y \mid sah, x)$, specified, respectively, as Poisson and negative binomial (NB2) with the usual exponential conditional mean. Table 4 presents estimation results from joint models for $f(y, sah \mid x)$. All computations were performed using TSP 5.0 (Hall and Cummins 2005)[7].

In what concerns $g(y \mid sah, x)$, a noticeable, often found, result, is the clear rejection of the Poisson model in favour of the NB2 model ($\hat{\alpha} = .619$, statistically significant). This result can be taken as indication of overdispersion in the data, with reference to the Poisson specification. Thus, even with correct specification of $E(y \mid sah, x)$, NB2's estimates seem preferable to those of the Poisson.

Table 3: Estimation Results – $g(y \mid sah, x)$: Poisson, NB2

	Poisson		NB2	
Variable	coefficient	st. error	coefficient	st. error
intercept	.859	.077	.872	.065
y	–	–	–	–
sah	-.339	.014	-.352	.011
age	-.103	.022	-.147	.017
agesq	.007	.002	.012	.002
rural	-.040*	.022	-.055	.020
education	.020	.003	.024	.003
female	-.009**	.020	.022**	.016
income	.012	.003	.014	.003
married	.100	.023	.121	.019
not_work	.120	.025	.100	.021
retired	.066	.026	.075	.025
limitation	.087	.043	.065*	.036
n_chronic	.128	.009	.151	.008
no_dental	-.094	.038	-.110	.033
ill	.479	.018	.488	.016
stress	.275	.023	.292	.021
smoke	-.125	.034	-.110	.027
sedentary	.085	.023	.089	.019
med_supply	.021	.004	.022	.003
nhs_only	-.050*	.027	-.042*	.023
α	–	–	.619	.013
Log-likelihood	-45198.4		-41440.8	
SBIC	45300.5		41547.9	
Sample size	27044			

* Not significant at the .05 level. ** Not significant at the .10 level.

[7] TSP codes for ML estimation of copula models with discrete marginals and MSL estimation of the mixture model are available on request from the authors.

Table 4 – Estimation Results: $f(y, sah \mid x)$

Model	Frank		FGM		Mixture	
Variable	coefficient	st. error	coefficient	st. error	coefficient	st. error
$f_1(y\|x_1)$						
intercept	-.543	.042	-.538	.040	-.755	.054
age	-.069	.013	-.071	.012	-.097	.018
agesq	.007	.002	.010	.001	.011	.002
rural	-.053	.018	-.035	.017	-.061	.021
education	.012	.002	.017	.002	.013	.003
female	.070	.014	.056	.013	.084	.017
income	.007*	.003	.004**	.003	.006**	.004
married	.129	.016	.096	.015	.150	.021
not_work	.160	.017	.147	.016	.132	.023
retired	.098	.023	.107	.021	.128	.026
limitation	.213	.032	.024**	.027	.172	.037
n_chronic	.221	.008	.236	.007	.230	.009
ill	.646	.014	.604	.013	.653	.016
stress	.378	.021	.352	.019	.402	.022
med_supply	.024	.003	.023	.003	.020	.004
nhs_only	-.031**	.020	-.051	.019	-.030**	.025
α	.684	.011	.593	.010	.179	.015
$f_2(sah\|x_2)$						
intercept	3.717	.041	3.815	.041	4.118	.055
age	-.243	.013	-.224	.013	-.260	.015
agesq	.013	.002	.010	.001	.013	.002
rural	.005**	.003	-.001**	.018	.012**	.020
education	.047	.002	.049	.002	.050	.002
female	-.110	.015	-.112	.015	-.123	.017
income	.041	.003	.034	.003	.045	.004
retired	-.236	.022	-.194	.022	-.265	.025
limitation	-.805	.035	-.800	.035	-.881	.043
n_chronic	-.327	.008	-.338	.008	-.365	.009
no_dental	-.148	.028	-.224	.028	-.151	.034
ill	-.702	.015	-.710	.015	-.777	.018
stress	-.355	.021	-.361	.021	-.398	.024
smoke	.037**	.023	-.006**	.023	.058	.027
sedentary	-.086	.018	-.120	.018	-.113	.020
nhs_only	-.037*	.022	-.047	.021	-.053	.026
λ_3	1.351	.017	1.403	.017	1.489	.021
λ_4	2.997	.020	3.062	.020	3.289	.031
λ_5	5.062	.025	5.161	.025	5.554	.046
σ^2	-	-	-	-	.686	.027
δ	-1.484	.048	-.688	.021	-.663	.023
Log-likelihood	-67,658.7		-67,750.0		-67,548.39	
SBIC	67,699.5		67,765.4		67,742.2	
Sample size	27,044					

