CRENoS

Centro Ricerche Economiche Nord Sud
Università di Cagliari
Università di Sassari

THE COPULA APPROACH TO SAMPLE SELECTION MODELLING: AN APPLICATION TO THE RECREATIONAL VALUE OF FORESTS

Margarita Genius
Elisabetta Strazzera

CONTRIBUTI DI RICERCA
03/08

Margarita Genius
Dept. of Economics, University of Crete, Greece

Elisabetta Strazzera
DRES and CRENoS, University of Cagliari, Italy

THE COPULA APPROACH TO SAMPLE SELECTION MODELLING: AN APPLICATION TO THE RECREATIONAL VALUE OF FORESTS

Abstract
The sample selection model is based upon a bivariate or a multivariate structure, and distributional assumptions are in this context more severe than in univariate settings, due to the limited availability of tractable multivariate distributions. While the standard FIML estimation of the selectivity model assumes normality of the joint distribution, alternative approaches require less stringent distributional hypotheses. As shown by Smith (2003), copulas allow great flexibility also in FIML models. The copula model is very useful in situations where the applied researcher has a prior on the distributional form of the margins, since it allows separating their modelling from that of the dependence structure. In the present paper the copula approach to sample selection is first compared to the semiparametric approach and to the standard FIML, bivariate normal model, in an illustrative application on female work data. Then its performance is analysed more thoroughly in an application to Contingent Valuation data on recreational values of forests.

JEL: C34, C51, H41, Q26
Keywords: Contingent Valuation, Selectivity Bias, Bivariate Models, Copulas

November 2003

1. Introduction

Endogenous sampling is a pervasive problem in applied microeconomics, especially in survey data analysis. Contingent Valuation surveys are no exception: it is often observed that only a sub-sample of respondents give information on their willingness to pay (WTP) for insuring provision of the good in the contingent market. When prices are blatantly over or understated, or when no answer is given at all, data are classified as "protest" responses. Selectivity effects could bias the estimates of WTP based on the truncated sample of valid responses, and in such case the valuation of the public good would be incorrect. Only recently this issue has been fully addressed in the Contingent Valuation literature: see Donaldson et al. (1998), Alvarez-Farizo et al. (1999), Kontoleon and Swanson (2002), Strazzera et al. (2002, 2003).

In an extensive survey on the topic of sample selection modelling, Vella (1998) affirms that "the ability to estimate and test econometric models over nonrandomly chosen sub-samples is unquestionably one of the more significant innovations in microeconometrics". While progress in the econometric analysis and treatment of sample selection cannot be denied, the debate is still open on what is the best procedure to be followed to obtain robust estimates from sample selection models. In general, Full Information Maximum Likelihood (FIML) estimates are recognized as the most efficient, as long as the underlying models are correctly specified. The proviso is important, since FIML sample selection models are typically based on the assumption of bivariate normality of the joint distribution, which implies that the marginals are themselves univariate normals. In many applications this assumption is unduly restrictive: in Contingent Valuation studies, the WTP distribution is often modelled as a non-normal: examples are the Logistic, the Weibull, the Gamma distribution.

In an effort to attain more flexibility in sample selection modelling, a conspicuous stream of research has focused on non-parametric or semi-parametric methods, which do not require stringent distributional assumptions. The problem is that these methods are

much more computationally burdensome than their parametric counterparts. Also, larger data sets are needed for these estimates to be reliable. Furthermore, the choice of the bandwidth can affect the resulting estimates: in particular, problems of overfitting have been reported when cross-validation techniques are used in conjunction with kernel estimates (Mroz and Savage 1999), and this is especially so in two-stage estimation problems. On the other hand, if no cross-validation or optimal criteria are used to select the bandwidth, then many estimation rounds using different bandwidths are needed to ensure the resulting estimates do not differ drastically across bandwidths. Another drawback of using semi-parametric methods to correct selectivity bias is that no estimate of the dependence is separately obtained. In some particular applications it might be of some importance to get estimates of the dependence between the participation and outcome decisions, and a parametric approach is well suited in these cases. Following results by Olsen (1980) and Lee (1982, 1983), 2-step parametric methods have been applied to sample selection models, which do not rely on distributional assumptions of joint normality. These models represent a flexible and simple method to correct selectivity. Unfortunately, the 2-step parametric estimator is especially susceptible to collinearity problems: see Nawata and Nagase (1996), Leung and Yu (1996, 2000), Puhani (2000). When a moderate level of collinearity is detected, the FIML method is recommended[1].

In order to loosen the restrictive BVN distributional assumption of the standard FIML model, Smith (2003) suggests use of the copula approach. Note that in addition to normal marginal distributions, the BVN specification imposes constraints on the type of dependence allowed between the two underlying error terms. Broadly speaking, a copula is a function that links separately specified marginals into a multivariate distribution on

[1] However, if collinearity is very high, two-part models, which maintain the outcome as conditionally independent of the participation choice, rather than sample selection models are preferable.

$[0,1]^n$. The copula representation of the multivariate distribution allows different specifications for the marginals and greater flexibility in the specification of the dependence, therefore bypassing some of the limitations of bivariate normality mentioned above. As it will be seen in the course of the paper, this is especially useful in situations where the researcher might have some prior knowledge of the marginal distributions and also when asymmetry and/or fat tails in the bivariate distribution are suspected.

A fairly well-known example of copula is the Lee (1983) inverse normal transformation: it consists in specifying non normal marginals, and transforming them into normal distributions by means of the inverse standard normal distribution function, so that a BVN can be used to model the joint distribution. Although this method allows great flexibility in the specification of the marginals, the type of dependence is restricted to linear correlation. Other copulas, allowing a wider range of dependency patterns, would be more suitable in many applications. Smith (cit.) indicates a special class of copulas, namely the Archimedean copulas, easy to implement and quite flexible to fit a variety of distributional shapes.

In this paper, we first show how the copula approach works in an illustrative example using previously published data (Martins, 2001) on female labour participation and wages. The copula parametric approach is compared to the semiparametric 2-step method that Martins suggests to correct selectivity bias in the wages estimates. Afterwards, we apply the copula approach to Contingent Valuation data on recreational values of forests. Several copula models, both Archimedean and non-Archimedean, are estimated, with the two-fold objective of checking different distributional hypotheses for the marginals, and different structures of dependency between them. It is shown that the joint distribution is well accommodated by an Archimedean copula (namely the Joe copula), which models a right-skewed joint distribution with logistic marginals.

The paper is organized as follows: the next section describes the copula models and their application to the sample selection

problem; section 3 shows how the copula approach works in comparison to the standard FIML, BVN model, and the semiparametric method on female labour data. The fourth section is devoted to the application of the copula approach to Contingent Valuation data on the recreational value of forests, characterized by selectivity bias due to protest responses to the WTP question. Several models are estimated, allowing testing of different dependence structures and distributional assumptions for the marginals. Section 5 concludes the paper.

