Zimmer, David M.; Trivedi, Pravin K.; Li, Tong; Cameron, A. Colin

Working Paper

Modeling the Differences in Counted Outcomes using Bivariate Copula Models: with Application to Mismeasured Counts

Working Paper, No. 04-3

Provided in Cooperation with:
University of California Davis, Department of Economics

Modeling the Differences in Counted Outcomes using Bivariate Copula Models: with Application to Mismeasured Counts*

A. Colin Cameron†, Tong Li‡,
Pravin K. Trivedi‡, and David M. Zimmer‡

April 2004

Abstract

This paper makes three contributions. First, it uses copula functions to obtain a flexible bivariate parametric model for nonnegative integer-valued data (counts). Second, it recovers the distribution of the difference in the two counts from a specified bivariate count distribution. Third, the methods are applied to counts that are measured with error. Specifically we model the determinants of the difference between the self-reported number of doctor visits (measured with error) and true number of doctor visits (also available in the data used).

* : We are grateful to Co-editor Frank Windmeijer and two referees for their helpful comments that greatly improve the paper. Helpful comments and suggestions for improvements on earlier versions were also received from Douglas J. Miller and Murray Smith. However, we retain responsibility for any remaining errors.

†: Department of Economics, University of Califormia - Davis. ‡ : Department of Economics, Wylie Hall 105, Indiana University, Bloomington, IN 47405, U.S.A..

1 Introduction

This article provides a new method for studying the distribution of the difference between two nonnegative correlated counts, y_1 and y_2, whose marginal distributions $F_1(y_1)$ and $F_2(y_2)$ are parametrically specified. This topic is motivated by some data situations. In one of these y_1 and y_2 are two measurements, perhaps replicated, of the same outcome. One or both might be contaminated by measurement error, and one's interest is in studying the distribution of the difference. For example, y_1 may be an observed variable, y_2 may be the corresponding value from a cross-validation study, and $(y_1 - y_2)$ is the measurement error. A second data situation is one in which y_1 and y_2 are paired observations that are jointly distributed. They could be data on twins, spouses, or paired organs (kidneys, lungs, eyes). The interest lies in studying and modeling the difference. For example, one may want to analyze the sources of differential utilization of health care, e.g. doctor visits, by two spouses. Another example from empirical industrial organization involves the difference between the number of firms entering and exiting an industry (Mayer and Chappell, 1992; Berglund and Brännäs, 2001).

When the bivariate distribution of (y_1, y_2) is known, standard methods can be used to derive the distribution of any continuous function of the variables, say $H(y_1, y_2)$. Indeed, there is a rich statistical literature that deals with this class of problems that includes the distribution of sums of independent random variables. A problem arises, however, when the bivariate distribution is either not available or available in an explicit form only under some restrictive assumptions. This situation arises in the case of many nonnormal discrete random variables. For example, most specifications of bivariate Poisson and Binomial distributions only admit positive dependence between counts, thus lacking generality. On the other hand, as in the case of entries and exits from an industry, dependence between the two variables may be positive or negative.

We propose a solution based on copula functions. Copulas, originally introduced by

Sklar in a 1959 article in French (see also Sklar, 1973),[1] have been suggested as a useful method for deriving joint distributions given the marginals, especially when one wants to work with nonnormal distributions. The approach is likely to be fruitful when the marginals can be specified with confidence, but the joint distribution is awkward to establish. The approach, though not new, has recently attracted considerable attention (Genest and Rivest, 1993; Joe, 1997; Nelsen, 1999; Capéraà, Fougères and Genest, 2000). To date several published articles (Miller and Liu, 2002; Smith, 2003) and working papers (Chen and Fan, 2002) using copulas in econometrics have focused mainly on continuous variables. Several econometrics papers have modeled sample selection using bivariate latent variable distributions that can be interpreted as specific examples of copula functions - Lee (1983), Prieger (2002) and van Ophem (1999, 2000). Other approaches for modeling correlated count variables, without explicitly using copulas, are developed in Cameron and Trivedi (1998), Munkin and Trivedi (1999), and Chib and Winkelmann (2001). The copula approach used in this paper, although relatively unexplored in applied statistics and econometrics, can be used to study the joint distributions of any set of discrete, continuous, or mixed discrete/continuous variables. The approach allows us to estimate the parameters of a bivariate distribution based on specific families of copulas. These estimates are used to recover the empirical cdf and/or the pmf of the difference, $y_1 - y_2$. The proposed method will generalize to continuous random variables, with or without dependence.

We carry out a case study using health care utilization data from Australia. In the empirical application, y_1 represents an individual's number of self-reported physician visits, and y_2 denotes his number of actual physician visits. Using a unique Australian data set that has both self-reported and independently observed measures of physician visits, we study the difference between the two measures to determine sources of misreporting. Results indicate a relationship between the number of visits and the extent of misreporting.

[1] Sklar (1996) clarifies in a brief note the contributions made by others such as Schweizer and Fréchet to the development of copulas. We owe this reference to a referee.

We measure the effect of key regressors on the difference in counts.

The remainder of the paper is organized as follows. Section 2 sketches the essentials of the copula-based approach, the research problem of interest, and our solution method. Section 3 briefly discusses other methods of obtaining joint distributions of count variables. Section 4 deals with an application that involves the distribution of measurement errors in recorded number of physician visits using an Australian data set. The fit of the copula models is also discussed in Section 4. Section 5 gives some concluding remarks.

2 The Copula Approach

In order that this paper should be reasonably self-contained, we begin by reviewing some basic properties of copulas.[2]

2.1 Properties of Copulas

To define a copula we begin with possibly dependent uniform random variables $U_1, ..., U_q$ on the $[0, 1]$-interval. The dependence relationship is described through their joint cdf

$$C(u_1, ..., u_q) = \Pr[U_1 \leq u_1, ..., U_q \leq u_q], \tag{1}$$

where the function $C(\cdot, \cdots, \cdot)$ is the copula, and u_j is a particular realization of U_j, $j = 1, ..., q$, where $q \geq 2$. Note that for a function $C(\cdot, \cdots, \cdot)$ to be a copula on $[0, 1]^q$, it must have the properties: its domain is $[0, 1]^q$; it is grounded, and increasing on the unit hypercube (see Nelsen (1999)).

Now for q marginal cdfs $F_1(\cdot), ..., F_q(\cdot)$ and arbitrary $(x_1, ..., x_q)$, we have from (1)

$$\begin{aligned} C(F_1(x_1), ..., F_q(x_q)) &= \Pr[F_1^{-1}(U_1) \leq x_1, \ldots, F_q^{-1}(U_q) \leq x_q] \\ &\equiv F(X_1, \cdots, X_q), \end{aligned} \tag{2}$$

[2] An excellent review of the copula literature is provided by Frees and Valdez (1998).

where $X_j = F_j^{-1}(U_j)$, $j = 1, \ldots, q$. Therefore, $F(\cdot, \cdots, \cdot)$ defines a joint cdf for the q variables $X_1, \ldots, X_q$. With a copula-based construction of a joint cdf, we select a set of marginals and combine them to generate a joint cdf. A given copula is a functional form for combining selected marginals. Sklar's Theorem states that for any multivariate cdf, there exists a copula function such that this cdf can be represented as a function of its marginal cdfs through this copula. Also, if this multivariate cdf is continuous, then the copula representation is unique. It is worth noting that for a joint distribution of multivariate discrete random variables, the associated copula representation is not unique. Such a non-uniqueness arises from the fact that a cdf of a discrete random variable does not map such a variable to the entire $[0, 1]$ interval, and thus the copula C need not be uniform over rectangles. See Joe (1997, p.14) for a detailed discussion on this issue. This result, however, does not create a serious problem from a modeling viewpoint, as while a copula is not unique for a joint distribution of discrete variables with marginals $F_k(\cdot)$, $k = 1, \ldots, q$, it is unique on $\prod_{k=1}^{q} Ran(F_k)$, where $Ran(F_k)$ denotes the range of the marginal distribution $F_k(\cdot)$ consisting of all the possible values of $F_k(\cdot)$ (see Nelsen, 1999, p.15).

