Wage Gaps in the New Zealand Labour Market

by

Murat Genç[1]
Department of Economics, University of Otago

and

Murray D. Smith
Health Economics Research Unit, University of Aberdeen

This Version: December 31, 2007

PRELIMINARY AND INCOMPLETE
WORK IN PROGRESS
NOT FOR QUOTATION

Abstract: Using data from the Statistics New Zealand's 2003 CURF (Confidentialised Unit Record File) data set, we estimate wage regressions taking account of sample selection bias arising from the exclusion of individuals with no market income. We use the "copula approach" in the specification of sample selection models. We find evidence of a statistically and economically significant female/male differential. Ethnicity, however, is found to matter for certain groups only, not for Maori.

Keywords: Self-selection model; Copula; Sklar's theorem; Dependence; Spearman's S_ρ.

JEL Classification Codes: C21, C24, C51

[1]Address for correspondence: Department of Economics, University of Otago, P.O.Box 56, Dunedin, New Zealand (E-mail: mgenc@business.otago.ac.nz).

1 Introduction

This article focuses on using the "copula approach" in the specification of sample selection models as applied to the labour market discrimination by gender and ethnicity in the labour force in New Zealand. Simple descriptive statistics indicate that there are significant wage gaps between males and females, and between Pakeha and people from other ethnic groups such as Maori and Pacific Islanders. The human capital models of Mincer and Polachek [21] and Polachek [25] provide an economic explanation for such wage gaps. They are explained as a result of different levels of acquired skills that lead to differences in productivity and hence in wages. The statistical discrimination literature deals with whether these gaps are fully accounted for by human capital variables such as age, experience and educational attainment. The usual approach is to estimate an earnings function by extending the standard form developed by Mincer [20] by including gender and ethnicity dummy variables. A typical problem in estimating these models is that no market wage is observed for individuals who do not work. Including only those individuals who work to form the sample on which the estimation is based could cause *sample selection* bias, since the decision to work may be systematically correlated to potential wages. Statistical techniques were developed to estimate these models following the work of Heckman [12]. One potential drawback of these techniques is the assumption of multivariate normality of the unobservable error terms.

The modelling technique used here derives from a representation theorem due to Sklar, see [28] and [29], in which the joint distribution of random variables can be expressed as a function of its univariate margins: that function being the copula. The copula represents the dependence structure between random variables, it captures entirely their joint behaviour. Whilst there exists an extensive statistical literature on copulas, they have received relatively less attention in econometrics. Applications include Dardanoni and Lambert [5] in economics, and Miller and Liu [19] and Smith [31] and [32] in econometrics. Surveys discussing the usefulness of copulas in the fields of econometrics and finance are respectively Trivedi and Zimmer [33] and Cherubini et al [3]. In regard to statistical modelling, Joe [14, Chapter 11] gives five studies in which copula functions are used to model various multivariate and longitudinal data sets. The specification method suggested by Lee [15] for modelling self-selection provides an example of the copula approach, as will be shown below.

The econometric context in which our wage gap problem is set is one of needing to apply binary models designed to account for data selectivity, should it be present. The last thirty to forty years have seen numerous contributions to the literature on the use of these binary models; see, for example, Vella [34] for a recent survey. However, the vast majority of analyses have depended on the statistical assumption of multivariate normality (Heckman [11]). Although ubiquitous throughout all facets of econometric modelling, the adequacy of inference based on the assumption of multivariate normality has often been questioned, and often found to be wanting in the context of sample selection models. Unfortunately, relaxing multivariate normality by replacing it with an alternative multivariate distribution has received relatively little attention. In the main, this was because of the additional computational burdens that were expected to arise. Instead, the literature developed by focusing on semi-parametric and non-parametric versions of these models, where modelling improvements might be brought about by the use of flexible functions of parameters and the covariates of the random variables; see, for example, the articles in the special edition by Härdle and Manski [10]. The aim of this article is to return to the issue of replacing multivariate normality with an alternative multivariate distribution (or, more precisely, a class of multivariate distributions). The adverse computational consequences are, if anything, mitigated under the proposed method of model specification: the so-called copula approach.

The copula approach is a modelling strategy whereby a joint distribution is induced by specifying marginal distributions, and a function that binds them together: the copula. The

copula parameterises the dependence structure of the random variables, thereby capturing all of the joint behaviour. This then frees the location and scale structures to be parameterised through the margins, one at a time. Most importantly, the copula approach permits specifications other than multivariate normality, although it does retain that distribution as a special case.

As all multivariate distributions have a copula representation (Sklar's Theorem; see Section 2), it might seem that the copula approach is nothing more than the reworking of an old theme. The advantage derived by the copula approach might simply be that econometricians are better practiced at modelling univariate distributions than they are multivariate ones. The ideal, of course, is to choose the right statistical model *a priori*, and hence the right copula. However, when working with empirical data it is rare to have such insight. The specification problem is further compounded in most sample selection models due to latency of the underlying utilitarian variables, and the presence of covariates. When faced with such difficulties, it is advantageous to have to hand a range of potential candidate models from which a preferred fit can emerge. Under a copula approach, families of models can be constructed according to classes of copula functions.

2 Copula Functions

2.1 Theory

The study of the copula function was initiated in the 1940s by Hoeffding [13], and further developed in the post-war period by Fréchet [8]. Particularly important was the work of Sklar, especially his representation theorem from which the copula approach to modelling is derived; see [28] and [29]. For histories of the development of copula theory see Dall'Aglio [4], Schweizer [26], Fisher [7] and Nelsen [24]; also of interest is Sklar [30]. Joe [14], Frees and Valdez [9] and Nelsen [24] present comprehensive surveys of the theory of copula functions.