2. The Copula Approach to Sample Selection

The structure of the sample selection model (in its simplest parametric form) is a two-equation system: the first equation is the *Selection equation*

$$\boldsymbol{Y}_{1i} = \begin{cases} 1 \ \boldsymbol{if} \ \boldsymbol{z}_i'\gamma + \varepsilon_i \geq 0 \\ 0 \ \boldsymbol{if} \ \ \boldsymbol{z}_i'\gamma + \varepsilon_i < 0 \end{cases} \qquad (1)$$

which determines the observability or not for all the members in the sample of the second equation, the
Outcome equation

$$\boldsymbol{Y}_{2i} = \boldsymbol{x}_i'\beta + \boldsymbol{u}_i \qquad (2);$$

where Y_{2i} is the dependent variable of principal interest, which is observed only when Y_{1i} =1; x_i and z_i are vectors of exogenous variables; β and γ are vectors of unknown parameters; ε_i and $\boldsymbol{u}_i$ are error terms with zero mean and with $E[\boldsymbol{u}_i \mid \varepsilon_i] \neq 0$.

Knowledge of the joint distribution of (u_i, ε_i), H, allows writing the log-likelihood of the full ML model as

$$\boldsymbol{l} = \sum_{\boldsymbol{Y}_{1i}=0} (1-\boldsymbol{I}_i) ln\, \boldsymbol{F}(-\boldsymbol{z}_i'\gamma) + \sum_{\boldsymbol{Y}_{1i}=1} \boldsymbol{I}_i\, ln \frac{1}{\sigma} \left(\boldsymbol{g}(\boldsymbol{y}) - \frac{\partial \boldsymbol{H}(0,\boldsymbol{y})}{\partial \boldsymbol{y}} \right)_{\boldsymbol{y}=\frac{\boldsymbol{y}_{2i}-\boldsymbol{x}_i'\beta}{\sigma}} \tag{3}$$

where g is the pdf of u_2. This model was originated in Gronau (1974) and Heckman (1974), who specified H as a Bivariate Normal. This distributional assumption is still the paradigm in FIML sample selection modelling, due to ease of implementation and relative flexibility in modelling correlation[2]. Unfortunately, distributional misspecification will, in general, produce inconsistent estimates of the parameters: see Vella (cit.) for a thorough discussion.

A recent trend is to relax the normality assumption by using semiparametric methods, which do not impose parametric forms on the error distribution. As explained in the introduction of this paper, this strategy imposes several costs. Lee (1982, 1983) suggests a different approach: even if the stochastic parts of the two equations are specified as non-normal, they can be transformed into random variables that are characterized by the bivariate normal distribution. This transform, which involves the use of the inverse standard normal distribution, is an example of a bivariate *copula function*, which is defined as follows:

Definition: A *2-dimensional copula* is a function $\mathbf{C : [0,1]^2 \rightarrow [0,1]}$, with the following properties:

For every $\boldsymbol{u} \in [0,1]$, $\boldsymbol{C(0,u) = C(u,0) = 0}$;
For every $\boldsymbol{u} \in [0,1]$, $\boldsymbol{C(u,1) = u}$ *and* $\boldsymbol{C(1,u) = u}$;
For every $(u_1, v_1), (u_2, v_2) \in [0,1] \times [0,1]$ *with* $u_1 \le u_2$ *and* $v_1 \le v_2$:

[2] As opposed, for example, to the bivariate logistic that restricts correlation to a narrow range: $\left[-3/\pi^2 , 3/\pi^2 \right]$.

$$C(u_2, v_2) - C(u_2, v_1) - C(u_1, v_2) + C(u_1, v_1) \geq 0 .$$

The last condition is the two-dimensional analogue of a nondecreasing one-dimensional function.
The theoretical basis of multivariate modeling by copulas is provided by a theorem due to Sklar (1959).

Sklar's Theorem

Let $\boldsymbol{H}$ *be a joint distribution function with margins* $\boldsymbol{F_1}$ *and* $\boldsymbol{F_2}$*, which are, respectively, the cumulative distribution functions of the random variables* $\boldsymbol{x_1}$ *and* $\boldsymbol{x_2}$*. Then there exists a function* $\boldsymbol{C}$ *such that* $\boldsymbol{H(x_1, x_2) = C(F_1(x_1), F_2(x_2))}$*, for every* $\boldsymbol{x_1, x_2} \in \overline{\boldsymbol{R}}$ *, where* $\overline{\boldsymbol{R}}$ *represents the extended real line. Conversely, if C is a copula and* $\boldsymbol{F_1}$ *and* $\boldsymbol{F_2}$ *are distribution functions, then the function H defined above is a joint distribution function with margins* $\boldsymbol{F_1}$ *and* $\boldsymbol{F_2}$*.*

Since the copula function "links a multidimensional distribution to its one-dimensional margins" (Sklar, 1996), the name "copula" (connection) is explained. This approach ensures a high level of flexibility to the modeler, since the specification of the margins $\boldsymbol{F_1}$ and $\boldsymbol{F_2}$ can be separated from the specification of the dependence structure through the function *C* and an underlying parameter θ, which governs the intensity of the dependence.
The aforementioned Lee's inverse normal transformation corresponds to specifying a bivariate normal copula with non-normal margins. Although it is computationally straightforward, and flexible in the specification of the marginals, its use in empirical works has been relatively scant: the reason may be that the type of dependence allowed for by this copula is restricted to linear correlation. Other copula functionals allow greater flexibility in the dependence structure. In consideration of their simple mathematical structure, Smith (cit.) advocates use of Archimedean copulas for application to selectivity models.
Archimedean copulas are functions generated by an additive continuous, convex decreasing function φ, with φ(1)=0. If, in

addition, $\varphi(0)=\infty$, the generator is *strict*. In general, Archimedean copulas have the following form:

$$\varphi(C_\theta(u,v)) = \varphi(u) + \varphi(v).$$

The additive structure of copulas in this class makes estimation of the maximum likelihood, and calculation of the score function, relatively easy. Furthermore, the family is sufficiently large so as to allow a wide range of distributional shapes (right or left skewness, fat or thin tails, etc.).

Another characteristic of copulas that can be valuable to the applied researcher is the capability of accommodating both positive and negative dependence. Copulas ranging from the lower Fréchet bound (perfect negative dependence as $\theta \to -\infty$) to the upper Fréchet bound (perfect positive dependence as $\theta \to \infty$) are said *comprehensive*. A measure of dependence commonly used in econometrics applications is linear correlation; however, this measure is valid only when dealing with elliptical copulas (such as the BVN). Alternative measures of dependence include Kendall's $\boldsymbol{\tau}$ (K_τ) and Spearman's $\boldsymbol{\rho}$ (S_ρ), which are measures of concordance[3]. The former is defined as follows:

$$\boldsymbol{K}_\tau = \boldsymbol{P}\left((\boldsymbol{X}-\widetilde{\boldsymbol{X}})(\boldsymbol{Y}-\widetilde{\boldsymbol{Y}})>0\right) - \boldsymbol{P}\left((\boldsymbol{X}-\widetilde{\boldsymbol{X}})(\boldsymbol{Y}-\widetilde{\boldsymbol{Y}})<0\right).$$

