Comparison of parametric, semi-parametric and nonparametric two-stage estimation methods for copula models

Haftom Temesgen Abebe

promotor :
dr. Goele MASSONNET

Eindverhandeling voorgedragen tot het bekomen van de graad
Master of Statistics Biostatistics

universiteit hasselt

Interuniversity Institute for Biostatistics and
Statistical Bioinformatics
Universiteit Hasselt

Comparison of Parametric, Semi-parametric and Nonparametric two-stage estimation Methods for copula models

By:
Haftom Temesgen Abebe

Supervisor: Dr. Goele Massonnet

A thesis submitted in partial fulfillment of award of Master of Biostatistics to the Center for Statistics, Universiteit Hasselt, Diepenbeek, Belgium

2007- 2009

Certification

This is to certify that this project was carried out by Haftom Temesgen Abebe under our thorough supervision and reflects his true research ability.

Haftom Temesgen Abebe

Student

……..……..……..……..

Signature

….……

Date

Supervisor: Dr. Goele Massonnet

…….…….……...

Signature

…….

Date

Acknowledgments

Thanks to the Almighty God for his unlimited blessing.
First and for most I would like to express my deeply indebted to my supervisor Dr. Goele Massonnet for her guidance, suggestions and for sharing me her previous working papers from the very early stage of the project. Goele is a very kind person to work with. It was a real pleasure to work with her.

I would like to express my heartfelt appreciation for my professors at Hasselt University, Martine Machiels for her administrative and secretariat support, VLIR for Scholarship, my family members and my friends. Specially, I would like to thank my wife Desta Abay. Her patient love enabled me to complete my studies.

Haftom Temesgen Abebe
Diepenbeek
August 28, 2009

Abstract

Introduction: Copulas are often used to model correlated survival data with small and equal cluster size. In copula models, the joint survival function of the different observational units in a cluster is modelled as a function, called the copula, of the marginal survival functions of the observational units. The copula determines the type of dependence. We consider parametric copula models; the parameter vector is called the dependence or association parameter vector. A nice feature of copulas is that the marginal survival functions in the copula model do not depend on the parameters of the copula. Therefore, the marginal survival functions and the dependence parameter vector can be estimated separately. This is the idea of the two-stage estimation approach.

In the first stage, the marginal survival functions are estimated. This can be done in a parametric, a semi-parametric and a nonparametric way, possibly taking into account the effect of covariates (Shih and Louis, 1995; Glidden, 2000; Andersen, 2005). In the second stage, the dependence parameter vector is estimated by maximizing the loglikelihood, with the marginal survival functions replaced by their estimates obtained in the first stage.

Objective: The objective of this thesis is to compare the parametric, semi-parametric and nonparametric two-stage estimation methods. Based on a simulation study, we would like to study the robustness of the estimation of the dependence vector against misspecification of the marginal survival functions.

Method: In this study, all approaches parametric, semi-parametric and nonparametric were compared for data sets which were generated from Clayton copula using simulations. Boxplots, estimated bias and estimated mean square error (MSE) of the copula parameter are the bases for the comparison between those three approaches.

Results and conclusion: The results provide evidence for the fact that parametric approach is more accurate for the correct model (multiplicative Weibull regression model) and for the misspecified models (additive Weibull model and multiplicative loglogistic regression model).

Key words: *copula function, joint survival function, two-stage estimation, Clayton copula*.

Contents

List of Tables

List of Figures

1. Introduction

In survival analysis the response of interest is the time from a well-defined time origin to the occurrence of a specific event. Some examples are the time from onset of disease to death, the time from recovery to the time of recurrence of a disease. This response is called the failure time, the survival time or the event time. A special difficulty that often occurs in the analysis of survival data is the possibility that some responses are known to have occurred only within certain intervals. Such incomplete observation of the event time is called (right) censoring. Right censoring occurs when only a lower bound for the time of interest is known. In many studies there is a natural clustering in the data; event times within the same cluster may be correlated. Such data are known as clustered survival data. In this thesis, we consider bivariate data with right censoring.

In the recent years, copula models became increasingly popular for modelling multivariate survival data. According to Li (2000), a copula is a function that links univariate marginals to their full multivariate distribution. In copula models the joint survival function of the event times in a cluster is modelled as a function, called the copula. Genest and Mackay (1986) and Shih and Louis (1995) offer a way of combining the marginal approach with a model for the association within units. The joint survival function is modelled through the marginal survival functions and the association parameter. Because of the way the copula model splits the problem in two with the marginal survival functions and an association parameter, a two-stage estimation procedure suggests itself by first estimating the margins and then using the estimated margins to estimate the association parameter.

A copula is used as a general way of formulating a multivariate distribution with small and equal cluster size in such a way that various general types of dependence can be represented. One of the advantages of copula models is their relative mathematical simplicity; the marginal survival functions do not depend on the parameters of the copula. Another advantage is the possibility to build a variety of dependence structures based on existing parametric or nonparametric models of the marginal distributions. Therefore, the marginal survival functions and the dependence parameter vector can be estimated separately. This is the idea of the two-stage

estimation approach. This two-stage estimation procedure was suggested by Hougaard (1987) and also studied by Shih and Louis (1995) in the case where each margin was modelled separately.

In the first stage, the marginal survival functions are estimated. This can be done in a parametric, a semi-parametric or a nonparametric way, possibly taking into account the effect of covariates (Shih and Louis, 1995; Glidden, 2000; Andersen, 2005). In the second stage, the dependence parameter is estimated by maximizing the loglikelihood, with the marginal survival functions replaced by their estimates obtained from the first stage.

The objective of this project is to compare the parametric, semi-parametric and the nonparametric two-stage estimation methods. Based on a simulation study, we study the robustness of the estimation of the dependence parameter vector against misspecification of the marginal survival functions.

In Section 2 we describe some general ideas on copulas and we give a definition of the Clayton copulas. We also discuss the parametric, semi-parametric and nonparametric two-stage estimation methods. In Section 3 we compare the performance of the three estimation methods based on a simulation study. Finally, the discussion and conclusions are presented in Section 4.

2. Methodology

2.1 Copula models

Copula models are typically used to model the joint survival function of clustered data with small and equal cluster size. The copula approach is often used for bivariate data (see. e.g., Shih and Louis, 1995; Andersen, 2005; Duchateau and Janssen, 2008). We study copula models for bivariate failure time data. Assume we have K different clusters, each cluster i having 2 subjects. Each subject is observed from time zero to a failure time T_{ij} or to a potential right censoring time C_{ij} independent of T_{ij}. Let X_{ij} = min (T_{ij} , C_{ij}) be the observed time and δ_{ij} be the censoring indicator which is equal to 1 if $X_{ij}=T_{ij}$ and 0 otherwise. Let (T_{i1}, T_{i2}) be bivariate of failure times the observational units in cluster i and let $\boldsymbol{S}_{ij}$, j=1,2, be the marginal survival functions of T_{ij}, where the index ij is used to indicate that the marginal survival function may depend on a covariate $\boldsymbol{z}_{ij}$. The survival copula is the function that links the marginal survival functions $\boldsymbol{S}_{ij}$ to generate the joint survival function, i.e.,

$$S_i(t_1, t_2) = C_\theta(S_{i1}(t_1), S_{i2}(t_2)) \qquad t_1, t_2 \geq 0 \tag{1}$$

for the bivariate distribution function $C_{\boldsymbol{\theta}}$ defined on $(u, v) \in [0, 1]^2$, taking values in [0, 1] and having uniform marginals. C_θ is called a survival copula with parameter θ and satisfying: $C_\theta(u,0) = 0 = C_\theta(0, v)$, $C_\theta(u,1) = u$ and $C_\theta(1,v) = v$. The existence (and uniqueness if the marginal survival functions are all continuous) follows from Sklar's theorem (Sklar, 1959). In this report we focus on an important class of copulas known as Archimedean copulas, which received considerable attention (Genest and Mackay, 1986) for the study of clustered survival data. Archimedean copulas are defined as

$$C(u, v) = \phi^{-1}[\phi(u) + \phi(v)], \tag{2}$$

where $u, v \in [0, 1]$ and where ϕ is a generator of the copula, satisfying: ϕ: $[0, 1] \rightarrow [0, \infty]$, is a continuous strictly decreasing function such that $\phi(0) = \infty$, $\phi(1)=0$, and $\phi^{-1}$is the pseudo-inverse of ϕ, continuous and non increasing on $[0, \infty]$ and strictly decreasing on $[0, \phi(0)]$ and ϕ is convex.