* Not significant at the .05 level. ** Not significant at the .10 level.

Expectably, the estimated coefficients of the regressors in x_1 are quite different under NB2 and Poisson models for $g(y \mid sah, x)$ and for $f_1(y \mid x_1)$ within joint models (table 4). Actually they are not comparable, as they do not refer to the same quantities. In the former case, each coefficient estimates the relative change of the conditional mean of y, given (sah, x), that is,

$$\frac{D_j E(y|sah,x)}{E(y|sah,x)},$$

with D_j denoting first-order partial derivative with respect to the j-th regressor of $E(y|sah,x)$. In the latter case each estimate refers to the relative change of the conditional expected value of y, marginal to *sah*, that is,

$$\frac{D_j E(y|x_1)}{E(y|x_1)}.$$

Only if $E(y|sah,x) = E(y|x_1)$, do these expressions coincide. Independence of y is assumed with respect to x covariates out of x_1 – *no_dental*, *smoke* and *sedentary* (which would qualify as instruments for *sah*, possibly useful in IV/GMM estimation of structural demand equations), but it is clearly not assumed with respect to *sah*.

Under full information approaches (table 4), most coefficients estimates from the three models are noted to be similar, in both magnitude and sign. The exception is provided by the dependence parameter estimates, a result that may be explained by functional form differences between models (with regard to the mixture model, $\hat{\sigma}^2$ possibly captures part of the effects of unobserved heterogeneity). The closeness of coefficients estimates from the three models seems to reflect the flexibility of copulas, able to discern dependence from the marginals. Meanwhile, the (small) differences to estimates from the mixture model may be due to the fact that the latter are not ML estimates, being obtained from maximization of an approximate log-likelihood function. MSL estimates for the mixture model are obtained using S = 100 draws of pseudo-random vectors from the bivariate normal. This number of draws is selected for computational convenience and from rough comparisons with results for larger S (*e.g.*, S = 250 yields almost the same estimates and standard errors). Results might be closer to those from Frank and FGM models with a significantly larger *S* but this would increase the computational burden. Here, instead of direct MC sampling, the so-called "quasi-MC" methods (*e.g.*, Halton sequences) may prove more efficient.[8]

With regard to regressors coefficients in table 4, some estimates point to a varying degree of relevance of the corresponding covariates in the two margins: the variable *income* is hardly relevant to explain health care use but seems clearly relevant in $f_2(sah \mid x_2)$; the opposite occurs with respect to *rural* (residence in a rural area), which is relevant in $f_1(y \mid x_1)$ and irrelevant in $f_2(sah \mid x_2)$. On the other hand, the estimated relevance of *nhs-only* (membership exclusively in the statutory public system) in $f_1(y \mid x_1)$ and $f_2(sah \mid x_2)$ seems to vary under different models (clearly relevant in both marginals only under the FGM copula).

Overall, estimation results for the joint models are in line with the usual findings in the literature. In general, worse-off individuals in terms of health status seek medical care more often (see the sign of covariates *limitation* (+), *n_chronic* (+), *ill* (+) and *stress* (+)) than those in better health. Nevertheless, higher income or education levels are both linked to an increase in demand for health care, though *income* shows little relevance.