2.2 Bivariate Copula Representation

For the bivariate case, suppose $F(y_1, y_2)$ is a joint distribution with corresponding marginal distributions $F_1(y_1)$ and $F_2(y_2)$. Then $F(y_1, y_2)$ can be expressed as

$$F(y_1, y_2) = C\left(F_1(y_1), F_2(y_2); \theta\right) \tag{3}$$

where C is a parametric copula function, and θ is a dependence parameter measuring dependence between the two random variables. The properties of copulas make them attractive for many empirical applications. A researcher might not know the joint distribution of two variables, or the joint distribution might be intractable, but if the marginal distributions are known and take a convenient form, then the copula approach provides a representation of the joint distribution.

Joe (1997) defines a bivariate copula associated with $F(\cdot)$, denoted by $C(u,v)$, as a two-dimensional probability distribution function defined on the unit square $[0,1]^2$, with univariate marginals uniform on $[0,1]$. For all $(u,v) \in [0,1]^2$, $C(u,0) = C(0,v) = 0$; $C(u,1) = u$, and $C(1,v) = v$. In this notation Sklar's Theorem states that, there exists a copula function C such that

$$F(x,y) = C(F_x(x), F_y(y)), \tag{4}$$

where $F(x,y) = \Pr[X \leq x, Y \leq y]$ is a bivariate distribution function of random variables X, Y, and $F_x(x)$ and $F_y(y)$ denote the marginal distribution functions.

If F is continuous, and if the univariate margins have corresponding quantile functions F_x^{-1} and F_y^{-1}, then the unique copula in equation (2) can be expressed as

$$C(u_1,u_2) = F\left(F_x^{-1}(u_1), F_y^{-1}(u_2)\right). \tag{5}$$

If F is discrete, then (5) gives a unique copula representation for F for $(u_1,u_2) \in Ran(F_x) \times Ran(F_y)$. The copula approach involves specifying marginal distributions of each random variable along with a function (copula) that binds them together. The copula function can be parameterized to include measures of dependence between the marginal distributions. If no dependence is detected, the two marginals are independent, and estimation can be performed on each variable separately. However, if dependence is present, improved estimates may be obtained by recovering a joint distribution by way of a copula function. Since a copula can capture dependence structures regardless of the form of the margins, a copula approach to modeling related variables is flexible and potentially very useful to statisticians.

The table below gives examples of some bivariate copula functions that have been used in the literature. Here ϕ and Φ denote the normal density and cdf respectively, and η equals $1 - e^{-\theta}$. Joe (1997) discusses the properties of these copulas.

Copula type	**Function** $C(u,v)$	**Dependence**
Product	uv	N.A.
Frank	$-\theta^{-1}\log\left((\eta-(1-e^{-\theta u})(1-e^{-\theta v}))/\eta\right)$	$-\infty<\theta<\infty$
Normal	$\Phi_B\left[\Phi^{-1}(u)\,\Phi^{-1}(v);\theta\right]$	$-1\leq\theta\leq+1$
Kimeldorf and Sampson	$\left(u^{-\theta}+v^{-\theta}-1\right)^{-1/\theta}$	$-1\leq\theta<0$ *or* $0<\theta<\infty$

The dependence parameter θ is not always easy to interpret because the relationship between the dependence parameters and familiar measures of association such as Spearman's 'rho' may not be transparent. Indeed most copulas do not require that $\theta\in[-1\,,\,1]$. Typically, when y_1 and y_2 are continuous variables, θ is converted into Kendall's 'tau' or Spearman's 'rho' which are both bounded on the interval $[-1\,,\,1]$. See Bouye et al. (2000) for a discussion of how to convert dependence parameters into Kendall's 'tau' and Spearman's 'rho'. However, when y_1 and y_2 are discrete, Marshall (1996) and Tajar et al. (2001) explain that Kendall's 'tau' and Spearman's 'rho' depend on the choice of marginal distributions, and, thus, they are not useful measures of dependence. The implication is that one must use caution when interpreting dependence parameters of copulas for discrete variables. Since our empirical applications consider discrete count variables, we do not use Kendall's 'tau' or Spearman's 'rho'.

2.3 Parametric Families of Copulas

Like all multivariate distribution functions, bivariate copulas must obey the Fréchet-Hoeffding lower and upper bounds, C^- and C^+ , defined as

$$C^-(u_1,u_2) = \max(u_1+u_2-1,0) \tag{6}$$

$$C^+(u_1,u_2) = \min(u_1,u_2), \tag{7}$$

for $(u_1,u_2)\in[0,1]^2$. Thus, by Sklar's theorem, for a joint cdf $F(\cdot,\cdot)$ of (y_1,y_2) with marginal distributions $F_1(\cdot)$ and $F_2(\cdot)$, respectively, we have the corresponding Fréchet-Hoeffding bounds as follows

$$\max(F_1(y_1)+F_2(y_2)-1,0)\leq F(y_1,y_2)\leq\min(F_1(y_1),F_2(y_2)).$$

Fréchet-Hoeffding bounds are important for interpreting dependence parameters θ. A desirable feature of a copula is that as θ approaches the lower (upper) bound of the permissible range, the copula corresponds to the lower (upper) Fréchet-Hoeffding bound. However, the parametric forms of some copulas place restrictions on the dependence structure such that one or both Fréchet-Hoeffding bounds are not included in the permissible range.

In preliminary analysis, we considered three different copulas: the Normal copula and the Frank copula include both Fréchet-Hoeffding bounds in their permissible ranges while the Kimeldorf and Sampson copula only includes the Fréchet-Hoeffding upper bound. The latter two are members of the Archimedean family, with the representation $C(u,v) = \xi\left(\xi^{-1}(u) + \xi^{-1}(v)\right)$ where ξ is a generator function; see Smith (2003) for an extensive discussion of this copula class as well as several generator functions. For a more extensive list of families of copulas, see Hutchinson and Lai (1990).

The question of comparing and selecting from a family of copulas is at present an open one. The Frank copula provides the best fit in terms of information criteria, so we focus on results for the Frank specification. However, results for the other two copulas were nearly identical.

Frank's copula (1979) is $C(u,v;\theta) = -\theta^{-1}\log\left((\eta - (1-e^{-\theta u})(1-e^{-\theta v}))/\eta\right)$, where $\eta = 1 - e^{-\theta}$. The dependence parameter θ can equal any value on the real domain $(-\infty\,,\,\infty)$ except zero. Values of $-\infty$, 0, and ∞ correspond to the Fréchet-Hoeffding lower bound, independence, and the Fréchet-Hoeffding upper bound. This copula permits both positive and negative association between the variables.

2.4 Modeling Differences in Counts

In this paper, we use the copula approach to represent $F(y_1, y_2)$, which will also allow us to derive the distribution of $y_1 - y_2$, where both y_1 and y_2 are nonnegative integer counts. Although to the best of our knowledge, no existing copula article attempts to model distributions of differences between variables, such an application is in principle

straightforward. If the joint distribution $F(y_1, y_2)$ is known, then standard methods can be used to derive the distribution of $y_1 - y_2$. However, no explicit form of bivariate count distribution with flexible dependence structure is available. There are attempts in the literature to develop a bivariate count distribution, but they suffer from shortcomings. Kocherlakota and Kocherlakota's (1992) trivariate reduction method and Marshall and Olkin's (1990) mixture method both restrict dependence between y_1 and y_2 to be positive. Gourieroux, Monfort and Trognon's (1984) moment based method ignores the integer value nature of the counts. Munkin and Trivedi (1999) and Chib and Winkelmann (2001) propose a bivariate count model with flexible dependence, but their method requires approximating integrals using either Gauss-Hermite approximation or simulations.