There are a great variety of families of copulas which belong to the Archimedean copulas. However, in this report we consider the Clayton copula. For the Clayton copula, $\phi^{-1}(s) = (1+\theta s)^{-1/\theta}$ is the Laplace transform of a gamma distribution with mean 1 and variance θ. The generator of the copula is the inverse of the Laplace transform, i.e.,

$$\phi(u) = \frac{1}{\theta}(u^{-\theta} - 1).$$

Hence, the Clayton copula is given by:

$$C_\theta(u, v) = (u^{-\theta} + v^{-\theta} - 1)^{-1/\theta}, \text{ where } \theta > 0. \tag{2}$$

The failure times T_{i1} and T_{i2} are positively associated when $\theta > 0$ and are independent when $\theta \to 0$. For investigating more deeply the dependence parameter we consider a measure of correlation, which is known as Kendall's τ. Kendall's τ is a rank correlation measure, and so measures the strength of the relationship between two variables. It is invariant under strictly increasing transformations of the underlying random variables. Like other measures of correlation Kendall's τ will take values between -1 and +1, with a positive correlation indicating that the ranks of both variables increase together while a negative correlation indicates that as the rank of the one variable increases the other one decreases.

Let T_{i1} and T_{i2} be failure times with an Archimedean copula C generated by ϕ, Kendall's τ is then given by:

$$\tau_\vartheta = 1 + 4\int_0^1 \frac{\varphi(t)}{\varphi'(t)} dt \tag{3}$$

When C is a Clayton copula (3), then Kendall's τ is given by:

$$\tau_\vartheta = 1+4\int_0^1 \frac{t^{-\theta}-1}{-\theta t^{-\theta-1}} dt = 1+4\int_0^1 \frac{t^{\theta+1}-t}{\theta} dt = \frac{\theta}{\theta+2}, \tag{4}$$

where θ is the copula parameter.

2.2 Two-stage estimation

The bivariate survival function S_i (. , .) in survival copula model (1) is characterized by the dependence function C_θ(. , .) and the two marginal survival functions S_{i1}(.) and S_{i2} (.). This special structure suggests that we may estimate the two margins and the copula parameter θ separately. The estimation of S_i (. , .) may be carried out in two-stages: in the first stage, the marginal survival functions S_{i1}(.) and S_{i2}(.) are estimated nonparametrically (Nelson-Aalen estimator), semi-parametrically or by using a parametric model (Weibull regression model). At the second stage, the estimated marginal survival functions are plugged in the likelihood and the association parameter is estimated via the maximum likelihood method. The likelihood, L, can be written as

$$\prod_{i=1}^{K} f(X_{i1},X_{i2})^{\delta_{i1}\delta_{i2}} \frac{\partial S(X_{i1},X_{i2})}{\partial X_{i1}}^{\delta_{i1}(1-\delta_{i2})} \frac{\partial S(X_{i1},X_{i2})}{\partial X_{i2}}^{\delta_{i2}(1-\delta_{i1})} S(X_{i1},X_{i2})^{(1-\delta_{i1})(1-\delta_{i2})}. \quad (5)$$

Using the relationship between the joint survival function and the survival copula (1) and using as notation $U_i = S_1(X_{i1})$ and $V_i = S_2(X_{i2})$, the likelihood in terms survival copula is given by:

$$\prod_{i=1}^{K} [f(X_{i1}) f(X_{i2})]^{\delta_{i1}\delta_{i2}} c_\theta(U_i,V_i)^{\delta_{i1}\delta_{i2}} \frac{\partial C_\theta(U_i,V_i)}{\partial u_i}^{\delta_{i1}(1-\delta_{i2})} \frac{\partial C_\theta(U_i,V_i)}{\partial V_i}^{\delta_{i2}(1-\delta_{i1})} C_\theta(U_i,V_i)^{(1-\delta_{i1})(1-\delta_{i2})}. \quad (6)$$

2.2.1 First stage: estimation of the marginal survival functions

In the first stage of the estimation, we estimate the marginal survival functions using a parametric, a semi-parametric and a nonparametric approach. A model for the marginal survival functions is fitted taking the clustering into account in the variance of the estimated parameters.

2.2.1.1 Parametric approach

In the first stage of the estimation the marginal survival functions are estimated by using a parametric approach proposed by Andersen (2005). We assume that the marginal survival function comes from a multiplicative Weibull regression model:

$$\widehat{S}_{ij}(t) = \exp\{-\hat{\lambda} t^{\rho} \exp(\hat{\beta} z_{ij})\}.$$

It follows from Andersen (2005) that the parameter estimators are consistent and asymptotically normal.

2.2.1.2 Semi-parametric approach

In the semi-parametric estimation approach, we model the marginal survival functions in the first step using a Cox proportional hazard model with covariate $z_{ij.}$, as proposed by Andersen (2005).
The Cox regression model is given by:

$$\lambda_{ij}(t) = \lambda_0(t)\exp(\beta z_{ij}),$$

where $\lambda_o(t)$ is the baseline hazard at time t and β is the fixed effect parameter.
In terms of the cumulative hazard function, we obtain

$$\Lambda_{ij}(t) = \Lambda_0(t)\exp(\beta z_{ij}).$$

The survival function is then given by:

$$\hat{S}_{ij}(t) = \exp\left\{-\hat{\Lambda}_0(t;\hat{\beta})\exp(\hat{\beta} z_{ij})\right\}.$$

It follows from Spiekerman and Lin (1998) that $\hat{\Lambda}_0(.;\hat{\beta})$ and $\hat{\beta}$ are consistent and asymptotically normal.

2.2.1.3 Nonparametric approach

In the nonparametric approach, we assume that we observe a deterministic binary covariate z_{ij} at the observational unit level. In the first stage the marginal survival

functions are estimated by using a Nelson-Aalen estimator for the group with $z_{ij}=0$ and $z_{ij}=1$ separately. For j=1,2, we have

$$\hat{S}_{ij}(t)=\begin{cases}\exp\left(-\hat{\Lambda}_1(t)\right), \text{ if } z_{ij}=0,\\ \exp\left(-\hat{\Lambda}_2(t)\right), \text{ if } z_{ij}=1.\end{cases}$$

Where $\hat{\Lambda}_1(.)$ and $\hat{\Lambda}_2(.)$ are the Nelson-Aalen estimators for the cumulative hazard function for the group with $z_{ij}=0$ and $z_{ij}=1$.

2.2.2 Second stage: estimation of the association parameter

In the second stage of estimation, we estimate the copula parameter θ by maximizing the loglikelihood, in which the marginal survival functions are replaced by their estimates obtained in the first stage. The derivations of the formulas are found in the Appendix I. Then the log likelihood function for the Clayton copula can be written as

$$\sum_{i=1}^{K}\delta_{i1}\delta_{i2}[\log(S_1^{-\theta}(X_{i1})+S_2^{-\theta}(X_{i2})-1)^{-1/\theta\ -2}\ S_1^{-\theta-1}(X_{i1})\,S_2^{-\theta-1}(X_{i2})]+$$
$$\sum_{j=1}^{2}\sum_{i=1}^{K}\delta_{ij}(1-\delta_{ij})\log[(S_1^{-\theta}(X_{i1})+S_2^{-\theta}(X_{i2})-1)^{-1/\theta\ -1}\ S_j^{-\theta-1}(X_{i1})]+$$
$$\sum_{i}^{K}\ (1-\delta_{i1})(1-\delta_{i2})\log[(S_1^{-\theta}(X_{i1})+S_2^{-\theta}(X_{i2})-1)^{-1/\theta}] \qquad . \qquad (7)$$

In the second stage of the estimation, we replace in this logliklihood expression $S_{ij}(X_{ij})$ by the estimated marginals $\hat{S}_{ij}(X_{ij})$ from the first stage. To estimate the copula parameter θ, the logliklihood expression (7) is maximized with respect to θ.

3. Simulation studies

3.1 Description of the simulations

This section presents a simulation study aimed at assessing the behaviour of the association and the regression parameters, for the parametric, semi-parametric and the nonparametric approaches presented in Section 2.2. A possible shortcoming of the parametric and semi-parametric methods of estimating the copula parameter θ is that they are likely to be inconsistent if the marginal distributions are misspecified at the first step. To study the robustness of the model, we generate 1000 data sets with 100 and 500 clusters that contain two observations each. The observations for each data set are generated in the following way. First, 100 and 500 pairs (U_{i1}, U_{i2}) are generated from Clayton copula using the algorithm in Appendix II. To obtain the failure times T_{ij} we assume that in the first step the marginal survival functions come from:

1. **A multiplicative Weibull regression model with a binary covariate (z_{i1} = 1 for the first level and z_{i2} = 0 for the second level)**

The marginal survival function is then given by:

$$U_{ij} = S_{ij}(T_{ij}) = \exp\{-\Lambda_0(T_{ij})\exp(z_{ij}\beta)\}, \tag{8}$$

with $\Lambda_0(T_{ij}) = \lambda T_{ij}^{\rho}$. The jth failure time in the ith cluster is then obtained as follows:

$$T_{ij} = \left[\frac{-\log(U_{ij})}{\lambda \exp(z_{ij}\beta)}\right]^{1/\rho},$$

for j = 1,2, and $i=1,2,\ldots,K$, where (U_{i1} , U_{i2}) are the pairs that we obtained by generating from a Clayton copula. This assumption corresponds to the correct specification of the marginal distributions.

