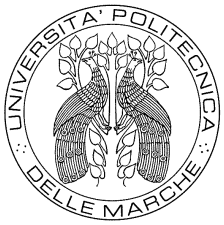

UNIVERSITÀ POLITECNICA DELLE MARCHE

Dipartimento di Scienze Economiche e Sociali

Of Butterflies and Caterpillars: Bivariate Normality in the Sample Selection Model

Claudia Pigini

QUADERNI DI RICERCA n. 377*

June 2012

(*) La numerazione progressiva continua dalla serie denominata "Quaderni di ricerca — Dipartimento di economia"

Collana curata da Massimo Tamberi

ISSN: 2279-9575

Abstract

Since the seminal paper by Heckman (1974), the sample selection model has been an essential tool for applied economists and arguably the most sensitive to sources of misspecification among the standard microeconometric models involving limited dependent variables. The need for alternative methods to get consistent estimates has led to a number of estimation proposals for the sample selection model under non-normality. There is a marked dichotomy in the literature that has developed in two conceptually different directions: the bivariate normality assumption can be either replaced, by using copulae, or relaxed/removed, relying on semi and nonparametric estimators.

This paper surveys the more recent proposals on the estimation of sample selection model that deal with distributional misspecification giving the practitioner a unified framework of both parametric and semi-nonparametric options.

JEL Class.: C14, C18, C24, C46
Keywords: Sample selection model, bivariate normality, maximum likelihood, copulae, semiparametric methods

Indirizzo: Dipartimento di Scienze Economiche e Sociali, Università Politecnica delle Marche. E-mail: c.pigini@univpm.it

Contents

Of Butterflies and Caterpillars: Bivariate Normality in the Sample Selection Model*

Claudia Pigini

What the caterpillar calls the end, the rest of the world calls a butterfly. Lao Tzu.

1 Introduction

Since the seminal papers by Gronau (1974) and Heckman (1974), (1976), (1979), the sample selection model has been an essential tool for applied economists[1] and arguably the most sensitive to sources of misspecification among the standard microeconometric models involving limited dependent variables. In particular, both maximum likelihood and the two-step estimators, put forward in Heckman (1974) and Heckman (1976) respectively, heavily rely on the assumption of bivariate normality. When this assumption fails, the maximum likelihood estimator is inconsistent (Arabmazar and Schmidt, 1982) and some of the weaker conditions ensuring consistency of the two-step estimator may also not hold.

Considerable theoretical effort has been put into dealing with non-normality in the sample selection model. Even recently, some of the research on this matter has returned to the issue of testing its distributional assumptions. There are extensions to proposals originally put forward in the 80's: van der Klaauw and Koning (2003) derived an LR test from the semiparametric approach of Gallant and Nychka (1987); Montes-Rojas (2011) runs a Monte Carlo study for the LM test statistic based on the Edgeworth series expansion that serves as a suitable alternative distribution in Lee (1984); Lucchetti

I am grateful to Jack Lucchetti for the useful suggestions and discussions in the process of writing this paper. I am also grateful to Stefano Staffolani, Alessandro Sembenelli, Lorenzo Cappellari, Alessandro Sterlacchini, Alessia Lo Turco, Chiara Gigliarano and one referee for their comments.

[1] According to Google Scholar at 5-19-2011, the three papers Heckman (1974), (1976) and (1979) count 20273 citations.

and Pigini (2011) also investigate the finite-sample properties of conditional moment tests for bivariate normality in models with sample selection as suggested in Smith (1985) and (1987). However, there is no comprehensive work that establishes the superiority in terms of performance of one test statistic or the other, so there is, to date, no common procedure to test bivariate normality in the sample selection model.

In any case, when evidence is at odds it is tempting to conclude that there is a non-normality issue. The need for alternative methods to get consistent estimates has led, in the last thirty years, to a number of estimation proposals for the sample selection model under non-normality. There is a marked dichotomy in the literature that has developed in two conceptually different directions: the bivariate normality assumption can be either replaced or relaxed/removed.

The first approach consists of replacing the assumption of bivariate normality with a specific alternative multivariate distribution, therefore remaining within the parametric framework. This way of proceeding leads to the use of alternative known distributions to model the marginals and the correlation structure. Among others, a fairly recent contribution to this approach has been made by the introduction of copulae to model sample selectivity. Of course this implies that the researcher has to select the appropriate distribution(s) and rewrite, in many cases, the log-likelihood function on a case by case-wise basis. However, if correct distributional assumptions have been made [2], these estimators will have all the desirable properties. The second approach is based on the idea of removing the distributional assumption either in part or altogether. This led to a number of contributions proposing semi or nonparametric estimators mainly of the two-step model with the aim of removing or generalising the correction of the selectivity bias. This approach often results in adding non-linear terms to the conditional mean especially in the main equation which contains the parameters of economic interest.

But what is really the cause of non-normality? The failure of a distributional assumption may be seen as a specification error due to wrong *a priori* conjectures on the specification of the *model* and not a problem with the data such model is imposed upon. Perhaps, there are other hypotheses that are being violated within the model specification as, for instance, the linearity of the conditional mean or wrong conditioning. Intuitively, omitted non-linear terms would be then included in the disturbances distorting the characteristics of their distribution. In this spirit is, for example, the RESET-like normality test of Pagan and Vella (1989) for univariate limited

[2] provided they can be tested

dependent variable models, where non-linear terms in the conditional mean serve as detectors of departures from normality in the error term. In other words, a linear univariate model with a non-normally distributed error term can be seen as a model with non-linearities in the conditional mean and a normally distributed error term. Thus, bivariate non-normality can be seen, from another perspective, as a consequence of wrong conditioning; in this sense, the joint distribution of the disturbances in the selection model will deviate from normality inasmuch as the conditional means are non-linear. Viewing non-normality as some kind of a "dual" problem basically gives the practitioner the opportunity of rewriting the model or modifying the specification for the bivariate normality assumption to hold; whenever possible, "inducing" bivariate normality ensures the maximum likelihood estimates are consistent and fully efficient.

Alternatively, the literature offers the practitioner two options: either replacing the distributional assumption or relying on semi-nonparametric methods, thus disposing of bivariate normality. The choice of following one of the various approaches available can be made on the basis of two main criteria: first the practitioner can compare non-nested models by means of traditional diagnostics such as a Vuong test (1989) and information criteria, when the likelihood is available, or relying on cross-validation to choose between semi or nonparametric models. Secondly, the choice can be driven by the purpose of the estimates: if the main interest is on the structural parameters, the practitioner may find more appropriate to change the stochastic structure of the model in order to get consistent estimates of the behavioural specification; alternatively, he can lean towards a more flexible specification, through semiparametric methods, whenever the aim of the model is forecast and statistical accuracy is crucial.

The aim of this paper is to survey the more recent proposals on the estimation of sample selection model that deal with distributional misspecification giving the practitioner a unified framework of both parametric and semi-nonparametric options. We add to previous surveys (Vella (1998), Pagan and Ullah (1999)) very recent approaches, as, for example, the use of copulae to overcome the non-normality problem.

The paper is organised as follows: section 2 briefly describes the sample selection model and its estimation by both two-step and maximum likelihood under correct specification; section 3 reviews those approaches that replace the assumption on the underlying distribution of the error terms; section 4 reviews semi-nonparametric proposals; section 5 contains an example using data from Mroz (1987) of some of the previous methods. Section 6 concludes.

2 The sample selection model

The problem of selectivity bias first arose as an empirical matter in Heckman (1974) and Gronau (1974) when estimating the wage distribution of working women conditional on facing the decision of entering the labour market. In Heckman (1974) we find the first maximum likelihood estimator while in Heckman (1976) the two-step estimator was first introduced.

2.1 "Why are the data missing?" [3]

Since the seminal paper by Heckman (1979), that clarifies and extends the work of Heckman (1976), selection bias has been referred to as a specification error caused by the presence of non-random missing data. Such misspecification results in the inconsistency of least squares estimates of the behavioural parameters of interest. Suppose you have an iid sample of n cross-sectional observations and you want to estimate the parameters of the following system of two simultaneous equations

$$Y_{1i} = X'_{1i}\beta_1 + U_{1i} \quad (1)$$
$$Y_{2i} = X'_{2i}\beta_2 + U_{2i} \quad (2)$$

where X_{1i} and X_{2i} are full-ranked sets of regressors with respectively k_1 and k_2 explanatory variables. Error terms have zero mean and cross-equation non-zero covariance. When data on response variables are available for all observations, consistent estimates of β_1 and β_2 can be trivially obtained by ordinary least squares.

Suppose now that Y_1 is censored, which means we have all the information only for a subsample of p observations while Y_1 is missing for the remaining $n-p$ observations. The bias in the estimation of β_1 arises when it is Y_2 that describes the mechanism determining which observations are selected into the subsample of complete data. In other words, for the p observations to be considered non-randomly selected, "we must know why the data are missing" (Heckman, 1976). A number of examples fall into this framework. As mentioned above, the most celebrated one is the distribution of the observed wage faced by working women who are, however, a subset of a larger sample of females (non-randomly) deciding whether or not to participate into the labour market (see Heckman (1974) and Gronau (1974)). Other examples are listed in Heckman (1979) such as wages of trainees, migrants and union members.

[3] Heckman (1979)

Following Heckman's notation, for equation (1), the population regression function would be

$$\mathrm{E}\left[Y_{1i}|X'_{1i}\beta_1\right] = X'_{1i}\beta_1 \quad \text{for} \quad i = 1...n \tag{3}$$

while the regression function of the selected sample would depend on a *selection rule*, rule that put those p observations in the subsample in the first place. Let us define (conventionally) this rule as follows: Y_{1i} is observed only if $Y_{2i} \geq 0$, where zero is a normalised threshold, and define a dummy variable d_i that is equal to 1 if $Y_{2i} \geq 0$ and zero otherwise. So now we can write the regression function for the selected subsample as

$$\mathrm{E}\left[Y_{1i}|X'_{1i}\beta_1, d_i = 1\right] = X'_{1i}\beta_1 + \mathrm{E}\left[U_{1i}|d_i = 1\right] \quad \text{for} \quad i = 1...p. \tag{4}$$

As pointed out in Manski (1989), in absence of selectivity bias, equation (3) equals equation (4):

$$\mathrm{E}\left[Y_{1i}|X'_{1i}\beta_1, d_i = 1\right] = \mathrm{E}\left[Y_{1i}|X'_{1i}\beta_1, d_i = 0\right] = \mathrm{E}\left[Y_{1i}|X'_{1i}\beta_1\right] \tag{5}$$

Using the definition of d_i we can rewrite equation (4) as

$$\mathrm{E}\left[Y_{1i}|X'_{1i}\beta_1, U_{2i} \geq -X'_{2i}\beta_2\right] = X'_{1i}\beta_1 + \mathrm{E}\left[U_{1i}|U_{2i} \geq -X'_{2i}\beta_2\right] \tag{6}$$

Equation (6) openly shows why the selectivity bias arises. It is clear that if we were to use ordinary least squares to estimate the parameters of equation (1), we would be omitting the conditional mean of U_{1i} on the right hand-side of equation (6) that is non-zero and a function of X_{2i}, so the estimates would be inconsistent. An endogeneity problem also appears as disturbances are correlated with regressors since U_{1i} depends on X_{2i}. In particular, the distribution of the latent variable, let's say Y^*_{1i}, does not depend on U_{2i}. However Y_{1i} does. Following the example of the labour force participation, we may separate the observed wage Y_{1i} for the selected subsample of p working women and the latent wage Y^*_{1i} that we dot not observe for those $n-p$ women as it does not exceed their reservation wage. Clearly, the distribution of the observed wage Y_{1i} depends on the decision of entering the labour market which, following the above normalisation, occurs whenever $Y_{2i} \geq 0$.

The maximum likelihood and two-step estimators currently used for models with sample selection bias were purposed respectively in Heckman (1974) and in Heckman (1976) (1979). The following sections briefly review these estimators.

2.2 Maximum likelihood estimator

The maximum likelihood estimator for the sample selection model, also known as Type II Tobit model (Amemiya, 1984), is based on the idea of estimating the parameters of the joint distribution of Y_{1i}, Y_{2i} for the whole sample instead of focusing on the conditional mean of Y_{1i} for the selected one.

The key assumption for the maximum likelihood estimation is the joint normality of the disturbances (U_{1i}, U_{2i}):

$$\begin{pmatrix} U_{1i} \\ U_{2i} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}; \begin{pmatrix} \sigma^2 & \rho\sigma \\ \rho\sigma & 1 \end{pmatrix} \right)$$

where σ^2 is the variance of U_{1i} and ρ the correlation coefficient between the two error terms.