The copula representation is used to model a joint bivariate distribution. The key is to recognize that the copula representation $C\left(F_1(y_1), F_2(y_2); \theta\right)$, or equivalently $C\left(u, v; \theta\right)$, can be used in place of the unknown joint cdf $F(y_1, y_2)$. In the case of two continuous random variables, the joint density is obtained from $\partial^2 C / \partial u \partial v$, denoted $c_{12}\left(\cdot\right)$. In the case of discrete random variables, the continuous derivatives are replaced by finite differences, as shown below.

Suppose, for the case of discrete random variables, the variable of interest is the difference $z = y_1 - y_2$. We present a simple approach using copulas to derive the distribution of z.

The joint probability mass function (pmf) is derived by taking finite differences:

$$\begin{aligned} c_{12}\left(F_1(y_1), F_2(y_2); \theta\right) = & \, C\left(F_1(y_1), F_2(y_2); \theta\right) - C\left(F_1(y_1 - 1), F_2(y_2); \theta\right) \\ & - C\left(F_1(y_1), F_2(y_2 - 1); \theta\right) + C\left(F_1(y_1 - 1), F_2(y_2 - 1); \theta\right), \end{aligned} \quad (8)$$

where lower-case "c" denotes the pmf.

With the transformation $z = y_1 - y_2$, the joint pmf can be equivalently expressed in terms of z and y_2 as,

$$c_{12}\left(F_1(z + y_2), F_2(y_2); \theta\right). \quad (9)$$

The pmf of z, denoted $g(z)$, is obtained by summing over all possible values of y_2,

$$g(z) = \sum_{y_2=0}^{\infty} c_{12} \left(F_1(z + y_2), F_2(y_2); \theta \right). \tag{10}$$

For any value of z, (10) gives the corresponding probability mass. The cdf of $g(z)$ is calculated by accumulating masses at each point z,

$$G(z) = \sum_{k=-\infty}^{z} g(k). \tag{11}$$

Both $g(z)$ and $G(z)$ characterize the full distribution of z so that inference can be made regarding the difference between two count variables. This method can also be applied to any discrete or continuous variables when the marginal distribution of the components of the differences is parametrically specified.

2.5 Estimation

The first step in the copula approach is to specify the marginal distributions. In our applications, y_1 and y_2 are nonnegative integer counts, so we specify $F_1(y_1)$ and $F_2(y_2)$ as cdfs of the negative binomial-2 distribution (NB2). This specification has been found to provide a flexible specification of a count regression in many different alternative situations.[3] Each marginal is specified conditional on vectors of exogenous covariates $\mathbf{X}_1$ and $\mathbf{X}_2$ with corresponding parameter vectors $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$. For each observation $i = 1, ..., N$, $F_1(y_{1i}|\mathbf{X}_{1i}, \boldsymbol{\beta}_1)$ and $F_2(y_{2i}|\mathbf{X}_{2i}, \boldsymbol{\beta}_2)$ are

$$F_j(y_{ji}|\mathbf{X}_{ji}, \boldsymbol{\beta}_j) = \sum_{k=0}^{y_{ji}} \frac{\Gamma(k + \psi_j)}{\Gamma(\psi_j)\Gamma(k+1)} \left(\frac{\psi_j}{\lambda_{ji} + \psi_j} \right)^{\psi_j} \left(\frac{\lambda_{ji}}{\lambda_{ji} + \psi_j} \right)^{k}, \tag{12}$$

[3] As pointed out by a referee, the copula approach requires specification of both the marginal distributions and the copula function. For our data the negative binomial provides a good model. It allows for the large overdispersion in the doctor visit data (the sample variances are roughly five times the sample mean). And there is no excess zeros problem for our data, since for both self-reported and actual numbers of doctor visits, there are only about 10% of individuals who have reported or had zero doctor visit, respectively. The methods developed here can, of course, be adapted to alternative models for the marginals.

for $j = 1, 2$, where $\lambda_{ji} = \exp(\mathbf{X}'_{ji}\boldsymbol{\beta}_j)$ is the conditional mean, and $\psi_j = 1/\alpha_j$, ($\alpha_j > 0$) is the overdispersion parameter in the conditional variance $\lambda_{ji}(1 + \psi_j\lambda_{ji})$.

Once the marginal distributions are specified, an appropriate copula function C is selected; in this paper, we use the Frank copula. Then

$$C\left(F_1(y_{1i}|\mathbf{X}_{1i}, \boldsymbol{\beta}_1), F_2(y_{2i}|\mathbf{X}_{2i}, \boldsymbol{\beta}_2); \theta\right)$$

provides a representation of the unknown joint distribution $F(y_{1i}, y_{2i}|\mathbf{X}_{1i}, \mathbf{X}_{2i}, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$.

The joint pmf is formed by taking differences as shown in equation (8). The log-likelihood function is formed by taking the logarithm of the pmf and summing over all observations. The log-likelihood is maximized using a quasi-Newton iterative algorithm requiring only first derivatives. Post-convergence, the variances of the estimates are obtained using the robust 'sandwich' formula.

Maximization of the log-likelihood using variants of the Newton-Raphson procedure that we used were found to be straightforward and computationally efficient even for a high-dimensional parameter space. However, establishing error bounds for some quantities of interest involves approximation that are discussed in the next section.

3 Other Approaches

It is useful to compare the results from a copula-based model with other methods of generating joint distributions. Therefore, we present results from two other similar approaches. One such model is the Marshall-Olkin bivariate negative binomial with marginals that are univariate negative binomial, generated as a "shared-frailty model", defined as

$$f(y_1, y_2|\lambda_1, \lambda_2) = \frac{\Gamma(y_1 + y_2 + \alpha)}{y_1!y_2!\Gamma(\alpha)} \left(\frac{\lambda_1}{\lambda_1 + \lambda_2 + 1}\right)^{y_1} \left(\frac{\lambda_2}{\lambda_1 + \lambda_2 + 1}\right)^{y_2} \left(\frac{1}{\lambda_1 + \lambda_2 + 1}\right)^{\alpha}, \tag{13}$$

where $\lambda_1, \lambda_2, \alpha$ are, respectively, the two univariate means and the overdispersion parameter.

Like the copula approach, the Marshall-Olkin model provides a closed-form likelihood function that is easily estimated. However, this approach also has several disadvantages. First, it only applies to applications where both marginals are negative binomials, whereas the copula approach accomodates any combination of marginal distributions. Second, it restricts heterogeneity to the identical component α for both count variables. Third, the correlation between the two count variables,

$$Corr(y_{1,}y_2) = \frac{\lambda_1\lambda_2}{\sqrt{(\lambda_1^2 + \alpha\lambda_1)(\lambda_2^2 + \alpha\lambda_2)}}, \tag{14}$$

must be positive. In our application, correlation between the number of self-reported visits and the number of actual visits is likely to be positive, but for many applications, such as entry and exit of firms into an industry, the assumption of positive dependence might not be plausible.

A second approach based on unobserved heterogeneity is presented by Munkin and Trivedi (1999). They assume that y_1 and y_2 are correlated even after controlling for $\mathbf{X}_1$ and $\mathbf{X}_2$ because of a common unobserved hetergeneity component w, and y_1 and y_2 are independent after controlling for $\mathbf{X}_1$, $\mathbf{X}_2$, and w. Therefore, they model y_1 and y_2 separately as Poisson with conditional means $\lambda_{1i} = \exp(\mathbf{X}'_{1i}\boldsymbol{\beta}_1 + \gamma_1 w_i)$ and $\lambda_{2i} = \exp(\mathbf{X}'_{2i}\boldsymbol{\beta}_2 + \gamma_2 w_i)$. Then the joint distribution of y_1 and y_2 is simply the product of the two independent marginal distributions.