2. **A multiplicative loglogistic regression model with a binary covariate (z_{i1} = 1 for the first level and z_{i2} = 0 for the second level)**

The marginal survival function is then given by:

$$\begin{aligned} U_{ij} &= S_{ij}(T_{ij}) = \exp\{-\Lambda_0(T_{ij})\exp(z_{ij}\beta)\} \\ &= [S_0(t)]^{\exp(z_{ij}\beta)} \end{aligned} \quad , \tag{9}$$

with $S_0(t) = \dfrac{1}{1+\exp(\alpha)t^k}$. The jth failure time in the ith cluster is obtained as follows:

$$T_{ij} = \left[\frac{1-(U_{ij})^{1/\exp(\beta Z_{ij})}}{\exp(\alpha)\,(U_{ij})^{1/\exp(z_{ij}\beta)}}\right]^{1/k},$$

for j = 1,2. and i = 1,2,..,K, where (U_{i1} , U_{i2}) are the pairs that we obtained by generating from a Clayton copula.

3. **An additive regression model with Weibull baseline hazard model with a binary covariate (z_{i1} = 1 for the first level and z_{i2} = 0 for the second level)**

$$\lambda_{ij}(t) = \lambda\rho t^{\rho-1} + \beta z_{ij}.$$

The marginal survival function is given by:

$$S_{ij}(t) = \exp(-\lambda\rho t^{\rho-1} - \beta z_{ij} t). \tag{10}$$

The computation of the T_{ij} is obtained by solving numerically such that equation (11) is satisfied.

$$\begin{aligned} &\log(U_{ij}) = -\lambda\rho t^{\rho-1} - \beta z_{ij} t \\ &\Rightarrow \log(U_{ij}) + \lambda\rho t^{\rho-1} + \beta z_{ij} t = 0, \end{aligned} \tag{11}$$

for j = 1,2, and i = 1,2,..,K, where (U_{i1} , U_{i2}) are the pairs that we obtained by generating from a Clayton copula.

A censoring time C_{ij} for the observations of each cluster is generated from a uniform distribution so that we obtained approximately 20% censoring. Thus we have (T_{i1}, T_{i2}) and (C_{i1}, C_{i2}), for i = 1, 2, … , K. The observed times are given by X_{ij} = min (T_{ij}, C_{ij}). For each simulated data set, we estimate the marginal survival functions and the association parameter using the parametric, semi-parametric and nonparametric two-stage estimation approaches discussed in Section 2.2.

3.2 Choice of the parameters

For each of the three model assumptions described in Section 3.1, we consider 100 and 500 clusters with 2 observations each. Three values of the association parameter θ were considered corresponding to the Kendall's τ of 0.2, 0.5 and 0.8 for the Clayton copula. Therefore, using (4) the values of θ that we choose for the Clayton copula are 0.5, 2.0, and 8.0, which corresponds to the Kendall's τ respectively.

The parameter values β, λ and ρ of the multiplicative Weibull regression model are chosen to resemble the diagnosis data set in Duchateau and Janssen (2008): β= 0.522 λ= 0.119 and ρ= 2.42. For the multiplicative loglogistic regression model and additive regression model with Weibull baseline hazard, we choose the parameters such that the means and the variances of the event times T_{ij} of these two models are (approximately) the same as the mean and variance of the event times T_{ij} in the case of the multiplicative Weibull regression model. The computation of the means and the variances of these three models are found in Appendix III. By solving these equations numerically, we obtain the values for the parameters of the multiplicative loglogistic regression model: α = -3.1061, k = *4.5725* and β =*0.3534*. For additive regression model with Weibull baseline hazard, we obtain λ= 0.119 β= 0.0391 and ρ= 2.4199.

The motivation for choosing the parameter values of the multiplicative loglogistic regression model and the additive regression model with Weibull baseline hazard can be seen from the plots of the density functions of T_{ij} that correspond to these three models to compare the shape of their densities.

From Figure 1 it was observed that the three density functions have approximately the same mean and variance.

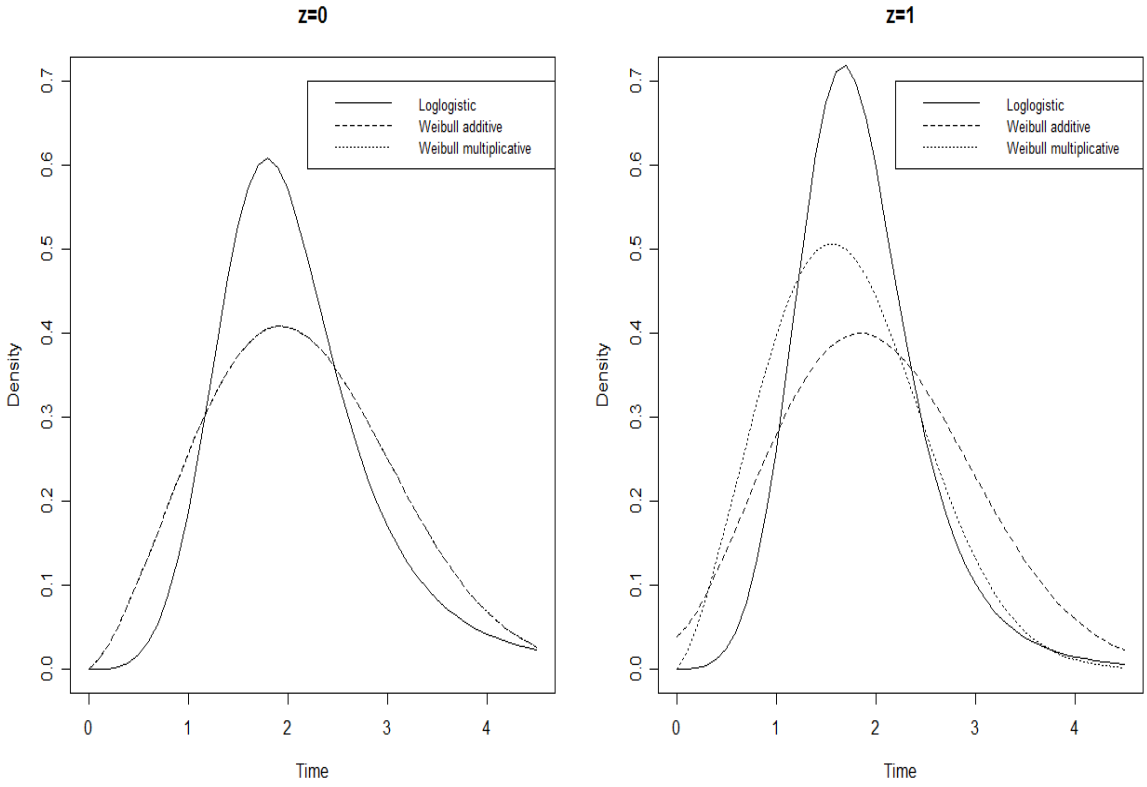

Figure 1: Density functions of the Loglogistic, Weibull regression and Additive regression model.

3.3 Results of the simulation study

To compare the estimated marginal survival functions and the estimated copula parameters obtained by the parametric, semi-parametric and nonparametric two-stage estimation approaches we consider:

(i) a box plot of $\widehat{\beta^l}$ *and* $\widehat{\theta^l}$ for the three approaches, where $\widehat{\beta^l}$ is the estimate of β for the lth data set and where $\widehat{\theta^l}$ is the estimate of θ for the lth data set, for $l = 1, \ldots,$ 1000. We only show the boxplots for $\theta = 0.5$ and $K = 100$ and $K = 500$ in this section. Similar plots can be made for the other values of θ.

(ii) the estimated bias and the estimated mean squared error (MSE) for the copula parameter are obtained as follows: estimated bias = $L^{-1} \sum\theta^{(i)} - \theta_0$ and estimated

MSE = $L^{-1} \sum \{\theta^{(i)} - \theta_0\}^2$, where θ_0 is the true value. A MSE close to zero means that the estimator $\hat{\theta}$ predicts the parameter θ with approximately perfect accuracy. The method with the smallest estimated bias and MSE is generally better.
The computations and the Figures were programmed in R Version 2.7 and in SAS Version 9.1.

3.3.1 Results for the multiplicative Weibull regression model

Table 1 and Table 2 present the estimated bias and the estimated MSE of the estimators for the association parameter obtained by the parametric, semi-parametric and nonparametric two-stage estimation approaches when the marginal survival functions are correctly specified.

From Table 1, we can observe that for all values θ except $\theta = 0.5$ if $K = 500$, the estimated absolute bias of $\hat{\theta}$ is smallest for parametric approach. In general, the absolute values for the estimated bias of $\hat{\theta}$ are larger for the nonparametric approach as compared to the semi-parametric approach. Also we observe that for $\theta = 8$, the estimated absolute bias of $\hat{\theta}$ for the three approaches are negative whereas for $\theta = 0.5$, the absolute estimated bias of $\hat{\theta}$ is positive for the three approaches. This means that most of 1000 estimated association parameters θ for the three approaches are less than the true value $\theta = 8$ and larger than the true value $\theta = 0.5$. We can conclude that the parametric approach performs better than the other two approaches if the marginal survival function is correctly specified whereas the semi-parametric and the nonparametric approaches give similar results.

