DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 32, No. 4, 2016, pp. 887–905, http://dx.doi.org/10.1515/JOS-2016-0046

The Use of Official Statistics in Self-Selection Bias Modeling

Luciana Dalla Valle[1]

Official statistics are a fundamental source of publicly available information that periodically provides a great amount of data on all major areas of citizens' lives, such as economics, social development, education, and the environment. However, these extraordinary sources of information are often neglected, especially by business and industrial statisticians. In particular, data collected from small businesses, like small and medium-sized enterprizes (SMEs), are rarely integrated with official statistics data.

In official statistics data integration, the quality of data is essential to guarantee reliable results. Considering the analysis of surveys on SMEs, one of the most common issues related to data quality is the high proportion of nonresponses that leads to self-selection bias.

This work illustrates a flexible methodology to deal with self-selection bias, based on the generalization of Heckman's two-step method with the introduction of copulas. This approach allows us to assume different distributions for the marginals and to express various dependence structures. The methodology is illustrated through a real data application, where the parameters are estimated according to the Bayesian approach and official statistics data are incorporated into the model via informative priors.

Key words: Bayes theorem; copulas; Heckman's two-step method; informative priors; small and medium-sized enterprizes.

1. Introduction

Official statistics are a fundamental source of information about many aspects of citizens' lives, about health, education, public and private services, as well as about the economic climate, the financial situation, and the environment.

Official statistics represent precious and rich data sources not only for public institutions, but also for firms that need to compare their performance against their competitors, measure the satisfaction of their customers, explore new markets and identify the most profitable locations to establish new subsidiaries.

However, the use of official statistics by firms, and in particular by medium-sized enterprizes (SMEs), is still rather limited.

Due to the recent growth of the number of available data sources and the increase in data quality, the use of innovative methods to aggregate results obtained from

[1] School of Computing, Electronics and Mathematics, Plymouth University, Drake Circus, PL4 8AA Plymouth, Devon, UK. Email: luciana.dallavalle@plymouth.ac.uk

Acknowledgments: I thank the anonymous referees for their insightful comments, which significantly improved the presentation of the article.

Disclaimer: The contents of this publication, including analysis and statistical elaborations, are the sole responsibility of the author and can in no way to be attributed to ISTAT. The views expressed in this article, as well as the information included in it, do not necessarily reflect the opinion or position of the European Commission and in no way commit the institution.

official statistics and from specific datasets is fundamental in order to obtain reliable analyses.

The issue of data quality may invalidate statistical results, in which case integrating different data sources and methods to improve data quality is needed.

According to the literature, the reliability of the results of a survey is reduced by the existence of nonsampling errors or errors related to data-collection methods.

The major types of nonsampling errors are measurement, coverage, and self-selection errors (see Nicolini and Dalla Valle 2012).

A *coverage error* is observed when the total number of subjects (target population) and their list (frame population), available to the creator of the sampling list used to select surveyed units, do not coincide.

A *measurement error* is given by the difference between the real value of an item related to a surveying unit and the corresponding observed value. This type of error frequently has been attributed to the presence of an interviewer.

Finally, a *self-selection error*, or unit nonresponse error, takes place when the selected unit does not answer or does not fill in the questionnaire form. This nonresponse may be caused by the inability to reach the subject or by his/her refusal to join the survey. The self-selected subjects who have provided answers to the questionnaire form a nonprobabilistic sample of the population.

In this article we focus on self-selection error, which is associated with subjects' independent decision to take part in the survey.

The main issue with self-selection is that the responders differ from nonresponders and therefore estimating an effect from only the responders might confound the effect and the choice to respond. Typically responders have common characteristics (i.e., they may all be young, middle-class women). In this case the sample is biased, since it does not represent the population it is related to, and the sample distribution of the variables differs from the same variables in the population.

The literature proposes some methods to correct the bias caused by self-selection. The Propensity Score Matching method was first introduced by Rubin (1974) and later developed by Rosenbaum and Rubin (1983), and suggests correcting the self-selection bias in probabilistic terms. According to this method, propensity scores are calculated using a multivariate logistic regression, and then each responder (from the so-called treatment group) is matched with a nonresponder (from the so-called control group) with the same score (for more details, please see Nicolini and Dalla Valle 2011). However, Propensity Score Matching requires large samples with substantial overlap between treatment and control groups.

The Heckman two-step Procedure, proposed by Heckman (Heckman 1979), considers two equations tied together by a latent factor that allows the missing data associated with the nonresponding subjects to be estimated. Heckman's method and its variants have been an essential tool for applied economics. Hamilton and Nickerson (2003) apply Heckman's method to strategic management and in particular to endogenous self-selection, according to which managers choose strategies and organizational forms with the expectation that they will yield high performance. The authors show that the use of corrections for endogeneity may yield more accurate estimates of the costs and benefits of alternative strategic choices. Lucchetti and Pigini (2013) use Heckman's self-selection model to propose a test for

bivariate normality in imperfectly observed models, based on the information matrix test for censored models with bootstrap critical values via Monte Carlo simulation. However, Heckman's approach requires restrictive assumptions that limit its flexibility and makes it difficult to adapt it to various dependence structures in the data.

We propose a novel approach allowing us on the one hand to correct self-selection bias and on the other hand to integrate specific data with official statistic data. This innovative approach combines the virtues of a flexible generalization of Heckman's two-step method using copulas and the Bayesian approach. The use of copulas to generalize Heckman's method relaxes the assumptions of normality and permits the accommodation of different types of dependencies, while the Bayesian approach allows the integration of official data by means of prior information. Moreover, our method can be applied successfully when dealing with small samples.

The remainder of this article is organized as follows: in Section 2 we introduce copulas and we present the main results of copula theory; Section 3 is devoted to the self-selection model as proposed by Heckman; Section 4 illustrates the characteristics of the proposed approach, using copulas within the self-selection model and integrating information with the Bayesian approach; Section 5 introduces an illustrative example and presents the results of the application of our model; finally, concluding remarks are given in Section 6.

2. Introduction to Copulas

2.1. Definition of Copula

The copula allows us to model the joint distribution of two or more random variables in a flexible way, incorporating their dependency effects. The word copula is derived from Latin, meaning to bind, tie, connect, and was first adopted by Sklar (Sklar 1959). In this context, the term refers to the ability of the copula to link the marginal distributions of random variables to a multivariate distribution, generating a stochastic dependence relationship. The main advantage of the copula is that it allows us to explicitly express the dependence structure of a set of random variables, whatever the distribution of these variables.