The log-likelihood function for model (1)-(2) can be written as

$$\sum_{d_i=1} \ln \int_{-Z_{2i}}^{+\infty} \varphi_2(Z_{1i}, U_{2i}) dU_{2i} + \sum_{d_i=0} \ln \int_{-\infty}^{+\infty} \int_{-\infty}^{-Z_{2i}} \varphi_2(U_{1i}, U_{2i}) dU_{1i} dU_{2i} \quad (7)$$

where φ_2 is the bivariate standard normal density function, $Z_{1i} = \frac{Y_{1i} - X'_{1i}\beta_1}{\sigma}$ and $Z_{2i} = X'_{2i}\beta_2$. To ease the notation, we parametrise the bivariate normal density by using hyperbolic functions so that instead of the correlation coefficient ρ we will be using $\alpha = \operatorname{atanh}(\rho)$ and the related $c_\alpha = \cosh(\alpha)$ and $s_\alpha = \sinh(\alpha)$. With this quantities and solving the integrals, (7) can be rewritten as

$$\ell = \sum_{d_i=1} \left(\ln \left[\Phi(c_\alpha Z_{2i} + s_\alpha Z_{1i}) \right] - 0.5 Z_{1i}^2 - T \right) + \sum_{d_i=0} \ln \left[1 - \Phi(Z_{2i}) \right] \quad (8)$$

where Φ is the univariate standard normal distribution function, $c_\alpha = \frac{1}{\sqrt{1-\rho^2}}$, $s_\alpha = \rho c_\alpha$ and $T = \ln \sigma\sqrt{2\pi}$. The $k_1 + k_2 + 2$ vector $\hat{\theta}' = (\hat{\beta_1}', \hat{\beta_2}', \hat{\sigma}, \hat{\alpha})$ that maximises (8) is a consistent, fully efficient and asymptotically normal estimator of the parameters of model (1)-(2). The usual rules of identification apply.

Despite its desirable properties, non-normality of the error terms causes the maximum likelihood estimator to be inconsistent (see, for example, Manski (1989) and Vella (1998)). This is, in general, a shortcoming of limited dependent variable models that heavily rely on distributional assumptions

for estimation (Smith, 1989). In particular, for the Tobit and sample selection models, the consequences of distributional misspecification have been investigated by Arabmazar and Schmidt (1982).

2.3 Two-step estimator

While the maximum likelihood estimator of model (1)-(2) is most efficient, in the 70's it was common sense to worry about finding methods that were not as expensive and extremely time consuming as was maximum likelihood estimation at the time. As reported in Heckman (1976), it cost 700$ to estimate the sample selection model by maximum likelihood, while only 15$ using the two-step procedure. It was mostly its manageability, in terms of costs and implementation, that gave Heckman's two-step estimator for the sample selection (1976 1979) such popularity. As mentioned in section 2.2, there is also the matter of finding an estimator that would offer robustness under a weaker specification of the distributional structure. Olsen (1980) and Newey (1999a) even proved that for the two-step estimator to be consistent, weaker assumptions are needed, namely the linearity of the selection correction term. Also, as argued in Vella (1998), the choice of an estimator that is "only" consistent over a maximum likelihood one, results in an acceptable efficiency loss if more robustness to distributional misspecification is gained.

In reviewing the two-step estimator, we follow, however, Heckman's original setting, assuming that the error terms of model (1)-(2) are jointly normal. The two-step estimator is based on the idea of estimating the conditional mean $\mathrm{E}\left[U_{1i}|U_{2i} \geq -Z_{2i}\right]$ that appears in equation (6).

Assuming that the covariance structure of the disturbances is the one used for maximum likelihood estimation, a known result of bivariate normality is

$$\mathrm{E}\left[U_{1i}|U_{2i} \geq -Z_{2i}\right] = \rho\sigma\lambda_i \tag{9}$$

where

$$\lambda_i = \frac{\varphi\left(Z_{2i}\right)}{1 - \Phi\left(Z_{2i}\right)}$$

is the inverse Mill's ratio which is monotone decreasing in the probability of being selected into the subsample. Using λ_i, equation (1) may be corrected for the bias selectivity as follows:

$$\begin{aligned} Y_{1i} &= X_{1i}'\beta_1 + \gamma\lambda_i + V_i & (10)\\ d_i &= X_{2i}'\beta_2 + U_{2i}. & (11) \end{aligned}$$

In practice, the first step consists of estimating a probit model for the selection equation (11 maximising

$$\ell = \sum_{d_i=1} \ln\left[\Phi\left(Z_{2i}\right)\right] + \sum_{d_i=0} \ln\left[1 - \Phi\left(Z_{2i}\right)\right]. \tag{12}$$

This allows to give a consistent estimate λ_i by taking the probit generalised residuals (Gourieroux, Monfort, Renault, and Trognon, 1987) only for the selected sample. The second step is the least square estimation of

$$Y_{1i} = X'_{1i}\beta_1 + \gamma\hat{\lambda}_i + V_i. \tag{13}$$

Despite the ease of implementation, this estimator does have some shortcomings. Other than the loss of efficiency, the disturbances are heteroscedastic by construction. Moreover, the estimator of the covariance matrix is downward biased (see Heckman (1976)). Identification of this model basically relies on the inverse Mill's ratio being sufficiently non-linear in Z_{2i}. For a more detailed discussion on the matter see Manski (1989) and Puhani (2000).

As mentioned earlier in this section, the consequences of distributional misspecification have been found to be less severe than for the maximum likelihood estimator. Newey (1999a) proved that as long as $\mathrm{E}\left[U_{1i}|U_{2i} \geq -Z_{2i}\right]$ is linear in the bias correction term λ_i and such correction term is non-linear in Z_{2i} (conditions that are met when bivariate normality is the correct specification), the two-step estimator is consistent.

However, with the advancements of numerical optimisation techniques, it has become easier and easier to estimate this model by maximum likelihood which ensures that estimates are consistent and most efficient. That is, if the underlying distribution is correctly specified.

3 Parametric estimators

In the previous sections, we have recalled that while the distributional misspecification in the sample selection model leads to the inconsistency of the maximum likelihood estimates consequences are not so severe for the two-step estimator. However, falling back on the latter still brings the efficiency loss and having to deal with an inconsistent covariance matrix estimator. For these reasons, since the early 80's, the literature has agreed upon the need of finding alternative estimators for the sample selection model when the assumption of bivariate normality is violated. As introduced in section 1, dealing with non-normality brought to a number of research works that

correct such misspecification either generalising the conditional mean specification or working on the assumption of the bivariate distribution of the disturbances.

This section surveys the main estimation proposals that replace the bivariate normality with an alternative distribution. Some approaches proceed in a two-step manner thus intervening on both the distributional assumption for the selection equation and on the functional form of the bias correction term while the re-specification of the bivariate stochastic structure of the disturbances in the sample selection model implies that the log-likelihood has to be rewritten for any given choice of specific alternative distributions of the disturbances.

In this section, we first briefly review the proposals of Olsen (1980) and Lee (1982) who a two-step estimation under alternative distributions of the selection equation error term. Then we turn to the copula approach. It first appeared in Lee (1983), although it reached wide popularity only in the next decade, with the contributions by Prieger (2002) and Smith (2003).

3.1 Two-step estimators under alternative assumptions

The literature that developed two-step estimators for the sample selection model under non-normality has mainly relied on semiparametric estimators. There are, however, two contributions from the early 80's using a two-step parametric estimator under specific alternative distributions of the selection equation error term: Olsen (1980) and Lee (1982).

As a matter of fact, Newey (1999a) proved the two-step estimator to be consistent under assumptions that are less restrictive than bivariate normality, namely the normal distribution of the selection equation error term U_{2i} and the linearity of the conditional expectation of U_{1i} on U_{2i} shown in (9) of section 2.3. Under these conditions, alternative distributions may be imposed on U_{2i}, in order to estimate the selection equation with maximum likelihood, as long as the conditional expectation remains of linear form.

Instead of assuming normality of the distribution of U_{2i}, Olsen (1980) proposes a two-step estimator in which U_{2i} is uniformly distributed. Clearly the term correcting the selectivity bias is no longer the inverse Mill's ratio. The last term in equation (4) becomes:

$$\mathrm{E}\left[U_{1i}|d_i = 1\right] = \rho\sigma\sqrt{3}\left(Z_{2i} - 1\right) \tag{14}$$

which is a transformation of the residuals of a linear probability model for the selection equation taken only for the sub-sample. Being the correction a linear transformation of the index function of the selection equation,

stronger exclusion restrictions are required for identification. According to Vella (1998), the resulting estimates of this specification are similar to those obtained form the two-step estimation under bivariate normality.

The strategy adopted by Lee (1982) is transforming a general distribution into a standard normal by means of the quantile function. He relaxes the assumption that U_{2i} is normally distributed but follows a specific alternative distribution F_2. Since linearity of $E[U_{1i}|U_{2i}]$ cannot be generalised to any distribution, he needs to assume it. Therefore (14) becomes

$$\mathrm{E}[U_{1i}|d_i = 1] = \rho\sigma \frac{\varphi(J_2(U_{2i}))}{F_2(U_{2i})} \tag{15}$$

where $J_2(U_2) = \Phi^{-1}F_2(U_2)$ is called a normalising transformation. However flexible, the consistency of this estimator requires that both F_2 is correctly specified and that the conditional expectation is linear.

3.2 Copulae

Another way to deal with the misspecification on the underlying distribution in the sample selection model, is to substitute the assumption of bivariate normality with a specific alternative bivariate distribution function. According to Smith (2003), this way was not much pursued due to computational difficulties that can be overcome, however, when the joint distribution is specified as a copula. This statistical tool is mostly used in financial econometrics often to model the tail dependence in risk management problems. In this section, we review the recent applications of the copula approach to sample selection modelling, giving minimum theoretical elements on bivariate copulae that we borrow from Smith (2003) and Bhat and Eluru (2009), in order to ease the exposition. For a detailed review of the copula approach in econometrics, see Trivedi and Zimmer (2007). Throughout this section, we will follow the notation of Smith (2003).

3.2.1 A few elements of Copula theory

A copula is basically a function that takes as arguments pre-specified univariate marginal distribution functions and produces their multivariate distribution. The foundations of copulae lie on Sklar's theorem (1959) which shows that given two marginal distribution functions $F_1(x_1)$ and $F_2(x_2)$ of X_1 and X_2 respectively, and their joint distribution function $F(x_1, x_2)$, then there exist a function C, the copula, such that

$$F(x_1, x_2) = C(F_1(x_1), F_2(x_2)). \tag{16}$$

The copula represents the joint distribution function of two random variables in terms of their distinct marginals separating them from the their dependence structure. Standard properties of bivariate copulae are $C(u,v) = 0$ if $u = v = 0$, $C(1,v) = v$ and $C(u,1) = u$ where $u, v \in [0,1]$. Three examples of simple copulae are

$$\Pi = uv \tag{17}$$

$$W = \frac{u+v-1+|u+v-1|}{2} = \max(u+v-1, 0) \tag{18}$$

$$M = \frac{u+v-|u-v|}{2} = \min(u,v) \tag{19}$$

where (17) is called the Product copula which is equivalent to $F(x_1, x_2) = F_1(x_1)F_2(x_2)$ and models stochastic independence of X_1 and X_2. Expressions (18) and (19) are, respectively, the lower and the upper Fréchet-Hoeffding bounds (see Kwerel (1988)) which have the property of containing any bivariate copula:

$$W(u,v) \leq C(u,v) \leq M(u,v) \tag{20}$$

As customary in these kind of settings, the dependence structure of the bivariate distribution is separately parametrised by a scalar θ whose possible values indexes a member of the family of copulae represented by the notation $C_\theta(u,v)$. For instance, families of bivariate copulae employed in modelling sample selection are Gaussian, Farlie-Gumbel-Morgenstern (FGM henceforth) and Archimedean. The ability that a family of copulae has to represent the degree of dependence between X_1 and X_2 is measured by the extent to which it is able to cover the closed interval delimited by the Fréchet-Hoeffding bounds which are approached by C as θ reaches its extreme values. As an example, in the Gaussian family $C_{-1}(u,v) = W$ and $C_1(u,v) = M$ as $\theta \in [-1,1]$. Furthermore, a family of copulae is comprehensive when is able to fully parametrise the range of dependence between W and M including also Π as a special or limiting case. For families of copulae that are not comprehensive[4] other measures of concordance(discordance) are needed to properly examine their coverage. Most traditional is Pearson's product-moment correlation coefficient ρ which measures linear dependence between X_1 and X_2. However, some of its characteristics, such as the lack of invariance with respect to the margins and the fact that $\rho = 0$ does not imply

[4]one example is the FGM family of copulae which covers Π but not the Fréchet-Hoeffding bounds

independence of X_1 and X_2, are not desirable in the context of copula modelling. Such limitations led to the use of other measures such as Kendall's τ and Spearman's ρ[5]. Among their properties, they do not depend on the marginal distribution functions but only on the copula, are bounded between $[-1, 1]$ and they take value -1 in W, 1 in M and 0 in Π.

