**Limburgs Universitair Centrum**

**Faculteit Wetenschappen**

# Random effects survival models applied to animal breeding data

Proefschrift voorgelegd tot het behalen van de graad van
**Doctor in de Wetenschappen, Groep Wiskunde**
aan het Limburgs Universitair Centrum te verdedigen door

**Rosemary Nguti**

Promotor:
Prof. dr. Paul Janssen

2003

# Acknowledgment

This work could not have been accomplished without the assistance and support of numerous persons. I take this opportunity to acknowledge their contributions.

First of all I would like to thank Steve, my dear husband, who is the pillar behind this achievement. Your support and encouragement was crucial to all my efforts.

I am deeply grateful to my supervisor Prof. Paul Janssen, whose support, advice, guidance and encouragement affected the entire work on this thesis. It has been a privilege and pleasure to work with you. I have learnt a lot from our collaboration. Thank you so much.

The parts that constitute this thesis are also in collaboration with several individuals: Gerda Claeskens, John Rowlands, Luc Duchateau and Tomasz Burzykowski. I have gained deep insight from each of you within the specific areas that I have worked with you. Thanks to you all. Special thanks to Dr. Leyden Baker for providing the data and for his contribution into the initiation of this study.

During the course of this work I have also had the opportunity to relate with many individuals. Members of the Biometrics Unit at ILRI, members of CENSTAT, current and past VLIR students. Knowing you all contributed to the pleasant atmosphere in which I worked. To Alice and Armand, I say thanks for your hospitality.

I acknowledge the financial support from the Belgian Technical Cooperation (BTC), the research assistantship support from ILRI, and the support of the Dept. of Mathematics, University of Nairobi.

Last but not the least, to my children, Sheila, Ian and Lillian. To you, I am sorry, for having to be away.

I commit this thesis to God who has carried me through each day.

Rosemary Nguti

November 28, 2003

# Contents

# List of Tables

# List of Figures

# Chapter 1

# Introduction

In the developing sub-Saharan countries, sheep and goats are kept mainly by small scale farmers for meat production. These animals graze on open natural pastures or in communal pastoral systems. Almost 30% of the animals however do not reach to maturity age due to high levels of infections from endoparasites. Although control methods that focus on reducing contamination of pastures through anthelmintic treatment are in place, their use is limited due to high cost. Due to this there is need for animals that are well adapted to the environment so as to increase productivity.

The data considered in this thesis come from a breeding program conducted from 1991-1996 at the International Livestock Research Institute (ILRI). The objective of the program was to assess genetic resistance to endoparasites among the Red Masaai, Dorper and their cross breeds. In each of the years, the lambs were observed for a period of at most 12 months. The genetic resistance of the lambs has been assessed by using linear mixed models on measurements of packed cell volume (PCV), faecel egg count (FEC) and body weight (BWT) collected at arbitrarily defined time points in the animal's life span. It has now been reported (Baker et al., 1994, 1999, 2003) that the Red Maasai has higher resilience (higher PCV) and higher resistance (lower FEC) than the Dorper.

From the experimental set-up, information of any lambs that died during the follow-up time was also available. These time-to-event measurements (time to death) are the main focus of this thesis. The genetic component was considered at the sire level. Thus the times of lambs from the same sire were assumed to be correlated. Such survival times,

fall in the class of multivariate time-to-event data. These are time-to-event data that are correlated within some cluster. For example the times to recurrent trypanosomiases infections for a lamb or the survival times of lambs from the same sire. In the last several years extensive research on multivariate time-to-event data has been carried out (Klein and Moeschberger, 1997, Hougaard, 2000, Therneau and Grambsch, 2000). To account for correlation in the times within a cluster or equivalently heterogeneity between clusters, a random effect term is used, resulting in what are known as frailty models. Models with one random effect per cluster are known as shared frailty models. Most shared frailty models are an extension of the semi-parametric Cox proportional hazard (PH) model (Therneau and Grambsch, 2000). On the other hand the Weibull baseline hazard has been the most widely used form for the parametric shared frailty model (Hougaard, 2000).

Many methods of estimation have been described in the literature to fit semi-parametric shared frailty models. In order to explore the common ground for these methods we carried out an extensive review within the context of a parallel study in a multicenter clinical trial setting (Duchateau et al., 2002). In this setting, the patients within a centre constitute a cluster. The aim of this latter study was to assess the relationship between the size of a multicenter trial, in terms of number of centres and patients per centre through simulations. Also studied was the bias and spread of the estimates of the heterogeneity parameter around its true value, as affected by the size of the trial. Here we report the reviewed methods from this study which constitute the main contents of Chapter 2. This review gives us a solid basis for working with the shared frailty models on the animal breeding data, in the second part of the thesis.

A pertinent question that may arise while using frailty models is whether indeed there is heterogeneity (correlation) among the event times of individuals across clusters. This heterogeneity is measured in terms of the variance of the random effects. Testing for heterogeneity is a non-standard testing problem as the variance parameter is on the boundary of the parameter space under the null hypothesis. Such problems have been studied in the recent past in the area of linear mixed models (Self and Liang, 1987, Stram and Lee, 1994, 1995, Verbeke and Molenberghs, 2003) but not in the context of frailty models. Based on the simulations in Duchateau et al. (2002), we conjectured that the likelihood ratio test for heterogeneity had an asymptotic distribution which was a 50:50 mixture of a point mass at zero and a chi-square distribution with one degree of freedom. In Chapter 3 we give a theoretical proof of this conjecture for a parametric shared frailty model. We also

consider the asymptotic distribution of the score test for this testing problem.

In the second part of the thesis, we give a detailed description of the animal breeding data set in Chapter 4 and also give an overview of the findings that are available from this experiment. We also show the need for more advanced techniques for analysing these data, such as using the shared frailty models. In modelling time-to-event data, the Cox PH model is often the model of choice. For this hazard model, the shape of the underlying population risk function (baseline hazard) is left unspecified. Non-parametric kernel estimation methods of the hazard function can however be used to determine the shape of the hazard function (Tanner and Wong, 1983, Cheng, 1987, Müller and Wang, 1994). In Chapter 5 we employ the method of Müller and Wang (1994). The variable patterns obtained for the estimated hazard in each of the six years do not support the use of any parametric form for the baseline hazard function. Thus semi-parametric shared frailty models are used to analyse the animal breeding data, and we report the findings also in Chapter 5.

From the longitudinal nature of the animal experiment, repeated measurements of PCV, BWT and FEC were also collected for each lamb. These traits were recorded on average within monthly time intervals as long as the lamb was under observation and before it was one year old. These traits were considered as time-varying covariates in the shared frailty model analysis, so as to assess their effect on the risk of mortality as they evolved over time. PCV, BWT and FEC are however informative of the survival of the animal. For instance low PCV , high FEC and low BWT may be associated with sick animals. In the last few years, methodologies that use the information available in both the time-to-event and such informative repeated measurements, have been proposed in medical research (De Grutolla and Tu, 1994, Tsiatis et al., 1995, Faucett and Thomas, 1996, Wulfsohn and Tsiatis, 1997, Henderson et al., 2000, Wang and Taylor, 2001). In Chapter 6 we describe briefly these models and then adapt the joint methodology to induce association between the time-to-event and separate repeated measurements of PCV, BWT and FEC. To this end, the repeated measurements are modelled using linear mixed models while the time-to-event measurements are in turn modelled using the Cox PH model. Association between the two models is induced through a Gaussian latent process that depends on the random effects of the repeated measurements. Thus, the time-to-event component of the joint model, can indeed be viewed as a random effects survival model with subject specific effects which are governed by the evolution of the repeated measurement process.

# Part I

# Methodology

# Chapter 2

# Frailty models for multivariate survival data

## 2.1 Description of time-to-event data

### 2.1.1 Introduction

Time-to-event data arise in studies involving the observation of individuals from some starting point to the principal end point when the event of interest occurs. Some examples are the time from onset of disease to death, the time from HIV infection to AIDS, the duration of strikes or periods of unemployment, the time from recovery to the time of recurrence of disease, the time taken by subjects to complete specified tasks or the lengths of tracks on a photographic plate. Depending on the field of application we speak about survival time in biometrics, failure time in engineering or duration time in economics.
To determine the time to an event precisely, there are three requirements that are necessary: the time of origin must be unambiguously defined, a scale for measuring the passage time must be agreed upon and finally a precise definition of the endpoint is needed. For example, in survival after heart attack, the time of origin could be the onset of illness, time of admission to hospital or time of treatment, thus depicting the need for a well defined time of origin for this study. Here the time in days may be used as the measuring scale for the time to the event of interest (such as death).

The time of origin is not usually at the same time point for each individual, leading to what is often referred to as staggered entry. In such a case, each individuals' time-to-event is measured from the date of entry to the time of the event. On the other hand, clock time is the most frequently used measure of the time-to-an-event, although the choice of a measuring scale depends on the area of application. For example, the number of kilometers covered by a car would be the measuring scale in assessing the reliability of a shock absorber. The need of a well defined end point is necessary to overcome any ambiguity. For example in shock absorber assessment, a precise definition of the endpoint is needed as there can be various modes of failure of the equipment.

A key analytical difficulty that occurs with time-to-event data is the presence of censored observations. Censored data arise when an individual has been followed up only to a certain time point before the event has taken place. For example, an individual may not experience an event within the study period and so the only information available is that the individual has not experienced the event at the maximum follow-up time. This type of censoring is known as right censoring. Thus the event of interest is only observed if it happens prior to some censoring time, otherwise the time is censored in that we know the event has not occurred up to this time. Censored data can also result when an individual is lost to follow-up or withdraws from the study before the event occurs. Most often the censoring process is assumed to be non-informative for time-to-event data. If we define T to be the time to the event and C to be the time at which censoring of an individual occurs then an individual is right censored if $T > C$ and uncensored if $T \leq C$. Non-informative censoring implies that the censoring of an observation does not provide any information regarding the prospects of survival. Informative censoring occurs when the censoring mechanism is related to the survival time. For example, informative dropout occurs if an individual withdraws from a study for reasons which are related to survival time (e.g. illness as may happen in AIDS studies) , or if an individual is lost to follow-up because they feel sufficiently recovered that they do not present for follow-up appointments. For the data sets considered in this work, the censoring mechanism is assumed to be non-informative.

### 2.1.2 Basic quantities

Consider a set of $n$ subjects where the $i^{\text{th}}$ subject, $i = 1, \ldots, n$, is observed from a time zero to a failure time $T_i$ or to a potential right censoring time $C_i$. Let $T_i^o = \min(T_i, C_i)$, be the observed time and $\delta_i$ be the censoring indicator which is equal to 1 if $T_i^o = T_i$ and 0 otherwise. Hence the observed data available for the $i^{\text{th}}$ subject is $(T_i^o, \delta_i)$. The basic analytical quantities for time-to-event data are the survival function

$$S(t) = Pr(T \geq t)$$

which is the probability of surviving beyond time t and the hazard function

$$\lambda(t) = \lim_{\Delta t \to 0^+} \frac{Pr(t \leq T < t + \Delta t | T \geq t)}{h} .$$

This gives the instantaneous failure rate or risk of failure, conditional on having survived up to time $t$. Another closely related basic quantity is the cumulative hazard function, defined as

$$\Lambda(t) = \int_0^t \lambda(u) du.$$

The survival function is often used in determining summary statistics such as the median time to event. A plot of the estimated survival function is a powerful tool for visualizing how the occurrence of the events is distributed over time especially in cases where individuals are grouped. The hazard function on the other hand is used in assessing how measured covariates are related to the risk of the event through a hazard model as discussed in the next section.

### 2.1.3 Modelling time-to-event data

In most time-to-event studies, interest focuses on how the risk of failure is affected by a set of $p$ explanatory variables $x_1, x_2, \ldots, x_p$, measured for all individuals over the course of the study period. The two approaches which have been used are through the effect of the variables on the hazard function (hazard model) or their effect on time (accelerated failure time model). In the former model, which is the most common approach, the hazard is expressed as a product of some baseline hazard and a function that explains how the risk depends on the covariate values. In the latter model, the explanatory variables are assumed to act multiplicatively on the time scale, thus affecting the rate at which an

individual proceeds on the time scale. In this thesis we only focus on hazard models. Let $\lambda(t|\boldsymbol{x}_i)$, $i = 1, \ldots, n$, be the hazard function at time $t$ for the $i^{\text{th}}$ individual with covariates $\boldsymbol{x}_i^T = (x_{i1}, x_{i2}, ....., x_{ip})$ . Then a hazard model has the form

$$\lambda(t|\boldsymbol{x}_i) = \lambda_0(t)\phi(\boldsymbol{x}_i)$$

where $\lambda_0(t)$ is the baseline hazard function, common to all individuals, and $\phi(\boldsymbol{x}_i)$ is a non-negative function of $\boldsymbol{x}_i$. The shape of the baseline hazard $\lambda_0(t)$ can be left unspecified (semi-parametric hazard models) or it may be assumed to have some specific parametric form (parametric hazard models). The most commonly used parametric form for $\lambda_0(t)$ is the Weibull distribution.

Cox (1972) proposed $\phi(\boldsymbol{x}_i) = \exp(\boldsymbol{x}_i^T\boldsymbol{\beta})$ leading to the following hazard model

$$\lambda_i(t) = \lambda_0(t)\exp(\boldsymbol{x}_i^T\boldsymbol{\beta}) \tag{2.1}$$

where $\lambda_i(t) = \lambda(t|\boldsymbol{x}_i)$ and $\lambda_0(t)$ is the unspecified baseline hazard which corresponds to the hazard for an individual whose covariate values are all zero. The vector $\boldsymbol{\beta}$ contains the unknown regression parameters associated with $\boldsymbol{x}_i$, the vector of the explanatory variables. The covariates may all be constant over time (e.g. gender, breed, institution), while some of them may be time-varying (e.g. blood pressure, body-weight), in which case the notation $\boldsymbol{x}_i(t)$ is used. Model (2.1) above is known as the Cox proportional hazards (Cox PH) model and has been used extensively in the last three decades in the analysis of time-to-event data. Its name comes from the fact that the ratio of the hazard for any two individuals (say $i^{\text{th}}$ and $j^{\text{th}}$) which is

$$\frac{\lambda(t|\boldsymbol{x}_i)}{\lambda(t|\boldsymbol{x}_j)} = \exp(\beta_1(x_{i1} - x_{j1}) + \ldots + \beta_p(x_{ip} - x_{jp}))$$

is constant and independent of time $t$ implying that the hazard ratios are proportional over time. Thus the parameter estimates for $\boldsymbol{\beta}$ in Model (2.1) are interpreted as the population average relative risk only if $\boldsymbol{x}$ is fixed over time as the proportionality assumption is violated for time-dependent covariates.

The survival function corresponding to (2.1) is

$$S(t|\boldsymbol{x}_i) = (S_0(t))^{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}$$

where $S_0(t) = \exp\left(-\int_0^t \lambda_0(u)du\right)$ is the survival function for an individual with all the covariates equal to zero.

With no ties among the event times, the estimators of the parameters $\boldsymbol{\beta}$ in (2.1) are obtained by maximizing the partial likelihood introduced by Cox,

$$PL(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left\{ \frac{\exp(\boldsymbol{x}_i^T \boldsymbol{\beta})}{\sum_{j \in R_i} \exp(\boldsymbol{x}_j^T \boldsymbol{\beta})} \right\}^{\delta_i} \tag{2.2}$$

where $R_i$ is the set (risk set) of individuals who have not yet experienced the event at time $t_i$ and are still under observation at that time. Although the partial likelihood is not a full likelihood, the estimators obtained from this maximization have been shown to be both consistent and have asymptotic normal properties (Cox, 1972).

## 2.2 Multivariate survival data

In modelling time-to-event data for a group of individuals the underlying assumption is that the failure times among the individuals are independent. However, this is often not the case, as the failure times may belong to related individuals from some group (twins or items from the same production line) resulting in times that are correlated. When this is the case, then the independence assumption is violated. Time-to-event data that are correlated are often referred to as multivariate survival data (Hougaard, 2000) otherwise they are said to be univariate.

There are two main types of multivariate survival data: *parallel* and *longitudinal* data. Parallel data consist of a number of clusters and each cluster (batch, family, centre) in turn contains several items/individuals. On the other hand, longitudinal data are a result of a stochastic process of events, e.g., the asthma seizures of an individual over time. The cluster is now the individual and within that individual the recurrence of asthma seizures is observed (recurrent data).

In both types of multivariate data, events within a cluster are correlated. The general idea is that there are some unobserved risk factors (e.g. genetic in twin studies) that explain the dependence. These unobserved factors are often assumed to be constant over time for longitudinal data and common between the individuals in a cluster for parallel data. Using standard hazard models such as the Cox PH in the presence of dependence produces biased estimates (Wei et al., 1989).

The main focus of this thesis will be on parallel data from an animal breeding program. In general, the settings of an animal breeding program give a natural source for clustered

data, if the response of interest is the time to an event (examples are death and time to first calving). In this setting the time-to-event for animals of the same father (or family) form a cluster if the aim of the experiment is to assess genetic variability.

## 2.3 The shared frailty model

### 2.3.1 Background

In the last several years, multivariate time-to-event data analysis has been studied extensively (see Klein and Moeschberger, 1997, Hougaard, 2000, Therneau and Grambsch, 2000). To account for the correlation of the failure times within a cluster a random effect term, commonly known as the 'frailty', is included in the hazard model, thus leading to what are known as frailty models. Frailty models are thus analogous to linear mixed effects models with the frailty term acting multiplicatively on the hazard function. These models were introduced by Vaupel (1979). The shared frailty model assumes that all the individuals within the same cluster 'share' the same frailty. Furthermore the frailties are assumed to be independent from cluster to cluster. Thus the underlying concept in these models is that the failure times of individuals in a cluster are dependent, while those across clusters are independent. However, conditional on the frailties, the failure times are independent. For example in a multicenter clinical trial with time-to-event as the response of interest, the times of individuals in each centre are dependent. This dependence can for instance be ascribed to the clinical practice at the centre which has an influence on the outcome. Finally, in the recurrent event setting, the times from the same individual are now assumed to be dependent thus having a common (shared) frailty.
Frailty models have also been considered for univariate data (see Aalen, 1994, Hougaard, 1995). When this is the case, the frailty term is assumed for each individual and is deemed to represent unmeasured covariates. These unmeasured covariates are thought to induce heterogeneity among the individuals after taking into account any measured covariates.
Shared frailty models are a special case of more general frailty models such as the correlated (Petersen, 1998) and multivariate frailty models (Vaida and Xu, 2000 , Ripatti et al., 2002). In the correlated frailty models, two random variables are used to characterize the frailty effect of each cluster. For example, in twin pairs one random variable can be

assigned for twin 1 and one for twin 2 so that they are no longer constrained to have a common frailty. These two random variables are associated and jointly distributed and therefore knowing one of them does not imply the other. Also, these two random variables can certainly be negatively associated, which would induce negative association between the survival times of the twins. In the multivariate frailty models, two or more frailty terms are assumed for each individual in the cluster, so as to induce a more elaborate association structure between the times for the individuals.

### 2.3.2 Model formulation

Assume we have a total of $n$ individuals that come from $G$ different groups, such that the $i^{\text{th}}$ group has $n_i$ individuals. For each individual the observed data is $y_{ij} = (T_{ij}^o, \delta_{ij})$, with $i = 1, \ldots, G$, and $j = 1, 2, \ldots, n_i$. Here $T_{ij}^o$ is the observed time and $\delta_{ij}$ is the censoring indicator both defined as in Section 2.1.2. It follows that the number of events from the $i^{\text{th}}$ group is $D_i = \sum_{j=1}^{n_i} \delta_{ij}$.
The frailty model is given by

$$\lambda_{ij}(t) = \lambda_0(t) \exp(\boldsymbol{x}_{ij}^T \boldsymbol{\beta} + w_i) \tag{2.3}$$

where $\lambda_{ij}(t)$ is the hazard function for the $j^{\text{th}}$ individual from the $i^{\text{th}}$ group, $\lambda_0(t)$ is the baseline hazard at time $t$, $\boldsymbol{x}_{ij}$ is the vector of $p$ covariates recorded for the individual and $w_i$ is the random effect for the $i^{\text{th}}$ group. In this model $\lambda_0(t)$ can be left arbitrary or be assumed to have some specific parametric distribution as before. The $w_i$'s, i =1,. . . ,G are a sample (independent and identically distributed) from a density $f_W(\cdot)$. The frailty model can be rewritten as follows:

$$\lambda_{ij}(t) = \lambda_0(t) \exp(w_i) \exp(\boldsymbol{x}_{ij}^T \boldsymbol{\beta}) = \lambda_0(t) u_i \exp(\boldsymbol{x}_{ij}^T \boldsymbol{\beta}) \tag{2.4}$$

where $u_i = \exp(w_i)$ is known as the frailty. Model (2.4) is a conditional hazard function given the independent $u_i$' s, i =1,. . . ,G which are assumed to have a common density $f_U(\cdot)$. Two classical choices for the density of the frailties are:
(a) The zero-mean normal density for W; then the density of U is log-normal, i.e.,

$$f_U(u) = \frac{1}{u\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\log u)^2}{2\sigma^2}\right) .' \tag{2.5}$$

with mean $e^{\sigma^2/2}$ and variance $e^{\sigma^2}(e^{\sigma^2}-1)$.

(b) The one-parameter gamma density for U,

$$f_U(u) = \frac{u^{\frac{1}{\gamma}-1}\exp\left(-\frac{u}{\gamma}\right)}{\gamma^{\frac{1}{\gamma}}\Gamma\left(\frac{1}{\gamma}\right)}. \tag{2.6}$$

Then the corresponding density for W is

$$f_W(w) = \frac{(\exp(w))^{\frac{1}{\gamma}}\exp\left(\frac{-\exp(w)}{\gamma}\right)}{\frac{1}{\gamma}\Gamma\left(\frac{1}{\gamma}\right)}.$$

which is the log-gamma density. We note that $E[W] = \psi(\frac{1}{\gamma})$ and $Var(W) = \psi^{(1)}(\frac{1}{\gamma})$ where $\psi(\cdot)$ and $\psi^{(1)}(\cdot)$ are the digamma and trigamma functions respectively.

Since U in (2.4) can be thought of as a mixing term, its density $f_U(\cdot)$ is also referred to as a mixing distribution. Typically $\text{Var}(W) = \sigma^2$ is used to describe the heterogeneity among the groups in the log-normal density case whereas $\text{Var}(U) = \gamma$ is used in the gamma density case. Below we use $\theta$ as generic notation for heterogeneity (meaning $\sigma^2$ for the log-normal density and $\gamma$ for the gamma density). If $\gamma$ is small then the gamma and log-normal distributions are similar (Kalbfleisch and Prentice, 1980, p. 26).

The gamma distribution has been used extensively, due to its mathematical convenience that results from the simple form of its Laplace transform. For example the Laplace transform corresponding to (2.6) is

$$L_u(s) = E\left[e^{-sU}\right] = (1+\gamma s)^{-1/\gamma}. \tag{2.7}$$

This leads to closed form expressions for the unconditional (marginal) survival and hazard functions. No closed form expression exists for the Laplace transform for the log-normal distribution. On the other hand, this latter distribution is more flexible than the gamma in creating correlated frailties, thus resulting in its use in multivariate frailty models.

Other frailty distributions which have been used in the literature are the stable distribution and the power variance functions (PVF) (Hougaard, 2000). The power variance function is a larger family of distributions which includes among others, the gamma and the positive stable and hence is less restrictive. The calculations for this larger family are however more difficult thus hindering the use of this distribution.

The dependence between the times of individuals in a cluster is often measured using measures of dependence such as the Spearman's correlation and the Kendall's coefficient

of concordance (Hougaard, 2000, p. 129) . These measures depend only on the frailty distribution and are independent of the regressor variables as well as the number of individuals in a cluster. For example corresponding to the gamma frailty model (2.4), the Kendall's coefficient is $\gamma/(2+\gamma)$.
In this thesis, only the gamma and the log-normal frailty distributions are considered and interest mainly focuses on the heterogeneity and regression parameters and not on the measures of dependence of the survival times.

### 2.3.3 Heterogeneity parameter

As seen in the previous section, the heterogeneity parameter is $\theta = \text{Var}(U) = \gamma$ for the gamma frailty or $\theta = \text{Var}(W) = \sigma^2$ for the log-normal density. If $\theta = 0$ then there is no heterogeneity between clusters. In this section we look at the effect of $\theta$ values on the median time to event in relation to a multicenter breast cancer clinical trial setting. Let $T$ be the time-to-event and assume that $T$ is exponential with parameter $\lambda_0$) and that $\theta = 0$. Then it follows that the median time to event is

$$T_M = \frac{\log 2}{\lambda_0}.$$

For $\lambda_0 = 0.07$ and $\lambda_0 = 0.22$ the median time to events are 9.9 and 3.15 years respectively. On the other hand if $\theta > 0$, $\lambda(t|u) = u\lambda_0$ and U is a gamma frailty, then the median time to event is

$$T_M = \frac{\log 2}{U\lambda_0}.$$

We have that the density $f_{T_M}(t)$ of the median time-to-event is

$$f_{T_M}(t) = \left(\frac{\log 2}{\theta\lambda_0)}\right)^{1/\theta} \left(\frac{1}{t}\right)^{1+1/\theta} \frac{1}{\Gamma(1/\theta)} \exp\left(-\frac{\log(2)}{t\theta\lambda_0}\right).$$

Based on this density function it can be derived that for instance, if $\theta = 0.1$, then 90% of the centres will have a median time to event between 2 and 5.6 years for $\lambda_0 = 0.22$ and between 6.4 and 17.6 for $\lambda_0 = 0.07$ as can be seen in Figure 2.1. On the other hand, if $\theta = 0.2$, then the median time to event is now between 2 and 7.3 years for $\lambda_0 = 0.22$ and between 4.5 and 18.2 years for $\lambda_0 = 0.07$ for 90 % of the centres. We thus observe that the median time to event becomes more spread as the value of the heterogeneity parameter increases. On the other hand the median time to event is observed to be sensitive to the event rate, with the spread decreasing with increase in the event rate.

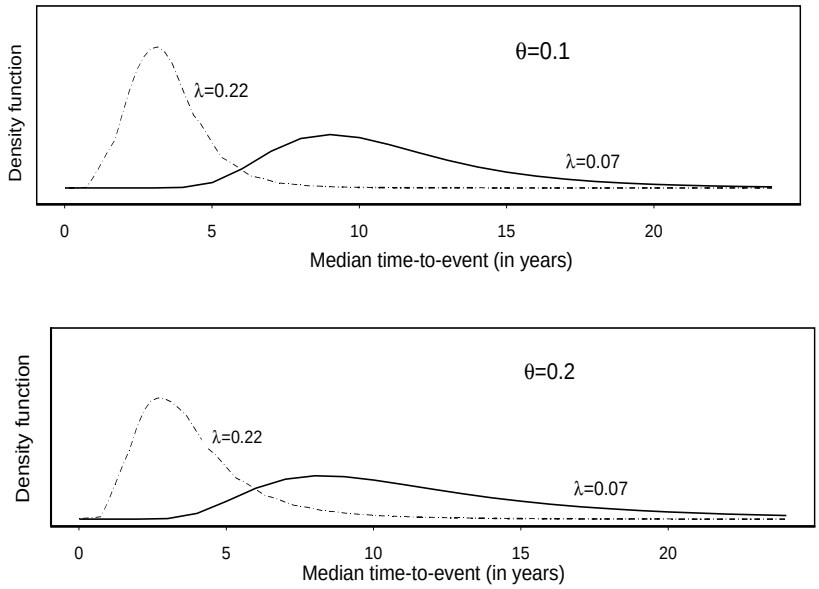



Figure 2.1: *The spread of the median time to event from centre to centre for* $\theta = 0.1$ *and* $0.2$ *and* $h_0 = 0.07$ *and* $0.22$.

## 2.4 Estimation in the shared frailty model

Before looking at the estimation methods that have been used for the frailty model, we briefly discuss identifiability of the model and also give a brief introduction of the various estimation procedures that have been used in the literature.

### 2.4.1 Introduction : identifiability and estimation

Prior to exploring estimation procedures of the unknown parameters in the shared frailty model, we start by briefly commenting on whether for this model, the parameters and related functions, such as the survival and cumulative hazard functions, are identifiable from a particular set of data. The aim of this section is to illustrate with examples issues that are discussed when identifiability of the frailty model is considered. Thus it is not a complete discussion.

Identifiability issues of the frailty model, especially in econometric applications, have been discussed by Elbers and Ridder (1982), Heckman and Singer (1984) and Lancaster (1990).

These issues are discussed in relation to complete non-censored data from the univariate frailty model whose hazard for the $i^{\text{th}}$ individual is given as

$$\lambda(t|u_i, \boldsymbol{x}_i) = u_i \psi(\boldsymbol{x}_i, t) \tag{2.8}$$

Elbers and Ridder(1982) have shown that the finite mean assumption (say $E[U] = 1$) plays the same role as the mean zero assumption in a linear regression model. They show that under this assumption and provided that $\psi(\boldsymbol{x}, t) = \lambda_0(t)\phi(\boldsymbol{x})$, i.e., factors into the product of a function of $\boldsymbol{x}$ and a function of $t$ then (2.8) is identifiable, as long as there are at least two distinct values of the covariates. However (2.8) is unidentifiable if $\psi(\boldsymbol{x}, t) \equiv \lambda_0(t)$, i.e., if $\phi(\boldsymbol{x}) \equiv 1$. For example consider the following two models:

Model 1: $\lambda(t|u) = u\frac{\beta}{\gamma(1+\beta t)}$ such that $f_U(u) = 1$ when $u = 1$.

Model 2: $\lambda(t|u) = u\frac{\beta}{\gamma}$ such that $f_U(u)$ has the gamma density (2.6) above.

If $\Lambda(t) = \int_0^t \lambda(s)ds$ denotes the cumulative baseline hazard then for both models the unconditional survival function is $S(t) = \int_0^\infty \exp(-u\Lambda(t))f(u)du = (1 + \beta t)^{-1/\gamma}$.

Thus indeed, starting from $S(t)$ which is observable, it is impossible to uniquely identify the density $f_U(\cdot)$ of the frailty and the hazard function $\lambda_0(\cdot)$.

Heckman and Singer (1984), on the other hand, discuss the identifiability of (2.8), when the mixing distribution has infinite mean but with restrictions on $\phi(\boldsymbol{x})$. Lancaster (1990) further shows that if $\lambda_0(t)$ has a Weibull or exponential distribution and the mixing distribution has finite mean, then model (2.8) is identifiable even in its simplest form when $\phi(\boldsymbol{x}) \equiv 1$. Lenstra et al. (1995) provide a constructive identification proof for (2.8) for the case that $\phi(\boldsymbol{x})$ has two distinct values, both for complete and censored data.

In relation to frailty models for multivariate event data, Honoré (1993) shows that for complete data (no censoring), Model (2.4) is identifiable under much weaker assumptions than the univariate model. He shows that this model is identifiable even when there are no observable covariates. For example the two models in the example above are easily seen to be identifiable, when adapted to multivariate event data with clusters of size two, for example in twin studies. This leads to the following unconditional survival functions:

Model 1: $S(t_1, t_2) = \prod_{j=1}^{2} (1 + \beta t_i)^{-1/\gamma}$

Model 2: $S(t_1, t_2) = (1 + \beta(t_1 + t_2))^{-1/\gamma}$

which are uniquely determined for each of the two models. For completeness, given these two unconditional survival functions we show that the baseline hazard $\lambda_0(\cdot)$ and frailty density $f_U(\cdot)$ can be uniquely determined.

Consider $S(t_1, t_2)$ for Model 1. We observe that

$$S(t_1, t_2) = S(t_1, 0)S(0, t_2) = S_1(t_1)S_2(t_2) \tag{2.9}$$

where $S_j(t_j)$ is the survival function for $j^{\text{th}}$ failure time in the cluster, $j = 1, 2$. From (2.9) it is easily seen that $f(t_1, t_2) = f_1(t_1)f_2(t_2)$ where

$$f_j(t_j) = -\frac{dS_j(t_j)}{dt_j} = \frac{\beta}{\gamma}(1 + \beta t_j)^{\frac{-1}{\gamma}-1}$$

is the density function for $T_j, j = 1, 2$. Thus $T_1$ and $T_2$ are independent which implies that $U \equiv 1$. It also follows that

$$\lambda_j(t_j) = \frac{\beta}{\gamma}(1 + \beta t_j)^{\frac{-1}{\gamma}}$$

for $j = 1, 2$, implying that both failure times in the cluster have a common baseline hazard function.

Now consider a general frailty model as in (2.4) but without covariates. Further assume that there is no censoring. Then the unconditional multivariate survival function for individuals in the $i^{\text{th}}$ group is

$$S(t_{i1}, \ldots, t_{in_i}) = L_u(\Lambda_0(t_{i1}) + \ldots + \Lambda_0(t_{in_i}))$$

where $L_u(s)$ is the Laplace transform of the distribution of $U$ and $\Lambda_0(\cdot)$ is the cumulative baseline hazard function (Hougaard, 2000, p. 222). Now consider the unconditional survival function for Model 2. Then we have that

$$\begin{aligned} S(t_1, t_2) &= L_u(\Lambda_0(t_1) + \Lambda_0(t_2)) \\ &= (1 + \beta(t_1 + t_2))^{-1/\gamma} \\ &= \left(1 + \gamma\left(\frac{\beta}{\gamma}(t_1 + t_2)\right)\right)^{-1/\gamma}. \end{aligned}$$

By the uniqueness property of the Laplace transform, this is the Laplace transform of a gamma random variable with parameter $\frac{1}{\gamma}$ (see (2.7)). Hence $\Lambda_0(t_j) = \frac{\beta}{\gamma}t_j$, $j = 1, 2$ which in turn implies that the common baseline hazard function for each of the failure times in the cluster is $\lambda_0(t) = \frac{\beta}{\gamma}$.

Thus we are able to uniquely identify from the unconditional survival function the baseline hazard function and distribution of the frailty for the two models.

In the second part of this section we briefly discuss some of the methods of estimation

for the shared frailty model, that have been discussed in recent literature. As noted above, the baseline hazard $\lambda_0(t)$ in the frailty model (2.4) can be specified explicitly or left unspecified. Under the parametric assumption, the parameters in the resulting model can be estimated using maximum likelihood estimation procedures. For example, suppose that $\lambda_0(t) = \rho\lambda t^{\,\rho-1}$ which is the Weibull distribution with $\lambda$ and $\rho$ as the scale and shape parameters respectively. Then the unconditional (observable) likelihood to be maximized is

$$L\left(\boldsymbol{\zeta}\right) = \prod_{i=1}^{G} \int \left[ \prod_{j=1}^{n_i} \left( \rho\lambda t_{ij}^{\rho-1} \exp(\boldsymbol{x}_{ij}^T\boldsymbol{\beta} + w_i) \right)^{\delta_{ij}} \exp\left( -\lambda t_{ij}^{\rho} \exp(\boldsymbol{x}_{ij}^T\boldsymbol{\beta} + w_i) \right) f_W(w_i) dw_i \right]$$

where $\boldsymbol{\zeta} = (\boldsymbol{\beta}, \theta, \rho, \lambda)$.

