Luechinger, Simon; Stutzer, Alois; Winkelmann, Rainer

Working Paper

Self-Selection and Subjective Well-Being: Copula Models with an Application to Public and Private Sector Work

SOEPpapers on Multidisciplinary Panel Data Research, No. 135

Provided in Cooperation with:
German Institute for Economic Research (DIW Berlin)

Suggested Citation: Luechinger, Simon; Stutzer, Alois; Winkelmann, Rainer (2008) : Self-Selection and Subjective Well-Being: Copula Models with an Application to Public and Private Sector Work, SOEPpapers on Multidisciplinary Panel Data Research, No. 135, Deutsches Institut für Wirtschaftsforschung (DIW), Berlin

This Version is available at:
https://hdl.handle.net/10419/150683

Deutsches Institut für
Wirtschaftsforschung

DIW BERLIN

www.diw.de

SOEP The German Socio-Economic Panel Study

SOEPpapers

on Multidisciplinary Panel Data Research

135

Simon Luechinger • Alois Stutzer • Rainer Winkelmann

Self-selection and subjective well-being: Copula models with an application to public and private sector work

Berlin, October 2008

SOEPpapers on Multidisciplinary Panel Data Research
at DIW Berlin

This series presents research findings based either directly on data from the German Socio-Economic Panel Study (SOEP) or using SOEP data as part of an internationally comparable data set (e.g. CNEF, ECHP, LIS, LWS, CHER/PACO). SOEP is a truly multidisciplinary household panel study covering a wide range of social and behavioral sciences: economics, sociology, psychology, survey methodology, econometrics and applied statistics, educational science, political science, public health, behavioral genetics, demography, geography, and sport science.

The decision to publish a submission in SOEPpapers is made by a board of editors chosen by the DIW Berlin to represent the wide range of disciplines covered by SOEP. There is no external referee process and papers are either accepted or rejected without revision. Papers appear in this series as works in progress and may also appear elsewhere. They often represent preliminary studies and are circulated to encourage discussion. Citation of such a paper should account for its provisional character. A revised version may be requested from the author directly.

Any opinions expressed in this series are those of the author(s) and not those of DIW Berlin. Research disseminated by DIW Berlin may include views on public policy issues, but the institute itself takes no institutional policy positions.

The SOEPpapers are available at
http://www.diw.de/soeppapers

ISSN: 1864-6689 (online)

German Socio-Economic Panel Study (SOEP)
DIW Berlin
Mohrenstrasse 58
10117 Berlin, Germany

Contact: Uta Rahmann | urahmann@diw.de

Self-selection and subjective well-being: Copula models with an application to public and private sector work

SIMON LUECHINGER
University of Zurich

ALOIS STUTZER
University of Basle

RAINER WINKELMANN*
University of Zurich

February 2008

Abstract

We discuss a new approach to specifying and estimating ordered probit models with endogenous switching, or with binary endogenous regressor, based on copula functions. These models provide a framework of analysis for self-selection in economic well-being equations, where assigment of regressors may be choice based, resulting from well-being maximization, rather than random. In an application to public and private sector job satisfaction, and using data on male workers from the German Socio-Economic Panel, we find that a model based on Frank's copula is preferred over two alternative models with independence and normal copula, respectively. The results suggest that public sector workers are negatively selected.

JEL Classification: I31, C23

Keywords: Ordered probit, switching regression, Frank copula, job satisfaction, German Socio-Economic Panel.

* Address for correspondence: University of Zurich, Socioeconomic Institute, Zurichbergstr. 14, CH-8032 Zurich, Switzerland, phone: +41 (0)44 634 22 92, fax: +41 (0)44 634 49 96, email: sluechinger@iew.unizh.ch, alois.stutzer@unibas.ch and winkelmann@sts.uzh.ch

1 Introduction

This paper addresses a methodological shortcoming in the existing subjective well-being (aka happiness) literature, namely the failure to adequately account for self-selection. This recent literature studies the determinants of people's well-being (Frey and Stutzer, 2002). Some of these determinants are choice variables, which means that they are not randomly assigned. Presumably, people choose a sector of employment, decide to go to university or not, give up life as bachelor, etc. because they expect to be better off than in the alternative state. If the effect of these circumstances on people's well-being is to be measured, self-selection needs to be taken into account. In this paper, a possible solution, based on copulas, is offered, and both the problem of self-selection and the solution are illustrated in an application to the estimation of well-being differentials between workers in the public sector and workers in the private sector, using data from the German Socio Economic Panel. It is shown that ignoring self-selection can bring about grossly misleading inferences regarding sectoral well-being.

The fact that the self-selection problem has not yet been directly confronted by the empirical well-being literature may come as a surprise, given that methods to correct for self-selection in a regression context have been developed more than 30 years ago (Heckman, 1979, Gronau, 1974) and have been refined ever since (see Vella, 1988, for a survey). The early developments occurred in the area of labor economics, where simple regressions were found wanting, for instance, to estimate the determinants of women's potential wages (Heckman, 1974), of union and non-union wages (Lee, 1978) or of public and private sector wages (van der Gaag and Vijverberg, 1988, Dustmann and Van Soest, 1998). Self-selection models have been adopted in other areas of empirical economics as well, including health (Holly et al. 1998), and migration (Borjas, 1987), to name but a few.