3.2.2 Modelling sample selection with copulae

As argued in Smith (2003), a sample selection model where the the joint distribution of the disturbances is built by means of a copula, can be seen as a generalised model that includes the original specification (1)-(2) under the assumption of bivariate normality as a special case. This is because the choices of the marginal distributions and the copula to employ are two entirely separate steps based solely on the analyst's preferences. The literature provides several examples of modelling sample selection using copulae.

A first generalisation of the normal model appeared in Lee (1983) who used a Gaussian copula to model the bivariate distribution of the error terms. Lee's model construction is based on the idea that given two arbitrary know non-normal marginals for U_1 and U_2 in (1)-(2), namely $F_1(U_1)$ $F_2(U_2)$, we can obtain two normally distributed error terms $J_1(U_1)$ and $J_2(U_2)$ applying the standard normal quantile function as follows:

$$J_1(U_1) = \Phi^{-1}(F_1(U_1)) \tag{21}$$
$$J_2(U_2) = \Phi^{-1}(F_2(U_2)). \tag{22}$$

After the inversion, the joint distribution of J_1 and J_2 is specified as a bivariate normal $F(U_1, U_2) = \Phi_2(J_1(U_1), J_2(U_2); \rho)$ which is equivalent to a Gaussian copula

$$C_\theta(u, v) = \Phi_2(\Phi^{-1}(u), \Phi^{-1}(v); \theta) \quad \text{where} \quad \theta \in [-1, 1]. \tag{23}$$

Therefore, the log-likelihood function for Lee's model is

$$\ell = \sum_{d_i=1} \left[\ln \Phi\left(c_\alpha \Phi^{-1}(v) + s_\alpha \Phi^{-1}(u)\right) + \ln f_1 - \ln \sigma\right] + \sum_{d_i=0} \ln v \tag{24}$$

where $u = F_1(U_{1i})$, $v = F_2(U_{2i})$ and f_1 is the density function of U_1. Lee's approach was the first attempt to allow some flexibility to the marginal distributions of the error while maintaining the fully parametrised context of

[5]see the aforementioned references for further details

maximum likelihood estimation. At that time, however, no applications followed and bivariate normality remained the standard distributional assumption. Later, the increasing use of copulae included this approach as a special case of a wider range of choices withing the copula modelling.

The FGM (Farlie - Gumbel - Morgenstern) copula was first used in Prieger (2002) to jointly model health care usage and spells of hospitalisation in presence of self-selection. Prieger's specification of the marginals involves the normal distribution of the selection equation error term for the whole sample and the Gamma distribution for the sub-sample of hospitalised individuals. The FGM family of copulae is specified as

$$C_\theta(u,v) = uv(1+\theta(1-u)(1-v)) \quad \text{where} \quad \theta \in [-1,1]. \tag{25}$$

In general, the log-likelihood for FGM model is

$$\ell = \sum_{d_i=1} \ln\left[(1-v)(1-\theta v(1-2u))f_1\right] + \sum_{d_i=0} \ln v \tag{26}$$

Basing model selection on Vuong tests (Vuong, 1989) and traditional information criteria, such as BIC and AIC, Prieger concludes that the FGM model provides a better fit of the data than both the original and Lee's model. Despite the very handy analytic form of the FGM copula, it has some limitations. For instance, the FGM is not a comprehensive copula and can only capture weak dependence between the marginal distributions (figure (1b) in Bhat and Eluru (2009) is enlightening) as both Spearman's ρ and Kendall's τ are further bounded[6].

To overcome such limitations and take advantage of a manageable analytical form, Smith (2003) introduces the class of Archimedean copulae to sample selection modelling. Archimedean copulae are functions characterised by a so-called generator denoted by $\varphi : [0,1] \to [0,\infty)$ with $\varphi(1) = 0$. The generator is an additive, continuous, convex and decreasing function:

$$\varphi'(t) < 0, \quad \varphi''(t) > 0 \quad \text{for} \quad 0 < t \leq 1$$

A bivariate Archimedean copula is generated by φ as follows:

$$\varphi(C_\theta(u,v)) = \varphi(u) + \varphi(u) \tag{27}$$

As pointed out in Smith (2003), a peculiarity of this structure is the reduction of the arguments' dimension: a bivariate copula function takes two arguments while the generator function takes single one. This is a feature of

[6]Spearman's $\rho \in \left[-\frac{1}{3}, \frac{1}{3}\right]$ and Kendall's $\tau \in \left[-\frac{2}{9}, \frac{2}{9}\right]$

the Archimedean class of copulae that considerably eases their applications especially when it comes to log-likelihood and score computation. In addition, if $\varphi(0) = \infty$ than the inverse φ^{-1} exists, so that bivariate Archimedean copulae can be built as:

$$C_\theta(u, v) = \varphi^{-1}\left(\varphi(u) + \varphi(v)\right) \tag{28}$$

The Archimedean class of copulae is large enough to accommodate a wide range of distributional shapes. In addition, this family includes both comprehensive and non-comprehensive copulae and accommodates a wide range of dependencies. Some of the copulae that belong to this family are listed in table 1.

Table 1: Examples of bivariate Archimedean copulae

Name	$C_\theta(u, v)$	Generator $\varphi(t)$	Range of θ
AHM	$\frac{uv}{(1-\theta(1-u)(1-v))}$	$\log\frac{1-\theta(1-t)}{t}$	$[-1, 1]$
Clayton	$\left[u^{-\theta} + v^{-\theta} - 1\right]^{-1/\theta}$	$(t^{-\theta} - 1)/\theta$	$(0, \infty)$
Frank	$-\frac{1}{\theta}\ln\left(1 + \frac{(e^{-\theta u}-1)(e^{-\theta v}-1)}{e^{-\theta}-1}\right)$	$-\log\frac{e^{\theta t}-1}{e^{\theta}-1}$	$(-\infty, +\infty)$
Gumbel	$\exp\left[-\left((-\ln u)^\theta + (-\ln v)^\theta\right)^{1/\theta}\right]$	$(-\log t)^\theta$	$[1, +\infty)$
Joe	$1 - \left((1-u)^\theta + (1-v)^\theta - (1-u)^\theta(1-v)^\theta\right)^{1/\theta}$	$-\log(1-(1-t)^\theta)$	$[1, +\infty)$

In general, the log-likelihood function of an Archimedean copula for the sample selection model is

$$\ell = \sum_{d_i=1} \ln\left(1 - \frac{\varphi'(u)}{\varphi'(C_\theta)}\right) f_1 + \sum_{d_i=0} \ln v \tag{29}$$

Smith provides applications of the copula approach to both the labour supply case (using data from the 1987 Michigan Panel Study of Income Dynamics) and re-proposing the estimation of the model for health care usage and spells of hospitalisation in Prieger (2002). In both cases, estimations are carried out using the original normal model (1)-(2), Lee's model, Prieger's FGM and Archimedean copulae. In modelling labour supply, he specifies normal marginals, while for the hospitalisation spells he follows the normal-gamma

marginal specification in Prieger (2002). Basing model selection on the final values of the log-likelihood functions, Smith concludes that the FGM copula seems to better fit the data on labour supply while in modelling time spent in hospitals, some of the Archimedean copulae and Prieger's FGM outperform both the original and Lee's models.

The three classes of copulae listed above, namely the Gaussian, the FGM and the Archimedean families, have become a well established framework for copula modelling of sample selectivity. One example is in Genius and Strazzera (2008). The paper contains an interesting comparison between the estimation of a copula structure and the semi-parametric two-step estimation using the female labour force participation data in Martins (2001)[7]. The authors specify a t_ν distribution for the wage equation and logistic marginal for the selection equation. The joint distribution is specified as a Joe copula. The results are similar to those obtained with the semi-parametric two-step estimation in Martins (2001), while the AIC information criterion and the Vuong test suggest that the Joe model provides a better fit than the standard bivariate normal. Bhat and Eluru (2009) model the effect self-selection from residential neighbourhood choice on travel behaviour. A number of copulae among those described above are estimated revealing that the best fit to the data is offered by the Frank copula on the basis of the BIC information criterion. More importantly, results underline that, differently from the normal model, the copulae are able to detect the presence of self-selection, implying that misleading conclusions can by drawn when the assumptions on the dependency structure are inappropriate. Finally, Eberth and Smith (2010) exploit the copula approach to model the sample selection that arises from the decision to participate in to sport activities and their duration. The authors specify the marginals as a normal distribution for the selection equation and gamma for the duration of sport activities. Those are then bound together by a Frank copula. The estimates of the Frank model are compared with those obtained under the assumption of independence confirming the presence of a selectivity bias as the dependence structure appears to be statistically significant.

3.2.3 Remarks on the Copula approach

The literature on copulae applied to sample selection modelling suggests some considerations. The mentioned authors offer a number of remarks we feel the need to organise.

The next section of this paper is entirely devoted to semi and non para-

[7]This paper will be reviewed in detail in the next section.

metric approaches. However, it is necessary, for the sake of argument, to report the discussions of copula authors on the choice of one approach or the other. In favour of copulae, Bhat and Eluru (2009) argue that the semi and non parametric approaches imply the estimation of a large number of parameters, the resulting estimates are inefficient relatively to fully parametrised models, may not allow diagnostics and are limited when it comes to including a large set of covariates. Also, all authors agree upon the computational simplicity of the copula approach that allows the practitioner to exploit familiar tools such as maximum likelihood without requiring simulation methods or numerical integration. As pointed out in Smith (2003) maximum likelihood allows for simultaneous estimation of all the parameters of the model, the analytical score is not difficult to compute and such method, if the usual regularity of conditions are met, ensures consistent, efficient and asymptotically normal estimators. In addition, the practitioner has the possibility of a piece-wise model specification, as marginal distributions are not constrained to belong to the same family of the bivariate copula distribution. Moreover, Genius and Strazzera (2008) argue that copula modelling allows for direct estimation of the dependence structure in the sample selection model while semi and non parametric methods do not.

As always, however, there are pros and cons. There is a series of well known drawbacks when estimating models specified by copulae. As remarked in Smith (2003), in these models Hessian computation is troublesome, so the optimisation falls back on BFGS or, alternatively, on numerical approximations of the log-likelihood second derivatives in order to use the Newton-Rapshon algorithm. This reflects on the estimation of the parameters covariance matrix as well. As a result, the OPG estimator must be used which is known to have disappointing properties in small samples. Also convergence in these models is not always easily achieved as the log-likelihood may, in some cases, turn out to be not concave or present multiple maxima.

Another crucial point stands on the correct specification of these models. Maximum likelihood estimators have all the desirable properties provided the (alternative) distributional assumptions are correct. Smith (2003) argues that it seems plausible that if the maximum likelihood estimator of the sample selection model under the null hypothesis of bivariate normality is inconsistent when the distribution is misspecified, the same result would arise when the null hypothesis of a specific non-normal distribution is incorrect. We add that testing the distributional assumptions is not straightforward. One could think of an Information Matrix test for general correct specification. For one thing, this procedure involves the analytical Hessian matrix and the problems described above would arise. Moreover, even if the test statistic's computation is managed, results are unreliable (Lucchetti and Pigini,

2011). To date, all that can be done is *a posteriori* model selection. Since copulae are non-nested, one could base his decisions on the traditional information criteria (AIC, BIC) or Vuong selection tests. As a result, there is never a best way to fit one's data but only a winner of a pair-wise comparison. Every copula author stresses that, in this context, the choice of a particular model is made on a case-wise basis relating to a given empirical problem and, as we saw, it involves *a priori* decisions not only on the joint distributions and dependence structures, but on the marginals as well. Hence, the choice of a particular model specification relies solely on the researcher's insight of the empirical problem under examination.

When the practitioner faces a non-normality problem in the sample selection model, the option offered by the copula approach is worth pursuing whenever the accuracy of structural parameters estimates is the priority. Well motivated conjectures on the stochastic structure of the phenomenon may lead to specifications better fitting the data than the traditional sample selection model. Moreover, replacing the assumptions on the bivariate distribution, as it happens with copulae, allows the specification of the conditional mean to remain intact, which is crucial to the interpretability of a model from a behavioural point of view.