Table 2 shows that the estimated MSE of $\hat{\theta}$ is smallest for the parametric approach, except for $\theta = 2$ if the number of clusters is 100. The semi-parametric and nonparametric approaches provide similar values for the estimated MSE of $\hat{\theta}$. From Table 1 and 2, if we increase the number of clusters from 100 to 500 for a particular value of θ, the estimated absolute bias and estimated MSE of $\hat{\theta}$ decrease. This means that if the number of clusters increases, the estimated values of the copula parameter θ are closer to the true value. Also if we increase the values of θ for a particular value of K, the estimated absolute bias and estimated MSE of $\hat{\theta}$ increase. In general, the

parametric approach is considerably better than the semi-parametric and the nonparametric approaches.

Table 1: Estimated bias of the association parameter θ, for three approaches when marginal distributions are correctly specified.

	$K = 100$			$K = 500$		
$\tau(\theta)$	0.20	0.50	0.80	0.20	0.50	0,80
Θ	0.50	2.00	8.00	0.50	2.00	8.00
Parametric	0.000224	0.007178	-0.577586	0.002565	-0.011945	-0.097097
Semi-parametric	0.001762	-0.049520	-1.489570	0.001134	-0.032438	-0.519660
Nonparametric	0.0067624	-0.057544	-1.509004	0.002766	-0.035616	-0.524235

Table 2: Estimated MSE of the association parameter θ, for three approaches when marginal distributions are correctly specified.

		$K = 100$			$K = 500$	
$\tau(\theta)$	0.20	0.50	0.80	0.20	0.50	0,80
θ	0.50	2.00	8.00	0.50	2.00	8.00
Parametric	0.046290	0.185942	2.170887	0.008945	0.041287	0.402654
Semi-parametric	0.047156	0.181159	3.583578	0.009101	0.041823	0.639727
Nonparametric	0.048608	0.184787	3.710306	0.009147	0.042361	0.649917

Figure 2 shows that the median of the estimated values for β is slightly larger for K = 500 than K = 100. The estimated median of β is approximately equal for the parametric and semi-parametric approaches, both for K = 100 and K = 500. The estimated median and the estimated mean of β are equal for K = 500 whereas for K = 100, the estimated mean of β is greater than the estimated median of β. The boxplots of the estimated values of β for the two approaches indicate that the interquartile range is about 0.20 for K = 100 whereas for K = 500, the interquartile range is smaller (0.10). This implies that the interquartile range decreases as the number of clusters increases from 100 to 500. Finally, for the number of cluster 500, the distributions of

the estimated values of β are symmetric and the estimated values of β are closer to the true value $\beta = 0.522$.

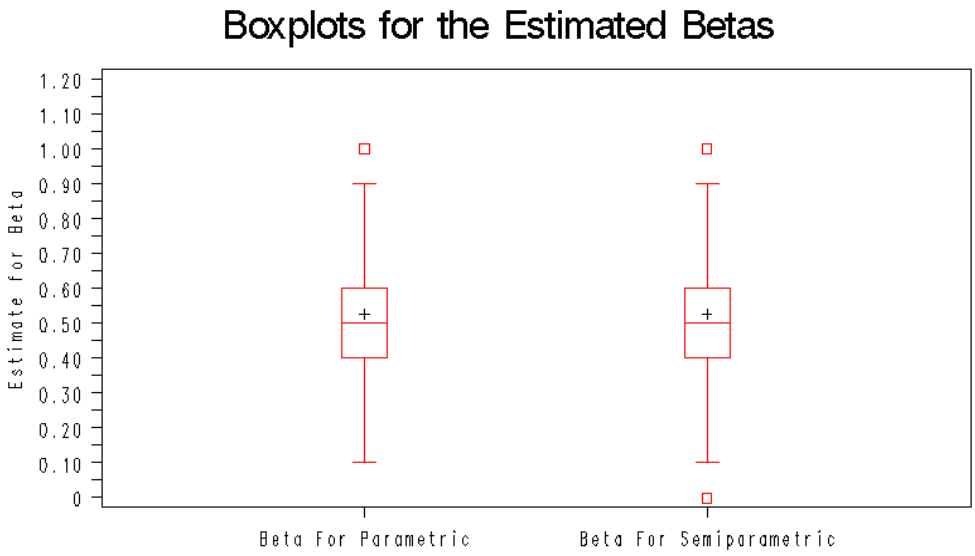

Figure 2a: Box plots for estimated β for the θ=0.5 and K=100 clusters.

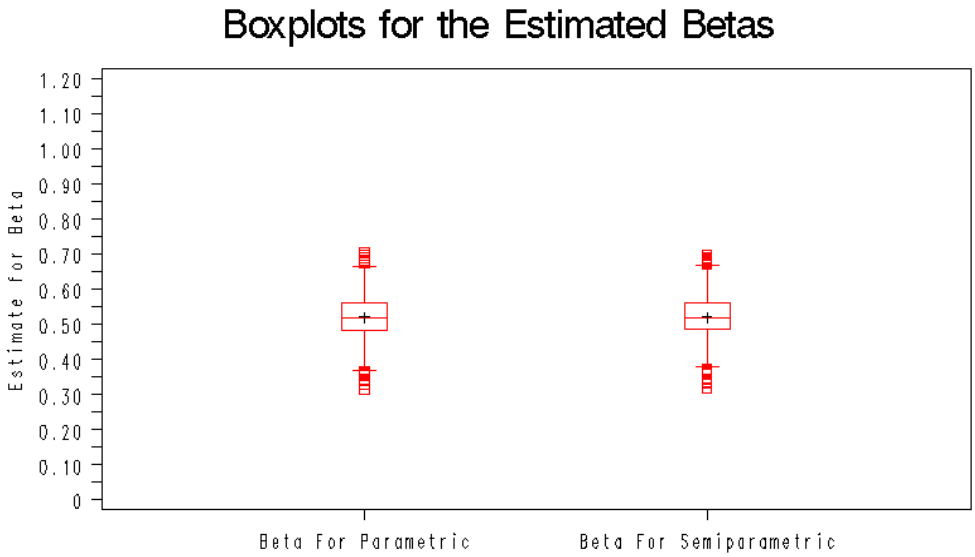

Figure 2b: Box plots for estimated β for the θ=0.5 and K=500 clusters.

From Figure 3, we can observe that the estimated median of θ for the three approaches is about 0.52 and 0.50 for K = 100 and 500, respectively. Many outliers are also observed in all approaches for K = 100 whereas for K = 500, few outliers are observed. The boxplots of the estimated values of θ for the three approaches indicate that the interquartile range decreases as the number of clusters increases from 100 to

500. In general, from the boxplots we observed that if we increase the number of clusters from 100 to 500, the estimated mean of θ is closer to the true value 0.5 and the distributions of the estimates values of θ are symmetric for all approaches.

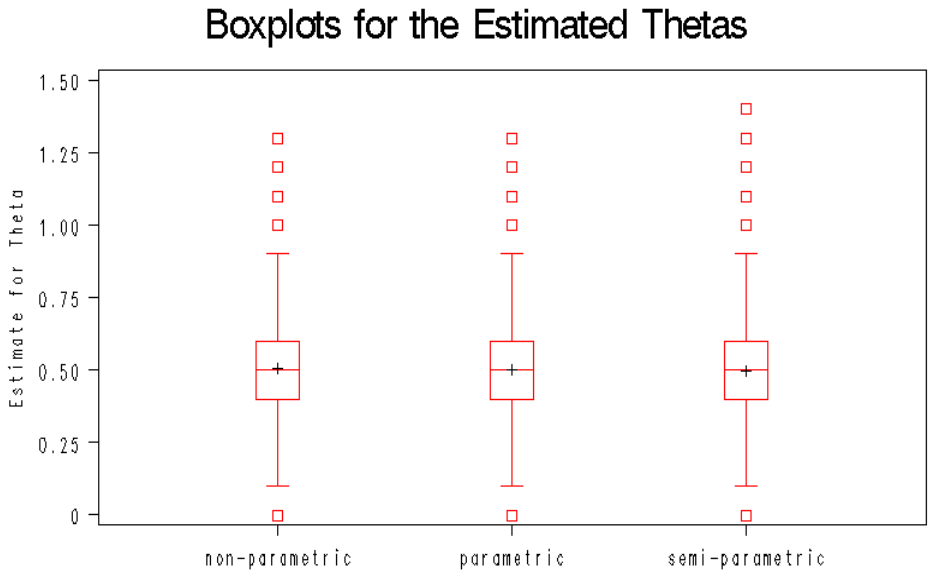

Figure 3a: Box plots for estimated θ with the value θ=0.5 and K=100 clusters.