In happiness research, self-selection arises naturally, since one can expect rational individuals to choose their life circumstances with a view of maximizing their happiness. This has to be recognized when attempting to estimate the effect of a choice variable on happiness. To fix ideas consider the choice between public and private sector employment, and its consequences for subjective well-

being. Let $U_i(1)$ be the subjective well-being (or job satisfaction) of a person while working in sector 1, the public sector. $U_i(0)$ is then the well-being of the same worker while working in sector 0, the private sector. The gain in well-being for that worker (of being in sector 1 rather than in sector 0) is $U_i(1)-U_i(0)$ which is inherently unobservable since, under the assumption of maximizing behavior, data can reveal only $U_i = \max[U_i(1), U_i(0)]$. If we consider population averages instead, the problem is that in the sample of sector 1 workers, we can identify $E[U_i(1)|U_i(1) > U_i(0)]$, but not $E[U_i(1)]$. In the sample of sector 0 workers, we can identify $E[U_i(0)|U_i(1) < U_i(0)]$, but not $E[U_i(0)]$. Ignoring this issue leads to biased inferences. For example, the coefficient of a sector 1 dummy variable in a regression model does not estimate the 'average sector gain' $E[U_i(1)]-E[U_i(0)]$. To overcome the problem, we need to introduce additional assumptions.

The suggestion of this paper is to address the self-selection issue in a general switching regression framework. Subjective well-being (the outcome variable), as elicited in single item survey questions, is a typical example for an ordered responses. With such ordered outcomes, it is natural to model the interdependence between outcome equations and selection equation using copula functions. The systematic use of copula functions in empirical economic research is a rather novel development (see, for example, van Ophem, 1999, Smith, 2005, Zimmer and Trivedi, 2006). Copula functions allow to generate joint distributions for two or more random variables with pre-specified marginal distributions in a very flexible manner. An excellent introduction to this method for empirical economists is provided by Trivedi and Zimmer (2007).

The copula approach can be implemented very easily and at low computational cost. Its main advantage is that it allows in a straightforward manner to incorporate departures from the standard trivariate normal assumption, an assumption that has often been used in prior research and equally often been criticized for being a choice of mere convenience lacking substantive justification. In this sense, our paper is a generalization of other recent implementations of switching regression models for ordered responses based on joint normality (DeVaro, 2006, Munkin and Trivedi, 2008). Within the copula approach, one can use a number of alternative joint distributions, and thus selection models, and thereby assess the sensitivity of the results to specific assumptions and es-

tablish robustness, without sacrificing the parsimonious parameterization and less demanding data requirements of a parametric model (as opposed to semi-parametric models).

The rest of the paper is organized as follows. The next section shows how copulas provide a natural framework for thinking about switching regression models for ordered responses. The general likelihood function is derived, and three specific cases are considered: independence copula, normal copula, and Frank's copula. Section 3 illustrated the proposed method in an application to job satisfaction of public and private sector workers. Tests show that the Frank copula dominates the other models in this application. Falsely ignoring such self-selection means that the effect of sector allocation on job satisfaction is underestimated. Section 4 concludes.

2 Modeling self-selection in well-being

2.1 A switching regression model of well-being

The topic of this paper is how to model the effect of a binary choice variable on subjective well-being. Consider two states, $s = 0, 1$, that are chosen by the individual rather than randomly assigned. We are interested in the well-being difference in the two states for a randomly selected member of the population, formally $E[U_i(1)] - E[U_i(0)]$. In parlance of treatment effect models, this is the "average treatment effect".

Formally, there is no major difference whether the outcome variable is earnings, as in much of the previous literature cited in the introduction, or "utility", proxied by some measure of subjective well-being. Hence, the well-established switching regression model is a natural starting point for any attempt to model the effect of a binary choice variable on subjective well-being. An adjustment is required since subjective well-being is usually measured on an ordered discrete scale, whereas the switching regression model in its standard form assumes continuous outcomes. We therefore formulate a switching regression model for latent well-being, which is then translated in a second step into observed outcomes by some threshold mechanism. In this spirit, let

$$y_0^* = x'\beta_0 + \varepsilon_0 \tag{1}$$

be the latent well-being index if $s = 0$, and

$$y_1^* = x'\beta_1 + \varepsilon_1 \tag{2}$$

be the latent well-being index if $s = 1$. x is a vector of explanatory variables that is the same in both equations, and β_0, β_1 are conformable sector-specific parameter vectors. We do not impose that $\beta_0 = \beta_1$, as the well-being returns to certain characteristics may be choice-specific. Individuals are observed either in state $s = 1$, or in state $s = 0$, but never in both. It is unreasonable to assume that individuals select themselves randomly into the two states. Rather, it is likely that there are idiosyncratic gains to well-being (one could call this in the current context "preference heterogeneity"), and that, for example, individuals who gain most from being in state 1 are actually the ones choosing $s = 1$ with highest probability. In its most extreme form, such (non-random) self-selection follows from the pure maximization hypothesis, whereby $s = 1$ whenever $y_1^* > y_0^*$, and $s = 0$ whenever $y_0^* \geq y_1^*$.