The simplest case will be set down here to illustrate the theory: the bivariate case. Consider a two-place function $C : \mathbb{I}^2 \to \mathbb{I}$, where $\mathbb{I}$ denotes the closed interval $[0, 1]$ of $\mathbb{R}$, the latter denoting the real line, while $\overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty, \infty\}$ will later be used to denote the extended real line. C is a copula function if it is 2-increasing with margins $C(1, y) = y$ and $C(x, 1) = x$, and grounded such that $C(0, y) = C(x, 0) = C(0, 0) = 0$, where the pair $(x, y) \in \mathbb{I}^2$. By 2-increasing it is meant

$$C(x_2, y_2) - C(x_2, y_1) - C(x_1, y_2) + C(x_1, y_1) \geq 0$$

for every x_1, x_2, y_1, y_2 in $\mathbb{I}$ such that $x_1 \leq x_2$ and $y_1 \leq y_2$.

Sklar's main result is that there exists a copula function which acts to represent the joint cumulative distribution function (cdf hereafter) of random variables in terms of its underlying one-dimensional margins. For example, let $F_1(z_1)$ and $F_2(z_2)$ denote, respectively, the cdf of the random variables Z_1 and Z_2; that is, $F_i(z_i) = \Pr(Z_i \leq z_i)$, where $z_i \in \overline{\mathbb{R}}$ $(i = 1, 2)$, and let $F(z_1, z_2) = \Pr(Z_1 \leq z_1, Z_2 \leq z_2)$ denote the joint cdf. Sklar's result is that the joint cdf can be represented according to

$$F(z_1, z_2) = C(F_1(z_1), F_2(z_2)). \tag{1}$$

The copula representation is a re-formulation of the joint cdf such that it separates the margins F_1 and F_2 from their interaction. So while the copula function takes as arguments the margins F_1 and F_2 in (1), the function itself is independent of those margins. The copula serves to capture the dependence between the random variables Z_1 and Z_2. When F_1 and F_2 are continuous functions, then (1) provides a unique representation of the cdf for any $(z_1, z_2) \in \overline{\mathbb{R}}^2$. Nelsen [24, Section 2.3] provides a proof of (1) that follows the method given in Schweizer and Sklar [27, Chapter 6] where a multivariate version of the theorem is proved. For alternate proofs (multivariate version) see Moore and Spruill [22, Lemma 3.2] and Carley and Taylor [2].

The copula density $c : I\!I^2 \to [0, \infty]$ of a copula C is defined as

$$c(x, y) = \frac{\partial^2}{\partial x \partial y} C(x, y).$$

It cannot be negative-valued as C is 2-increasing. The copula density occurs in the expression for the joint probability density function (pdf hereafter) of continuous random variables. Assuming that F_1 and F_2 are continuous functions, then, from (1), the joint pdf of Z_1 and Z_2 is given by

$$\frac{\partial^2}{\partial z_1 \partial z_2} F(z_1, z_2) = f_1(z_1) f_2(z_2) c(F_1(z_1), F_2(z_2))$$

where $f_i(z_i) = \frac{\partial}{\partial z_i} F_i(z_i)$ denotes the pdf of Z_i, $i = 1, 2$. Fang et al [6] term $c(F_1(z_1), F_2(z_2))$ the "density weighting function".

2.2 Examples

Consider the Product copula:

$$\Pi = xy$$

where here, and in the following examples, $(x, y) \in I\!I^2$. In light of (1), under Π the joint cdf of Z_1 and Z_2 must be given by $F(z_1, z_2) = F_1(z_1)F_2(z_2)$, implying that Z_1 and Z_2 are independent. Thus, the Product copula represents (bivariate) independence. The copula density of Π is obviously equal to unity.

Two further examples are the (bivariate) Fréchet lower bound

$$\begin{aligned} W &= \frac{x + y - 1 + |x + y - 1|}{2} \\ &= \max(x + y - 1, 0) \end{aligned}$$

and the (bivariate) Fréchet upper bound

$$\begin{aligned} M &= \frac{x + y - |x - y|}{2} \\ &= \min(x, y). \end{aligned}$$

These copulas are important in that the closed interval $[W, M]$ has the property of containing all bivariate copulas; namely, $W \leq C(x, y) \leq M$.

For the purposes of statistical modelling it is essential to parameterise the copula function so that data can be used to shed light on the extent of dependence between the random variables of interest. Let

$$C_\theta(x, y)$$

denote a family of copulas, where the members are indexed according to values assigned to θ (possibly vector valued). Provided that the margins F_1 and F_2 do not depend on θ, Sklar's representation (1) holds for all members of a given family.

There are numerous examples of families of bivariate copulas $C_\theta(x, y)$ given in Joe [14] and Nelsen [24], see also Table 1. For example, the family of Bivariate Normal copulas is given by

$$\Phi_2(\Phi^{-1}(x), \Phi^{-1}(y); \theta) \tag{2}$$

where $-1 \leq \theta \leq 1$; setting $\theta = 0$ yields Π. Here, $\Phi(\cdot)$ denotes the cdf of a standard Normal variable, and $\Phi_2(\cdot, \cdot; \theta)$ the cdf of a bivariate standard Normal variable with Pearson's product moment correlation coefficient θ. Note that setting $x = \Phi(z_1)$ and $y = \Phi(z_2)$ in (2) recovers the

bivariate standard Normal cdf, $\Phi_2(z_1, z_2; \theta)$. The copula density of the Bivariate Normal family is given by

$$\frac{\phi_2\left(\Phi^{-1}(x), \Phi^{-1}(y); \theta\right)}{\phi\left(\Phi^{-1}(x)\right)\phi\left(\Phi^{-1}(y)\right)}$$

where $\phi(\cdot)$ denotes the pdf of a standard Normal variable, and $\phi_2(\cdot, \cdot; \theta)$ the pdf of a bivariate standard Normal variable with Pearson's product moment correlation coefficient θ. Replacing x and y with, for example, the cdfs $F_1(z_1)$ and $F_2(z_2)$, respectively, as per (1), yields the bivariate Meta-Gaussian distribution. The family of bivariate Normal copulas nests the Product copula Π as the special case corresponding to $\theta = 0$.