Another expression for K_τ is in terms of copulas (see Nelsen, cit., p. 129):

$$\boldsymbol{K}_\tau = 4\iint_{I0,1I^2} \boldsymbol{C}(\boldsymbol{u},\boldsymbol{v})\boldsymbol{dC}(\boldsymbol{u},\boldsymbol{v}) - 1,$$

that is the expression we will use to compute it when a closed form expression is not available. The measure proposed by Spearman is given by

$$\boldsymbol{S}_\rho = 3\left(\boldsymbol{P}\left((\boldsymbol{X}-\widetilde{\boldsymbol{X}})(\boldsymbol{Y}-\boldsymbol{Y}')>0\right) - \boldsymbol{P}\left((\boldsymbol{X}-\widetilde{\boldsymbol{X}})(\boldsymbol{Y}-\boldsymbol{Y}')\right)<0\right)$$

where $(\boldsymbol{X},\boldsymbol{Y}),(\widetilde{\boldsymbol{X}},\widetilde{\boldsymbol{Y}})$ and $(\boldsymbol{X}',\boldsymbol{Y}')$ are three independent random vectors with a common distribution function H whose margins are F and G.

---

[3] Other measures of dependence rely on the criterion of dependence between random variables: for a definition, see Nelsen (cit.) p. 170.

Also in this case we have a copula expression:

$$\boldsymbol{S}_{\rho} = 12\iint_{[0,1]^2} \boldsymbol{uvdC(u,v)} - 3$$

For continuous random variables the above measures are measures of concordance, which implies that they take values in [-1,1], taking the value zero when we have independence (see Nelsen, cit., p. 136 for a definition of concordance measure). Spearman's ρ can be interpreted as a correlation coefficient between the cdfs of the two variables. We recall that the linear (or Pearson) correlation is not a measure of dependence: for example, $\boldsymbol{\rho(x,y)} = 0$ does not imply independence of the two variables.

The table below gives the functional form of selected copulas:

Table 1. Functional form of Copulas

Family	$\boldsymbol{C(u,v)}$	**Range of θ**	**Range of K_τ and S_ρ**	**θ_{indep}**
Product	uv		0	
Lee	$\Phi_2(\Phi^{-1}(u),\Phi^{-1}(v);\theta)$	[-1,1]	[-1,1]	0
Clayton	$[u^{-\theta}+v^{-\theta}-1]^{-1/\theta}$	$(0,\infty)$	[0,1]	0^+
Frank	$-\frac{1}{\theta}\ln\left(1+\frac{(e^{-\theta u}-1)(e^{-\theta v}-1)}{e^{-\theta}-1}\right)$ uv	$(-\infty,\infty)\setminus\{0\}$ 0	[-1,1]\0 0	0
Gumbel	$\exp\left(-\left((-\ln u)^{\theta}+(-\ln v)^{\theta}\right)^{1/\theta}\right)$	$[1,\infty)$	[0,1]	1
Joe	$1-\left((1-u)^{\theta}+(1-v)^{\theta}-(1-u)^{\theta}(1-v)^{\theta}\right)^{1/\theta}$	$[1,\infty)$	[0,1]	1
AMH	$\frac{uv}{(1-\theta(1-v)(1-u))}$	$[-1,1)$	[-0.18,0.33] [-0.27,0.47]	0
FGM	$uv(1+\theta(1-u)(1-v))$	$[-1,1]$	[-0.22,0.22] [-0.33,0.33]	0
Plackett	$\frac{[1+(\theta-1)(u+v)]}{2(\theta-1)} - \frac{\sqrt{[1+(\theta-1)(u+v)]^2-4uv\theta(1-\theta)}}{2(\theta-1)}$ uv	$(0,\infty)$ 0	[-1,1]	0

It can be observed that the FGM copula allows only for a limited degree of dependence (Kendall's τ is restricted to [-2/9,2/9] and Spearman's ρ to [-1/3,1/3]), which reduces its appeal for use in applications. Similar considerations hold also for the AMH, whose range for Kendall's τ is restricted to [-0.181,0.333] and for Spearman's ρ to [-0.271,0.478]. In contrast, the Frank and Plackett copulas are comprehensive, including the lower and upper Fréchet bounds and the independent copula. They both are symmetric, with thinner (Plackett) or fatter (Frank) tails than the BVN. In some applications symmetry may be an undesirable feature, and asymmetric copulas may be preferred. The Clayton copula exhibits asymmetry in the sense that there is a clustering of values in the left tail of the joint distribution: exactly the opposite to the Joe copula, which exhibits a strong clustering of values in the right tail. The Gumbel copula is similar to the Joe, but with a thinner tail. Unfortunately, the last three copulas, just as the most part of Archimedean copulas (one exception is the Frank copula), are monotonic: they cannot accommodate for negative dependence. Figures 1 and 2 show the plots of some copulas (Clayton, Lee, Gumbel, Joe) based on standard Normal and Logistic marginals, and the BVN standard model.

3. An illustrative example: sample selection modelling on female labour supply data

In a study published by the Journal of Applied Econometrics (2001) Martins applies both parametric and semiparametric methods to the estimation of the participation and wage equations for married women in Portugal. The author shows that the 2-step semiparametric estimator is more efficient than the parametric ML estimator. The parametric model is based on a wrong assumption of bivariate normality for the joint distribution function: testing for normality of residuals in the participation equation leads to rejection of the hypothesis. Estimation of a 2-step semiparametric

model is shown to produce more efficient estimates. In the following we show how the copula approach works in this context.

The data set is a sample from the Portuguese Employment Survey, interview year 1991. The sample used in the analysis consists of 2339 observations on married women, 1400 of whom were employed. Martins estimates a participation equation, regressing the dependent variable (which takes a value 1 if the woman participates in the labour force, and zero otherwise) on the following regressors: AGE (age in years), AGE2 (age squared), EDU (years of education), CHILD (the number of children under 18 in the household), YCHILD (number of children under the age of 3) LHUSWG (log of husband's wage). The outcome equation regresses the log of wages on the following variables: PEXP (potential experience years, calculated as age-edu-6), PEXP2 (PEXP squared), PEXPCHD (PEXP multiplied by CHILD), PEXPCHD2 (PEXP2 multiplied by CHILD). The results are summarized in table 2: the first two columns contain Martins' estimates of the parametric (FIML, BVN) model and of the 2-step semiparametric model, respectively in the first and in the second column. The standard errors reported in table 2 for the BVN model are calculated from the inverse of the computed Hessian, and differ slightly from those reported by Martins, calculated from the cross product of the first derivatives. In the selection equation, the husband's wage seems to have no significant effect on the decision to participate in the labour market, while in the wage equation the only coefficient that is significant at the 5% level is the educational attainment. Martins shows that the HH test (Horowitz and Härdle, 1994) rejects the Probit for the participation equation at the 5% level at bandwidth greater than 0.55, and argues that a semiparametric approach can be useful to overcome the misspecification problem. The estimates of the selection equation parameters in the semiparametric model can be obtained up to a factor of proportionality (i.e. one of the coefficients is normalized to one), so they are not directly comparable to the competing models; it can be noticed however

that the coefficient of the husband wage becomes significant in the semiparametric model. Focusing on the wage equation, significant estimates are obtained for the educational level and the two variables related to potential experience, while the 5% level of significance is not attained for the two interaction terms between potential experience and children.