If however $\lambda_0(t)$ is left unspecified, then the unknown parameters in the shared frailty model have been estimated using various approaches such as the EM algorithm (Klein, 1992), penalized partial likelihood approach (Therneau and Grambsch, 2000), Monte Carlo Markov Chain (MCMC) methods (Vaida and Xu, 2000), Monte Carlo EM approach (MCEM) (Ripatti et al., 2002) and different methods using Laplace approximation (Ripatti and Palmgren, 2000, Cortinas Abrahantes and Burzykowski, 2003). The choice of the estimation method in most cases is basically determined by the frailty distribution. We saw in Section 2.3.2 that closed forms of the unconditional survival and hazard functions are easily determined under the gamma frailty. This allows the use of maximum likelihood estimation procedures such as the EM algorithm as the unconditional likelihood is easily determined. This estimation procedure for the gamma frailty is discussed in Section 2.4.2. Unfortunately under the log-normal frailty distribution, explicit expressions of the unconditional survival and hazard functions do not exist. Consequently, estimation strategies for this frailty distribution are often based on numerical integration methods such as the Laplace approximation methods. We briefly comment on these procedures in Section 2.4.3. The penalized partial likelihood method of estimation is discussed in Section 2.4.4, while in Section 2.4.5 we show that the estimates obtained from the EM and penalized partial likehood estimation procedures are the same for the shared gamma frailty model.

### 2.4.2 The EM algorithm for the shared gamma frailty model

Typical for the frailty model is that we have observed information $\boldsymbol{y} = (y_{11}, \dots y_{Gn_G})^T$ and unobserved (latent) information $\boldsymbol{w} = (w_1, \dots w_G)^T$. Here $y_{ij} = (T^o_{ij}, \delta_{ij})$, $(i = 1, \dots, G,$ and $j = 1, 2, \dots, n_i)$ as defined before in Section 2.3.2. The conditional likelihood for the $i^{\text{th}}$ group is

$$L_i(y_{i1} \dots y_{in_i} \mid w_i) = \prod_{j=1}^{n_i} \left(\lambda_0(t_{ij}) \exp(\boldsymbol{x}_{ij}^T\boldsymbol{\beta} + w_i)\right)^{\delta_{ij}} \exp\left(-\Lambda_0(t_{ij}) \exp(\boldsymbol{x}_{ij}^T\boldsymbol{\beta} + w_i)\right)$$

If we let $\boldsymbol{\zeta} = (\theta, \boldsymbol{\beta})$ it follows that the observable likelihood $L_{obs,i}(\boldsymbol{\zeta})$ for the $i^{\text{th}}$ group is

$$\begin{aligned} L_{obs,i}(\boldsymbol{\zeta}) &= \int_0^\infty \prod_{j=1}^{n_i} \left(\lambda_0(t_{ij}) \exp(\boldsymbol{x}_{ij}^T\boldsymbol{\beta} + w_i)\right)^{\delta_{ij}} \exp\left(-\Lambda_0(t_{ij}) \exp(\boldsymbol{x}_{ij}^T\boldsymbol{\beta} + w_i)\right) \\ &\qquad \times \frac{(\exp(w_i))^{\frac{1}{\theta}}}{\theta^{\frac{1}{\theta}}\Gamma(1/\theta)} \exp\left(-\frac{\exp(w_i)}{\theta}\right) dw_i \\ &= \frac{\Gamma(1/\theta + D_i)}{\theta^{\frac{1}{\theta}}\Gamma(1/\theta)\left[\frac{1}{\theta} + \sum_{j=1}^{n_i} \Lambda_0(t_{ij}) \exp(\boldsymbol{x}_{ij}^T\boldsymbol{\beta})\right]^{\frac{1}{\theta}+D_i}} \prod_{j=1}^{n_i} \left(\lambda_0(t_{ij}) \exp(\boldsymbol{x}_{ij}^T\boldsymbol{\beta})\right)^{\delta_{ij}} \end{aligned}$$

where $D_i = \sum_{j=1}^{n_i} \delta_{ij}$. To estimate $\boldsymbol{\zeta}$ we would like to base the likelihood maximization and statistical inference on the observed data log-likelihood $l_{obs}(\boldsymbol{\zeta})$ (Klein, 1992) given by

$$\begin{aligned} l_{obs}(\boldsymbol{\zeta}) &= \sum_{i=1}^{G} \log L_{obs,i}(\boldsymbol{\zeta}) \\ &= \sum_{i=1}^{G} \left(D_i \log\theta - \log\Gamma\left(1/\theta\right) + \log\Gamma\left(1/\theta + D_i\right)\right) \\ &\quad - \sum_{i=1}^{G} (1/\theta + D_i) \log\left[1 + \theta \sum_{j=1}^{n_i} \Lambda_0(t_{ij}) \exp(\boldsymbol{x}_{ij}^T\boldsymbol{\beta})\right] \\ &\quad - \sum_{i=1}^{G} \sum_{j=1}^{n_i} \delta_{ij} [\log(\lambda_0(t_{ij}) + \boldsymbol{x}_{ij}^T\boldsymbol{\beta})]. \end{aligned} \tag{2.10}$$

This log-likelihood is however difficult to maximize as it contains, apart from $\boldsymbol{\zeta}$, also the unspecified baseline hazard. We therefore rely on the EM algorithm (Dempster et al., 1977) to estimate $\boldsymbol{\zeta}$.

Define $\ell_{full}(\boldsymbol{\zeta}) = \log f(\boldsymbol{y}, \boldsymbol{w}; \boldsymbol{\zeta})$ as the complete data log-likelihood and $\ell_{pred}(\boldsymbol{\zeta}) = \log f(\boldsymbol{w}|\boldsymbol{y}; \boldsymbol{\zeta})$ as the predictive log-likelihood. Then the algorithm works by alternating between two steps : an expectation step (E-step) in which the expected value of the unobserved part ($\boldsymbol{w}$) given the observed data and current parameter estimates is determined and a maximization step (M-step) which involves maximizing the complete data likelihood using the expected values of the unobserved part from the previous E-step.

In the EM algorithm framework we write

$$\ell_{obs}(\boldsymbol{\zeta}) = \ell_{full}(\boldsymbol{\zeta}) - \ell_{pred}(\boldsymbol{\zeta})$$

We here note that $\ell_{full}(\boldsymbol{\zeta}) = \log \left[ \prod_{i=1}^{G} L_i(y_{i1} \ldots y_{in_i} \mid w_i) f(w_i) \right]$.

Taking the conditional expectation with respect to $\boldsymbol{y}$ and with $\boldsymbol{\zeta}^{(k-1)}$ as a provisional value of $\boldsymbol{\zeta}$ at iteration step $k-1$ in the EM algorithm, we obtain

$$\log f(\boldsymbol{y}; \boldsymbol{\zeta}) \quad = \quad E_{\boldsymbol{\zeta}^{(k-1)}} \left[ \log f(\boldsymbol{y}, \boldsymbol{w}; \boldsymbol{\zeta}) | \boldsymbol{y} \right] - E_{\boldsymbol{\zeta}^{(k-1)}} \left[ \log f(\boldsymbol{w}|\boldsymbol{y}; \boldsymbol{\zeta}) | \boldsymbol{y} \right]$$

or

$$\ell_{obs}(\boldsymbol{\zeta}) \quad = \quad Q\left(\boldsymbol{\zeta}|\boldsymbol{\zeta}^{(k-1)}\right) - H\left(\boldsymbol{\zeta}|\boldsymbol{\zeta}^{(k-1)}\right) \tag{2.11}$$

where

$$Q\left(\boldsymbol{\zeta}|\boldsymbol{\zeta}^{(k-1)}\right) = E_{\boldsymbol{\zeta}^{(k-1)}} \left[ \log f(\boldsymbol{y}, \boldsymbol{w}; \boldsymbol{\zeta}) | \boldsymbol{y} \right]$$

and

$$H\left(\boldsymbol{\zeta}|\boldsymbol{\zeta}^{(k-1)}\right) = E_{\boldsymbol{\zeta}^{(k-1)}} \left[ \log f(\boldsymbol{w}|\boldsymbol{y}; \boldsymbol{\zeta}) | \boldsymbol{y} \right].$$

Instead of maximizing $\ell_{obs}(\boldsymbol{\zeta})$ for $\boldsymbol{\zeta}$ rather $Q\left(\boldsymbol{\zeta}|\boldsymbol{\zeta}^{(k-1)}\right)$ is maximized. It is a general result from EM methodology that if $\boldsymbol{\zeta}^{(k)}$ maximizes $Q\left(\boldsymbol{\zeta} \mid \boldsymbol{\zeta}^{(k-1)}\right)$ then $\ell_{obs}\left(\boldsymbol{\zeta}^{(k)}\right) \geq \ell_{obs}\left(\boldsymbol{\zeta}^{(k-1)}\right)$, i.e., $\boldsymbol{\zeta}^{(k)}$ is 'better' than $\boldsymbol{\zeta}^{(k-1)}$. This property is central in Dempster et al. (1977). As starting values for the algorithm, we use an initial guess $\boldsymbol{\zeta}^{(0)} = \left(\theta^{(0)}, \boldsymbol{\beta}^{(0)}\right)$. The details of the $k^{\text{th}}$ step of the algorithm are as follows.

Expectation step:

The expected value $Q\left(\boldsymbol{\zeta}|\boldsymbol{\zeta}^{(k-1)}\right)$ can be obtained by plugging in the conditional expectations for $w_i$ and $\exp(w_i)$ given $\boldsymbol{y}$ in $\log f(\boldsymbol{y}, \boldsymbol{w}; \boldsymbol{\zeta})$. The conditional density of $w_i$ given $\boldsymbol{y}$

is

$$f(w_i|\boldsymbol{y}) = \frac{L_i(y_{i1}\ldots y_{in_i} \mid w_i) f(w_i)}{L_{obs,i}(\boldsymbol{\zeta})}$$

$$= \frac{\left[\frac{1}{\theta} + \sum_{j=1}^{n_i} \Lambda_0(t_{ij}) \exp(\boldsymbol{x}_{ij}^T \boldsymbol{\beta})\right]^{\frac{1}{\theta}+D_i}}{\Gamma\left(1/\theta + D_i\right)} \left(\exp(w_i)\right)^{\frac{1}{\theta}+D_i}$$

$$\times \exp\left(-\exp(w_i)\left[\frac{1}{\theta} + \sum_{j=1}^{n_i} \Lambda_0(t_{ij}) \exp(\boldsymbol{x}_{ij}^T \boldsymbol{\beta})\right]\right).$$

which is a log-gamma distribution with parameters $\frac{1}{\theta} + D_i$ and $\frac{1}{\theta} + \sum_{j=1}^{n_i} \Lambda_0(t_{ij}) \exp(\boldsymbol{x}_{ij}^T \boldsymbol{\beta})$. With $\psi(.)$ as the digamma function, it then follows that

$$E_k(w_i) = E_{\boldsymbol{\zeta}^{(k-1)}}\left[w_i|\boldsymbol{y}\right] = \psi\left(D_i + 1/\theta^{(k-1)}\right) - \log\left(1/\theta^{(k-1)} + \Lambda_i^{(k-1)}\right)$$

and

$$E_k(\exp(w_i)) = E_{\boldsymbol{\zeta}^{(k-1)}}\left[\exp(w_i)|\boldsymbol{y}\right] = \frac{1/\theta^{(k-1)} + D_i}{1/\theta^{(k-1)} + \Lambda_i^{(k-1)}} \tag{2.12}$$

where

$$\Lambda_i^{(k-1)} = \sum_{j=1}^{n_i} \Lambda_0^{(k-1)}(t_{ij}) \exp\left(\boldsymbol{x}_{ij}^T \boldsymbol{\beta}^{(k-1)}\right) \tag{2.13}$$

with $\Lambda_0^{(k-1)}(\cdot)$ as defined in (2.17). These expected values need to be inserted in

$$\log f(\boldsymbol{y}, \boldsymbol{w}; \boldsymbol{\zeta}) = \sum_{i=1}^{G} \log f_W(w_i) + \sum_{i=1}^{G}\sum_{j=1}^{n_i} \delta_{ij}\left[\log \lambda_0(t_{ij}) + \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + w_i\right]$$

$$- \sum_{i=1}^{G}\sum_{j=1}^{n_i}\left[\Lambda_0(t_{ij}) \exp(\boldsymbol{x}_{ij}^T\boldsymbol{\beta} + w_i)\right]. \tag{2.14}$$

Plugging in the density for $f_W(\cdot)$, corresponding to the gamma frailty and replacing $w_i$ and $\exp(w_i)$ in (2.14) by their conditional expectations and adding and subtracting the term $\sum_{j=1}^{n_i} \delta_{ij} \log E_k(\exp(w_i))$, we obtain after rearranging some of the terms

$$Q\left(\boldsymbol{\zeta} \mid \boldsymbol{\zeta}^{(k-1)}\right) = Q_1\left(\theta \mid \boldsymbol{\zeta}^{(k-1)}\right) + Q_2\left(\boldsymbol{\beta} \mid \boldsymbol{\zeta}^{(k-1)}\right)$$

with

$$Q_1\left(\theta \mid \boldsymbol{\zeta}^{(k-1)}\right) = \sum_{i=1}^{G}\left[(1/\theta + D_i)E_k(w_i) - E_k(\exp(w_i))/\theta - D_i \log E_k(\exp(w_i))\right]$$

$$-G\left[\log\theta/\theta + \log\Gamma(1/\theta)\right] \tag{2.15}$$

which is a function of $\theta$ only and

$$Q_2\left(\boldsymbol{\beta} \mid \boldsymbol{\zeta}^{(k-1)}\right) = \sum_{i=1}^{G}\sum_{j=1}^{n_i}\delta_{ij}\left[\log\lambda_0(t_{ij}) + \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \log E_k(\exp(w_i))\right]$$

$$-\sum_{i=1}^{G}\sum_{j=1}^{n_i}\Lambda_0(t_{ij})\exp(\boldsymbol{x}_{ij}^T\boldsymbol{\beta})E_k(\exp(w_i)) \tag{2.16}$$

which is a function of $\boldsymbol{\beta}$ and $\lambda_0(\cdot)$.

Maximization step:

The reason why we artificially include the term $\sum_{j=1}^{n_i}\delta_{ij}\log E_k(\exp(w_i))$ in the definitions of $Q_1\left(\theta \mid \boldsymbol{\zeta}^{(k-1)}\right)$ and $Q_2\left(\boldsymbol{\beta} \mid \boldsymbol{\zeta}^{(k-1)}\right)$, with respectively a minus-sign and a plus-sign, is that we now can interpret $Q_2\left(\boldsymbol{\beta} \mid \boldsymbol{\zeta}^{(k-1)}\right)$ as a log-likelihood for censored data with $(\boldsymbol{\beta}^T, 1)^T$ as regression coefficients and $(x_{ij1}, \ldots, x_{ijp}, \log E_k(\exp(w_i)))$ as risk variables.
To maximize $Q_2\left(\boldsymbol{\beta} \mid \boldsymbol{\zeta}^{(k-1)}\right)$ with respect to $\boldsymbol{\beta}$ and $\Lambda_0(\cdot)$ a profile likelihood idea is used. The estimate for $\boldsymbol{\beta}$ is first obtained from the profile likelihood and then the full censored data log-likelihood is maximized as a function of the baseline hazard only, as detailed below.

For $t_{(1)} < \ldots\ldots < t_{(r)}$ as the ordered event times ($r$ denotes the number of distinct event times), $N_{(l)}$ as the number of events at time $t_{(l)}$, $l = 1, \ldots, r$ and using Breslow's method for handling ties (see Sections 8.2 and 8.3 in Klein and Moeschberger, 1997), the partial likelihood corresponding to $Q_2\left(\boldsymbol{\beta} \mid \boldsymbol{\zeta}^{(k-1)}\right)$ is

$$\ell_{part}\left(\boldsymbol{\beta} \mid \boldsymbol{\zeta}^{(k-1)}\right) = \sum_{l=1}^{r}\left[\sum_{t_{ij}=t_{(l)}}\eta_{ij}^{(k-1)} - N_{(\ell)}\log\left(\sum_{t_{qs}\geq t_{(\ell)}}\exp\left(\eta_{qs}^{(k-1)}\right)\right)\right]$$

where

$$\eta_{ij}^{(k-1)} = \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \log E_k(\exp(w_i)).$$

This is maximized to get an estimate for $\boldsymbol{\beta}$. Corresponding to the censored data log-likelihood ($Q_2\left(\boldsymbol{\beta} \mid \boldsymbol{\zeta}^{(k-1)}\right)$) the estimate for $\Lambda_0^{(k-1)}$ in this step is given by

$$\Lambda_0^{(k-1)}(t) = \sum_{t_{(l)} \leq t} \lambda_{l0}^{(k-1)} \tag{2.17}$$

where

$$\lambda_{l0}^{(k-1)} = \frac{N_{(l)}}{\sum_{t_{qs} \geq t_{(l)}} \exp(\boldsymbol{x}_{qs}^T \boldsymbol{\beta}^{(k-1)}) E_{k-1}(\exp(w_s))}.$$

We note that (2.17) is the Breslow estimator for the cumulative hazard in presence of ties which depends on $\boldsymbol{\beta}^{(k-1)}$ and $\theta^{(k-1)}$ through $E_{k-1}(\exp(w_s)$.
Hence in general in the M-step, we maximize,

$$Q_1\left(\theta \mid \boldsymbol{\zeta}^{(k-1)}\right) + \ell_{part}\left(\boldsymbol{\beta} \mid \boldsymbol{\zeta}^{(k-1)}\right). \tag{2.18}$$

with respect to $\boldsymbol{\zeta} = (\theta, \boldsymbol{\beta})$ and $\Lambda_0(\cdot)$. The new estimate for $\boldsymbol{\zeta}^{(k)}$ in the $k^{\text{th}}$ iteration is then used to obtain $Q(\boldsymbol{\zeta} \mid \boldsymbol{\zeta}^{(k)})$, the update of the conditional expectation, and so on. This process continues until the difference between the two consecutive values $E_{\boldsymbol{\zeta}^{(k-1)}}\left[\log f(\boldsymbol{y}, \boldsymbol{w}; \boldsymbol{\zeta}) \mid \boldsymbol{y}\right]$ and $E_{\boldsymbol{\zeta}^{(k)}}\left[\log f(\boldsymbol{y}, \boldsymbol{w}; \boldsymbol{\zeta}) \mid \boldsymbol{y}\right]$ becomes smaller than some prespecified value $\epsilon$ (see Flow Chart 1 in the Section 2.6 ).

**Remark 1.** Nielsen et al. (1992) propose a modified profile likelihood EM algorithm that leads to the same results but with much faster convergence.

**Remark 2.** Klein (1992) determines the standard errors of the estimates of $\boldsymbol{\beta}$, $\theta$ and $\lambda_{l0}(\cdot)$ from the inverse of the observed information matrix of the observable log-likelihood $\ell_{obs}(\boldsymbol{\zeta})$ given in (2.10). This information matrix is a square matrix of size $r + p + 1$ where $r$ is the number of distinct failure times. For large data sets this procedure is quite numerically intensive.

### 2.4.3 The EM algorithm for the shared log-normal frailty

Under the log-normal frailty model, the observed likelihood $L_{obs,i}(\boldsymbol{\zeta})$ for the $i^{\text{th}}$ group is

$$L_{obs,i}(\boldsymbol{\zeta}) = \int_{-\infty}^{\infty} \prod_{j=1}^{n_i} \left[\lambda_0(t_{ij}) \exp(\boldsymbol{x}_{ij}^T \boldsymbol{\beta} + w_i)\right]^{\delta_{ij}} \exp\left[-\Lambda_0(t_{ij}) \exp(\boldsymbol{x}_{ij}^T \boldsymbol{\beta} + w_i)\right]$$

$$\times (2\pi\theta)^{-\frac{1}{2}} \exp\left(-\frac{w_i^2}{2\theta}\right) dw_i$$

$$= \left[ \int_{-\infty}^{\infty} (\exp(w_i))^{D_i} \exp\left( -exp(w_i) \sum_{j=1}^{n_i} \Lambda_0(t_{ij}) \exp(\boldsymbol{x}_{ij}^T \boldsymbol{\beta}) \right) \left( \frac{1}{2\pi\theta} \right)^{-\frac{1}{2}} \exp\left( -\frac{w_i^2}{2\theta} \right) dw_i \right]$$

$$\times \prod_{j=1}^{n_i} \left( \lambda_0(t_{ij}) \exp(\boldsymbol{x}_{ij}^T \boldsymbol{\beta}) \right)^{\delta_{ij}} . \tag{2.19}$$

The integrand above has no explicit solution and apart from containing $\boldsymbol{\zeta}$ also depends on the baseline hazard $\lambda_0(\cdot)$. The EM methodology is thus adapted and the details are as follows.

Expectation step:

As in (2.11), $\ell_{obs}(\boldsymbol{\zeta})$ can be written in terms of $Q\left(\boldsymbol{\zeta}|\boldsymbol{\zeta}^{(k-1)}\right)$ and $H\left(\boldsymbol{\zeta}|\boldsymbol{\zeta}^{(k-1)}\right)$ where now

$$\begin{aligned} Q\left(\boldsymbol{\zeta}|\boldsymbol{\zeta}^{(k-1)}\right) &= E_{\boldsymbol{\zeta}^{(k-1)}} \left[\log f(\boldsymbol{y}, \boldsymbol{w}; \boldsymbol{\zeta}) | \boldsymbol{y}\right] \\ &= \sum_{i=1}^{G} \sum_{j=1}^{n_i} \delta_{ij} \left[ log\lambda_0(t_{ij}) + \boldsymbol{x}_{ij}^T \boldsymbol{\beta} + E_{k-1}[w_i | \boldsymbol{y}, \boldsymbol{\zeta}] \right] \end{aligned}$$

$$- \sum_{i=1}^{G} \sum_{j=1}^{n_i} \Lambda_0(t_{ij}) \exp(\boldsymbol{x}_{ij}^T \boldsymbol{\beta}) E_{k-1} \left[\exp(w_i) | \boldsymbol{y}, \boldsymbol{\zeta}\right] \tag{2.20}$$

and

$$H\left(\boldsymbol{\zeta}|\boldsymbol{\zeta}^{(k-1)}\right) = E_{\boldsymbol{\zeta}^{(k-1)}} \left[\log f(\boldsymbol{w}|\boldsymbol{y}; \boldsymbol{\zeta}) | \boldsymbol{y}\right]$$

$$= -\frac{1}{2} \sum_{i=1}^{G} \left( \log(2\pi) + \log\theta + \frac{1}{\theta} E_{k-1} \left[ w_i^2 | \boldsymbol{y}, \boldsymbol{\zeta} \right] \right) . \tag{2.21}$$

The computations of the expectation in (2.20) and (2.21) are of the type $E\left[h(w|\boldsymbol{y}, \boldsymbol{\zeta})\right] = \int h(w|\boldsymbol{y}, \boldsymbol{\zeta}) f(w|\boldsymbol{y}, \boldsymbol{\zeta}) dw$ and are not available in closed form. These integrals can however be approximated by numerical methods (e.g. Gaussian quadratures) or Monte Carlo simulation methods. Both Vaida and Xu (2000) and Ripatti et al. (2002) have looked at the multivariate log-normal frailty model of which the shared frailty is a special case. In the approaches used by these authors, $\boldsymbol{w}^T = (w_1, ...., w_G)$ is assumed to have a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{D}$, which is not necessarily diagonal. Vaida and Xu (2000) use a Gibbs sampler to draw samples from the posterior distribution of the frailties ($f(w|\boldsymbol{y})$) . These are used to approximate the conditional expectation given the data and the current values of the parameters in the model. Following

the same ideas, Ripatti et al. (2002) on the other hand use rejection sampling to draw the samples from $f(w|\boldsymbol{y})$.

Maximization step:

The estimates from the E-step, are used to replace the expected values in (2.20) and (2.21). In this step, the estimation of $\boldsymbol{\beta}$ and $\lambda_0(\cdot)$ is conveniently separated from the variance term $\theta$. Expression (2.20) has the same form as the log-likelihood for censored data, containing $E_{k-1}\left[\exp(w_i)|\boldsymbol{y}, \boldsymbol{\zeta}\right]$ as fixed offset terms. A profile likelihood approach is hence adopted to maximize $Q\left(\boldsymbol{\zeta}|\boldsymbol{\zeta}^{(k-1)}\right)$ by first fixing $\boldsymbol{\beta}$ in a similar manner as discussed in the maximization step of Section 2.4.2 to obtain the estimates for $\boldsymbol{\beta}$ and $\Lambda(\cdot)$.

**Remark 3.** Ripatti and Palmgren (2000) use Laplace approximation to obtain an approximation to the observed log-likelihood corresponding to (2.19) for a multivariate log-normal frailty distribution. They then use maximum likelihood estimation methods to maximize the approximated likelihood. For the shared frailty model (independent frailties), this approximation approach has been shown to result in an estimation procedure equivalent to the penalized partial likelihood method (Therneau et al., 2003).

**Remark 4.** Cortinas Abrahantes and Burzykowski (2003) also use a modified EM approach for a multivariate normal frailty model. Using similar approach to Vaida and Xu (2000) they use the Laplace approximation in the E-step.

**Remark 5.** Vaida and Xu (2000), Ripatti et al. (2002) and Cortinas Abrahantes and Burzykowski (2003) use the formula derived by Louis (1982) to obtain the observed information matrix

$$I(\boldsymbol{\zeta}_\lambda) = \mathrm{E}\left[\frac{-\partial^2 \ell_{full}(\boldsymbol{\zeta})}{\partial \boldsymbol{\zeta}_\lambda \partial \boldsymbol{\zeta}_\lambda^T} \mid \boldsymbol{y}, \hat{\boldsymbol{\zeta}_\lambda}\right] - \mathrm{Var}\left[\frac{-\partial \ell_{full}(\boldsymbol{\zeta})}{\partial \boldsymbol{\zeta}_\lambda} \mid \boldsymbol{y}, \hat{\boldsymbol{\zeta}_\lambda}\right]$$

where $\boldsymbol{\zeta}_\lambda = (\boldsymbol{\zeta}, \lambda_{10}, \ldots, \lambda_{r0})$. This information matrix can easily be estimated as a by-product of the EM-algorithm. The dimension of this matrix is equal to the sum of the number of parameters and distinct event times in the data. Cortinas Abrahantes and Burzykowski (2003) show through simulations that the bias from estimating the standard errors of the random effects using only the sub-matrix of $I(\boldsymbol{\zeta}_\lambda)$ corresponding to these effects, is greatly reduced with increase in the sample size.

Parner (1998) showed the consistency of such an estimator for gamma distributed frailties, while Andersen et al. (1997) studied its performance in several applications.

### 2.4.4 The penalized partial likelihood for the shared frailty model

Penalized partial likelihood estimation originates from cubic splines regression in the Cox PH model. In a general Cox PH model (2.1) a linear relationship is assumed between the covariates and the log of the hazard. In spline smoothing this linear relationship is relaxed and a flexible function of the covariates $g(\cdot)$ is used to establish the appropriate functional form of the relationship. Thus we have

$$\lambda_i(t) = \lambda_0(t) \exp\{g(\boldsymbol{x}_i)\}.$$

Sleeper and Harrington (1990) studied regression splines while Gray (1992) used cubic splines, to approximate the covariate transformation. Gray (1992) substracts a penalty term $\left(\xi \int [g''(z)]^2 dz\right)$ where $g''(z) = \frac{d^2}{dz^2} g(z)$ from the log partial likelihood and considers

$$\ell_{pp\ell}(\boldsymbol{\beta}) \quad = \quad \log \prod_{i=1}^{n} \left\{ \frac{\exp(\boldsymbol{x}_i^T \boldsymbol{\beta})}{\sum_{j \in R_i} \exp(\boldsymbol{x}_j^T \boldsymbol{\beta})} \right\}^{\delta_i} - \xi \int \left[g''(z)\right]^2 dz$$

which is a penalized partial log-likelihood. The term $\xi$ is known as the tuning parameter which controls the amount of smoothing.

In relation to the frailty model (2.4) and following ideas from generalized mixed models, McGilchrist (1993) maximizes the log-likelihood $\ell_1 + \ell_2$ where

$\ell_1$= the log of the partial likelihood for (2.4) treating the $w$'s as fixed

and

$\ell_2 = \sum_{i=1}^{G} \log(f_W(w_i))$.

Thus $\ell_1 + \ell_2$ is a penalized partial likelihood, with $-\ell_2$ as the penalty function. From (2.14) and (2.18) it is thus clear that for the problem at hand, a logical proposal for the penalized partial likelihood to use for the estimation of $\boldsymbol{\zeta} = (\theta, \boldsymbol{\beta})$ is

$$\ell_{pp\ell}(\boldsymbol{\zeta}, \boldsymbol{w}) = \ell_{part}(\boldsymbol{\zeta}, \boldsymbol{w}) - \ell_{pen}(\theta, \boldsymbol{w}) \tag{2.22}$$

where,

$$\ell_{part}(\boldsymbol{\zeta}, \boldsymbol{w}) = \sum_{\ell=1}^{r} \left[ \sum_{t_{ij}=t_{(\ell)}} \eta_{ij} - N_{(\ell)} \log \left( \sum_{t_{qs} \geq t_{(\ell)}} \exp(\eta_{qs}) \right) \right]$$

with $\eta_{ij} = \boldsymbol{x}_{ij}^T \boldsymbol{\beta} + w_i$, and

$$\ell_{pen}(\theta, \boldsymbol{w}) = -\sum_{i=1}^{G} \log f_W(w_i).$$

In (2.22), the parameter $\theta$ plays the role of the tuning parameter in smoothing splines. But, unlike in the latter where the parameter is set by the user, $\theta$ has to be estimated from the data.
The details of this estimation approach under the log-normal and gamma frailty distributions are discussed next .

**Zero-mean normal density**

For the random effects $w_i$, $i = 1, \ldots, G$, we have (now think and write $\theta$ for the heterogeneity, i.e., Var(U)=$\sigma^2$=$\theta$ ):

$$\ell_{pen}(\theta, \boldsymbol{w}) = \frac{1}{2} \sum_{i=1}^{G} \left[ \frac{w_i^2}{\theta} + \log(2\pi\theta) \right]$$

and (2.22) becomes the penalized partial likelihood studied in McGilchrist (1993).
The random effects are thus in both parts of the penalized partial likelihood. The second term penalizes random effects that are far away from the mean value zero by reducing the penalized partial likelihood. This corresponds to shrinking the random effects towards the zero-mean.
The maximization of the penalized partial log-likelihood consists of an inner and an outer loop. In the inner loop, for a provisional value of $\theta$, the Newton-Raphson procedure is used to maximize $\ell_{pp\ell}(\boldsymbol{\zeta}, \boldsymbol{w})$ for $\boldsymbol{\beta}$ and $\boldsymbol{w}$ (best linear unbiased predictors, BLUP's). In the outer loop, the restricted maximum likelihood estimator for $\theta$ is obtained using the BLUP's. The process is iterated until convergence.
The details are as follows. Let $\ell$ denote the outer loop index and $k$ the inner loop index. Further let $\theta^{(\ell)}$ be the estimate for $\theta$ at the $\ell^{th}$ iteration in the outer loop. Given $\theta^{(\ell)}$, $\boldsymbol{\beta}^{(\ell,k)}$ and $\boldsymbol{w}^{(\ell,k)}$ are the estimates and predictions for $\boldsymbol{\beta}$ and $\boldsymbol{w}$ at the $k^{\text{th}}$ iterative step in the inner loop.
Starting from initial values $\boldsymbol{\beta}^{(1,0)}$, $\boldsymbol{w}^{(1,0)}$ and $\theta^{(1)}$ the $k^{\text{th}}$ iterative step for Newton-Raphson, given $\theta^{(\ell)}$, is given by

$$\begin{bmatrix} \boldsymbol{\beta}^{(\ell,k)} \\ \boldsymbol{w}^{(\ell,k)} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\beta}^{(\ell,k-1)} \\ \boldsymbol{w}^{(\ell,k-1)} \end{bmatrix} - \boldsymbol{V}^{-1} \begin{bmatrix} \mathbf{0} \\ \left[\theta^{(\ell)}\right]^{-1} \boldsymbol{w}^{(\ell,k-1)} \end{bmatrix} + \boldsymbol{V}^{-1} \begin{bmatrix} \boldsymbol{X} & \boldsymbol{Z} \end{bmatrix} \frac{d\ell_{part}(\boldsymbol{\zeta})}{d\boldsymbol{\eta}}$$

where

$$\boldsymbol{V} = \begin{bmatrix} \boldsymbol{V}_{11} & \boldsymbol{V}_{12} \\ \boldsymbol{V}_{21} & \boldsymbol{V}_{22} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}^T \\ \boldsymbol{Z}^T \end{bmatrix} \left( \frac{-d^2 \ell_{part}(\boldsymbol{\zeta})}{d\boldsymbol{\eta}\, d\boldsymbol{\eta}^T} \right) \begin{bmatrix} \boldsymbol{X} & \boldsymbol{Z} \end{bmatrix} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \left[\theta^{(\ell)}\right]^{-1} \boldsymbol{I}_G \end{bmatrix},$$

$\boldsymbol{X} = [\boldsymbol{x_{11}}, \ldots, \boldsymbol{x_{Gn_G}}]^T$ is an $n \times p$ covariate matrix with $n = \sum_{i=1}^{G} n_i$,
$\boldsymbol{Z} = \text{diag}\,(\mathbf{1}_{n_1}, \ldots, \mathbf{1}_{n_G})$ with $\mathbf{1}_{n_i}$ as a column vector of size $n_i$ with all entries one,
and $\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{w}$, such that $\boldsymbol{\eta}^T = (\eta_{11}, \ldots, \eta_{Gn_G})$.
Once the Newton Rahpson procedure has converged for the current value of $\theta^{(\ell)}$, a REML estimate for $\theta$ is given by

$$\theta^{(\ell+1)} = \frac{\sum_{i=1}^{G} \left( w_i^{(\ell,k)} \right)^2}{G - q}$$

where $q = \text{trace}\left[ \left( \boldsymbol{V}^{-1} \right)_{22} \right] / \theta^{(\ell)}$.
This outer loop is iterated until convergence based on the difference between the sequential values $\theta$ is obtained.