More formally, the copula is a multivariate distribution function defined over the unit cube $[0,1]^d$ (where d is the dimension of the copula), $C : [0,1]^d \rightarrow [0,1]$, linking two or more marginals distributed as uniforms. In the bivariate case, our focus in the remainder of the article, $d = 2$ and the copula is expressed as

$$C_\theta(u_1, u_2) = \Pr(U_1 < u_1, U_2 < u_2), \tag{1}$$

where C is the bivariate copula, U_1, U_2 are uniformly distributed random variables, with support belonging to the set $[0, 1]^2$, and θ is the copula dependence parameter vector.

The most important result in copula theory is Sklar's theorem (Sklar 1959), stating that if F is a joint bivariate distribution function with marginals F_1 and F_2, then there exists a bivariate copula C such that for (x_1, x_2)

$$F(x_1, x_2) = C_\theta(F_1(x_1), F_2(x_2)). \tag{2}$$

If F_1 and F_2 are continuous functions, then the copula is unique for any $(x_1, x_2) \in \mathbb{R} \cup \{-\infty, +\infty\}$. Thus, although the marginals are arguments of the copula, it

is independent of them, since it separates the distributions of the marginals from their dependence structure, parametrized by θ.

Nelsen's (1999) corollary suggests a method of generating copulas via the inversion method. If F is a continuous bivariate joint distribution function with univariate marginals F_1 and F_2 and generalized inverses F_1^{-1} and F_2^{-1}, then for (u_1, u_2) there exists a unique copula C such that

$$C_\theta(u_1, u_2) = F\left(F_1^{-1}(u_1), F_2^{-1}(u_2)\right). \tag{3}$$

2.2. *Types of Copulas*

The two main families of copulas are the Elliptical and Archimedean copulas (see Joe 1993, 1997).

Elliptical copulas are the copulas of elliptical distributions and their form is generally obtained using Nelsen's corollary (3). They are multivariate distributions sharing many of the tractable properties of the multivariate normal distribution.

The most popular elliptical copula is the Normal or Gaussian copula, whose characteristics are summarized in Table 1.

Another example of a copula that is particularly useful for its mathematical simplicity is the Farlie-Gumbel-Morgenstern (FGM) copula (Morgenstern 1956; Gumbel 1960; Farlie 1960).

The Archimedean family includes copulas expressed in a simple form based on the mathematical theory of associativity, and covers a variety of dependence structures. Archimedean copulas are constructed based on a generator function $\varphi : [0, 1] \rightarrow [0, \infty]$, with the properties of being a continuous, convex, and decreasing function (i.e., $\varphi(1) = 0$, $\varphi'(t) < 0$ and $\varphi''(t) > 0$ for $0 < t < 1$). The function $\varphi(t)$ generates the copula, in the bivariate case, as follows

$$\varphi(C_\theta(u_1, u_2)) = \varphi(u_1) + \varphi(u_2). \tag{4}$$

When the generator is strict (i.e., $\varphi(0) = \infty$), then the inverse $\varphi^{-1}(\cdot)$ exists and the copula is expressed as

$$C_\theta(u_1, u_2) = \varphi^{-1}[\varphi(u_1) + \varphi(u_2)],$$

otherwise a pseudoinverse function $\varphi^{[-1]}$ is used.

Some of the most popular Archimedean copulas are the AMH, Clayton, Gumbel and Frank copula (Ali et al. 1978; Clayton 1978; Gumbel 1960; Frank 1979). The main characteristics of these types of copulas are listed in Table 1 and they will be used in Section 5 to fit our model to real data. The range of Kendall's τ is reported for comparison purposes. This concordance measure is generally preferred to the copula's dependence parameter θ, since τ is invariant with respect to the marginals and to strictly increasing transformations of the variables. For more details about transforming the copula parameter θ into Kendall's τ, please see Smith (2003).

Figure 1 shows the bivariate contour plots of the different types of copulas illustrated in this section, all with standard normal margins and $\tau = 0.5$.

Table 1. Characteristics of different copulas.

Copula	Distribution	Generator function $\varphi(t)$	θ range	Kendall's τ range	Type of dependence
Gaussian	$\Phi_2(\Phi^{-1}(u_1), \Phi^{-1}(u_2); \theta)$	-	$-1 \leq \theta \leq 1$ $\theta = 0$ independence	$-1 \leq \tau \leq 1$	both negative and positive equal in upper and lower tails
FGM	$u_1 u_2[1 + \theta(1 - u_1)(1 - u_2)]$	-	$-1 \leq \theta \leq 1$ $\theta = 0$ independence	$-2/9 \leq \tau \leq 2/9$	both negative and positive equal and weak dependence in tails
AMH	$u_1 u_2/[1 - \theta(1 - u_1)(1 - u_2)]$	$\ln\left(\frac{1-\theta(1-t)}{t}\right)$	$-1 \leq \theta < 1$ $\theta = 0$ independence	$-0.1817 \leq \tau < 1/3$	both negative and positive unequal and weak dependence in tails
Clayton	$\left(u_1^{-\theta} + u_2^{-\theta} - 1\right)^{-1/\theta}$	$(1/\theta)(t^{-\theta} - 1)$	$0 < \theta < \infty$ $\theta \rightarrow 0$ independence	$0 < \tau < 1$	only positive strong left tail and weak right tail
Gumbel	$\exp\left(-\left[(-\ln u_1)^\theta + (-\ln u_2)^\theta\right]^{1/\theta}\right)$	$(-\ln t)^\theta$	$1 \leq \theta < \infty$ $\theta = 1$ independence	$0 \leq \tau < 1$	only positive strong right tail and weak left tail
Frank	$-\frac{1}{\theta}\ln\left(1 + \frac{(e^{-\theta u_1} - 1)(e^{-\theta u_2} - 1)}{e^{-\theta} - 1}\right)$	$-\ln\left[\left(e^{-\theta t} - 1\right)\right] / \left(e^{-\theta} - 1\right)$	$\{\theta \in (-\infty; 0) \cup (0; \infty)\}$ $\theta \rightarrow 0$ independence	$-1 \leq \tau \leq 1$	both positive and negative equal in upper and lower tails

Note that $\Phi_2(\cdot, \cdot; \theta)$ denotes the bivariate cumulative distribution function (cdf) of the standard normal with Pearson's correlation parameter θ and $\Phi^{-1}(\cdot)$ denotes the inverted cdf of the univariate standard normal.