The semiparametric estimator imposes a heavy computational load in comparison to the FIML method. We show now how the copula approach allows fairly easy estimations while relaxing the constraints imposed by the standard BVN model. As a first step, the margins should be specified, based on some explorative analysis of the data, or theoretical priors. For the selection equation, applying the HH test to the Logit specification, we observe that it is not rejected at the 5% level up to bandwidth h=0.9, and is not rejected at 10% level for bandwidth h=1: the Logistic could be a candidate for the error distribution in the participation model. For the wage equation, a Pagan-Vella (1989) test indicates a strong departure from normality. Heckman et al. (2001), considering that wage distributions are often fat tailed, argue that "the family of Student-t_ν distributions offers an attractive and potentially more appropriate class of models for the treatment parameters than those implied by the benchmark Normal model". We then choose a logistic distribution for the participation equation, and a Student-t_ν distribution for the wage equation, and estimate different copula models based on these marginals. In the last column of table 4 we report the estimates obtained from the Joe copula model. The parameter ν of the t_ν distribution is estimated along with the other parameters. Its value, about 3, indicates very heavy tails in the distribution: we recall that for ν=1 the t distribution is a Cauchy, while for ν >30 it approximates a Normal. In the selection equation, the husband's wage is significant at the 5% level; in the wage equation the two interaction terms between potential experience and children are not statistically significant, while all the other estimates are significant at the 1% level. These results are close to those

obtained with the 2-step semiparametric estimator, but they have been obtained with less computations than those required by the semiparametric approach, since the latter entails approaching the estimation as a two-step procedure and trying several bandwidths both for the first step estimates and for the constant term of the wage equation. In addition, the copula approach allows estimation of the dependence structure, which is not estimated in the semiparametric model. The approach using copulas can very easily be implemented using any software that allows for user specified likelihood functions such as GAUSS, LIMDEP, STATA, or even EVIEWS. Model selection criteria such as Akaike or tests such as Vuong (1989) can be used as an aid in selecting between any two competing models. In the example above, the Akaike and Schwarz information criteria which use a penalization for the number of parameters in a model as well as the Vuong test favor the Joe copula with logistic and t_ν marginals over the standard bivariate normal model (Vuong's statistic is 8.7 and the test is asymptotically normal).

When the hypothesis of bivariate normality for the joint distribution is not satisfied, and collinearity problems prevent from using the parametric 2-step procedure, the copula approach can be a useful alternative to the semiparametric method. In cases where departures from the marginals specified in the copula function are minor, small losses in consistency are traded-off for bigger efficiency. If larger departures are detected, the copula approach allows a better fitting model to be chosen among a wide range of marginal distributions and dependence structures.

4. Contingent Valuation Analysis of Recreational Values of Forests

In the following we present an application of the copula approach to the analysis of data on recreational benefits provided by forests and woodlands in Scotland. The study was conducted by the

Queens University Belfast[4]: a detailed description of the survey can be found in Strazzera et al. (cit.), so we report here only a brief summary.

The questionnaires were administered on-site in selected forest and woodlands sites used for recreation, through face-to-face interviews. Individuals were asked various questions aimed at conveying information about their demographic and socio-economic characteristics, interests and hobbies, previous excursions to forests, and details on the present visit. Afterwards, they were asked if they would be willing to pay a given entry fee (bid) to the forest, were this the only possibility to maintain public access to the forest. The fee was supposed to be paid by the respondent for each person in the party. The initial bid amounts t used were uniformly distributed across visitors, and were chosen on the basis of initial estimates of the WTP distribution obtained from extensive pilot studies. Next, individuals were asked the exact amount they would be willing to pay as an entry charge to the forest for each component of the party.

Table 3 gives summary statistics for the data used in this analysis: mean and standard deviation of the covariates for the full sample, and for the sub-sample of non protesters. Full descriptions of these variables are given in Appendix. It can be seen that there are 535 protest responses, which amounts to 18% of the sample.

The models are estimated using different covariate specifications related to the effect of socio-economic or personal characteristics, such as income, education, age, sex; or features of the visit, such as the number and age of components of the party, expenses for parking or food, activities engaged in during the visit, previous visit experiences. We first estimate a standard FIML model, based on the assumption of bivariate normality of the joint distribution: column 1 of Table 4 reports the parameter estimates for the best fitting regressions for the two equations (participation and

[4] We are grateful to George Hutchinson for kind permission to use the data for further analysis.

valuation), selected by means of likelihood ratio tests for nested specifications from more comprehensive models.

The explanatory variables in the participation equation are: the amount the individual was asked to pay at the first stage of the elicitation process (i.e. the bid multiplied by the number of people in the party); the number of visits to the forest where the interview took place, or to other forest sites during the past year; time spent in the forest; parking expenditure; income (class 2); and a dummy variable indicating whether the individual was alone or in a party when visiting the forest. It can be observed that higher tendered bids induce a higher probability of a protest response. People who frequently visit forests are also more probably protesters, and this can be explained as a reaction to the reallocation of their property rights (in the Coasian sense). On the other hand, people who spent more time in the forest are less likely to protest, as well as people who paid a parking fee for the current visit, while the effect of income is not clear-cut.

The valuation equation specifies log WTP as the dependent variable. The results indicate that more frequent visitors to the forest are willing to pay less (as it is obvious for a downward sloping demand curve). Time spent at the site and the appreciation of the recreational benefits given by the forest have, as expected, a positive effect. Also parking expenditures are positively correlated with stated WTP, and this can be easily explained by considering that the object of the elicitation question was a ticket inclusive of parking fees. Income has also the expected effect since the lower income categories are willing to pay less on average; males are willing to pay more than females. The negative estimate for the coefficient of Children seems to indicate that respondents placed lower values for children in their party; but the effect must be somehow counter-balanced, since the coefficient estimate for party size close to one indicates that there is some proportionality between the total amount the respondent is willing to pay and the number of people in the pool.

Although this model does not show evident symptoms of misspecification (namely, instability of the coefficient estimates, and the correlation coefficient close to its boundary), we wish to investigate the tenability of the assumption of bivariate normality for the joint distribution. We first maintain the hypothesis of normal marginals, and check the structure of dependence between the two equations. In column 2 of table 4 we only report results for the three best fitting copulas: Frank, Gumbel and Joe, but all the copula models included in Table 1 were estimated, except the Lee copula which in this case of normal marginals is equivalent to the BVN. We could observe that all copulas have a better fit than the benchmark BVN model, which suggest a dependence structure between the two equations more complex than just linear correlation. Since the estimation results suggest positive dependence (all comprehensive copulas estimate positive dependence), monotonic Archimedean copulas are applicable. The performance of the Gumbel and Joe copulas suggests that the joint distribution is probably skewed to the right. The three selected copulas give similar estimates for the covariate coefficients, but it can be observed that the estimated dependence (as measured by the Kendall's or the Spearman's parameters) is higher in the Joe copula, which also has the better fit according to the Akaike's criterion.