**Gamma density**

For random effects $w_i$, $i = 1, \ldots, G$, with corresponding gamma density (2.6) for the frailties, we have (now $\gamma = \theta$):

$$\ell_{pen}(\theta, \boldsymbol{w}) = -\sum_{i=1}^{G} \left( \frac{w_i - \exp(w_i)}{\theta} \right) - G \left( \frac{\log \theta}{\theta} - \log \Gamma \left( \frac{1}{\theta} \right) \right). \tag{2.23}$$

To maximize the penalized partial (log)likelihood we still can use an inner and outer loop. The inner loop is identical to the one described in the normal density case. Therefore, in the outer loop, a log-likelihood similar to $\ell_{obs}(\cdot)$ is maximized for $\theta$ as in the case of the EM-algorithm. This likelihood, for fixed value of $\theta^{(\ell)}$, also corresponds to $Q(\boldsymbol{\zeta}|\boldsymbol{\zeta}^{(k-1)}) - H(\boldsymbol{\zeta}|\boldsymbol{\zeta}^{(k-1)})$ evaluated at $(\theta^{(l)}, \hat{\boldsymbol{\beta}}_{\theta^{(l)}}, \hat{\boldsymbol{w}}_{\theta^{(l)}})$, where $\hat{\boldsymbol{\beta}}_{\theta^{(l)}}$ and $\hat{\boldsymbol{w}}_{\theta^{(l)}}$ are the estimates obtained by maximizing the penalized partial likelihood for fixed value of $\theta^{(l)}$. This expression is in terms of the partial likelihood containing the expectation of the frailties as fixed offset terms.
We therefore apply, in the outer loop, the golden section method (Brent, 1973) on the following modified version of the log-likelihood:

$$\ell_{obs}^{part}(\boldsymbol{\zeta},\boldsymbol{w}) \;=\; \ell_{part}(\boldsymbol{\zeta},\boldsymbol{w})$$

$$+\sum_{i=1}^{G}\left[\log\frac{\Gamma(D_i+1/\theta)}{\Gamma(1/\theta)}+\frac{1}{\theta}\log\left(\frac{1/\theta}{\Lambda_i+1/\theta}\right)-D_i\log(D_i+1/\theta)+D_i\right]. \quad (2.24)$$

That $\ell_{obs}^{part}(\boldsymbol{\zeta},\boldsymbol{w})$ is a modified version of $\ell_{obs}(\boldsymbol{\zeta})$ can be seen as follows: With $\Lambda_i = \sum_{j=1}^{n_i}\Lambda_0(t_{ij})\exp(\boldsymbol{x}_{ij}^T\boldsymbol{\beta})$ then the observable log-likelihood $\ell_{obs}(\boldsymbol{\zeta})$ given by (2.10) can be written as

$$\begin{aligned}\ell_{obs}(\boldsymbol{\zeta}) \;=\; & \sum_{i=1}^{G}\sum_{j=1}^{n_{ij}}\left[\delta_{ij}\left[\log\lambda_0(t_{ij})+\boldsymbol{x}_{ij}^T\boldsymbol{\beta}+\log E(\exp(w_i))\right]\right.\\ & \left.-\left[\Lambda_0(t_{ij})\exp(\boldsymbol{x}_{ij}^T\boldsymbol{\beta})E(\exp(w_i))\right]\right]\\ & +\sum_{i=1}^{G}\left[D_i\log\theta+\log\frac{\Gamma(D_i+1/\theta)}{\Gamma(1/\theta)}-(D_i+1/\theta)\log[\theta(\Lambda_i+1/\theta)]\right.\\ & \left.-D_i\log E(\exp(w_i))+\Lambda_i E(\exp(w_i))\right]. \quad (2.25)\end{aligned}$$

Now modify $\ell_{obs}(\boldsymbol{\zeta})$ by replacing the first (double) sum on the r.h.s of (2.25) by $\ell_{part}(\boldsymbol{\zeta},\boldsymbol{w})$. Using the relations $E(\exp(w_i)) = (D_i+1/\theta)/(\Lambda_i+1/\theta)$ and $\sum_{i=1}^{G}D_i = \sum_{i=1}^{G}\Lambda_i E(\exp(w_i))$ (martingale residuals sum to zero) it easily follows that the second sum in the r.h.s of (2.25) equals the second sum in the r.h.s of (2.24). This second relationship comes from the fact that corresponding to the frailty model (2.4) the quantity $M_{ij} = \delta_{ij}-\Lambda_0(t_{ij})\exp(\boldsymbol{x}_{ij}^T\boldsymbol{\beta}+w_i)$ is a martingale residual and $\sum_{i=1}^{G}\sum_{j=1}^{n_i}M_{ij}=0$ (Fleming and Harrington, 1991). The details for the penalized partial likelihood approach for the gamma are given in Flow Chart 2 in Section 2.6.

**Remark 6.** For the variance-covariance matrix of the estimates of $\boldsymbol{\beta}$ and $\boldsymbol{w}$, Therneau and Grambsch (2000) propose two estimates; $V_1 = H^{-1}IH^{-1}$ and $V_2 = H^{-1}$ where

$$H = \begin{pmatrix} \frac{\partial^2\ell_{ppl}(\boldsymbol{\zeta},\boldsymbol{w})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T} & \frac{\partial^2\ell_{ppl}(\boldsymbol{\zeta},\boldsymbol{w})}{\partial\boldsymbol{\beta}\partial\boldsymbol{w}^T} \\ \frac{\partial^2\ell_{ppl}(\boldsymbol{\zeta},\boldsymbol{w})}{\partial\boldsymbol{\beta}^T\partial\boldsymbol{w}} & \frac{\partial^2\ell_{ppl}(\boldsymbol{\zeta},\boldsymbol{w})}{\partial\boldsymbol{w}\partial\boldsymbol{w}^T} \end{pmatrix},$$

and $I$ is the usual Cox PH model information matrix. Under the log-normal frailty distribution McGilchrist (1993) proposes the estimate $2\hat{\theta}^2\left[G-2q+\hat{\theta}^{-2}\text{trace}\left[\left(\boldsymbol{V}^{-1}\right)_{22}\right]^2\right]^{-1}$ for the variance of the REML estimator for $\theta$ . No estimate is available under the gamma frailty distribution.

### 2.4.5 Equivalence of the EM and the penalized partial likelihood approach for the shared gamma frailty model

In this section we show that for the shared gamma frailty model, the estimates for $(\boldsymbol{\beta}, \theta, \Lambda_0(t))$ and predictions for $\boldsymbol{w}$ from the EM and penalized partial likelihood estimation procedures are the same. This equivalence is discussed in Therneau and Grambsch (2000, p.254) and Therneau et al. (2003). To achieve our objective, we shall reformulate $\ell_{part}(\boldsymbol{\zeta}, \boldsymbol{w})$ in (2.22) in the counting process notation. Under this formulation, the pair $(T_i^o, \delta_i)$ is replaced by $(N_i(t), R_i(t))$, where

$$N_i(t) \quad = \quad \text{the number of observed events in [0,t] for the } i^{th} \text{ individual}$$

$$R_i(t) = \begin{cases} 1 & \text{if } i^{\text{th}} \text{ unit is under observation and at risk at time t} \\ 0 & \text{otherwise.} \end{cases}$$

For right censored data $N_i(t) = I\{T_i^o \leq t, \delta_i = 1\}$ and $R_i(t) = I\{T_i^o \geq t\}$, such that $N_i(t)$ makes a jump of size 1 in case of an event at time $t$ while $R_i(t)$ changes from one to zero. In case of a censored event at time $t$, $N_i(t) \equiv 0$ while $R_i(t)$ still changes from one to zero. The pair $(N_i(t), R_i(t))$ is used in defining a martingale process for right censored data (Gill, 1984, Fleming and Harrington, 1991).

Let $n = \sum_{j=1}^{G} n_j$ be the total number of individuals and define $Z_{ij}$ to be unity if the $i^{\text{th}}$ individual comes from the $j^{\text{th}}$ group and zero otherwise. Further let $dN_i(t)$ be the change in $N_i(t)$ over the infinitesimal time interval $[t, t+dt)$, such that for non-tied data, $dN_i(t) = 1$ if an event occurs at time $t$ and $dN_i(t) = 0$ otherwise. It then follows that, for untied data, the partial likelihood is

$$PL(\boldsymbol{\beta}, \boldsymbol{w}) = \prod_{i=1}^{n} \prod_{t \geq 0} \left\{ \frac{R_i(t) \exp(\boldsymbol{x}_i^T \boldsymbol{\beta} + \boldsymbol{Z}_i^T \boldsymbol{w})}{\sum_{k=1}^{n} R_k(t) \exp(\boldsymbol{x}_k^T \beta + \boldsymbol{Z}_k^T \boldsymbol{w})} \right\}^{dN_i(t)}$$

where $\boldsymbol{Z}_i^T = (Z_{i1}, \ldots, Z_{iG})$ is a vector with a single entry one and all other entries zero and $\boldsymbol{w}^T = (w_1, \ldots w_G)$ and the denominator $\sum_k R_k(t) \exp(\boldsymbol{x}_k^T \boldsymbol{\beta} + \boldsymbol{Z}_k^T \boldsymbol{w})$ is the sum over all the individuals who are at risk at time $t$. Note that $i = 1, \ldots, n$ is a single index associated with all the individuals, unlike in the previous sections. The partial log-likelihood now becomes

$$\ell_{part}(\boldsymbol{\beta}, \boldsymbol{w}) = \sum_{i=1}^{n} \int_0^{\infty} R_i(t)(\boldsymbol{x}_i^T \boldsymbol{\beta} + \boldsymbol{Z}_i^T \boldsymbol{w}) dN_i(t)$$

$$-\sum_{i=1}^{n}\int_0^{\infty} log\left\{\sum_{k=1}^{n} R_k(t)\exp(\boldsymbol{x}_k^T\boldsymbol{\beta}+\boldsymbol{Z}_k^T\boldsymbol{w})\right\}dN_i(t). \tag{2.26}$$

For the penalized partial likelihood approach as seen above, one needs to maximize in the inner loop

$$\ell_{pp\ell}(\boldsymbol{\zeta},\boldsymbol{w}) = \ell_{part}(\boldsymbol{\beta},\boldsymbol{w}) - \ell_{pen}(\theta,\boldsymbol{w})$$

with respect to $\boldsymbol{\zeta}$ where $\ell_{part}(\boldsymbol{\beta},\boldsymbol{w})$ is as in (2.26) and $\ell_{pen}(\theta,\boldsymbol{w})$ is the penalty function given in (2.23). To obtain the estimate for $w_j$ we need to solve

$$\frac{\partial \ell_{pp\ell}(\boldsymbol{\zeta},\boldsymbol{w})}{\partial w_j} \equiv \frac{\partial \ell_{part}(\boldsymbol{\beta},\boldsymbol{w})}{\partial w_j} - \frac{\partial \ell_{pen}(\theta,\boldsymbol{w})}{\partial w_j} = 0.$$

Now

$$\frac{\partial \ell_{part}(\boldsymbol{\beta},\boldsymbol{w})}{\partial w_j} = \sum_{i=1}^{n}\int_0^{\infty}\left[Z_{ij} - \bar{Z}_j(t)\right]dN_i(t) \tag{2.27}$$

where

$$\bar{Z}_j(t) = \frac{\sum_{l=1}^{n} Z_{lj}R_l(t)\exp(\boldsymbol{x}_l^T\boldsymbol{\beta}+\boldsymbol{Z}_l^T\boldsymbol{w})}{\sum_{k=1}^{n} R_k(t)\exp(\boldsymbol{x}_k^T\boldsymbol{\beta}+\boldsymbol{Z}_k^T\boldsymbol{w})}$$

is the weighted mean of the $Z_{lj}$'s for all individuals in the $j^{\text{th}}$group who are at risk at time $t$, with $\frac{R_l(t)\exp(\boldsymbol{x}_l^T\boldsymbol{\beta}+\boldsymbol{Z}_l^T\boldsymbol{w})}{\sum_{k=1}^{n} R_k(t)\exp(\boldsymbol{x}_k^T\boldsymbol{\beta}+\boldsymbol{Z}_k^T\boldsymbol{w})}$ as the weights. On substituting this in (2.27), the second term $\sum_{i=1}^{n}\int_0^{\infty}\bar{Z}_j(t)dN_i(t)$ simplifies to $\sum_{l=1}^{n}\int_0^{\infty} R_l(t)Z_{lj}\exp(\boldsymbol{x}_l^T\boldsymbol{\beta}+\boldsymbol{Z}_l^T\boldsymbol{w})d\Lambda_0(t;\boldsymbol{\beta},\boldsymbol{w})$ where

$$d\Lambda_0(t;\boldsymbol{\beta},\boldsymbol{w}) = \frac{\sum_{i=1}^{n} dN_i(t)}{\sum_{k=1}^{n} R_k(t)\exp(\boldsymbol{x}_k^T\boldsymbol{\beta}+\boldsymbol{Z}_k^T\boldsymbol{w})} \tag{2.28}$$

is the Breslow estimator for the cumulative intensity function for given values of $\boldsymbol{\beta}$ and $\boldsymbol{w}$. It then follows on simplification that

$$\frac{\partial \ell_{part}(\boldsymbol{\beta},\boldsymbol{w})}{\partial w_j} = \sum_{i=1}^{n}\left[Z_{ij}\delta_i - Z_{ij}\Lambda_0(t_i;\boldsymbol{\beta},\boldsymbol{w})\exp(\boldsymbol{x}_i^T\boldsymbol{\beta}+\boldsymbol{Z}_i^T\boldsymbol{w})\right]$$

On taking into account that $Z_{ij}$ is a binary covariate we obtain

$$\frac{\partial \ell_{part}(\boldsymbol{\beta},\boldsymbol{w})}{\partial w_j} = [D_j - A_j\exp(w_j)]$$

where $A_j = \sum_{k=1}^{n_j} \Lambda_0(t_{kj};\boldsymbol{\beta},\boldsymbol{w})\exp(\boldsymbol{x}_k^T\boldsymbol{\beta})$ and $D_j = \sum_{i=1}^{n} Z_{ij}\delta_i$ is the number of events in the $j^{\text{th}}$ family. We note here that $A_j$ is equivalent to $\Lambda$ defined in (2.13) without the iteration step notation.

On the other hand

$$\frac{\partial \ell_{pen}(\theta,\boldsymbol{w})}{\partial w_k} = -\frac{1-\exp(w_k)}{\theta}.$$

Thus the score equation to be solved is

$$S(\boldsymbol{\zeta},\boldsymbol{w_j}) = \frac{\partial \ell_{pp\ell}(\boldsymbol{\zeta},\boldsymbol{w})}{\partial w_j} = [D_j - A_j\exp(w_j)] + \frac{1-\exp(w_j)}{\theta}.$$

On solving we get that

$$\exp(w_j) \;=\; \frac{D_j + \frac{1}{\theta^*}}{A_j + \frac{1}{\theta^*}} \tag{2.29}$$

which depends on $\theta^*$, a fixed value of $\theta$ (solution from outer loop) .

To get the estimator for $\boldsymbol{\beta}$ we only need to maximize $\ell_{part}(\boldsymbol{\zeta},\boldsymbol{w})$ given in (2.26) . This is a profile log-likelihood for $\boldsymbol{\beta}$ with $\boldsymbol{Z}^T\boldsymbol{w}$ as fixed offsets and standard methods for maximizing the Cox partial likelihood can be used.

Next we consider the estimates obtained from the EM approach which is discussed in Section 2.4.2. In the $k^{\text{th}}$ iteration of the E-step of the algorithm we saw that we needed to determine $E_k(\exp(w_i))$ which is given in (2.12). At convergence we have that the estimates for $\boldsymbol{\beta}$, $\theta$ and $\Lambda_0(\cdot)$ (say $\hat{\boldsymbol{\beta}},\hat{\theta},\hat{\Lambda}_0(\cdot)$) will satisfy

$$\exp(\hat{w}_i) \cong \frac{1/\hat{\theta} + D_i}{1/\hat{\theta} + \hat{\Lambda}_i} \tag{2.30}$$

where $\hat{\Lambda}_i$ is given in (2.13). We note that this is equivalent to (2.29), the solution from the penalized partial likelihood for fixed values of $\boldsymbol{\beta}$ and $\theta$ and with $\hat{\Lambda}_i$ replaced with $\hat{A}_i$. In the M-step of the algorithm, the estimator for $\boldsymbol{\beta}$ was obtained by maximizing the profile likelihood $Q_2(\boldsymbol{\beta} \mid \boldsymbol{\zeta})$ as given in (2.16). At the convergence of the algorithm, this will contain $\log(\exp(\hat{w}_i)) = \boldsymbol{Z}_i^T\hat{\boldsymbol{w}}$ as a fixed offset. Thus indeed the solution for $\boldsymbol{\beta}$ from the EM-algorithm will also be equivalent to that from the penalized partial likelihood approach. Finally, the estimate for $\Lambda_0(\cdot)$ is obtained by substituting the estimates for $\boldsymbol{\beta}$ and $E[\exp(w_s)]$ in (2.17). We thus get that

$$\Lambda_0(t) = \sum_{t_{(l)}\le t} \frac{N_{(l)}}{\sum_{t_{qs}\ge t_{(l)}} \exp(\boldsymbol{x}_{qs}^T\hat{\boldsymbol{\beta}} + \hat{w}_s)}.$$

which is the Breslow estimator and is equivalent to (2.28).

## 2.5 Discussion

In this chapter we have looked at the various methods of estimation that have been described in the literature to fit semi-parametric shared frailty models. We have focused on the EM-algorithm and penalized partial likelihood estimation methods. Above we only present the methodological aspects of the estimation procedures. We do not however compare the performance of the various approaches. Indeed for a specific frailty distribution, it would be interesting to assess the performance of each of the various approaches, through simulations. This would give a guideline on what bias and variability to expect when using a specific approach.

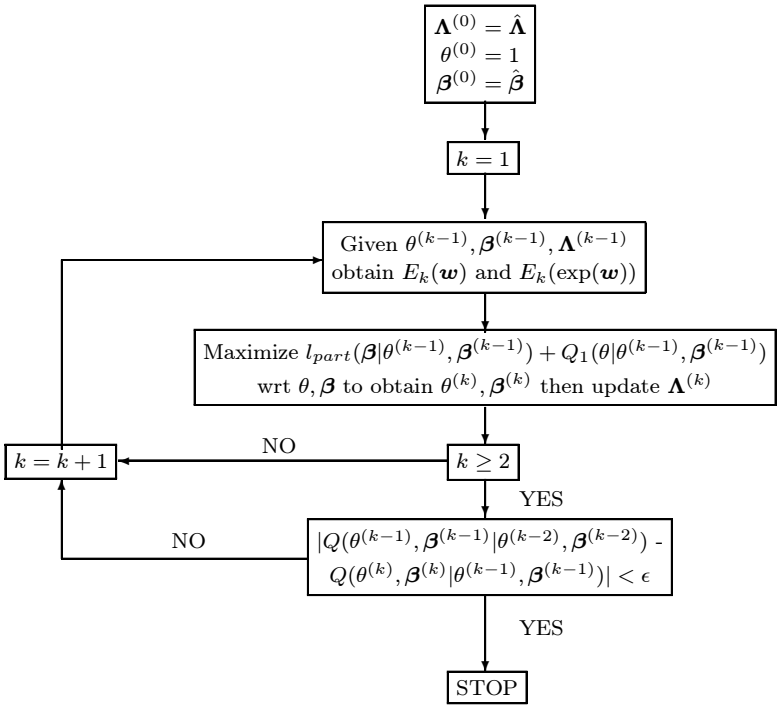
Currently only few standard softwares can be used to fit frailty models. The penalized partial likelihood is the method of estimation implemented in S-plus . Recently an additional subroutine that runs on a UNIX port has been implemented in S-plus for the semi-parametric frailty model. This routine fits correlated frailties for both log-normal and gamma frailties (Therneau, 2003). Parametric frailty models on the other hand can be fitted also in S-plus as well as in Stata.

Above we have only considered the gamma and log-normal frailty distributions. The choice of the frailty distribution is often governed by the problem at hand in terms of the model implications. Each distribution has its own desirable properties. The gamma distribution is more often than not used due to its mathematical convinience while the log-normal distribution is frequently used in situations with correlated frailties. The latter is a popular choice in animal-breeding studies where clusters may not necessarily be independent. The stable distribution on the other hand (though not considered here) has the desireable property that the marginal hazard model (frailties integrated out) retains the proportionality assumption. This is not the case with the gamma and the form of the deviation from propotionality is unknown under the log-normal frailty model (Hougaard, 2000, p. 245).

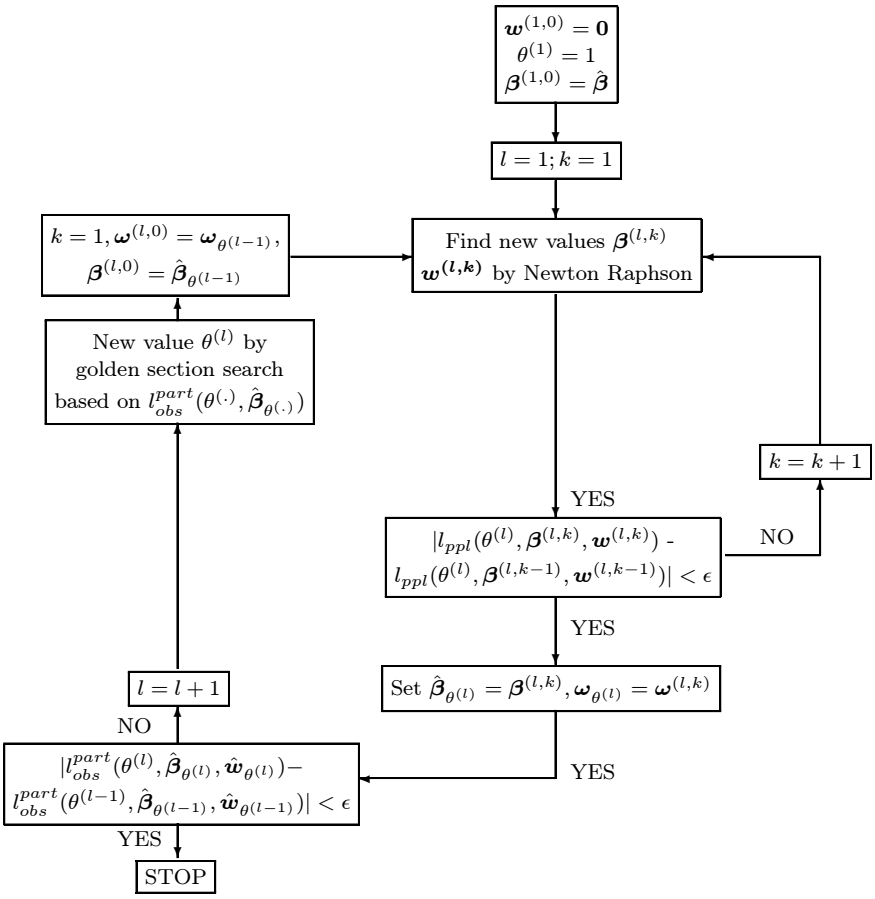
We have also highlighted the identifiability issues that are considered for the frailty model especially for univariate data. In the specific examples presented, a gamma frailty distribution is used. For identifiability to hold, the mean of the frailty is restricted to be unity. Under the log-normal frailty model, this restriction (i.e. $E[U] = 1$) would mean that $E[W] = -\sigma^2/2$ for the model to be identifiable (see Hougaard, 2000, p. 244, Colosimo and Oliveira, 2002).

Lastly, we remark that full Bayesian estimation approaches for the frailty models exist although we do not consider them in this text. The various approaches in the literature within this context, differ mostly in the modelling aspect of the baseline hazard or as above, in the frailty distribution. A detailed review of Bayesian estimation of frailty models is given by Ibrahim et al. (2001).

## 2.6 Appendix




**Flow Chart 1.** $\widehat{\boldsymbol{\beta}}$ are the estimates for the regression coefficients in the classical Cox regression model (without frailties); and $\widehat{\boldsymbol{\Lambda}} = (\widehat{\Lambda}_1, \ldots, \widehat{\Lambda}_G)$ where $\widehat{\Lambda}_i$ is obtained from (2.17) with $\boldsymbol{\beta^{(0)}}$ and $E_0(\exp(w_i))$=1. $E_k(\boldsymbol{w}) = (E_k(w_1), \ldots, E_k(w_G))$, etc .




**Flow Chart 2.** The penalized likelihood approach for the semi-parametric gamma frailty model. For details, see Section 2.4.4.

# Chapter 3

# Likelihood ratio and score tests for a shared frailty model: a non-standard problem

## 3.1 Introduction

In spite of the fact that shared frailty models are very useful to describe and to model multivariate survival data, the inferential properties are not yet well examined. The main reason is that, due to the complexity of the modelling, it is very hard to derive statistical properties (e.g. asymptotic properties) for frailty models in general, see e.g. Murphy (1994, 1995). As seen in Chapter 2 the complexity lies in the fact that the likelihood expressions needed for the inference are implicit and difficult, so that in many situations numerical algorithms are needed to obtain estimates and standard errors.

We saw in Section 2.3 that the underlying concept in the frailty models is that the failure times of individuals in a cluster are dependent, while those across clusters are independent. One of the important methodological questions is to provide information on the asymptotic distributional behavior of the likelihood ratio test for heterogeneity (between cluster variability). To test for heterogeneity we consider the following hypotheses testing problem. Assume that the random effect, present in the shared frailty model, has variance

$\theta$. The relevant hypotheses are:

$$H_0 : \theta = 0 \quad \text{versus} \quad H_a : \theta > 0. \tag{3.1}$$

which is a one-sided hypotheses testing problem. The theory of testing such hypotheses has a long history, going back to Chernoff (1954) and has been extensively studied in linear mixed models. By now it is well known that for linear mixed models, the asymptotic distribution theory for the likelihood ratio (see Self and Liang, 1987, Stram and Lee, 1994, 1995) and score statistic (Verbeke and Molenberghs, 2003) for this testing problem does not follow the classical chi-square limit theory. The reason is that, under the null hypothesis, the parameter of interest is at the boundary of the parameter space (in the alternative hypothesis the heterogeneity parameter is subject to an inequality constraint). As a consequence the classical conditions needed for the likelihood ratio theory are not satisfied. We therefore need to develop "likelihood ratio theory under non-standard conditions".

This phenomenon has been recognized in the literature on frailty models. Vaida and Xu (2000, p. 3322) write "for the likelihood ratio test a correction for the null distribution, which is no longer a chi-square distribution, is needed as discussed in similar set-ups by Stram and Lee (1994) and Self and Liang (1987) in the context of mixed effects models". Duchateau et al. (2002) simulate the limit distribution of the likelihood ratio test and conjecture that the simulated distribution is a 50:50 mixture of a $\chi^2_0$ and a $\chi^2_1$ distribution.

In most of the estimation procedures considered in the last chapter we have seen that we need to estimate the cumulative baseline hazard in addition to the other parameters of the semi-parametric shared frailty model. As our main interest in this chapter is the heterogeneity parameter we shall derive the asymptotic null distribution for the likelihood ratio test and the score test for heterogeneity for the shared gamma frailty model with a Weibull baseline hazard. In Section 3.2, a precise description of the model is given. In Section 3.3 we give, for complete data (no censoring), the asymptotic distribution of the likelihood ratio statistic for this testing problem. We first consider, for transparency of the proof, bivariate complete data without covariates (Section 3.3.1). Bjarnason and Hougaard (2000) use this model to study the Fisher information matrix. The idea behind the simplification is to fully understand the statistical properties for a simple, though relevant, model. In Section 3.3.2 we consider the more general case that includes covariates.

The definition and distribution of the score statistic is dealt with in Section 3.4. The proofs are given in Section 3.5; key references are Vu and Zhou (1997) and Silvapulle and Silvapulle (1995). A short introductory discussion on possible extensions to censored data is given in Section 3.6. We finish the chapter with a discussion section (Section 3.8).

## 3.2 Complete data model

We observe a set of $n$ independent random vectors $\boldsymbol{T}_i = (T_{i1}, T_{i2})$, $i = 1, \ldots, n$. Each vector is considered as a cluster of size two, as in twin studies where $T_{i1}$ and $T_{i2}$ are the observed times for the first and second twin in the $i^{\text{th}}$ cluster. We assume that, conditional on the frailty variables $U_i = u$, the lifetimes $T_{i1}$ and $T_{i2}$ are independent with a Weibull distribution, i.e., the conditional hazard is

$$\lambda(t \mid u, x) = u\lambda\rho t^{\rho-1} \exp(\boldsymbol{x}^T\boldsymbol{\beta})$$

with $\lambda > 0$ and $\rho > 0$. Further $U_i$ is taken to have the gamma density

$$f_{U_i}(u) = \frac{u^{\frac{1}{\theta}-1} \exp\left(-\frac{u}{\theta}\right)}{\theta^{\frac{1}{\theta}}\Gamma\left(\frac{1}{\theta}\right)}.$$

as in (2.6) but with $\text{Var}(U_i) = \theta$.

The key idea is that within cluster dependence is caused by the frailty variables $U_1, \ldots, U_n$ representing unobserved common risk factors. For example, unobserved genetic and environmental effects in twin studies or subject effects when the eyes of an individual are considered. Given $U_i = u$, the conditional survival function of $(T_{i1}, T_{i2})$ is

$$\begin{aligned} S(t_1, t_2 \mid u, \boldsymbol{x}_1, \boldsymbol{x}_2) &= P(T_{i1} > t_1, T_{i2} > t_2 \mid u, \boldsymbol{x}_1, \boldsymbol{x}_2) \\ &= \exp\left[-u\left\{\lambda_1 t_1^\rho \exp(\boldsymbol{x}_1^T\boldsymbol{\beta}) + \lambda_2 t_2^\rho \exp(\boldsymbol{x}_2^T\boldsymbol{\beta})\right\}\right]. \end{aligned}$$

The (unconditional) survival function is

$$\begin{aligned} S(t_1, t_2 \mid \boldsymbol{x}_1, \boldsymbol{x}_2) &= E\left[\exp\{-u\left(\lambda_1 t_1^\rho \exp(\boldsymbol{x}_1^T\boldsymbol{\beta}) + \lambda_2 t_2^\rho \exp(\boldsymbol{x}_2^T\boldsymbol{\beta})\right)\}\right] \\ &= \{1 + \theta(\lambda_1 t_1^\rho \exp(\boldsymbol{x}_1^T\boldsymbol{\beta}) + \lambda_2 t_2^\rho \exp(\boldsymbol{x}_2^T\boldsymbol{\beta})\}^{-\frac{1}{\theta}}. \end{aligned}$$

The corresponding joint density is

$$f(t_1, t_2 \mid \boldsymbol{x}_1, \boldsymbol{x}_2) = \frac{(1+\theta)\lambda_1\lambda_2\rho^2 t_1^{\rho-1} \exp(\boldsymbol{x}_1^T\boldsymbol{\beta}) t_2^{\rho-1} \exp(\boldsymbol{x}_2^T\boldsymbol{\beta})}{\{1 + \theta\left[\lambda_1 t_1^\rho \exp(\boldsymbol{x}_1^T\boldsymbol{\beta}) + \lambda_2 t_2^\rho \exp(\boldsymbol{x}_1^T\boldsymbol{\beta})\right]\}^{\frac{1}{\theta}+2}}.$$

For $\theta > 0$ (heterogeneity between clusters) the components of the vector $(T_{i1}, T_{i2})$ are correlated (within cluster correlation). To quantify the within cluster dependence we can use Kendall's coefficient of concordance as mentioned in Section 2.3.2. For our model Kendall's coefficient of concordance is $\theta/(2+\theta)$ which is zero for $\theta = 0$. Moreover we easily obtain that

$$\lim_{\theta \underset{>}{\to} 0} f(t_1, t_2 \mid \boldsymbol{x}_1, \boldsymbol{x}_2) = \prod_{j=1}^{2} \lambda_j \rho t_j^{\rho-1} \exp(\boldsymbol{x}_j^T \boldsymbol{\beta}) \exp\left(-\lambda_j t_j^{\rho} \exp(\boldsymbol{x}_j^T \boldsymbol{\beta})\right),$$

i.e., $T_{i1}$ and $T_{i2}$ are independent Weibull distributed random variables with scale and shape parameters $\lambda_j \exp(\boldsymbol{x}_{ij}^T \boldsymbol{\beta})$ and $\rho$, respectively.

The marginal likelihood for the data is given by

$$\prod_{i=1}^{n} \frac{(1+\theta)\lambda_1 \lambda_2 \rho^2 T_{i1}^{\rho-1} \exp(\boldsymbol{x}_{i1}^T \boldsymbol{\beta}) T_{i2}^{\rho-1} \exp(\boldsymbol{x}_{i2}^T \boldsymbol{\beta})}{\left\{1 + \theta\left[\lambda_1 T_{i1}^{\rho} \exp(\boldsymbol{x}_{i1}^T \boldsymbol{\beta}) + \lambda_2 T_{i2}^{\rho} \exp(\boldsymbol{x}_{i2}^T \boldsymbol{\beta})\right]\right\}^{\frac{1}{\theta}+2}}.$$

with corresponding log-likelihood

$$\begin{aligned} L_n = \sum_{i=1}^{n} \{ & \log(1+\theta) + \log\lambda_1 + \log\lambda_2 + 2\log\rho + \boldsymbol{\beta}^T(\boldsymbol{x}_{i1} + \boldsymbol{x}_{i2}) + (\rho-1)(\log T_{i1} + \log T_{i2}) \\ & - \left(\frac{1}{\theta} + 2\right) \log\left(1 + \theta\left[\lambda_1 T_{i1}^{\rho} \exp(\boldsymbol{x}_{i1}^T \boldsymbol{\beta}) + \lambda_2 T_{i2}^{\rho} \exp(\boldsymbol{x}_{i2}^T \boldsymbol{\beta})\right]\right)\}. \end{aligned} \tag{3.2}$$

To test the within cluster dependence we consider the hypothesis testing problem stated in (3.1). To focus attention on the main ideas we assume $\lambda_1 = \lambda_2 = \lambda$ and take $\boldsymbol{x}$ as univariate, i.e., $\boldsymbol{x} = x$. Moreover, for further discussion it is convenient to work with the following transformed Weibull parameters: $\eta = -\log\lambda$ and $\alpha = -\log\rho$. With this transformation then each of the nuisance parameters $(\eta, \alpha, \beta)$ is in $\mathbb{R}$. We use $\tau$ as shorthand notation for the set of model parameters $(\theta, \eta, \alpha, \beta)$ and $\nu$ for $(\eta, \alpha, \beta)$. In terms of $\tau$ the parameter space is $\Theta = [0, \infty) \times \mathbb{R}^3$ and the testing problem can be written as

$$H_0 : \tau \in \Theta_0 = \{0\} \times \mathbb{R}^3 \text{ against } H_a : \tau \in \Theta_1 = (0, \infty) \times \mathbb{R}^3. \tag{3.3}$$

## 3.3 The likelihood ratio statistic

The likelihood ratio statistic for testing (3.3), i.e., for testing cluster dependence is

$$d_n = 2\left\{\sup_{\tau \in \Theta} L_n(\tau) - \sup_{\tau \in \Theta_0} L_n(\tau)\right\}$$

Under the null hypothesis the parameter vector of interest is at the boundary of the parameter space. Therefore the standard asymptotic distribution theory for likelihood ratio tests does not work. In Section 3.3.1 we obtain the asymptotic distribution for $d_n$ for the special case when there are no covariates. In Section 3.3.2 we consider complete data with covariates. Different settings of the observed covariates will be considered.

### 3.3.1 Complete survival data, no covariates

For complete bivariate data with no covariates the testing problem (3.3) reduces to the more simple form

$$H_0 : (\theta, \eta, \alpha) \in \Theta_0 = \{0\} \times \mathbb{R}^2 \text{ against } H_a : (\theta, \eta, \alpha) \in \Theta_1 = (0, \infty) \times \mathbb{R}^2 \tag{3.4}$$

The corresponding log-likelihood is

$$L_n(\theta, \eta, \alpha) = \sum_{i=1}^{n} \left[ -2(\eta + \alpha) + \log(1 + \theta) + (e^{-\alpha} - 1)(\log T_{i1} + \log T_{i2}) \right] - \left( \frac{1}{\theta} + 2 \right) \sum_{i=1}^{n} \log \left\{ 1 + \theta e^{-\eta} (T_{i1}^{e^{-\alpha}} + T_{i2}^{e^{-\alpha}}) \right\}.$$

**Theorem 1.** *The likelihood ratio statistic $d_n$ for testing the one-sided heterogeneity hypothesis (3.4) in the shared gamma frailty model with Weibull baseline hazard has an asymptotic null distribution which is an equal mixture of a point mass at zero and a chi-square distribution with one degree of freedom, that is, $d_n \to_d \frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$ as $n \to \infty$.*

We give the proof of this theorem in Section 3.5. The first and second derivatives of $L_n(\tau)$ needed for the proof are given in Section 3.9, where $\tau = (\theta, \eta, \alpha, \beta)$ and $\nu = (\eta, \alpha, \beta)$. The structure of the proof makes clear how to deal with boundary parameters, as well as with the one-sided aspect of the testing problem.

Vu and Zhou (1997) give a set of conditions under which a general result holds on the asymptotic behaviour of likelihood ratio tests where, under the null hypothesis, the true values are allowed to lie on the boundary of the parameter space. These conditions are outlined with the proof in Section 3.5.

This result has an immediate impact on how to determine (asymptotic) critical values and $P$-values for likelihood ratio tests for heterogeneity. Erroneously relying on chi-square

distribution theory as for two-sided tests, leads to a conservative strategy in rejecting the null hypothesis of no heterogeneity.

