Journal of Statistical Planning and Inference 139 (2009) 3865 – 3877

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

Modelling udder infection data using copula models for quadruples

Goele Massonnet[a], Paul Janssen[a,*], Luc Duchateau[b]

[a] Interuniversity Institute for Biostatistics and statistical Bioinformatics, Hasselt University, Agoralaan 1, B-3590 Diepenbeek, Belgium
[b] Department of Physiology and Biometrics, Ghent University, Ghent, Belgium

ARTICLE INFO

Available online 22 May 2009

MSC:
62N01
62F40

Keywords:
Bootstrap
Copula models
Pseudo-likelihood ratio test
Quadruples
Two-stage estimation

ABSTRACT

We study copula models for correlated infection times in the four udder quarters of dairy cows. Both a semi-parametric and a nonparametric approach are considered to estimate the marginal survival functions, taking into account the effect of a binary udder quarter level covariate. We use a two-stage estimation approach and we briefly discuss the asymptotic behaviour of the estimators obtained in the first and the second stage of the estimation. A pseudo-likelihood ratio test is used to select an appropriate copula from the power variance copula family that describes the association between the outcomes in a cluster. We propose a new bootstrap algorithm to obtain the *p*-value for this test. This bootstrap algorithm also provides estimates for the standard errors of the estimated parameters in the copula. The proposed methods are applied to the udder infection data. A small simulation study for a setting similar to the setting of the udder infection data gives evidence that the proposed method provides a valid approach to select an appropriate copula within the power variance copula family.

1. Introduction

Copulas are often used to model the dependence structure of bivariate survival data. In this paper, we study two-stage estimation and goodness-of-fit for copulas for four-dimensional survival data. Our motivation for this is a data set on the correlated infection times in the four udder quarters in dairy cows.

In copula models the joint survival function of the four event times in a cluster is modelled as a function, called the copula, of the marginal survival functions of the four event times. The copula determines the type of dependence. We consider parametric copulas, 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 association parameter vector can be estimated separately. This is the idea of the two-stage estimation approach (see, e.g., Shih and Louis, 1995a; Glidden, 2000; Andersen, 2005). In the first stage, the marginal survival functions are estimated. We consider both a semi-parametric and a nonparametric approach to estimate the marginal survival functions.

In the semi-parametric approach, we estimate the marginal survival functions using a marginal Cox model with a deterministic binary covariate at the observational unit level. This type of covariate corresponds to the situation we have in the udder infection data (see Section 2). Spiekerman and Lin (1998) prove the consistency and the asymptotic normality of the parameter estimators in the marginal Cox model in the case of stochastic covariates. Since we deal with a deterministic covariate, the proofs in their paper need adaptations.

In the nonparametric approach, observational units having the same covariate value are pooled. We then estimate the marginal survival functions by using a Nelson–Aalen estimator for the pooled data. It is important to note that the pooled observational

[*] Corresponding author. Tel.: +32 11 26 82 36/05; fax: +32 11 26 82 99.
E-mail addresses: goele.massonnet@uhasselt.be (G. Massonnet), paul.janssen@uhasselt.be (P. Janssen), luc.duchateau@ugent.be (L. Duchateau).

0378-3758/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2009.05.025

units are correlated. The consistency and the asymptotic normality results that we obtain for the pooled data are new and extend results in Shih and Louis (1995a).

In the second stage of the estimation, the association parameter vector is estimated by maximising the loglikelihood, with the marginal survival functions replaced by their estimates obtained in the first stage. The proofs of the asymptotic behaviour of the estimator of the association parameter vector use the asymptotic properties of the estimators obtained in the first stage. Further, we study four-dimensional copula models. Our results therefore extend results obtained by Shih and Louis (1995a), Glidden (2000) and Andersen (2005).