The dependence parameter estimate is both negative and significant across the three joint models. As expected, all three models point to a negative dependence between *y* and *sah*, after conditioning on observed factors. The precision of the estimates can help fuel the suspicion of simultaneity of both variables, which, as previously mentioned, can cause endogeneity of *sah* within regression models for health care utilization.

Negative dependence between *y* and *sah* also seems visible in the estimates of the average effects, included in table 5. The figures reported in the table refer to the estimated conditional mean of *y* for, respectively, the five admissible values of *sah*, *nhs* = 0,1, and remaining covariates at sample averages. Let these remaining covariates be termed x_*; then, under the Poisson and NB2 models, the values in the table are computed as $\exp\left((sah, nhs, \bar{x}_*')\hat{\beta}\right)$. For joint models, the conditional moment

[8] According to, *e.g.*, Bhat (2001) and Train(2003), Halton draws can be computationally much more efficient than direct MC sampling.

$$E\left(y|sah,nhs,\bar{x}_*\right)=\sum_{y\geq 0} yf\left(y|sah,nhs,\bar{x}_*\right)=$$

$$\sum_{y\geq 0} y\frac{f\left(y,sah\,|\,nhs,\bar{x}_*\right)}{f\left(sah\,|\,nhs,\bar{x}_*\right)}=\frac{\sum_{y\geq 0} yf\left(y,sah\,|\,nhs,\bar{x}_*\right)}{\sum_{y\geq 0} f\left(y,sah\,|\,nhs,\bar{x}_*\right)}$$

is estimated by

$$\frac{\sum_{y=0}^{30} y\hat{f}\left(y,sah\,|\,nhs,\bar{x}_*\right)}{\sum_{y=0}^{30} \hat{f}\left(y,sah\,|\,nhs,\bar{x}_*\right)},$$

where $\hat{f}$ denotes evaluation of f at the estimated parameters (for $y > 30$ both summands in the fraction are negligible). Under the mixture model, $f\left(y,sah\,|\,nhs,\bar{x}_*\right)$ is estimated by (8), with $\hat{f}_1$ and $\hat{f}_2$ evaluated at MSL estimates, and $\left(\varepsilon_1^s,\varepsilon_2^s\right)$, $s=1,\ldots,100$, random draws from the bivariate normal with parameters $\left(0,\hat{\sigma}^2,\hat{\delta}\right)$.

Table 5
Average Effects – $E\left(y\,|\,sah,nhs,\bar{x}_*\right)$

	nhs_only = 0				
sah	Poisson	NB2	Frank	FGM	Mixture
1	2.646	2.682	2.866	2.828	2.349
2	1.887	1.885	2.793	2.772	1.670
3	1.345	1.325	2.387	2.426	1.237
4	.959	.932	1.920	1.960	.967
5	.683	.655	1.771	1.782	.714
	nhs_only = 1				
sah	Poisson	NB2	Frank	FGM	Mixture
1	2.519	2.572	2.814	2.743	2.252
2	1.795	1.808	2.737	2.684	1.603
3	1.280	1.271	2.334	2.344	1.191
4	.912	.894	1.884	1.900	.932
5	.650	.628	1.745	1.737	.687

Results indicate that estimates are only slightly higher for *nhs_only* = 0 than for *nhs_only* = 1. In line with most results in the literature, this suggests that, for an "average" individual, being covered only by the NHS has a small downward impact on the influence of his own *sah* on mean health care utilization. As expected, this mean decreases as sah increases, a result that is consistently obtained across the various specifications. The Frank and FGM models yield somewhat similar results, distinctly higher than the Poisson, NB2 and mixture models, namely for higher values of *sah*.