### 3.3.2 Complete survival data, including covariates

It is possible to generalize Theorem 1 to the situation of non-identically distributed observations. This allows the distribution of the lifetimes $T_{ij}, j = 1, 2$ to depend on covariate information. The log-likelihood now is

$$L_n(\tau) = \sum_{i=1}^{n} \{\log(1+\theta) - 2(\eta+\alpha) + \beta(x_{i1}+x_{i2}) + (e^{-\alpha}-1)(\log T_{i1} + \log T_{i2}) \\ - \left(\frac{1}{\theta}+2\right)\log\left(1+\theta e^{-\eta}\left[\, T_{i1}^{e^{-\alpha}}\exp(\beta x_{i1}) + T_{i2}^{e^{-\alpha}}\exp(\beta x_{i2})\right]\right)\},$$

with $\tau = (\theta, \eta, \alpha, \beta)$.

Let $\boldsymbol{F}_n(\tau)$ be the matrix of the negative of the second derivatives of $L_n(\tau)$ with respect to $\tau$and define $\boldsymbol{G}_n(\nu) = E[\boldsymbol{F}_n(0,\nu)]$ $(\nu = (\eta, \alpha, \beta))$. In the covariate-free case $(\nu = (\eta, \alpha))$ the Fisher information matrix $\boldsymbol{G}_n(\nu)$ only depends on the parameter $\eta$ and its determinant is independent of the parameters (see proof of Theorem 1 in Section 3.5). However in the presence of covariates the matrix $\boldsymbol{G}_n(\nu)$ depends on the coefficient $\beta$. Restrictions on the allowable range of values for $\beta$ are needed for the asymptotic theory to go through. These restrictions depend on the type of covariates considered.

To explain the issue we take the same modelling situation as in the previous section, yet with a covariate added. Denote $M_{k,l} = \sum_{i=1}^{n}\sum_{j=1}^{2} x_{ij}^k \exp(l\beta x_{ij})$ and
$N_{k,l} = \sum_{i=1}^{n}\exp\{\beta(kx_{i1} + lx_{i2})\}$. Further let $\psi$ be the digamma function and $\zeta(2,q) = \int_0^\infty \frac{te^{-qt}}{1-e^{-t}}dt$. With these notations the entries of the symmetric matrix $\boldsymbol{G}_n(\nu)$ on suppressing the dependence on subscript $n$ are:

$G_{1,1} = n + 4(M_{0,3} - M_{0,2} + N_{2,1} + N_{1,2} - N_{1,1})$

$G_{2,2} = M_{0,1}$

$G_{3,3} = -2n(\psi(1)+\eta) + (\psi(2)+\eta)M_{0,1} + \{(\psi(2)+\eta)^2 + \zeta(2,2)\}M_{0,1}$

$G_{4,4} = M_{2,1}$

$G_{1,2} = 2N_{1,1} + 2M_{0,2} - 2M_{0,1}$

$G_{1,3} = -2(\psi(2)+\eta)M_{0,1} + 2M_{0,2}(\psi(3)+\eta) + 2(\psi(2)+\eta)N_{1,1}$

$G_{1,4} = 2M_{1,1} - 2M_{1,2} - \sum_{i=1}^{n}(x_{i1}+x_{i2})\exp\{\beta(x_{i1}+x_{i2})\}$

$G_{2,3} = (\psi(2)+\eta)M_{0,1}$

$G_{2,4} = -M_{1,1}$

$G_{3,4} = -(\psi(2) + \eta)M_{1,1}$

A direct calculation shows that the determinant of this matrix is independent of the value of $\eta$, but does depend on $\beta$. We now consider two applications.

*Example 1.* Suppose $n$ is even, $x_{i1}$ is a binary covariate, half of the observations $x_{i1} = 1$ with corresponding $x_{i2} = 0$. This situation occurs for example for a treatment of the eyes when only one eye, either left or right, is treated and the other serves as a control. A specification of the matrix $\boldsymbol{G}_n(\nu)$ gives that the entries of $\tilde{\boldsymbol{G}} \equiv \tilde{\boldsymbol{G}}(\nu) = \boldsymbol{G}_n(\nu)/n$ are given by

$\tilde{G}_{1,1} = 1 + 4\exp(3\beta)$

$\tilde{G}_{2,2} = 1 + \exp(\beta)$

$\tilde{G}_{3,3} = -2(\psi(1) + \eta) + \{\psi(2) + \eta + (\psi(2) + \eta)^2 + \zeta(2,2)\}\{1 + \exp(\beta)\}$

$\tilde{G}_{4,4} = \exp(\beta)$

$\tilde{G}_{1,2} = 2\exp(2\beta)$

$\tilde{G}_{1,3} = -2(\psi(2) + \eta)\{1 + \exp(\beta)\} + 2(\psi(3) + \eta)\{1 + \exp(2\beta)\} + 2(\psi(2) + \eta)\exp(\beta)$

$\tilde{G}_{1,4} = \exp(\beta) - 2\exp(2\beta)$

$\tilde{G}_{2,3} = (\psi(2) + \eta)\{1 + \exp(\beta)\}$

$\tilde{G}_{2,4} = -\exp(\beta)$

$\tilde{G}_{3,4} = -(\psi(2) + \eta)\exp(\beta).$

Direct calculation shows that the determinant of $\tilde{\boldsymbol{G}}(\nu)$ depends on both $\beta$ and $\eta$ and can take on both positive and negative values. It is a linear function of $\eta$ but depends on $\beta$ in a more complex manner. In particular we have

$$\det(\tilde{G}) = -e^{5\beta} + e^{4\beta}(3\eta + 3.203) - e^{3\beta}(4\eta - 3.599) + e^{2\beta}(2\eta - 1.154) - e^{\beta}(\eta - 1.222)$$

The allowable range of $\beta$ values for the matrix $\boldsymbol{G}_n(\nu)$ to be positive definite increases with increasing values of $\eta$ (decreasing $\lambda = \exp(-\eta)$). See Figure 3.1. In particular, when $\eta = 2$, then we require that $\beta < 2.1674$ which corresponds to a risk difference between treatment and non-treatment of size 8.736.

*Example 2.* Fixed covariate design. For $v_i = \frac{(i-1)}{(n-1)}$, $i = 1, \ldots, n$, generate the covariate values $(x_{i1}, x_{i2})$ as follows: $x_{i1} = F_1^{-1}(v_i)$ and $x_{i2} = F_2^{-1}(v_i)$ where $F_1$ and $F_2$ are given distribution functions. It is now easy to compute limiting expressions for the entries of the matrix $\boldsymbol{G}_n$. For example, $n^{-1}M_{k,l} \to \sum_{j=1}^{2} E[X_j^k \exp(l\beta X_j)]$ where $X_j \sim F_j$. Also in this case, restrictions on the regression coefficient are needed for the matrix $\boldsymbol{G}_n$ to be

positive definite.

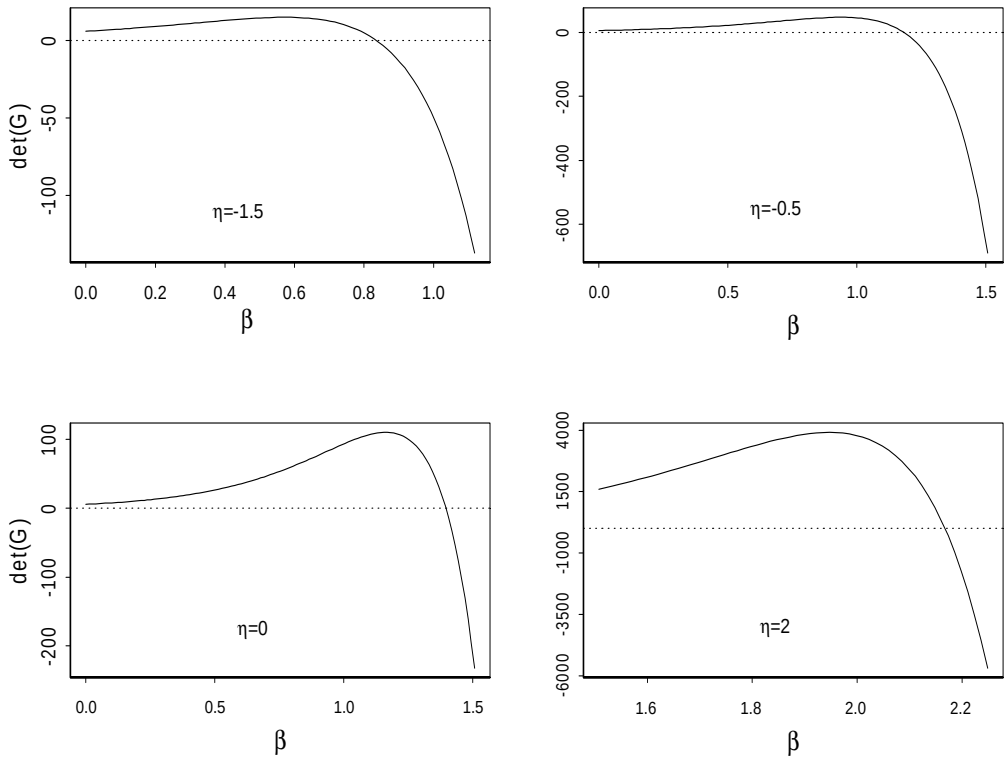



Figure 3.1: *Determinant of $\tilde{G}$ as a function of the regression coefficient $\beta$ and $\eta$ for a gamma frailty model with a binary covariate as in Example 1.*

**Theorem 2.** *Assume that the Fisher information matrix $\boldsymbol{G}_n$ is positive definite and that $\|\beta x\|$ is finite uniformly over the coefficient $\beta$ and the covariate space of $x$. Then the likelihood ratio statistic $d_n$ for testing the one-sided heterogeneity hypothesis (3.3) in the shared gamma frailty model with Weibull baseline hazard has an asymptotic null distribution which is an equal mixture of a point mass at zero and a chi-square distribution with one degree of freedom, that is, $d_n \to_d \frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$ as $n \to \infty$.*

## 3.4 The score tests

In this section we formulate the asymptotic distribution of the score test statistic for complete survival data with no covariates. If parameters constrained under the null hypothesis

belong to the interior of the parameter space, it is well known that likelihood ratio, Wald and score statistics have asymptotically the same distribution under the null hypothesis. Under inequality constraints in the alternative hypothesis, a score statistic is no longer uniquely defined, see Silvapulle and Silvapulle (1995). These authors propose a different score-type statistic which only requires estimation under the null hypothesis. Under mild regularity conditions, they obtain that under the null hypothesis asymptotically the score statistic follows the same mixture distribution as the likelihood ratio statistic.
In general the score test statistic for a two-sided testing problem is

$$S_n(\tau_0)G_n^{-1}(\tau_0)S_n^T(\tau_0)$$

where $\tau_0$ is the true value of $\tau$ and $S_n(\tau_0)$ is the score vector evaluated at $\tau = \tau_0$.
The explicit expressions for the components of the score vector $S_n(\tau) = (S_{n,\theta}(\tau), S_{n,\nu}(\tau))$ with $S_{n,\theta}(\tau) = \partial L_n(\tau)/\partial\theta$ and $S_{n,\nu}(\tau) = (\partial L_n(\tau)/\partial\eta, \partial L_n(\tau)/\partial\alpha, \partial L_n(\tau)/\partial\beta)^T$ are given in Section 3.9.
Via a Taylor series expansion Silvapulle and Silvapulle (1995) rewrite the score statistic as the difference of the minimum of two quadratic forms, of which the minimisation of the first one is under the null hypothesis which can be performed exactly. We state the resulting score statistic in the following theorem. Let $\hat{\nu}$ be the maximum likelihood estimator of the nuisance parameters under the null hypothesis and let $S_{n,\theta}(0,\hat{\nu})$ denote the score vector evaluated at $(0,\hat{\nu})$.

**Theorem 3**

*(i) For a shared gamma frailty model with exponential baseline hazard the score statistic for testing the heterogeneity hypothesis (3.4) is given by*

$$\mathcal{S}_n = \frac{1}{3n^2}\{S_{n,\theta}(0,\hat{\eta})\}^2 - 3n \inf_{b\geq 0}\left\{\left(\frac{1}{3n^{3/2}}S_{n,\theta}(0,\hat{\eta}) - b\right)^2\right\}.$$

*(ii) For a shared gamma frailty model with Weibull baseline hazard function a score statistic for testing the heterogeneity hypothesis (3.4) is given by*

$$\mathcal{S}_n = \frac{1}{3n^2}\frac{\pi^2}{\pi^2-4}\{S_{n,\theta}(0,\hat{\nu})\}^2 - 3n\left(1-\frac{4}{\pi^2}\right)\inf_{b\geq 0}\left\{\left(\frac{1}{3n^{3/2}}\frac{\pi^2}{\pi^2-4}S_{n,\theta}(0,\hat{\nu}) - b\right)^2\right\}.$$

*For both models the corresponding score statistic has, under the null hypothesis, asymptotic distribution* $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$.

The proof is also given in the next section. Note that in the exponential baseline hazard model when $S_{n,\theta}(0,\hat{\eta}) \geq 0$, the expression of the score statistic simplifies to

$$\mathcal{S}_n = \frac{1}{3n^2}\{S_{n,\theta}(0,\hat{\eta})\}^2.$$

A similar simplification holds for the Weibull baseline hazard model.

## 3.5 Proofs

Vu and Zhou (1997) in their Theorem 2.2 give a set of conditions under which a general result holds on the asymptotic behaviour of likelihood ratio tests where, under the null hypothesis, the true values are allowed to be on the boundary of the parameter space. For the model specified in Section 3.3.1 , we will show that their set of conditions is satisfied. First define $\boldsymbol{D}_n(\nu) = E[S_n^T(0,\nu)S_n(0,\nu)]$ and let $\boldsymbol{G}_n(\nu)$ and $\boldsymbol{F}_n(\tau)$ be as defined in Section 3.3.2. but with $\tau = (\theta,\eta,\alpha)$ and $\nu = (\eta,\alpha)$. As derived in Section 3.9 we have for the shared frailty model with Weibull baseline hazard that

$$\boldsymbol{G}_n(\nu) = n\boldsymbol{G}(\eta) = n\begin{pmatrix} 5 & 2 & 2(2-\rho_e+\eta) \\ 2 & 2 & 2(1-\rho_e+\eta) \\ 2(2-\rho_e+\eta) & 2(1-\rho_e+\eta) & \pi^2/3+2(1-\rho_e+\eta)^2 \end{pmatrix} \tag{3.5}$$

with $\gamma_e$ the Euler constant. From Property 2 in Section 3.9 we know that $\boldsymbol{D}_n(\nu) = \boldsymbol{G}_n(\nu)$, a property that we expect since our likelihood is based on a sample of independent and identically distributed vectors.

The following are the necessary conditions of Vu and Zhou (1997). To avoid ambiguity in notation, they are presented in terms of an unknown parameter $\boldsymbol{\vartheta} \in \Theta \subset \mathbb{R}^k$, with $\Theta$ as the parameter space. Further $\boldsymbol{\vartheta}_0$ is the true unknown parameter value.

(A1) For a neighborhood $\mathcal{N}$ of $\boldsymbol{\vartheta}_0$ the log-likelihood function $L_n(\boldsymbol{\vartheta})$ is continuous on $\boldsymbol{\vartheta} \cap \aleph$ and the first and second derivatives of $L_n(\boldsymbol{\vartheta})$ exist, are finite and continuous

(A2) For a subset $\Omega$ of $\Theta$ there exists a closed cone $C_\Omega$ with vertex at $\boldsymbol{\vartheta}_0$ such that

$$\inf_{\boldsymbol{x}\in C_\Omega}\left|G_n^{T/2}(\boldsymbol{x}-\boldsymbol{y})\right| \leq \mu(\boldsymbol{y})\left|G_n^{T/2}(\boldsymbol{y}-\boldsymbol{\vartheta}_0)\right| \quad \text{for } \mu(\boldsymbol{y}) \in \Omega$$

and

$$\inf_{\boldsymbol{y}\in\Omega}\left|G_n^{T/2}(\boldsymbol{x}-\boldsymbol{y})\right| \leq \nu(\boldsymbol{x})\left|G_n^{T/2}(\boldsymbol{x}-\boldsymbol{\vartheta}_0)\right| \quad \text{for } \nu(\boldsymbol{x}) \in C_\Omega$$

with $\mu(\boldsymbol{y}) \to \mathbf{0}$ as $\boldsymbol{y} \to \boldsymbol{\vartheta}_0$ and $\nu(\boldsymbol{x}) \to \mathbf{0}$ as $\boldsymbol{x} \to \boldsymbol{\vartheta}_0$.

(A3) For a subset $\Omega$ of $\Theta$ there exists a closed cone $\widetilde{C}_\Omega$ with vertex at 0, not depending on $n$ such that the sets of transformed cones

$$\widetilde{C}_{\Omega_n} = \left\{\widetilde{\boldsymbol{\vartheta}} : \widetilde{\boldsymbol{\vartheta}} = \boldsymbol{G}_n^{T/2}(\boldsymbol{\vartheta} - \boldsymbol{\vartheta}_0), \boldsymbol{\vartheta} \in C_\Omega\right\}$$

asymptotically coincide with $\widetilde{C}_\Omega$ in the sense that as $n \to \infty$

$$\sup_{|\boldsymbol{b}|=1} \left| \inf_{\boldsymbol{\vartheta} \in \widetilde{C}_{\Omega_n}} |\boldsymbol{b} - \boldsymbol{\vartheta}|^2 - \inf_{\boldsymbol{\vartheta} \in \widetilde{C}_\Omega} |\boldsymbol{b} - \boldsymbol{\vartheta}|^2 \right| \to 0.$$

(B1) $E[S_n(\boldsymbol{\vartheta}_0)] = \mathbf{0}$, and $\boldsymbol{D}_n(\boldsymbol{\vartheta}_0)$ and $\boldsymbol{G}_n(\boldsymbol{\vartheta}_0)$ are finite.

(B2) $\kappa_{\min}(\boldsymbol{G}_n(\boldsymbol{\vartheta}_0)) \to \infty$, where $\kappa_{\min}(\boldsymbol{A})$ denotes the smallest eigenvalue of a symmetric positive definite matrix $\boldsymbol{A}$

(B3) $\sup_{\boldsymbol{\vartheta} \in N_n(A)} \|\boldsymbol{G}_n^{-1/2}(\boldsymbol{\vartheta}_0)\boldsymbol{F}_n(\boldsymbol{\vartheta})\boldsymbol{G}_n^{-T/2}(\boldsymbol{\vartheta}_0) - \boldsymbol{I}_k\|_1 = o_P(1)$
where $N_n(A) = \left\{\boldsymbol{\vartheta} : (\boldsymbol{\vartheta} - \boldsymbol{\vartheta}_0)^T \boldsymbol{G}_n(\boldsymbol{\vartheta} - \boldsymbol{\vartheta}_0) \leq A^2, \boldsymbol{\vartheta} \in \Theta, for A > 0\right\}$
and $\|\boldsymbol{W}\|_1$ is the sum of the absolute values of the elements of a matrix $\boldsymbol{W}$.

(B4) For some positive definite matrix $\boldsymbol{V}$, $\|\boldsymbol{G}_n^{-1/2}(\boldsymbol{\vartheta}_0)\boldsymbol{D}_n(\boldsymbol{\vartheta}_0)\boldsymbol{G}_n^{-T/2}(\boldsymbol{\vartheta}_0) - \boldsymbol{V}_k\|_1 \to 0$.

(B5) $(\boldsymbol{D}_n(\boldsymbol{\vartheta}_0))^{-1/2} S_n(\boldsymbol{\vartheta}_0) \to_d N(0, \boldsymbol{I}_k)$.

**Note 1**. If

$$\liminf_{n\to\infty} \frac{\kappa_{\min}(\boldsymbol{G}_n(\boldsymbol{\vartheta}_0)}{\kappa_{\max}(\boldsymbol{G}_n(\boldsymbol{\vartheta}_0)} > 0.$$

where $\kappa_{\max}(\boldsymbol{A})$ denotes the largest eigenvalue of a symmetric positive definite matrix $\boldsymbol{A}$, then (A2) is equivalent to the Chernoff regularity (Chernoff, 1954). Geyer (1994) shows that each convex set is Clarke regular (Clarke, 1983) and that Clarke regularity is stronger than Chernoff regularity.

**Note 2**. When (B2) holds, $\boldsymbol{G}_n(\boldsymbol{\vartheta}_0)$ is positive definite for $n$ large enough.

**Note 3**. When (B2) and (B4) hold, $\boldsymbol{D}_n(\boldsymbol{\vartheta}_0)$ is positive definite for $n$ large enough.

### 3.5.1 Proof of Theorem 1

The main issue is to check the conditions (A1)-(A3) and (B1)-(B5) needed for the validity of Theorem 2.2 of Vu and Zhou (1997). For the problem at hand, the parameter of interest is $\tau$ and the subsets under the null and alternative hypothesis of the parameter space $\Theta$ are $\Theta_0$ and $\Theta_1$ respectively.

(A1) The log-likelihood function, score vector and components of the matrix of second derivatives of the log-likelihood (see Section 3.9 for explicit expressions) are continuous and finite on a neighborhood of the true parameter value $\tau_0 = (0, \nu_0)$. For the score component and second derivatives of the log-likelihood with respect to $\theta$ the boundedness is shown by an expansion of the logarithmic function in the second term of equation (3.17) in Section 3.9. For the other derivatives the result is straightforward to obtain.

(A2) The closed cones with vertex at $\tau_0$ for $\Theta_0$ and $\Theta_1$ are $C_{\Theta_0} \equiv \Theta_0$ and $C_{\Theta_1} \equiv \Theta$.
From Property 1 in Section 3.9 we have that

$$\liminf_{n\to\infty} \frac{\kappa_{\min}(\boldsymbol{G}_n(\nu))}{\kappa_{\max}(\boldsymbol{G}_n(\nu))} = \frac{\kappa_{\min}(\boldsymbol{G}(\eta))}{\kappa_{\max}(\boldsymbol{G}(\eta))} > 0.$$

It therefore suffices to show the Chernoff regularity, which is satisfied since the parameter space $\Theta_1 = (0, \infty) \times \mathbb{R}^3$ is convex (Geyer, 1994).
(A3) The transformed cones, used to obtain the asymptotic distribution of the likelihood ratio test, are for $j = 0, 1$

$$\widetilde{C}_{n,\Theta_j} = \left\{ (\widetilde{\theta}, \widetilde{\eta}, \widetilde{\alpha}) = \boldsymbol{G}_n^{T/2}(\nu) \begin{pmatrix} \theta \\ \eta \\ \alpha \end{pmatrix} \text{ with } \begin{pmatrix} \theta \\ \eta \\ \alpha \end{pmatrix} \in C_{\Theta_j} \right\}$$

with $\boldsymbol{G}_n^{1/2}(\nu)$ and $\boldsymbol{G}_n^{T/2}(\nu)$ the left and the corresponding right Cholesky square root of $\boldsymbol{G}_n(\nu)$. A direct calculation shows that $\boldsymbol{G}_n^{T/2}(\nu) = n^{1/2}\boldsymbol{G}^{T/2}(\eta)$ with

$$\boldsymbol{G}^{T/2}(\eta) = \begin{bmatrix} \sqrt{5} & 2/\sqrt{5} & 2(2 - \gamma_e + \eta)/\sqrt{5} \\ 0 & \sqrt{6/5} & \sqrt{2}(1 - 3\gamma_e + 3\eta)/\sqrt{15} \\ 0 & 0 & \sqrt{(\pi^2 - 4)/3} \end{bmatrix}$$

We therefore have

$$\begin{aligned} \widetilde{C}_{n,\Theta_0} &= \left\{ (\widetilde{\theta}, \widetilde{\eta}, \widetilde{\alpha}) : \widetilde{\theta} - \sqrt{\frac{2}{3}}\widetilde{\eta} - \frac{2\sqrt{5}}{\sqrt{3(\pi^2 - 4)}}\widetilde{\alpha} = 0 \right\} \equiv \widetilde{C}_{\Theta_0} \\ \widetilde{C}_{n,\Theta_1} &= \left\{ (\widetilde{\theta}, \widetilde{\eta}, \widetilde{\alpha}) : \widetilde{\theta} - \sqrt{\frac{2}{3}}\widetilde{\eta} - \frac{2\sqrt{5}}{\sqrt{3(\pi^2 - 4)}}\widetilde{\alpha} \geq 0 \right\} \equiv \widetilde{C}_{\Theta_1}. \end{aligned}$$

Since $\widetilde{C}_{n,\Theta_j} \equiv \widetilde{C}_{\Theta_j}$, condition (A3) holds.

(B1) Based on the expressions in Section 3.9 we can easily show that $E[S_n(0,\nu)] = 0$ and we know that $\boldsymbol{D}_n(\nu) = \boldsymbol{G}_n(\nu)$ are finite matrices.

(B2) $\kappa_{\min}(\boldsymbol{G}_n(\nu)) = n\kappa_{\min}(\boldsymbol{G}(\eta)) \to \infty$, $n \to \infty$. This follows from Property 1 in Section 3.9.

(B3) For $(0,\nu_0)$, the true parameter value, define

$$N_n(A) = \left\{ \tau = (\theta,\eta,\alpha) : (\theta, \eta-\eta_0, \alpha-\alpha_0)\boldsymbol{G}_n(\nu_0) \begin{pmatrix} \theta \\ \eta-\eta_0 \\ \alpha-\alpha_0 \end{pmatrix} \leq A^2, \tau \in \Theta \right\}.$$

To prove (B3) we need to show that

$$\sup_{\tau \in N_n(A)} \|\boldsymbol{G}_n^{-1/2}(\nu_0)\boldsymbol{F}_n(\tau)\boldsymbol{G}_n^{-T/2}(\nu_0) - \boldsymbol{I}_3\|_1 = o_P(1). \tag{3.6}$$

with $P$ a shorthand notation for $P_{\tau_0}$ ($\tau_0 = (0,\nu_0) \in \Theta_0$, the true value of the parameter under the null hypothesis). Note that (recall equation (3.5))

$$\begin{aligned} &\boldsymbol{G}_n^{-1/2}(\nu_0)\boldsymbol{F}_n(\tau)\boldsymbol{G}_n^{-T/2}(\nu_0) - \boldsymbol{I}_3 \\ &\quad = \boldsymbol{G}^{-1/2}(\eta_0)\left(\frac{\boldsymbol{F}_n(0,\nu_0) - \boldsymbol{G}_n(\nu_0)}{n}\right)\boldsymbol{G}^{-T/2}(\eta_0) \\ &\qquad + \boldsymbol{G}^{-1/2}(\eta_0)\left(\frac{\boldsymbol{F}_n(\tau) - \boldsymbol{F}_n(0,\nu_0)}{n}\right)\boldsymbol{G}^{-T/2}(\eta_0). \end{aligned} \tag{3.7}$$

Note that, for matrices $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$, $\|\boldsymbol{W}_1\boldsymbol{W}_2\|_1 \leq \|\boldsymbol{W}_1\|_1\|\boldsymbol{W}_2\|_1$. Since $\|\boldsymbol{G}^{-1/2}(\eta_0)\|_1 = \|\boldsymbol{G}^{-T/2}(\eta_0)\|_1 \leq C(\eta_0)$, with $0 < C(\eta_0) < \infty$, (3.6) follows by showing that

$$\left\| \frac{\boldsymbol{F}_n(0,\nu_0)}{n} - \boldsymbol{G}(\eta_0) \right\|_1 = o_P(1) \tag{3.8}$$

and

$$\sup_{\tau \in N_n(A)} \left\| \frac{\boldsymbol{F}_n(\tau) - \boldsymbol{F}_n(0,\nu_0)}{n} \right\|_1 = o_P(1). \tag{3.9}$$

To establish the validity of (3.8) we need the entries of $\boldsymbol{F}_n(0,\nu_0)$ given in Section 3.9. For each entry we apply the law of large numbers to obtain

$$\left| \left( \frac{\boldsymbol{F}_n(0,\nu_0)}{n} \right)_{[i,j]} - (\boldsymbol{G}(\eta_0))_{[i,j]} \right| = o_P(1).$$

Hence (3.8) is valid.

To establish (3.9) we need the entries of $\boldsymbol{F}_n(\tau)$, the negative of the second derivatives of $L_n$, which are also given in Section 3.9. For each entry we need to show that

$$\sup_{\tau \in N_n(A)} \left| \left( \frac{\boldsymbol{F}_n(\tau) - \boldsymbol{F}_n(0, \nu_0)}{n} \right)_{[i,j]} \right| = o_P(1). \tag{3.10}$$

We show how to prove (3.10) for the $[2,2]$-entry.

$$\left( \frac{\boldsymbol{F}_n(\tau) - \boldsymbol{F}_n(0, \nu_0)}{n} \right)_{[2,2]} = \frac{1}{n} \sum_{i=1}^{n} H_{22}(T_i, \tau)$$

where, with $U_i = T_{i1}^{e^{-\alpha}} + T_{i2}^{e^{-\alpha}}$ (as defined in Section 3.9 but with $\beta = 0$),

$$H_{22}(T_i, \tau) = -\theta e^{-2\eta}(1 + 2\theta) \frac{U_i^2}{(1 + \theta e^{-\eta} U_i)^2} + e^{-\eta}(1 + 2\theta) \frac{U_i}{1 + \theta e^{-\eta} U_i} - e^{-\eta_0} U_i.$$

Note that $H_{22}(T_i, \tau_0) \equiv 0$. There exists a fixed positive integer $n_0$ such that for all $n \geq n_0$

$$\sup_{\tau \in N_n(A)} e^{-\alpha} \leq K \equiv 2(e^{-\alpha_0} + 1)$$

and, for some constant $D > 0$,

$$|H_{22}(T_i, \tau)| < D(T_{i1}^{2K} + T_{i2}^{2K}).$$

With $\mu(\tau) = E_{\tau_0} H(T_i, \tau)$ we have by the dominated convergence theorem that $\lim_{\tau \to \tau_0} \mu(\tau) = \mu(\tau_0) \equiv 0$. Now the proof of (3.9) follows since

$$\begin{aligned} &\sup_{\tau \in N_n(A)} \left| \frac{1}{n} \sum_{i=1}^{n} H_{22}(T_i, \tau) \right| \\ &\quad \leq \sup_{\tau \in N_n(A)} \left| \frac{1}{n} \sum_{i=1}^{n} H_{22}(T_i, \tau) - \mu(\tau) \right| + \sup_{\tau \in N_n(A)} |\mu(\tau)| = o_P(1). \end{aligned} \tag{3.11}$$

An application of Theorem 16(a) in Ferguson (1996), p. 108 implies indeed that the first term in the right-hand side of (3.11) is $o_P(1)$ (uniform law of large numbers); elementary analysis implies that the second term in the right-hand side of (3.11) is $o(1)$.

Similar proofs hold for all the other entries of $(\boldsymbol{F}_n(\tau) - \boldsymbol{F}_n(0, \nu_0))/n$.

(B4) The matrix $\boldsymbol{V} = \boldsymbol{I}_3$ and (B4) holds since $\boldsymbol{G}_n(\nu) = \boldsymbol{D}_n(\nu)$.

(B5) Since $(T_{i1}, T_{i2})$, $i = 1, \ldots, n$, are independent and identically distributed vectors, an

application of classical multivariate central limit theorem theory gives that $\boldsymbol{G}_n^{-1/2}(\nu_0)S_n(0,\nu_0) \to_d N = (N_1, N_2, N_3)$ which has a multivariate normal with mean vector zero and covariance matrix $\boldsymbol{I}_3$.

Since the Vu and Zhou (1997) conditions (A1)-(A3) and (B1)-(B5) are valid, an application of their Theorem 2.2 gives that the asymptotic null distribution of $d_n$, the likelihood ratio statistic, is the same as the distribution of

$$\inf_{\widetilde{\tau}\in\widetilde{C}_{\Theta_0}} |N-\widetilde{\tau}|^2 - \inf_{\widetilde{\tau}\in\widetilde{C}_{\Theta_1}} |N-\widetilde{\tau}|^2 \tag{3.12}$$

where $\widetilde{\tau} = (\widetilde{\theta}, \widetilde{\eta}, \widetilde{\alpha})$.

From the definitions $\widetilde{C}_{\Theta_0}$ and $\widetilde{C}_{\Theta_1}$ we have

$$\inf_{\widetilde{\tau}\in\widetilde{C}_{\Theta_0}} |N-\widetilde{\tau}|^2 = \left(\frac{N_1 + aN_2 + bN_3}{\sqrt{1+a^2+b^2}}\right)^2 \tag{3.13}$$

with $a = -\sqrt{\frac{2}{3}}$ and $b = -\frac{2\sqrt{5}}{\sqrt{3(\pi^2-4)}}$ (see the proof of condition (A3)), i.e., the random variable in (3.13) has a chi-square distribution with one degree of freedom. We further have

$$\inf_{\widetilde{\tau}\in\widetilde{C}_{\Theta_1}} |N-\widetilde{\tau}|^2 = \begin{cases} 0 & N \in \widetilde{C}_{\Theta_1} \\ \left(\dfrac{N_1 + aN_2 + bN_3}{\sqrt{1+a^2+b^2}}\right)^2 & N \notin \widetilde{C}_{\Theta_1} \end{cases} \tag{3.14}$$

Moreover we have $P(N \in \widetilde{C}_{\Theta_1}) = 0.5$. This, together with (3.12) - (3.14) implies that the asymptotic distribution of the likelihood ratio test is $0.5\chi_0^2 + 0.5\chi_1^2$.

### 3.5.2 Proof of Theorem 2

This follows along the lines of the proof of Theorem 1. The main differences arises only from the fact that $\boldsymbol{G}_n$ depends on both $\eta$ and $\beta$. Write $\boldsymbol{G}_n(\eta,\beta) = n\left(n^{-1}\boldsymbol{G}_n(\eta,\beta)\right)$. Then the conditions on $x$ and $\beta$ assure that $\boldsymbol{G}(\eta,\beta) = \lim_{n\to\infty} n^{-1}\boldsymbol{G}_n(\eta,\beta)$ exists and that $\boldsymbol{G}(\eta,\beta)$ is a symmetric positive definite matrix. Its Cholesky decomposition leads to cones $\widetilde{C}_{n,\Theta_j}(j=0,1)$ of which the limiting cones are defined using the matrix $\boldsymbol{G}(\eta,\beta)$. The remaining part of the proof holds under the boundedness assumption on $\beta$ and on the covariate $x$.

### 3.5.3 Proof of Theorem 3

We first state the general form of the score statistic to test the heterogeneity hypothesis. Partition the Fisher information matrix $\boldsymbol{G}_n(\nu)$ such that the upper left block corresponds to the parameter $\theta$ constrained to zero under the null hypothesis and the lower right block is defined by the nuisance parameters $\nu$. Specifically,

$$\boldsymbol{G}_n(\nu) = \begin{pmatrix} G_{n,00}(\nu) & G_{n,01}(\nu) \\ G_{n,01}^T(\nu) & G_{n,11}(\nu) \end{pmatrix}.$$

Further, define $G_n^{00}(\nu) = (\boldsymbol{G}_n^{-1}(\nu))_{00} = (G_{n,00}(\nu) - G_{n,01}(\nu)G_{n,11}^{-1}(\nu)G_{n,01}^T(\nu))^{-1}$, let $\hat{\nu}$ be the maximum likelihood estimator of the nuisance parameters under the null hypothesis and let $S_{n,\theta}(0,\hat{\nu})$ denote the score vector evaluated at $(0,\widehat{\nu})$.