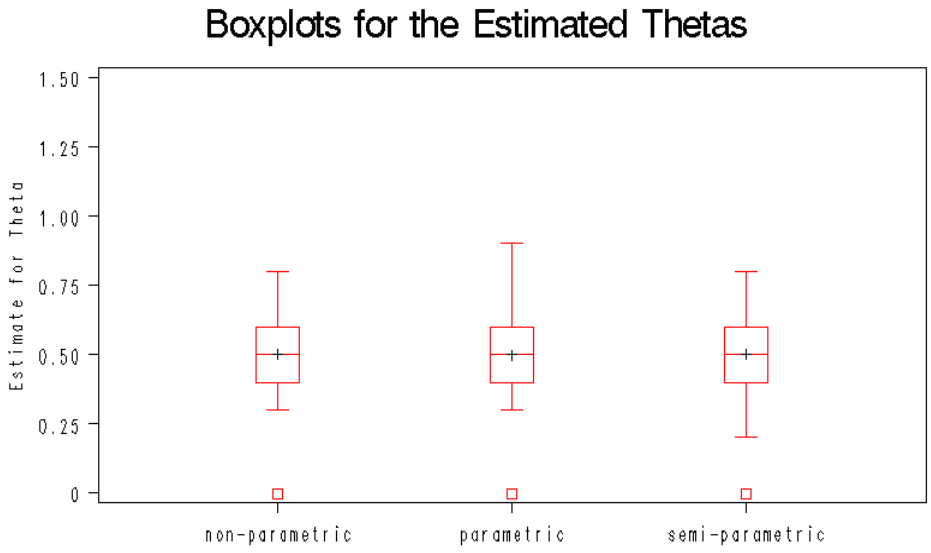

Figure 3b: Box plots for estimated θ with the value θ=0.5 and K=500 clusters.

3.3.2 Results for the multiplicative loglogistic regression model

Table 3 and Table 4 present the estimated bias and the estimated MSE of the estimators for the association parameter obtained by the parametric, semi-parametric and nonparametric two-stage estimation approaches when the marginal survival functions are misspecified.

From Table 3, we can observe that for all values of θ except $\theta = 0.5$ the estimated absolute bias of $\hat{\theta}$ is smallest for the parametric approach and largest for the semi-parametric approach whereas for $\theta = 0.5$, the estimated absolute bias of $\hat{\theta}$ is smallest for the semi-parametric approach both for $K = 100$ and $K = 500$. Also it was observed that, except for $\theta = 0.5$, all the estimated biases of $\hat{\theta}$ are negative for the three approaches. This means that except for $\theta = 0.5$, most of 1000 estimated values of θ for the three approaches are smaller than the true value of θ.

Table 4 shows that for all values of θ except for $\theta = 0.5$, the estimated MSE of $\hat{\theta}$ is smallest for the parametric approach and largest for the semi-parametric approach whereas for $\theta = 0.5$, the estimated MSE of $\hat{\theta}$ is smallest for the semi-parametric approach both for $K = 100$ and $K = 500$. Further more, from Table 3 and 4, if we increase the number of clusters from 100 to 500 for a particular value of θ, the estimated absolute bias and estimated MSE of $\hat{\theta}$ decrease. This means that if the number of clusters increases, the estimated values of the copula parameter θ are closer to the true value. Also if we increase the values of θ for a particular value of K, the estimated absolute bias and estimated MSE of $\hat{\theta}$ increase. Finally, we can conclude that, the parametric approach performs well compared to the nonparametric and semi-parametric approaches.

Table 3: Estimated bias of the association parameter θ, for three approaches when marginal distributions are misspecified (loglogistic).

	$K = 100$			$K = 500$		
$\tau(\theta)$	0.2	0.5	0.8	0.2	0.5	0.8
θ	0.5	2	8	0.5	2	8
Parametric	0.036517	-1.465873	-7.458529	0.036684	-1.459934	-7.44575
Semi-parametric	0.003195	-1.503504	-7.497158	-0.00190	-1.49997	-7.486951
Nonparametric	0.013563	-1.492388	-7.487823	0.008436	-1.489482	-7.477714

Table 4: Estimated MSE of the association parameter θ, for three approaches when marginal distributions are misspecified (loglogistic).

	$K = 100$			$K = 500$		
$\tau(\theta)$	0.2	0.5	0.8	0.2	0.5	0.8
θ	0.5	2	8	0.5	2	8
Parametric	0.042196	2.190054	55.66938	0.042507	2.173872	55.47872
Semi-parametric	0.041045	2.301334	56.24469	0.041047	2.291349	56.09066
Nonparametric	0.042657	2.269198	56.10715	0.041693	2.260499	55.95368

From Figure 4, it was observed that for the two approaches the estimated mean of β is smaller than the estimated median of β both for $K = 100$ and $K = 500$. The estimated mean of β for the semi-parametric approach is about 0.37 which is slightly larger than the estimated mean of β for the parametric approach both for $K = 100$ and $K = 500$. The boxplots of the estimated values of β for the two approaches indicate that the interquartile range is approximately 0.10, both for $K = 100$ and $K = 500$.

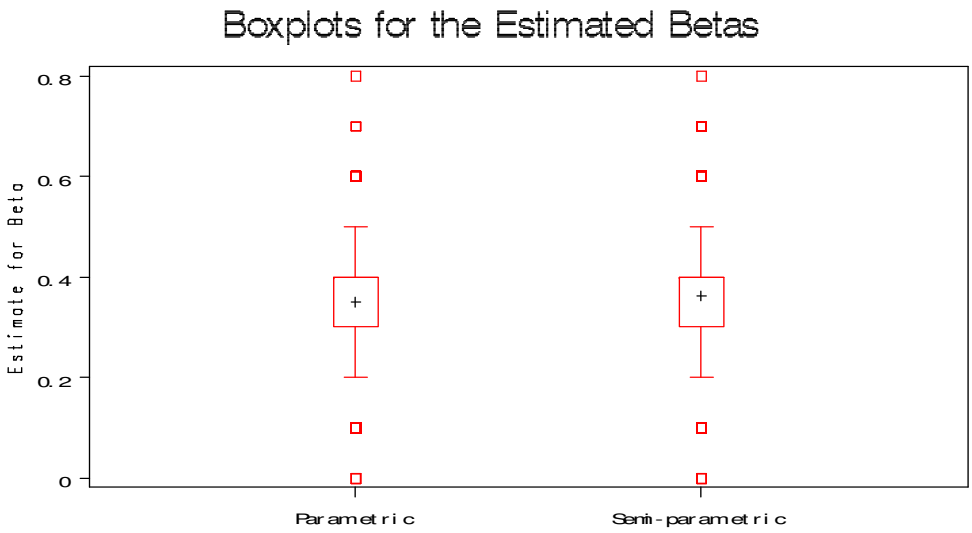

Figure 4a: Box plots for estimated β with the value θ = 0.5 and K = 100 clusters.

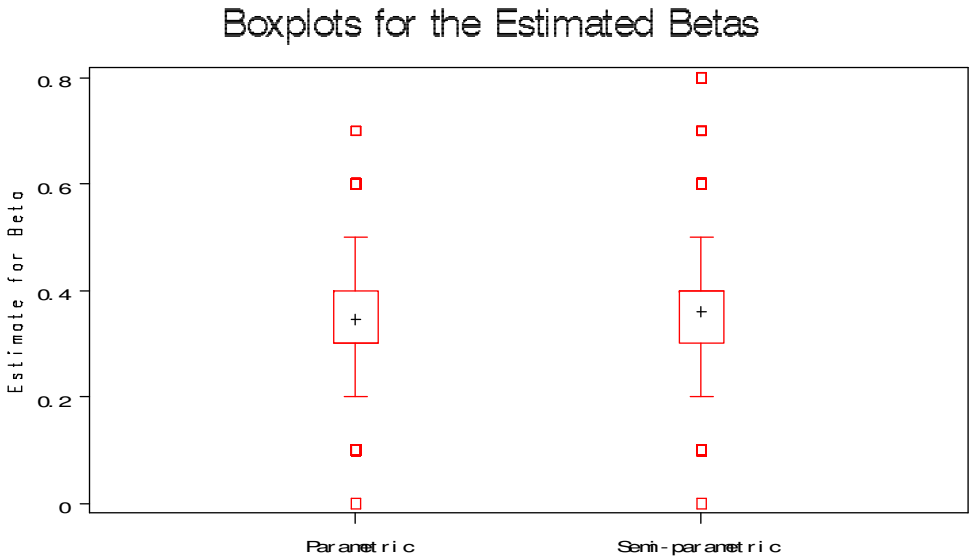

Figure 4b: Box plots for estimated β with the value θ = 0.5 and K = 500 clusters.