A further example is the Plackett family of copulas:

$$\begin{cases} \frac{1}{2(\theta-1)}\left(s - \sqrt{s^2 - 4xy\theta(\theta-1)}\right) & \text{where } \theta > 0,\ \theta \neq 1 \text{ and} \\ & \quad s = 1 + (x+y)(\theta-1), \\ \Pi & \text{when } \theta = 1. \end{cases} \tag{3}$$

The copula density of the Plackett family is given by

$$\theta\left(s - 2xy(\theta-1)\right)t^{-3}$$

where $t = \sqrt{s^2 - 4xy\theta(\theta-1)}$.

Both of the aforementioned families are examples of comprehensive copulas, in that they include W, Π and M as special cases: for the bivariate Normal this trio is obtained respectively by setting $\theta = -1, 0, +1$, while for the Plackett family the trio is found corresponding to the limits $\theta \to 0^+, 1, \infty$, respectively.

2.3 Measures of Association

The ability of a given family of (bivariate) copulas to represent differing degrees of dependence between random variables can be examined in terms of the extent to which it covers, for every $(x, y) \in \mathbb{I}^2$, the interval between the lower and upper Fréchet bounds for copulas, $[W, M]$. This is generally determined at the extremes of the parameter space. For example, the Bivariate Normal family (2) has full coverage as $\Phi_2(\Phi^{-1}(x), \Phi^{-1}(y); -1) = W$ and $\Phi_2(\Phi^{-1}(x), \Phi^{-1}(y); 1) = M$. Furthermore, this family is comprehensive because it also includes $\Pi = \Phi_2(\Phi^{-1}(x), \Phi^{-1}(y); 0)$. Comprehensive families of copulas contain the full range of dependence structures. However, despite the existence of such families it may nevertheless be counterproductive in modelling contexts to confine attention to comprehensive families because there are typically many other features of data that are of interest.

Many copula families are not comprehensive, one example being the Farlie-Gumbel-Morgenstern family of copulas (FGM hereafter)

$$xy(1 + \theta(1-x)(1-y)) \qquad \text{where } -1 \leq \theta \leq 1 \tag{4}$$

it includes Π, when $\theta = 0$, but it fails to contain either W or M. For such families it is desirable to assess coverage in terms of measures of association. In this respect, most familiar is Pearson's product moment correlation coefficient. However, this measure suffers from a lack of invariance with respect to the margins. In which case it is advisable to seek alternatives to Pearson's measure that satisfy invariance.

Two measures of association that are invariant are Kendall's τ and Spearman's S_ρ. Both are measures of probabilistic concordance and are bound to $[-1, 1]$: if the dependence structure is given by W both measures are equal to -1, if the structure is M both take value $+1$, while for Π both are equal to 0. Moreover, both measures are functions only of the copula of the joint

distribution. For independent pairs (Z_{1i}, Z_{2i}), $i = 1, 2, 3$, that are copies of (Z_1, Z_2), τ and S_ρ are defined as:

$$\tau = \Pr((Z_{11} - Z_{12})(Z_{21} - Z_{22}) > 0) - \Pr((Z_{11} - Z_{12})(Z_{21} - Z_{22}) < 0)$$

and

$$S_\rho = 3\left(\Pr((Z_{11} - Z_{12})(Z_{21} - Z_{23}) > 0) - \Pr((Z_{11} - Z_{12})(Z_{21} - Z_{23}) < 0)\right)$$

Should (Z_1, Z_2) be a pair of continuous random variables, with the copula of the joint distribution given by C, then τ and S_ρ may be simplified (see Nelsen [23]):

$$\begin{aligned} \tau &= 4\int\int_{\boldsymbol{I}^2} C(x,y)dC(x,y) - 1 \\ &= 4E[C(X,Y)] - 1 \end{aligned}$$

and

$$\begin{aligned} S_\rho &= 12\int\int_{\boldsymbol{I}^2} xydC(x,y) - 3 \\ &= 12E[XY] - 3 \end{aligned}$$

Here, X and Y denote standard uniform random variables with joint cdf C. For the FGM family of copulas $\tau = 2\theta/9$ and $S_\rho = \theta/3$, with coverage in terms of these measures: $-2/9 \leq \tau \leq 2/9$ and $-1/3 \leq S_\rho \leq 1/3$.

2.4 The Copula Approach to Model Construction

For the purposes of statistical modelling, it is the converse of the copula representation of the joint cdf given by Sklar's theorem that is relevant. In other words, given models for the margins and a copula function that binds them together, this then has the effect of constructing a statistical model for the random variables of interest, as a joint cdf is specified. Consider, for example, a bivariate setting in which Z_1 and Z_2 denote the variables of interest. Required is a statistical model for the true, but unknown joint distribution of Z_1 and Z_2; naturally, this distribution may depend on parameters and covariates. Under a copula approach, models for the margins $F_1(z_1)$ and $F_2(z_2)$ are proposed, as well as a selection of a copula family C_θ. Then, by (1), these selections have the effect of specifying the joint cdf of Z_1 and Z_2. Intuitively, the copula approach determines each component of the overall model, then engineers them together using a copula function.

An added boon for modelling that results by adopting a copula approach concerns the freedom to specify each margin; for example, identicality in distribution of the margins need not be imposed. Indeed, because the copula representation is unique on the domain of support of the random variables in question, multivariate models can be constructed using a copula approach whose margins can be either continuous or discrete, or mixtures of both.

3 The Self-Selection Model

3.1 Model and Likelihood

Sample stratification, or sample selection, is commonplace amongst microeconometric data, whereby underlying individual choices can themselves influence the observations collected on the random variables of interest. Models of increasing complexity have been constructed to account for stratification in its various guises, should it be present, and a number of these are discussed in texts such as Amemiya [1, Secs.10.6-10.10], Maddala [17] and [18, Part III], and Lee [16, Sec.5.6]. In this section, attention focuses on the self-selection model based on a

binary indicator S that governs whether or not an observation is generated on a second random variable Y. In economics, one often-studied example of this type of self-selectivity is labour force participation, where data generated on labour supply from non-participants is unable to reflect their true market wage.