A less extreme proposition is obtained from a generalized selection rule, a third latent equation for the selection of states,

$$s^* = z'\gamma + \nu \tag{3}$$

and

$$s = \begin{cases} 1 & \text{if } s^* > 0 \\ 0 & \text{if else} \end{cases} \tag{4}$$

In this model, the absence of self-selection is equivalent to the statistical independence of ν and ε_0 and ε_1, respectively. The nature of self-selection correspondingly hinges on the joint distributions $f(\nu, \varepsilon_0)$ and $f(\nu, \varepsilon_1)$.

Regardless of how these two bivariate distributions are specified, the model needs to be adjusted to account for the discrete and ordinal scale of *observed* responses. In particular, we follow standard practice and assume a threshold mechanism. The ordered discrete responses $y_s = 0, \ldots, J$, i.e., people's judgments about their subjective well-being, are determined as

$$y_s = j \quad \text{if and only if} \quad \kappa_{s,j} < y_s^* \leq \kappa_{s,j+1} \tag{5}$$

where $s = 0, 1$, $j = 0, 1, \ldots, J$ are the observed ordered discrete responses, and the threshold values $\kappa_{s,j}$, form a partition of the real line i.e., $\kappa_0 = -\infty$, $\kappa_{11} = \infty$, and $\kappa_{s,j+1} > \kappa_{s,j}$ $\forall j$. This is not an standard ordered response model since the probability of observing $y_s = j$ depends on the outcome of the selection variable s, and s and y_s are not necessarily independent. We have

$$\begin{aligned} P(y_0 = j, s = 0|x, z) &= P(\kappa_{0,j+1} - x'\beta_0 < \varepsilon_0 \leq \kappa_{0,j} - x'\beta_0, \nu \leq -z'\gamma) \\ &= P(\varepsilon_0 < \kappa_{0,j+1} - x'\beta_0, \nu \leq -z'\gamma) - P(\varepsilon_0 < \kappa_{0,j} - x'\beta_0, \nu \leq -z'\gamma) \quad (6) \end{aligned}$$

$$\begin{aligned} P(y_1 = j, s = 1|x, z) &= P(\kappa_{1,j} - x'\beta_1 < \varepsilon_1 \leq \kappa_{1,j+1} - x'\beta_1, \nu > -z'\gamma) \\ &= P(\varepsilon_1 < \kappa_{1,j+1} - x'\beta_1, -\nu < z'\gamma) - P(\varepsilon_1 < \kappa_{1,j} - x'\beta_1, -\nu < z'\gamma) \quad (7) \end{aligned}$$

Under independence of ε_0, ε_1 and ν, the joint probabilities can be factored into their marginals, and one obtains univariate ordered and binary response models. The independence scenario provides a useful hint how the modeling of the joint distribution of the three stochastic terms should be approached. One can simply follow the lead of the empirical literature, where ordered probit (or logit) models are routinely employed for ordinal responses (see e.g. McKelvey and Zavoina, 1975), whereas the simple probit (or logit) model is applied to binary responses. The choice between probit and logit is inconsequential, and we select the probit as benchmark.

Arguably then, a natural starting point for an ordered response switching regression model is a class of models that preserves the probit structure at the marginal level. In other words, the joint distributions $f(\nu, \varepsilon_0)$, and $f(\nu, \varepsilon_1)$ should be such that all three error terms are normally distributed. This does not tell us much yet, since there are many joint distribution with normal marginals. The leading example is that of the bivariate normal (which is in fact a special case of a copula). The copula method provides a general approach to generate joint distribution functions for given marginals, and thus a way to specify many ordered probit models with endogenous switching in a unified framework. Copulas are formulated in terms of cumulative distribution functions (cdf), rather then joint densities, which is particularly appealing, since cdfs are needed to evaluate the joint probabilities in (6) and (7). A brief overview of the technique is given in the next section, before we return to the specific implementation of a model for well-being under self-selection.

2.2 Copula Functions

In statistics, a copula is a multivariate joint distribution function defined on the n-dimensional unit cube $[0,1]$ such that every marginal distribution is uniform on the interval $[0,1]$. For example, the normal, or Gaussia, copula, for $n = 2$, is

$$P(U \leq u, V \leq v) = C(u,v) = \Phi_2(\Phi^{-1}(u), \Phi^{-1}(v); \theta) \tag{8}$$

where Φ and Φ_2 are the uni- and bivariate cdf of the standard normal distribution, and θ is the coefficient of correlation. Two other examples are Clayton's copula

$$C(u,v) = (u^{-\theta} + v^{-\theta} - 1)^{-1/\theta}$$

and the Frank copula

$$C(u,v) = -\theta^{-1} \log \left\{ 1 + \frac{(e^{-\theta u} - 1)(e^{-\theta v} - 1)}{(e^{-\theta} - 1)} \right\} \tag{9}$$

The marginal distributions implied by bivariate copulas are

$$F(u) = P(U \leq u, V \leq 1) = C(u,1)$$

and

$$F(v) = P(U \leq 1, V \leq v) = C(1,v)$$

respectively. It is easy to verify that all three copulas have the key property that their marginal distributions are uniform, as $C(u,1) = u$ and $C(1,v) = v$.