Finally, it is interesting to see how the estimated models fit the data. To give an idea of the goodness of fit of the models, table 6 gives the true and fitted frequencies of the number of visits to the doctor. The fitted frequencies distribution is obtained as the average over observations of the predicted probabilities fitted for each count. Formally, $n^{-1}\sum_{i=1}^{n} f_1(y \mid sah_i, x_i)$, for $y = 0, 1, 2, \ldots$; under models of the joint p.f. of (y, sah) this is computed as $n^{-1}\sum_{i=1}^{n}\left(f(y, sah_i \mid x_i)/f_2(sah_i \mid x_{2i})\right)$. Both the joint models and the NB2 model fit the data relatively well, being particularly good at predicting the number of individuals with few visits (up to 2). For $y = 3$ these four models under-predict the actual frequency, while the reverse occurs for $y > 3$.

Table 6
Actual and Fitted Frequencies

		Model				
Visits	Actual	Poisson	NB2	Frank	FGM	Mixture
0	.413	.313	.414	.414	.400	.408
1	.247	.311	.256	.254	.260	.253
2	.138	.189	.140	.140	.145	.136
3	.104	.097	.076	.077	.080	.075
4	.038	.047	.043	.044	.045	.043
> 4	.060	.043	.071	.071	.070	.085

The statistical significance of the differences between actual and fitted frequencies can be assessed using a test for the joint moment conditions

$$\begin{cases} E\left(d_j(y) - \Pr(y = j \mid sah, x)\right) = 0, \; j = 0, \ldots, 4, \\ E\left(d_5(y) - \Pr(y > 4 \mid sah, x)\right) = 0, \end{cases}$$

with the binary variables d_j defined as $d_j(y) = 1$, if $y = j$, j = 0, …, 4, and $d_5(y) = 1$, if $y > 4$.[9] In order to try and reduce the effects of a large sample size on

[9] For details on how to implement this type of tests and a simulation on their performance, see Cameron and Trivedi (1998).

the outcome of the test, it is carried out with a sub-sample of about 25% of the initial size (6436 observations). For each specification, the results of the test, asymptotically distributed as a chi-squared distribution with 5 degrees of freedom, are the following:

Poisson	NB2	Frank	FGM	Mixture
589.75	144.62	123.51	147.50	198.98

The null hypothesis is clearly rejected across all specifications, in spite of the reduced sample size. Nevertheless, the outcomes of the test suggest an ordering of the models, with the Poisson displaying a worse result than the remaining models, NB2 included. According to the results of table 6, the NB2 model competes well with joint models, performing even better than the FGM and mixture models. Among joint models, in accordance with the results in table 6, the Frank copula seems to provide the best fit to the observed dataset.

6. Conclusion

The study of the relevant factors influencing health care utilization constitutes one of the main research interests in health economics. In this context, the measurement of the impact of self-assessed health on the demand for health care stands as an important issue that requires careful methodological approaches. In particular, the possible endogeneity of *sah* within regression models for health care utilization is usually met with GMM-type methods, requiring available instruments.

Alternatively, one can turn to a full information methodology, by modelling the joint conditional p.f. of *sah* and a measure of health care utilization (y), given a set of regressors. This is the route taken in the present paper, through the use of copula functions, that provide the means for flexible and tractable specification of the desired p.f.. ML estimation of the adopted models naturally enables recovery of several entities of interest, such as features of the conditional p.f. of either endogenous variable, given the other. Among such features, the value of the conditional expectation of y, for different *sah* values constitutes a prominent example with potential economic interest, indicative of the influence of the individual's health status on medical care utilization. In itself, this entity constitutes a useful and legitimate alternative to the more conventional measure of causal effect of *sah* on *y*, within uni- or multi-equation regression models for health care utilization. In any case, if one wants to move along conventional routes, then GMM/NLIV seem more advisable methods than NLS or conditional ML, which, either treat *sah* as exogenous or simply remove it from the regression model. In the present paper, this suspicion is reinforced by the precision with which the various adopted joint models are able to gauge the dependence parameter between *sah* and *y*. Nevertheless, a formal test of *sah*'s endogeneity within a limited information approach (*e.g*., Poisson, NB2 or simultaneous equations model) would require use of a Hausman-type test (Grogger 1990), which is not within present purposes.