3. The Self-Selection Model

The self-selection model we are proposing is also known as the *Tobit-2* model (as introduced by Tobin 1958). This is a censored regression model where the dependent variable is only observed in a selected sample that is not representative of the population. Censoring occurs when the value of the dependent variable is only partially known. It is a defect in the sample, because if there were no censoring, then the data would be a representative sample from the population of interest.

In 1979, Heckman proposed a model for self-selection, which is made by two linear equations related to each other: the substantial equation and the selection equation.

Supposing that data are missing for $N - n$ observations (the number of nonresponders), we define the selection equation (that represents participation) for individual i, $i = 1, \ldots, N$, as follows:

$$y_{1i}^{*} = \mathbf{x}_{1i}\beta_1 + \varepsilon_{1i}, \tag{5}$$

where y_{1i}^{*} is an unobserved latent random variable such that $y_{1i}^{*} > 0$ corresponds to responders, while $y_{1i}^{*} \leq 0$ corresponds to nonresponders; $\mathbf{x}_{1i}$ is the ith vector of variables known for all N subjects, β_1 is a vector of parameters, and ε_{1i} is the error.

The substantial equation (that is observed for participants) for individual i is:

$$y_{2i}^{*} = \mathbf{x}_{2i}\beta_2 + \varepsilon_{2i}, \tag{6}$$

where y_{2i}^{*} denotes the latent continuous variable of interest, $\mathbf{x}_{2i}$ is the ith vector of variables known for all N subjects, β_2 is a vector of parameters, and ε_{2i} is the error.

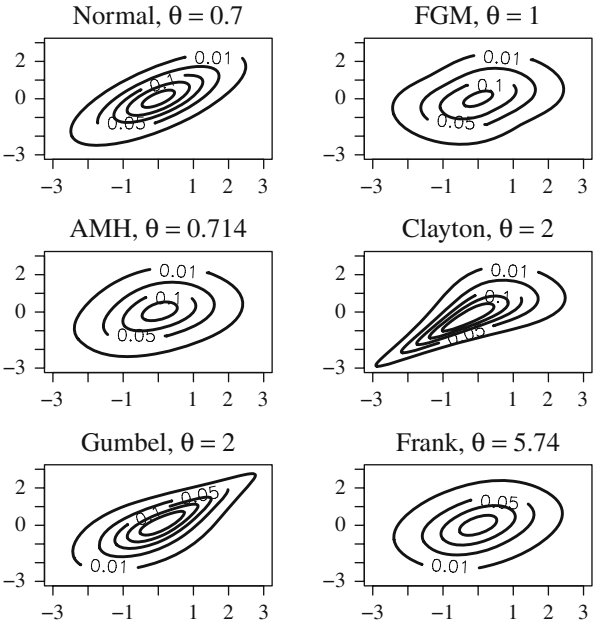

Fig. 1. Bivariate contour plots of different copulas, with standard normal margins and $\tau = 0.5$. From the top left figure: Normal with $\theta = 0.7$, FGM with $\theta = 1$, AHM with $\theta = 0.714$, Clayton with $\theta = 2$, Gumbel with $\theta = 2$, Frank with $\theta = 5.74$.

Note that the explanatory variables $\mathbf{x}_1$ and $\mathbf{x}_2$ for the selection and substantial equation may or may not be equal. However, the model is well identified if the exclusion restriction is fulfilled, that is, if $\mathbf{x}_1$ includes a component that has substantial explanatory power but that is not present in $\mathbf{x}_2$ (see Heckman 1979). If the exclusion restriction is not fulfilled, the consequence is perfect multicollinearity and the equations cannot be estimated.

We can now define the observed variables

$$y_{1i} = \begin{cases} 0 & \text{if } y_{1i}^* \leq 0 \\ 1 & \text{if } y_{1i}^* > 0 \end{cases}$$

and

$$y_{2i} = \begin{cases} 0 & \text{if } y_{1i} = 0 \\ y_{2i}^* & \text{if } y_{1i} = 1, \end{cases}$$

where $y_{1i} = 1$ corresponds to a responder and $y_{1i} = 0$ to a nonresponder, and we observe the outcome y_{2i} only if the latent selection variable y_{1i}^* is positive.

Note that the self-selection model can be alternatively written such that the selection equation becomes

$$y_{1i} = \mathbb{I}\left[y_{1i}^* > 0\right]$$

and the substantial equation becomes

$$y_{2i} = \mathbb{I}\left[y_{1i}^* > 0\right] y_{2i}^*,$$

where $\mathbb{I}[\cdot]$ is the indicator function.

Hence, the likelihood function of the self-selection model is

$$L = \prod_{i=1}^{N} \left\{\Pr\left[y_{1i}^* \leq 0\right]\right\}^{1-y_{1i}} \left\{f_{2|1}\left(y_{2i}|y_{1i}^* > 0\right) \cdot \Pr\left[y_{1i}^* > 0\right]\right\}^{y_{1i}} \tag{7}$$

where the first term is the contribution of nonresponders and the second term is the contribution of responders. In other words, the density of y_{2i} is the same as that of y_{2i}^* for $y_{1i} = 1$ and is equal to the probability of observing $y_{1i}^* \leq 0$ if $y_{1i} = 0$.

The conditional density in Equation (7) can be written as follows

$$f_{2|1}\left(y_{2i}|y_{1i}^* > 0\right) = \frac{1}{1 - F_1(0)}\left[f_2(y_{2i}) - \frac{\partial}{\partial y_2} F(0, y_{2i})\right]$$

where $F_1(0) = \Pr\left\{y_{1i}^* \leq 0\right\} = \Pr\{y_{1i} = 0\}$ and $F(\cdot, \cdot)$ is the bivariate joint cdf (cumulative distribution function). Substituting the conditional density form into (7) yields

$$L = \prod_{i=1}^{N} \{F_1(0)\}^{1-y_{1i}} \left\{f_2(y_{2i}) - \frac{\partial}{\partial y_2} F(0, y_{2i})\right\}^{y_{1i}}. \tag{8}$$

4. Copulas Applied to Self-Selection

The likelihood function of the self-selection model (8) can be re-expressed in a more flexible way using copulas. In particular, in (8) the derivative of the joint cdf, following Sklar's theorem and its corollary, can be written as

$$\frac{\partial}{\partial y_2} F(0, y_{2i}) = \frac{\partial}{\partial v} C_\theta(F_1(0), v)\Big|_{v \to F_2} \cdot \frac{\partial F_2}{\partial y_2}.$$