4 Semiparametric estimators

As mentioned in section 1, a key point of this paper is to join different literature threads that dealt with the failure of the normality assumption in the sample selection model. As we saw, the copula approach allows to substitute such assumption with a specific bivariate stochastic structure leading to the familiar maximum likelihood estimator.

On the contrary, there is a great number of contributions that has gone in the direction of removing distributional assumptions altogether. Since the late 80's, semiparametric methods have been proposed as an alternative approach to the estimation of the sample selection model. This is, however, only one among the numerous applications of semi and nonparametric econometrics. According to Racine (2008), the first published paper on nonparametric methods dates back to Rosenblatt (1956), while theoretical contributions to this research thread started to proliferate in the 60's. But what does semiparametric mean? Let us start with the definition given by Powell (1994):

> Semiparametric modelling is, as its name suggests, a hybrid of the parametric and nonparametric approaches to construction, fitting and validation of statistical models.

The key feature of nonparametric methods is not requiring the specification of some functional forms that might, instead, play an important role in fully parametric models. As a matter of fact, nonparametric methods come in handy whenever the parametric specification of a certain model turns out to be inappropriate and there is no way to assess whether a specific alternative might be better suited. Following Racine (2008), when we use parametric methods we are implicitly assuming that there is a pre-determined data generating process of which a possible realization is the data we are working on. We then use these data to approximate the parameters of the dgp. On the contrary, nonparametric methods let the data determining their joint distribution and a possible appropriate model without any *a priori* conjectures. The term semiparametric stands for the application of a nonparametric method to approximate some unknown function with, however, some parametric restrictions on the quantities of interest or on the argument the unknown function takes.

An extensive review of the estimation proposals for the sample selection model is given in both Vella (1998) and in chapter 8 of Pagan and Ullah (1999). In this section, we first briefly review the semiparametric maximum likelihood of Gallant and Nychka (1987) as it has already been analysed in detail in Vella (1998) and Pagan and Ullah (1999); we then turn to semiparametric approaches to the estimation of the two-step sample selection model with the aim of explaining the options available to the practitioner without giving extensive and complete theory.

4.1 Semiparametric maximum likelihood

A way to overcome non-normality in the sample selection model, is to approximate the joint density function of the disturbances. The semi-parametric method of Gallant and Nychka (1987) consists of specifying the bivariate density function of the error terms by means of a Hermite series expansions that include the bivariate normal density as a special case. Following a recent application of this method by van der Klaauw and Koning (2003), the density approximation can be written as

$$b\left(U_{1i}, U_{2i}\right) = \varphi_2\left(U_{1i}, U_{2i}\right) \sum_{k=0}^{K} \sum_{j=0}^{J} \pi_{k,j} U_{1i}^{k} U_{2i}^{j} \tag{30}$$

where φ_2 is the bivariate standard normal density function, $\pi_{k,j}$ are additional parameters of the bivariate density and π_{00} is normalised to one. Using $b\left(U_{1i}, U_{2i}\right)$ instead of the bivariate normal density, equation (7) becomes

$$\sum_{d_i=1} \ln \int_{-Z_{2i}}^{+\infty} b\left(Z_{1i}, U_{2i}\right) dU_{2i} + \sum_{d_i=0} \ln \int_{-\infty}^{+\infty} \int_{-\infty}^{-Z_{2i}} b\left(U_{1i}, U_{2i}\right) dU_{1i} dU_{2i}. \quad (31)$$

Let us define

$$S\left(U_{1i}, U_{2i}\right) = \sum_{k=0}^{K} \sum_{j=0}^{J} \pi_{k,j} U_{1i}^{k} U_{2i}^{j} - 1$$

so that the log-likelihood function for observation i can be written as

$$\ell = \sum_{d_i=1} \left(\ln \left[\Phi(c_\alpha Z_{2i} + s_\alpha Z_{1i}) + \int_{-Z_{2i}}^{+\infty} S\left(U_{1i}, U_{2i}\right) \varphi_2\left(U_{1i}, U_{2i}\right) dU_{2i} \right] - 0.5 Z_{1i}^2 - T \right) +$$

$$\sum_{d_i=0} \ln \left[1 - \Phi(Z_{2i}) + \int_{-\infty}^{+\infty} \int_{-\infty}^{-Z_{2i}} S\left(U_{1i}, U_{2i}\right) \varphi_2\left(U_{1i}, U_{2i}\right) dU_{1i} dU_{2i} \right].$$

The estimates of β_1 and β_2 are consistent as long as the number of approximating terms increases with the sample size. Melenberg and Van Soest (1993) provide an application of this method to estimate the wage equation of Dutch women conditional on labour force participation. They find that the approximation of Gallant and Nychka (1987) succeeds in detecting the distributional misspecification. van der Klaauw and Koning (2003) exploit this flexible density as an alternative distributional hypothesis to test for bivariate normality. Since $b\left(U_{1i}, U_{2i}\right)$ contains the bivariate normal density as a special case, they test the joint significance of the π_{kj}s (except π_{00}) by means of an LR test statistic.

4.2 Semiparametric two-step estimators

With the exception of the semiparametric maximum likelihood of Gallant and Nychka (1987), over the last thirty years the contributions to the semiparametric estimation of the sample selection model relied mostly on the two-step procedure. In this context, the focus of semiparametric methods is on the estimation of the parameters of the main equation relaxing the assumption of bivariate normality. As mentioned in section 4, while some parametric restrictions are maintained, the distributional assumption on U_{2i} and the linearity in (4) are removed.

First of all, equation (11) can no longer be estimated by maximum likelihood. Basically, univariate normality of U_{2i} assumed in section 2.3 implied that

$$\mathrm{E}\left[d_i = 1 | X_{2i}\right] = \Phi\left(Z_{2i}\right). \tag{32}$$

If we remove the distributional assumption, the conditional mean of d_i on X_{2i} becomes

$$\mathrm{E}\left[d_i = 1 | X_{2i}\right] = F\left(Z_{2i}\right) \tag{33}$$

where $F\left(\right)$ is an unknown function. Secondly, to consistently estimate β_1 in (10), we had to add the inverse Mill's ratio λ_i to the linear specification of the main equation. The inverse Mill's ratio had a specific functional form and entered linearly the main equation. When distributional assumptions are removed, the conditional expectation of U_{1i} on d_i is no longer the inverse Mill's ratio and equation (4) becomes

$$\mathrm{E}\left[U_{1i} | d_i = 1\right] = g\left(Z_{2i}\right) \tag{34}$$

where $g()$ is again some unknown function. The use of nonparametric methods allows, in this case, for some way to approximate that function in order to still obtain an estimate of our behavioural parameters β_1 in (10). In semiparametric models for sample selection, the arguments of both $F()$ and $g()$ are often subject to parametric restrictions. In most cases, it is required that the index function of the selection equation is linear in β_2[8]. This assumption is the so-called *index restriction* (see Vella (1998)).

In the next sections we will list some semiparametric methods for the sample selection model that may serve as feasible options to the practitioner to overcome the non-normality issue. We will first see alternative estimators for the selection equation and then turn to the main equation and to the selectivity bias correction. We must stress that this paper gives by no means a comprehensive theoretical framework for nonparametric methods nor this is its intent. For extensive reviews on nonparametric and semiparametric econometrics the reader is referred to Powell (1994), Pagan and Ullah (1999), Li and Racine (2007) and Racine (2008). The aim is to give the practitioner a guided tour of the options available while giving only very few theoretical elements necessary to a smooth reading. The rest of this section contains some elements on Kernel estimation theory that will ease the exposition of the estimation proposals for the sample selection model.

Kernel density estimator

A standard tool of nonparametric methods is the Kernel density estimator that was first put forward by Rosenblatt (1956). We give a brief review of this

[8]Therefore, we will keep the previous notation $Z_{2i} = X_{2i}'\beta_2$ for the index function.

estimator following Racine (2008) and leaving out its asymptotic properties. For an intuitive description, think of a nonparametric density approximation by means of an histogram. Given a random variable X, an histogram estimation of the unknown density is built as the relative frequency of the x_i that lie in a bin, centred in x_0, $b = [x_0 - h/2; x_0 + h/2]$ where h is the width of b:

$$\hat{f}_i = \frac{1}{nh} \sum_{i=1}^{n} \mathbf{I}\left(x_i \in b\right) \tag{35}$$

Though very straightforward, the histogram is not a particularly efficient tool. It is, for one thing, discontinuous. In principle, Kernel density estimation uses the same mechanism as the histogram. The indicator function $\mathbf{I}$ is basically replaced by a smoothing function $K()$ which, along with some other desirable properties, ensures continuity of the estimated density:

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right) \tag{36}$$

is called the Rosenblatt-Parzen estimator (Rosenblatt (1956) and Parzen (1962)). $K()$ is a non-negative symmetric function with

$$\int_{-\infty}^{+\infty} K(t)dt = 1 \quad \int_{-\infty}^{+\infty} tK(t)dt = 0 \quad \int_{-\infty}^{+\infty} t^2K(t)dt < \infty$$

A Kernel exhibiting these properties is called a second-order Kernel. Among common types of Kernel density estimators, a simple example is the Gaussian Kernel

$$K(t) = \varphi(t) = (2\pi)^{-1/2}\exp\left[-\frac{1}{2}t^2\right]$$

It is well known that Kernel estimators are biased in finite samples. Such bias can be reduced using higher order Kernels although they may take negative values (Parzen, 1962). An example of higher order Kernel is the third-order Gaussian Kernel $K_3(t) = 0.5(3 - t^2)\varphi(t)$. Conditions for consistency of Kernel estimators are that $h \to 0$ for $n \to +\infty$; on the other hand, the variance of the Kernel estimator $V \to 0$ for $nh \to +\infty$. Both the bias and the variance depend on h, in a positive and negative manner respectively. There is, therefore, a trade-off between the estimator being unbiased or efficient that depend on the choice of h which affects the Kernel estimator's finite-sample properties. There are well established criteria to choose optimally the bandwidth h (see the above mentioned references for a detailed exposition) aimed to contain bias and variance at the same time. For instance, for

the second-order Gaussian Kernel the optimal bandwidth is $h = 1.06\hat{\sigma}n^{-1/2}$ where $\hat{\sigma}$ is the sample standard deviation.

A well known shortcoming of Kernel density estimators is the so-called "curse-of-dimensionality" problem which occurs when we want to estimate the joint density of more continuous variables. The problem arises as the convergence properties of the estimator worsens as the number of continuous variables increases. Silverman (1986) shows that with more than a few variables, the curse of dimensionality makes the kernel estimator not as useful as in the univariate case. Racine (2008) argues that, despite its flaws, the nonparametric estimators are consistent even under misspecification of the underlying model and turn out to be, therefore, more useful than parametric estimators. One the contrary, the latter outperform nonparametric estimators in terms of efficiency, but may lose consistency under misspecification.

4.2.1 Selection equation as a single index model

As we are trying to estimate the parameters of the sample selection model in a two-step manner, we first need to focus on the selection equation. Following Pagan and Ullah (1999), let us write

$$d_i = F\left(Z_{2i}\right) + U_{2i}. \tag{37}$$

When we make the traditional normality assumption $U_{2i} \sim N\left(0,1\right)$, $F()$ is the standard normal distribution function and, under correct specification, $\hat{\beta}_2$ obtained by maximum likelihood is consistent and fully efficient. However, when we want to avoid any kind of imposition of a specific distribution on U_{2i}, we must turn to semiparametric methods in order to let $F()$ to be of an unknown form, however monotone, and find some approximation[9]. Most of the estimation proposals reviewed later in this section, maintain the index restriction $Z_{2i} = X'_{2i}\beta_2$ such that the index function for each observation is a scalar. Under this assumption, (37) is called a *single index* model.