The significance of copulas lies in the fact that by way of transformation, any joint distribution function can be expressed as a copula applied to the marginal distributions. This result is due to Sklar (1959). Sklar's theorem states that given a joint distribution function $F(y_1, \ldots, y_k)$, and respective marginal distribution functions, there exists a copula C such that the copula binds the margins to give the joint distribution.

For the bivariate case, Sklar's theorem can be stated as follows. For any bivariate distribution function $F(y_1, y_2)$, let $F_1(y_1) = F(y_1, \infty)$ and $F_2(y_2) = F(\infty, y_2)$ be the univariate marginal probability distribution functions. Then there exists a copula C such that

$$F(y_1, y_2) = C(F_1(y_1), F_2(y_2))$$

Moreover, if the marginal distributions are continuous, the copula function C is unique. We see, that the copula is now expressed as a function of cdf's. But a standard result in statistics states that cdf's are uniform distributed over the interval $[0, 1]$. Since the marginal distributions of a copula are uniformly distributed, it follows that the marginal distribution of $y_1 = F_1^{-1}(u)$ and $y_2 = F_2^{-1}(v)$ are F_1 and F_2, as stated.

The practical significance of copula functions in empirical modeling stems from the fact that they can be used to build new multivariate models for given univariate marginal component cdf's. If the bivariate cdf $F(y_1, y_2)$ is unknown, but the univariate marginal cdf's are of known form, then one can choose a copula function and thereby generate a representation of the unknown joint distribution function. The key is that this copula function introduces dependence, captured by additional parameter(s), between the two random variables (unless the independence copula $C(u, v) = uv$ is chosen). The degree and type of dependence depends on the choice of copula. There is a large literature on this topic (Trivedi and Zimmer, 2007). For our purposes, it is essential that the copula allows for positive *and* negative correlation, since we do not want to restrict the selection pattern *a priori*: we want to learn from the data whether individuals observed in state 1 are more, less or equally happy / satisfied in comparison to a randomly selected individual in that state ceteris paribus, i.e., for a given set of explanatory variables. The Clayton copula is unattractive for that reason, as it allows only positive dependence.

We therefore consider three copula functions in the following application, the normal copula, the Frank copula, and the independence copula $C(u, v) = uv$. In the normal case, $-1 \leq \theta \leq 1$, with -1 signifying perfect negative correlation, 0 signifying independence, and $+1$ signifying perfect positive correlation. Since copulas in general do not impose linear dependence structures,

correlation measures have only limited information value when moving away from the normal copula. There are a number of other indicators of a copula's ability to generate dependence (see Trivedi and Zimmer, 2007, for a detailed discussion). One is the question whether it can reach the Fréchet upper and lower bounds. The Fréchet upper bound for any bivariate distribution is given by $F_u(y_1, y_2) = \min[F_1(y_1), F_2(y_2)]$, where F_1 and F_2 are the marginal cdfs. $F(y_1, y_2) = F_u$ requires F to be the most positive dependent bivariate distribution in any possible sense. The lower bound is given by $F_l(y_1, y_2) = \max[0, F_1(y_1) + F_2(y_2) - 1]$, representing greatest possible negative dependence. Both normal and Frank copula can reach F_l and F_u, and thus span the full range of dependence. For the Frank copula, the dependence parameter may assume any real value. Values of $-\infty$, 0, and ∞ correspond to the Fréchet lower bound, independence, and the Fréchet upper bound, respectively. Like the normal copula, the Frank copula is symmetric in both tails.

2.3 Implementation

For any given copula, the two required joint probabilities, $P(y_0 = j, s = 0|x, z)$ and $P(y_1 = j, s = 1|x, z)$ in (6) and (7) are fully determined. The assumption of ordered probit and probit marginals requires that $\nu \sim Normal(0, 1)$, $\varepsilon_1 \sim Normal(0, 1)$, $\varepsilon_0 \sim Normal(0, 1)$, where the variances are normalized to unity for identification. Thus,

$$P(y_0 = j, s = 0|x, z) = C(\Phi(\kappa_{0,j+1} - x'\beta_0), \Phi(-z'\gamma), \theta_0) - C(\Phi(\kappa_{0,j} - x'\beta_0), \Phi(-z'\gamma), \theta_0) \quad (10)$$

and

$$\begin{aligned} P(y_1 = j, s = 1|x, z) = {} & C(\Phi(\kappa_{1,j+1} - x'\beta_1), 1, \theta_1) - C(\Phi(\kappa_{1,j} - x'\beta_1), 1, \theta_1) \\ & - C(\Phi(\kappa_{1,j+1} - x'\beta_1), \Phi(-z'\gamma), \theta_1) + C(\Phi(\kappa_{1,j} - x'\beta_1), \Phi(-z'\gamma), \theta_1) \quad (11) \end{aligned}$$