Silvapulle and Silvapulle (1995) define the score statistic

$$\begin{aligned}\mathcal{S}_n &= n^{-1}S_{n,\theta}^T(0,\hat{\nu})G_n^{00}(0,\hat{\nu})S_{n,\theta}(0,\hat{\nu}) \\ &\quad - \inf_{b\geq 0}\left\{\left(n^{-1/2}G_n^{00}(0,\hat{\nu})S_{n,\theta}(0,\hat{\nu}) - b\right)^T\{G_n^{00}(0,\hat{\nu})\}^{-1}\left(n^{-1/2}G_n^{00}(0,\hat{\nu})S_{n,\theta}(0,\hat{\nu}) - b\right)\right\}.\end{aligned}$$

They show that assuming the existence of a matrix $\boldsymbol{H} = \boldsymbol{H}(\tau)$ such that for $n \to \infty$

(C1) $n^{-1/2}S_n(\tau) \to_d N(0, \boldsymbol{H}(\tau))$,

and for any $a > 0$

(C2)

$$\sup_{\|h\|\leq a}\left[n^{-1/2}\{S_n(\tau + n^{-1/2}h) - S_n(\tau)\} + \boldsymbol{H}(\tau)h\right] = o_p(1),$$

the likelihood ratio and score statistic for testing (3.4) have asymptotically the same distribution. Condition (C2) ensures the existence of the score statistic within a small neighbourhood of $\tau_0$. It is also clear from (C1) and (C2) that $\boldsymbol{H}(\tau)$ is essentially $\boldsymbol{G}_n(\tau)$. For the case of a shared gamma frailty model with an exponential baseline hazard there is only the nuisance parameter $\eta$ (or $\lambda = \exp(-\eta)$). For this special case, with Fisher information matrix

$$\boldsymbol{G}_n = n\begin{pmatrix} 5 & 2 \\ 2 & 2 \end{pmatrix},$$

we have that $G_n^{00} = (3n)^{-1}$, not dependent on any nuisance parameters, and hence we obtain the following score statistic

$$\mathcal{S}_n = \frac{1}{3n^2}\{S_{n,\theta}(0,\hat{\eta})\}^2 - 3n\inf_{b\geq 0}\left\{\left(\frac{1}{3n^{3/2}}S_{n,\theta}(0,\hat{\eta}) - b\right)^2\right\}$$

For the Weibull baseline hazard the nuisance parameter is $\nu = (\eta, \alpha)$ and the Fisher information matrix is $\boldsymbol{G}_n(\nu)$ as given in (3.5), from which it is deduced that $G_n^{00} = \pi^2/(3n(\pi^2 - 4))$. Hence the resulting score statistic is obtained as given in Theorem 3(ii).

## 3.6 Extension to censored data

In this section we set an initial step to the study of the likelihood ratio test for heterogeneity in the case of censored data. Like in the preceeding sections we assume that we have bivariate data. For each of the $n$ individuals we observe $\boldsymbol{T}_i^o = (T_{i1}^o, T_{i2}^o)$, $i = 1, \ldots, n$, where $T_{ij}^o = \min(T_{ij}, C_{ij})$, $j = 1, 2$, and $T_{ij}$ and $C_{ij}$ are the time to failure and censoring time respectively. Also observed are the censoring indicators $\delta_{ij} = 1$ if $T_{ij}^o = T_{ij}$ and zero otherwise. For simplicity we assume that $\lambda_1 = \lambda_2 = \rho = 1, (\eta = \alpha = 0)$ and no covariates. Since the heterogeneity parameter is the only remaining unknown parameter, the likelihood ratio test for $H_0 : \theta = 0$ versus $H_a : \theta > 0$ is

$$d_n = 2\{\sup_{\theta \geq 0} L_n(\theta) - L_n(0)\}. \tag{3.15}$$

The log-likelihood $L_n(\tau)$ reduces to

$$\begin{aligned} L_n(\theta) \;=\; & -\frac{1}{\theta} \sum_{\{i:D_i=0\}} \log\{1 + \theta(T_{i1}^o + T_{i2}^o)\} - \frac{1}{\theta+1} \sum_{\{i:D_i=1\}} \log\{1 + \theta(T_{i1}^o + T_{i2}^o)\} \\ & -\frac{1}{\theta+2} \sum_{\{i:D_i=2\}} \log\{1 + \theta(T_{i1}^o + T_{i2}^o)\} + N_2 \log(1+\theta) \end{aligned} \tag{3.16}$$

where $D_i = \sum_{j=1}^{2} \delta_{ij}$ and $N_2 = \sharp\{i : D_i = 2\}$.

This log-likelihood is, a particular example of the likelihood and the log-likelihood expression given on pages 1476 and 1490 of Murphy and van der Vaart (1997) with $\lambda(t) \equiv 1$ and $\tau = \infty$. Indeed an equivalent way to describe the information contained in $(T_{ij}^o, \delta_{ij})$, $i = 1, \ldots, n$, $j = 1, 2$, is through the counting process formulation introduced in Section 2.4.5. We now have

$$N_i(t) = \sum_{j=1}^{2} N_{ij}(t) = \sum_{j=1}^{2} 1\{T_{ij}^o \leq t, \delta_{ij} = 1\}$$

and the risk process

$$R_i(t) = \sum_{j=1}^{2} R_{ij}(t) = \sum_{j=1}^{2} 1\{T_{ij}^o \geq t\}.$$

It then easily follows that

$$\frac{[\{1+\theta N_i(t-)\}R_i(t)]^{\Delta N_i(t)}}{(1+\theta\int_0^\infty R_i(t)dt)^{1/\theta+N(\infty)}} = \begin{cases} \{1+\theta(T_{i1}^o+T_{i2}^o)\}^{-1/\theta-D_i} & \text{for } D_i=0 \text{ or } 1 \\ (1+\theta)\{1+\theta(T_{i1}^o+T_{i2}^o)\}^{-1/\theta-D_i} & \text{for } D_i=2 \end{cases}$$

and hence $L_n(\theta)$ is a special case of the log-likelihood expression in Murphy and van der Vaart (1997).

**Theorem 4.** *The likelihood ratio statistic $d_n$ for testing the one-sided heterogeneity hypothesis (3.15) in the shared gamma frailty model with constant baseline hazard, has an asymptotic null distribution which is an equal mixture of a point mass at zero and a chi-square distribution with one degree of freedom, that is, $d_n \to_d \frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$ as $n \to \infty$.*

**Proof.**

It can easily be shown from (3.16) that

$$\begin{aligned} \frac{d^2}{d\theta^2}L_n(\theta) &= -\frac{2}{\theta^3}\sum\log(1+\theta U_i) + \frac{2}{\theta^2}\sum_{i=1}^n \frac{U_i}{1+\theta U_i} + \frac{1}{\theta}\sum_{i=1}^n \frac{U_i^2}{(1+\theta U_i)^2} \\ &\quad + \sum_{\{i:D_i=1\}} \frac{U_i^2}{(1+\theta U_i)^2} + 2\sum_{\{i:D_i=2\}} \frac{U_i^2}{(1+\theta U_i)^2} - \frac{N_2}{(1+\theta)^2} \end{aligned}$$

where $U_i = T_{i1}^o + T_{i2}^o$.

On expanding the logarithmic term and simplifying we have

$$\frac{d^2}{d\theta^2}L_n(0) = -\frac{2}{3}\sum U_i^3 + \sum_{\{i:D_i=1\}} U_i^2 + 2\sum_{\{i:D_i=2\}} U_i^2 - N_2.$$

It then follows that since $G = \lim_{n\to\infty} E\left[-\frac{1}{n}\frac{d^2}{d\theta^2}L_n(0)\right]$ is positive, the log-likelihood function $L_n(\theta)$ is concave in a closed neighbourhood $\mathcal{N}$ of zero. Let $\hat{\theta}_n = \arg\max_{\theta\in\mathcal{N}} L_n(\theta)$. For $\hat{\theta}_n$ the following properties are immediate from the more general Theorem 2 in Murphy (1994) and Theorem 1 in Murphy (1995): under the null hypothesis the likelihood estimator $\hat{\theta}$ converges strongly to zero and the distribution of $\sqrt{n}\hat{\theta}_n$ tends to a normal limit with zero mean and bounded limit variance $G^{-1}$. As a result we have $\lim_{n\to\infty} P(\hat{\theta}_n \le 0) = 0.5$. We wish to determine $P(d_n \le c)$ under the null hypothesis. If $c < 0$ then $P(d_n \le c) = 0$. Next we consider the case when $c \ge 0$. Write

$$P(d_n \le c) = P(d_n \le c \mid \hat{\theta}_n > 0)P(\hat{\theta}_n > 0) + P(d_n \le c \mid \hat{\theta}_n \le 0)P(\hat{\theta}_n \le 0).$$

If $\hat{\theta}_n \leq 0$ then $d_n = 0$ and therefore under the null hypothesis we have $P(d_n \leq c \mid \hat{\theta}_n \leq 0) = 1$. For $\hat{\theta}_n > 0$ we consider the Taylor expansion of $d_n$ which yields

$$d_n = 2(L_n(\hat{\theta}_n) - L_n(0) = -\hat{\theta}_n^2 \frac{d^2}{d\theta^2} L_n(\tilde{\theta}),$$

with $\tilde{\theta}_n$ as an intermediate point in $(0, \hat{\theta}_n)$. Since $\hat{\theta}_n \to_P 0$ then also $\tilde{\theta}_n \to_P 0$. By the weak law of large numbers with estimated parameters (Iverson and Randles, 1989), it suffices that

$$-\frac{1}{n}\frac{d^2}{d\theta^2} L_n(\tilde{\theta}) \to_P G$$

as $n \to \infty$. It then follows that $Z_n = \sqrt{n}\hat{\theta}_n \left(-\frac{1}{n}\frac{d^2}{d\theta^2} L_n(\tilde{\theta})\right) \to_d Z$, where $Z \sim N(0,1)$. Thus

$$P(d_n \leq c \mid \hat{\theta}_n > 0) P(\hat{\theta}_n > 0) = P(0 < Z_n \leq \sqrt{c}) \to \frac{1}{2} P(Z^2 \leq c).$$

Hence under the null hypothesis

$$P(d_n \leq c) \to \frac{1}{2} P(Z^2 \leq c) + \frac{1}{2}.$$

## 3.7 Data example

We consider a data set from the Diabetic Retinopathy Study (Huster et al. 1989) to test for heterogeneity by considering time to blindness in each eye of 197 patients with diabetic retinopathy. One eye of each patient was randomly selected for treatment and the other was observed without treatment. Of the 197 patients, 80 patients the times to event were censored for both eyes. For these patients, both eyes were censored at the same time, but the times varied from patient to patient. Further, blindness was observed only in the untreated eye for 63 patients, and only in the treated eye for 16 patients. The remaining 44 patients went blind and out of this simultaneous blindness in both eyes was observed in 6 patients.

We fitted a Weibull baseline hazard model with a gamma frailty distribution. A binary covariate indicating a treated/non-treated eye was included in the model. This model is more general than that considered in Section 3.6 as we use a Weibull baseline hazard ($\lambda_0(t) = \lambda\rho t^{\rho-1}$) and also include covariates. Explorative analysis (Figure 3.2) shows that the Weibull baseline hazard is an acceptable parametric model for these data. The results are given in Table 3.1 below.

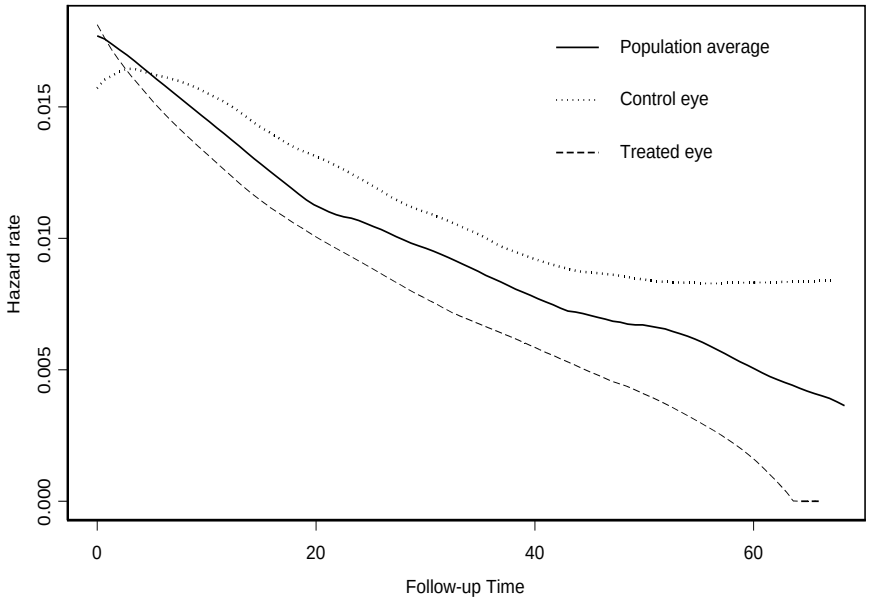



Figure 3.2: *Non-parametric hazard estimate for the Diabetic Retinopathy data.*

Table 3.1: *Parameter estimates (s.e.) and likelihood values for a Weibull gamma frailty model for the Diabetic Retinopathy data.*

| | Parameter | | | | Log | LRT |
|---|---|---|---|---|---|---|
| | $\theta$ | $\lambda$ | $\rho$ | $\beta$ | likelihood | p-value |
| Full model | 0.712±0.145 | 0.011±0.190 | 0.888±0.006 | 0.382±0.046 | -841.272 | 0.0006 |
| Reduced model | | 0.015±0.126 | 0.799±0.005 | 0.280±0.027 | -846.499 | |

Based on the above results we reject the null hypothesis $H_0 : \theta = 0$ for no heterogeneity. We thus conclude that the time to blindness was affected by some unobserved additional patient characteristics after taking into account the treatment. The presence of heterogeneity is confirmed by the construction of a profile likelihood based confidence interval for $\theta$. For a given value of $\theta$ (say $\theta^f$) we maximize $L_n(\theta^f, \lambda, \rho, \beta)$ with respect to $\lambda$, $\rho$ and $\beta$. Denote these maximizers as $\lambda(\theta^f)$, $\rho(\theta^f)$ and $\beta(\theta^f)$. We then obtain the profile likelihood $L_n(\theta^f, \lambda(\theta^f), \rho(\theta^f), \beta(\theta^f))$ and follow the method explained in Morgan (1992) to determine the profile likelihood based 95 % confidence interval (c.i) for $\theta$. From Figure 3.3 the estimated approximate 95% c.i is (0.245,1.323). Hence, heterogeneity is present in these data.

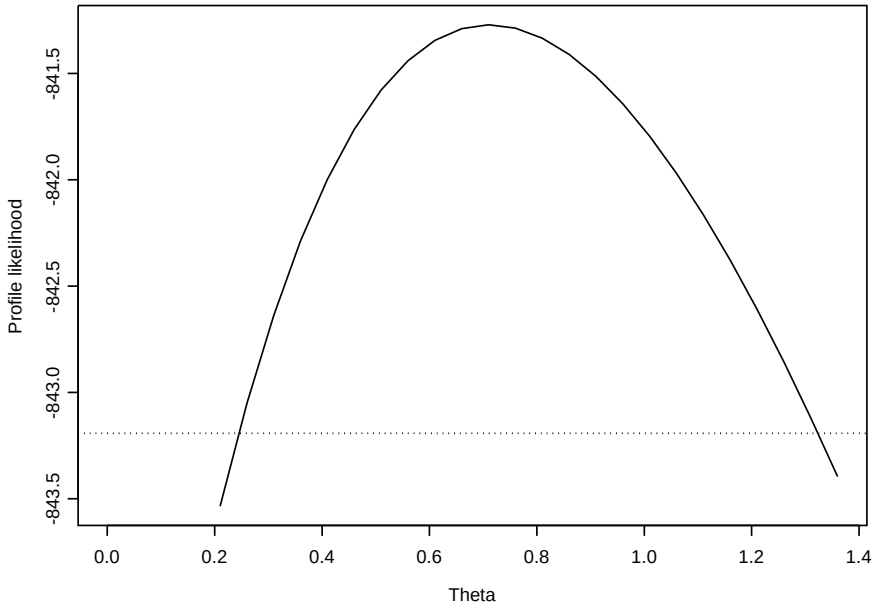



Figure 3.3: *Profile likelihood* $L_n(\theta, \lambda(\theta), \rho(\theta), \beta(\theta))$ *for the Diabetic Retinopathy data.*

## 3.8 Discussion

In Section 3.6 we discussed one-sided heterogeneity tests in the frailty model for censored data. For clarity we have restricted attention to the simple situation where the heterogeneity parameter is the single unknown parameter. The likelihood expressions given in Murphy (1995) and Murphy and van der Vaart (1997) will be useful to extend Theorem 4 to more complex parametric and semi-parametric frailty models. The related study of score tests for censored data is a further interesting topic. The discussion will be slightly more complicated than the one given in Section 3.5 since the censored data score vector does not have mean zero (see for example Theorem 2 in Murphy, 1995). One way to define the score statistic then would be to start with a centered score vector by subtracting its mean and then proceed in a similar manner as in Section 3.5.

Wald-type test statistics for testing hypothesis (3.1) may be employed as well. Robertson, Wright and Dykstra (1988) construct a Wald statistic for the situation where the alternative hypothesis is described by inequalities. Their test statistic requires estimation of model parameters under both the null and alternative hypothesis. Sen and Silvapulle

(2002) state a Wald test statistic as a difference of the minimum of two quadratic forms which has under the null hypothesis the same asymptotic distribution as the score and likelihood ratio statistic. For more details, see the recent review paper by Sen and Silvapulle (2002).

An interesting topic of further research is to study the distributional behaviour of the test for heterogeneity under local alternatives converging to the null hypothesis at rate $n^{-\frac{1}{2}}$. As in the two-sided testing problems, it is expected that the test statistics will have the same power characteristics under the local circumstances.

A further relevant issue for further study is to provide information on good finite sample approximation of the mixing properties, i.e., can we improve the asymptotic 50:50 mixture of the $\chi_0^2$ and $\chi_1^2$ by finding mixing proportions that depend on the information of the sample size? In a setting of regression spline mixed models, Claeskens (2002) calculates finite sample approximations to the mixing probabilities. In the frailty models currently under consideration the situation is more complex by the presence of nuisance parameters under the null hypothesis and censoring. Bootstrapping the distribution of the test statistic can provide another alternative to the asymptotic distribution.

## 3.9 Appendix

Define the following random variables:
$U_i = U_{i1} + U_{i2}$=$T_{i1}^{e^{-\alpha}} \exp(\beta x_{i1}) + T_{i2}^{e^{-\alpha}} \exp(\beta x_{i2})$
$V_i = \log T_{i1} + \log T_{i2}$
$W_i = W_{i1} + W_{i2} = T_{i1}^{e^{-\alpha}} \log T_{i1} \exp(\beta x_{i1}) + T_{i2}^{e^{-\alpha}} \log T_{i2} \exp(\beta x_{i2})$.
and $M_i = x_{i1}U_{i1} + x_{i2}U_{i2}$
Calculation of the score vector $S_n(\theta, \eta, \alpha, \beta)$:

$$\begin{aligned}
\frac{\partial}{\partial\theta}L_n(\tau) &= \frac{n}{(1+\theta)} + \frac{1}{\theta^2}\sum_{i=1}^{n}\log(1+\theta e^{-\eta}U_i) - (\frac{1}{\theta}+2)e^{-\eta}\sum_{i=1}^{n}\frac{U_i}{1+\theta e^{-\eta}U_i} \qquad (3.17)\\
\frac{\partial}{\partial\eta}L_n(\tau) &= -2n + (1+2\theta)e^{-\eta}\sum_{i=1}^{n}\frac{U_i}{(1+\theta e^{-\eta}U_i)}\\
\frac{\partial}{\partial\alpha}L_n(\tau) &= -2n - e^{-\alpha}\sum_{i=1}^{n}V_i + (1+2\theta)e^{-(\eta+\alpha)}\sum_{i=1}^{n}\frac{W_i}{1+\theta e^{-\eta}U_i}.\\
\frac{\partial}{\partial\beta}L_n(\tau) &= \sum_{i=1}^{n}(x_{i1}+x_{i2}) - (1+2\theta)e^{-\eta}\sum_{i=1}^{n}\frac{M_i}{1+\theta e^{-\eta}U_i}.
\end{aligned}$$

Under the null hypothesis, using an expansion of the logarithm in the second term of (3.17), it follows that:

$$\begin{aligned}
\frac{\partial}{\partial\theta}L_n(0,\nu) &= n - 2e^{-\eta}\sum_{i=1}^{n}U_i + \frac{1}{2}e^{-2\eta}\sum_{i=1}^{n}U_i^2\\
\frac{\partial}{\partial\eta}L_n(0,\nu) &= -2n + e^{-\eta}\sum_{i=1}^{n}U_i\\
\frac{\partial}{\partial\alpha}L_n(0,\nu) &= -2n - e^{-\alpha}\sum_{i=1}^{n}V_i + e^{-(\eta+\alpha)}\sum_{i=1}^{n}W_i.\\
\frac{\partial}{\partial\beta}L_n(0,\nu) &= \sum_{i=1}^{n}(x_{i1}+x_{i2}) - e^{-\eta}\sum_{i=1}^{n}M_i.
\end{aligned}$$

The components needed in the calculation of the Fisher information matrix $\boldsymbol{F}_n(\tau)$ are :

$$[1,1]: \quad \frac{\partial^2}{\partial\theta^2}L_n(\tau) = -\frac{n}{(1+\theta)^2} - \frac{2}{\theta^3}\sum_{i=1}^{n}\log(1+\theta e^{-\eta U_i}) + \frac{2e^{-\eta}}{\theta^2}\sum_{i=1}^{n}\frac{U_i}{1+\theta e^{-\eta U_i}} + (\frac{1}{\theta}+2)e^{-2\eta}\sum_{i=1}^{n}\frac{U_i^2}{(1+\theta e^{-\eta U_i})^2}$$

$$[2,2]: \quad \frac{\partial^2}{\partial\eta^2}L_n(\tau) = \theta e^{-2\eta}(1+2\theta)\sum_{i=1}^{n}\frac{U_i^2}{(1+\theta e^{-\eta U_i})^2} - e^{-\eta}(1+2\theta)\sum_{i=1}^{n}\frac{U_i}{1+\theta e^{-\eta U_i}}$$

$$[3,3]: \quad \frac{\partial^2}{\partial\alpha^2}L_n(\tau) = e^{-\alpha}\textstyle\sum_{i=1}^{n}V_i - (1+2\theta)e^{-(\eta+\alpha)}\displaystyle\sum_{i=1}^{n}\frac{W_i}{1+\theta e^{-\eta U_i}} - (1+2\theta)e^{-(\eta+2\alpha)}\sum_{i=1}^{n}\frac{W_{i1}\log T_{i1} + W_{i2}\log T_{i2}}{1+\theta e^{-\eta U_i}} + \theta(1+2\theta)e^{-2(\eta+\alpha)}\sum_{i=1}^{n}\frac{W_i^2}{(1+\theta e^{-\eta U_i})^2}$$

$$[4,4]: \quad \frac{\partial^2}{\partial\beta^2}L_n(\tau) = -(1+2\theta)e^{-\eta}\sum_{i=1}^{n}\frac{x_{i1}^2U_{i1} + x_{i2}^2U_{i2}}{1+\theta e^{-\eta U_i}} + \theta(1+2\theta)e^{-2\eta}\sum_{i=1}^{n}\frac{M_i^2}{(1+\theta e^{-\eta U_i})^2}$$

$$[1,2]: \quad \frac{\partial^2}{\partial\theta\partial\eta}L_n(\tau) = -e^{-2\eta}(1+2\theta)\sum_{i=1}^{n}\frac{U_i^2}{(1+\theta e^{-\eta U_i})^2} + 2e^{-\eta}\sum_{i=1}^{n}\frac{U_i}{1+\theta e^{-\eta U_i}}$$

$$[1,3]: \quad \frac{\partial^2}{\partial\theta\partial\alpha}L_n(\tau) = 2e^{-(\eta+\alpha)}\sum_{i=1}^{n}\frac{W_i}{1+\theta e^{-\eta U_i}} - (1+2\theta)e^{-(2\eta+\alpha)}\sum_{i=1}^{n}\frac{U_iW_i}{(1+\theta e^{-\eta U_i})^2}$$

$$[1,4]: \quad \frac{\partial^2}{\partial\theta\partial\beta}L_n(\tau) = -2e^{-\eta}\sum_{i=1}^{n}\frac{M_i}{1+\theta e^{-\eta U_i}} + (1+2\theta)e^{-2\eta}\sum_{i=1}^{n}\frac{U_iM_i}{(1+\theta e^{-\eta U_i})^2}$$

$$[2,3]: \quad \frac{\partial^2}{\partial\eta\partial\alpha}L_n(\tau) = -(1+2\theta)e^{-(\eta+\alpha)}\sum_{i=1}^{n}\frac{W_i}{(1+\theta e^{-\eta U_i})^2} + (1+2\theta)\theta e^{-(2\eta+\alpha)}\sum_{i=1}^{n}\frac{U_iW_i}{(1+\theta e^{-\eta U_i})^2}$$

$$[2,4]: \quad \frac{\partial^2}{\partial\eta\partial\beta}L_n(\tau) = (1+2\theta)e^{-\eta}\sum_{i=1}^{n}\frac{M_i}{1+\theta e^{-\eta U_i}} - (1+2\theta)\theta e^{-2\eta}\sum_{i=1}^{n}\frac{U_iM_i}{(1+\theta e^{-\eta U_i})^2}$$

$$[3,4]: \quad \frac{\partial^2}{\partial\alpha\partial\beta}L_n(\tau) = (1+2\theta)e^{-(\eta+\alpha)}\sum_{i=1}^{n}\frac{x_{i1}W_{i1} + x_{i2}W_{i2}}{1+\theta e^{-\eta U_i}} - (1+2\theta)\theta e^{-(2\eta+\alpha)}\sum_{i=1}^{n}\frac{W_iM_i}{(1+\theta e^{-\eta U_i})^2}.$$

Under the null hypothesis, and using an expansion of the logarithm in the second term of the second derivative with respect to $\theta$, it follows that:

$$[1,1]: \quad \frac{\partial^2}{\partial\theta^2}L_n(0,\nu) = -n + 2e^{-2\eta}\sum_{i=1}^{n}U_i^2 - \frac{2}{3}e^{-3\eta}\sum_{i=1}^{n}U_i^3$$

$$[2,2]: \quad \frac{\partial^2}{\partial\eta^2}L_n(0,\nu) = -e^{-\eta}\sum_{i=1}^{n}U_i$$

$$[3,3]: \quad \frac{\partial^2}{\partial\alpha^2}L_n(0,\nu) = e^{-\alpha}\sum_{i=1}^{n}V_i - e^{-(\eta+\alpha)}\sum_{i=1}^{n}W_i - e^{-(\eta+2\alpha)}\sum_{i=1}^{n}\{W_{i1}\log T_{i1} + W_{i2}\log T_{i2}\}$$

$$[4,4]: \quad \frac{\partial^2}{\partial\beta^2}L_n(0,\nu) = -e^{-\eta}\sum_{i=1}^{n}x_{i1}^2U_{i1} + x_{i2}^2U_{i2}$$

$$[1,2]: \quad \frac{\partial^2}{\partial\theta\partial\eta}L_n(0,\nu) = -e^{-2\eta}\sum_{i=1}^{n}U_i^2 + 2e^{-\eta}\sum_{i=1}^{n}U_i$$

$$[1,3]: \quad \frac{\partial^2}{\partial\theta\partial\alpha}L_n(0,\nu) = 2e^{-(\eta+\alpha)}\sum_{i=1}^{n}W_i - e^{-(2\eta+\alpha)}\sum_{i=1}^{n}U_iW_i$$

$$[1,4]: \quad \frac{\partial^2}{\partial\theta\partial\beta}L_n(0,\nu) = -2e^{-\eta}\sum_{i=1}^{n}M_i + e^{-2\eta}\sum_{i=1}^{n}U_iM_i$$

$$[2,3]: \quad \frac{\partial^2}{\partial\eta\partial\alpha}L_n(0,\nu) = -e^{-(\eta+\alpha)}\sum_{i=1}^{n}W_i$$

$$[2,4]: \quad \frac{\partial^2}{\partial\eta\partial\beta}L_n(0,\nu) = e^{-\eta}\sum_{i=1}^{n}x_{i1}U_{i1} + x_{i2}U_{i2}$$

$$[3,4]: \quad \frac{\partial^2}{\partial\alpha\partial\beta}L_n(0,\nu) = e^{-(\eta+\alpha)}\sum_{i=1}^{n}x_{i1}W_{i1} + x_{i2}W_{i2}.$$

When $\beta = 0$ (no covariates) the expected values are

$$\begin{aligned}
E[-\frac{\partial^2}{\partial\theta^2}L_n(\tau)] &= \frac{(5+9\theta+6\theta^2)}{(1+\theta)^2(1+2\theta)(1+3\theta)} \\
E[-\frac{\partial^2}{\partial\eta^2}L_n(\tau)] &= \frac{2}{(1+3\theta)} \\
E[-\frac{\partial^2}{\partial\alpha^2}L_n(\tau)] &= 2n + 2n\zeta(2,2) + \frac{2n}{(1+3\theta)}[(\psi(2)+h(\theta))^2 + \zeta(2,\frac{1}{\theta}) \\
&\quad -2\theta[\psi(3)^2 - \psi(2)^2 + 2(\psi(3)-\psi(2))h(\theta) + \zeta(2,3)]] \\
E[-\frac{\partial^2}{\partial\theta\partial\eta}L_n(\tau)] &= \frac{2}{(1+3\theta)(1+2\theta)} \\
E[-\frac{\partial^2}{\partial\theta\partial\alpha}L_n(\tau)] &= \frac{4n\psi(3)}{(1+3\theta)} + \frac{2nh(\theta)}{(1+2\theta)(1+3\theta)} - \frac{n(2+8\theta)\psi(2)}{(1+2\theta)(1+3\theta)} \\
E[-\frac{\partial^2}{\partial\eta\partial\alpha}L_n(\tau)] &= \frac{2n}{(1+3\theta)}[\psi(2)+\eta-\psi(\frac{1}{\theta}+1)-\log(\theta)]
\end{aligned}$$

To obtain the matrix $\boldsymbol{G}_n(\nu)$ we need these expected values under $H_0$ which yields the matrix $\boldsymbol{G}_n(\nu)$ given in Section 3.5. Note that $\det\{\boldsymbol{G}(\eta)\} = 2\pi^2 - 8$. Since the submatrices (5) and $\begin{pmatrix} 5 & 2 \\ 2 & 2 \end{pmatrix}$ have positive determinants it follows that $\boldsymbol{G}(\eta)$ is positive definite (see e.g. Artin (1991), p. 242). Moreover since $\boldsymbol{G}(\eta)$ is symmetric its eigenvalues are real. For $\kappa$ an eigenvalue of $\boldsymbol{G}(\eta)$ and $x_\kappa$ the corresponding eigenvector we have

$$x_\kappa^T G(\eta) x_\kappa = \kappa x_\kappa^T x_\kappa > 0$$

which implies that the eigenvalues of $\boldsymbol{G}(\eta)$ are strictly positive.

**Property 1.** *The symmetric matrix $\boldsymbol{G}(\eta)$ is positive definite and therefore has for every fixed value of $\eta$, three positive eigenvalues.*

Now

$$\begin{aligned}
E[\frac{\partial}{\partial\theta}L_n(0,\nu)\frac{\partial}{\partial\theta}L_n(0,\nu)] &= 5n \\
E[\frac{\partial}{\partial\eta}L_n(0,\nu)\frac{\partial}{\partial\eta}L_n(0,\nu)] &= 2n \\
E[\frac{\partial}{\partial\alpha}L_n(0,\nu)\frac{\partial}{\partial\alpha}L_n(0,\nu)] &= n(4(\psi(3)+\eta)^2 - 6(\psi(2)+\eta)^2 \\
&\quad +4(\psi(2)+\eta)(\psi(1)+\eta) + 2\zeta(2,1) + 4\zeta(2,3) - 4\zeta(2,2)) \\
E[\frac{\partial}{\partial\theta}L_n(0,\nu)\frac{\partial}{\partial\eta}\ L_n(0,\nu)] &= 2n \\
E[\frac{\partial}{\partial\theta}L_n(0,\nu)\frac{\partial}{\partial\alpha}\ L_n(0,\nu)] &= n(6\psi(4) - 6\psi(3) + 2\psi(2) + 2\eta) \\
E[\frac{\partial}{\partial\eta}L_n(0,\nu)\frac{\partial}{\partial\alpha}\ L_n(0,\nu)] &= n(4\psi(3) - 4\psi(2) + 2\psi(1) + 2\eta)
\end{aligned}$$

**Property 2.** $\boldsymbol{G}_n(\nu) = \boldsymbol{D}_n(\nu) = n\boldsymbol{G}(\eta)$.

**Proof.** Using the recursive property of the digamma function, $\psi(\nu+1) = \psi(\nu) + \frac{1}{\nu}$, we have $\psi(3) = \psi(2) + \frac{1}{2}$ and $\psi(1) = \psi(2) - 1$. From this we obtain that

$$4(\psi(3)+\eta)^2 - 6(\psi(2)+\eta)^2 + 4(\psi(2)+\eta)(\psi(1)+\eta) \;=\; 2(\psi(2)+\eta)^2 + 1 \quad (3.18)$$

A direct calculation also shows that

$$2\zeta(2,1) + 4\zeta(2,3) - 4\zeta(2,2) \;=\; 2\zeta(2,2) + 1 \quad (3.19)$$

From (3.18) and (3.19) we obtain $(\boldsymbol{G}_n(\eta))_{33} = (\boldsymbol{D}_n(\eta))_{33}$ where $\boldsymbol{D}_n(\eta) = \boldsymbol{D}_n(\eta,\alpha)$. Equality of the other entries can be showed in a similar (more easy) way. Direct calculation of the matrix elements yields the simplified expression for $\boldsymbol{G}_n$ in (3.5).

# Part II

# Advanced models for analysing animal breeding data

# Chapter 4

# Diani lamb data

In this chapter we give a detailed description of the dataset that is used in this second part of the thesis. The dataset comes from an animal breeding program that was carried out by the International Livestock Research Institute (ILRI) from 1991 to 1996. The objective of the experiment was to study genetic resistance to naturally acquired gastro-intestinal nematodes in different breeds of sheep, namely the Red Maasai, Dorper and their crossbreeds. In Section 4.1 we highlight some of the factors that motivated the need for the breeding program. The experimental setup of the study is briefly outlined in Section 4.2 while we report the measurements that were collected in Section 4.3. In total there were 1785 lambs from this breeding experiment. Of these, 696(38%) lambs died while 94 (5%) were lost or stolen before they were one year old. The main causes of death experienced over the six years are reported in Section 4.4. Finally in Section 4.5 we give the motivation for the developments of the new techniques to analyse these data. These new methodologies are discussed in Chapters 5 and 6.

## 4.1 Background

In the tropics small ruminants (sheep and goats) are an important source of income for many smallholder farmers. They are primarily kept for meat production. However, the productivity of sheep in the tropics is often low compared with animals in temperate regions. An important factor contributing to this low productivity is high mortality rate;

in tropical environments between 20% and 50% of the lambs born can die before weaning (Gatenby, 1986, Mukasa-Mugerwe et al., 2000, Wilson et al., 1993). Sheep in these regions primarily graze natural pastures or utilise crop residues. It is then not surprising that infections with gastrointestinal (GI) nematode parasites (endoparasites) are commonly one of the major causes of mortality (Over et al., 1992).

Current control methods for GI nematode parasites focus on reducing contamination of pastures through anthelmintic treatment and/or controlled grazing (Barger, 1999). In Africa, the use of these control methods is limited by the high cost of anthelmintics, their uncertain availability and increasing frequency of drug resistance (Waller, 1997). There is also limited scope in many communal pastoral systems for controlled grazing. It appears unlikely that new broad-spectrum anthelmintics will be available in the near future because of the major costs associated with the development of new products. To date, no commercial vaccines are available to control GI nematode parasites (Smith, 1999). The characterization and utilization of host genetic variation for resistance to endoparasites is thus an alternative approach to control endoparasites.