A central issue of semiparametric estimation of single index models is identification, namely the intercept cannot be estimated and slope coefficients can be determined up to a scale factor. Since $F()$ is unknown, it absorbs any constant term that we may add to Z_{2i}. Moreover, we could rescale $F\left(Z_{2i}\right)$ by any constant that $\Pr(d_i = 1)$ and $\Pr(d_i = 0)$ would remain the same. Therefore, if we treat the selection equation as a single index model, we only have $k_2 - 2$ estimable parameters, as we must leave out the intercept term and a slope coefficient needs to be normalised. As in Pagan and Ullah (1999),

[9]Semiparametric estimation of binary choice models is treated extensively in Pagan and Ullah (1999) chapter 7.

let us define a vector θ of estimable parameters and decompose our index function Z_{2i} as

$$Z_{2i} = X'_{2i}\beta_2 = \gamma + \beta_2^2 \left(X_{2i}^2 + X_{2i}^3\theta_1 + ... + X_{2i}^{k_2}\theta_{k_2-2}\right) = \gamma + \beta_2^2 \left(\bar{X}'_{2i}\theta\right)$$

where γ is the intercept term, β_2^2 is the coefficient taken to be the scale factor and $\bar{X}_{2i}$ is a matrix containing the regressors X_{2i}^j, for $j = 1...k_2$, excluding the intercept term. With this specification, identification is achieved if the following conditions are met. First of all, for θ_q to exist, with $q = 1...k_2 - 2$, β_2^2 must be non-zero. This becomes clear if we write θ_q as $\theta_q = \beta_2^{q+2}/\beta_2^2$ [10]. Secondly, X_{2i}^2 must be a continuous variable. Let us consider the case where we normalise β_2^2 to unity. Then X_{2i}^2 cannot be rescaled without changing β_2^2 and consequently θ. This uniquely identifies the value of the index function. If we were to use a dummy variable as X_{2i}^2, we would have that the value of the index function does not change to order-invariant transformations.

Now that we have set the conditions for identification of the single index model, we turn to the estimation of the unknown objects in the selection equation, $F()$ and θ. Proposals for a consistent semiparametric estimation of binary choice models are numerous and differentiated on the base of objective functions and approximations of $F()$. Some popular estimators have been put forward by Manski (1975), Cosslett (1983), Ruud (1986), Gallant and Nychka (1987), Cavanagh and Sherman (1998), Powell, Stock, and Stoker (1989), Horowitz (1992), Gabler, Laisney, and Lechner (1993), Ichimura (1993), Klein and Spady (1993), Gozalo and Linton (1994), Froelich (2006). Among the others, Klein and Spady's estimator has been the most analysed and used in empirical applications, presents some theoretical advantages and it is included in all the software packages that perform nonparametric estimation.

In Klein and Spady (1993), the estimator $\widehat{\theta}$ of the estimable parameters of the single index model maximises the traditional log-likelihood function for binary choice models

$$\ell_i = d_i \ln\left(\widehat{F}_i\left(\bar{X}'_{2i}\theta\right)\right) + (1 - d_i)\ln\left(1 - \widehat{F}_i\left(\bar{X}'_{2i}\theta\right)\right) \tag{38}$$

where $\widehat{F}()$ is the leave-one-out nonparametric kernel estimator of the conditional mean $\mathrm{E}\left[d_i|\bar{X}_{2i}\right]$ written as

[10] We add that β_2^2 should preferably be positive for ease of interpretation although it is not necessary for identification.

$$\widehat{F}_i\left(\bar{X}'_{2i}\theta\right) = \frac{\sum_{j=1}^{n} d_i K\left(\frac{\bar{X}'_{2j}\theta - \bar{X}'_{2i}\theta}{h}\right)}{\sum_{j=1}^{n} K\left(\frac{\bar{X}'_{2j}\theta - \bar{X}'_{2i}\theta}{h}\right)} \quad \text{for} \quad j \neq i \tag{39}$$

This is basically the Nadaraya-Watson (Nadaraya (1964a), Nadaraya (1964b) and Watson (1964)) estimator for the regression model, where the form of conditional mean of the dependent variable on the regressors is unknown. Klein and Spady's estimator, along with Ichimura's, has arguably been the most employed tool for the estimation of discrete choice models without distributional assumptions. Among nonparametric estimators, Klein and Spady's is most efficient and therefore often present in empirical applications. For details on its asymptotic distribution and efficiency bounds, see Pagan and Ullah (1999) and, of course, Klein and Spady (1993)

Once we have estimated θ semiparametrically, we need to compute the quantities that will enter the main equation to correct the selection bias instead of the inverse Mill's ratio. We will see in detail in the next section that semiparametric approaches to sample selection use either the estimated index function $\widehat{Z}_{2i}$ or the propensity score $P_i = \Pr[d_i = 1|X_{2i}]$.

4.2.2 Main equation as a partially linear model

Once the quantities that serve as bias correction have been determined by the semiparametric estimation of the selection equation, we turn to the main equation to estimate the parameters of interest from a behavioural point of view. As we saw in section 4.2, a consequence of removing the bivariate normality assumption is that we can no longer write the bias correction as the inverse Mill's ratio so that

$$Y_{1i} = X'_{1i}\beta_1 + g(X_{2i}, \beta_2) + v_i \tag{40}$$

where the function $g()$ is now left undetermined. Depending on estimation proposals, the argument taken by $g()$ can either be the estimated index function $\widehat{Z}_{2i}$ or the propensity score $P_i = \Pr[d_i = 1|X_{2i}]$, that is the estimated probability of being selected into the subsample. Equation (40) is a special case of what in semiparametric econometrics is known as a "partially linear" model. The are mainly two differences between the traditional partially linear model and (40). Firstly, in may cases equation (40) is subject to the index restriction we have seen in 4.2, namely the argument taken by $g()$ must be linear in the parameters so that we often write $g(Z_{2i})$ which is not necessarily required in partially linear models. Secondly, there are different identification requirements. The specification of a partially linear model is

such that the assumption of a linear conditional expectation of Y_{1i} on the regressors is partly relaxed. A part of the model remains linear in a subset of regressors, X_{1i}, while the other is allowed to take any form. As $g()$ is unknown (and possibly linear), X_{1i} and X_{2i} are chosen to be disjoint sets of regressors to ensure the identification of β_2. This condition creates no problem if we look at the partially linear model as a generalisation of the fully linear one. However, in sample selection models it is reasonable to think that variables determining the presence in the subsample may also directly affect the outcome Y_{1i} and hence taken to be a part of X_{1i}. In this sense, complying with identification requirements by using two completely different sets of regressors poses a limit to the desired specification. In any case, strong exclusion restrictions on the regressors are needed to achieve identification. Such restrictions are not required in the parametric model since $g()$ is the inverse Mill's ratio which is a known nonlinear function of Z_{2i}. As pointed out in Pagan and Ullah (1999), the precision of the estimates of β_1 depends on the extra elements contained in X_{2i}. Also, as well as for the single index model, an intercept term can not be consistently estimated as it gets absorbed into the unknown function $g()$. A few pointers on how to estimate the intercept term in this model are given later in this section. Methods for the estimation of partially linear models are described in detail in chapter 5 of Pagan and Ullah (1999) and Racine (2008). The semiparametric literature on sample selection offers several methods to consistently estimate the parameters of interest β_1 while dealing with the unknown form of $g()$.

One of the earliest proposals can be found in Powell (1987) and later reviewed in Powell (1994) which is based on differencing across observations in (40) in order to eliminate the selection bias. This strategy is similar to the one adopted by Robinson (1988) for the estimation of the partially linear model. We will often refer to this estimator as "difference" estimator as it resembles the within group estimator for panel data. The idea is, for the selected subsample, to take two observations i and j

$$\begin{aligned} Y_{1i} &= X_{1i}'\beta_1 + g\left(Z_{2i}\right) + v_i \\ Y_{1j} &= X_{1j}'\beta_1 + g\left(Z_{2j}\right) + v_j \end{aligned}$$

and compute their difference

$$Y_{1i} - Y_{1j} = \left(X_{1i} - X_{1j}\right)'\beta_1 + g\left(Z_{2i}\right) - g\left(Z_{2j}\right) + v_i - v_j$$

Then a pairwise comparison is performed on the value of the index function. For those observations for which $Z_{2i} - Z_{2j} \simeq 0$ and so $g\left(Z_{2i}\right) - g\left(Z_{2j}\right) \simeq 0$, β_1 can be estimated by a least square regression of $Y_{1i} - Y_{1j}$ on $X_{1i} - X_{1j}$ since the selection bias has been eliminated by differencing. In practice,

there are weights being applied that are higher for those pairs of observation that exhibit a small difference in the values of their index function. Thus, after some parametric estimation of β_2 in the selection equation, Powell's estimator may be written as a weighted least squares estimator

$$\widehat{\beta}_1 = \left(\sum_{i=1}^{n} \sum_{j=1 j \neq i}^{n} w_{ij} \left(X_{1i} - X_{1j} \right) \left(X_{1i} - X_{1j} \right)' \right)^{-1} \sum_{i=1}^{n} \sum_{j=1 j \neq i}^{n} w_{ij} \left(X_{1i} - X_{1j} \right) \left(Y_{1i} - Y_{1j} \right)$$

where w_{ij} are weights obtained by

$$w_{ij} = K \left(\frac{\widehat{Z}_{2i} - \widehat{Z}_{2j}}{h_w} \right)$$

which is a kernel estimation, $\widehat{Z}_{2i}$ is equal to $X'_{2i}\widehat{\beta}_2$ and h_w is the chosen bandwidth for the weights computation. Powell's strategy has the practical advantage of not having to approximate $g()$ and allows for identification of all elements in β_1 including the intercept term. However, since the elimination of the selection bias is based on a pairwise comparison of the scalars Z_{2i} and Z_{2j}, the index restriction on the selection equation is necessary. This assumption is then relaxed in Ahn and Powell (1993) by using a nonparametric estimation of the propensity score $\widehat{P}_i$ instead of $\widehat{Z}_{2i}$ to compute the weights. The authors choose the Kernel estimator

$$\widehat{P}_i = \frac{\sum_{j=1 j \neq i}^{n} d_i K \left(\frac{X_{2i} - X_{2j}}{h} \right)}{\sum_{j=1 j \neq i}^{n} K \left(\frac{X_{2i} - X_{2j}}{h} \right)} \tag{41}$$

and, therefore, the weights become

$$w_{ij} = K \left(\frac{\widehat{P}_i - \widehat{P}_j}{h_w} \right)$$

Other estimators are based on the approximation of the unknown function $g()$ rather than elimination of the selection bias. One example was proposed by Cosslett (1991). Maintaining the index restriction he divides the range of $\widehat{Z}_{2i}$ into K intervals and defines K dummy variables D_{ij} with $j = 1...k$ that take value one when the index function for observation i falls into interval j and zero otherwise. He then estimates β_1 by ordinary least squares

$$Y_{1i} = X'_{1i}\beta_1 + \sum_{j=1}^{K} \alpha_j D_{ij} + v_i \tag{42}$$

The number of dummy variables employed should increase with the sample size. Cosslett (1991) proved the estimator is consistent but not asymptotically normal. However, Pagan and Ullah (1999) point out that it should still be used for some preliminary analysis due to its simple computation.

Another estimator can be found in Heckman and Robb (1985) where the authors suggest relaxing the assumption on the inverse Mill's ratio being of a known form. The first step consists of a parametric estimation of β_2; Then they use the estimated propensity score $\widehat{P}_i = \Pr(d_i = 1|X_{2i})$ to correct the selectivity bias in equation (40):

$$Y_{1i} = X'_{1i}\beta_1 + \hat{g}(\hat{P}_i) + v_i \tag{43}$$

where $\hat{g}()$ is the approximation of the unknown form of the selection bias correction in terms of a Fourier series expansion as in Gallant (1981).

Although the above mentioned proposals are all feasible options to overcome the non-normality issue, probably the most popular semiparametric estimator for the sample selection model is Newey's two-step series estimator. It was first put forward by Newey (1988) and then re-proposed in Newey (1999b) and Newey (2009).

The parameters of the selection equation that are needed to compute the index function can be estimated semiparametrically by using one of the methods we mentioned in the previous section. As a fact, Newey (1988) even allows estimation of the selection equation by a linear probability model. However, in Newey (1999b) and (2009) the use of Klein and Spady's estimator is recommended as it exhibits the best performance in terms of efficiency. As mentioned in section 4.2.1, estimating the selection equation with Klein and Spady's method does not allow identification of the intercept term and one of the parameter must be normalised to one. After estimation, one can compute the index function, with the estimated subset of parameter, as linear, $\bar{X}'_{2i}\widehat{\theta}$, or by means of another function that must, however, be known. Following Newey's notation, we indicate this generic function as $\nu(\bar{X}_{2i}, \widehat{\theta}) = \widehat{\nu}_i$. Since $\widehat{\theta}$ does not contain an intercept and the elements are scaled because of the normalisation, Newey suggests some re-proportioning of the index function running a probit estimation on the on the model

$$d_i = \gamma_0 + \gamma_1 \widehat{\nu}_i + \eta_i$$

where $\eta_i \sim N(0,1)$. The index function that will correct the selection bias is then computed as $\widetilde{Z}_{2i} = \widetilde{\gamma}_0 + \widetilde{\gamma}_1 \widehat{\nu}_i$. The second step is then the least squares estimation of the equation

$$Y_{1i} = X'_{1i}\beta_1 + \widehat{g}\left(h_i\right) + v_i \tag{44}$$

where $g()$ is unknown and h_i is some known transformation of $\widetilde{Z}_{2i}$. Newey suggests some possibilities: $h_i = \widetilde{Z}_{2i}$; $h_i = 2\Phi(\widetilde{Z}_{2i}) - 1$; $h_i = \phi(\widetilde{Z}_{2i})/\Phi(\widetilde{Z}_{2i})$. $\widehat{g}\left(h_i\right)$ is the approximation of the unknown function $g()$ by means of a power series in h_i of k terms. That is $\hat{g}\left(h_i\right) = \sum_{j=1}^{k} \alpha_j h_i^j$. An alternative to power series for the approximation of $g()$ is the use of splines which may help avoid problems caused by the presence of outliers and singularities in h_i.