where $C(u, v)$ is either the normal copula (8), Frank's copula (9), or the independence copula. The parameters of the model, $\xi = (\kappa_0, \kappa_1, \beta_0, \beta_1, \gamma, \theta_0, \theta_1)'$, can be estimated by maximum likelihood

without much difficulty. Given an independent sample of observation tuples (y_i, s_i, x_i, z_i), the likelihood function is simply

$$L(\xi; y, s, x, z) = \prod_{i=1}^{n} P(y_s, s|x, z) \tag{12}$$

In our application, the log likelihood function was maximized using the MAXLIK routine in GAUSS with numerical first and second derivatives. No convergence problems were encountered. Under the assumptions of the model, the maximum likelihood estimator has the desirable large sample properties. The two specifications are non-nested and information criteria can be used to select among competing models. Alternatively, Vuong (1989) provides a framework for formal testing. Since the two models are overlapping, both including the independence copula as a special case, the two-step procedure should be applied.

The estimated ordered probit coefficients have the usual interpretation related to such models (see, for instance, Boes and Winkelmann, 2006). In particular, they can be used to compute marginal effects for a randomly selected person in the two states, net of selection bias. A comparison of the outcome distribution of a randomly selected person in the two states provides an estimate of the average treatment effect.

The dependence parameters θ_s inform about the direction of the selection bias. The null hypothesis of no self-selection implies that $\theta_s = 0$, an hypothesis that can be tested directly. If rejected, an interesting quantification of the selection effects can be obtained by comparing the outcome distribution of self-selected workers, for instance $p_{00} = P(y_0 = j|s = 0, x, z)$, with the *counterfactual* predicted distribution $p_{01} = P(y_0 = j|s = 1, x, z)$ of a person who chose state 1 but is (hypothetically) allocated to state 0. For instance, positive selection is defined as a situation where p_{00} lies to the right of p_{01}, in the sense that the probability of reporting high levels of well-being in state 0 is higher for persons who actually chose that state, relative to the others.

3 Application: Well-Being of Public and Private Sector Workers

In this section, the copula methodology is applied to a model of sectoral well-being in a sample of West German male workers. We distinguish between two sectors, the private sector and the public (or government) sector. Rather than studying life satisfaction, or other more broadly defined well-being indicators, we focus on a more natural and immediate concept in the study of employment related well-being, namely job satisfaction. By doing so, the influence of partially unobservable variation in circumstances in other domains of life is reduced, and more precise estimates can be expected. Technically, job satisfaction is an ordinal variable as all the other subjective well-being constructs, and the modeling considerations of the previous section fully apply.

The question of empirical interest in this application is whether sector specific job satisfaction and sector choice are jointly determined. If so, public (and private) sector workers are not representative of the entire population. As a consequence, estimating a model of public sector job satisfaction using public sector workers, or of private sector job satisfaction using private sector workers, does not recover the underlying population relationships. For instance, such sub-sample estimates would misrepresent the job satisfaction difference between the two sectors for an average worker. Specifically, we suspect selection based on comparative gain, whereby public sector workers are those who gain most from that type of work environment, whereas private sector workers are those whose preferences and values are better matched in private sector jobs.

The selection effects we are interested in are always conditional on other observed determinants. The general latent variable model was formulated in equations (1) and (2) as

$$y_s^* = x'\beta_s + \varepsilon_s \qquad s = 0, 1$$

where y_s^* is the latent job satisfaction index in the private ($s = 0$) and public ($s = 1$) sector, respectively, and x is a vector of explanatory variables that affects job satisfaction, similar to those found in related papers on the topic of job satisfaction (e.g. Clark, 1997). Details are given in the next section, where we describe the dataset drawn from the German Socio-Economic Panel, as well as the particular variables employed.

3.1 German Socio-Economic Panel

The data are extracted from the German Socio-Economic Panel, 2004. We base our illustration on that particular year because it includes a relatively rich menu of questions that are potentially related to a person's preference for public and private sector employment. These questions were not included in other years of the survey. The analysis is based on a sample of male workers in West Germany, between the ages of 25 and 60. Accounting for (relatively few) observations with missing values on any of the dependent or independent variables, the final sample comprises 4181 records.

Table 1 presents variable definitions and summary statistics. As mentioned earlier, we use job satisfaction, rather than overall satisfaction with life, as outcome variable. Originally, this variable is measured on a 0-10 scale. Because of low response frequencies in the 0-2 range, we combined them into a single outcome. The average job satisfaction in the sample is 5.1 on the 0 to 8 scale.

– – – – – – – – –

Table 1 about here

– – – – – – – – –

Among the standard socio economic controls, AGE, EDUCATION, MARRIED and HEALTH, only the last deserves additional comment as it is an "objective" measure of health, a caseness score generated from an eight item list of ailments (difficulties of climbing stairs, impairment in daily activities, job, or social contacts due to physical or emotional problems, strong pain).