Figure 5 shows that the estimated mean and the estimated median of θ are approximately equal for the nonparametric and the semi-parametric approaches whereas for the parametric approach the estimated mean of θ is larger than the estimated median of θ both for $K = 100$ and $K = 500$. The boxplots of the estimated values of θ indicate that the interquartile range is about 0.25 for the nonparametric and the semi-parametric approaches whereas for the parametric approach the interquartile range is about 0.32 both for $K = 100$ and $K = 500$.

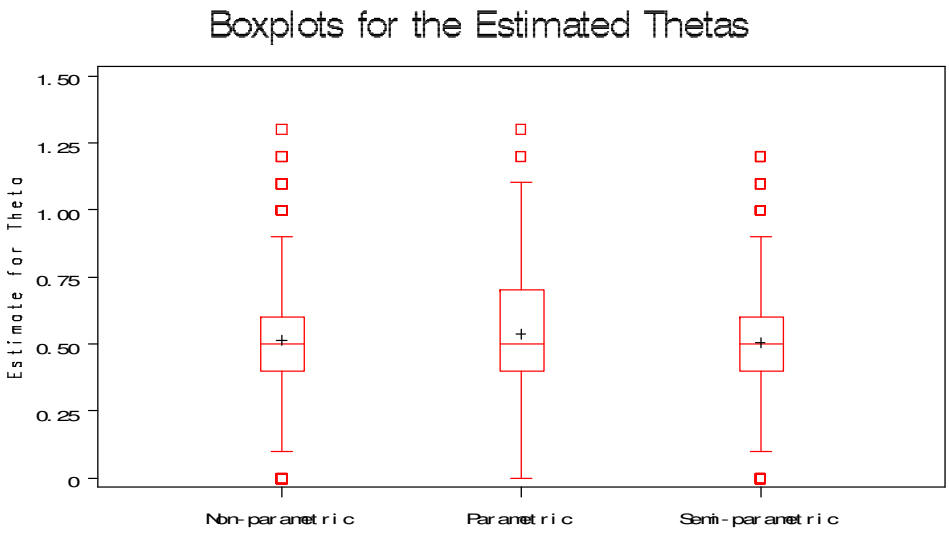

Figure 5a: Box plots for estimated θ *with the value θ=0.5 and K=100 clusters.*

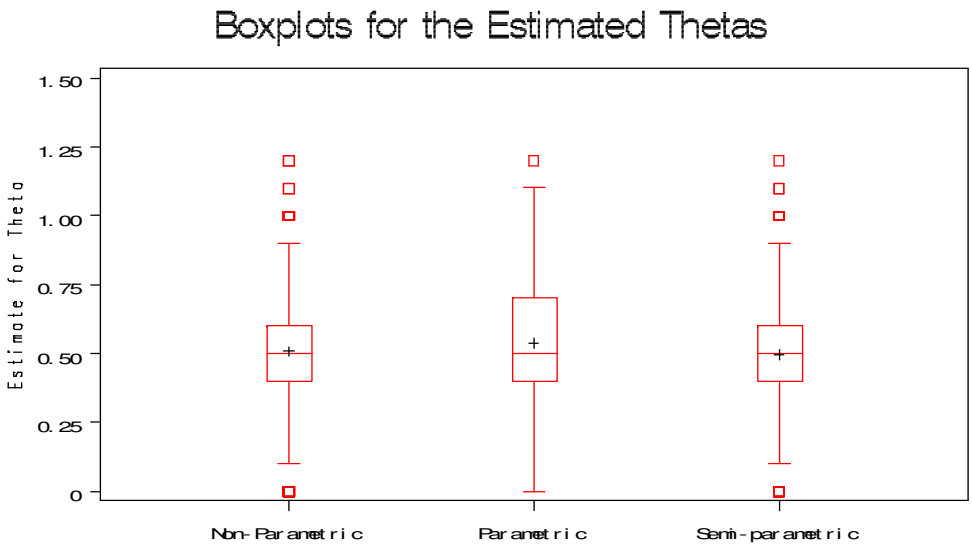

Figure 5b: Box plots for estimated θ with the value θ=0.5 and K=500 clusters.

3.3.3 Results for additive regression model with Weibull baseline hazard

Table 5 and Table 6 present the estimated bias and the estimated MSE of the estimators for the association parameter obtained by the parametric, semi-parametric and nonparametric two-stage estimation approaches when the marginal survival functions are misspecified.

From Table 5, we can observe that for all values of θ except for $\theta = 0.5$, the estimated absolute bias of $\hat{\theta}$ is smallest for the parametric approach. Table 6 shows that for large values of θ, the estimated MSE of $\hat{\theta}$ is smallest for the parametric approach whereas for $\theta = 0.5$, the estimated MSE of $\hat{\theta}$ is smallest for the semi-parametric approach. The estimated MSEs are similar for the semi-parametric and the nonparametric approaches. Finally, Table 5 and 6 indicate that the estimated absolute bias and estimated MSE of $\hat{\theta}$ decrease if the number of clusters increases from 100 to 500 for a particular value of θ. This means that as the number of clusters increases, the estimated values of the copula parameter θ are closer to the true value. Also if we increase the values of θ for a particular value of K, the estimated absolute bias and estimated MSE of $\hat{\theta}$ increase.

In general, we can conclude that the parametric approach performs better than the other two approaches. This may be due to the fact that, when $z_{ij} = 0$, the marginal survival function reduces to the multiplicative Weibull regression model. Hence, we have only misspecification when the covariate z_{ij} is equal to 1.

Table 5: Estimated bias of the association parameter θ, for three approaches when marginal distributions are misspecified (additive Weibull).

	$K = 100$			$K = 500$		
$\tau(\theta)$	0.2	0.5	0.8	0.2	0.5	0.8
Θ	0.5	2	8	0.5	2	8
Parametric	0.026194	0.041769	-0.718988	0.012919	0.018030	-0.336448
Semi-parametric	0.001187	-0.089611	-1.410522	-0.007847	-0.041654	-0.527627
Nonparametric	0.015604	-0.042229	-1.486290	-0.004010	-0.029579	-0.556186

Table 6: Estimated MSE of the association parameter θ, for three approaches when marginal distributions are misspecified (additive Weibull).

	$K = 100$			$K = 500$		
$\tau(\theta)$	0.2	0.5	0.8	0.2	0.5	0.8
Θ	0.5	2	8	0.5	2	8
Parametric	0.048642	0.209617	2.330342	0.010051	0.040862	0.498952
Semi-parametric	0.046031	0.194920	3.556593	0.009529	0.042627	0.654911
Nonparametric	0.049764	0.202136	3.883541	0.009569	0.042385	0.698909

From Figure 6, we can say that for the two approaches, the estimated mean of β is approximately equal for the parametric and the semi-parametric approaches both for $K = 100$ and $K = 500$. Also negative value of outlier is observed for $K = 100$, whereas for $K = 500$, positive outliers are observed for the both approaches. The boxplots of the estimated values of β for the two approaches indicate that the interquartile range is about 0.20 for $K = 100$ whereas for $K = 500$, the interquartile range is smaller (0.10).

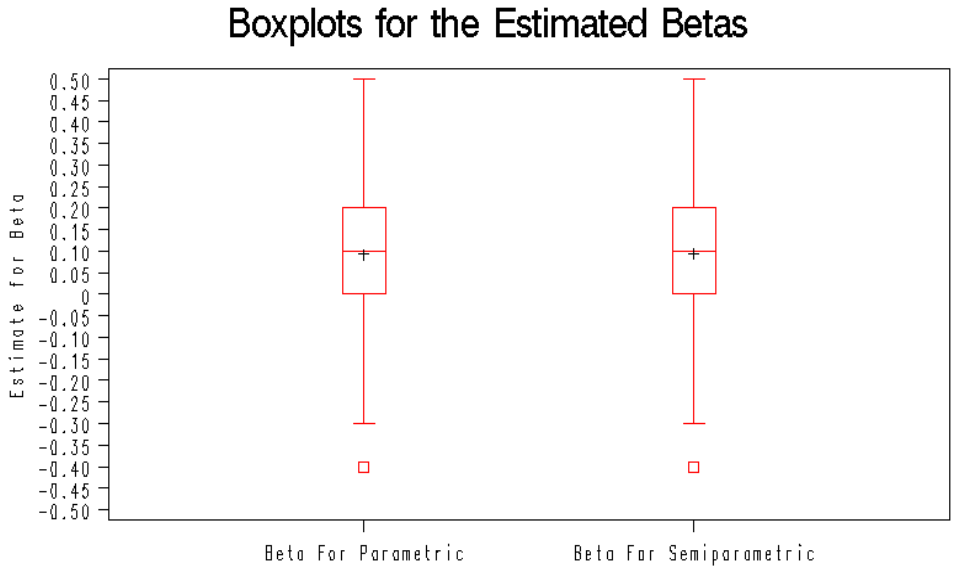

Figure 6a: Box plots for estimated β with the value θ=0.5 and K=100 clusters.