The foregoing text has suggested some ideas for future research. One of these consists on a systematic enquiry into the identification possibilities of the present full information approach, as it relates to limited information modelling. What, if any, is the relationship between the parameters identified by the latter, and those identified by the former, namely moments of the conditional density of each endogenous variable? Is there a mapping between the two groups of parameters when discrete variables are involved? To the best of our knowledge, no such enquiry has yet been produced.

Another idea consists on the extension of copula models to the trivariate case, so as to deal with the possibility of sample selection within the NHSur dataset, apart from the present endogeneity issue. In this sense, the present study can constitute a first step in that direction, which, in itself a complex issue, may well benefit from some of the ideas and methods set forth in the present work.

References

Bago D'Uva T (2006) Latent Class Models for Utilization of Health Care. Health Economics 15:329-343.

Barros PP, Machado MP, Sanz-De-Galdeano, A (2008) Moral hazard and the demand for health services: A matching estimator approach. Journal of Health Economics 27(4):1006-1025.

Barros PP, Almeida Simões, J (2007) Portugal: Health System Review. Health Systems in Transition 9(5):1–140.

Bhat CR (2001) Quasi-random Maximum Simulated Likelihood Estimation of the Mixed Multinomial Logit Model. Transportation Research: Part B: Methodological 35:677-693.

Bouyé E, Durrleman V, Nikeghbali A, Riboulet G, Roncalli T (2000) Copulas for Finance: A Reading Guide and Some Applications. Unpublished Manuscript, London, Financial Econometrics Research Centre, City University Business School.

Cameron AC, Trivedi PK (1998) Regression analysis of count data. New York, Cambridge University Press.

Dancer D, Rammohan A, Smith MD (2008) Infant mortality and child nutrition in Bangladesh. Health Economics 17(9):1015-1035.

Deb P, Trivedi PK (1997) Demand for Medical Care by the Elderly: A Finite Mixture Approach. Journal of Applied Econometrics 12:313-336.

Deb P, Trivedi PK (2002) The Structure of Demand for Health Care: Latent Class Versus Two-part Models. Journal of Health Economics 21:601-625.

Denuit M, Lambert P (2005) Constraints on concordance measures in bivariate discrete data. Journal of Multivariate Analysis 93:40–57.

Frank MJ (1979) On the Simultaneous Associativity of $F(x,y)$ and $x + y - F(x,y)$. Aequationes Mathematicae 19:194-226.

Gouriéroux C, Monfort A (1991) Simulation Based Econometrics in Models with Heterogeneity. Annales d'Economie et de Statistique 20:69-107.

Grogger J (1990) A Simple Test for Exogeneity in Probit, Logit and Poisson Regression Models. Economics Letters 33: 329-332.

Grossman M (1972) Concept of Health Capital and Demand for Health. Journal of Political Economy 80(2): 223-225.

Hall BH, Cummins C (2005) TSP User's Guide, Version 5. TSP International, Palo Alto, Ca.

Heckman J (1976) The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models. Annals of Economics and Social Measurement 5:475-492.

Hoeffding W (1940) Masstabinvariante Korrelationstheorie. Schriften des Matematischen Instituts und des Instituts für Angewandte Mathematik der Universität Berlin, 5, Heft 3:179-233. [Reprinted as "Scale-invariant Correlation Theory", in N. I. Fisher, P. K. Sen (eds.), *The Collected Works of Wassily Hoeffding*, New York, Springer.]

Joe H (1997) Multivariate Models and Dependence Concepts. New York, Chapman & Hall.

Jones AM, O'Donnell O (2002) Econometric analysis of health data. Chichester, Wiley.

Jurges H (2007) True Health vs Response Styles: Exploring Cross-country Differences in Self-reported Health. Health Economics 16:163-178.

Kenkel DS (1995) Should You Eat Breakfast - Estimates from Health Production-Functions. Health Economics 4(1):15-29.