In addition, we observe a number of indicators of attitudes towards risk, social responsibility and career orientation. In particular, survey participants were asked about the importance they place on the following three aspects of life: having a successful career; helping other people; being engaged in social and political activities. The importance questions are asked on a four point scale, with responses "unimportant / not very important / important / very important", and we define dummy variables taking the value 1 for outcomes "important" or "very important". The risk variable is also a self-assessment, measured on an 0-10 scale. Our conjecture is that career oriented

individuals and those willing to take higher risks are more likely to be found in the private sector, whereas individuals who put more importance on helping and public service tend to be matched to the public sector.

3.2 Results

The estimated coefficients for three job-satisfaction ordered probits are shown in Table 2. The three models on display use the independence copula, the normal copula, and the Frank copula, respectively. For each model, there are two columns. The first shows the estimated parameters for the public sector job satisfaction equation, while the second does the same for the private sector equation. The table doesn't report the parameters of the selection equation, nor the estimated threshold parameters, but these are available on request.

The coefficients can be interpreted in a number of ways. One is in terms of implied changes of event probabilities such as $\partial P(y_0 > 5|x)/\partial x_j$, or $\partial P(y_1 > 6|x)/\partial x_j$. Given the ordered probit structure, these changes depend on specific κ values as well as on the point in the covariate space x. If one defines a "typical" individual as one, where $P(y_s > j|x) = \pi_{js}$, and π_{js}, $j = 0, \ldots, 7$ is the unconditional complementary cumulative distribution function of y_s, then these marginal effects are simply $\partial P(y_s > j|x)/\partial x_j = w_{js}\beta_j$ with weights $w_{js} = \phi(\Phi^{-1}(\pi_{js}))$. For example, in the case of public sector job satisfaction, the weights can be computed as 0.067, 0.118, 0.161, 0.254, 0.329, 0.397, 0.308, and 0.150, respectively. Accordingly, $\partial P(y_1 > 6)/\partial \text{health} = 0.308 \times (-0.198) = -0.061$, based on the public sector health coefficient from the independence copula. Thus, a unit increase in the health caseness score is predicted to decrease the probability of a job satisfaction response of at least "7" by about 6 percentage points, for a typical worker.

– – – – – – – – –

Table 2 about here

– – – – – – – – –

An alternative, and perhaps simpler possibility for interpreting the coefficients is to look at

relative magnitudes, i.e. at trade-off ratios. For example, the estimated coefficient of being married tends to be of opposite sign and between 1/2 and 2/3 of the absolute value of the health coefficient. Thus, being married rather than single compensates (in terms of keeping the job satisfaction distribution unchanged) for a 1/2 to 2/3 point increase in the health caseness score, reflecting the substantial importance of health for job satisfaction. In either case, we find the differences in the effects across the three models, while existant, nevertheless not to be of major magnitude.

As typically found in the literature, job-satisfaction is u-shaped in age (ceteris paribus, controlling for health and other factors that typically also vary with age). Education has no effect on job satisfaction. Married workers have a higher job satisfaction than others, although the effect is statistically significant only in the private sector. Risk tolerant and career oriented workers have higher job satisfaction than others, but only if employed in the private sector. In the public sector, there is a positive effect of having an attitude towards helping others.

Both copula functions with dependence nest the independence copula, and are thus amenable to formal likelihood ratio tests. Based on such tests, we conclude that the independence copula cannot be rejected against the normal copula, but it is rejected when compared to the Frank copula (the p-value is 0.016). As a consequence, a direct comparison between the two dependence copulas clearly favors the Frank copula. More importantly, in this application, substantive conclusions regarding the presence of self-selection and correlated errors between selection and outcome equations depend on whether one uses the normal copula, the approach advocated in the previous literature (DeVaro, 2006, Munkin and Trivedi, 2008), or the Frank copula. No evidence for self-selection is found in the normal case; by contrast, the specification based on the Frank copula shows that accounting for self-selection leads to an improved model, and that workers in the public and private sector are not randomly drawn from the underlying population of all workers.

The nature of the selection process can be inferred from the estimates of θ_1 and θ_0. In the Frank model, both are negative, indicating a negative dependence between the error of selection into the public sector, and the errors in both latent outcome equations: public sector workers tend to be less satisfied than an average worker in either of the two sectors, but the effect is much larger

(and statistically significant only) in the private sector. The fact that the difference between the two dependence parameters is not statistically significant suggests that in this application, it is not comparative advantage that is at work (public sector workers would be much worse off in the private sector than in their sector of choice) but rather absolute advantage of private sector workers.

– – – – – – – – –

Table 3 about here

– – – – – – – – –

The consequences of ignoring such self-selection can be seen in Table 3, where predicted job satisfaction distributions in the two sectors are shown net of self-selection (i.e. for a randomly selected worker). The difference between the two distributions represent what one would commonly refer to as "average treatment effect". Under the independence assumption, job satisfaction is somewhat greater in the public sector than in the private sector. The difference is not very large, however. For example, the probability of a response greater than 6 is by 2.9 percentage points higher for the public sector. With self-selection, the gap increases to an estimated 19.9 percentage points. The reason for this discrepancy is that the independence model assumes that, for given x, public and private sector workers are alike, whereas the Frank model suggests that public sector workers are intrinsically less satisfied. Ignoring this heterogeneity leads to an underestimation of public sector job satisfaction. This bias is avoided in a model where heterogeneity, and correlation between selection and outcome, is taken into account.