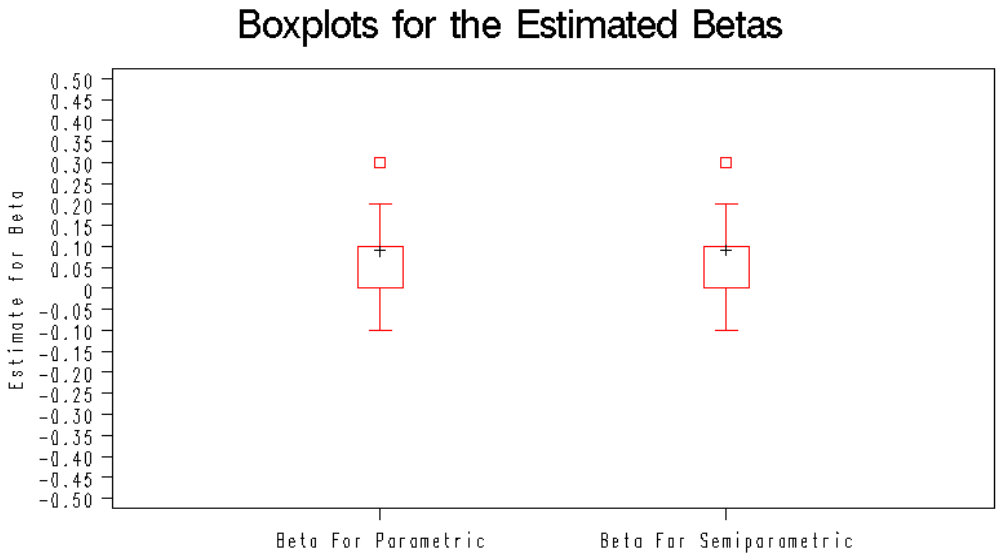

Figure 6b: Box plots for estimated β with the value θ=0.5 and K=500 clusters.

From Figure 7, it was observed that the estimated mean and estimated median of θ are approximately equal for the nonparametric and the semi-parametric approaches both for K = 100 and K = 500 whereas for the parametric approach the estimated mean of θ is larger than the estimated median of θ both for K = 100 and K = 500. Many outliers are also observed in all approaches for K = 100 whereas for K = 500, few outliers are observed. The boxplots of the estimated values of θ for the three approaches indicate that the interquartile range decreases as the number of clusters increases from 100 to 500.

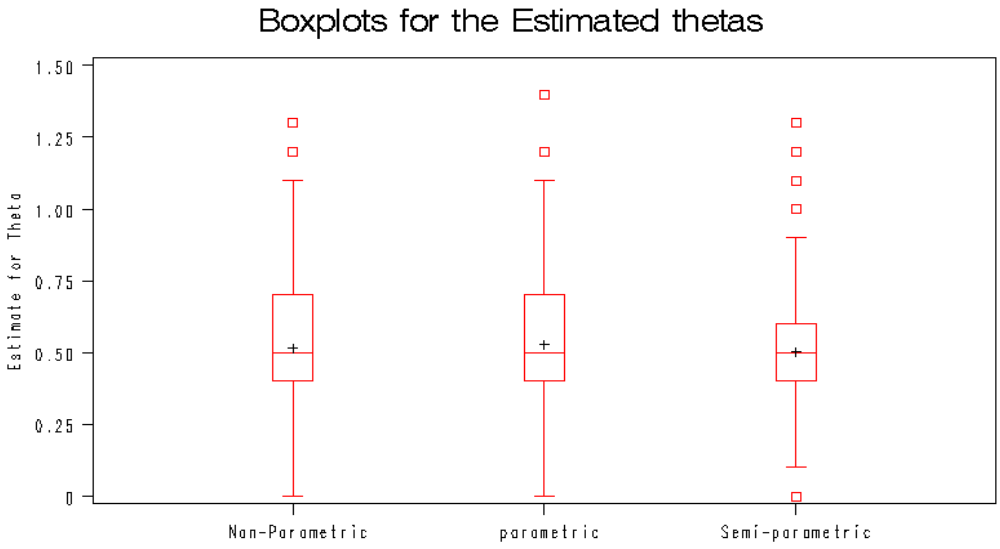

Figure 7a: Box plots for estimated θ with the value θ=0.5 and K=100 clusters.

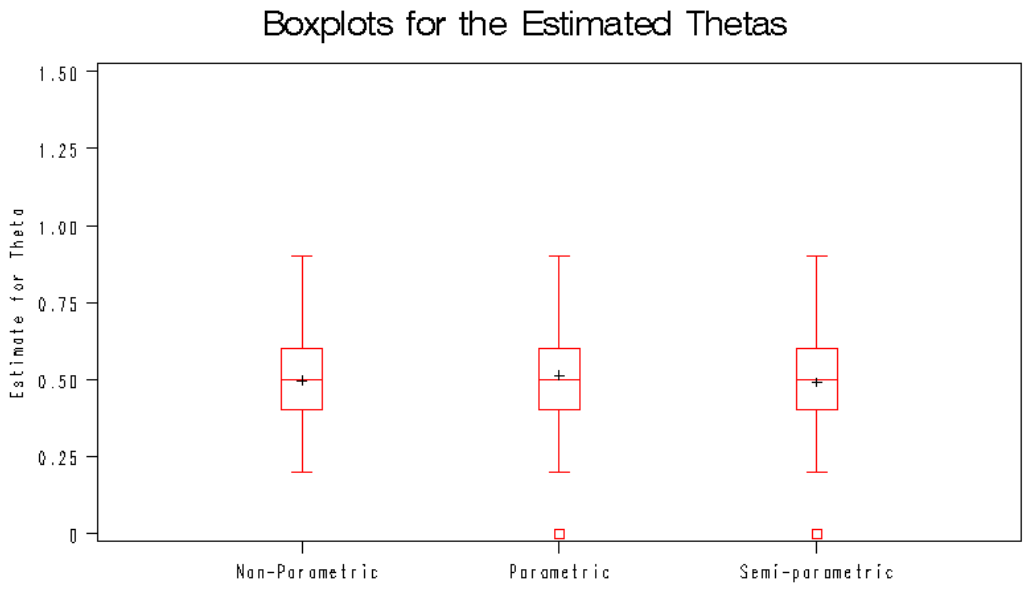

Figure 7b: Box plots for estimated θ with the value θ=0.5 and K=500 clusters.

3.3.4 Model comparison

We compare the performance of the estimation methods for the three model assumptions made to generate the data. If the data are generated assigning a multiplicative Weibull regression model (correct model), it was observed that the estimated absolute bias and the estimated MSE of $\hat{\theta}$ are smallest for the parametric approach. If the data are generated assigning a misspecified models (an additive regression with Weibull baseline hazard or multiplicative loglogistic regression model), it was observed that for all values of θ except $\theta = 0.5$, the estimated absolute bias and the estimated MSE of $\hat{\theta}$ are smallest for the parametric approach. For the multiplicative Weibull regression model, the estimated absolute bias and the estimated MSE of $\hat{\theta}$ are largest for the nonparametric approach whereas for the multiplicative loglogistic regression model, the estimated absolute bias and the estimated MSE of $\hat{\theta}$ are largest for the semi-parametric approach. If we increase the number of clusters from 100 to 500 for a particular value of θ, the estimated absolute bias and the estimated MSE of $\hat{\theta}$ decrease for the three models. And if we increase

the true value of θ for a particular value of K, the estimated absolute bias and the estimated MSE of $\hat{\theta}$ increase for the three models.

For the multiplicative Weibull regression model (correct model), the distributions of the estimated values of β are symmetric and the estimated values of β are closer to the true value if the number of clusters increases from 100 to 500 whereas for the misspecified models, the distributions of the estimated values of β are asymmetric and the estimated values of β are not closer to the true values if the number of clusters increases from 100 to 500. For the multiplicative Weibull regression model and additive regression model with Weibull baseline hazard, the boxplots of the estimated values of θ for the three approaches indicate that the interquartile range decreases as the number of clusters increases from 100 to 500 whereas for multiplicative loglogistic regression model, the interquartile range is constant both for $K = 100$ and $K = 500$.

4. Discussion and Conclusion

This study focused on the comparison of parametric, semi-parametric and nonparametric two-stage estimation methods for copula models. A nice property of copulas is that the marginal survival functions do not depend on the parameters of copula. Therefore, the marginal survival functions and the dependence parameter vector can be estimated separately.

In the first stage, we estimate the marginal survival functions using parametric and semi-parametric approaches proposed by Andersen (2005), and using a nonparametric approach. In the second stage, the dependence parameter is estimated by maximizing the loglikelihood, with the marginal survival functions replaced by their estimates. To compare the performance of the parametric, semi-parametric and the nonparametric estimation approaches, we compute the estimated biases and the estimated MSEs for the estimators of the copula parameters based on 1000 simulated data sets.