The following step involves the analysis of the distributional specification of the two margins. Following Martins (cit.), we use both the Horowitz (1993) and Horowitz and Härdle (cit.) tests for the normality assumption for the selection equation. For the valuation equation we apply the Pagan-Vella test for normality. While the results of the latter (F-statistic: 2.81, p-value 0.037) would lead to rejection of the hypothesis of normality for the valuation equation, the outcome from the former tests is not so clear-cut. The HH test does not reject the probit model for the participation equation at all selected bandwidths; the Horowitz test at bandwidth h=1 rejects the Probit (Figure 1), while at the same bandwidth the Logit is not rejected (Figure 2).

After estimating the model under different distributional specifications (Normal, Logistic, Extreme Value) for either margin, we select the logistic-logistic specification as the one giving the best fit as measured by the Akaike and Schwarz criteria. The last columns of Table 4 report results for the best fitting model, i.e. the Joe copula, which under all distributional assumptions performed better than the competing models. Its opposite, the Clayton copula, is also reported for demonstrative purposes. We also show results for the Lee copula, since it is fairly well known in the econometrics literature: recent applications in sample selection modelling include Von Ophem (2000) and Heckman et al. (cit.). Parameter estimates do not change dramatically across copulas, but it can be observed that for most parameters the Joe and the Clayton copulas show departures in opposite directions from the benchmark estimates. The estimate of θ in the Clayton copula, and its associated standard error, would indicate lack of dependence; however, this is due to the fact that the type of left tail clustering assumed by this copula is not compatible with our data, and the value of the log-likelihood confirms the relatively bad fit. The parameter θ is not directly comparable across copulas, but Kendall's τ and Spearman's ρ are. Again, the Akaike and Schwarz criteria indicate the Joe copula, which exhibits the highest degree of dependence, as the best fitting model

Table 5 reports the estimates and confidence intervals for the measures of central tendency obtained from the benchmark BVN and the alternative copula models referred to above. Since the parameter estimates do not differ much across models, the mean and median values estimates obtained from them are also very close. It can be observed that the Clayton copula estimates are slightly biased upward, and less precise than all competing models (wider confidence intervals both for the mean and median values). It is remarkable that the mean and median estimates produced by the Joe copula with logistic marginals, i.e. the best fitting copula, are very close to those produced by the BVN model, but with

tighter confidence intervals. The plots reported in Figure 5 are useful to explain this result: while the fitted Joe copula exhibits some skewness and fatter tails with respect to the fitted BVN, yet the divergence is not dramatic. The advantage of using the copula approach in this application is the gain in the precision of the estimates. In cases where departures from the bivariate normal assumption are more serious than in the present application, more conspicuous differences in the punctual estimates are to be expected.

5. Conclusions

The copula representation of the bivariate distribution underlying the sample selection model allows different specifications for the marginals and great flexibility in the specification of the dependence. In a recent paper Smith (cit.) suggests the use of copula functions, and in particular Archimedean copulas, to correct selectivity bias in data affected by endogenous sampling. In this paper we show that copula models are indeed efficient, flexible and easy tools to deal with sample selection. First, we compared the copula approach to the standard FIML method and to the semiparametric method. Using data published by Martins (cit.), we could show that the copula approach produces estimates for the covariate coefficients similar to those obtained from the semiparametric approach, while giving more information on the dependence structure, and requiring less computational effort. We then applied the copula approach to Contingent Valuation data, collected to assess the use value of forests for recreation. This data had been modelled in a previous paper (Strazzera et al., cit.) by means of standard parametric sample selection models: it was found that, given the moderate level of collinearity present in the data, the FIML model was to be preferred to the Heckman's 2-step parametric model. Here, the tenability of the assumption of bivariate normality implicit in the standard FIML model is checked, and it is found that, while no clear-cut results are obtained for the participation equation, the hypothesis of

normality for the distribution of errors in the outcome equation is rejected. Since this is sufficient to reject the BVN hypothesis, the copula approach is applied to analyse and test different hypotheses on both the dependence structure and the distributional shape of the margins. Several copula models were estimated, and the best fitting model was selected according to the Akaike and the Schwartz criteria: it is a Joe copula, i.e. a model suitable for asymmetric, right-tailed joint distributions, which links two logistic distributions. It is shown that the copula model produces more precise estimates, even though it can be noticed that the punctual estimates are not much different from those obtained from the benchmark model. We argue that in circumstances where the misspecification of the BVN model is stronger than in the present application, it can be expected that the punctual estimates would differ more markedly across models.

References

Alvarez-Farizo B., N. Hanley, R.E. Wright and D. MacMillan (1999), "Estimating the Benefits of Agri-Environmental Policy: Econometric Issues in Open-Ended Contingent Valuation Studies," *Journal of Environmental Planning and Management,* 42(1): 23-43.

Donaldson, C., A.M. Jones, T. Mapp and J.A. Olson (1998), "Limited Dependent Variables in Willingness to Pay Studies: Applications in Health Care," *Applied Economics*, 30: 667-77.

Gronau, R. (1974), "Wage Comparisons - A Selectivity Bias," *Journal of Political Economy,* 82:
1119-1144.

Heckman, J. (1974), "Shadow Prices, Market Wages, and Labor Supply," *Econometrica*, 42: 679-693.

Heckman, J., J.L. Tobias and E. Vytlacil (2001), "Simple Estimators for Treatment Parameters in a Latent Variable Framework with an Application to Estimating the Returns to Schooling," *NBER working paper* W7950.

Horowitz JL. (1993), "Semiparametric estimation of a work-trip mode choice model," *Journal of Econometrics* 58: 49–70.

Horowitz JL., Härdle W. (1994), "Testing a parametric model against a semiparametric alternative," *Econometric Theory* 10: 821–848.

Kontoleon A. and T.M. Swanson (2003), "The WTP for Property Rights for the Giant Panda: Can a Charismatic Species be an Instrument for Nature Conservation?," forthcoming in Land Economics.

Lee, L.-F. (1983), "Generalized Econometric Models with Selectivity," *Econometrica* 51: 507–12.

Lee, L. F. (1982), "Some Approaches to the Correction of Selectivity Bias," *Review of Economic Studies*, 49: 355-372.

Leung, S. F. and S. Yu (1996), "On the Choice between Sample Selection Models and Two-part Models," *Journal of Econometrics*, 72: 197-229.

Leung, S. F. and S. Yu (2000), "Collinearity and Two-Step Estimation of Sample Selection Models: Problems, Origins, and Remedies," *Computational Economics*, 15: 173-199.

Martins Fraga M. (2001), "Parametric and Semiparametric Estimation of Sample Selection Models: An Empirical Application to the Female Labour Force in Portugal," *Journal of Applied Econometrics,* 16: 23-39.