Thus the likelihood function (8) can be written in terms of copulas as follows

$$L = \prod_{i=1}^{N} \{F_1(0)\}^{1-y_{1i}} \left\{ \left[1 - \frac{\partial}{\partial F_2} C_\theta(F_1, F_2) \right] \cdot f_2(y_{2i}) \right\}^{y_{1i}}. \tag{9}$$

4.1. Heckman's Model

Heckman's model is also called the Normal model. He supposes that the marginal latent variables Y_1^* and Y_2^* are distributed according to Gaussian models, such that:

$$Y_1^* \sim N(\mathbf{x}_1\beta_1, 1) \qquad Y_2^* \sim N\left(\mathbf{x}_2\beta_2, \sigma_2^2\right),$$

where $\sigma_1^2 = 1$. As a consequence the error terms follow a bivariate normal distribution:

$$\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \theta \\ \theta & \sigma_2^2 \end{pmatrix} \right).$$

The likelihood function in this case takes the form

$$L = \prod_{i=1}^{N} \{1 - \Phi(\mathbf{x}_{1i}\beta_1)\}^{1-y_{1i}} \left\{ \frac{1}{\sigma_2} \phi\left(\frac{y_{2i} - \mathbf{x}_{2i}\beta_2}{\sigma_2} \right) \Phi\left(\frac{\mathbf{x}_{1i}\beta_1 + \frac{\theta}{\sigma_2}(y_{2i} - \mathbf{x}_{2i}\beta_2)}{\sqrt{1-\theta^2}} \right) \right\}^{y_{1i}}, \tag{10}$$

where the left term corresponds to non self-selection, while the right term corresponds to self-selection.

Heckman's assumption of a joint normal distribution for the error terms is overly restrictive, limiting the applicability of his approach. As pointed out by Lee (1983), the copula approach can be used to relax the traditional assumption that the marginal distributions are normal. Indeed, the marginals are very often not normally distributed, especially financial variables. Smith (2003) provides a general copula-based framework for Heckman's model by demonstrating that copulas can be used to extend the standard analysis to any bivariate distribution with given marginals (see also Smith 2005). The use of normal marginals and normal copula leads us to the traditional Heckman's method, as is shown by comparing Equations (10) and (8) (see Bhat and Eluru 2009). However, with significant departures from normality for the marginals and/or the copula, the traditional Heckman's approach is no longer sufficiently general and the use of the copula approach is essential to provide the flexibility necessary for modelling the data and the dependencies in the correct way. The following sections will demonstrate how the Bayesian approach

allows us to incorporate different sources of information into the generalized Heckman's model, and how this technique can be applied to nonresponse modelling.

According to the copula approach, the likelihood has to be calculated using the (8). The expressions of the derivatives $\frac{\partial}{\partial F_2} C_\theta(F_1, F_2)$ for each type of copula are listed in Table 2.

4.2. The Bayesian Approach

In order to integrate specific data with official data sources, we use the Bayesian approach, specifying informative priors using official information. The Bayesian approach is based upon the idea that the interviewer begins with some prior beliefs about the system and then updates these beliefs on the basis of observed data. This updating procedure is based upon Bayes' Theorem:

$$\pi(\eta|\text{data}) \propto f(\text{data}|\eta)p(\eta) \qquad (\text{posterior} \propto \text{likelihood} \times \text{prior}),$$

where η is the parameter vector. Generally, parameter estimates are determined employing the Markov Chain Monte Carlo (MCMC) method, which uses algorithms to sample observations from the posterior distribution based on the construction of a Markov chain that has the posterior as its equilibrium distribution. The state of the chain after a number of steps is then used as a sample of the posterior distribution. If the prior distributions are conjugate, general MCMC algorithms are not needed, but simpler techniques, like the Gibbs Sampler, may be used (see Albert and Chib 1993; Gamerman and Lopes 2006; Armero et al. 2008).

In order to apply the Bayesian approach to the generalized Heckman's model, we specify prior distributions for the vectors of parameters of the selection equation β_1 and of the substantial equation β_2, for the variance parameter σ_2^2, and for the copula dependence parameter θ. Then, we sample from the posterior distribution by implementing a Metropolis-within-Gibbs algorithm.

We assume a multivariate normally distributed vague prior for the selection equation parameter vector $\beta_1 \sim N(\mu_1, \Sigma_1)$ where μ_1 is a $(n_1 + 1)$-dimensional vector of zeros and $\Sigma_1 = 100 I_{n_1+1}$, with I_{n_1+1} the $(n_1 + 1)$-dimensional identity matrix. Like for the parameter vector β_1, we consider a multivariate normal prior for the substantial equation parameter vector β_2, but we used information from official statistics to define informative prior distributions. Hence, $\beta_2 \sim N(\mu_2, \Sigma_2)$, where μ_2 is a $(n_2 + 1)$-dimensional vector and

Table 2. Expressions for the copula derivatives $\frac{\partial}{\partial F_2} C_\theta(F_1, F_2)$.

Copula	Expression for $\frac{\partial}{\partial F_2} C_\theta(F_1, F_2)$
Gaussian	$\Phi\left(\frac{\Phi^{-1}(u_1) - \theta\Phi^{-1}(u_2)}{\sqrt{1-\theta^2}}\right)$
FGM	$u_1[1 + \theta(1 - u_1)(1 - 2u_2)]$
AMH	$\frac{(1-\theta)u_1 + \theta u_1^2}{(1-\theta(1-u_1)(1-u_2))^2}$
Clayton	$u_2^{-(\theta+1)}\left(u_1^{-\theta} + u_2^{-\theta} - 1\right)^{-\left(\frac{1+\theta}{\theta}\right)}$
Gumbel	$u_2^{-1}(-\ln(u_2))^{\theta-1} \cdot C_\theta(u_1, u_2)\left[(-\ln(u_1))^\theta + (-\ln(u_2))^\theta\right]^{\left(\frac{1}{\theta}-1\right)}$
Frank	$\left[1 - e^{\theta C_\theta(u_1,u_2)}\right](1 - e^{\theta u_2})^{-1}$

Σ_2 is the $(n_2 + 1)$-dimensional prior covariance matrix. For σ_2^2 we consider the vague prior $\sigma_2^2 \sim \Gamma^{-1}(a, b)$ where $a = 0.001$ and $b = 0.001$. As prior distribution for τ we consider the vague prior $\tau \sim Beta(\alpha, \beta)$ extended to the range $[-1, 1]$ (Huard et al. 2006), where $\alpha = 1$ and $\beta = 1$.