Variation among sheep breeds in resistance to GI nematode parasites has been extensively studied over the past half century (Gray et al., 1995). The reported findings of the above mentioned experiment (Baker et al., 1999, 2003) now show strong evidence that the Red Maasai (R) are both more resistant and resilient to naturally acquired and artificial infections with GI nematode parasites than other breeds notably the Dorper (D). Resilience (or tolerance) is defined as the ability of the host to survive and be productive in the face of parasite challenge while resistance is defined as the initiation and maintenance of responses provoked in the host to suppress the establishment of parasites and/or eliminate parasite burdens (Baker et al, 2003).

The Red Maasai is an East African fat-tailed sheep breed, which is associated with the Maasai tribe found in northern Tanzania and south-central Kenya (Wilson, 1991). The Dorper breed was developed in South Africa in the 1940s by interbreeding the Dorset and the Blackhead Persian breeds (Milne, 2000). The Dorper has a reputation for being well adapted to harsh, arid conditions (Cloete et al., 2000) and was first imported into Kenya from South Africa in the 1960s. It is also a popular breed for meat production.

## 4.2 Experimental design

In the first year of the study (1991) Dorper and RxD ewes were mated to 12 Dorper and 12 Red Maasai rams in single sire mating groups of about 18 animals in a partial diallel design. In a diallel design purebred and half-crossbred dams are mated with purebred sires from each of the breeds. In the subsequent years all three ewe genotypes (Dorper, Red Maasai, RxD) were mated to 12 Dorper and 12 Red Maasai rams to generate six lamb breeds or crosses (Table 4.1). A total of 264 Dorper, 312 RxD and 138 Red Maasai ewes were used in the six years. About 7-8 new rams of each breed were used at each mating so that 35 different Red Maasai rams and 41 different Dorper rams were used over the entire study. All Dorper and Red Maasai rams purchased were as unrelated as possible to ensure a representative genetic sample.

Table 4.1: *Numbers of lambs born by year of birth and genotype (D=Dorper, R=Red Maasai), numbers treated for GI nematode parasites at one and two months of age prior to weaning, and numbers weaned at about three months of age and the deaths after weaning.*

| Genotype | Year of Birth | | | | | | |
|---|---|---|---|---|---|---|---|
| (Sire breed x Dam Breed) | 1991 | 1992 | 1993 | 1994 | 1995 | 1996 | Total |
| D x D | 93 | 65 | 54 | 30 | 39 | 30 | 311 |
| D x (R x D) | 92 | 77 | 93 | 70 | 65 | 35 | 432 |
| D x R | 0 | 7 | 38 | 24 | 27 | 27 | 123 |
| R x D | 83 | 58 | 57 | 13 | 14 | 9 | 234 |
| R x (R x D) | 99 | 81 | 96 | 61 | 69 | 67 | 473 |
| R x R | 0 | 8 | 34 | 45 | 64 | 61 | 212 |
| Total | 367 | 296 | 372 | 243 | 278 | 229 | 1785 |
| Number treated pre weaning | 221 | 213 | 283 | 60 | 40 | 25 | 842 |
| Number weaned | 310 | 242 | 347 | 170 | 202 | 171 | 1442 |
| Deaths or stolen | | | | | | | |
| Pre-weaning | 58 | 54 | 25 | 73 | 75 | 58 | 343 |
| Post-weaning | 61 | 38 | 175 | 71 | 75 | 27 | 447 |

## 4.3 Data collected

Measurements of packed red cell volume (PCV), faecel egg count (FEC) and body weight (BWT) were taken from lambs up to about one year of age in batches of lambs born in each of the years 1991 to 1996. All lambs were weighed at birth and their BWT, PCV and FEC were subsequently recorded at one and two months of age. On either of these latter occasions, when individual lambs had a FEC greater than or equal to 2,000 eggs per gram (epg) and/or a PCV less than or equal to 20%, they were treated (drenched) with an anthelmintic drug. Packed cell volume and FEC were measured according to methods reported by Baker et al. (1999). PCV is the percentage of red blood cells to the total volume of a blood sample and in general gives a good indication of how anemic an animal is. Thus it is a good indicator of how well the animal is managing to cope with the pathogenic effects of the blood-sucking parasite *Haemonchus contortus*[1] which was the main parasite species that was found in this study. Faecel egg counts on the other hand are known to be highly correlated with worm counts (Woolaston and Baker, 1996). Packed cell volume is often used to measure resilience while FEC is used to measure resistance (Baker et al., 2003).

At about three months of age, the time of weaning, the lambs were again weighed, and blood and faecal samples collected for PCV and FEC, respectively. All lambs were then drenched. The lambs were then left to graze on pasture, separately from the ewes and rams. Every week a monitor group of about 50 lambs, made up of approximately equal numbers of lambs of each genotype and gender, was sampled and their mean FEC recorded. If the mean FEC was over 2,000 epg then, during two consecutive days, all lambs were weighed, faeces and blood samples taken for FEC and PCV respectively and the lambs were then drenched. This procedure was followed until the lambs reached on average one year of age resulting in five drenchings in each year except 1994 and 1996. In 1994 the lambs were drenched eight times post-weaning, while in 1996 six drenchings occurred. A sample of a few data lines is given in the appendix of this chapter while Figures 4.1 to

[1] Adult worms live in the stomach of ruminant animals, females depositing upto 10,000 eggs per day which pass out of the host in the faeces. After a day or two, first stage larvae (L1) hatch. The larvae feed on microorganisms in faeces and developing into the L2 larvae and then L3 larvae stages. L3 larvae are ingested with grass while grazing. Within the stomach the larvae molts two or three more times. In favorable conditions, adult worms develop and feed on the host's blood.

4.3 show scatter plots of the measurements recorded from weaning to 12 months for PCV and FEC and those from birth to 12 months for BWT, across the six years. In each plot individual profiles for a randomly selected sample of 15 lambs are highlighted. In these plots we see that although the lambs were all measured on the same day, the individual measurements are clustered around a particular age. This is due to the fact that in each year, lambs were born within a period ranging from 20-40 days. For example in Figure 4.1 the measurements are clustered around the age of 90 days in 1991 but around the age of 100 days in 1992.

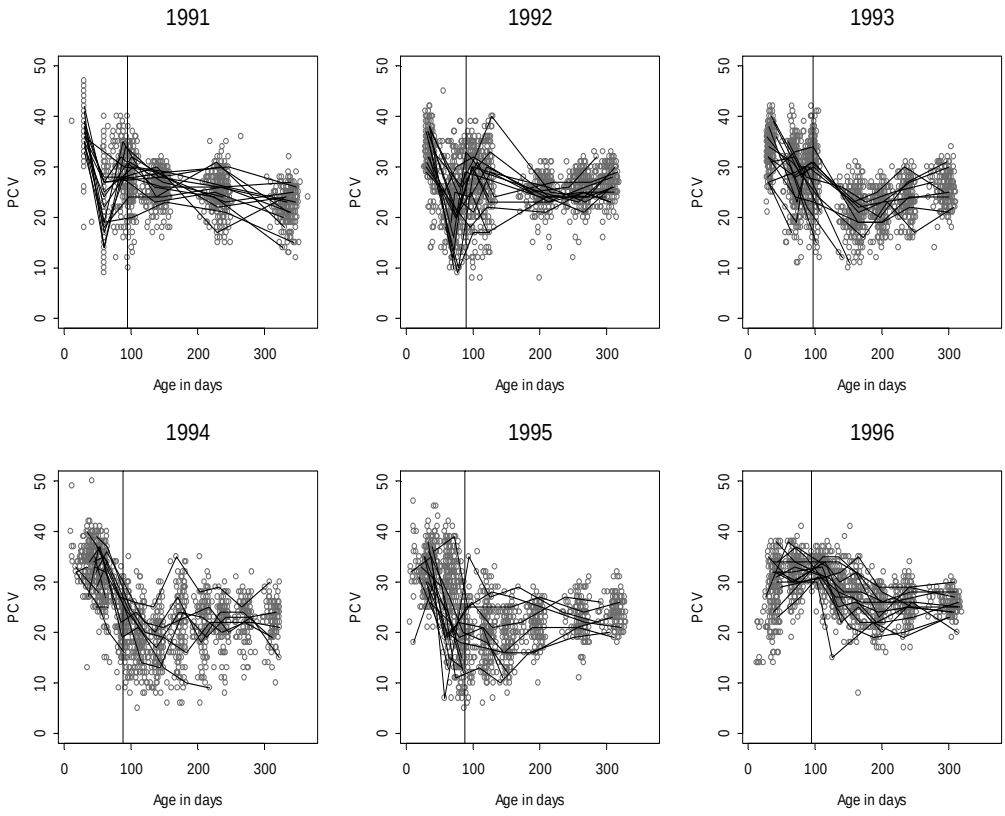



Figure 4.1: *PCV measurements for the years 1991 to 1996. Bold vertical line indicates when the lambs were weaned.*

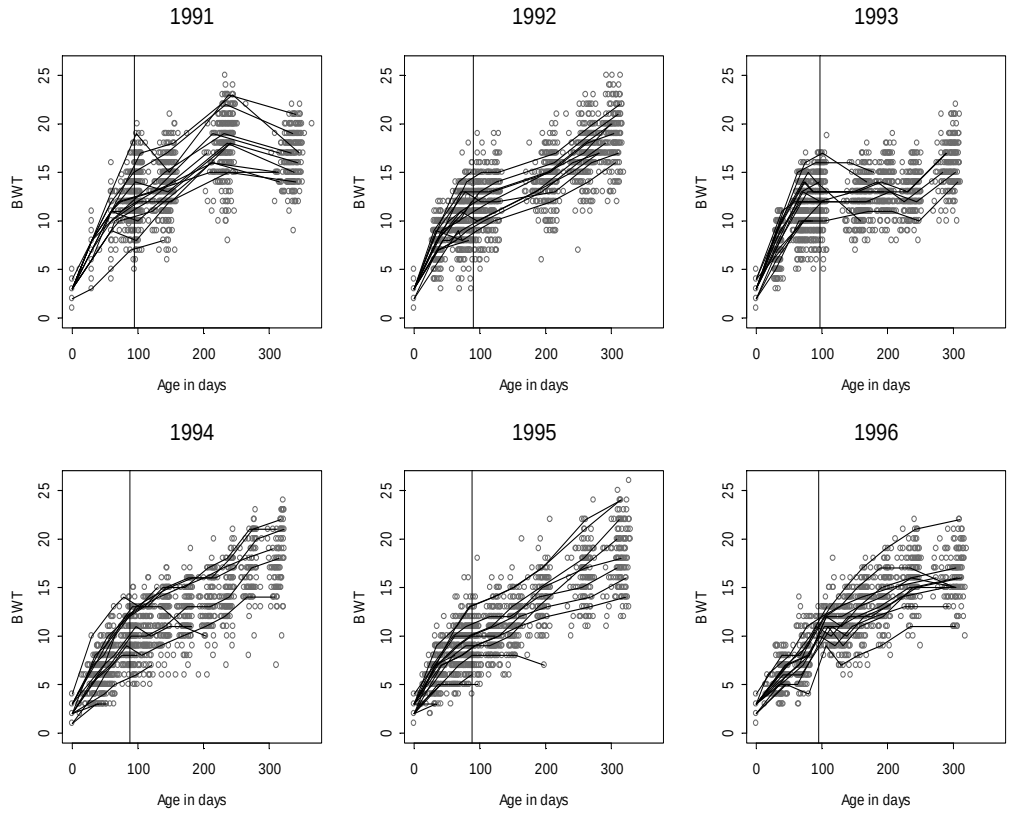



Figure 4.2: *Body weight measurements for the years 1991 to 1996. Bold vertical line indicates when the lambs were weaned.*

## 4.4 Causes of mortality

As noted in Section 4.1, mortality is the main contributing factor for low productivity of sheep in the tropics. Various causes of mortality were observed in the study. These were grouped into nine categories:

1) still births

2) mis-mothering, e.g. death from suffocation, starvation and weakness at birth

3) endoparasites

4) pneumonia

5) digestive disorders such as bloat and enteritis

6) accidents which included those killed by predators or from plant poisoning

7) lost or stolen

8) miscellaneous diseases such as coccidiosis, dystocia or foot rot

9) unknown causes.

Overall, the deaths resulting from accidents were included in the lost or stolen category since they were few, whilst deaths associated with digestive disorders were included in the miscellaneous category. Further, only two lambs in the pre-weaning period died from an unknown cause. These were also included in the miscellaneous category within this period. This resulted in six and five categories defining the cause of death in the pre-weaning and post-weaning periods, respectively. The break down of the these causes in the pre-weaning and post-weaning periods by genotype is given in Tables 4.2 and 4.3, respectively.

Overall, there were 1785 lamb from this experiment. Of these, 696 (39%) died while 94 (5%) were lost or stolen before they were one year old. Among the deaths, 343 (44%) occurred before weaning, of which about a third were associated with mis-mothering (Table 4.2). Pre-weaning endoparasite infections accounted for about a fifth of the deaths. The major cause of mortality in the post-weaning period was associated with endoparasite

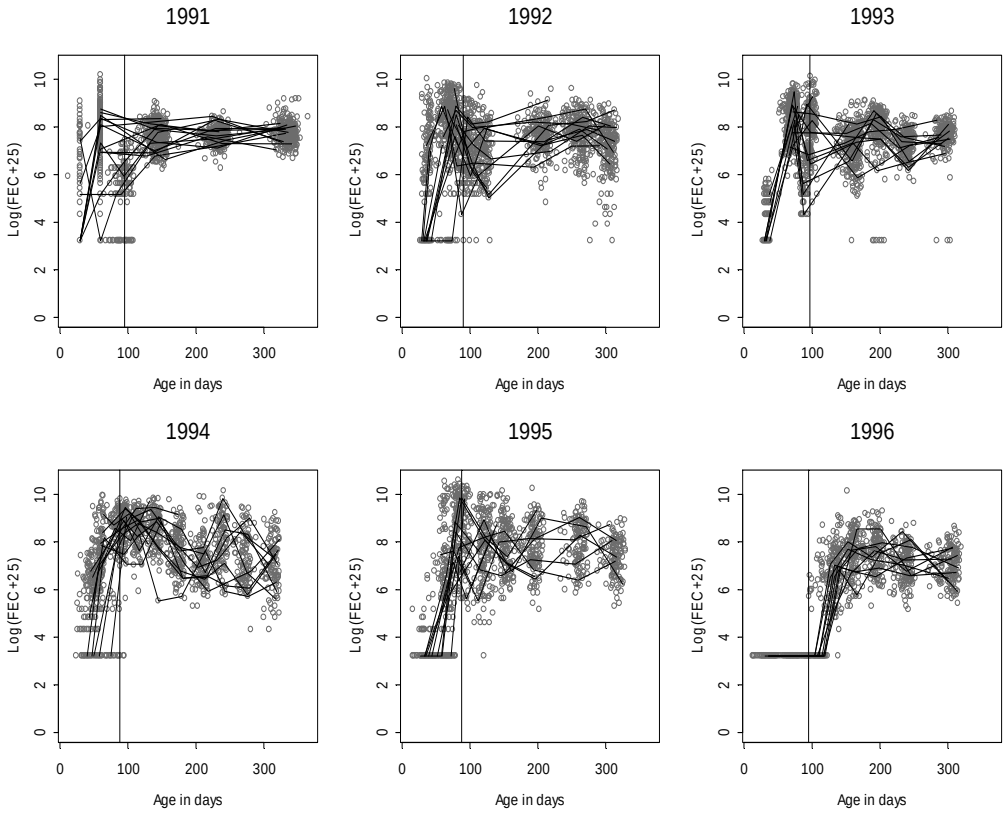



Figure 4.3: *Transformed* log*(FEC+25) measurements for the years 1991 to 1996. Bold vertical line indicates when the lambs were weaned.*

Table 4.2: *Numbers of deaths (proportions) in the pre-weaning (from birth to approximately 90 days) period by cause of death for lambs of different genotypes (D=Dorper, R=Red Maasai) and corresponding overall average mortality rates.*

| | D x D | D x (R x D) | D x R | R x D | R x (R x D) | R x R | Total |
|---|---|---|---|---|---|---|---|
| Number born | 311 | 432 | 123 | 234 | 473 | 212 | 1785 |
| Cause of death | | | | | | | |
| Still births | 7(0.08) | 13(0.14) | 2(0.09) | 3(0.09) | 7(0.09) | 8(0.25) | 40(0.12) |
| Mis-mothering | 29(0.33) | 29(0.31) | 7(0.32) | 10(0.29) | 22(0.29) | 10(0.31) | 107(0.31) |
| Endoparasites | 21(0.24) | 22(0.24) | 6(0.27) | 7(0.21) | 12(0.16) | 3(0.09) | 71(0.21) |
| Pneumonia | 6(0.07) | 10(0.11) | 1(0.04) | 5(0.15) | 14(0.18) | 1(0.03) | 37(0.11) |
| Lost/Stolen | 6(0.07) | 10(0.11) | 0(0.00) | 6(0.18) | 8(0.11) | 2(0.06) | 32(0.09) |
| Miscellaneous | 19(0.22) | 8(0.09) | 6(0.26) | 3(0.09) | 12(0.16) | 8(0.25) | 56(0.16) |
| Total | 88(0.28) | 92(0.21) | 22(0.18) | 34(0.15) | 75(0.16) | 32(0.15) | 343(0.19) |

infections, accounting for 212 (47%) of the deaths (Table 4.3), while 62 (14%) were lost or stolen. Pre-weaning, average mortality was between 15 and 28% with the Dorper lambs having the highest mortality. The average mortality post weaning was between 17 and 47% which showed a decreasing trend with increasing proportion of Red Maasai in the genotype.

Table 4.3: *Numbers of deaths (proportions) post-weaning (from 90 to 365 days) period by cause of death for lambs of different genotypes (D=Dorper, R=Red Maasai) and corresponding overall average mortality rates.*

| | D x D | D x (R x D) | D x R | R x D | R x (R x D) | R x R | Total |
|---|---|---|---|---|---|---|---|
| Number weaned | 224 | 340 | 100 | 200 | 398 | 180 | 1442 |
| Cause of death | | | | | | | |
| Endoparasites | 49(0.46) | 71(0.52) | 21(0.48) | 21(0.45) | 28(0.42) | 22(0.48) | 212(0.47) |
| Pneumonia | 16(0.15) | 19(0.14) | 4(0.09) | 6(0.13) | 13(0.20) | 3(0.07) | 61(0.14) |
| Lost/Stolen | 14(0.13) | 20(0.15) | 2(0.05) | 8(0.17) | 13(0.20) | 5(0.11) | 62(0.14) |
| Miscellaneous | 19(0.18) | 23(0.17) | 6(0.14) | 8(0.17) | 11(0.17) | 9(0.20) | 76(0.18) |
| Cause unknown | 8(0.08) | 5(0.04) | 11(0.25) | 4(0.08) | 1(0.01) | 7(0.15) | 36(0.07) |
| Total | 106(0.47) | 138(0.41) | 44(0.44) | 47(0.24) | 66(0.17) | 46(0.26) | 447(0.31) |

## 4.5 Motivation

To assess the genetic resistance of the sheep, which was the objective of the experiment, BWT, PCV and FEC measurements collected at the individual time points (e.g. at weaning and 8 months) have been analysed using the classical linear mixed model (Baker et al., 1994, 1998, 2003, Baker, 1998). The estimated variance components from these analyses have been used to determine heritability estimates.

Heritability helps to explain the degree to which genes control expression of a trait such as BWT, PCV or FEC. In our case, it would be a measure of the degree (0 to 100%) to which the lambs resemble their father(or mother) for the specific trait of interest. For instance, let $Y_{ij}$ denote the observed measurement of the trait of interest for the $j^{\text{th}}$ lamb from the $i^{\text{th}}$ sire, $i = 1, \ldots, G$ and $j = 1, \ldots, n_i$ at some time point (say 3 months). Then the simple linear mixed model used in this analysis is

$$Y_{ij} = \boldsymbol{x}_{ij}^T \boldsymbol{\beta} + s_i + \epsilon_{ij} \tag{4.1}$$

where $\boldsymbol{x}_{ij}$ is the incidence vector for the fixed effects, $\boldsymbol{\beta}$ is the vector of associated parameters, $s_i$ is the random effect of the $i^{\text{th}}$ sire and $\epsilon_{ij}$ is the random error term. The assumptions on the random terms are that the $s_i$ are identically and independent distributed (i.i.d) $N(0, \sigma_s^2)$ terms and the $\epsilon_{ij}$'s are i.i.d $N(0, \sigma_e^2)$. Further $s_i$ and $\epsilon_{ij}$ are assumed to be independent. Model (4.2) is referred to as a sire model. The sire genetic contribution is then estimated using the heritability measure defined as

$$h_s^2 = \frac{4\sigma_s^2}{\sigma_s^2 + \sigma_e^2}. \tag{4.2}$$

where the numerator is an estimate of the genetic variance while the denominator estimates phenotypic variance. The phenotype of an animal corresponds to the observed measurements e.g., PCV, and is the combined effect of all genetic and environmental influences.

Baker et al. (2003) reports the heritability estimates for the traits BWT, PCV and log(FEC +25) (LFEC) at each measurement time (i.e., birth, 1 month, 2 months, weaning and all post-weaning time points). These analyses were carried out with the linear mixed model (4.2) using the restricted maximum likelihood (REML) estimation method. The ASREML programme of Gilmour et al. (1999) was used to estimate simultaneously both fixed effects and variance components. Their results show that the Red Massai have

higher resistance (lower FEC) and higher resilience (higher PCV) than Dorpers. Baker et al. (1994, 1998) report preliminary results of this study using similar methods of analysis as described above. The performance of the ewes that were dams of the lambs whose data are used in the current study was evaluated by Baker et al. (1999), also using linear mixed models. In that analysis, it was clearly shown that the Red Maasai ewes were more resistant and resilient to GI nematode parasites than the Dorper ewes in a sub-humid tropical environment.

In carrying out the linear mixed model analysis as that described above, only the animals that survived to the time points of interest were utilised. Methodologies such as survival analysis that use the information available for all animals up to the time they die or are lost to follow up can alternatively be used in this assessment. In Chapter 5, survival analysis using shared frailty models is used to assess variations in time to death of the genotypes. In this analysis BWT, PCV and LFEC are considered as time-varying covariates. Kalbfleisch and Prentice(1980) distinguish between two types of time-dependent covariates; *external* and *internal* covariates. External covariates are those whose values do not depend on the survival status of the individual, for example, the monthly rainfall amount. Internal covariates on the other hand are only measured as long as the individual is under observation, like BWT, PCV and FEC. Unlike the external covariates, the internal covariates carry information about the survival pattern of the individuals. For example, high PCV and low FEC values may be associated with higher chances of survival as these are indicators of the health status of the animal.

In the recent past, methodologies, which simultaneously use the information available in survival and such time-varying covariates, have been proposed in medical research. In Chapter 6 we describe and adapt this joint modelling methodology to the animal breeding data, where the time to death of the lamb is modelled jointly with either PCV, BWT or FEC.

| VARIABLE | Description |
|---|---|
| NUMB | Animal number |
| DAM_ID | Dam identification number |
| DAM_BRD | Dam breed; 1-DxD, 2-RxD, 5-RxR |
| SIRE_ID | Sire identification number |
| SIRE_BRD | Sire breed |
| YEARB | Year of birth |
| BREED | Breed of lamb; 1-DxD, 2-RxD, 3-Dx(DxR), 4-Rx(RxD), 5-RxR, 6-DxR |
| SEX | Gender: 1:-Female, 2:-Male |
| BIRTHWT | Weight at birth |
| BIRTH_DT | Date of birth |
| DISP_DT | Date of death |
| DACTION | Action at disposal |
| DREASON | Death reason |
| DATE30 | Date of one month |
| AGE30 | Age at one month |
| WT30 | Weight at one month |
| PCV30 | PCV at one month |
| FEC30 | FEC at one month :-99999 indicates that FEC was not recorded |
| DATE60 | Date of two months |
| AGE60 | Age at two months |
| WT60 | Weight at two months |
| PCV60 | PCV at two months |
| FEC60 | FEC at two months |
| WEAN_DT | Date of weaning |
| AGEWEAN | Age at weaning |
| BIRTH_TY | Type of birth: 1:-single birth, 2:-twin |
| BIRTHDAY | Day of birthday in the calendar year from 1st January |
| DAMAGE | Age of the dam |
| WWT1 | Day 1 BWT measurement at weaning |
| WWT2 | Day 2 BWT measurement at weaning |
| WEANWT | Average of two weaning BWT measurements |
| WPCV1 | Day 1 PCV measurement at weaning |
| WPCV2 | Day 2 PCV measurement at weaning |
| WEANPCV | Average of two weaning PCV measurements |
| WFEC1 | Day 1 FEC measurement at weaning |
| WFEC2 | Day 2 FEC measurement at weaning |
| WEANFEC | Average of two weaning FEC measurements |
| DATE1 | Date of 1st post-weaning measurements |
| AGE1 | Age at the 1st post-weaning measurements |
| WT1A | Day 1 measurement of BWT at 1st post-weaning time point |
| WT1B | Day 2 measurement of BWT at 1st post-weaning time point |
| AVWT1 | Average of the two BWT measurements at 1st post-weaning time-point |
| PCV1A | Day 1 measurement of PCV at 1st post-weaning time point |
| PCV1B | Day 2 measurement of PCV at 1st post-weaning time point |
| AVPCV1 | Average of the two PCV measurements at 1st post-weaning time-point |
| FEC1A | Day 1 measurement of FEC at 1st post-weaning time point |
| FEC1B | Day 2 measurement of FEC at 1st post-weaning time point |
| AVFEC1 | Average of the two FEC measurements at Day 1 post-weaning time-point |
| DATE2 | Date of 2nd post-weaning measurements |
| AGE2 | Age at the 2nd post-weaning measurements |
| WT2A | Day 1 measurement of BWT at 2nd post-weaning time point |
| WT2B | Day 2 measurement of BWT at 2nd post-weaning time point |
| AVWT2 | Average of the two BWT measurements at 2nd post-weaning time-point |
| . | |
| . | |

| Obs | NUMB | DAM_ID | DAM_BRD | SIRE_ID | SIRE_BRD | YEARB | BREED | SEX |
|---|---|---|---|---|---|---|---|---|
| 1 | 3225 | 5189 | 2 | 1974 | 1 | 91 | 3 | 1 |
| 2 | 3226 | 1682 | 1 | 1980 | 1 | 91 | 1 | 2 |
| 3 | 3227 | 5162 | 2 | 1972 | 1 | 91 | 3 | 2 |

| Obs | BIRTHWT | BIRTH_DT | DISP_DT | DACTION | DREASON | DATE30 | AGE30 | WT30 |
|---|---|---|---|---|---|---|---|---|
| 1 | 2 | 06/13/91 | 11/05/92 | 5 | 12 | 07/13/91 | 30 | 6 |
| 2 | 3 | 05/28/91 | 10/26/92 | 2 | 25 | 06/27/91 | 30 | 8 |
| 3 | . | 05/31/91 | 05/31/91 | 6 | 31 | . | . | . |

| Obs | PCV30 | FEC30 | DATE60 | AGE60 | WT60 | PCV60 | FEC60 | WEAN_DT |
|---|---|---|---|---|---|---|---|---|
| 1 | 39 | 0 | 08/12/91 | 60 | 8 | 14 | 7900 | 09/30/91 |
| 2 | 45 | 0 | 07/27/91 | 60 | 11 | 32 | 8150 | 09/30/91 |
| 3 | . | 99999 | | . | . | . | 99999 | . |

| Obs | AGEWEAN | BIRTH_TY | BIRTHDAY | DAMAGE | WWT1 | WWT2 | WEANWT | WPCV1 |
|---|---|---|---|---|---|---|---|---|
| 1 | 109 | 1 | 164 | 3 | 10 | . | 10 | 32 |
| 2 | 125 | 1 | 148 | 2 | 16 | . | 16 | 28 |
| 3 | . | 1 | 151 | 3 | . | . | . | . |

| Obs | WPCV2 | WEANPCV | WFEC1 | WFEC2 | WEANFEC | DATE1 | AGE1 | WT1A |
|---|---|---|---|---|---|---|---|---|
| 1 | . | 32 | 350 | 99999 | 350 | 11/19/91 | 159 | 12 |
| 2 | . | 28 | 99999 | 99999 | 99999 | 11/19/91 | 175 | 19 |
| 3 | . | . | 99999 | 99999 | 99999 | | | . |

| Obs | WT1B | AVWT1 | PCV1A | PCV1B | AVPCV1 | FEC1A | FEC1B | AVFEC1 |
|---|---|---|---|---|---|---|---|---|
| 1 | 12 | 12 | 26 | 27 | 27 | 2300 | 1500 | 1900 |
| 2 | 18 | 19 | 30 | 25 | 28 | 1800 | 700 | 1250 |
| 3 | . | . | . | . | . | 99999 | 99999 | 99999 |

| Obs | DATE2 | AGE2 | WT2A | WT2B | AVWT2 | PCV2A | PCV2B | AVPCV2 |
|---|---|---|---|---|---|---|---|---|
| 1 | 02/17/92 | 249 | 15 | 15 | 15 | 20 | 26 | 23 |
| 2 | 02/17/92 | 265 | 22 | 21 | 22 | 34 | 37 | 36 |
| 3 | . | . | . | . | . | . | . | . |

| Obs | FEC2A | FEC2B | AVFEC2 | DATE3 | AGE3 | WT3A | WT3B | AVWT3 |
|---|---|---|---|---|---|---|---|---|
| 1 | 2600 | 1700 | 2150 | 05/29/92 | 351 | 14 | 14 | 14 |
| 2 | 2950 | 1600 | 2275 | 05/29/92 | 367 | 20 | 20 | 20 |
| 3 | 99999 | 99999 | 99999 | . | . | . | . | . |

| Obs | PCV3A | PCV3B | AVPCV3 | FEC3A | FEC3B | AVFEC3 | DATE4 | AGE4 |
|---|---|---|---|---|---|---|---|---|
| 1 | 22 | 24 | 23 | 3000 | 3600 | 3300 | 07/21/92 | 404 |
| 2 | 19 | 29 | 24 | 4700 | 4300 | 4500 | 07/21/92 | 420 |
| 3 | . | . | . | 99999 | 99999 | 99999 | . | . |

| Obs | WT4A | WT4B | AVWT4 | PCV4A | PCV4B | AVPCV4 | FEC4A | FEC4B |
|---|---|---|---|---|---|---|---|---|
| 1 | 17 | 17 | 17 | 28 | 28 | 28 | 950 | 1600 |
| 2 | 23 | 22 | 22 | 31 | 29 | 30 | 750 | 1150 |
| 3 | . | . | . | . | . | . | 99999 | 99999 |

| Obs | AVFEC4 | DATE5 | AGE5 | WT5A | WT5B | AVWT5 | PCV5A | PCV5B |
|---|---|---|---|---|---|---|---|---|
| 1 | 1275 | 09/21/92 | 466 | 18 | 18 | 18 | 22 | 23 |
| 2 | 950 | 09/21/92 | 482 | 25 | 25 | 25 | 33 | 32 |
| 3 | 99999 | . | . | . | . | . | . | . |

| Obs | AVPCV5 | FEC5A | FEC5B | AVFEC5 | DATE6 | AGE6 | WT6A | WT6B |
|---|---|---|---|---|---|---|---|---|
| 1 | 22 | 99999 | 450 | 450 | | . | . | . |
| 2 | 33 | 350 | 400 | 375 | | . | . | . |
| 3 | . | 99999 | 99999 | 99999 | | . | . | . |

# Chapter 5

# Application of shared frailty models

## 5.1 Introduction

In the previous chapter, we mentioned that alternative methods exist that can be used to model more adequately the information available for all animals up to the time they die or are lost to follow up. One such approach is using survival analysis. The main objectives under this approach are (1) to investigate the variation in lamb mortality among breeds and their crosses (genotypes); and (2) to investigate genetic variation for lamb mortality within genotypes.

Frailty models have been used for other species of livestock, for example in assessing the length of productive life in dairy cattle (Ducrocq et al., 1988), to assess viability of laying hens (Ducrocq, 2000), to obtain estimates of longevity of Swedish horses (Wallin et al., 2000) and sows (Yazdi et al., 2000) and to assessing genetic variation for disease resistance in growing pigs (Henryon et al., 2001). Both Cox and Weibull hazard models have been used in these studies although the parametric model has been used more extensively as it is less computer intensive than the Cox model.

Prior to carrying out any survival analysis using hazard models, we first investigated the shape of the hazard function. To this end non-parametric kernel density estimation methods were utilised. We used the estimator derived in Müller and Wang (1994) which

is described briefly in Section 5.2. From this assessment no parametric distribution was found to be appropriate for the hazard model and further analyses were carried out using the Cox PH and its extension, the shared frailty model. In Section 5.3 we discuss the analyses that we carried out and report the results in Section 5.4.
In most previous studies that use frailty models in animal research, heritability estimates for survival have been reported (Ducrocq et al., 1988, Ducrocq, 2000). In Section 5.5 we discuss briefly heritability estimation in survival analysis. The concluding remarks are given in Section 5.6.

## 5.2 Hazard function estimation

Consider the general hazard model of Section 2.1.3 given as

$$\lambda(t|\boldsymbol{x}_i) = \lambda_0(t)\phi(\boldsymbol{x}_i)$$

where the baseline hazard function $\lambda_0(t)$ may be left unspecified or it may be assumed to have some specific parametric form, thus determining the shape of the hazard function $\lambda(\cdot)$. We used non-parametric kernel based methods to determine this shape.
Non-parametric kernel estimation methods of the hazard function for right-censored data have received considerable attention in the statistical literature (Watson and Leadbetter, 1964, Ramlau-Hansen, 1983, Cheng, 1987, Müller and Wang, 1994). The main assumption made in these estimators about the unknown survival distributions is that the hazard functions vary smoothly over time.
In general a kernel estimator of a function $f$ at a given point $t$ is essentially a locally weighted average of the data from the interval $[t-b, t+b]$, where $b$ is the bandwidth or window size. A critical factor in the performance of the kernel estimator is the choice of the bandwidth which determines the degree of smoothness. The larger the bandwidth, the greater the smoothness. More smoothness leads to lower variability but also generally leads to increased bias. Watson and Leadbetter (1964) introduced the kernel estimator for the hazard function for uncensored data and Ramlau-Hansen (1983) extended the kernel hazard function to right censored data. The most widely used estimator for the hazard function from right-censored data has been the fixed-bandwidth kernel-smoothed estimator.

Let $T_i^o$, $i = 1, \ldots, n$, be the observed time-to-event and let $\delta_i$ be the censoring indicator as defined in Section 2.1.2. Let $(T_{(i)}^o, \delta_{(i)})$ be the ordered observations where the ordering is according to $T_i^o$. The fixed bandwidth estimator is then defined as

$$\hat{\lambda}(t) = \frac{1}{b} \sum_{i=1}^{n} K\left(\frac{t - T_{(i)}^o}{b}\right) \frac{\delta_{(i)}}{n - i + 1} \tag{5.1}$$

where $K(\cdot)$ is a kernel function and $b$ is the global bandwidth. Kernel functions are generally chosen to be symmetric probability density functions such as the normal density. The so-called Epanechnikov kernel $\left(K(x) = 0.75(1 - x^2) \text{for} -1 \leq x \leq 1\right)$ is a popular choice. The fixed-bandwidth kernel estimator however cannot adapt to unevenness in the distribution of the data. It tends to over-smooth in regions with many observations and under-smooth in regions with few observations. In the recent past, more flexible bandwidths such as the nearest neighbour (Tanner and Wong, 1984) and varying (local) bandwidths (Müller and Wang, 1990) have been suggested as alternatives to (5.1) in order to overcome this drawback associated with the fixed bandwidth. In addition bias problems have been found for fixed kernel estimators when estimating near the endpoints of the data. These problems arise when the support of the kernel exceeds the available data range. Owing to these boundary effects, varying kernel estimators that are more robust at the endpoints have been proposed (Hougaard et al., 1989, Müller and Wang, 1994). These type of kernels are often said to be boundary corrected or varying kernels.