The absence of strong sector differences in the selection of public sector workers suggests that the job satisfaction data may be sufficiently well described by a simpler model where $\varepsilon_0 = \varepsilon_1$, $\beta_0 = \beta_1$, and the outcome equation includes a single dummy variable for PUBLIC SECTOR, while we allow for dependence between ε and ν, again using Frank's copula. This is an instance of an ordered probit model with binary endogenous regressor. For the sample of 4818 male workers, and with the same explanatory variables as before, the log-likelihood value of this more restrictive model at the maximum is -9963.9, with an estimated θ parameter of -2.19 and standard error of

0.58. Two conclusions emanate: first, the ordered probit model with binary endogenous regressor and Frank copula cannot be rejected against the more general switching regression model; second, the nature of self-selection is person, rather than person-choice specific: workers with inherently lower job satisfaction tend to work in the public sector. Of course, ignoring such heterogeneity will then lead to an underestimation of the public-private sector job satisfaction differential, as shown above.

4 Conclusions

In this paper, we have proposed a new class of estimators for ordered probit models with self selection. The class has two main features. First, it preserves the marginal probit distributions for the ordered outcome and binary selection models, and thus generalizes the standard econometric methods for such variables that ignore self selection. Second, it accounts for the joint determination of outcome and selection in a simple, yet flexible parametric framework. Thus, implementation of these methods does not require any estimation and inferential methods beyond those of maximum likelihood.

Our proposal for generating a class of jointly determined models with probit marginals is based on copula functions. These functions generate joint distributions for multivariate random variables with predetermined marginals. Different copula functions induce alternative dependence structures. Flexibility arises since copulas are easily exchanged, and it therefore becomes feasible for a practitioner to empirically determine the best copula from a given set, and, perhaps equally important, to assess the robustness of key conclusions with respect to the choice of copula.

The new model was applied to an analysis of job satisfaction among public sector and private sector workers in Germany. The preferred estimates were based on the Frank copula function which allows for both negative and positive dependence. We found statistical evidence for self-selection, although the implied pattern was not one of idiosyncratic satisfaction gains from being in one sector, as hypothesized, but rather one of individual heterogeneity, with public sector workers manifesting

lower satisfaction overall. Relatedly, a formal comparison between an ordered probit model with binary endogenous variable and a fully-blown switching regression ordered probit model did not lead to a rejection of the former.

Although the methodological developments in the paper were motivated by a substantive issue related to well-being research, it is clear that they are applicable in other areas of empirical economics as well, whenever a joint model for an ordered outcome variable and a binary selection process is needed. Future research should pursue some obvious extensions of these methods, including an integration of additional copula functions beyond the three considered in this paper, and more general, multinomial selection mechanisms. Also, in well-being research, the endogeneity of choice variables should be taken more seriously. The methods proposed in this paper provide a framework of analysis.

References

Boes, S. and R. Winkelmann (2006) "Ordered Response Models", *Advances in Statistical Analysis*, 90(1), 165-180.

Borjas, G. (1987) "Self-Selection and the Earnings of Immigrants", *American Economic Review*, 77, 531-553.

Clark, A. (1997) "Job Satisfaction and Gender: Why are Women so Happy at Work?", *Labour Economics*, 4, 341-372.

DeVaro, Jed (2006) "Teams, Autonomy, and the Financial Performance of Firms", *Industrial Relations*, 45, 217-269.

Dustmann, C. and Van Soest, A. (1998) "Public and private sector wages of male workers in Germany", *European Economic Review*, 42, 1417-1441.

Frey, Bruno S. and Alois Stutzer (2002) "What Can Economists Learn from Happiness Research?", *Journal of Economic Literature* 40(2), 402-435.

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

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

Heckman, J. (1979) "Sample Selection as a Specification Error", *Econometrica*, 47, 153-161.

Holly, A., L. Gardiol, G. Domenighetti and B. Bisig, (1998) "An econometric model of health care utilization and health insurance in Switzerland", *European Economic Review*, 42, 513-522.

Lee, L.-F. (1978) "Unionism and Wage Rates: A Simultaneous Equations Model with Qualitative and Limited Dependent Variables", *International Economic Review*, 19, 415-433.

McKelvey, R.D. and Zavoina, W. (1975) "A statistical model for the analysis of ordinal level dependent variables", *Journal of Mathematical Sociology*, 4, 103-120.

Munkin, M.K. and Trivedi, P.K. (2008) "Bayesian Analysis of the Ordered Probit Model with Endogenous Selection", *Journal of Econometrics*, forthcoming.

Sklar, A. (1959) "Fonctions de répartition à n dimensions et leurs marges", *Publications de l'Institut de Statistique de L'Université de Paris*, 8, 229-231.