5. Innovation Survey Data

The methodology illustrated in the previous section was tested using two datasets: a national-level survey and an official EU-level survey dataset.

The first dataset is available on the ISTAT (Italian National Institution of Statistics) website and it contains data collected through a survey on innovations introduced and innovative activities undertaken by a sample of Italian firms between 2008 and 2010.

The Italian Innovation Survey, carried out on a two-year basis, collects information about new or significantly improved goods or services (product innovations) and new or significantly improved processes, logistics or distribution methods (process innovations), as well as about organizational and marketing innovation. The original data were perturbed by ISTAT, in order to guarantee the privacy of respondents (see ISTAT 2013).

From the original ISTAT dataset, we only selected SMEs, that, according to the definition provided by the European Union, include enterprises which employ fewer than 250 persons and which have an annual turnover not exceeding 50 million euros, and/or an annual balance sheet total not exceeding 43 million euros.

Moreover, we restricted our attention to the reference period of 2010, hence limiting the number of firms in the dataset to 4,266.

Therefore, from a total number of 3.8 millions of Italian SMEs in 2010, we only considered survey information of a small sample of about 4,200 firms.

The variables we used from the innovation survey dataset are described in Table 3.

We integrated the ISTAT innovation survey data with a second dataset, the 2010 Innovation Union Scoreboard (IUS) provided by the European Union (see European Commission 2010). IUS provides a comparative assessment of the research and innovation performance of the EU Member States and the relative strengths and weaknesses of their research and innovation systems.

In particular, we used data about human resources, firms' activities, and outputs, considering the following variables:

- human resources who completed tertiary education,
- business R&D firm expenditure,
- non-R&D innovation firm expenditure,
- firms introducing product or process innovations,
- firms introducing marketing/organizational innovations,
- knowledge-intensive services exports,
- sales of new-to-market and new-to-firm innovations.

We used the IUS variables listed above to define informative prior distributions for the substantial equation parameters β_2 of the generalized Heckman's model, described in Section 4. The parameters of these informative priors were defined based on the empirical distributions of the corresponding IUS variables. This approach allows us to integrate the

Table 3. Description of the innovation survey dataset variables.

Innovation survey dataset	
Variable names	Variable label
turn	turnover
rrdinx	expenses for activities of R&D
rrdexx	expenses for acquisitions of R&D services
rmacx	expenses for acquisition of machinery and equipment
roekx	expenses for acquisition of other external technologies
rdsgx	expenses for design activities
rprex	expenses for other innovative activities
rtrx	expenses for education on innovative activities
rmarx	expenses for marketing of innovative products
empdeg	number of employees with a university degree
turnmar	turnover coming from new products or services (or significantly improved products and services) for the reference market
turnin	turnover coming from new products or services (or significantly improved products and services) for the firm only

ISTAT national data source with the more general IUS international data source, provided by the European Commission.

5.1. The Model

We suppose the firms that did not respond to the questionnaire are those belonging to the business and other services and nonmarketed services NACE macrosectors. The percentage of respondent firms is 85.07%, while the percentage of nonrespondent firms is 14.93%.

We assume a Normal distribution for the marginal Y_1^* (selection equation)

$$Y_1^* \sim N(\mathbf{x}_1\beta_1, 1)$$

and a log-normal distribution for Y_2^* (substantial equation), after a graphical examination of the variable and the application of the Kolmogorov-Smirnov test, which accepts log-normality:

$$\log Y_2^* \sim N\left(\mu_l, \sigma_l^2\right),$$

where $\mu_l = e^{\mathbf{x}_2\beta_2+\sigma^2/2}$ and $\sigma_l^2 = \left(e^{\sigma^2} - 1\right)e^{2\mathbf{x}_2\beta_2+\sigma^2}$. Figure 2 shows the histogram of the variable Turnover.

In the model, the target variable y_2 is *turn*; the vector $\mathbf{x}_1$ comprises the above eleven variables listed in Table 3. The model is well identified if the exclusion restriction is fulfilled, that is, if $\mathbf{x}_1$ includes a component (*empdeg*) that has substantial explanatory power but that is not present in $\mathbf{x}_2$.

5.2. Results

We run the Metropolis-within-Gibbs algorithm for 10,000 iterations and discarded the first 2,000 iterations as the burn-in period. Because of space considerations, we here analyze the MCMC traceplots of the model using the Clayton copula, since the results obtained

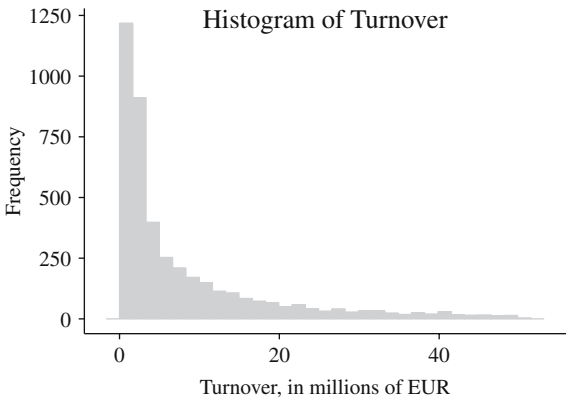

Fig. 2. Histogram of Turnover.

with the other choices of copulas are very similar to those presented. The trace plots of the parameters β_1, β_2, σ^2 and θ are listed in the Appendix. The sample paths show that the chains are well mixing, freely exploring the sample space.

Parameter estimates for the selection and substantial equations are very stable, as shown in Figures 3 and 4, representing credible intervals for β_1 and β_2, respectively. A credible interval is computed from the posterior distribution and is the interval within which the probability of the parameter of interest falling in is given by the level of credibility. The credible intervals are all very similar for the different choices of copula. The only exceptions are the credible intervals of the β_2 parameters modeled with the independence copula. However, this was expected, since the independence copula assumes no association between the selection and substantial equations. The results of the β_1 parameters indicate which variables are associated with response. From Figure 3, the variables with a significant negative influence on the response are *rrdinx*, *roekx* and *empdeg*, while the variable with a significant positive influence on the response is *rmarx*. This means that firms that invest in R&D and external technologies, do not invest in marketing, and employ several graduates, are nonrespondents. The β_2 parameters indicate which variables explain the firms' turnover. Figure 4 suggests that the variables with a significant positive

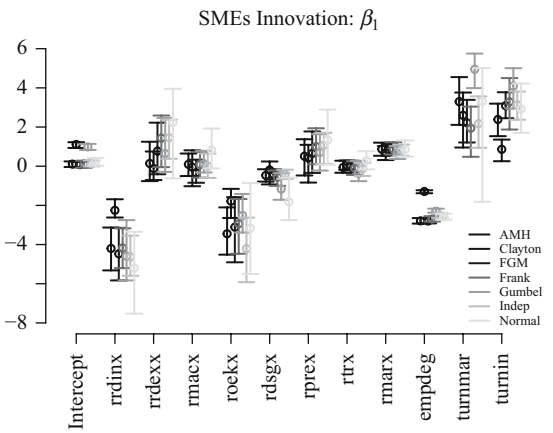

Fig. 3. Credible intervals of β_1 for all copulas considered at 95% level.