Typically, the self-selection model is embedded within an utilitarian framework according to a pair of underlying latent random variables Y_1^* and Y_2^*; selectivity arises if these unobservables are mutually dependent. Here it is assumed that the cdf of Y_i^* $(i = 1, 2)$, denoted by $F_i(y_i^*) = \Pr(Y_i^* \leq y_i^*)$, where $y_i^* \in \overline{I\!R}$, depends on the linear function $x_i'\beta_i$ and a scaling factor σ_i, where $X_i = x_i$ $(k_i \times 1)$ is a vector of covariates of Y_i^*, and β_i $(k_i \times 1)$ and scalar σ_i are unknown parameters. The joint cdf of (Y_1^*, Y_2^*) is denoted by $F(y_1^*, y_2^*) = \Pr(Y_1^* \leq y_1^*, Y_2^* \leq y_2^*)$, and it depends on all covariates and parameters.

The purpose of Y_1^* is to represent participation. In the classic self-selection models of microeconometrics, Y_1^* is assumed to be a continuous random variable, however, this can be relaxed without loss of generality. In the self-selection model, Y_2^* is observed for participants. In this section, it is assumed that Y_2^* is a continuous random variable with pdf $f_2(y) = \frac{\partial}{\partial y} F_2(y)$, for all real y in the support of Y_2^*.

The self-selection model arises when observations on a pair of random variables (S, Y) are generated according to the following observation rules:

$$S = 1\{Y_1^* > 0\} \qquad \text{and} \qquad Y = 1\{Y_1^* > 0\} Y_2^*$$

where $1\{A\}$ denotes the indicator function, taking value 1 if event A holds, and 0 otherwise. In effect, Y_2^* can be observed only when $Y_1^* > 0$. The participation mechanism is represented by the Bernoulli variable S, and it derives its properties from those of Y_1^*. Note that when $S = 0$, Y_2^* cannot be observed, and Y is assigned a dummy value of 0.

Let $s_1, ..., s_n$ denote n observations generated on S $(s_j \in \{0, 1\},\ j = 1, ..., n)$, and $y_1, ..., y_n$ the corresponding n observations generated on Y $(y_j \in I\!R,\ j = 1, ..., n)$. For a random sample of n observations, the likelihood function for the self-selection model is given by (c.f. Amemiya [1, eq.(10.7.3)])

$$L = \prod_0 \Pr(Y_{1j}^* \leq 0) \prod_1 f_{2|1}(y_j \,|Y_{1j}^* > 0\,) \Pr(Y_{1j}^* > 0) \tag{5}$$

where $\prod_0$ indicates the product over those observations for which $s_j = 0$, and $\prod_1$ the product over those observations for which $s_j = 1$. The function $f_{2|1}$ denotes the pdf of Y_2^*, given event $Y_1^* > 0$. Its functional form can be derived as follows:

$$\begin{aligned} f_{2|1}(y\,|Y_1^* > 0\,) &= \frac{1}{1 - F_1(0)} \frac{\partial}{\partial y} \left(F_2(y) - F(0, y)\right) \\ &= \frac{1}{1 - F_1(0)} \left(f_2(y) - \frac{\partial}{\partial y} F(0, y)\right) \end{aligned}$$

where $F_1(0) = \Pr(Y_1^* \leq 0) = \Pr(S = 0)$. Substitution into (5) yields

$$\begin{aligned} L &= \prod_0 F_1(0) \prod_1 \left(f_2(y) - \frac{\partial}{\partial y} F(0, y)\right) \\ &= \prod_0 F_1 \prod_1 \left(f_2 - \frac{\partial}{\partial y} F(0, y)\right) \end{aligned} \tag{6}$$

where, for convenience, the index j has been dropped in the first line. Additional simplified notation appears in the second line of (6): F_1 will be used from now on to denote $F_1(0) = \Pr(Y_{1j}^* \leq 0) = \Pr(S_j = 0)$, as too, from now on F_2 denotes $F_2(y_j) = \Pr(Y_{2j}^* \leq y_j)$, and f_2 denotes $f_2(y_j)$.

The component of (6) that presents the most difficulties to evaluate is $\frac{\partial}{\partial y}F(0,y)$. However, should Y_1^* and Y_2^* be independent, for example, then $\frac{\partial}{\partial y}F(0,y) = F_1 f_2$, and L can be separated as per $\left(\prod_0 F_1 \prod_1 (1-F_1)\right) \times \left(\prod_1 f_2\right)$. The likelihood (6) is the general form for the self-selection model. Particular likelihood functions arise from specifications assumed for F etc, a number of which are examined below.

4 Data Set and Empirical Results

4.1 The Data Set

The data set we use in this study is Statistics New Zealand's CURF (Confidentialised Unit Record File) for 2003. The CURF contains unit record level data from the June 2003 quarter Household Labour Force Survey (HLFS) and its supplement the New Zealand Income Survey (IS). It contains 28,982 observations. The information in the CURF has been confidentialised to protect the identity of respondents. In the first place, all household linkages have been removed, although there is the potential still for some household level analysis since variables have been added which identify household types, including variables representing numbers of children, numbers of adults, and weekly household (as well as individual) income. It is, however, impossible to identify, for example, married couples, so that joint estimation of household labour supply is not possible.

Other methods used to ensure the confidentiality of the data include the collapsing of categories for some variables into a smaller number of categories (for example, country of birth has been collapsed to a simple indicator as to whether an individual was born in New Zealand or not), the top-coding of some variables (for example, income has been top-coded to mask outliers amongst high income earners) and some minor degree of data swapping in the case of "unique" individuals whose combination of responses could potentially identify them.

There are many variables provided in the CURF for each individual, including actual and total earnings from the primary and any other wage and salary jobs, income from other sources broken down by source, indicators of receipt of various transfer payments, age, country of birth (and years in New Zealand), ethnicity, employment and labour force status, occupation and industry group (for the employed), local government region, marital status, qualifications, sex, household type, and numbers of dependent children in various age groups. Our analysis is limited to individuals who are in the labour force and who are aged between 15 and 64 years. The resulting sample data set contains 14,360 observations.