Mroz, T. A. and T. H. Savage (1999), "Overfitting and Biases in Nonparametric Kernel Regressions Using Cross Validated Bandwidths: a Cautionary Note," *working paper, University of North-Carolina, Chapel Hill,* 99-10.

Nawata, K. and N. Nagase (1996), "Estimation of Sample Selection Bias Models," *Econometric Reviews*, 15(4): 387-400.

Nelsen, R. B. (1999). *An Introduction to Copulas*. Lecture Notes in Statistics. New York: Springer-Verlag.

Olsen, R. (1980), "A Least Squares Correction for Selectivity Bias," *Econometrica* 48: 1815-1820.

Pagan, A. R. and F. Vella (1989), "Diagnostic Tests for Models Based on Individual Data: A Survey," *Journal of Applied Econometrics*, 4 Supplement: S29-S59.

Puhani, P.A. (2000), "The Heckman Correction for Sample Selection and its Critique," *Journal of Economic Surveys,* 14(1): 53-67.

Sklar, A. (1959), "Fonctions de Répartition à N Dimensions et Leurs Marges," *Publ. Inst. Statis. Univ. Paris,* 8: 229-231.

Sklar A. (1996) *Random Variables, Distribution Functions, and Copulas - a Personal Look Backward and Forward,* published in: *Distributions with Fixed Marginals and Related Topics*, edited by L. Rüschendorf, B. Schweizer, and M.D. Taylor, Institute of Mathematical Statistics, Hayward, CA, pages 1-14.

Smith M.D. (2003), "Modelling Sample Selection Using Archimedean Copulas", *Econometrics Journal*, 6: 99-123.

Strazzera, E., R. Scarpa, P. Calia, G. Garrod, K. Willis (2003): "Modelling Zero Values and Protest Responses in Contingent Valuation Surveys," *Applied Economics*, 35(2): 133-38.

Strazzera E., M. Genius, R. Scarpa and G. Hutchinson (2003), "The Effect of Protest Votes on the Estimates of Willingness to Pay for Use Values of Recreational Sites," *Environmental and Resource Economics,* 25: 461-476.

Vella, F. (1998), "Estimating Models with Sample Selection Bias: A Survey," *The Journal of Human Resources*, 33(1): 127-169.

Vuong, Q.H., 1989, "Likelihood Ratio Tests for Model Selection and Non-Nested Hypothesis," *Econometrica***,** 57(2): 307-333.

Table 2: Estimates of BVN, Semiparametric and Copula Models for Female Labour Participation and Wages

Variables	BVN		2-Step Semiparametric		Joe: Logistic & t-Student	
	Coeff.	(S.E.) p-value	Coeff.	(S.E.) p-value	Coeff.	(S.E.) p-value
CONST	-0.570	(0.937) 0.539			-0.740	(1.395) 0.596
CHILD	-0.120	(0.028) 0.000	-0.097	(0.012) 0.000	-0.187	(0.045) 0.000
YCHILD	-0.090	(0.074) 0.223	-0.018	(0.04) 0.653	-0.113	(0.109) 0.301
LHUSWG	-0.100	(0.077) 0.181	-0.078	(0.03) 0.009	-0.232	(0.112) 0.039
EDU	0.150	(0.010) 0.000	0.086	(0.012) 0.000	0.289	(0.018) 0.000
AGE	0.810	(0.253) 0.001	1		1.394	(0.389) 0.000
AGE2	-0.120	(0.031) 0.000	-0.145	(0.003) 0.000	-0.206	(0.048) 0.000
CONST	4.480	(0.089) 0.000	4.800	(1.700) 0.005	4.139	(0.075) 0.000
EDU	0.110	(0.005) 0.000	0.090	(0.015) 0.000	0.133	(0.003) 0.000
PEXP	0.130	(0.058) 0.087	0.410	(0.133) 0.002	0.379	(0.060) 0.000
PEXP2	-0.003	(0.014) 0.875	-0.060	(0.030) 0.045	-0.055	(0.012) 0.000
PEXPCHD	0.032	(0.035) 0.148	0.040	(0.026) 0.124	-0.000	(0.015) 0.977
PEXPCHD2	-0.010	(0.011) 0.078	-0.017	(0.010) 0.089	-0.003	(0.004) 0.489
σ	0.550	(0.015) 0.000			0.347	(0.019) 0.000
θ	0.350	(0.100) 0.000			2.782	(0.254) 0.000
K_τ	0.231					
S_ρ	0.340					
ν					2.953	(0.320) 0.000
Log-lik	-2488					-2334

Table: 3. Means and standard deviations (in parenthesis) by groups of respondents

	FULL SAMPLE	NON PROTESTERS
Mean WTP (£)	…	4.23(3.6)
Median WTP	…	3
Children	0.88 (1.08)	0.88 (1.076)
Alone	0.07 (0.26)	0.06 (0.23)
Time	4.71 (0.75)	4.77 (0.73)
Parking	0.23 (0.48)	0.26 (0.51)
Past	1.51 (1.35)	1.39 (1.23)
Other	1.40 (1.26)	1.35 (1.22)
Improved	0.92 (0.27)	0.92 (0.26)
Income		
1: <16000	0.32 (0.47)	0.31 (0.46)
2: 16000-30000	0.47 (0.50)	0.49 (0.50)
Male	0.65 (0.48)	0.65 (0.48)
Sample size	2964	2429