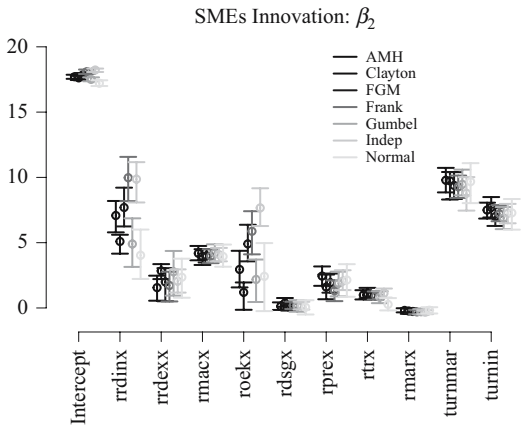

Fig. 4. Credible intervals of β_2 for all copulas considered at 95% level.

influence on the firms' turnover are *rrdinx*, *turnmar*, *turnin*, *rmacx*, *rrdexx*, and *rprex*. This means that firms investing in R&D, machinery equipment, new products and services, and other innovative activities show a high turnover.

Figures 5 and 6 show the boxplots of the posterior distributions of the parameters θ and τ. As can be seen from the plots, the dependence parameters τ are positive, meaning that the nonrespondent SMEs (firms that did not fill in the questionnaire) are those with high turnover. The values of Kendall's τ denote a moderate degree of dependence for almost all the different types of copulas.

5.2.1. Model Comparison

We compare the performances of the different copula models using the Deviance Information Criterion (DIC), which has the following expression

$$DIC = \bar{D} + p_D$$

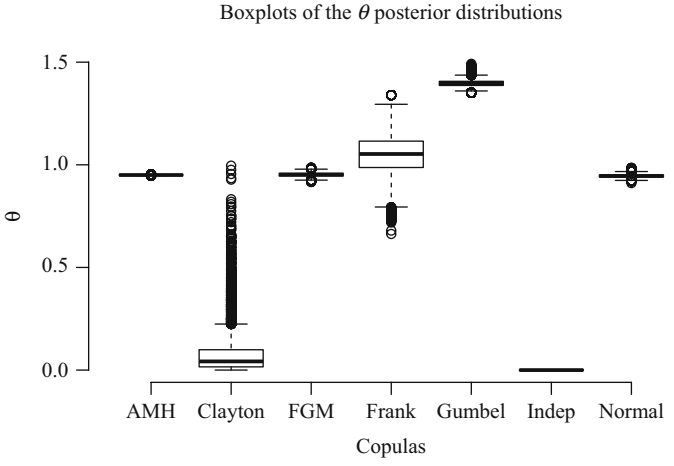

Fig. 5. Boxplots of the posterior distributions of θ for the different copulas.

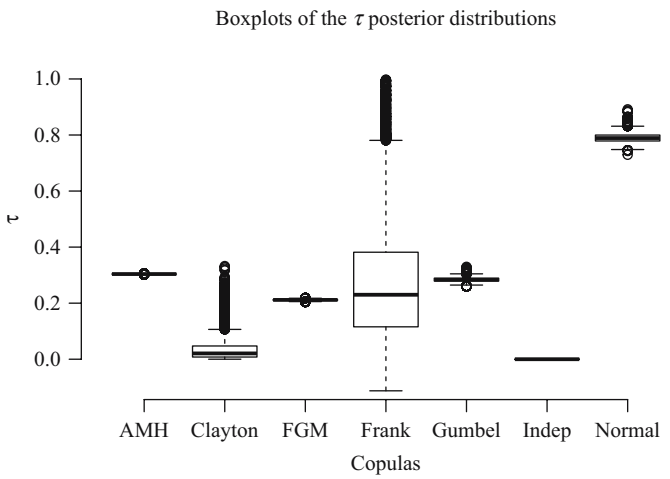

Fig. 6. Boxplots of the posterior distributions of τ for the different copulas.

where $\bar{D} = E(-2\log[L(\text{data}|\eta)])$ is the average of the log-likelihoods calculated at the end of each MCMC iteration, $p_D = \bar{D} - \hat{D}$ and $\hat{D} = -2\log[L(\text{data}|\eta^*)]$ is the log-likelihood calculated using the parameter posterior means. Models with smaller DIC are better supported by the data.

Table 4 lists the DIC results for the different copulas. The Clayton copula model outperforms the others, since it has the lowest DIC value. Therefore, the Clayton copula is the one that best models the relationship between Heckman's equations. The main advantage that the Clayton copula offers over the Normal is that the unequal tail dependence, which is stronger in the left tail, is properly accounted for, leading to more accurate results.

Finally, in order to correctly estimate our target variable, that is the turnover of the SMEs, we need to consider the dependence value estimated through the most suitable copula for our data. The mean turnover can be calculated as

$$E\left[Y|Y_1^* > 0\right] = \int_0^\infty y f_{2|1}\left(y|Y_1^* > 0\right)dy = \frac{1}{1 - F_1(0)}\left(E(Y) - \int_0^\infty y \frac{\partial}{\partial F_2} C_\theta(F_1, F_2) f_2 dy\right)$$

where the result was evaluated at $\mathbf{x} = \bar{\mathbf{x}}$, the covariate averages across the total number of firms. Figure 7 shows the histogram of the mean turnover value for the SMEs,

Table 4. Model comparison.