Following a similar classification used by Statistics NZ, we categorize the individuals into five ethnic groups: Pakeha/European, Maori, Mixed Maori, Pacific Islanders, and Other. The Mixed Maori represent the survey respondents who ticked both Maori and at least one other ethnic group, which is an option offered in the survey. Thus, Maori represent the individuals who identify themselves solely as Maori. The Pacific Islanders are made up of Samoan, Cook Islanders, Tongan, Niuean, Tokelauan, and Fijians. The ethnic group Other refers to all those not identifying themselves as European, Maori or Pacific Islander. This classification is dominated by Asian people.

The sample means of the variables used in our analysis are given in Table 2. The mean wage of female full-time employees is 89% of the mean wage of male full-time employees. Full-time Maori employees earn 15% less than the full-time Pakeha employees. The difference in mean wages is 25% in comparing the Pacific Islanders with the Pakeha.

4.2 Empirical Specification and Results

We estimate a self-selection model where Y_2^* is the logarithm of wages earned by the employed individuals. Our econometric specification is as follows.

Margins: (i) *Normal*. Of the entire sample of 14360, a total of 922 individuals reported themselves to be unemployed, while 13438 reported undertaking paid employment. Letting Y_1^* denote the propensity of undertaking paid employment, we assume normality for this random variable: $Y_1^* \sim N(x_1'\beta_1, 1)$, thus:

$$F_1 = 1 - \Phi(x_1'\beta_1).$$

(ii) *Shifted-Gamma*. Our specification for Y_2^*, the log-hourly wage earnings of those in paid employment (the self-selected sub-population), follows a Shifted-Gamma distribution. The pdf of $Y_2^* = y > \gamma$ is given by

$$f_2 = \frac{1}{\Gamma(\alpha)} \left(\frac{\nu}{\alpha}\right)^{-\alpha} \exp\left(-\frac{\alpha}{\nu}(y-\gamma)\right) (y-\gamma)^{\alpha-1}$$

where $\alpha > 0$, $\nu = \exp(x_2'\beta_2)$ and here the shift parameter γ is known, assigned the value $\log 6.39$ as the legal minimum hourly wage in New Zealand in 2003 was NZ\$6.40. For $y > \gamma$, the cdf is given by

$$F_2 = 1 - \frac{\Gamma\left(\alpha, \frac{\alpha}{\nu}(y-\gamma)\right)}{\Gamma(\alpha)}$$

where $\Gamma(a, b)$ denotes the incomplete gamma function, $\int_b^\infty \exp(-t)\, t^{a-1} dt$. Treating ν as a constant for the moment, fits to log-hourly wage amongst the 13438 individuals reported undertaking paid employment appear in Figure 1. Black represents the data. Blue the fit of a Normal distribution, and Green the fit of the Shifted-Gamma. It is clearly evident that the Shifted-Gamma provides the superior fit.

Table 3 lists maximum likelihood estimates for the preferred modelling outcome, corresponding to use of the Plackett family of copulas (3). The Plackett model (maximised log-likelihood -6711.34) strongly rejects the nested Independence (Product copula) model (maximised log-likelihood -6779.57) via the single restriction $\theta = 1$, the LR statistic being 136.5. Non-nested comparisons on the basis of AIC (cf. Joe [14, Sec.10.3]) with a number of other copula families lead to the same preferred outcome, namely, the Plackett model.

The point estimate of the dependence parameter θ of the Plackett family is 0.089, with associated standard error 0.011. The Wald test of the hypothesis $\theta = 1$, representing independence between participation and wage earnings, is strongly rejected, $t = -82$. Re-parameterising to Kendall's τ yields the point estimate -0.498, and for the alternate measure Spearman's S_ρ the estimate is -0.678. The strong negative relationship between the job participation and wage earnings variables is consistent with standard reservation wage theory: those individuals not in paid employment have not received a wage offer substantial enough to exceed their reservation wage (the lower is Y_1^* the greater must be Y_2^*).

The results suggest that age, number of children (both pre-school and school-age), marital status, education, and ethnicity are important factors in the job participation decision. Existence of children lowers the probability of employment, while married people are more likely to be employed. Any level of education has a positive effect on the probability of being employed. Although gender does not seem to effect the probability of employment, Maori and Mixed Maori are less likely to be employed compared with Pakeha. The coefficients of Paci and Other are not statistically significant. Migrants are also less likely to be employed in comparison with individuals born in New Zealand.

The age variable enters the log wage regressions in a quadratic form that permits calculation of a turning point, representing the age at which the effect of an extra year becomes negative. The turning point can be computed as the negative of the coefficient on age divided by twice the coefficient on age-squared. The results suggest that the turning point is approximately 47 years of age.

Any form of qualification is found to have a significant positive effect on the wages earned. Even the lowest form of qualification in the form of school level qualification is found to increase

the wages earned by 12.4%, with this effect being as high as 41.7% for university level qualification. The effect of qualification is found to be lower for Maori, but the coefficients of the other ethnicity-qualification interaction terms are not statistically significant.

The immigration status does not seem to have an effect on the wages earned, apart from individuals who have been in New Zealand between 5 and 9 years. These individuals are found to earn 8.2% more then people born in New Zealand. However, if these people belong to other ethnic groups (non- Maori and non-Pacific Islander), they earn 3.6% less compared to people born in New Zealand. Immigrants from Pacific Islands who have been in New Zealand for 10-14 years are also found to earn less.

The coefficients of the ethnicity dummy variables in the log wage equation are statistically significant for the Pacific Islanders and the others only. The Pacific Islanders earn about 11% less than Pakeha, and individuals from the Other group (mostly Asians) earn 5% less. The finding that the difference in the wages of Maori people (sole or Mixed) is not statistically significant is consistent with previous empirical studies in New Zealand.

The results suggest that the wage gaps between males and females are significant both statistically and economically. Females are found to earn 12.1% less than males. It is interesting that gender does have such a significant effect on wages even though its coefficient in the participation equation is not statistically significant.