Table 4. Estimates of BVN and Copula Models for Protest and WTP data for Forests

Variables	BVN	F and G normal			F and G logistic		
		Frank	Gumbel	Joe	Lee	Clayton	Joe
Constant	0.743	0.679	0.698	0.695	1.213	1.194	1.156
	(0.201)	(0.201)	(0.204)	(0.205)	(0.355)	(0.353)	(0.361)
Bid1	-0.354	-0.348	-0.357	-0.358	-0.629	-0.595	-0.637
	(0.036)	(0.035)	(0.035)	(0.035)	(0.064)	(0.066)	(0.063)
Alone	-0.636	-0.590	-0.606	-0.597	-1.086	-1.106	-1.049
	(0.107)	(0.106)	(0.106)	(0.105)	(0.182)	(0.183)	(0.179)
Time	0.193	0.202	0.201	0.201	0.345	0.341	0.354
	(0.039)	(0.039)	(0.039)	(0.040)	(0.070)	(0.069)	(0.071)
Park	0.584	0.577	0.580	0.583	1.227	1.208	1.231
	(0.094)	(0.945)	(0.094)	(0.093)	(0.209)	(0.207)	(0.207)
Past	-0.134	-0.132	-0.135	-0.133	-0.237	-0.231	-0.240
	(0.021)	(0.021)	(0.021)	(0.021)	(0.037)	(0.037)	(0.037)
Other	-0.070	-0.062	-0.063	-0.061	-0.116	-0.126	-0.104
	(0.021)	(0.021)	(0.021)	(0.021)	(0.038)	(0.038)	(0.038)
Inc2	0.168	0.158	0.165	0.162	0.282	0.284	0.278
	(0.057)	(0.057)	(0.057)	(0.057)	(0.102)	(0.102)	(0.101)
Constant	-0.666	-0.734	-0.717	-0.717	-0.632	-0.543	-0.647
	(0.113)	(0.115)	(0.114)	(0.114)	(0.113)	(0.113)	(0.112)
Children	-0.074	-0.077	-0.076	-0.078	-0.077	-0.074	-0.080
	(0.018)	(0.018)	(0.018)	(0.018)	(0.018)	(0.018)	(0.018)
Time	0.184	0.194	0.192	0.194	0.181	0.171	0.187
	(0.019)	(0.019)	(0.019)	(0.019)	(0.019)	(0.019)	(0.019)
Park	0.267	0.282	0.277	0.273	0.283	0.265	0.282
	(0.028)	(0.028)	(0.028)	(0.028)	(0.026)	(0.026)	(0.025)
Past	-0.115	-0.123	-0.121	-0.121	-0.121	-0.111	-0.124
	(0.012)	(0.012)	(0.012)	(0.012)	(0.012)	(0.012)	(0.012)
Male	0.067	0.069	0.068	0.068	0.078	0.078	0.080
	0.028)	(0.028)	(0.028)	(0.027)	(0.027)	(0.027)	(0.027)
Party	0.937	0.934	0.936	0.938	0.938	0.940	0.940
	(0.046)	(0.046)	(0.046)	(0.047)	(0.045)	(0.045)	(0.045)
Improved	0.190	0.189	0.190	0.186	0.166	0.160	0.161
	(0.050)	(0.051)	(0.050)	(0.050)	(0.052)	(0.052)	(0.052)
Inc1	-0.181	-0.182	-0.181	-0.181	-0.183	-0.185	-0.183
	(0.037)	(0.035)	(0.037)	(0.037)	(0.037)	(0.037)	(0.037)

Table 4 Continuous							
Inc2	-0.142 (0.035)	-0.134 (0.035)	-0.136 (0.035)	-0.137 (0.035)	-0.140 (0.034)	-0.152 (0.034)	-0.140 (0.034)
σ	0.649 (0.011)	0.658 (0.012)	0.652 (0.012)	0.639 (0.010)	0.367 (0.007)	0.364 (0.008)	0.356 (0.006)
θ	0.287 (0.074)	3.203 (0.727)	1.455 (0.130)	1.954 (0.308)	0.337 (0.078)	0.115 (0.109)	1.760 (0.193)
K_τ	0.185	0.325	0.313	0.345	0.219	0.054	0.297
S_ρ	0.275	0.473	0.449	0.491	0.323	0.081	0.428
Log-lik	-3606	-3601	-3601	-3600	-3590	-3596	-3584

Table 5: Means and Standard Deviations from BVN and Copula Models

	BVN	F and G normal			F and G logistic		
		Frank	Gumbel	Joe	Lee	Clayton	Joe
Mean WTP	3.518	3.424	3.446	3.444	3.591	3.738	3.550
C.I. Mean							
>	3.392	3.300	3.323	3.323	3.464	3.601	3.433
<	3.645	3.549	3.568	3.566	3.717	3.875	3.667
Median WTP	2.851	2.757	2.786	2.808	2.848	2.973	2.855
C.I. Med.							
>	2.739	2.640	2.673	2.700	2.736	2.843	2.762
<	2.962	2.874	2.900	2.916	2.959	3.103	2.949

Figure 1. Plots of BVN, Gumbel, Joe and Clayton Copulas: Normal marginals.

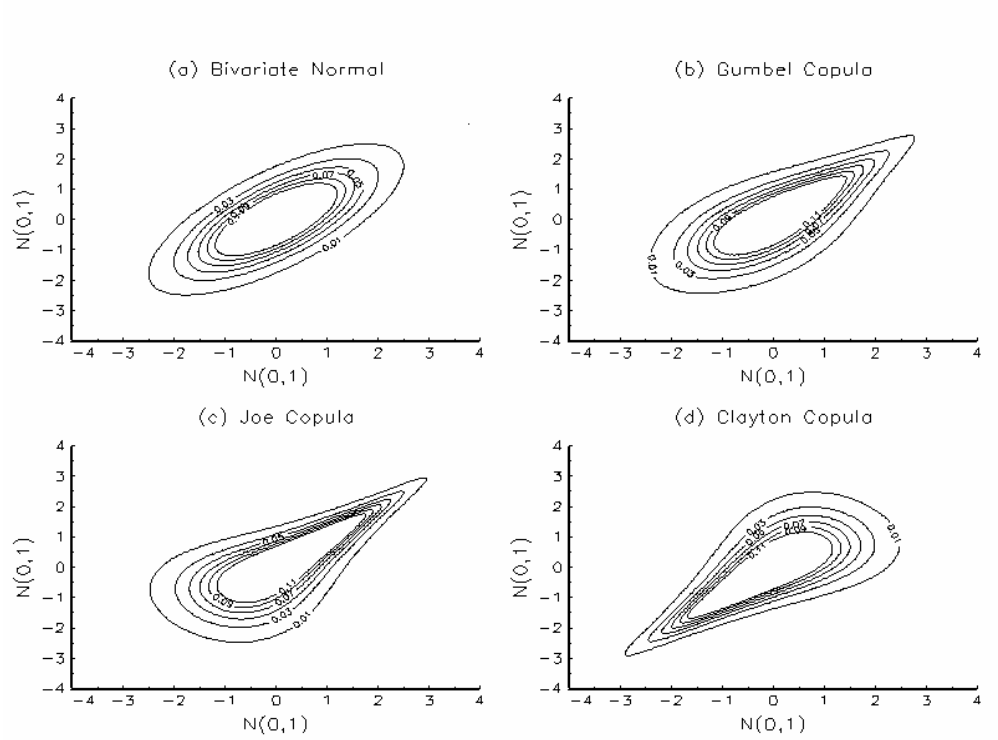


Figure 2. Plots of Gaussian, Gumbel, Joe and Clayton Copulas: Normal and Logistic marginals