	DIC
AMH	−43599.06
Clayton	**−50993.80**
FGM	−43678.94
Frank	−43681.95
Gumbel	−45594.26
Indep	−43767.92
Normal	−43593.09

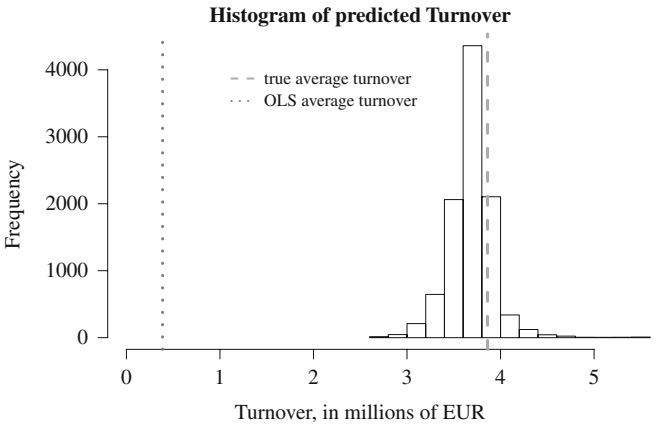

Fig. 7. Histogram of Turnover predicted via the Bayesian generalized Heckman approach. The plot compares the copula estimate for the average turnover with the biased OLS estimates.

calculated from the MCMC simulations. The dashed line represents the true average value of turnover for the observed dataset, while the dotted line represents the average value of turnover predicted by the traditional OLS model, which is based on the Normal copula and the log-transformation of y_2. Please note that the true value of turnover is available, since self-selection was artificially introduced in the Innovation survey dataset, as explained in Subsection 5.1. This result shows that the use of the OLS model in presence of self-selection is completely unrealistic and underestimates the true value of the target variable. The generalized Heckman's model using the Clayton copula performs well and accurately predicts the true value of turnover, since the predicted turnover is very close to its true value. The Clayton copula in this case is more flexible than the traditional Normal copula in capturing asymmetric tail dependence, and it gives more reliable predictions.

6. Concluding Remarks

This article illustrated the application of the Bayesian generalized Heckman approach to correct the self-selection bias integrating different sources of information.

This approach has a number of potential applications, especially where survey data are employed. The use of official statistics in sector and marketing analysis by firms is one of them. However, this approach can be successfully implemented in education, medical, and social studies.

A limitation of the study could be the computational complexity in some cases. However, the main advantage is the accuracy of the results compared to traditional approaches.

Further studies may include the analysis of additional families of copulas and their rotated versions.

Appendix

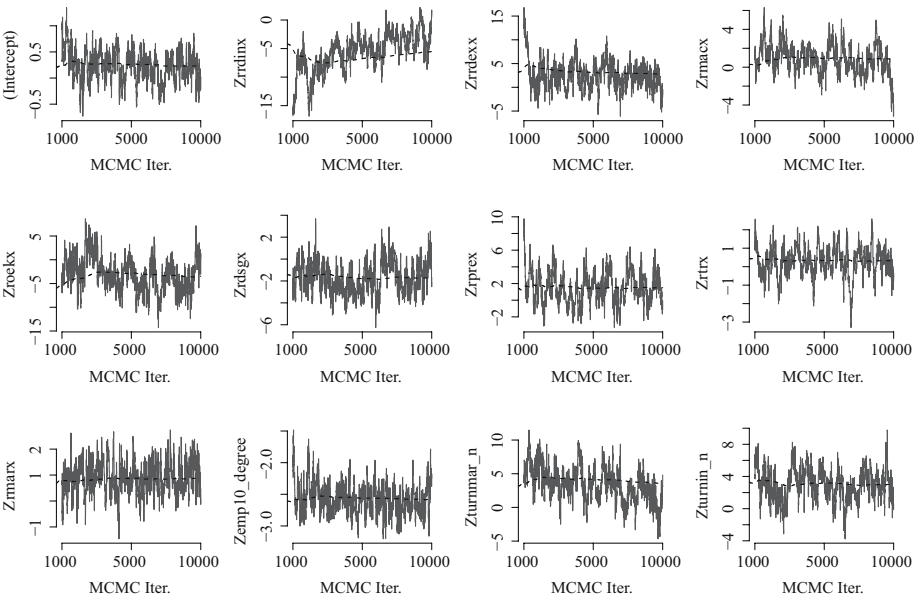

Fig. 8. Trace plots of the β_1 parameters for the Clayton copula model. The labels on the vertical axes refer to the names of the variables.

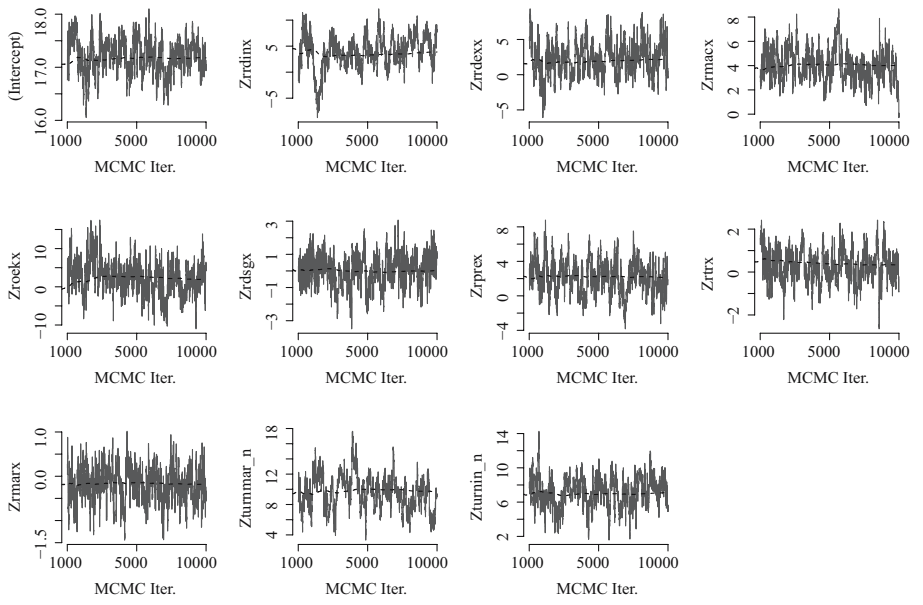

Fig. 9. Trace plots of the β_2 parameters for the Clayton copula model. The labels on the vertical axes refer to the names of the variables.

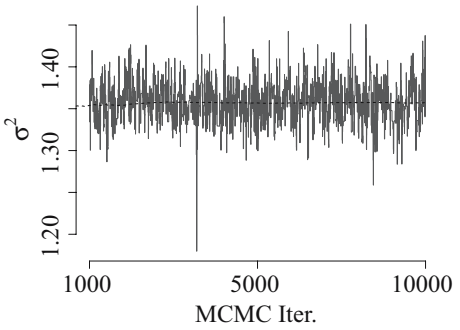

Fig. 10. Trace plots of the σ^2 parameter for the Clayton copula model.