Table 3 also lists the maximum likelihood estimation results of the standard Heckman selection model. Comparison of the maximised log-likelihood of the two methods indicates that the Plackett model clearly outperforms the Heckman model. (-6711 for Plackett versus -7347 for Heckman.) Although most coefficient estimates and the statistical significance of them are very similar, there are some differences, especially in the log wage equation. For example, the coefficient estimates of the immigration status variables are all statistically significant now, with immigrants of 5-9 years earning 10.4% and immigrants of 10-14 years earning 7.7% more then people born in New Zealand. Perhaps more importantly, the coefficient estimate of the ethnic group Other is not statistically significant when the Heckman method is applied. These are important differences in terms of policy implications or understanding the dynamics of the New Zealand labour market.

5 Conclusions

The distributional assumptions are important in applied sample selection models. This article focuses on using the copula approach in the specification of the sample selection models. Although vast majority of the applied microeconometric analyses are based on the statistical assumption of multivariate normality, the copula approach permits specifications of alternative multivariate distributions. We estimate various specifications to analyse the data from New Zealand labour market by using different copulas and distributional assumptions. We find the Plackett copula performs the best, and report the results. Our results indicate the wage gaps between male and female workers are quite substantial, and cannot be explained by differences in qualification or individual demographics. We also find that ethnicity does not matter for Maori, the indigenous people of New Zealand. However, we find that Pacific Islanders and people from other ethnic groups (most of whom are Asians) earn significantly less than Pakeha even after controlling for qualification and individual demographics.

The results from the standard Heckman model give different conclusions in terms of the effects of immigration status and ethnicity. This demonstrates the importance of the distributional assumptions in applied work. The advantage of taking an approach as suggested here is the ability to compare different models, one of which is the standard Heckman model.

References

[1] Amemiya, T. (1985). *Advanced Econometrics*. Harvard: Cambridge MA.

[2] Carley, H., and Taylor, M. D. (2002). A new proof of Sklar's theorem. In Cuadras, C. M., Fortiana, J., and Rodriguez-Lallena, J. A. (eds.), *Distributions With Given Marginals*, Kluwer: Dordrecht, 29-34.

[3] Cherubini, U., Luciano, E., and Vecchiato, W. (2004). *Copula Methods in Finance*. Wiley: New York.

[4] Dall'Aglio, G. (1991). Frechet classes: the beginnings. In Dall'Aglio, G., Kotz, S., and Salinetti, G. (eds.), *Advances in Probability Distributions with Given Marginals: Beyond the Copulas*, Kluwer: Dordrecht, 13-50.

[5] Dardanoni, V., and Lambert, P. (2001). Horizontal inequity comparisons. *Social Choice and Welfare*, 18, 799-816.

[6] Fang, H.-B., Fang, K.-T., and Kotz, S. (2002). The meta-elliptical distributions with given marginals. *Journal of Multivariate Analysis*, 82, 1-16.

[7] Fisher, N. I. (1997). Copulas. In Kotz, S., Read, C. B., and Banks, D. L. (eds.), *Encyclopedia of Statistical Sciences*, *Update Vol. 1*, Wiley: New York, 159-163.

[8] Fréchet, M. (1951), Sur les tableaux de correlation dont les marges sont donnés, *Ann. Univ. Lyon, Section A, Series 3*, 14, 53-77.

[9] Frees, E. W., and Valdez, E. A. (1998). Understanding relationships using copulas. *North American Actuarial Journal*, 2, 1-25.

[10] Härdle, W., and Manski, C. F. (eds.) (1993). Nonparametric and Semiparametric Approaches to Discrete Response Analysis. *Annals of the Journal of Econometrics*, 58, 1-274.

[11] Heckman, J. J. (1974). Shadow prices, market wages and labor supply. *Econometrica*, 42, 679-694.

[12] Heckman, J. J. (1979). Sample selection as a specification error. *Econometrica*, 47, 153-161.

[13] Hoeffding, W. (1940), Masstabinvariante Korrelationtheorie, *Schriften des Mathematischen Instituts und des Instituts für Angewandte Mathematik der Universität Berlin*, 5, 181-233. (Translated in Fisher, N. I., and Sen, P. K. (1994), *The Collected Works of Wassily Hoeffding*, Springer-Verlag: New York.)

[14] Joe, H. (1997). *Multivariate Models and Dependence Concepts*. Chapman and Hall: London.

[15] Lee, L.-F. (1983). Generalized econometric models with selectivity. *Econometrica*, 51, 507-512.

[16] Lee, M.-J. (1996). *Methods of Moments and Semiparametric Econometrics for Limited Dependent Variable Models*. Springer-Verlag: New York.

[17] Maddala, G. S. (1983). *Limited-Dependent and Qualitative Variables in Econometrics*. Cambridge University Press: Cambridge UK.

[18] Maddala, G. S. (ed.) (1994). *Econometric Methods and Applications* (Volume II). Edward Elgar: Aldershot.

[19] Miller, D. J., and Liu, W.-H. (2002). On the recovery of joint distributions from limited information. *Journal of Econometrics*, 107, 259-274.

[20] Mincer, J. (1974). *Schooling, Experience and Earnings*. Columbia University Press for the National Bureau of Economic Research: New York.

[21] Mincer, J., and Polachek, S. W. (1974). Family investments in human capital: earnings of women. *Journal of Political Economy*, 82, S76-S108.

[22] Moore, D. S., and Spruill, M. C. (1975). Unified large-sample theory of general chi-squared statistics for tests of fit. *The Annals of Statistics*, 3, 599-616.

[23] Nelsen, R. B. (1991). Copulas and Association. In Dall'Aglio, G., Kotz, S., and Salinetti, G. (eds.), *Advances in Probability Distributions with Given Marginals: Beyond the Copulas*, Kluwer: Dordrecht, 51-74.

[24] Nelsen, R. B. (2006). *An Introduction to Copulas*. 2nd edition. Springer-Verlag: New York.

[25] Polachek, S. W. (1981). Occupational self-selection: a human capital approach to sex differences in occupational structure. *Review of Economics and Statistics*, 63, 60-69.