Das, Newey, and Vella (2003) further relax the semiparametric specification of the sample selection model generalising the estimator of Newey (1988). They suggest estimating nonparametrically both equations and then using the propensity score $\widehat{P}_i = \Pr\left[d_i = 1 | X_{2i}\right]$ to correct the selectivity bias in place of the estimated index function. Differently from the partially linear model (40), Das, Newey, and Vella (2003) allow also the main equation to contain an unknown function of some covariates of behavioural interest in X_{1i}. For exposition purposes, let us partition the vector of covariates of the main equation as $X_{1i} = [W_i, Q_i]$ for which $E[Y_{1i}|Q_i] = \lambda(Q_i)$ where $\lambda()$ is an unknown function that needs to be approximated. They give the example of a sample selection model for wages and labour force participation where Q_i is the education taken to be endogenous. Their extension allows for the endogenous variable to enter the main equation non-linearly. The main equation (1) becomes

$$Y_{1i} = W'_i\delta + \lambda\left(Q_i\right) + U_{1i} \tag{45}$$

The parametric restriction $W'_i\delta$ helps avoid the curse of dimensionality problem when estimating (45) semiparametrically. Differently from Newey (1988), the term plugged in (45) to correct the selection bias is an unknown function of the probability of being selected. They assume that $\mathrm{E}[U_{1i}|d_i = 1, X_{1i}, X_{2i}] = g\left(P_i\right)$ where $P_i = \Pr\left[d_i = 1 | X_{2i}\right]$ is the propensity score which means that the bias correction term depends only on the probability of selection where $g()$ is again an unknown function. Hence, the main equation (45) can be rewritten as

$$Y_{1i} = W'_i\delta + \lambda\left(Q_i\right) + g\left(P_i\right) + v_i \tag{46}$$

As $\lambda()$ and $g()$ are unknown functions, the model would require strong exclusion restrictions, namely X_{1i} and X_{2i} should be two disjoint sets of regressors

since P_i and therefore $g(P_i)$ depend on X_{2i}, $P_i(X_{2i})$. Assuming that $g()$, $\lambda()$ and P are smooth functions and that the distributions are continuous, they allow $\lambda()$ to be identified up to a constant (similarly to what happens in Klein and Spady). Also, the intercept term cannot be estimated.

The two-step estimation of (46) employs power series for the nonparametric estimation of both the selection and the main equation. The first step consists of a linear probability model where the explanatory variables are the regressors X_{2i} plus their powers and interactions. To write in a synthetic way a power series for a row vector of regressors, let us define the operator

$$\mathrm{p}_k\left(X_i\right) = \left[X_i' \otimes \mathrm{p}_{k-1}\left(X_i\right)\right]_+ \quad \text{where} \quad \mathrm{p}_1 = X_i'.$$

The operator p_k indicates the by-row Kronecker or horizontal direct product where the symbol $[\quad]_+$ basically eliminates duplicate elements from the Kronecker product [11]. The selection equation can now be written as

$$d_i = \mathrm{p}_k\left(X_{2i}\right)\gamma + \nu_i \tag{47}$$

After estimation by means of a linear probability model, the fitted propensity score for the selected sample $\widehat{P}_i = \Pr\left[d_i = 1 | X_{2i}\right]$ needs to be computed as $\hat{P}_i = \mathrm{p}_k\left(X_{2i}\right)\widehat{\gamma}$ Approximating $\lambda()$ and $g()$ (46) again by means of power series, the second step consists of a least square estimation of

$$Y_{1i} = W_i'\delta + \mathrm{p}_k\left(Q_i\right)\pi + \sum_{s=1}^{S}\alpha_s h_i^s + v_i \tag{48}$$

where h_i is a function of $\widehat{P}_i$; As in Newey (1988), Das, Newey, and Vella (2003) use the identity $h_i = \widehat{P}_i$ or $h_i = \phi\left(\widehat{P}_i\right)/\Phi\left(\widehat{P}_i\right)$ with the leading term reproducing the original parametric specification.

As we saw, a common characteristic of the estimators that are based on approximation of unknown functions in the main equation is the impossibility of estimating the intercept term due to identification issues. We now turn to how to estimate the intercept as it is the last step in chronological order when we estimate semiparametric models of this kind. Estimators of the intercept have been proposed by Heckman (1990), Andrews and Schafgans (1996) and (1998).

Let us call the intercept term of the main equation (40) β_0. The estimator proposed by Heckman (1990) is

[11] For instance, if we had a two-dimensional vector of regressors $X_i = X_{mi}, X_{ni}$, the operator p_k with $k = 2$ (for simplicity) would produce $\mathrm{p}_2 = \left[X_{mi}^2, X_{mi}X_{ni}, X_{ni}^2\right]$

$$\widehat{\beta}_0 = \frac{\sum_{i=1}^{n} \left(Y_{1i} - X'_{1i}\widehat{\beta}_1 \right) d_i \mathbf{I} \left(\widehat{Z}_{2i} > w \right)}{\sum_{i=1}^{n} d_i \mathbf{I} \left(\widehat{Z}_{2i} > w \right)} \tag{49}$$

where we consider X_{1i} not including an intercept term and w is a smoothing parameter. Andrews and Schafgans (1996) and (1998) substitute $\mathbf{I}$ with a smoothing function as follows:

$$\widehat{\beta}_0 = \frac{\sum_{i=1}^{n} \left(Y_{1i} - X'_{1i}\widehat{\beta}_1 \right) d_i S \left(\widehat{Z}_{2i} - \nu_n \right)}{\sum_{i=1}^{n} d_i S \left(\widehat{Z}_{2i} - \nu_n \right)} \tag{50}$$

where ν_n is a smoothing parameter increasing in the sample size. S is a non-decreasing functions as, for instance, the function indicated in Pagan and Ullah (1999):

$$\begin{aligned} S(x) &= 0 && \text{for} && x \leq 0 \\ S(x) &= e^{-x/(\nu_n - x)} && \text{for} && 0 < x < b \\ S(x) &= 1 && \text{for} && x \geq b \end{aligned}$$

The literature offers some application of semiparametric estimation of sample selection models mostly for comparison purposes. Newey, Powell, and Walker (1990) replicate the estimation of female annual work hours selected by labour force participation with endogenous wage rate first shown in Mroz (1987). They re-propose the two-stage least squares and Heckman's two-step estimation comparing them with Powell (1987) "difference" estimator and Newey's two-step series. For the selection equation, the hypothesis of normality of the error term cannot be rejected based on the comparison between the probit model and the semiparametric estimation of Ichimura (1993) and Klein and Spady (1993). The selection equation is therefore estimated by means of the traditional probit model. Estimated parameters are then used to compute the index function. h_i inverse Mill's ratio times the powers of the index function: $h_i = \left[\phi \left(\widehat{Z}_{21} \right) / \Phi \left(\widehat{Z}_{2i} \right) \right] \widehat{Z}_{2i}^{j}$ with $j = 0, 1$. A formal test is derived for non-nested models by means of quadratic forms of the differences in the estimated coefficient on the inverse asymptotic covariance matrix of the differences. Results show that both semiparametric estimation do not differ strongly from the parametric one.

The comparison is extended in Ahn and Powell (1993) who show alongside the previous estimates the results of the "difference" estimator with their nonparametric selection equation. The results of this last estimation are

somewhat closer to the two-stage least squares rather than to two-step or semiparametric estimation.

Martins (2001) analyses data on wage and labour force participation for married women in Portugal. The choice of using Newey's two step series estimator is driven by the inadequacy of the probit model for the selection equation within the two-step estimation. Non-normality is assessed by means of Horowitz and Härdle (1994)'s specification test. Following Newey (1988), Martins estimates the selection equation with the estimator proposed in Klein and Spady (1993) using a fourth-order Kernel. She then computes the estimated single index $\widehat{Z}_{2i}$ which enters the main equation by the unknown function $g()$. Such function is then approximated by the power series $\hat{g}_i\left(\widehat{Z}_{2i}\right) = \alpha_1\widehat{Z}_{2i} + \alpha_2\widehat{Z}_{2i}^2$. The constant term for the main equation is estimated as in Andrews and Schafgans (1998). Results show that the compared models, parametric two-step and Newey's estimator, differ greatly in the value of the education coefficients and in the effect of the husband's wage on the participation decision. Martins finds out that policy implications are quite sensitive to estimation methods since the semiparametric model reveals a substantially different behaviour in participation decisions.

Finally, Hussinger (2008) studies the effect of public R&D subsidies on firms private R&D investment in the German manufacturing sector. The selection equation determines whether the subsidy is received and, if so, the amount of funding. Therefore, the first step is estimated by either parametric probit or Tobit models. For the second step estimation, Hussinger compares the parametric two-step, the estimator in Cosslett (1991), Newey's two-step series estimator and the "difference"-like estimator in Robinson (1988) [12]. He uses 21 dummy variables to implement Cosslett's estimator and a power series in the estimated index function h_i^j with $j = 0, 1$, where $h_i = 2\Phi\left(\widehat{Z}_{2i}\right) - 1$, for Newey's second step. The estimation of the intercept for Newey's and Robinson's models is carried out by means of both Heckman (1990) and Andrews and Schafgans (1998). His results show that the estimated coefficients are robust throughout the different approaches concluding that a possible misspecified distribution is sometimes not crucial in empirical applications.

4.2.3 Remarks on the semiparametric approach

The strongest point of the semiparametric approach applied to the sample selection model is the property of maintaining consistency of these estimators even disposing, in part or altogether, of distributional assumptions.

[12] It is a traditional semiparametric estimator of partially linear models similar to the estimator proposed in Powell (1987)

Some simple versions of these methods are also rather easy to implement (for instance, a linear probability model for the selection equation as in Das, Newey, and Vella (2003) and a power series approximation for the bias correction). Moreover, the simple structure of the two-step estimation maintained in the applications of semiparametric methods facilitates the generalisation of these approaches to cases, for instance, of endogeneity as in Das, Newey, and Vella (2003) and to more elaborate censoring schemes (see Pagan and Ullah (1999)).

However, these estimators do have some shortcomings and therefore need to be handled with some caution.

First of all, as for the traditional two-step estimator recalled in section 2.3, the bias of the covariance matrix estimator remains and so does the problem of correcting it. As we also mentioned in section 4.2, it is well known that semiparametric methods that rely on kernel estimators are prone to suffer from the "curse-of-dimensionality" problem which is likely to occur whenever more than a few continuous variables are included in the specification. Moreover, semiparametric estimators are, in general, less efficient than parametric ones.

Aside from statistical problems, there are other considerations that need to be made. We have seen, for instance, that with semiparametric methods we can remove bivariate normality. However, there are often other assumptions that need to be made that are almost as restrictive as a distributional assumption such as the index restriction and the monotonicity of the link function F in the selection equation (see sections 4.2.1 and 4.2.2). Also, there are many other constraining conditions imposed on these models for the purpose of identification therefore posing a limit to the desired specification. Moreover, semiparametric methods, with the exception of Gallant and Nychka (1987), will estimate in a two-step manner a bivariate model whose dependence structure, which we could be interested in, is neither estimated directly nor easy to extract.

The practitioner who wants to estimate a sample selection model using semiparametric methods may also encounter some technical difficulties. To date, packages implementing more complex semiparametric procedures (such as Klein and Spady (1993)) have become quite standard in software[13]. However, the available versions are rather CPU-intensive and the set of options the software provides, such as the choice of the bandwidth or the order and type of kernels, is sometimes quite limited. In cases where this kind of alternative options are more appropriate (for instance, to ensure convergence by

[13]for instance, the `snp` and `sml` packages in Stata (De Luca, 2008) and the `np` package in `R` (Hayfield and Racine, 2008).

means of higher order kernels or simply for replicability purposes), the user should re-write the entire code to implement the estimation. In addition, convergence problems are likely to occur when using more elaborate models such as a specification containing many discrete variables (country dummies, categorical variables for example education, time dummies for cross-sectional waves, geographical provenance of individuals to name a few) that are quite common when dealing with micro-data.