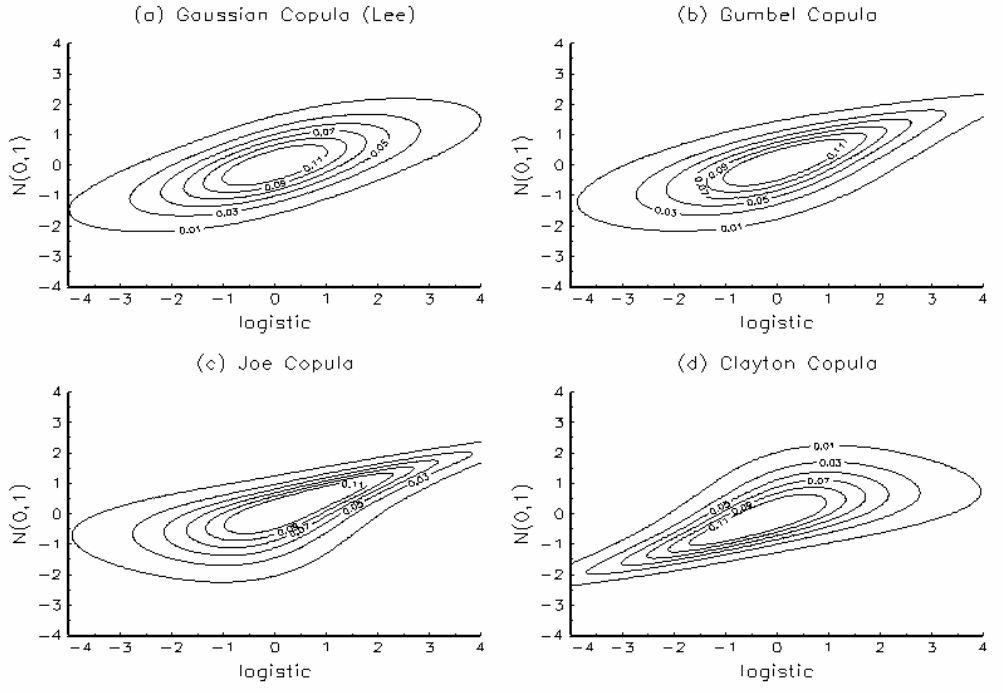


Figure 3. Horowitz test, Probit specification, bandwidth h=1

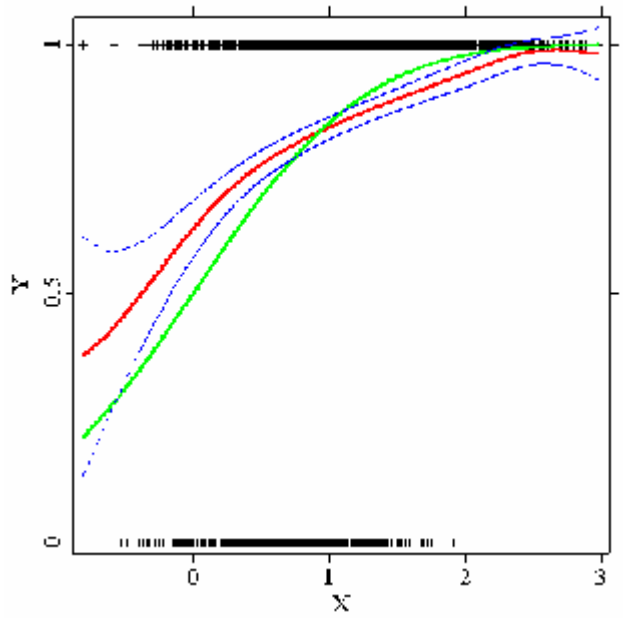


Figure 4. Horowitz test, Logit specification, bandwidth h=1

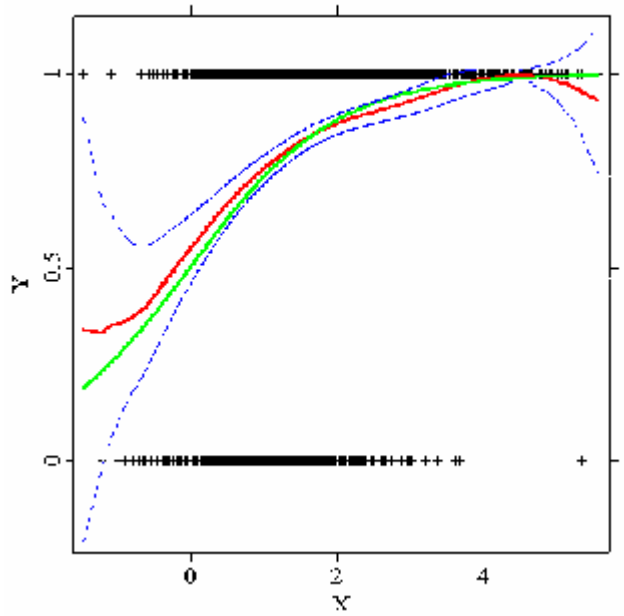
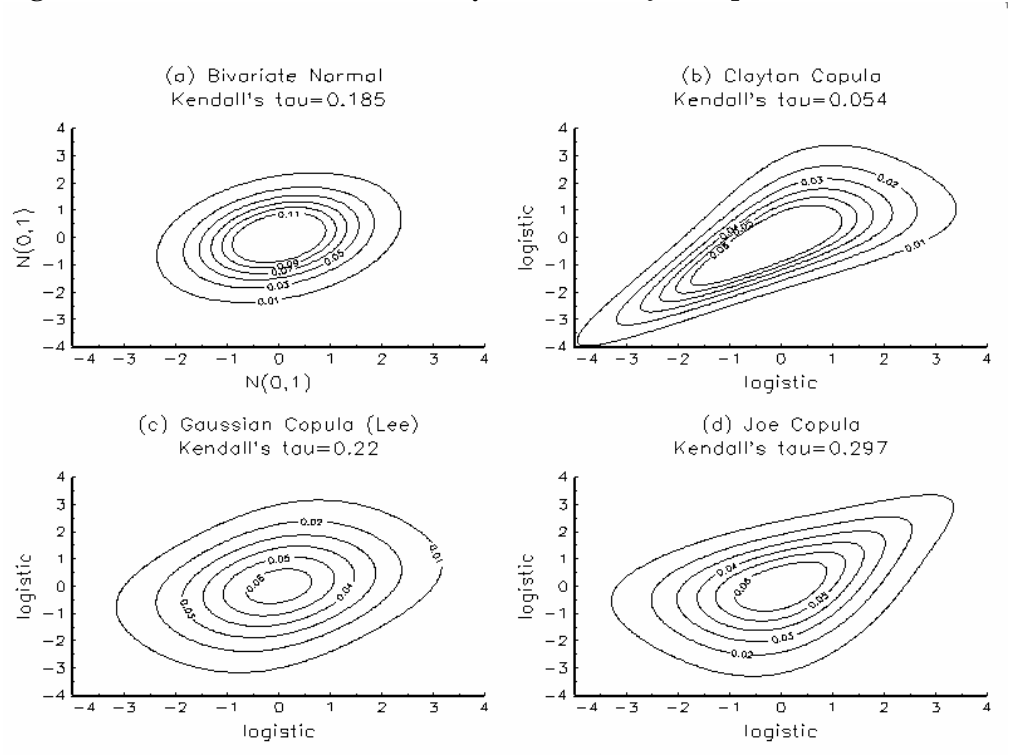


Figure 5. Plots of estimated BVN, Clayton, Lee and Joe Copula models



Appendix

List of variables

Wtp: total amount the respondent is willing to pay for the party, i.e. amount per party
Bid1: (log of) first bid presented to respondent
Nparty: (log of) size of the party
Children: number in party younger than 18
Adults: number of adults in party
Alone the respondent has visited the forest alone
Male the respondent is male
Time: (log of) time passed in the forest (minutes)
Parking: (log of) cost of parking (£)
Past: (log of) number of visits to the forest in the past year
Others: (log of) number of visits to other forests in the past year
Improved: the forest has improved recreation: 1-yes; 0-no
Income: Household income (£)

1 <15999
2 16000<30000
3 30000 and above