[26] Schweizer, B. (1991). Thirty years of copulas. In Dall'Aglio, G., Kotz, S., and Salinetti, G. (eds.), *Advances in Probability Distributions with Given Marginals: Beyond the Copulas*, Kluwer: Dordrecht, 13-50.

[27] Schweizer, B., and Sklar, A. (1983). *Probabilistic Metric Spaces*. North-Holland: New York.

[28] Sklar, A. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institute de Statistique de l'Universite de Paris*, 8, 229-231.

[29] Sklar, A. (1973). Random variables, joint distributions, and copulas. *Kybernetika*, 9, 449-460.

[30] Sklar, A. (1996). Random variables, distribution functions, and copulas - a personal look backward and forward. In Rüschendorf, L., Schweizer, B., and Taylor, M. D. (eds.), *Distributions With Fixed Marginals And Related Topics*. Lecture Notes - Monograph Series, Volume 28, Institute of Mathematical Statistics: CA, 1-14.

[31] Smith, M. D. (2003). Modelling sample selection using Archimedean copulas. *The Econometrics Journal*, 6, 99-123.

[32] Smith, M. D. (2008). Stochastic frontier models with dependent error components. Forthcoming in *The Econometrics Journal*.

[33] Trivedi, P. K., and Zimmer, D.M. (2005). Copula modeling: an introduction for practitioners. *Foundations and Trends in Econometrics*, 1, 1-111.

[34] Vella, F. (1998). Estimating models with sample selection bias: a survey. *The Journal of Human Resources*, 33, 127-143.

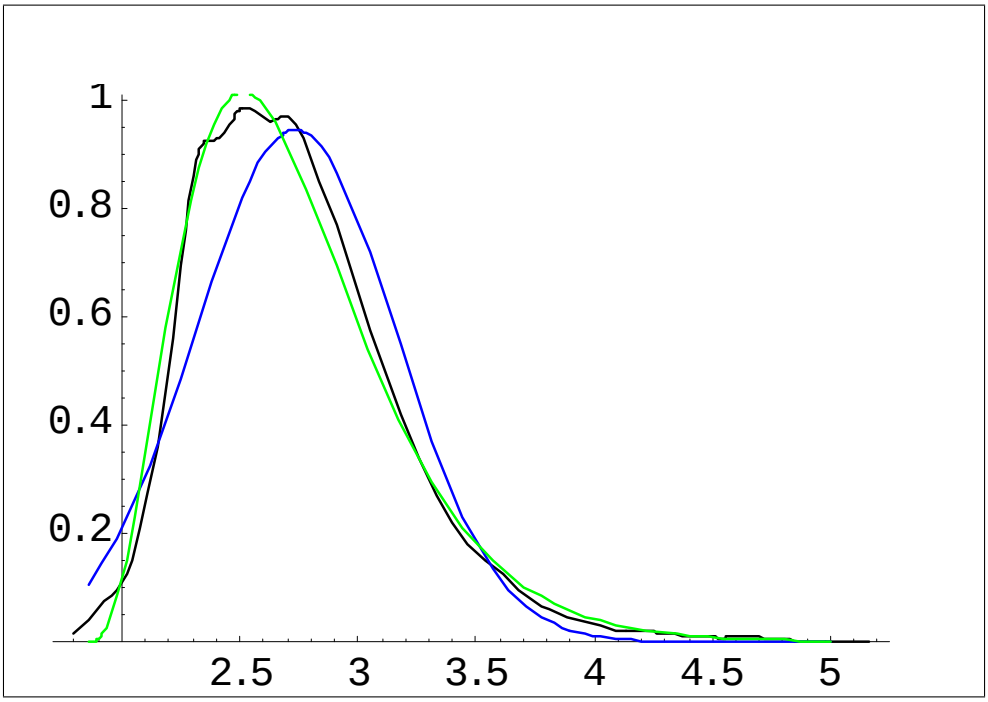

Figure 1: Kernel Smooth Distribution of Log Wages (black), with Normal (blue) and Shifted Gamma (green) fits

Table 1: Examples of Bivariate Copula Families

AMH [(a)]	
Copula C_θ	$\frac{xy}{1-\theta(1-x)(1-y)}$
Density c_θ	$[1+\theta\,(xy+x+y-2+\theta\,(1-x)\,(1-y))]\,[1-\theta\,(1-x)\,(1-y)]^{-3}$
Parameter θ	$-1 \leq \theta \leq 1$
Spearman's S_ρ [(b)]	$\frac{12(1+\theta)}{\theta^2}\,\mathrm{dilog}(1-\theta) - \frac{24(1-\theta)}{\theta^2}\log(1-\theta) - \frac{3(\theta+12)}{\theta}$
FGM [(a)]	
Copula C_θ	$xy(1+\theta(1-x)(1-y))$
Density c_θ	$1+\theta(1-2x)(1-2y)$
Parameter θ	$-1 \leq \theta \leq 1$
Spearman's S_ρ	$\frac{\theta}{3}$
Frank	
Copula C_θ	$-\theta^{-1}\log\left(1+\frac{(e^{-\theta x}-1)(e^{-\theta y}-1)}{e^{-\theta}-1}\right)$
Density c_θ	$\theta\left(1-e^{-\theta}\right)e^{-\theta(x+y)}\left[1-e^{-\theta}-(e^{-\theta x}-1)(e^{-\theta y}-1)\right]^{-2}$
Parameter θ	$-\infty < \theta < \infty$
Spearman's S_ρ [(c)]	$1-\frac{12}{\theta}(D_1(\theta)-D_2(\theta))$
Plackett	
Copula C_θ [(d)]	$\frac{1}{2(\theta-1)}\,(s-t)$
Density c_θ	$\theta\,(s-2xy(\theta-1))\,t^{-3}$
Parameter θ	$\theta > 0$
Spearman's S_ρ	$\frac{\theta+1}{\theta-1} - \frac{2\theta}{(\theta-1)^2}\log\theta$
Normal	
Copula C_θ	$\Phi_2(\Phi^{-1}(x),\Phi^{-1}(y);\theta)$
Density c_θ	$\frac{\phi_2(\Phi^{-1}(x),\Phi^{-1}(y);\theta)}{\phi(\Phi^{-1}(x))\;\phi(\Phi^{-1}(y))}$
Parameter θ	$-1 \leq \theta \leq 1$
Spearman's S_ρ	$\frac{6}{\pi}\arcsin\left(\frac{\theta}{2}\right)$

Notes: (a) AMH denotes Ali-Mikhail-Haq. FGM denotes Farlie-Gumbel-Morgenstern.