Since we do not observe w, we draw a pseudo random number from an assumed standard normal distribution and calculate the joint distribution as the product of the two marginals. We repeat this exercise 400 times to obtain a simulated likelihood function, which is then estimated in the usual way. This approach is referred to as unobserved heterogeneity (UH). Identification requires that either γ_1 or γ_2 be normalized to unity, as without such a restriction on the factor loading parameters, one cannot identify the scales. We set $\gamma_2 = 1$, the marginal corresponding to actual doctor visits, so that γ_1 is a general measure of correlation between the two measures of utilization. Results are very similar if we instead set $\gamma_1 = 1$.

Unlike the Marshall-Olkin model, the UH approach allows flexibility in choosing functional forms for the marginals, and in the choice of distribution of w. Moreover, the method allows for positive and negative correlation, as measured by the variable γ_1, between the outcome variables. The main disadvantage of the UH approach is that the numerical integration can be extremely time consuming, especially for large models or large datasets.

Despite disadvantages associated with the Marshall-Olkin and UH methods, they are helpful in comparing the performance of the copula approach. Therefore, results from these two models are presented below along with results from the Frank copula.

4 Application: Measurement Error in Self-reported Counts

The maximum likelihood estimation procedure described in Section 2.5 produces parameter estimates ($\hat{\boldsymbol{\beta}}_1$, $\hat{\boldsymbol{\beta}}_2$, $\hat{\theta}$) and corresponding covariance matrix $\hat{\Omega}$. Substituting the estimated parameters into the joint pmf of y_1 and y_2 yields

$$\hat{c}_{12}\left(F_1(y_1|\bar{\mathbf{X}}_1,\hat{\boldsymbol{\beta}}_1),F_2(y_2|\bar{\mathbf{X}}_2,\hat{\boldsymbol{\beta}}_2);\hat{\theta}\right) \tag{15}$$

where the covariates $\mathbf{X}_1$ and $\mathbf{X}_2$ have been set to their mean values.[4] Following the technique presented in section 2.4, the transformation $z = y_1 - y_2$ gives the expression

$$\hat{c}_{12}\left(F_1(z+y_2|\bar{\mathbf{X}}_1,\hat{\boldsymbol{\beta}}_1),F_2(y_2|\bar{\mathbf{X}}_2,\hat{\boldsymbol{\beta}}_2);\hat{\theta}\right). \tag{16}$$

The estimated density of z is obtained by summing over all possible values of y_2 in (9) as follows

$$\hat{g}(z)=\sum_{y_2=0}^{\infty}\hat{c}_{12}\left(F_1(z+y_2|\bar{\mathbf{X}}_1,\hat{\boldsymbol{\beta}}_1),F_2(y_2|\bar{\mathbf{X}}_2,\hat{\boldsymbol{\beta}}_2);\hat{\theta}\right). \tag{17}$$

[4] In principle, one could set the covariates to any values of interest. In the empirical applications, we set all covariates to their means while adjusting one variable at a time to determine its impact on $y_1 - y_2$.

For the empirical applications in this paper, the area of interest for $\hat{g}(z)$ lies mostly in the region $z \in [-8\,,\,8]$, but for other applications, a different range might be more appropriate. The calculation in expression (17) requires summation from zero to infinity, which is in practice replaced by a truncated sum in which the upper bound of the summation can be a sufficiently large finite value. For count variables, as the count approaches infinity, the probability mass approaches zero. Very large counts are associated with a pmf that is close to zero. For the applications in this paper, we calculated the summation from zero to 50. Values at 50 are indistinguishable from zero. Different values for the upper bound of the summation, provided they are large enough, did not affect the results. The estimated cdf $\hat{G}(z)$ is calculated by accumulating masses as in expression (11). Alternatively, one can rewrite (17) as an expectation using importance sampling techniques, and approximate this expectation using the random draws from the importance function.

Error bounds on $\hat{g}(z)$ and $\hat{G}(z)$ are obtained by a Monte Carlo technique based on the asymptotic normal distribution of $(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2, \hat{\theta})'$. Simulated parameters $\left(\tilde{\boldsymbol{\beta}}_1, \tilde{\boldsymbol{\beta}}_2, \tilde{\theta}\right)'$ are drawn from $N\left((\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2, \hat{\theta})', \hat{\boldsymbol{\Omega}}\right)$, and $\hat{g}(z)$ and $\hat{G}(z)$ are recalculated using $\left(\tilde{\boldsymbol{\beta}}_1, \tilde{\boldsymbol{\beta}}_2, \tilde{\theta}\right)'$. This is repeated for 500 replications. The 2.5 and 97.5 percentiles of the 500 replications provide error bounds.

4.1 NCEPH Data

As an illustration of our approach, we study the distribution of measurement errors using data from the Record Linkage Study conducted by the National Centre of Epidemiology and Population Health (NCEPH) at the Australian National University. The data set is unique in that it contains both self-reported physician visits and actual physician visits as recorded by the Health Insurance Commission, although it is probably not nationally representative. We treat the latter as a cross-validation sample, as it is commonly believed that the number of physician visits recorded by the Commission is accurate. The availability of both mis-reported and accurate number of visits in this data set makes

the data interesting in studying the measurement error in counted outcomes.[5] However, analyzing the measurement error problem in these data is complicated by the lack of a statistical model for the difference of counts, which is a discrete variable taking on integer values that can be negative, zero, or positive. Our approach is indeed motivated by such a difficulty in modeling and provides a useful solution.[6]

As with many microeconomic data sets, errors in the measurement of self-reported variables is a potential problem (Li, 2002; Li, Trivedi and Guo, 2003; Guo and Li, 2001). In health economics, a related concern is whether individuals accurately report their level of health care utilization. McCallum et al. (1993) report that evidence from the United States suggests individuals tend to underreport their actual level of usage. They find similar results for Australian data. What factors might cause a person to misreport his level of utilization? According to McCallum et al.(1994),

> "The most obvious factor that might lead to error in self reports is the actual utilization rate. The more similar events that people have to recall, the more their memory is likely to 'decay' to generic memories. Similarly, factors associated with high use, older age, female gender and health status may be associated with error in reporting."

Using their Australian data set, we test their hypotheses using a copula approach. We delete individuals for whom either self reported physician visits or actual physician visits are missing. The final sample size is 502. Summary statistics are provided in Table 1. The utilization variables highlight the problem of misreporting: on average, individuals have 6.5 physician visits (ACTUAL) but report only 4.4 visits (SELF). Explanatory variables include age, sex, income, unemployment status, health status, number of chronic health

[5] A previous regression analysis of these data, but with the main focus on health insurance decision, was given in Cameron and McCallum (1995).

[6] The only possible alternative model for the differences in counts that we are aware of and permits negative counts, is the ordered probit. In our example, the thresholds might be $\leq -8, -7, -6, \ldots, 6, 7, \geq 8$. An example of this is Hausman, Lo and Mackinlay (1992) cited in Cameron and Trivedi (1998, p.88).

conditions, level of education, number of kids under age 6 and age 18, an indicator of private insurance, and an indicator of whether the person works a shift job. We also include an interaction term FEMALE*AGE.

A reasonable assumption is that SELF and ACTUAL are correlated and jointly distributed as $F(\text{SELF}, \text{ ACTUAL})$. The copula approach discussed above provides a representation of $F(\text{SELF}, \text{ACTUAL})$. We use the technique of section 2 to derive the distribution of $z = \text{SELF} - \text{ACTUAL}$ in order to determine sources of misreporting. Both marginal distributions are negative binomial-2 as specified in equation (12). Since both marginals correspond to the same individual, each marginal includes the same set of explanatory variables.