We used the varying kernel and varying bandwidth estimator of Müller and Wang (1994) given as

$$\hat{\lambda}(t) = \frac{1}{b(t)} \sum_{i=1}^{n} K_t\left(\frac{t - T_{(i)}^o}{b(t)}\right) \frac{\delta_{(i)}}{n - i + 1}. \tag{5.2}$$

For this estimator both the bandwidth and the kernel function depend on the time point. Müller and Wang (1994) proposed the following polynomial boundary kernel that generalizes the Epanechnikov kernel

$$K_t(z) = \begin{cases} K_+(\frac{t}{b(t)}, z) & \text{if} \quad \{t : 0 \leq t < b(t)\} \\ \frac{3}{4}(1 - z^2) & \text{if} \quad \{t : b(t) \leq t \leq R - b(t)\} \\ K_-(\frac{R-t}{b(t)}, z) & \text{if} \quad \{t : R - b(t) \leq t \leq R\} \end{cases}$$

where $R$ is the right endpoint of the data, $q = t/b(t)$ and on $[-1, q]$

$$K_+(q, z) = \frac{12(1 + z)}{(1 + q)^4} \left[(1 - 2q)z + \frac{(3q^2 - 2q + 1)}{2}\right],$$

while on $[-q, 1]$, $K_-(q, z) = K_+(q, -z)$. This correction allows the shape of the kernel to change on the boundary. This estimator has been implemented as an executable function in S-Plus (Mathsoft, 1999) and can be downloaded from http://www.stats.ox.ac.uk/pub/SWin/.
We estimated the hazard function separately for each of the six years (Figure 5.1). Within each year we also obtained the hazard function estimate for the Red Maasai and Dorper genotypes. As there were no Red Maasai lambs in 1991, the estimated hazard function for the Red Maasai was obtained only for the years 1992-1996.
This plot shows that the risk of mortality at any point in time is much lower for the Red Maasai than that of the Dorper. Although not shown, the risks of death for the other groups lay in between those for the pure breeds. The figure demonstrates the variable pattern in the hazard estimate across years with risk of mortality generally lower in 1991 and 1992 than in the other years. Due to this variability, further analysis was carried out using the Cox proportional hazards model where the baseline hazard is not restricted to be of a particular shape. In trying to understand the variable patterns of the estimated hazard function we also looked at the rainfall patterns across the six years (Figure 5.1). Peak rainfall was higher in 1993-1996 compared with 1991 and 1992. Except for 1996, when lambs were born earlier than in other years, peak rainfall patterns in the years appeared to be followed by a rise in risk of mortality as estimated by the hazard function. The month of birth of the lambs varied across the years as matings took place at intervals of 10 to 12 months. Further, 1994 had the highest average rainfall post-weaning. This could explain the higher number of post-weaning measurements collected in this year, as mentioned in Section 4.3.

## 5.3 Statistical analysis

Statistical analysis was done separately in the pre-weaning and post-weaning periods as the critical period for assessing genetic resistance to endoparasites in lambs is between weaning and 12 months of age, during which the immune system of the young animal is developing.
We saw in Section 4.4 that there were various causes of mortality. As an initial step, we investigated in both the pre-weaning and post-weaning periods whether the ranking of causes of death varied across breeds and also across the years. This assessment is

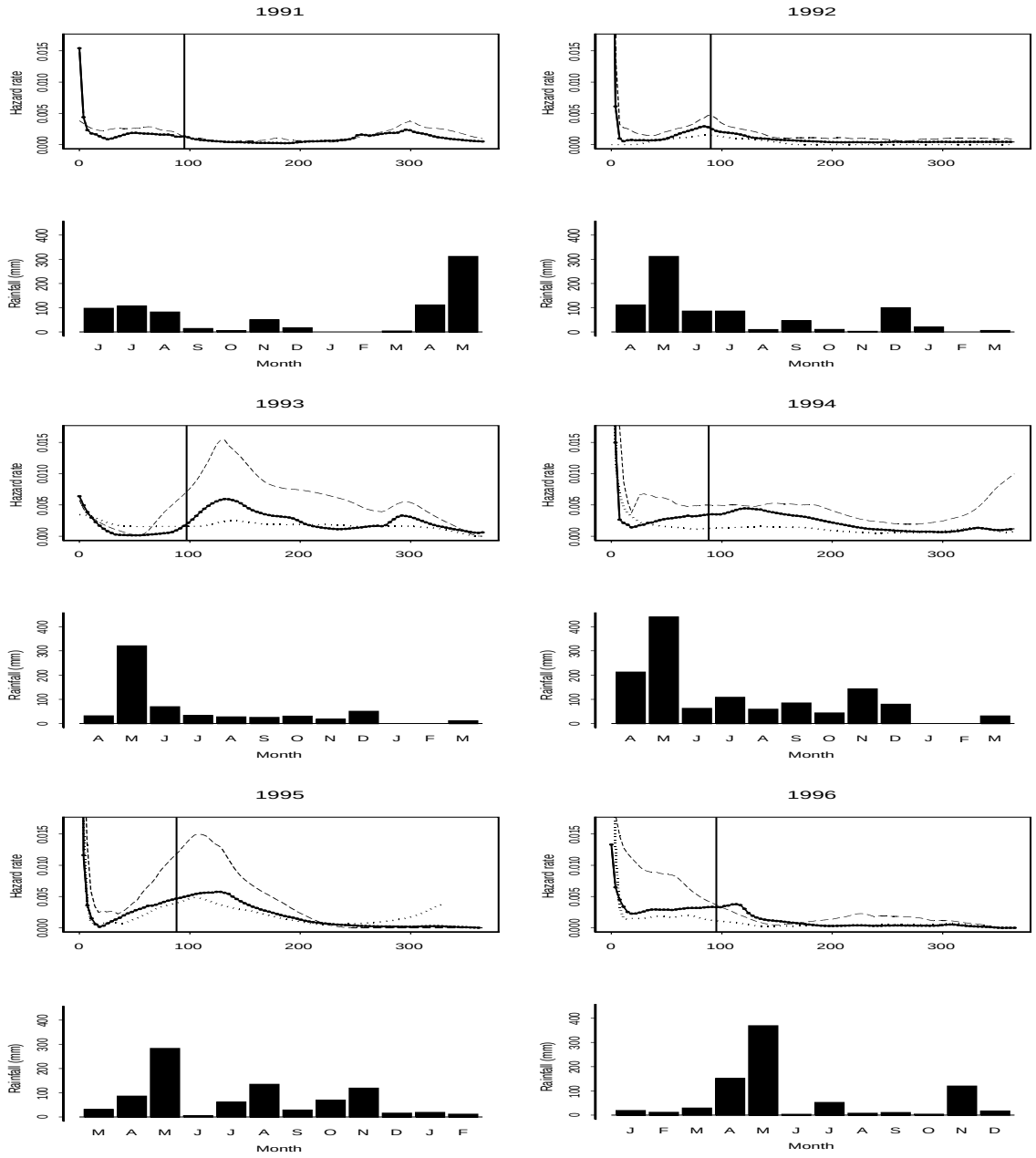



Figure 5.1: *Non-parametric estimate of the hazard function: dark line:-population average; dashed line:-Dorpers; dotted line:-Red Maasai and the associated rainfall patterns for a period of 12 months from the time of lambing. Bold vertical line indicates when the lambs were weaned.*

described in Section 5.3.1 while in Section 5.3.2 we give details of the survival analysis that was carried out.

### 5.3.1 Preliminary analysis

To assess whether the ranking of the various causes of mortality varied across genotype, a Poisson regression model (log-linear model, see Agresti, 1990) with genotype and cause of mortality as fixed effects was used. If $Y_{ij}$ is the number of lambs experiencing the $j^{\text{th}}$ cause of mortality from the $i^{\text{th}}$ genotype, then $Y_{ij}$ is a Poisson random variable with mean $\mu_{ij}$. The observed values $y_{ij}$ are the cell counts in Tables 4.2 and 4.3 for the pre- and post-weaning periods. The model used in this analysis is

$$\log(\mu_{ij}) = \mu + \vartheta_i^A + \vartheta_j^B \tag{5.3}$$

where

$\mu$= the overall mean

$\vartheta_i^A$=the breed main effects (6 levels)

$\vartheta_j^B$=the cause of death main effects (6 levels pre-weaning and 5-levels post-weaning).

Model (5.3) is a log-linear model and its goodness of fit can be assessed using the deviance statistic

$$G^2 = 2\sum\sum y_{ij} \log\left(\frac{y_{ij}}{\hat{\mu}_{ij}}\right)$$

where $\hat{\mu}_{ij}$ are the estimated cell counts. The deviance statistic has an approximate chi-square distribution with $(i-1)(j-1)$ degree of freedom. If a breed by cause interaction term is included in (5.3) then the deviance is zero as the number of parameters to be estimated is equal to the number of cells in the table. The ratio of the estimated deviance for Model (5.3) to its degrees of freedom can also be used to check for overdispersion in the model. For a Poisson model as that postulated here the ratio should be close to unity. In this analysis the calculated deviance was 32.14 with 24 degrees of freedom (df) for the pre-weaning period, whilst that of the post-weaning period was 33.87 with 20 df. In both periods the ratio of the deviance to the degree of freedom was close to unity, indicating only a small over-dispersion. This implies that the interactions between genotype and cause of mortality in both periods were not significant (at 5%) when averaged over year. In total, the number of lambs that died or were lost in the six years were as follows (from Table 4.1): 118 (15%) in 1991, 92 (12%) in 1992, 200 (25%) in 1993, 144 (18%) in 1994, 151 (19%) in 1995 and 85 (11%) in 1996. Using a Poisson regression model (Model (5.3))

we also assessed the distribution of the various causes of these deaths across the years. The resulting deviances were 64.6 on 23 df and 137.1 on 19 df for the pre- and post-weaning periods respectively. This shows that the ranking of causes of mortality varied across years, particularly in the post-weaning period indicating a significant interaction between year and cause of mortality. For example, there was a large incidence of 'lost/stolen' lambs (22%) in 1993, and only one lamb in 1996 was diagnosed as dying due to endoparasites. The majority of deaths in this year had an unknown cause. Prior to weaning there was a higher proportion of deaths due to endoparasites in 1991 (43%) than the other years. There was a higher proportion of still births (26%) than average in 1995.

Based on these preliminary findings, further analysis was carried out with age of lamb at time of death (regardless of cause) as the response variable. The time of birth and time of weaning were taken as the time of origin in the pre-weaning and post-weaning periods respectively. Thus in the post-weaning period the time-to-event of the lamb was the residual life from weaning. The relevant event to the biologist was disposal of an animal, which included animals that either died or were stolen/lost. Thus all causes of mortality as described above were regarded as 'events'. For the pre-weaning period analysis, lambs that were weaned were censored, while those that were weaned and lived beyond 365 days (from birth) were censored in the post-weaning analysis. Still born lambs were however excluded. Analysis was carried out in S-Plus (Mathsoft, 1999) where the penalized likelihood approach (Section 2.4.4) is the implemented method of estimation for the Cox proportional hazards model.

### 5.3.2 Survival analysis

In all the analysis carried out terms for the fixed effects for genotype (6 levels), year of birth (6 levels), gender (2 levels) and age of the dam (5 levels) were always included. For ease of reference we will refer to these as the baseline covariates.

Subsequently the weight at birth was included as a curvilinear term (with linear and quadratic terms) in the pre-weaning period while the weight at weaning was included for the post-weaning period. The effect of time-varying body weight as an alternative to either birth weight or weaning weight was also assessed. PCV and FEC (post weaning) and anthelmintic treatment (pre weaning) were then additionally considered also as time-varying covariates. These analyses with time-varying BWT, PCV and FEC were carried

out in order to utilize more adequately all the measurements of these traits which were repeatedly recorded as described in Section 4.3. Packed cell volume and FEC could not be used for the pre-weaning analysis as they were not measured until one month of age. The effect of treatment in the pre-weaning period was assessed using a binary covariate whose values changed at the time of treatment. Thus the associated parameter estimate is the relative change in risk due to treatment. No first order interaction terms were found to be significant in the two periods and hence these were not considered further. The analysis carried out can be summarised as follows:

Baseline covariates

- Genotype
- Year of birth
- Gender
- Age of dam

Pre-weaning covariates

- Birth weight or time-varying body weight
- Treatment

Post-weaning covariates

- Weight at weaning or time-varying body weight
- Time dependent PCV
- Time dependent FEC

For each of the settings above two analyses were undertaken in each period. The first analysis used the Cox proportional hazards model

$$\lambda_i(t) = \lambda_0(t) \exp(\boldsymbol{x}_i^T(t)\boldsymbol{\beta}) \tag{5.4}$$

where $\lambda_i(t)$ is the hazard function for the $i^{\text{th}}$ lamb, $i = 1, \ldots, n$, $\lambda_0(t)$ is the unspecified baseline hazard, $\boldsymbol{x}_i(t)$ is the incidence vector of the fixed effects for this lamb at time $t$ and $\boldsymbol{\beta}$ the vector of associated parameters.

In the second analysis, a shared frailty model with sire included as the random effect term was used. Suppose that the $i^{\text{th}}$ sire $i = 1, \ldots, G$ has $n_i$ lambs then

$$\lambda_{ij}(t) = \lambda_0(t) \exp(\boldsymbol{x}_{ij}^T(t)\boldsymbol{\beta} + w_i) \tag{5.5}$$

where $\lambda_{ij}(t)$ is the hazard function for the $j^{\text{th}}$ lamb, $j = 1, \ldots, n_i$ from the $i^{\text{th}}$ sire, $\boldsymbol{x}_{ij}(t)$ is the incidence vector of the fixed effects at time $t$ and $w_i, i = 1, \ldots, G$ is the

random sire effect. For models (5.4) and (5.5), the covariate vectors $\boldsymbol{x}_i(t)$ and $\boldsymbol{x}_{ij}(t)$ are only time-dependent if time-varying covariates such as FEC and PCV are considered. These time-varying covariates change values only at the measurement times as described in Section 4.3 and in between these time points the last observed value (LVCF) is used, resulting in a piecewise constant profile. In such cases the hazards are then proportional only between intervals in which the covariates remain constant (e.g. between 2 months and weaning). Thus the $\boldsymbol{\beta}$ parameter estimates associated with a factor variable for any model containing a time-varying covariate cannot be interpreted as overall relative risks across the pre- or post-weaning period. In such a model, these estimates can be thought as the relative risks only within intervals in which the covariate is constant (see Klein and Moeschberger, 1997, p. 275).

For the frailty term $u_i = \exp(w_i)$, we used both the log-normal and gamma frailty distributions given by (2.5) and (2.6) respectively.

## 5.4 Results

### 5.4.1 Effect of baseline covariates

The estimated survival curves for the different genotypes in the pre-weaning and post-weaning periods are shown in Figures 5.2 and 5.3 respectively. These curves are adjusted for the other factors (covariates), namely gender, age of dam and year of birth. To get these curves in each of the periods, a stratified Cox PH model (5.4) was first fitted to the data with breed as the stratification variable while the other baseline covariates were included in the model. Thus the hazard function for the $j^{\text{th}}$ lamb from the $k^{\text{th}}$ breed is

$$\lambda_{j(k)}(t) = \lambda_{0(k)}(t) \exp(\boldsymbol{x}_j^T \boldsymbol{\beta})$$

where $\lambda_{0(k)}$ is the common baseline hazard for the lambs from this breed and $\boldsymbol{\beta}$ is the vector of unknown parameters corresponding to year, gender and age of the dam effects. If we let $\hat{\boldsymbol{\beta}}$ be the vector of estimated parameters, then the estimated survival function at time $t$ for lambs from the $k^{\text{th}}$ breed is $\hat{S}_k(t) = \exp(-\hat{\Lambda}_0(t))$ where

$$\hat{\Lambda}_0(t) = \sum_{t_{(l)} \leq t} \frac{N_{(l)}}{\sum\limits_{j \in R(t_{(l)})} \exp(\boldsymbol{x}_j^T \hat{\boldsymbol{\beta}})}$$

is the Breslow estimator of the cumulative baseline (see Section 2.4.2). This is evaluated with year of birth, age of dam and gender equal to the mean values for the data within each strata.

From Figure 5.2 it is observed that the Dorpers had the highest mortality at any point in

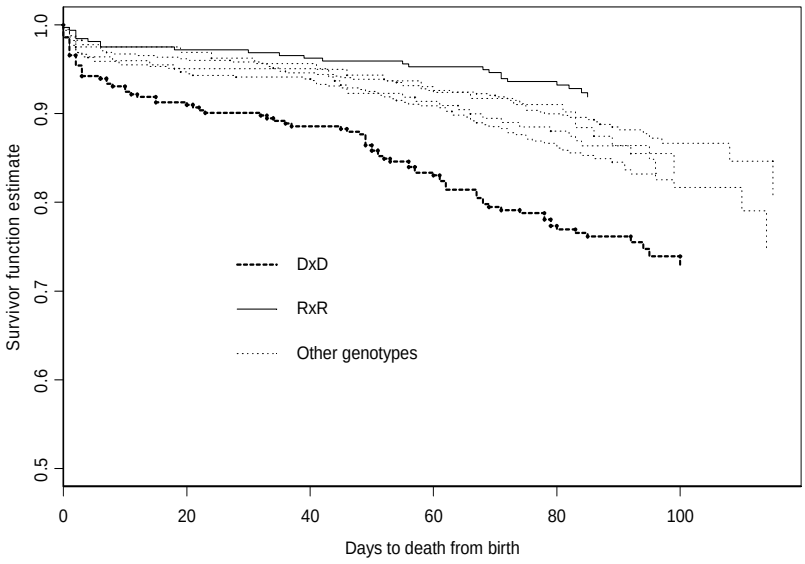



Figure 5.2: *Estimated survival curves for lambs of different genotypes in the pre-weaning period.*

time during the pre-weaning period, while Red Maasai had the lowest mortality. Throughout the post-weaning period (Figure 5.3) the Dorper had the highest average mortality while the R x (R x D) and the Red Maasai had the lowest mortality.

The $\log(-\log(\hat{S}_k(t))$ versus time plots for the different genotypes adjusted for the other factors show approximately parallel curves in the pre-weaning period (Figure 5.4). Some curves in the plots for the post-weaning period (Figure 5.5) show a tendency to cross between day 1 and day 40 after weaning, but are approximately parallel thereafter, thus the proportionality assumption holds during most of the study time.

The results of the fitted Cox PH and shared frailty hazard models are shown in Tables 5.1 and 5.2 for the pre-weaning and post-weaning periods, respectively. In both periods the shared frailty models gave parameter and standard error estimates for fixed effects which were essentially the same as those of the simpler Cox PH model. Further the parameter estimates from the gamma and log-normal frailty models were similar, with difference only

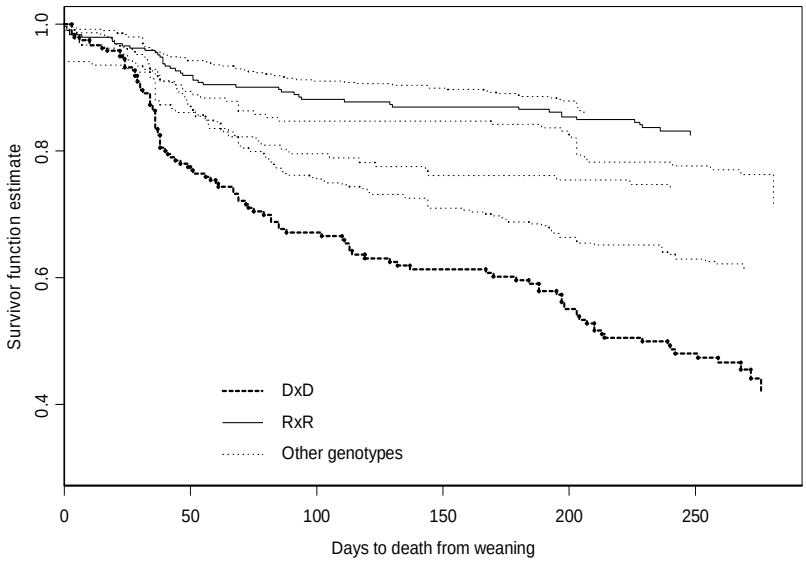



Figure 5.3: *Estimated survival curves for lambs of different genotypes in the post-weaning period.*

arising in the value of the estimated variance of the random term. Due to this and the fact that under the penalized partial likelihood approach there exists an explicit analytical formula for estimating the standard error of $\text{Var}(W) = \sigma^2$ for the log-normal distribution (Remark 6 in Section 2.4.4) only the results for this frailty distribution are tabulated.
A detailed interpretation of the parameter estimates for both periods from the shared frailty model are given below.

*Genotype*

The Dorper lambs were observed to have a higher relative risk of mortality than all the other genotypes. The relative risk of mortality for the D x (R x D) genotype relative to the Dorper lambs was 0.61 ($P< 0.01$) and this declined to 0.27 for the Red Maasai in the pre-weaning period ($P< 0.001$). A decreasing trend in the risk of mortality was observed with increasing proportion of Red Maasai in the genotype. In the post-weaning period the risk of mortality for the D x (R x D) genotype relative to the Dorper was 0.61 ($P<0.001$) and that for the Red Maasai 0.25 ($P<0.001$). The overall trends in the pre-weaning and the post-weaning periods were similar, with the exception that R x (R x D) lambs had the lowest risk of mortality post weaning. These trends are also observed in the estimated survival curves (Figs 5.2 and 5.3).

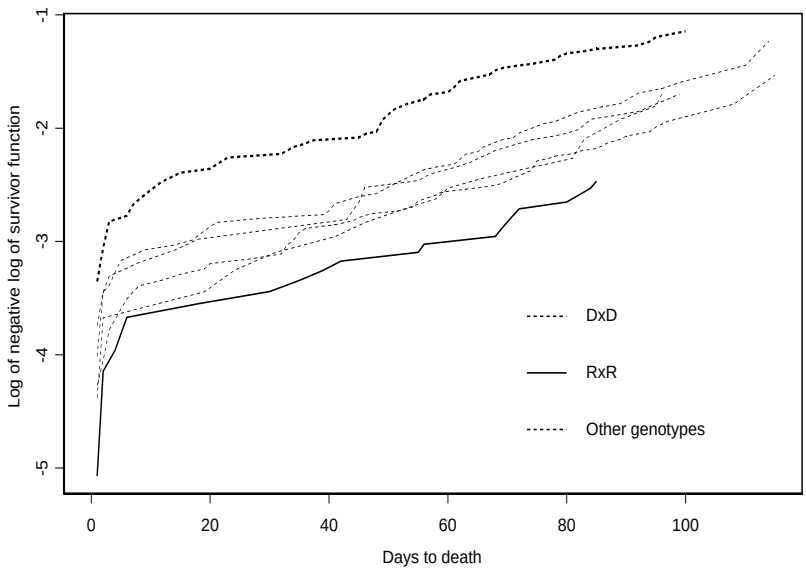



Figure 5.4: *Log of negative log of survival versus time in the pre-weaning period.*

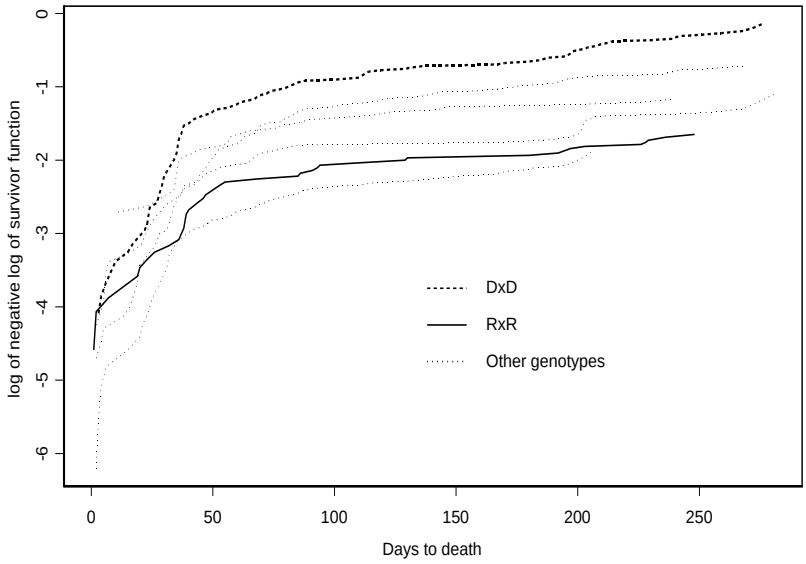



Figure 5.5: *Log of negative log of survival versus time in the post-weaning period.*

*Year*

Survival rates were different among years and appeared to be associated to some degree

Table 5.1: *Parameter estimates and hazard ratios (95% c.i.) from the Cox proportional hazards and the shared frailty models applied to survival time (regardless of cause) in the pre-weaning period.*

| Effect | No. of lambs | Cox Proportional hazards model | | Shared frailty hazards model | |
|---|---|---|---|---|---|
| | N | est±s.e. | HR(c.i.) | est± s.e. | HR(c.i.) |
| Genotype | | | | | |
| DxD | 304 | ref | 1.00 | ref | 1.00 |
| Dx(DxR) | 419 | -0.48±0.16 | 0.62(0.43,0.81) | -0.49±0.16 | 0.61(0.42,0.81) |
| DxR | 121 | -0.72±0.26 | 0.49(0.24,0.73) | -0.74±0.26 | 0.48(0.23,0.72) |
| RxD | 231 | -0.62±0.21 | 0.54(0.31,0.77) | -0.63±0.22 | 0.53(0.30,0.77) |
| Rx(RxD) | 466 | -0.83±0.17 | 0.44(0.29,0.58) | -0.85±0.18 | 0.43(0.2,0.58) |
| RxR | 204 | -1.29±0.25 | 0.28(0.14,0.41) | -1.32±0.26 | 0.27(0.13,0.40) |
| Year of birth | | | | | |
| 1991 | 363 | ref | 1.00 | ref | 1.00 |
| 1992 | 293 | 0.26±0.20 | 1.29(0.78,1.81) | 0.23±0.21 | 1.26(0.74,1.78) |
| 1993 | 365 | -1.05±0.28 | 0.35(0.16,0.54) | -1.07±0.29 | 0.34(0.15,0.53) |
| 1994 | 236 | 0.99±0.20 | 2.70(1.62,3.78) | 0.99±0.21 | 2.70(1.57,3.83) |
| 1995 | 262 | 0.75±0.21 | 2.13(1.26,2.99) | 0.74±0.22 | 2.10(1.19,3.01) |
| 1996 | 226 | 0.64±0.22 | 1.90(1.07,2.73) | 0.65±0.24 | 1.91(1.03,2.79) |
| Gender | | | | | |
| Females | 837 | ref | 1.00 | ref | 1.00 |
| Males | 908 | 0.08±0.12 | 1.08(0.84,1.33) | 0.07±0.12 | 1.07(0.83,1.32) |
| Age of dam | | | | | |
| <=2yrs | 183 | ref | 1.00 | ref | 1.00 |
| =3 yrs | 403 | 0.08±0.24 | 1.09(0.58,1.60) | 0.10±0.24 | 1.10(0.58,1.63) |
| =4 yrs | 386 | -0.08±0.25 | 0.92(0.47,1.37) | -0.07±0.25 | 0.93(0.48,1.39) |
| =5 yrs | 383 | 0.09±0.24 | 1.09(0.58,1.61) | 0.10±0.24 | 1.10(0.58,1.62) |
| >=6yrs | 390 | 0.11±0.25 | 1.12(0.58,1.66) | 0.11±0.25 | 1.12(0.58,1.66) |
| Sire | | | | | |
| variance (s.e.) | | 0.00 | | 0.052(0.049) | |

with variation in rainfall (Figure 5.1). In the pre-weaning period the risk of mortality was least in 1993, which was on average 34(%) that observed in 1991 (P<0.001) and highest in 1994, approaching three times that in 1991 (P<0.001). In the post-weaning period the risk of mortality was about 5 fold higher in 1993, 1994 and 1995 compared with 1991 and 1992. On the other hand the risk of mortality in 1996 was reduced by 8% (P<0.001) in the pre-weaning period when compared to 1991 but was similar in the post-weaning period. These findings concur with the hazard estimates shown in Figure 5.1 with 1993 having the lowest average risk in the pre-weaning period and the highest in the post-weaning period. Further the estimated hazard function was higher in the pre-weaning period in 1996 than

in 1991 but post weaning, the estimates are of the same magnitude.

Table 5.2: *Parameter estimates and hazard ratios (95% c.i.) from the Cox proportional hazards and the shared frailty models applied to survival time (regardless of cause) in the post-weaning period.*

| Effect | No. of lambs | Cox Proportional hazards model | | Shared frailty hazards model | |
|---|---|---|---|---|---|
| | N | est±s.e. | HR(c.i.) | est± s.e. | HR(c.i.) |
| Genotype | | | | | |
| DxD | 223 | ref | 1.00 | ref | 1.00 |
| Dx(DxR) | 340 | -0.48±0.13 | 0.62(0.46,0.78) | -0.51±0.13 | 0.60(0.44,0.76) |
| DxR | 101 | -0.78±0.19 | 0.46(0.28,0.63) | -0.79±0.19 | 0.45(0.28,0.62) |
| RxD | 200 | -1.00±0.18 | 0.37(0.24,0.50) | -1.02±0.19 | 0.36(0.23,0.49) |
| Rx(RxD) | 398 | -1.58±0.16 | 0.21(0.14,0.27) | -1.62±0.17 | 0.20(0.13,0.27) |
| RxR | 180 | -1.34±0.19 | 0.26(0.16,0.36) | -1.39±0.20 | 0.25(0.15,0.35) |
| Year of birth | | | | | |
| 1991 | 309 | ref | 1.00 | ref | 1.00 |
| 1992 | 242 | 0.01±0.21 | 1.01(0.59,1.42) | -0.04±0.22 | 0.96(0.55,1.38) |
| 1993 | 347 | 1.58±0.16 | 4.87(3.36,6.39) | 1.60±0.17 | 4.95(3.29,6.60) |
| 1994 | 170 | 1.45±0.19 | 4.27(2.66,5.87) | 1.48±0.21 | 4.40(2.64,6.16) |
| 1995 | 203 | 1.44±0.19 | 4.21(2.62,5.79) | 1.45±0.21 | 4.25(2.52,5.98) |
| 1996 | 171 | 0.50±0.25 | 1.64(0.84,2.45) | 0.54±0.26 | 1.72(0.83,2.61) |
| Gender | | | | | |
| Females | 695 | ref | 1.00 | ref | 1.00 |
| Males | 747 | 0.31±0.10 | 1.36(1.11,1.62) | 0.32±0.10 | 1.38(1.12,1.65) |
| Age of dam | | | | | |
| <=2yrs | 158 | ref | 1.00 | ref | 1.00 |
| =3 yrs | 340 | -0.36±0.17 | 0.70(0.47,0.92) | -0.36±0.17 | 0.70(0.47,0.93) |
| =4 yrs | 330 | -0.63±0.17 | 0.53(0.35,0.71) | -0.64±0.18 | 0.53(0.35,0.71) |
| =5 yrs | 314 | -0.49±0.17 | 0.62(0.41,0.82) | -0.49±0.17 | 0.61(0.41,0.81) |
| >=6yrs | 300 | -0.79±0.18 | 0.46(0.29,0.62) | -0.79±0.19 | 0.46(0.29,0.62) |
| Sire | | | | | |
| variance (s.e.) | | 0.00 | | 0.054(0.037) | |

*Gender*

The risk of mortality for male lambs was about a third higher than that for female lambs during the post-weaning period ($P<0.01$), but there was no significant gender effect in the pre-weaning period.

*Age of dam.*

There was no significant effect of age of the dam on the risk of mortality in the pre-weaning period implying that mothering capability was independent of age. In the post-weaning period, however, the risk of mortality of lambs born to mothers that were two years of

age or younger was higher that of lambs born to older ewes. Compared with lambs born to 2-year old mothers the relative risk was 0.69 for lambs born to 3-year old mothers and 0.43 for lambs born to mothers aged six years or more (P<0.001).