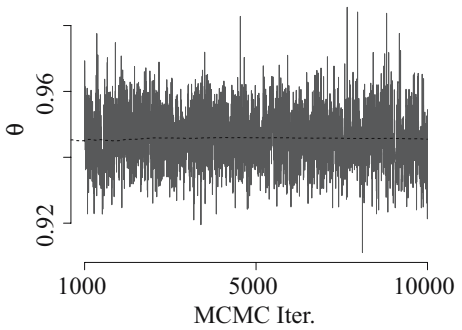

Fig. 11. Trace plots of the θ parameter for the Clayton copula model.

7. References

Albert, J.H. and S. Chib. 1993. "Bayesian Analysis of Binary and Polychotomous Response Data." *Journal of the American Statistical Association* 88: 669–679. Doi: http://dx.doi.org/10.2307/2290350.

Ali, M.M., N.N. Mikhail, and M.S. Haq. 1978. "A Class of Bivariate Distributions Including the Bivariate Logistic." *Journal of Multivariate Analysis* 8: 405–412. Doi: http://dx.doi.org/10.1016/0047-259X(78)90063-5.

Armero, C., A. López-Quílez, and R. López-Sánchez. 2008. "Bayesian Assessment of Times to Diagnosis in Breast Cancer Screening." *Journal of Applied Statistics* 35: 997–1009. Doi: http://dx.doi.org/10.1080/02664760802191397.

Bhat, C.R. and N. Eluru. 2009. "A Copula-Based Approach to Accommodate Residential Self-Selection Effects in Travel Behavior Modeling." *Transportation Research Part B* 43: 749–765. Doi: http://dx.doi.org/10.1016/j.trb.2009.02.001.

Clayton, D.G. 1978. "A Model for Association in Bivariate Life Tables and its Application in Epidemiological Studies of Family Tendency in Chronic Disease Incidence." *Biometrika* 65: 141–151. Doi: http://dx.doi.org/10.1093/biomet/65.1.141.

European Commission. 2010. "Innovation Union Scoreboard 2010." Available at: http://ec.europa.eu/enterprise/policies/innovation/files/ius/ius-2010_en.pdf (accessed February 2014).

Farlie, D.J.G. 1960. "The Performance of Some Correlation Coefficients for a General Bivariate Distribution." *Biometrika* 47: 307–323. Doi: http://dx.doi.org/10.2307/2333302.

Frank, M.J. 1979. "On the Simultaneous Associativity of F(x, y) and x + y − F(x, y)." *Aequationes Mathematicae* 19: 194–226.

Gamerman, D. and H. Lopes. 2006. "Markov Chain Monte Carlo." *Texts in Statistical Science*, 2nd ed. New York: Chapman & Hall.

Gumbel, E.J. 1960. "Bivariate Exponential Distributions." *Journal of the American Statistical Association* 55: 698–707.

Hamilton, B.H. and J.A. Nickerson. 2003. "Correcting for Endogeneity in Strategic Management Research." *Strategic Organization* 1: 51–78. Doi: http://dx.doi.org/10.1177/1476127003001001218.

Heckman, J.J. 1979. "Sample Selection Bias as a Specification Error." *Econometrica* 47: 153–161.

Huard, D., G. Evin, and A.-C. Favre. 2006. "Bayesian Copula Selection." *Computational Statistics and Data Analysis* 51: 809–822. Doi: http://dx.doi.org/10.1016/j.csda.2005.08.010.

ISTAT. 2013. *Italian Innovation Survey*. Available at: http://www.istat.it (accessed February 2014).

Joe, H. 1993. "Parametric Families of Multivariate Distributions with Given Marginals." *Journal of Multivariate Analysis* 46: 262–282. Doi: http://dx.doi.org/10.1006/jmva.1993.1061.

Joe, H. 1997. *Multivariate Models and Dependence Concepts*. London: Chapman & Hall.

Lee, L.-F. 1983. "Generalized Econometric Models with Selectivity." *Econometrica* 51: 507–512. Doi: http://dx.doi.org/10.2307/1912003.

Lucchetti, R. and C. Pigini. 2013. "A Test for Bivariate Normality with Applications in Microeconometric Models." *Statistical Methods and Applications* 22: 535–572. Doi: http://dx.doi.org/10.1007/s10260-013-0236-5.

Morgenstern, D. 1956. "Einfache Beispiele zweidimensionaler Verteilungen." *Mitteilungsblatt für Mathematische Statistik* 8: 234–235.

Nelsen, R.B. 1999. *An Introduction to Copulas*. New York: Springer.

Nicolini, G. and L. Dalla Valle. 2011. "Errors in Customer Satisfaction Surveys and Methods to Correct Self-Selection Bias." *Quality Technology & Quantitative Management* 8: 167–181. Doi: http://dx.doi.org/10.1080/16843703.2011.11673254.

Nicolini, G. and L. Dalla Valle. 2012. "Census and Sample Surveys." In *Modern Analysis of Customer Surveys, with Applications Using R*, edited by Ron S. Kenett and Silvia Salini, 37–53. Statistics in Practice Series. Hoboken, NJ: John Wiley & Sons.

Rosenbaum, P. and D. Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70: 41–55. Doi: http://dx.doi.org/10.1093/biomet/70.1.41.

Rubin, D.B. 1974. "Estimating Causal Effects of Treatments in Randomized and Non-Randomized Studies." *Journal of Educational Psychology* 66: 688–701. Doi: http://dx.doi.org/10.1037/h0037350.

Sklar, A. 1959. "Fonctions de répartition a *n* dimensions et leurs marges." *Publications de l'Institut de Statistique de l'Universite de Paris* 8: 229–231.

Smith, M. 2003. "Modelling Sample Selection Using Archimedean Copulas." *Econometrics Journal* 6: 99–123. Doi: http://dx.doi.org/10.1111/1368-423X.00101.

Smith, M.D. 2005. "Using Copulas to Model Switching Regimes with an Application to Child Labour." *The Economic Record* 81: S47–S57.

Tobin, J. 1958. "Estimation of Relationships for Limited Dependent Variables." *Econometrica* 26: 24–36. Doi: http://dx.doi.org/10.2307/1907382.

Received April 2015
Revised October 2015
Accepted March 2016