As a preliminary excercise, we attempt to model the effects of covariates on the measurement error in a linear regression context with the actual number of visits as a right hand side variable. The inclusion of this variable on the right hand side of the regression is to address the issue of the relative difference as suggested by a co-editor. Because of the endogeneity of the actual number of visits, we use the predicted value of the actual number of visits from the NB2 as an instrument. Table 2 presents the estimates from the IV estimation. It turns out that no variable is statistically significant: the linear regression model does not help us understand better on the problem of interest.[7]

4.1.1 Results

Results for the Frank copula model are given in Table 3. Utilization for both SELF and ACTUAL is positively and significantly related to age (AGE) and being female (FEMALE), but the interaction term FEMALE*AGE is significantly negative indicating that increased utilization associated with being female diminishes with age. Unhealthy individuals (low value for HEALTH) and those with chronic conditions (CHRONIC) have

[7] We have also tried different specifications such as using log(self/actual) as a dependent variable for the subsample with both positive self and actual reported numbers. None of these specifications give us meaningful results.

more physician visits. Being employed at a shift job (SHIFT) is associated with fewer physician visits.

Having private health insurance (PHINS=1) does not appear to affect utilization. Whereas this result may come as a surprise, since one expects having insurance as a stimulus to additional health care usage, it is possible to rationalize the result in the Australian institutional context. Specifically, we note that "the Australian health system provides universal access to needed health care, regardless of the ability to pay" (Hall, 1999: p. 97). Hall's (1999) overview of the Australian system makes clear that a major role of private insurance is to provide higher quality of care in public hospital. Hence we should not expect that private insurance has any impact on an individual's use of primary care services that we model in this section. The dependence parameter θ is 5.829. Allowing for dependence leads to higher log-likelihood relative to the independence assumption, which suggests that the two outcomes ACTUAL and SELF are jointly determined.[8]

Because of the positive (unconditional and conditional) correlation between the two counts, our benchmark Marshall-Olkin model should also perform well on this sample. However, it is interesting to note that the models have quite large differences in log-likelihood with the Frank copula model having the largest log-likelihood, given only 500 observations, and essentially the same number of parameters. As a result, on any information criteria the Frank copula model will be preferred. For example, using BIC $= -2*ln(L)+K*ln(N)$ where K is the number of parameters estimated and N is the sampe size, the criteria values are 5123.9 for the Marshall-Olkin model and 4893.79 for the Frank copula. Moreover, both the copula approach and the Marshall-Olkin model outperform the UH approach by a larger margin. The implication is that the copula model provides the best fit of the three approaches. This could be due to the fact that the Frank copula model is the least restrictive one among the three models.

To analyze the effect of particular covariates on $z =$ SELF $-$ ACTUAL, we calculate

[8] We have estimated the model under the assumption that the two counts are independent, and found that our copula model provides a better fit to the empirical frequency distribution.

pmfs and display them graphically in Figures 1 - 3, as it is more informative to look at the results in this way. For each graph, covariates are set to their mean values while a particular covariate of interest is adjusted to determine its impact on z. For example, the first graph of the first rows of Figures 1 - 3 shows two pmfs: one for females (FEMALE = 1) and one for males (FEMALE = 0) where the other covariates are set to their means.[9] The pmf for females has a lower peak at zero than the pmf for males. The interpretation is that females tend to misreport their true number of physician visits more than males. The pmfs for both females and males have fatter left tails than their corresponding right tails, which suggests that overall misreporting is mostly due to underreporting rather than overreporting. The left tail of the pmf for females is fatter than that for males, which indicates that females tend to underreport more than males. The right tails are not statistically different from each other.

The second graph of the first rows of Figures 1 - 3 compare thirty year old individuals (AGE = 3.0) and sixty year olds (AGE = 6.0) where other covariates are set to their sample averages. Sixty year olds have a lower peak at zero indicating that they tend to misreport more than thirty year olds. Both graphs have fatter left tails indicating that when individuals misreport their number of physician visits, they then to underreport. Sixty year olds tend to underreport more than thirty year olds.

The first graph of the second rows of Figures 1 - 3 compare those with median health (HEALTH = 8.5) to those at the 25th percentile of health (HEALTH = 2.5). Unhealthy individuals tend to misreport their number of physician visits more than healthy people, while both groups tend to underreport more than overreport.

The second graph of the second rows of Figures 1 - 3 compare shift workers (SHIFT = 1) to nonshift workers (SHIFT = 0). The differences are less pronounced than in the other cases, but nonshift workers show a slight tendency to misreport compared to shift workers.

[9] The interaction term FEMALE*AGE is set to the average of AGE for the female's graph and zero for the male's graph.

In summary, females, sixty year olds, unhealthy individuals, and nonshift workers tend to misreport their number of physician visits when compared to their counterparts. These groups also have higher levels of utilization, as indicated by the coefficients of FEMALE, AGE, HEALTH, and SHIFT. This is consistent with the results of McCallum et al. (1993). Evidently, the more physician visits a person must recall, the less accurate are the self-reported numbers. Moreover, as McCallum et al. (1993, 1994) find, overall misreporting is mostly underreporting, as reflected in the thick left tails of the pmf graphs.

4.2 Measures of Fit

We employ two techniques to gauge the fit of the copula approach. The first measure of fit is Andrews' GoF test (Andrews, 1988). The GoF test is calculated as $S = (\mathbf{f} - \hat{\mathbf{f}})'\hat{\mathbf{\Sigma}}^{-1}(\mathbf{f} - \hat{\mathbf{f}})$ where $(\mathbf{f} - \hat{\mathbf{f}})'$ is an $(N \times q)$ matrix of differences between sample and fitted cell frequencies, q is the number of cells, and $\hat{\mathbf{\Sigma}}$ is its estimated covariance matrix. Under the null hypothesis of no misspecification, the test has an asymptotic $\chi^2(q-1)$ distribution. When the statistic is formed using maximum likelihood estimates, computation is simplified. Let $\mathbf{A}$ be a $(N \times q)$ matrix with ith row given by $(\mathbf{f}_i - \hat{\mathbf{f}}_i)$, and let $\mathbf{B}$ be a $(N \times K)$ matrix with ith row given by $(\partial/\partial\Psi)\log f_i(y_i|\mathbf{\Psi})$, where $\mathbf{\Psi}$ is the vector of K parameters. Defining $\mathbf{H} = [\mathbf{A}\ \ \mathbf{B}]$, the test statistic is

$$\tau_{GoF} = \mathbf{1}'\mathbf{H}(\mathbf{H}'\mathbf{H})^{-1}\mathbf{H}'\mathbf{1} \tag{18}$$

where $\mathbf{1}$ is a column vector of ones. We calculate $q = 10$ cells.[10]

The objective of the test is to determine the fit of the marginal distributions. For the sample we use, the test statistics are 11.57 for self-reported visits and 8.64 for the actual number of visits. These values favor the null hypothesis of no misspecification for

[10] There is a shortcoming of the τ_{GoF} test. Classical tests with fixed significance levels tend to overreject the null hypothesis in large samples. The GoF test suffers from the same problem (Deb and Trivedi, 1997; Cameron and Trivedi, 1998). Despite this caveat, the GoF test serves as a useful indictor of fit with smaller values indicating better fit.

both measures of utilization. Therefore, our marginals are well-specified. The marginal corresponding to actual physician visits has a better fit (lower GoF statistic). This is intuitive because self reported physician visits include measurement error contamination.

We also compared fitted versus empirical cell frequencies of the pmf $g(z)$. For each observation i, $\hat{g}_i(z)$ is calculated for cells $z = -8, ..., 8$. Averaging each cell across all observation produces an estimated pmf $\hat{g}(z)$. Figure 4 shows estimated pmfs of z, denoted by the dashed lines, compared to the actual pmfs of z, denoted by the solid lines. The estimated pmfs of the copula approach and the Marshall-Olkin model appear to match well with the actual pmfs. However, the fit of the UH is not as close to the actual distribution.

5 Conclusion

This paper presents a new method for studying the distribution of the difference between two nonnegative integer counts variables. Estimation is complicated by the lack of availability of a convenient representation for the bivariate distribution of the two counts. The proposed method uses copulas to express the bivariate distribution so that the distribution of the difference between the counts can be recovered. The technique is fully parametric and straightforward to implement.