The estimated absolute bias and the estimated MSE of the estimator for the copula parameter are smallest for the parametric approach for the three models, except for θ = 0.5. For the multiplicative Weibull regression model, the estimated absolute bias and the estimated MSE of $\hat{\theta}$ are largest for the nonparametric approach whereas for the multiplicative loglogistic regression model, the estimated bias and the estimated MSE of $\hat{\theta}$ are largest for the semi-parametric approach. For the misspecified models (an additive regression with Weibull baseline hazard and multiplicative loglogistic regression model), the estimated bias and the estimated MSE of the copula parameter are larger for parametric approach when $\theta = 0.5$. This implies the parametric approach performs worse than the nonparametric and the semi-parametric approach when θ = 0.5 for the misspecified models. If we increase the number of clusters from 100 to 500 for a particular value of θ, the estimated absolute bias and the estimated MSE of $\hat{\theta}$ decrease for all simulation settings and for the three models. For the multiplicative Weibull regression model (correct model), the distributions of the estimated values of β are symmetric and the estimated values of β are closer to the true value if the number of clusters increases from 100 to 500 whereas for the misspecified models, the distributions of the estimated values of β are asymmetric and the estimated values

of β are not closer to the true values if the number of clusters increases from 100 to 500.

In conclusion, for the correct model, estimated bias of the parameter estimators of the marginal survival functions is small whereas for the misspecified models, the parameter estimators of the survival functions are more biased. In estimating the copula parameter, the parametric approach is more accurate for the correct model and for the misspecified models for large values of θ.

The following aspect is recommended for further research. We considered 100 and 500 clusters, θ*=0.5, 2 or 8,* and about 20% of censoring. It might be interesting to consider values of θ which are smaller than 0.5 and to vary the percentages of censored observations to obtain conclusions in a more general setting.

References

1. Andersen, E.W. (2005). Two-Stage Estimation in Copula models used in Family Studies, Lifetime Data Analysis 11,333-350.

2. Clayton, D.G. (1978). A model for association bivariate tables and its application in epidemiological studies of family tendency in chronic disease incidence, Biometrika, 65, 141-151.

3. Duchateau, L. and Janssen, P. (2008). The Frailty Model. Springer, New York.

4. Genest, C., and Rivest, L.-P. (1993). Statistical inference procedures for bivariate Archimedean copulas. J. Amer. Statist. Assoc. 88, 423, 1034-1043.

5. Glidden, D. (2000). A Two-Stage Estimator of the Dependence Parameter for the Clayton-Oakes Model, Lifetime Data Analysis 6,141-156.

6. Mario R. Melchiori (2003). Which Archimedean Copula is the right one? Universidad Nacional del Litoral Argentina.

7. Massonnet, G. (2008). Contributions to frailty and copula modeling with applications to clinical trails and dairy cows data. PhD thesis, Hasselt University.

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

9. Shih, J.H. and Louis, T.A. (1995). Inferences on the Association Parameter in Copula Models for Bivariate Survival Data, Biometrics 51, 1384-1399.

10. Shemyakin, A. and Youn, H. (2000). Statistical aspects of joint-life insurance pricing, 1999 Proceedings of Amer. Stat. Assoc., 34-38.

Appendix

I. Contributions to the likelihood

When the two observations are censored, the contribution to the likelihood is given by

$$S_i(X_{i1}, X_{i2}) = (S_{i1}{}^{-\theta}(X_{i1}) + S_{i2}{}^{-\theta}(X_{i2}) - 1)^{\frac{-1}{\theta}}$$

When the two observations are events, the contribution to the likelihood is given by

$$f(\mathrm{x_{i1}}, \mathrm{x_{i2}}) = \frac{\partial^2 S(x_{i1}, x_{i2})}{\partial x_{i1} \partial x_{i2}} = \frac{\partial^2 C(u,v)}{\partial u \, \partial v} f(x_{i1}) f(x_{i2})$$

$$= c_\theta \ (u,v) f(x_{i1}) f(x_{i2})$$

$$= (1+\theta)(S_1^{-\theta}(x_{i1}) + S_2^{-\theta}(x_{i2}) - 1)^{-1/\theta - 2} \quad S_1^{-\theta-1}(x_{i1}) S_2^{-\theta-1}(x_{i2})$$

where c_θ is the copula density.

II. Algorithm to generate pairs (u_1, u_2) from a Clayton copula with association parameter θ

1. Simulate v_1 from U [0; 1]

Take $u_1 = v_1$

2. Simulate v_2 from U [0; 1]

Set

$$v_2 = \frac{d}{du_1} C_\theta(u_1, u_2) \Big|_{u_1} = \frac{d}{du_1} \left[(u_1{}^{-\theta} + u_2{}^{-\theta} - 1)^{-1/\theta} \right] \Big|_{u_1}$$

$$= -\frac{1}{\theta}(u_{11}^{-\theta} + u_2^{-\theta} - 1)^{\frac{-1}{\theta}-1}(-\theta u_{11}^{-\theta-1})$$

$$v_{21} = u_{11}^{-\theta-1}(u_{11}^{-\theta} + u_2^{-\theta} - 1)^{\frac{-1}{\theta}-1}$$

Then solve for u_2 we get,

$$u_2 = \left(\left(\frac{v_2}{u_1^{-\theta-1}}\right)^{-\frac{\theta}{(1+\theta)}} + 1 - u_1^{-\theta}\right)^{-1/\theta} = \left\{1 + u_1^{-\theta}(v_2^{-\frac{\theta}{\theta+1}} - 1)\right\}^{\frac{-1}{\theta}}$$

Thus we have (u_1, u_2). Repeat step 1 and 2, K times to generate pairs (u_{i1}, u_{i2}), for i = 1, . . . , K, using the above algorithm. Where K is number of clusters.

III. The means, densities and variances of the three models are given as follows:

1. Weibull regression model:

The hazard and marginal survival functions of this model is

$$\lambda_{ij}(t) = \lambda\rho t^{\rho-1}\exp(\beta z_{ij})$$
$$S_{ij}(t) = \left\{\exp(-\lambda t^{\rho})\right\}^{\exp(\beta z_{ij})}$$

The density function is

$$f_{ij}(t) = \lambda\rho t^{\rho-1}\exp(-\lambda t^{\rho})\exp(\beta z_{ij})\left\{\exp(-\lambda t^{\rho})\right\}^{\exp(\beta z_{ij})-1}$$

The first and the second moments are given by:

$$E(T_{ij}) = \Gamma(1+\frac{1}{\rho})\left\{\frac{1}{\lambda\exp(\beta z_{ij})}\right\}^{1/\rho}$$

$$E(T_{ij}^2) = \Gamma(\frac{2}{\rho}+1)\left\{\frac{1}{\lambda\exp(\beta z_{ij})}\right\}^{2/\rho}$$

The variance is

$$Var(T_{ij}) = \Gamma(1+\frac{1}{\rho})\{\ \frac{1}{\lambda\exp(\beta z_{ij})}\ \}^{1/\rho} - \left[\Gamma(\frac{2}{\rho}+1)\{\ \frac{1}{\lambda\exp(\beta z_{ij})}\ \}^{2/\rho}\right]^2$$

2. Loglogistic regression model:

The density function is

$$f_{ij}(t) = e^{\alpha}kt^{k-1}\{\ \frac{1}{1+e^{\alpha}t^{k}}\ \}^{\exp(\beta z_{ij})+1}\exp(\beta z_{ij})$$

The first and the second moments are given by:

$$E(T_{ij}) = \int_0^{+\infty} e^{\alpha}kt^{k}\{\ \frac{1}{1+e^{\alpha}t^{k}}\ \}^{\exp(\beta z_{ij})+1}\exp(\beta z_{ij})\,dt$$

$$(= e^{-\alpha/k}\,\frac{\pi}{k}\frac{1}{\sin(\frac{\pi}{k})} \quad if\ z_{ij}=0\ and\ k>1)$$

$$E(T_{ij}^2) = \int_0^{+\infty} e^{\alpha}kt^{k+1}\{\ \frac{1}{1+e^{\alpha}t^{k}}\ \}^{\exp(\beta z_{ij})+1}\exp(\beta z_{ij})\,dt$$

The variance is given by:

$$Var(T_{ij}) = E(T_{ij}^2) - \left\{E(T_{ij})\right\}^2$$

3. Additive regression model with Weibull baseline hazard:

The density function is

$$f_{ij}(t) = (\lambda\rho t^{\rho-1} + \beta z_{ij})\exp(-\lambda t^{\rho} - \beta z_{ij}t)$$

The first and the second moments are given by:

$$E(T_{ij}) = \int_0^{+\infty} (\lambda\rho t^{\rho} + \beta z_{ij}t)\exp(-\lambda t^{\rho} - \beta z_{ij}t)dt$$

$$E(T_{ij}) = \int_0^{+\infty} (\lambda\rho t^{\rho+1} + \beta z_{ij}t^2)\exp(-\lambda t^{\rho} - \beta z_{ij}t)dt$$

The variance is

$$Var(T_{ij}) = E(T_{ij}^2) - \{\ E(T_{ij})\ \}^2$$