(b) $\mathrm{dilog}(z) = \int_1^z \log(t)\,(1-t)^{-1}\,dt$ is the dilogarithm function.

(c) the Debye function $D_k(z) = kz^{-k}\int_0^z t^k\left(e^t-1\right)^{-1}dt$, for k any positive integer.

(d) $s = 1+(\theta-1)(x+y)$ and $t = \sqrt{s^2-4xy\theta(\theta-1)}$.

Table 2: Sample Means of the Variables

	Employed (n=13438)		Unemployed (n=922)	Total
	Full time (n= 10374)	Part time (n= 3064)		
Variables	Mean	Mean	Mean	Mean
Wage (overall)	18.10	13.84		16.02
Wage (pakeha)	18.74	14.26		16.85
Wage (maori)	15.99	11.54		13.21
Wage (mixd)	16.29	13.29		13.54
Wage (paci)	13.95	11.97		12.49
Wage (other eth)	18.15	12.88		15.04
Wage (male)	18.99	13.55		17.27
Wage (female)	16.88	13.91		14.80
Age	38.98	36.11	33.32	38
% female	42.21	78.72	50.98	50.56
% married	66.13	53.07	37.96	61.53
% separated	8.60	10.41	11.50	9.17
% with uni degree	15.39	9.40	9.00	13.70
% with post-school qualification	44.58	31.92	34.27	41.21
% with school qualification	21.92	34.40	24.95	24.78
% immigrant 0-4 years	4.08	3.62	10.20	4.37
% immigrant 5-9 years	2.79	2.81	4.66	2.91
% immigrant 10-14 years	2.11	2.02	3.15	2.16
No. of school-age children	0.40	0.52	0.41	0.43
No. of children under 5 years	0.58	0.71	0.61	0.61
% in top two occupatioal groups	0.38	0.25		
% in middle five occupatioal groups	0.58	0.64		
% maori	9.24	8.09	18.43	9.59
% mixed maori	3.42	3.52	7.70	3.72
% pacific islander	5.55	2.90	6.94	5.08
% other ethnic groups	6.84	7.77	12.15	7.38
% main city resident	48.26	43.31	43.71	46.91

Table 3: Maximum Likelihood Estimates: Plackett Copula

	Participation		**Wage Earnings**	
	Plackett	Heckman	Plackett	Heckman
Constant	0.273	0.107	−1.616**	1.700**
Individual Demographics				
Age (years)	0.051**	0.063**	0.060**	0.040**
Age-squared (years/100)	−0.062**	−0.077**	−0.064**	−0.042**
Gender (female=1)	0.009	−0.010	−0.130**	−0.127**
No. school-age children	−0.105**	−0.099**		
No. children under 5 years	−0.077*	−0.109**		
Married or cohabit	0.527*	0.497**		
Divorced, widowed or separated	0.034	0.039		
Main city resident	0.053	0.070	0.088**	0.083**
In top two occupational groups			0.312**	0.285**
In middle five occupational groups			0.098**	0.082**
Employed part-time			−0.199**	−0.126**
Education (ref: no qualification)				
University degree	0.347**	0.389**	0.349**	0.340**
Post-school qualification	0.202**	0.228**	0.187**	0.145**
School leaving qualification	0.274**	0.286**	0.117**	0.097**
Ethnicity (ref: Pakeha/European)				
Maori	−0.414**	−0.438**	0.042	0.003
Mixed Maori	−0.463**	−0.465**	0.002	−0.020
Pacific	−0.081	−0.132	−0.118**	−0.110**
Other	−0.052	−0.078	−0.053*	−0.040
Immigrant status				
0-4 years in NZ	−0.377**	−0.407**	−0.031	−0.044*
5-9 years in NZ	−0.353*	−0.338*	0.079*	0.099**
10-14 years in NZ	−0.524**	−0.544**	0.063	0.074*
Ethnicity-Education Interactions				
Maori - University degree			−0.114*	−0.114**
Maori - Post-school qualification			−0.086**	−0.060*
Maori - School leaving qualification			−0.086**	−0.051
Mixed - University degree			0.062	0.110
Mixed - Post-school qualification			−0.013	−0.012
Mixed - School leaving qualification			−0.011	0.017
Pacific - University degree			0.015	−0.047
Pacific - Post-school qualification			−0.052	−0.059
Pacific - School leaving qualification			0.012	−0.004
Other - University degree			−0.004	−0.055
Other - Post-school qualification			0.063*	0.043
Other - School leaving qualification			−0.035	−0.026
Ethnic-Immigrant Interactions				
Pacific – 0-4 years in NZ	−0.497*	−0.470*	0.007	−0.020
Pacific – 5-9 years in NZ	−0.149	−0.034	−0.110	−0.119*
Pacific – 10-14 years in NZ	0.057	0.043	−0.121*	−0.131**
Other – 0-4 years in NZ	−0.472**	−0.468**	−0.042	−0.055
Other – 5-9 years in NZ	−0.269	−0.356*	−0.116*	−0.137**
Other – 10-14 years in NZ	0.572*	0.531*	−0.123*	−0.115*

Note: Amongst the covariates ** and * indicate significance at 1% and 5% levels respectively

Table 3 (Continued)

Other estimates and statistics		
	Plackett	Heckman
Gamma shape α	5.98^{**}	
Normal std deviation		0.334^{**}
Dependence θ	0.089^{**}	-0.220^{**}
Spearman's rho	-0.678^{**}	
Kendall's τ	-0.498^{**}	-0.141^{**}
Maximised log-likelihood	-6711.34	-7347.2