The approach is demonstrated for an empirical application of determinants of misreporting of physician visits by Australian citizens. Results indicate that the more physician visits an individual must recall, the more likely he is to misreport his number of visits. The elderly, the unhealthy, nonshift workers, and females are groups that have more physician visits than their counterparts, and these groups tend to misreport their true number of doctor visits more than their counterparts. Results also show that misreporting is primarily due to underreporting rather than overreporting.

While this paper focuses on the difference between two counts, the approach can be applied to any situation in which the difference between two outcomes is of interest and

data on both outcomes are available, but a convenient expression for the joint distribution of two outcomes is not available. Furthermore, the method can be extended to model other functions of two outcomes rather than the difference.

References

Andrews, D.W.K. (1988). "Chi-Square Diagnostic Tests for Econometric Models: Introduction and Applications". *Journal of Econometrics*, 37, 135-156.

Berglund E. and K. Brännäs (2001). "Plants' Entry and Exit in Swedish Municipalities". *The Annals of Regional Science*, 35, 431-448.

Bouye, E., V. Durrleman, A. Nikeghbali, G. Riboulet, T. Roncalli, (2000). "Copulas for Finance: A Reading Guide and Some Applications". Unpublished Manuscript, Financial Econometrics Research Centre, City University Business School: London.

Cameron, A.C. and J. McCallum (1995), "Private Health Insurance Choice in Australia: The Role of Long-term Utilisation of Health Services" in H. Lapsley ed., *Economics and Health: 1995, Proceedings of the Seventeenth Australian Conference of Health Economists*, pp. 143-162, Australian Studies in Health Service Administration No. 79, School of Health Services Management, University of New South Wales.

Cameron, A.C. and P.K. Trivedi (1998). *Regression Analysis of Count Data*. Econometric Society Monographs 30, Cambridge University Press, New York.

Capéraà, P., A. Fougères and C. Genest (2000). "Bivariate Distributions with Given Extreme Value Attractor". *Journal of Multivariate Analysis*, 72, 30-49.

Chen, X and Y. Fan (2002). "Estimation of Copula-Based Semiparameteric Time Series Models". Working paper, Vanderbilt University.

Chib, S. and R. Winkelmann (2001). "Markov Chain Monte Carlo Analysis of Correlated Count Data". *Journal of Business and Economic Statistics*, 19(4), 428-435.

Clayton, D.G. (1978). "A Model for Association in Bivariate Life Tables and its Application in Epidemiological Studies of Familial Tendency in Chronic Disease Incidence". *Biometrika*, 65(1), 141-151.

Deb, P. and P.K. Trivedi (1997). "Demand for Medical Care by the Elderly: A Finite Mixture Approach". *Journal of Applied Econometrics*, 12, 313-326.

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

Fréchet, M. (1951). "Sur les Tableaux de Correlation Dont les Marges Sont Donnees". *Annals of the University of Lyon, Section A*, 14, 53-77.

Frees, E.W. and E.A. Valdez (1998). "Understanding Relationships Using Copulas". *North American Actuarial Journal*, 2(1), 1-26.

Genest, C. and L. Rivest (1993). "Statistical Inference Procedures for Bivariate Archimedean Copulas". *Journal of the American Statistical Association*, 88(423), 1034-1043.

Gourieroux, C., A. Monfort, and A. Trognon (1984). "Pseudo Maximum Likelihood Methods: Applications to Poisson Models". *Econometrica*, 52, 701-720

Guo, J. and T. Li (2001). "Simulation-Based Estimation of the Structural Errors-in-Variables Negative Binomial Regression Model with an Application". *Annals of Economics and Finance*, 2, 101-122.

Hall, J. (1999). "Incremental Change In the Australian Health Care System". *Health Affairs*, 18(3), 95-110.

Hausman, J.A., A.W. Lo, and A.C. Mackinlay (1992). "An Ordered Probit Analysis of Transaction Stock Prices". *Journal of Financial Economics*, 31, 319-379.

Hutchinson, T.P. and C.D. Lai (1990). *Continuous Bivariate Distributions, Emphasising Applications*. Rumsby, Sydney, Australia.

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

Kimeldorf, G. and A.R. Sampson (1975). "Uniform Representations of Bivariate Distributions". *Communications in Statistics*, 4, 617-627.

Kocherlakota, S. and K. Kocherlakota (1992). *Bivariate Discrete Distributions*. New York: Marcel Dekker.

Lee, L. (1983). "Generalized Econometric Models with Selectivity". *Econometrica*, 51, 2, 507-512.

Li, T. (2002). "Robust and Consistent Estimation of Nonlinear Errors-in-Variables Models". *Journal of Econometrics*, 110(1), 1-26.

Li, T., P.K. Trivedi and J. Guo (2003). "Modeling Response Bias in Count: A Structural Approach with an Application to the National Crime Victimization Survey Data". *Sociological Methods & Research*, 31, 514-544.

Marshall, A. (1996). "Copulas, Marginals, and Joint Distributions" in *Distributions with Fixed Marginals and Related Topics*, ed. by L. Ruschendorf, B. Schweizer, and M.D. Taylor, Institute of Mathematic Statistics, Hayward (CA), 213-222.

Marshall, A.W. and I. Olkin (1990). "Multivariate Distributions Generated from Mixtures of Convolution and Product Families". In H.W. Block, A.R. Sampson, And T.H. Savits, *Topics in Statistical Dependence*. Pages 371-393. IMS Lecture Notes-Monograph Series, Volume 16.

Mayer, W.J. and W.E. Chappell (1992). "Determinants of Entry and Exit: An Application of the Compound Bivariate Poisson Distribution to the U.S. Industries, 1972-1977". *Southern Economic Journal*, 58(3), 770-778.

McCallum, J., J. Lonergan and C. Raymond (1994). "The NCEPH Record Linkage Pilot Study: A Preliminary Examination of Individual Health Insurance Commission Records with Linked Data Sets". Working paper, National Center for Epidemiology and Population Health, Australian National University.

McCallum, J., C. Raymond, and C, McGilchrist (1993). "How Accurate Are Self Reports of Doctor Visits?". Working Paper, Australian National University.

Miller, D.J., and W-h. Liu (2002). "On the Recovery of Joint Distributions from Limited Information". *Journal of Econometrics*, 107, 259-274.

Munkin, M. and P.K. Trivedi (1999). "Simulated Maximum Likelihood Estimation of Multivariate Mixed-Poisson Regression Models, with Application". *Econometrics Journal*, 2, 29-48.

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

Prieger, J. (2002). "A Flexible Parametric Selection Model for Non-Normal Data with Application to Health Care Usage". *Journal of Applied Econometrics*, 17(4), 367-392.

Sklar, A. (1959). "Fonctions de Repartition a n Dimensions et Leurs Marges". Publications de l'Institute de Statistique de l'Universite de Paris, 8, 229-231.

Sklar, A. (1973). "Random Variables, Joint Distributions, and Copulas". *Kybernetica*, 9, 449-460.

Sklar, A. (1996). "Random Variables, Distribution Functions, and Copulas - A Personal Look Backward and Forward" in *Distributions with Fixed Marginals and Related Topics*, ed. by L. Ruschendorf, B. Schweizer, and M.D. Taylor, Institute of Mathematic Statistics, Hayward (CA), 1-14.

Smith, M. (2003). "Modeling Selectivity Using Archimedean Copulas". *Econometrics Journal*, 6, 99-123.

Tajar, A., M. Denuit, Ph. Lambert (2001). "Copula-Type Representation for Random Couples with Bernoulli Margins". Working paper, Universite Catholique de Louvain.

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

Van Ophem, H. (2000). "Modeling Selectivity in Count Data Models". *Journal of Business and Economic Statistics*, 18, 503-511.