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

Trivedi, P.K and Zimmer, D.M. (2007), "Copula Modeling: An Introduction for Practitioners", *Foundations and Trends in Econometrics*, Volume 1, Issue 1.

Vella, F. (1998) "Estimating Models with Sample Selection Bias: A Survey." *Journal of Human Resources*, 33(1), 127-172.

van der Gaag, J. and Vijverberg, W.P.M. (1988) "A Switching Regression Model for Wage Determinants in the Public and Private Sectors of a Developing Country", *Review of Economics and Statistics*, 70, 244-252.

van Ophem, H. (1999) "A General Method To Estimate Correlated Discrete Random Variables", *Econometric Theory*, 15, 228-237.

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

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

Table 1: Definition of Variables

Variable	Definition	Mean
JOB SATISFACTION	Originally coded on a 0,1,...,10 scale. In the analysis, a transformed 0-8 scale is used, with outcomes 0,1,2 grouped together.	5.1
PUBLIC SECTOR	1 if current employment in public sector (includes civil service), else 0	0.22
AGE	Age, in years	42.9
EDUCATION	Years of formal schooling	12.8
MARRIED	1 if person is currently married, else 0	0.72
HEALTH	A caseness score* between 0 (perfect health) and 8 (poor health)	1.59
SUCCESS	Importance References ı ones career (very important/importa	0.75
HELP	Importance ɔr others (very important/importa	0.90
ENGAGEMENT	Importance (. engagement (very important/importa	0.30
RISK	Willingness to take risks (0 = "none"; 10 = "full")	5.17

* The caseness score is based the following eight indicators: Frequency (always/often/sometimes = 1) of strong physical pains; underachievement or limitations at work or during everyday tasks due to physical health problems; underachievement or limitations due to physical health problems; social limitations due to impaired health; affect of state of health (greatly/slightly=1) on climbing stairs; on other tiring everyday tasks.

Table 2. Self-Selection Ordered Probit Models of Sector-Specific Job Satisfaction (German Socio-Economic Panel 2004, N=4181)

	Independence copula		Normal copula		Frank copula	
Sector:	*Public*	*Private*	*Public*	*Private*	*Public*	*Private*
AGE$*10^{-1}$	-0.937*	-0.236	-0.961*	-0.172	-0.820*	-0.147
	(0.361)	(0.189)	(0.383)	(0.193)	(0.389)	(0.189)
AGE SQUARED$*10^{-2}$	1.155*	0.279	1.189*	0.187	1.004*	0.153
	(0.407)	(0.218)	(0.441)	(0.225)	(0.449)	(0.219)
EDUCATION$*10^{-1}$	0.044	0.026	0.084	-0.073	-0.060	-0.124
	(0.119)	(0.066)	(0.210)	(0.097)	(0.152)	(0.084)
MARRIED	0.094	0.130*	0.088	0.144*	0.096	0.148*
	(0.084)	(0.043)	(0.089)	(0.044)	(0.080)	(0.043)
HEALTH	-0.198*	-0.164*	-0.196*	-0.163*	-0.189*	-0.164*
	(0.017)	(0.009)	(0.019)	(0.009)	(0.022)	(0.009)
CAREER	0.057	0.086*	0.049	0.101*	0.082	0.112*
	(0.079)	(0.043)	(0.090)	(0.043)	(0.077)	(0.043)
HELP	0.405*	0.112	0.403*	0.111	0.381*	0.107
	(0.117)	(0.060)	(0.115)	(0.059)	(0.117)	(0.059)
ENGAGEMENT	0.023	0.055	0.041	0.010	-0.027	-0.003
	(0.073)	(0.041)	(0.111)	(0.051)	(0.080)	(0.046)
RISK	0.009	0.019*	0.007	0.023*	0.010	0.024*
	(0.016)	(0.008)	(0.019)	(0.008)	(0.016)	(0.008)
θ			0.077	-0.314	-1.885	-2.748*
			(0.349)	(0.212)	(1.661)	(1.062)
Log-Likelihood	-9'957.61		-9'956.43		-9'953.40	

Notes: Standard errors in parentheses; * indicates statistical significance at the 5% level; West-German men; not shown are the estimated threshold parameters, as well as the results for the probit selection equation. The selection equation includes two variables, German citizenship as well as father's empluyment in the public sector, in addition to those listed here.

Table 3. Average treatment effects with and without accounting for self-selection (predicted probabilities in percent)

	Independence copula		Frank copula	
Sector:	*Public*	*Private*	*Public*	*Private*
Job Satisfaction				
0	3.0	3.4	1.7	4.8
1	2.9	3.7	1.7	4.9
2	3.0	3.8	1.8	4.7
3	8.4	10.5	5.3	12.3
4	9.5	10.3	6.5	11.1
5	19.5	18.8	14.9	18.6
6	30.4	29.1	30.2	26.3
7	15.3	12.7	21.9	10.8
8	8.1	7.8	15.9	6.5

Note: The table shows the averaged (over all x) predicted outcome distribution for a randomly selected worker in the respective sectors.