Finally, when applying semiparametric methods, there may be some "interpretation" issues. We have seen in section 4.2.2 that some of these estimators use unknown functional forms for the selection bias correction which implies that non-linear and interaction terms are plugged into the main equation in order to approximate such functions by series expansions. While this approach is of rather simple computation, it often results in changing the original specification of behavioural interest.

When facing a non-normality problem in the sample selection model, the semiparametric literature offers the practitioner a number of alternatives to maintain consistency even when the distribution is misspecified. On the one hand, in those cases where the practitioner is interested in more general results, such as the signs and the significance of the effects of the covariates, even some simple version of semiparametric methods presented offer a rather manageable solution to the misspecification problem. On the other hand, there is a trade off between the flexibility of the functional forms adopted, in order to keep the assumptions to a minimum, and both the computation effort and the readability of the estimation output.

5 Butterflies and Caterpillars

The literature reviewed in this chapter contains a wide range of both theoretical and empirical contributions on how to deal with the failure of the bivariate normality assumption in the sample selection model. In simple terms, one may see the numerous estimation proposals as divided into two large groups of parametric and semiparametric estimators. The first one, including two-step parametric estimators and copulae, offers, in practice, the possibility of replacing bivariate normality with an alternative stochastic structure. The second relaxes either the bivariate normality assumption adopting a flexible bivariate density function, which is the case of Gallant and Nychka (1987), or more often removes it completely as with the semiparametric two-step estimators.

Within each of these groups, model selection is based on standard techniques. As the copula approach remains in the familiar maximum likelihood

framework, the model selection uses the traditional comparison methods for non-nested models based on the log-likelihood, such as information criteria and Vuong test (Vuong, 1989). When using semiparametric techniques, model selection is instead carried out by means of the cross-validation method (for extensive theory see Pagan and Ullah (1999)). It is, however, not clear how to choose between the two alternatives. To the best of our knowledge, the available criteria are not commonly used to decide whichever is best between the parametric alternatives with substituted probabilistic structure and semiparametric methods with the distributional assumption removed. However, the considerations made on both proposals suggest other, even though only subjective, criteria the practitioner can rely on to choose whether to employ parametric or semiparametric estimation options. We have seen that such "criteria" are mostly based on the purpose of the estimation. When there are well motivated conjectures on the stochastic structure of the phenomenon under examination, the copula approach may lead to specifications better fitting the data. If the alternative distribution is correctly specified[14], this way of proceeding models the departures from normality within the stochastic part of the model and therefore maintains the specification of the conditional mean intact, which is crucial to the interpretability of a model from a behavioural point of view and for the computation of quantities, such as elasticities, instrumental to policy conclusions. This process may become non-trivial when semiparametric methods are used as they often employ flexible functional forms that are not easy to re-conduct to a behavioural framework. However, in those cases where the purpose of the estimation is to draw more general policy conclusions without the need for a solid behavioural structure, semiparametric methods unburden the practitioner of the distributional assumption ensuring, at the same time, consistency without the need for further diagnosis on the stochastic part of the model.

However sophisticated, the computation of these methods is rather time consuming and they are often expensive to implement in software. So before investing in the use of these approaches, it might be worth exploring if the non-normality problem can be fixed while remaining within the traditional estimation of the sample selection model.

The idea of somehow "inducing" bivariate normality is suggested by viewing the problem from another perspective: as heteroskedasticity, non-normality may be seen as a specification error due the poor fitting of the chosen model and not a problem with the data the model is imposed upon. From this point of view, bivariate non-normality may just be the consequence

[14] testing the distributional assumption in these cases would not be trivial.

of the violation of other hypotheses[15] within the model specification as, for instance, sub-sampling or the linearity of the conditional mean. This way of approaching the failure of the bivariate normality assumption justifies the attempt of rewriting the model or modifying the specification for the bivariate normality assumption to hold.

In the univariate probit model, seeing non-normality as a specification error is straightforward. Assume we have the following probit model

$$Y_i^* = X_i'\beta + \upsilon$$

where we observe Y_i following $\mathbf{I}(Y_i^* > 0)$ and where υ is non-normal. The probability of $Y_i = 1$ is given by

$$\Pr(Y_i = 1|X_i) = \mathrm{F}(X_i'\beta) \tag{51}$$

where F is the distribution function of υ. Using the standard normal quantile function, equation (51) may be rewritten as

$$\Pr(Y_i = 1|X_i) = \Phi[h(X_i'\beta)]$$

where $h()$ is called the *normalising transformation*

$$h = \Phi^{-1}[\mathrm{F}(X_i'\beta)]$$

that can be approximated by means of a series expansion. In this spirit, Pagan and Vella (1989) proposed the RESET-like normality test for a univariate model. The test basically consists in adding the terms of this series expansion to the conditional mean specification as

$$Y_i^* = X_i'\beta + \gamma_0(X_i'\hat{\beta})^2 + \gamma_1(X_i'\hat{\beta})^3 + \varepsilon \tag{52}$$

with $\varepsilon \sim N(0,1)$ and then testing the hypothesis $\gamma_0 = \gamma_1 = 0$. We may see the non-linear terms inserted in (52) as corrections of the departures from normality or, conversely, the distribution of the disturbances will deviate from normality inasmuch as the conditional mean is non-linear. In this sense, a linear univariate model with a non-normal distribution can be seen as a model with a normally distributed error term and non-linearities in the conditional mean.

This argument does not apply to the bivariate case since the normalising transformation is not invertible. Still, there may be cases in which corrections to the model specification, in order to "induce" bivariate normality, are worth trying such as the insertion of non-linear terms, transformations of the data

[15] Hence the Lao Tzu citation

Table 2: Model specification, Greene (2002)

Main equation (1)		
Y_{1i}	WW	wife's average hourly earnings in 1975 dollars
	CONST	intercept
	AX	wife's actual years of previous labour market experience
X_{1i}	AX2	square of AX
	WE	wife's educational attainment in years
	CIT	= 1 if the wife lives in a large city
Selection equation (2)		
$\mathbf{I}(Y_{2i} > 0)$	LFP	= 1 if the wife worked in 1975
	CONST	intercept
	WA	wife's age
X_{2i}	FAMINC	family income in 1975 dollars
	KIDS	= 1 if the wife has children below the age of 19
	WE	wife's educational attainment in years

or sub-sampling if, for instance, we are in presence of structural breaks. One of these cases is the example presented next.

We estimate a sample selection model using the popular example of wage equation and labour force participation on the dataset from Mroz (1987) composed of 753 married women in 1975 between the age of 30 and 60 of which 428 are working. For both the wage equation and the labour force participation, we use the specification in Greene (2002) summarised in table 2.

The estimation results are displayed in table 3. The first two columns contain the traditional maximum likelihood and two-step estimator recalled in sections 2.2 and 2.3 respectively. The third column contains the semi-parametric two step series estimator first proposed in Newey (1988) and reviewed in section 4.2.2. As argued in Newey (1999b) and Newey (2009), any consistent estimator can be used for the selection equation. The univariate normality test[16] does not reject the hypothesis of U_{2i} being normally distributed so we keep the probit estimation of the selection equation. To estimate semiparametrically the parameters of the main equation (44), we follow Newey, Powell, and Walker (1990) using the series approximation

[16]the test is performed as in Chesher and Irish (1987) and Skeels and Vella (1999).

$h_i = \left[\phi\left(\widehat{Z}_{21}\right)/\Phi\left(\widehat{Z}_{2i}\right)\right]\widehat{Z}_{2i}^{j}$ with $j = 0, 1$. The fourth column reports the result of a Gaussian copula estimation as in Lee (1983) recalled in section 3.2.2. The marginal distribution of the selection equation error term F_2 is taken to be normal while we assume that U_{1i} is distributed as a Student's t_ν. The degrees of freedom ν are estimated by maximum likelihood along with the other parameters. Finally, we try a data transformation taking as the dependent variable for the main equation ln(WW) instead of WW and estimate the traditional sample selection model by maximum likelihood as in section 2.2. The lower part of table 3 contains the results of the bivariate normality test proposed in Lucchetti and Pigini (2011). The null hypothesis of bivariate normality is highly rejected by both versions of the test statistic [17] when the original specification proposed by Greene (2002) is maintained. However, the values of the test statistics drastically decrease when we take ln(WW) instead of WW as the dependent variable of the wage equation and the test statistic based on the selected moment conditions does not even reject the hull hypothesis of bivariate normality. As mentioned earlier in this section, non-normality can be seen as e consequence of a poor modelling choice. The results presented in table 3 show that in this case modifying the specification solves the non-normality problem. Notice also that the estimation by both two-step and maximum likelihood of the original specification reveals the inconsistency problem with non-normality: both the traditional estimators fail to capture the presence of the selectivity bias which is instead confirmed by the semiparametric estimator, consistent under non-normality, and the maximum likelihood estimator under correct specification of the distributional assumption.

Thus, the non-normality problem in the sample selection model justifies the need for some alternative estimation options in order to avoid severe misspecification. Moreover, there are cases where a transformation as simple as ln(WW) or some other re-specifications, that correct the non-normality problem, are possible without the need for alternative parametric of semiparametric estimators. This solution is therefore preferable since maximum likelihood estimator is consistent under correct specification and most efficient. Moreover, we are able to remain within the traditional and familiar framework of the sample selection model and maximum likelihood estimation.

[17] The test statistics based on four and the two selected moment conditions have decent finite sample properties while the other versions highly over-reject the null hypothesis of bivariate normality.

6 Final remarks

This chapter reviews some of the proposals on the estimation of the sample selection model when the bivariate normality assumption does not hold. Alternative estimation options come from both the parametric and the semiparametric literature. The first has most recently focused on the application of copulae to the sample selection model. This approach basically allows for the replacement of the probabilistic structure of the model and relies on maximum likelihood estimation. The proposed semiparametric methods aim instead to maintain consistent estimates while disposing of the distributional assumption. With the exception of the semiparametric maximum likelihood estimator of Gallant and Nychka (1987), the semiparametric literature on the sample selection model proposes a number of two-step estimators.

Still, both these two groups of estimators have some well known flaws and are often based on assumptions as arbitrary as bivariate normality. Moreover, the practitioner may find they are non-trivial to implement in software and there are no objective criteria to choose between parametric and semiparametric estimators when dealing with non-normality in the sample selection model. On this matter, the considerations made throughout this chapter suggest that the choice of the alternative estimation options could also rely on the estimation purpose.

In conclusion, when facing a non-normality problem in the sample selection model, alternative estimators are needed to avoid severe misspecification as it happens in the example proposed in section 5. However, such specification errors may be the consequence of a poor modelling choice and not of non-normality itself. On the contrary, non-normality may be seen as a consequence of a misspecified model and not a problem with the data. In this perspective, whenever possible, attempts to re-specify the model may eliminate the problem causing the distributional misspecification while remaining in the familiar parametric sample selection model.

References

Ahn, H., and J. L. Powell (1993): "Semiparametric estimation of censored selection models with a nonparametric selection mechanism," *Journal of Econometrics*, 58(1-2), 3–29.

Amemiya, T. (1984): "Tobit models: A survey," *Journal of Econometrics*, 24(1-2), 3–61.

Andrews, D. W., and M. A. Schafgans (1996): "Semiparametric Esti-

mation of a Sample Selection Model," Cowles Foundation Discussion Papers 1119, Cowles Foundation for Research in Economics, Yale University.

Andrews, D. W. K., and M. M. A. Schafgans (1998): "Semiparametric Estimation of the Intercept of a Sample Selection Model," *Review of Economic Studies*, 65(3), 497–517.

Arabmazar, A., and P. Schmidt (1982): "An Investigation of the Robustness of the Tobit Estimator to Non-Normality," *Econometrica*, 50(4), 1055–1063.

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: Methodological*, 43(7), 749–765.

Cavanagh, C., and R. P. Sherman (1998): "Rank estimators for monotonic index models," *Journal of Econometrics*, 84(2), 351–381.

Chesher, A., and M. Irish (1987): "Residual Analysis in the Grouped and Censored Normal Linear Model," *Journal of Econometrics*, 34, 33–61.

Cosslett, S. (1991): "Semiparametric estimation of a regression model with sample selectivity," in *Nonparametric and Semiparametric Methods in Econometrics and Statistics*, ed. by G. T. W. A. Barnett, J. L. Powell. Cambridge: Cambridge University Press.