Table 1 – Summary Statistics for Australian Application

Variable	De…nition	Mean	St. Dev.
Utilization			
ACTUAL	# of actual physician visits	6.52	5.92
SELF	# of self reported physician visits	4.37	4.55
Demographic			
AGE	age/10	4.59	1.24
FEMALE	=1 if female	0.48	0.50
FEMALE*AGE	female*age interaction	2.19	2.45
INCOME	income/10000	3.01	2.42
KIDS18	# of kids younger than 18	0.96	1.23
KIDS6	# of kids younger than 6	0.30	0.64
EDUC1	=1 if …rst level education	0.03	0.18
EDUC2	=1 if second level education	0.26	0.44
EDUC3	=1 if third level education	0.29	0.46
EDUC4	=1 if fourth level education	ommitted	
UNEMP	= 1 if umemployed	0.03	0.18
SHIFT	= 1 if shift worker	0.07	0.26
PHINS	= 1 if holds private health insurance	0.73	0.44
Health			
HEALTH	health score	7.39	2.10
CHRONIC	# of chronic conditions	1.80	1.76

Sample Size = 502

Table 2 - IV Results
Dependent Variable = (SELF ¡ ACTUAL)

Variable	Coe¤.	St. Err.
INTERCEPT	1.528	1.375
AGE	-0.434	0.587
INCOME	0.086	0.060
FEMALE	-0.501	2.993
FEMALE*AGE	0.157	0.513
KIDS6	0.288	0.382
KIDS18	-0.135	0.171
UNEMP	1.034	1.076
SHIFT	0.451	1.114
EDUC1	-0.420	1.892
EDUC2	-0.778	0.690
EDUC3	-0.814	0.965
HEALTH	-0.033	0.264
COND	0.113	0.450
PHINS	0.168	0.309
ACTUAL	-0.261	0.475
R-squared = 0.37		

* signi…cant at 5 percent level
Note: Fitted values of ACTUAL from a …rst stage negative binomial regression are used as instruments.

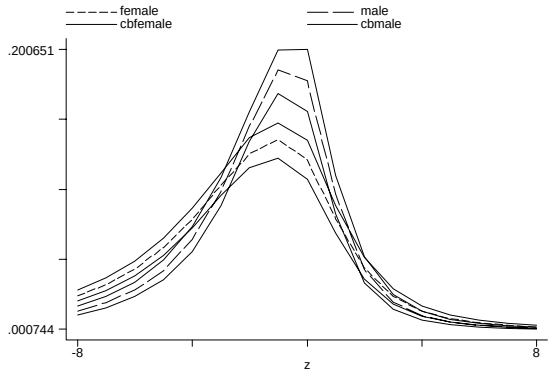
Table 3 – Results of Estimation

	Frank Copula		Marshall-Olkin		UH	
Variable	Coe¤.	St. Err.	Coe¤.	St. Err.	Coe¤.	St. Err.
			Self-Reported Visits			
Intercept	0.899*	0.308	-1.974*	0.768	0.473	0.307
AGE	0.108*	0.052	0.313*	0.127	0.206*	0.051
INCOME	0.024	0.017	0.047*	0.016	0.020	0.017
FEMALE	1.017*	0.468	1.967*	0.696	1.305*	0.297
FEMALE*AGE	-0.140	0.093	-0.291*	0.125	-0.193*	0.060
KIDS6	0.148	0.077	0.236	0.131	0.047	0.078
KIDS18	-0.108*	0.039	-0.102	0.072	-0.047	0.040
UNEMP	0.225	0.308	-0.099	0.362	-0.186	0.278
SHIFTW	-0.360*	0.132	-0.468*	0.176	-0.446*	0.165
EDUC1	0.150	0.250	0.238	0.224	0.383*	0.161
EDUC2	0.020	0.143	0.075	0.102	0.273*	0.109
EDUC3	0.141	0.104	0.121	0.140	0.394*	0.099
HEALTH	-0.088*	0.017	-0.084*	0.018	-0.095*	0.018
COND	0.168*	0.022	0.129*	0.020	0.165*	0.024
PHINS	0.077	0.096	0.050	0.121	0.040	0.098
$®_{self}$	0.443*	0.052	–	–	1e-4	0.006
			Actual Visits			
Intercept	0.700*	0.159	-1.036*	0.407	0.620*	0.311
AGE	0.217*	0.038	0.222*	0.063	0.122*	0.053
INCOME	0.016	0.018	0.016	0.014	0.034	0.018
FEMALE	1.363*	0.293	1.269*	0.357	1.075*	0.321
FEMALE*AGE	-0.211*	0.058	-0.200*	0.063	-0.140*	0.065
KIDS6	0.098	0.077	0.071	0.063	0.149	0.088
KIDS18	-0.047	0.041	-0.032	0.035	-0.100*	0.042
UNEMP	0.029	0.314	-0.258	0.262	0.040	0.315
SHIFTW	-0.523*	0.145	-0.401*	0.166	-0.335*	0.156
EDUC1	0.413*	0.187	0.243*	0.111	0.276	0.207
EDUC2	0.192	0.136	0.165*	0.072	0.094	0.112
EDUC3	0.298*	0.103	0.220*	0.077	0.221*	0.101
HEALTH	-0.068*	0.018	-0.049*	0.012	-0.108*	0.018
COND	0.131*	0.023	0.083*	0.015	0.192*	0.024
PHINS	0.008	0.092	0.005	0.069	0.060	0.102
$®_{actual}$	0.487*	0.053	–	–	0.067*	0.025
µ;®; $°_1$ (respectively)	5.829*	0.492	1.870*	0.126	0.846*	0.046
	-2440.68		-2465.61		-2825.77	

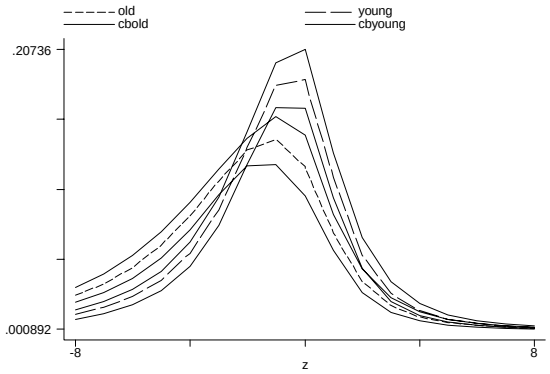
* signi…cant at 5 percent level

Figure 1 – Australian PMFs for the Frank Copula
95 percent confidence bands (CB) indicated by solid lines