To choose a copula model that describes the dependence of the outcomes within a cluster, we start with a large parametric copula family (the power variance copula family). We consider three copula models that are nested in the power variance copula family: Clayton, positive stable and inverse Gaussian copulas. Each of these three copulas model a different type of dependence. We use a pseudo-likelihood ratio test to select a copula in the power variance copula family that provides a good description of the type of dependence between the outcomes in a cluster. This pseudo-likelihood ratio test was also proposed by Glidden (2000) and Andersen (2005). Because the Clayton copula and the positive stable copula are on the boundary of the parameter space, the classical asymptotic theory is not valid, see Andersen (2005). Therefore, we propose a bootstrap algorithm to obtain the p-value for this test. Using this bootstrap algorithm, we also obtain estimates for the standard errors of the estimated parameters in the copula.

Section 2 contains the details on the udder infection data. In Section 3 we summarise some general ideas on copulas and give the precise definitions of important copula models that are nested in the power variance copula family. There we also discuss the relation between copula models and frailty models. Both models provide an appropriate way to describe the within cluster dependence of the outcomes. For small clusters with equal cluster size we prefer copula models above frailty models for different reasons. First, copula models provide a more direct way to describe the dependence between observations within the same cluster (see Section 3.1 for details). Second, fitting frailty models such as, e.g., semi-parametric positive stable or inverse Gaussian frailty models, requires specific software that is not available in the standard statistical software packages. In contrast, the use of the two-stage estimation approach for copula models makes that we can use standard statistical and numerical software.

In Section 4 we obtain the likelihood expression used for the pseudo-likelihood ratio test. We consider a semi-parametric and a nonparametric approach to estimate the marginal survival functions and we give the asymptotic behaviour of the estimators for the marginal survival functions and the cumulative hazard functions as well as for the association parameter vector. The pseudo-likelihood ratio test is discussed in Section 5, where we also propose a bootstrap algorithm that is used to obtain the p-value of the pseudo-likelihood ratio test. In Section 6 we analyse the udder infection data. In Section 7 we give a simulation based evaluation of the type I error and the power of the pseudo-likelihood ratio test in a setting similar to the udder infection data. Finally, main conclusions and possible further extensions are given in Section 8.

2. Udder infection data

We consider a data set on mastitis, an infection of the udder of a dairy cow. Mastitis can be caused by many organisms, most of them are bacteria, such as *Escherichia coli, Streptococcus uberis* and *Staphylococcus aureus*. Since each udder quarter is separated from the three other udder quarters, one quarter might be infected while the other quarters are infection-free. In this study, 100 cows are followed up for infections. From each quarter, a milk sample is taken monthly and is screened for the presence of different bacteria. Due to the periodic follow up, the infection time is defined as the average of the time of the last milk sample that indicates that there is no infection and the time of the first milk sample that indicates an infection. Observations can be right censored if no infection occurs before the end of the lactation period, which is roughly 300 days but different for every cow, or if the cow is lost to follow up during the study, for example due to culling. Note that this implies that there is a common censoring time for the four udder quarters of a cow (i.e., for all units in the cluster). We model the time to infection with any bacteria, with cow being the cluster and udder quarter the observational unit within the cluster. The correlation between the infection times of the four udder quarters of a cow is an important parameter to take preventive measures. With high correlation, a lot of attention should be given to the uninfected udder quarters of a cow that has an infected quarter. Further, the difference in teat end condition between front and rear quarters has been put forward to explain the difference in infection status (Adkinson et al., 1993). Therefore, we take into account the effect of the location of the udder quarter (front or rear) in the analysis. The covariate indicating the location is a binary covariate at the udder quarter level.

3. Copula models

3.1. Definitions and properties

Let $(T_{i1}, T_{i2}, T_{i3}, T_{i4})$ be a quadruple of failure times of the observational units in cluster i and let $S_{ij},\ j=1,\dots,4$, be the marginal survival function of T_{ij}, where the index ij is used to indicate that the marginal survival function may depend on a covariate z_{ij}. The survival copula is the function that links the marginal survival functions S_{ij} to generate the joint survival function, i.e.,

$$S_i(t_1, t_2, t_3, t_4; \zeta) = C_\zeta\{S_{i1}(t_1), S_{i2}(t_2), S_{i3}(t_3), S_{i4}(t_4)\} \tag{1}$$