Lee LF (1983) Generalized Econometric Models With Selectivity. Econometrica 51:507-512.

Lourenço OD, Quintal C, Ferreira P, Barros PP (2007a) A equidade na utilização de cuidados de saúde em Portugal: Uma avaliação baseada em modelos de contagem. Notas Económicas 25:6-26.

Lourenço OD (2007b) Unveiling health care consumption groups: a latent class approach in the Portuguese health data context. Dissertation, University of Coimbra Faculty of Economics.

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

Manning WG, Newhouse JP, Ware JE Jr (1982) The status of health in demand estimation, or beyond excellent, good, fair and poor. In: Fuchs VR (ed) Economic aspects of health, The University of Chicago Press, Chicago.

Marshall A (1996) Copulas, Marginals and Joint Distributions. in Ruschendorf L, Schweizer B, Taylor MD eds., Distributions With Fixed Margins and Related Topics, Hayward, Ca., Institute of Mathematic Statistics, 213-222.

Ministério da Saúde - Instituto Nacional de Saúde (1999) INS 1998/1999. Continente. Dados Gerais. INSA - Instituto Nacional de Saúde.

Morgenstern D (1956) Einfache Beispiele Zweidimensionaler Verteilungen. Mitteilingsblatt für Mathematische Statistik 8:234-235.

Muurinen JM (1982) Demand for health: A generalised Grossman model. Journal of Health Economics 1(1): 5-28.

Nelsen RB (2006) An Introduction to Copulas. 2nd. ed., New York, Springer.

Newhouse JP, Phelps CE, Marquis MS (1980) On Having Your Cake and Eating It Too - Econometric Problems in Estimating the Demand for Health-Services. Journal of Econometrics 13(3): 365-390.

OECD Health Data 2006, Organization for the Economic Co-operation and Development, Paris.

Patton AJ (2005) Estimation of Multivariate Models for Time Series of Possibly Different Lengths. Journal of Applied Econometrics 21(2): 147-173.

Quinn C (2007a) Using Copulas to Estimate Reduced-form Systems of Equations. *HEDG* Working Papers *07/25*, University of York.

Quinn C (2007b) The Health-economic Applications of Copulas: Methods in Applied Econometric Research. HEDG Working Papers 07/22, University of York.

Sklar A (1959) Fonctions de Répartition à *n* Dimensions et Leurs Marges. Publications de l'Institut de Statistique de l'Université de Paris 8:229-231.

Smith M (2003) Modeling Selectivity Using Archimedean Copulas. Econometrics Journal 6:99-123.

Train KE (2003) Discrete Choice Methods with Simulation. New York, Cambridge University Press.

Trivedi PK, Zimmer DM (2006) Copula Modeling: An Introduction for Practitioners. Foundations and Trends® in Econometrics 1(1):1-111.

Van Ourti T (2004) Measuring Horizontal Inequity in Belgian Health Care Using a Gaussian Random Effects Two-part Count Data Model. Health Economics 13:705-724.

Van Ophem H (1999) A general method to estimate correlated discrete random variables. Econometric Theory 15:228-237.

Wagstaff A (1986) The Demand for Health - Some New Empirical-Evidence. Journal of Health Economics 5(3):195-233.

Wagstaff A (1989) Econometric Studies in Health Economics - a Survey of the British Literature. Journal of Health Economics 8(1): 1-51.

Windmeijer FA, Santos-Silva JMC (1997) Endogeneity in count data models: An application to demand for health care. Journal of Applied Econometrics 1:281-294.

Winkelmann R (2004) Health Care Reform and the Number of Doctor Visits – An Econometric Analysis. Journal of Applied Econometrics 19:455-472.

Zimmer DM, Trivedi PK (2006) Using Trivariate Copulas to Model Sample Selection and Treatment Effects: Application to Family Health Care Demand. Journal of Business and Economic Statistics 24:63-76.