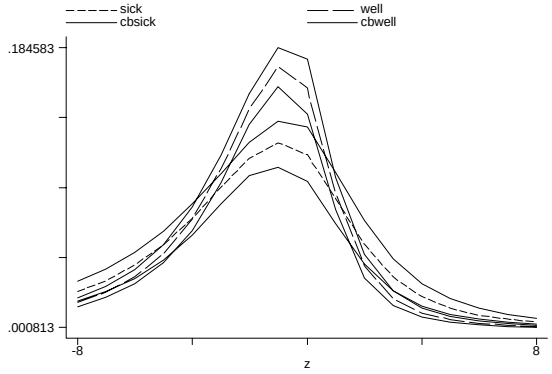
–female versus male–



–sixty-year-olds (OLD) versus thirty-year-olds (YOUNG)–



–health=2.5 (SICK) versus health=8.5 (WELL)–



–shift worker (SHIFT) versus nonshift worker (NOSHIFT)–

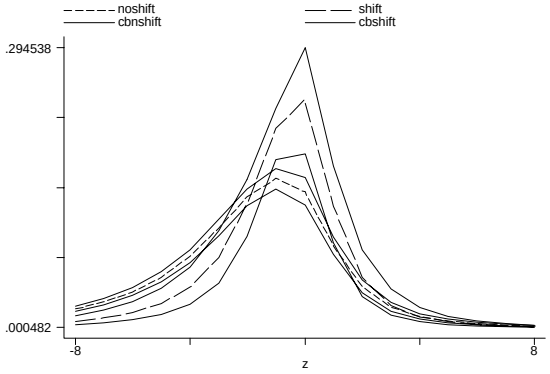
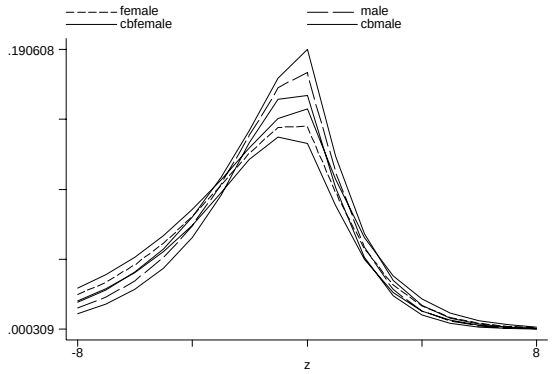
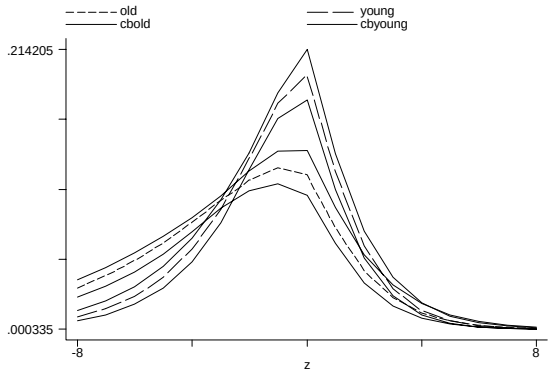


Figure 2 – Australian PMFs for the Marshall-Olkin Model
95 percent confidence bands (CB) indicated by solid lines

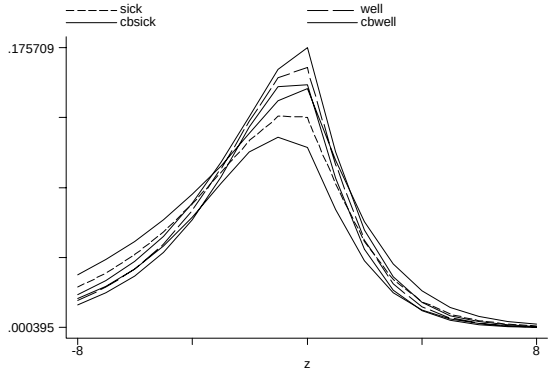
–female versus male–



–sixty-year-olds (OLD) versus thirty-year-olds (YOUNG)–



–health=2.5 (SICK) versus health=8.5 (WELL)–



–shift worker (SHIFT) versus nonshift worker (NOSHIFT)–

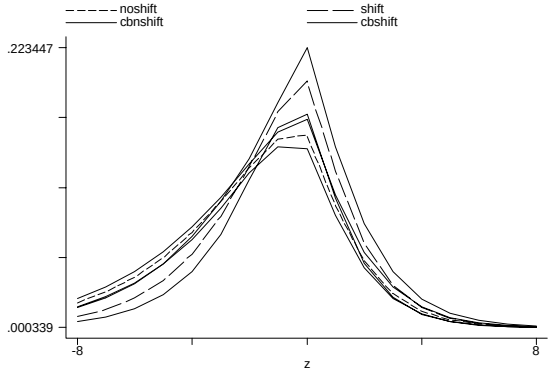
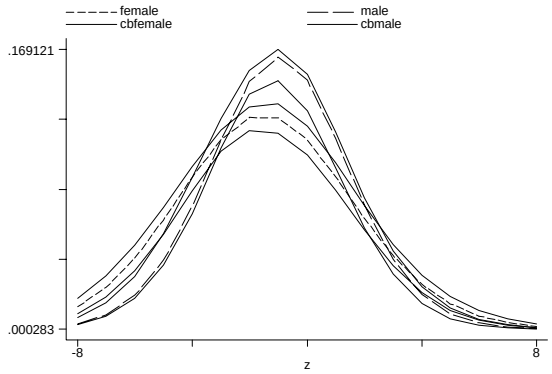
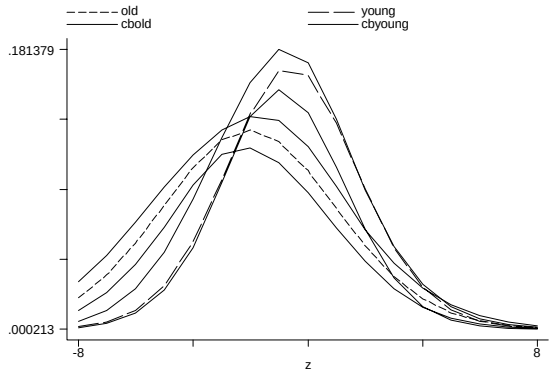


Figure 3 – Australian PMFs for the UH Model
95 percent confidence bands (CB) indicated by solid lines

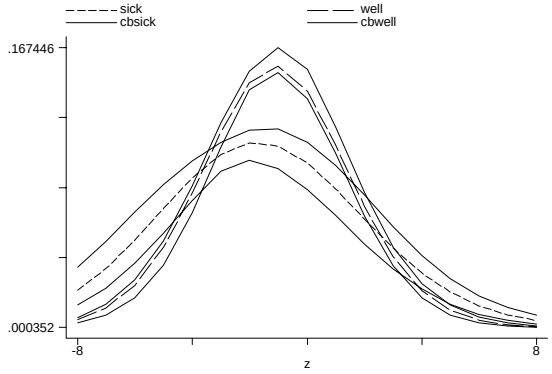
–female versus male–



–sixty-year-olds (OLD) versus thirty-year-olds (YOUNG)–



–health=2.5 (SICK) versus health=8.5 (WELL)–



–shift worker (SHIFT) versus nonshift worker (NOSHIFT)–

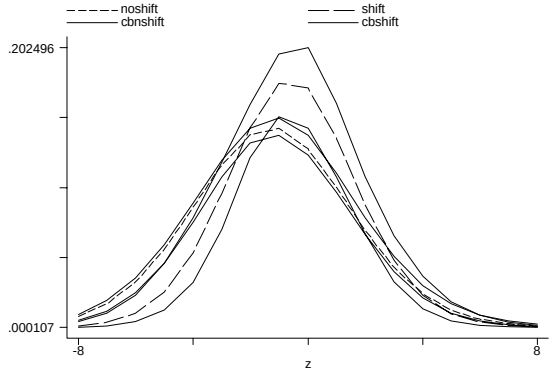
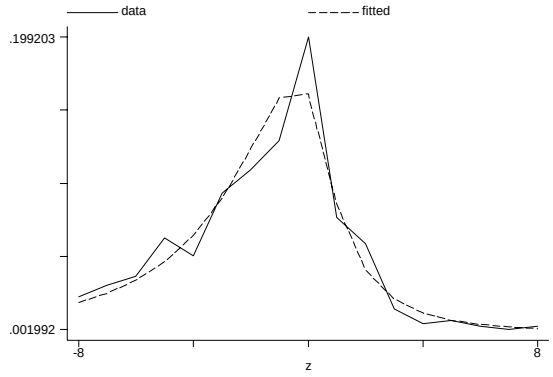
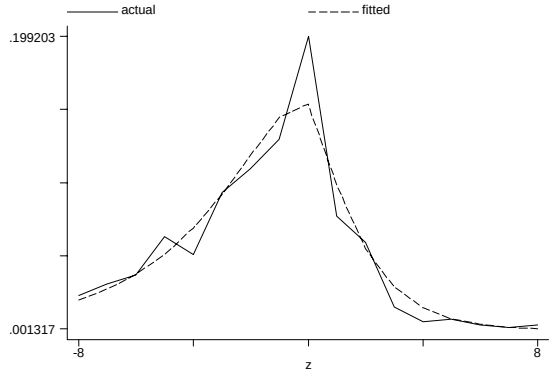


Figure 4 – Fitted pmfs of $z = y_1 - y_2$

–FRANK–



–MARSHALL OLKIN–



–UH–