Cosslett, S. R. (1983): "Distribution-Free Maximum Likelihood Estimator of the Binary Choice Model," *Econometrica*, 51(3), 765–782.

Das, M., W. K. Newey, and F. Vella (2003): "Nonparametric Estimation of Sample Selection Models," *Review of Economic Studies*, pp. 33–58.

De Luca, G. (2008): "SNP and SML estimation of univariate and bivariate binary-choice models," *Stata Journal*, 8(2), 190–220.

Eberth, B., and M. D. Smith (2010): "Modelling the participation decision and duration of sporting activity in Scotland," *Economic Modelling*, 27(4), 822–834.

Froelich, M. (2006): "Non-parametric regression for binary dependent variables," *Econometrics Journal*, 9(3), 511–540.

Gabler, S., F. Laisney, and M. Lechner (1993): "Seminonparametric estimation of binary choice models with an application to labor force participation," *Journal of Business and Economic Statistics*, 11, 61–80.

Gallant, A. R. (1981): "On the bias in flexible functional forms and an essentially unbiased form : The fourier flexible form," *Journal of Econometrics*, 15(2), 211–245.

Gallant, A. R., and D. W. Nychka (1987): "Semi-nonparametric Maximum Likelihood Estimation," *Econometrica*, 55(2), 363–390.

Genius, M., and E. Strazzera (2008): "Applying the copula approach to sample selection modelling," *Applied Economics*, 40(11), 1443–1455.

Gourieroux, C., A. Monfort, E. Renault, and A. Trognon (1987): "Generalised residuals," *Journal of Econometrics*, 34(1-2), 5–32.

Gozalo, P., and O. Linton (1994): "Local Nonlinear Least Squares Estimation: Using Parametric Information Nonparametrically," Cowles Foundation Discussion Papers 1075, Cowles Foundation for Research in Economics, Yale University.

Greene, W. H. (2002): *Econometric Analysis*. Prentice Hall, 5 edn.

Gronau, R. (1974): "Wage Comparisons-A Selectivity Bias," *Journal of Political Economy*, 82(6), 1119–1143.

Hayfield, T., and J. S. Racine (2008): "Nonparametric Econometrics: The np Package," *Journal of Statistical Software*, 27(5), 1–32.

Heckman, J. J. (1974): "Shadow Prices, Market Wages, and Labor Supply," *Econometrica*, 42(4), 679–694.

——— (1976): "The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models," *Annals of Economic and Social Measurement*, 5(4), 120–137.

——— (1979): "Sample Selection Bias as a Specification Error," *Econometrica*, 47(1), 153–161.

——— (1990): "Varieties of Selection Bias," *American Economic Review*, 80(2), 313–318.

Heckman, J. J., and R. J. Robb (1985): "Alternative methods for evaluating the impact of interventions," in *Longitudinal Analysis of Labor Market Data*, ed. by J. Heckman, and B. Singer, chap. 4, pp. 156–245. Cambridge: Cambridge University Press.

Horowitz, J. L. (1992): "A Smoothed Maximum Score Estimator for the Binary Response Model," *Econometrica*, 60(3), 505–531.

Horowitz, J. L., and W. Härdle (1994): "Testing a Parametric Model Against a Semiparametric Alternative," *Econometric Theory*, 10, 821–848.

Hussinger, K. (2008): "R&D and subsidies at the firm level: an application of parametric and semiparametric two-step selection models," *Journal of Applied Econometrics*, 23(6), 729–747.

Ichimura, H. (1993): "Semiparametric least squares (SLS) and weighted SLS estimation of single-index models," *Journal of Econometrics*, 58, 71–120.

Klein, R. W., and R. H. Spady (1993): "An efficient semiparametric estimator for binary response models," *Econometrica*, 61, 387–421.

Kwerel, S. M. (1988): "Frechet bounds," in *Encyclopedia of Statistical Sciences*, ed. by S. Kotz, and N. e. Johnson, pp. 202–209. Wiley and Sons, New York.

Lee, L.-F. (1982): "Some Approaches to the Correction of Selectivity Bias," *Review of Economic Studies*, 49(3), 355–372.

——— (1983): "Generalized Econometric Models with Selectivity," *Econometrica*, 51(2), 507–512.

——— (1984): "Tests for the Bivariate Normal Distribution in Econometric Models with Selectivity," *Econometrica*, 52(4), 843–863.

Li, Q., and J. S. Racine (2007): *Nonparametric Econometrics: Theory and Practice*. Princeton University Press.

Lucchetti, R., and C. Pigini (2011): "Conditional Moment Tests for Normality in Bivariate Limited Dependent Variable Models: a Monte Carlo Study," Working Papers 357, Universita' Politecnica delle Marche, Dipartimento di Scienze Economiche e Sociali.

Manski, C. (1975): "Maximum score estimation of the stochastic utility model of choice," *Journal of Econometrics*, 3(3), 205–228.

Manski, C. F. (1989): "Anatomy of the Selection Problem," *Journal of Human Resources*, 24(3), 343–360.

Martins, M. F. O. (2001): "Parametric and semiparametric estimation of sample selection models: an empirical application to the female labour force in Portugal," *Journal of Applied Econometrics*, 16(1), 23–39.

Melenberg, B., and A. Van Soest (1993): "Semi-Parametric Estimation on the Sample Selection Model," Working papers 9334, Tilburg - Center for Economic Research.

Montes-Rojas, G. V. (2011): "Robust misspecification tests for the Heckman's two-step estimator," *Econometric Reviews*, 30(2), 1–19.

Mroz, T. A. (1987): "The sensitivity of an empirical model of married women's hours of work to economic and statistical assumptions," *Econometrica*, 55(4), 765–799.

Nadaraya, E. A. (1964a): "On estimating regression," *Theory of Probability and Its Applications*, 9(1), 141–142.

——— (1964b): "On nonparametric estimation of density functions and regression curves," *Theory of Probability and Its Applications*, 10, 134–137.

Newey, W. K. (1988): "Two-Step Series Estimation of Sample Selection Models," Unpublished Manuscript.

——— (1999a): "Consistency of two-step sample selection estimators despite misspecification of distribution," *Economics Letters*, 63(2), 129 – 132.

——— (1999b): "Two-Step Series Estimation of Sample Selection Models," Working papers 99–04, Massachusetts Institute of Technology (MIT), Department of Economics.

——— (2009): "Two-Step Series Estimation of Sample Selection Models," *Econometrics Journal*, 12(s1), 217–229.

Newey, W. K., J. L. Powell, and J. R. Walker (1990): "Semiparametric Estimation of Selection Models: Some Empirical Results," *American Economic Review*, 80(2), 324–328.

Olsen, R. J. (1980): "A Least Squares Correction for Selectivity Bias," *Econometrica*, 48(7), 1815–1820.

Pagan, A., and A. Ullah (1999): *Nonparametric econometrics*. Cambridge University Press, New York, NY, USA.

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

Parzen, E. (1962): "On estimation of a probability density function and mode," *The Annals of Mathematical Statistics*, 33, 1065–1076.

Powell, J. L. (1987): "Semiparametric estimation of bivariate latent variable models," Working Paper 8704, Social Systems Research Institute, University of Wisconsin-Madison.

——— (1994): *Estimation of Semiparametric Models*vol. 4 of *Handbook of Econometrics*, chap. 41, pp. 2443–2521. North-Holland.

Powell, J. L., J. H. Stock, and T. M. Stoker (1989): "Semiparametric Estimation of Index Coefficients," *Econometrica*, 57(6), 1403–1430.

Prieger, J. E. (2002): "A flexible parametric selection model for non-normal data with application to health care usage," *Journal of Applied Econometrics*, 17(4), 367–392.

Puhani, P. (2000): "The Heckman Correction for Sample Selection and Its Critique," *Journal of Economic Surveys*, 14(1), 53–68.

Racine, J. S. (2008): *Nonparametric Econometrics: A Primer*, vol. 3. Foundations and Trends in Econometrics.

Robinson, P. M. (1988): "Root- N-Consistent Semiparametric Regression," *Econometrica*, 56(4), 931–954.

Rosenblatt, M. (1956): "Remarks on some nonparametric estimates of a density function," *Annals of Mathematical Statistics*, 27(3), 832–837.

Ruud, P. A. (1986): "Consistent estimation of limited dependent variable models despite misspecification of distribution," *Journal of Econometrics*, 32(1), 157–187.

Silverman, B. W. (1986): *Density Estimation for Statistics and Data Analysis*. New York: Chapman and Hall.

Skeels, C. L., and F. Vella (1999): "A Monte Carlo investigation of the sampling behavior of conditional moment tests in Tobit and Probit models," *Journal of Econometrics*, 92(2), 275–294.

Sklar, A. (1959): "Fonctions de répartition à n dimensions et leurs marges," *Publ. Inst. Statist. Univ. Paris*, 8, 229–231.

Smith, M. D. (2003): "Modelling sample selection using Archimedean copulas," *Econometrics Journal*, 6(1), 99–123.

Smith, R. J. (1985): "Some Tests for Misspecification in Bivariate Limited Dependent Variables Models," *Annales de l'INSÉÉ*, 59/60, 97–123.

——— (1987): "Testing the Normality Assumption in Multivariate Simultaneous Limited Dependent Variable Models," *Journal of Econometrics*, 34(1-2), 105–123.

——— (1989): "On the Use of Distributional Mis-Specification Checks in Limited Dependent Variable Models," *The Economic Journal*, Vol. 99, No. 395, Supplement: Conference Papers, 178–192.

Trivedi, P. K., and D. M. Zimmer (2007): *Copula modeling: an introduction for practitioners*, Foundations and Trends in Econometrics. Now Publishers.

van der Klaauw, B., and R. H. Koning (2003): "Testing the Normality Assumption in the Sample Selection Model with an Application to Travel Demand," *Journal of Business & Economic Statistics*, 21(1), 31–42.

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

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

Watson, G. S. (1964): "Smooth regression analysis," *Sankhya*, A(26), 359–372.

Table 3: Sample selection model for wages and labor force participation using Green's Specification on the dataset from Mroz (1987)

	MLE	Two-step	Newey (1989)	Lee (1983)	ln(WW)
Main equation					
CONST	-1.963(1.20)	-0.971(2.06)	-3.833(3.02)[(a)]	-4.500 (0.65)***	0.556(0.25)**
AX	0.028 (0.06)	0.021(0.06)	0.194(0.04)***	0.116(0.03)***	0.023(0.01)*
AX2	-0.000(0.00)	0.000(0.00)	-0.003(0.00)**	-0.002 (0.00)**	-0.000(0.00)
WE	0.457 (0.07)***	0.417(0.10)***	0.385(0.04)***	0.460 (0.05)***	0.065(0.02)***
CIT	0.446(0.32)	0.444(0.32)	0.091(0.23)	-0.037 (0.17)	0.056(0.06)
Selection equation					
CONST	-4.120(1.40)***	-4.157(1.40)***	-4.157(1.40)***	-2.079(1.01)**	-2.999(1.20)**
WA	0.184 (0.06)***	0.185(0.07)***	0.185(0.07)***	0.044 (0.05)	0.121(0.06)**
WA2	-0.002(0.00)***	-0.002(0.00)***	-0.002(0.00)***	-0.001 (0.00)	-0.002(0.00)**
FAMINC	0.000 (0.00)	0.000(0.000)	0.000(0.000)	-0.000 (0.00)**	0.000(0.00)***
KIDS	-0.451(0.13)***	-0.449(0.13)***	-0.449(0.13)***	-0.186 (0.10)*	-0.285(0.11)***
WE	0.095(0.02)***	0.098(0.02)***	0.098(0.02)***	0.149 (0.02)***	0.076(0.02)***
λ	-0.410(0.52)	-1.098(1.27)	-5.23(0.71)***	-	-0.686(0.06)***
λ_1	-	-	-3.85(0.92)***	-	-
σ	3.11	3.20		2.08 (0.11)***	0.83
ρ	-0.13	-0.34		0.93	-0.82
α	-0.13	-0.36		1.678 (0.18)***	
df	-	-	-	2.6(0.71)***	
Log-likelihood	-1581.26	-	-	-1415.10	-911.67
Univariate normality test					
Chi-square(2)		3.06	3.06		
Bivariate normality test					
Fourth m.c., Chi-square(3)	224.62***				7.29*
Selected m.c., Chi-square(2)	109.43***				4.13

[(a)] Intercept term and standard error are computed as in Andrews and Schafgans (1998). $v = \sqrt{n}$ and $b = 0.1$ as in Martins (2001)