*Sire variance*

The estimated sire frailty variance and its standard error under the log-normal frailty for the pre- and post-weaning periods are also shown in Tables 5.1 and 5.2. We also obtained the estimated frailties ($u_i$'s) for the 76 sires for both the pre-weaning and post-weaning periods. To get a better visualisation of the frailty effect, we plotted the estimated survival

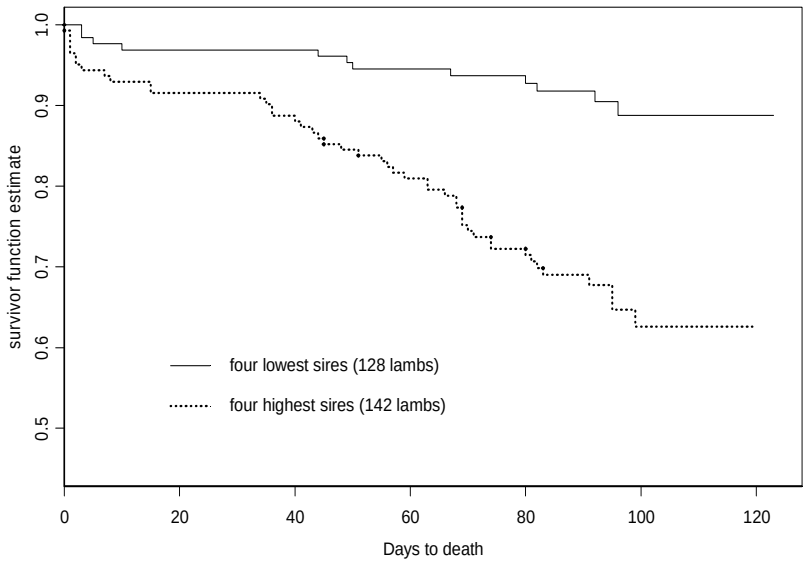



Figure 5.6: *Estimated survival curves for lambs from four sires with largest and four sires with lowest frailty estimates in the pre-weaning period.*

curves of lambs from the four sires with the highest and the four sires with the lowest frailty values (namely top and bottom 5% of the 76 sires). These curves are shown in Figs 5.6 and 5.7 for the two periods. In the pre-weaning period, of the four sires with the highest estimates of $u_i$, two were Dorpers and two were Red Maasai while those with the lowest estimates were all Dorpers. Post weaning, there were three Dorpers and one Red Maasai sire for both the high and the low estimates of the frailties ($u_i$'s). None of these sires were the same in the two periods. It is observed in these plots that the lambs from the sires with high values experienced events earlier than the lambs from sires with

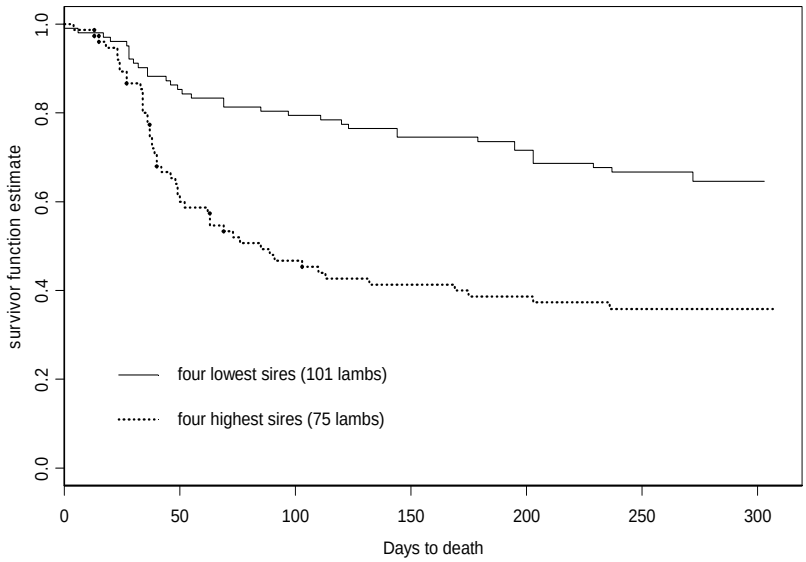
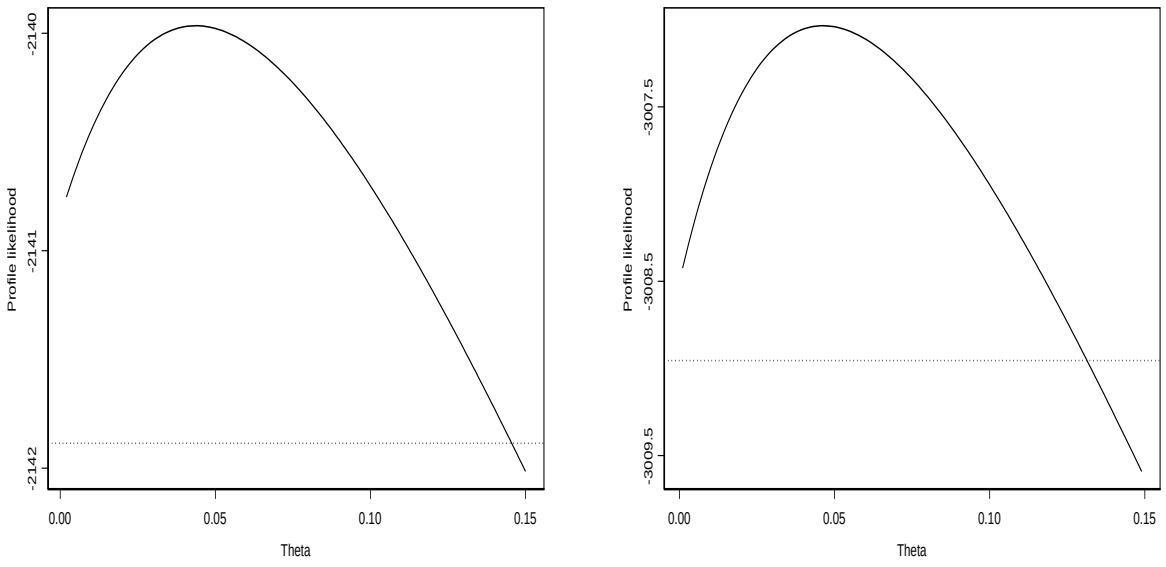



Figure 5.7: *Estimated survival curves for lambs from four sires with largest and four sires with lowest frailty estimates in the post-weaning period.*

low values. This not-withstanding, the estimated sire variance in the two periods was non-significant (at 5%). Figure 5.8 shows the profile log-likelihood for $\theta$ (sire variance) for the pre- and post-weaning periods. In each of this, the approximate 95% confidence interval (c.i.) for $\theta$ includes zero. This confidence interval is obtained by taking two values of $\theta$ for which the profile log-likelihood lies 1.92 units (dotted line in Figure 5.8) below the maximum profile log-likelihood value within each period. The respective log-likelihood values for the model with the baseline covariates with and without frailty were -2139.97 and -2140.34 for the pre-weaning period and -3007.35 and -3008.50 for the post-weaning period. If we conjecture that the theoretical results derived in Chapter 3 also hold for the current model (semi-parametric log-normal frailty model), then the likelihood ratio test statistics for the two periods are 0.74 and 2.3 with P-values 0.20 ($= Pr(\chi^2_1 > 0.74)/2$) and 0.07 ($= Pr(\chi^2_1 > 2.3)/2$) respectively. Thus, using either the approximate 95% c.i. or the likelihood ratio test, the null hypothesis ($H_0 : \theta = 0$) of no heterogeneity is not rejected in the two periods. This implies that the time-to-event of the lambs from the different sires are homogeneous.




(a) (b)

Figure 5.8: *Profile log-likelihood for $\theta$ (sire variance) for the (a) pre- weaning and (b) post-weaning periods.*

### 5.4.2 Effect of time-independent body weight

In this analysis the weight at birth or the weight at weaning was used additionally with the baseline covariates in the pre- and post-weaning periods respectively. In both periods there was a curvilinear relationship between body weight and risk of mortality. This relationship had significant linear and quadratic terms for birth weight in the pre-weaning period and weaning weight in the post-weaning period ($P<0.001$) (see first part of Tables 5.3 and 5.4). In the two periods the relative risk of mortality decreased quadratically with increased birth weight and weaning weight (Figure 5.9). This implies that lambs that were heavy in body weight at birth or weaning had lower risk of mortality when compared to lighter lambs. The detailed results for the baseline covariates are given in the first column of Tables 5.7 (pre weaning) and 5.8 (post weaning) at the end of this chapter.

Notably, in the pre-weaning period there was now no difference in the risk of mortality for lambs born in 1994 and 1995, when compared to those born in 1991. The higher risk observed earlier in the unadjusted model (Table 5.1) could be due to the fact that the

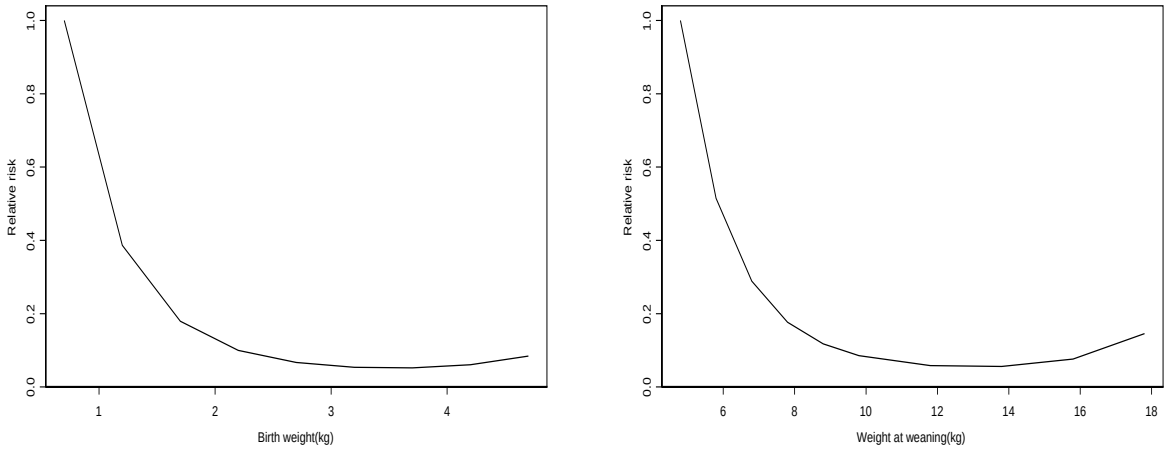
Table 5.3: *Parameter estimates from a shared frailty hazard model applied to survival time (regardless of cause), in the pre-weaning period for birth weight and time-varying body weight alternatively, and treatment, adjusted for baseline covariates.*

| Covariate | Parameter estimate±s.e. | Covariate | Parameter estimate ±s.e |
|---|---|---|---|
| Birth weight (kg) | | Time varying body weight (kg) | |
| - Linear | -2.38 ± 0.52 | - Linear | -1.06 ± 0.11 |
| - Square | 0.323 ± 0.102 | - Square | 0.038 ± 0.008 |
| Sire variance (s.e.) | 0.066 (0.053) | Sire variance | 0.067 (0.053) |
| Birth weight (kg) | | Time varying body weight (kg) | |
| - Linear | -2.36 ± 0.52 | - Linear | -1.04 ± 0.12 |
| - Square | 0.318 ± 0.102 | - Square | 0.035 ± 0.008 |
| Treatment | | Treatment | |
| - Not treated | reference | - Not treated | reference |
| - Treated | -0.51 ± 0.19 | - Treated | -0.92 ± 0.20 |
| Sire variance (s.e.) | 0.067 (0.053) | Sire variance | 0.063 (0.052) |

lambs born in these two years were much lighter than those born in 1991 (see Figure 4.2). Further, when the weight at weaning was taken into account, the Rx(RxD) and RxR genotypes now had similar relative risk of mortality when compared to the Dorpers. In addition there was a non-significant effect of the age of dam after taking into account the weight of the lamb. This could be due to the biological fact that lambs born to young dams are lighter in body weight, possibly due to lower milk production of the dam in her first parity. This difference in body weight could have lead to the significant age of dam effect in the unadjusted model (Table 5.2).

### 5.4.3 Effect of time-dependent covariates

In Section 4.3, we reported that BWT, PCV and FEC were periodically recorded until the lambs were almost one year old. In this section we assess the effect of these traits on the risk of mortality as they evolve over time. The effect of anthelmintic treatment in the pre-weaning period was also assessed in the presence of both birth weight and time-varying body weight (Table 5.3). These time-varying covariates were additionally included in models with the baseline covariates. The detailed results for the baseline covariates for the pre- and post-weaning periods are given in Tables 5.7 and 5.9 in the appendix of this chapter.




(a) (b)

Figure 5.9: *The decrease in risk of mortality with increasing body weight at (a) birth in the pre-weaning period and (b) weaning in the post-weaning period, with other factors of genotype, gender, year and age of dam held constant in lambs raised.*

*Body weight*

In both the pre- and post-weaning periods, the risk of mortality associated with body weight decreased as lambs gained weight. Additionally, at each time point heavier lambs had a lower risk of mortality than lighter lambs (Figure 5.10). Pre weaning, the relative hazards of the other genotypes relative to Dorper were slightly decreased when adjusted for time-varying body weight and now ranged from 0.55 to 0.17 (Table 5.7).

In the post-weaning, the hazard ratios for other genotypes were also decreased after adjusting for body weight. As in the analysis with time-invariant body weight (Section 5.4.2), the Rx(RxD) and RxR genotype had the lowest but similar relative mortality when compared to the Dorper. There was also no difference in the risk of mortality for lambs born in 1994 and 1995 in both the pre- and post-weaning periods and the age of dam was not significant (Table 5.9).

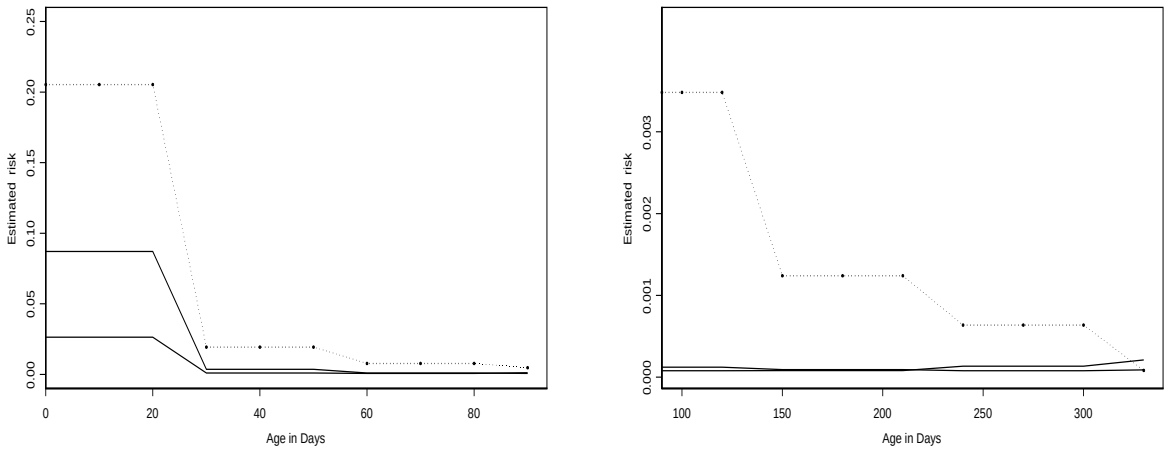
*Treatment*

The effect of treatment was only considered in the pre-weaning period since all lambs were treated together during the post-weaning period. The risk of mortality during the subsequent month for two lambs of the same weight was reduced by 0.40($P<0.001$) during

the next month when treated $(1 - \exp(-0.51))$ (Table 5.3).

*Packed cell volume and faecal egg count*

The means (standard deviations) of time-varying PCV and time-varying natural logarithm of FEC post-weaning were 25 (5.4) percent and 7.36 (1.27) log(epg+25), respectively. Both these time-dependent covariates had significant relationship with the risk of mortality when introduced in the model for the post-weaning period (Table 5.4). As noted in Section 5.3.2, these covariates are assumed to be piecewise constant. Between any two post-weaning sampling times for any two similar lambs from the same sire, and with PCV differing by one standard deviation, the risk of mortality of the lamb with the higher PCV relative to that of the lamb with the lower PCV ranged from 0.96 to 0.31 when the PCV ranged from 35 to 20 percent. On the other hand, if the natural logarithm of FEC for a lamb increased by one standard deviation, the relative risk of mortality of the lamb with the lower FEC relative to that of the lamb with the higher FEC ranged from 0.37 to 0.98 when the natural logarithm of FEC ranged from 10 (corresponding to 22,000 epg) to 7 (1,100 epg). When both variables were included simultaneously in the




(a) (b)

Figure 5.10: *Changes in risk of mortality with time-varying body weight in the (a) pre-weaning and (b) post-weaning periods for three lambs selected at random each month with body weights corresponding to the 2.5% (dashed line ), 50% (thin line ) and 97.5% (thick line ) percentiles in the distribution of body weight. Ranges of body weight were: at birth (1.5 to 3.8 kg), 30 days (4.2 to 9.9 kg), 60 days (5.5 to 13.3 kg), weaning (6.3 to 16.7 kg), 150 days(7.5 to 17.8 kg), 240 days (9.1 to 21.3 kg) and 330 days (16.0 to 24.5 kg).*

Table 5.4: *Parameter estimates from a shared frailty hazard model applied to survival time (regardless of cause), in the post-weaning period for birth weight and time-varying body weight alternatively, and with time-varying PCV (%) and LFEC(*$\log e.p.g$*) adjusted for baseline covariates.*

| Covariate | Parameter estimate±s.e. | Covariate | Parameter estimate ±s.e |
|---|---|---|---|
| Weaning weight | | Time varying body weight | |
| - Linear | -1.36 ± 0.10 | - Linear | -1.10 ± 0.07 |
| - Square | 0.050± 0.004 | - Square | 0.032 ± 0.003 |
| Sire variance (s.e) | 0.121 (0.051) | Sire variance (s.e.) | 0.183 (0.064) |
| Weaning weight | | Time varying body weight | |
| - Linear | -0.93 ± 0.11 | - Linear | -0.79 ± 0.07 |
| - Square | 0.034± 0.005 | - Square | 0.024 ± 0.003 |
| PCV | | PCV | |
| - Linear | -0.45 ± 0.04 | - Linear | -0.44 ±0.04 |
| - Square | 0.007 ± 0.001 | - Square | 0.007 ±0.001 |
| Sire variance (s.e.) | 0.102 (0.047) | Sire variance (s.e.) | 0.108 (0.049) |
| Weaning weight | | Time varying body weight | |
| - Linear | -1.27 ±0.11 | - Linear | -1.04 ± 0.08 |
| - Square | 0.047± 0.005 | - Square | 0.030± 0.003 |
| $\log(FEC+25)$ | | $\log(FEC+25)$ | |
| - Linear | -0.07 ± 0.01 | - Linear | -0.07 ± 0.01 |
| - Square | 0.001 ±0.0001 | - Square | 0.102 ± 0.050 |
| Sire variance (s.e.) | 0.093 (0.048) | Sire variance (s.e.) | 0.105 (0.051) |
| Weaning weight | | Time varying body weight | |
| - Linear | -0.90 ± 0.11 | - Linear | -0.75 ± 0.08 |
| - Square | 0.033 ± 0.005 | - Square | 0.023 ± 0.003 |
| PCV | | PCV | |
| - Linear | -0.46 ± 0.04 | - Linear | -0.45 ± 0.04 |
| - Square | 0.007 ± 0.001 | - Square | 0.007 ± 0.001 |
| $\log(FEC+25)$ | | $\log(FEC+25)$ | |
| - Linear | -0.07 ± 0.01 | - Linear | -0.06 ± 0.01 |
| - Square | 0.001± 0.0001 | - Square | 0.001± 0.0001 |
| Sire variance (s.e.) | 0.077 (0.045) | Sire variance (s.e.) | 0.087 (0.047) |

model the significance of both effects was maintained, i.e., PCV and FEC tended to have additive effects on survival (Table 5.4).

The difference in the relative risk of mortality was slightly decreased for lambs with more than 50% Red Maasai in the genotype when adjusted for PCV. Including FEC had minimal effect.

*Sire variance after adjusting for varying covariates.*

Inclusion of body weight in the model, increased the value of the sire variance post weaning

Table 5.5: *Parameter estimates and hazard ratios (95% c.i.) for genotype from a shared frailty hazard model applied to survival time with cause of death only restricted to mis-mothering in the pre-weaning period and to endoparasite in the post-weaning period, adjusted for the baseline covariates.*

| Effect | Mis-mothering only | | | Endoparasites only | | |
|---|---|---|---|---|---|---|
| | No. of lambs | Parameter estimate±s.e. | Hazard ratio (c.i.) | No. of lambs | Parameter estimate±s.e. | Hazard ratio (c.i.) |
| Genotype | | | | | | |
| DxD | 304 | ref | 1.00 | 210 | ref | 1.00 |
| Dx(DxR) | 419 | -0.50±0.27 | 0.61(0.29,0.92) | 314 | -0.38±0.19 | 0.68(0.43,0.94) |
| DxR | 121 | -0.82±0.44 | 0.44(0.06,0.82) | 77 | -0.78±0.28 | 0.46(0.21,0.71) |
| RxD | 231 | -0.57±0.38 | 0.57(0.14,0.99) | 194 | -1.01±0.28 | 0.36(0.17,0.56) |
| Rx(RxD) | 466 | -0.92±0.30 | 0.40(0.16,0.64) | 346 | -1.70±0.25 | 0.18(0.09,0.27) |
| RxR | 204 | -1.14±0.40 | 0.32(0.07,0.57) | 130 | -1.31±0.29 | 0.27(0.12,0.42) |

Year 1996 is excluded from the analysis for endoparasites because there was only one case of death due to endoparasites in this year. For comparative purposes hazard ratios for all causes corresponding to those in Table 5.2 but excluding 1996 were 1.00, 0.59, 0.35, 0.32, 0.20 and 0.25, respectively.

(Table 5.4), but the variance decreased again when PCV and FEC were added. Alterations to the model pre weaning (Table 5.3), however, had no significant influence on the sire variance.

### 5.4.4 Effect of baseline covariates for mis-mothering and endoparasite deaths

Since 31% of the deaths pre weaning were due to mis-mothering and 47% were associated with endoparasites in the post-weaning period, analysis with the baseline covariates was repeated for these causes of death only within each of the respective periods. Lambs that died from other causes were censored. The endoparasite analysis excluded the year 1996 since only one death in this year was diagnosed post weaning as associated with endoparasites. The parameter estimates for genotypes in these model are given in Table 5.5. The relative hazard for the genotypes for mis-mothering deaths had same trend as that observed for all causes in the pre-weaning period (cf Table 5.1). Similar relative hazards for the genotypes were obtained for deaths due to endoparasites alone compared with deaths due to all causes (shared frailty column Table 5.2).

### 5.4.5 Effect of baseline covariates with lambs stolen/lost as censored records

In all the analyses that have been carried out above, the lambs that were stolen or lost were treated as events as mentioned in Section 5.3.1. The analysis with the baseline covariates for both pre and post weaning periods, was now repeated with records for lamb that were stolen/lost now censored. The results for genotype and year of birth estimates from the shared frailty model are given in Table 5.6. Similar relative hazard estimates were obtained for genotype in both the pre- and post-weaning periods (see shared frailty column Tables 5.1 and 5.2). The estimates for the years 1993-1996 were each slightly increased by about 38% in the pre-weaning period. Post weaning there was a slight decrease in the relative risk for 1993. A slight increase was observed for the years 1994-1996. This can be attributed to the fact that pre weaning, the years 1991-1992 accounted for 23(72%) of the lambs that were stolen/lost. In the post-weaning period lambs lost/stolen were: 9(15%) in 1991, 5(8%) in 1992, 39(63%) in 1993 and a total of 9(15%) for the years 1994-1996.

Table 5.6: *Parameter estimates and hazard ratios (95% c.i.) for genotype and year of birth adjusted for gender and age of dam from a shared frailty hazard model applied to survival time in the pre- and post-weaning periods, with stolen lambs censored.*

| | Pre-weaning | | | Post-weaning | | |
|---|---|---|---|---|---|---|
| Effect | No. of lambs | Parameter estimate±s.e. | Hazard ratio (c.i.) | No. of lambs | Parameter estimate±s.e. | Hazard ratio (c.i.) |
| Genotype | | | | | | |
| DxD | 304 | | | 223 | | |
| Dx(DxR) | 419 | -0.57±0.17 | 0.56(0.38,0.75) | 340 | -0.52±0.14 | 0.59(0.43,0.76) |
| DxR | 121 | -0.72±0.26 | 0.49(0.24,0.74) | 101 | -0.71±0.20 | 0.49(0.29,0.68) |
| RxD | 231 | -0.73±0.24 | 0.48(0.25,0.71) | 200 | -1.00±0.21 | 0.37(0.22,0.52) |
| Rx(RxD) | 466 | -0.90±0.19 | 0.41(0.25,0.56) | 398 | -1.68±0.19 | 0.19(0.12,0.25) |
| RxR | 204 | -1.40±0.27 | 0.25(0.12,0.38) | 180 | -1.38±0.22 | 0.25(0.15,0.36) |
| Year of birth | | | | | | |
| 1991 | 363 | | | 309 | | |
| 1992 | 293 | 0.08±0.24 | 1.09(0.58,1.59) | 242 | -0.01±0.24 | 0.99(0.53,1.45) |
| 1993 | 365 | -0.90±0.29 | 0.41(0.17,0.64) | 347 | 1.49±0.19 | 4.45(2.81,6.09) |
| 1994 | 236 | 1.13±0.23 | 3.08(1.70,4.46) | 170 | 1.60±0.22 | 4.97(2.86,7.08) |
| 1995 | 262 | 0.91±0.24 | 2.48(1.34,3.62) | 203 | 1.54±0.22 | 4.66(2.63,6.68) |
| 1996 | 226 | 0.75±0.25 | 2.13(1.07,3.18) | 171 | 0.60±0.28 | 1.82(0.81,2.82) |

### 5.4.6 Heterosis

An additional analysis to assess for evidence of heterosis was also carried out in the pre- and post-weaning periods. This was achieved by substituting the genotype term in the model with baseline covariates with appropriate linear contrasts as given in Baker et al. (2003). In general heterosis is defined as the superior performance of crossbred animals relative to the average performance of the purebreds involved in the cross. This could be due to combining genes from different breeds thus concealing the effects of inferior genes. Two types of heterosis are the individual and maternal heterosis. Individual heterosis is the better performance of a crossbred animal over the average of the pure breeds. For example, a RxD lamb may perform better than the average of (DxD) and (RxR) lambs. Maternal heterosis is the better performance of a crossbred mother (such as increased litter size) relative to the average of the pure bred mothers.
In both periods there was no evidence of heterosis, either as a direct individual or maternal effect.

## 5.5 Heritability estimates

Often, when frailty proportional hazard models are fitted in animal studies, the frailty variance is utilised in the calculation of heritability estimates (Ducrocq et al., 1988, Ducrocq, 2000, Henryon et al., 2001, Southey et al., 2001, Yazdi et al., 2000). From the linear mixed models perspective in Chapter 4.5, we saw that the heritability estimate is the ratio of genetic variation to phenotypic variation given by (4.2). In the frailty model approach, Ducrocq and Casella (1996) derived the heritability estimate defined as

$$h_s^2 = \frac{4\sigma_s^2}{\sigma_s^2 + \frac{\pi^2}{6}} \tag{5.6}$$

where $\sigma_s^2$ is the estimated variance of the frailty term (sire variance in this case) and which provides an estimate of the genetic variation. The denominator as before is an estimate of the phenotypic variance, where $\frac{\pi^2}{6}$ is the variance of the standard extreme value distribution (Lawless, 1982). This latter variance comes from the relationship of the extreme value distribution with the Weibull distribution, when the response variable

(time-to-event) is transformed on to a log scale as shown below.
Consider the following frailty model

$$\lambda_{ij}(t) = \lambda_0(t)\exp(\boldsymbol{x}_{ij}^T\boldsymbol{\beta} + w_i) \tag{5.7}$$

where $w_i$ is the random effect of the $i^{\text{th}}$ sire and $\lambda_0(t)$ is assumed to follow a Weibull distribution. This implies that

$$\lambda_0(t) = \lambda\rho t^{\rho-1}$$

where $\lambda$ and $\rho$ are the scale and shape parameters, respectively. Thus the time-to-event has density

$$f_T(t) = \lambda\rho t^{\rho-1}\exp(\boldsymbol{x}_{ij}^T\boldsymbol{\beta} + w_i)\exp\left(-\lambda t^{\rho}\exp(\boldsymbol{x}_{ij}^T\boldsymbol{\beta} + w_i)\right).$$

Let Y be the log transformation of T (i.e., $Y = \log T$). Then the density of Y is

$$\begin{aligned} f_Y(y) &= \lambda\rho e^{y\rho}\exp(\boldsymbol{x}_{ij}^T\boldsymbol{\beta} + w_i)\exp\left(-\lambda e^{y\rho}\exp(\boldsymbol{x}_{ij}^T\boldsymbol{\beta} + w_i)\right) \\ &= \rho\exp\left[\left(\log\lambda + y\rho + \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + w_i\right) - \exp\left(\log\lambda + y\rho + \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + w_i\right)\right] \end{aligned}$$

If we let $\varepsilon = \log\lambda + Y\rho + \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + w_i$ then $\varepsilon$ has the density

$$f_\varepsilon(\varepsilon) = \exp(\varepsilon - \exp(\varepsilon))$$

which is the extreme value density, with variance as $\frac{\pi^2}{6}$ (Lawless, 1982). Hence Var(Y)=$\sigma_s^2 + \frac{\pi^2}{6}$ is taken as an estimate of the phenotypic variance. Thus heritabilility can be estimated as in (5.6). We note that this estimate is on a log-scale ($\log T$) and needs to be transformed back to the original scale. Using Taylor series expansion Ducrocq (1999a) derived the approximation

$$h_{sT}^2 = (\exp(\nu\rho))^2\,\frac{4\sigma_s^2}{\sigma_s^2 + \frac{\pi^2}{6}} \tag{5.8}$$

for original time scale, where $\nu = E[\varepsilon]$.
Korsgaard et al. (1999) proposes a modified expression of the heritability estimate given as

$$h_s^2 = \frac{4\sigma_s^2}{\sigma_s^2 + \sigma_e^2 + \frac{\pi^2}{6}} \tag{5.9}$$

where $\sigma_e^2$ is the residual variability. For this model, the random effect term used in the frailty model is $w_{ij} = s_i + e_{ij}$ where $s_i$ is the sire effect and $e_{ij}$ is a residual effect of

the $j^{\text{th}}$ animal from the $i^{\text{th}}$ sire. We saw in Section 2.3 that conditional on the frailty, there is independence of the individuals in a cluster. Thus, the underlying assumption in Model (5.7) is that conditional on the sire effect, the lambs from the same sire are independent. This has the implication that either all the lambs from the same sire come from different mothers or that the maternal genetic effect is insignificant. In the random effect formulation proposed by Korsgaard et al. (1999), this independence assumption is relaxed. Nevertheless, this latter model is not within the class of shared frailty models as more than one random effect is used per cluster.

In a more recent paper Yazdi et al. (2002) use $\sigma_s^2 + 1$ as an estimate for the phenotypic variance, based on a model similar to (5.7) but without covariates (i.e., $\boldsymbol{x}_{ij} \equiv \mathbf{0}$). This is motivated by the fact that (5.8) had been observed to be sensitive to the choice of the Weibull shape parameter $\rho$.

All the above heritability estimate expressions have been derived from parametric frailty models with a Weibull baseline hazard. The use of $\frac{\pi^2}{6}$ or 1 in a Cox proportional hazards model is an open question under discussion. For this reason heritability estimates were not calculated in this study.

## 5.6 Discussion

Previous analyses of the lamb data used in this study has shown Red Maasai sheep to be more resistant and resilient to gastro-intestinal parasites and more productive than Dorper sheep (Baker, 1998, Baker et al., 1999, 2003). In the current analysis the overall lamb mortality averaged 19% in the pre-weaning period which is within the 12% to 50% range reported for lamb mortality for tropical sheep (Traore et al., 1985, Wilson et al., 1993). In the post-weaning period it was 31%. As in the previous analysis, the Red Maasai are shown to perform better than the Dorper in terms of survival. The Red Maasai had about a three quarter lower risk of mortality than the Dorper in both the pre- and post-weaning periods.

Although there were year-to-year variations in the proportions of deaths caused by endoparasites, the ranking of the frequency of this disease across genotypes was remarkably constant. Furthermore, when differences in survival post-weaning across breeds due to endoparasites were compared with corresponding differences for all causes of mortality,

similar results were obtained. This implies that the differences in survival manifested between Dorper and Red Maasai breeds were associated with a variety of causes of death, not only endoparasites. Indeed, prior to weaning endoparasites accounted for only one fifth of all deaths. This suggests that Red Maasai sheep are manifesting a degree of general adaptability to tropical conditions, which includes enhanced resistance to or torelance to specific diseases such as haemonchosis.

Survival of animals has often been analysed as binary traits (0/1 for alive or dead) at arbitrarily defined time points in the animal's life span. Only the overall mortality to specific time points is of interest in such an analysis. However, in survival analysis all the information available in the lifespan of an animal can be used efficiently, since censored observations and uncensored observations are combined in a single analysis. The loss in information when failure time is analyzed through a logistic rather than a survival analysis approach was assessed by Yazdi et al. (2002). One major advantage of this approach over that of logistic regression is the ability to incorporate covariates that vary with time such as treatment, body weight, PCV and FEC. Lambs with low PCV or high FEC on a given sampling occasion were more susceptible than others to high mortality during the next month, despite treatment. This association with mortality appears to be independent of body weight. This is an important result because it demonstrates that animals that have already been infected, resulting in a high FEC and low PCV, have a greater risk of mortality than those with more normal values. This is despite treatment that occurred on average every 5-6 weeks post-weaning. The risk of mortality was substantially lowered in the pre-weaning period by treatment, demonstrating the impact of treatment generally on mortality to weaning.

Time-varying body weight was also strongly associated with mortality. When introduced into the model post-weaning a large increase occurred in the value of the sire variance component. This implies that disease and body weight affected the chances of survival independently and that by adding body weight to the model the direct genetic influence of sire on survival associated with disease could be seen more clearly. By introducing PCV and FEC, variables associated with disease, into the model the sire variance was again reduced, confirming the indication of genetic differences in PCV and FEC among sires (Baker, 2003).

Frailty models have been used for other species of livestock, for example in assessing the length of productive life in dairy cattle (Ducrocq et al., 1988), to assess viability of lay-

ing hens (Ducrocq, 2000), in obtaining estimates of longevity of Swedish horses (Wallin et al., 2000) and sows (Yazdi et al., 2000) and for assessing genetic variation for disease resistance in growing pigs (Henryon et al., 2001). The frailty variance in this studies has been utilised in the calculation of heritability estimates (Ducrocq et al., 1988, Ducrocq (2000), Henryon et al., 2001, Southey et al., 2001, Yazdi et al., 2000). The definitions of heritability for survival proposed in the literature have been derived based on a parametric hazard model, with a Weibull baseline hazard. The use of these heritability definitions in a semi-parametric frailty model is an open question that has not been resolved. Due to this fact no heritability estimates were calculated in this study.

Finally, above we have only presented the results from the log-normal frailty model. As noted earlier, the results from the gamma frailty model were similar to those tabulated above. Corresponding to the shared frailty models fitted in Tables 5.1 and 5.2, the sire variance estimates from the gamma frailty model were 0.005 and 0.0326 respectively. These estimates translate to estimates for the parameter $\gamma$ in the frailty model (2.6). Thus, the similarity of the results from this model to those from the log-normal frailty model may be due to the fact that the gamma tends to the log-normal when $\gamma$ is small (see Section 2.3.2). No explicit expression exists for calculating standard errors for these variance estimates from the gamma model.

## 5.7 Appendix

Table 5.7: *Parameter estimates for baseline covariates in the pre-weaning period from a shared frailty hazard model with birth weight and time-varying body weight alternatively, and treatment.*

| | Birth weight | | Time-varying body weight | |
|---|---|---|---|---|
| | only | with Treatment | only | with Treatment |
| | est±s.e. | est±s.e. | est±s.e. | est±s.e. |
| Genotype | | | | |
| DxD | ref | ref | ref | ref |
| Dx(DxR) | -0.55±0.16 | -0.56±0.16 | -0.59±0.16 | -0.60±0.16 |
| DxR | -0.78±0.27 | -0.77±0.27 | -0.83±0.26 | -0.83±0.26 |
| RxD | -0.71±0.23 | -0.71±0.23 | -0.75±0.23 | -0.76±0.23 |
| Rx(RxD) | -1.00±0.19 | -0.99±0.19 | -1.12±0.18 | -1.11±0.18 |
| RxR | -1.49±0.26 | -1.49±0.26 | -1.77±0.26 | -1.76±0.26 |
| Year of birth | | | | |
| 1991 | ref | ref | ref | ref |
| 1992 | -0.02±0.22 | -0.01±0.22 | -0.26±0.22 | -0.23±0.22 |
| 1993 | -1.21±0.29 | -1.19±0.29 | -1.23±0.29 | -1.18±0.29 |
| 1994 | 0.47±0.24 | 0.50±0.24 | 0.01±0.23 | 0.08±0.23 |
| 1995 | 0.34±0.24 | 0.39±0.24 | 0.06±0.23 | 0.13±0.23 |
| 1996 | 0.67±0.24 | 0.73±0.25 | -0.04±0.24 | 0.07±0.25 |
| Gender | ref | ref | ref | ref |
| Females | | | | |
| Males | -0.17±0.12 | -0.17±0.12 | -0.26±0.12 | -0.27±0.12 |
| Age of dam | | | | |
| <=2yrs | ref | ref | ref | ref |
| =3 yrs | 0.19±0.24 | 0.20±0.24 | 0.36±0.24 | 0.39±0.24 |
| =4 yrs | 0.14±0.25 | 0.15±0.25 | 0.42±0.25 | 0.46±0.25 |
| =5 yrs | 0.33±0.25 | 0.34±0.25 | 0.55±0.24 | 0.59±0.25 |
| >=6yrs | 0.34±0.25 | 0.35±0.25 | 0.46±0.25 | 0.50±0.25 |

see Table 5.3

Table 5.8: *Parameter estimates for the baseline covariates in the post-weaning period from a shared frailty model with weight at weaning and also time-varying PCV and LFEC.*

| | Weight at weaning | | | |
|---|---|---|---|---|
| | only | with PCV | with LFEC | with PCV & LFEC |
| | est±s.e. | est±s.e. | est±s.e. | est±s.e. |
| Genotype | | | | |
| DxD | ref | ref | ref | ref |
| Dx(DxR) | -0.52±0.13 | -0.52±0.14 | -0.43±0.14 | -0.47±0.15 |
| DxR | -0.88±0.20 | -0.82±0.20 | -0.76±0.21 | -0.75±0.21 |
| RxD | -1.19±0.20 | -0.99±0.20 | -1.22±0.22 | -1.06±0.22 |
| Rx(RxD) | -1.69±0.19 | -1.35±0.18 | -1.58±0.19 | -1.32±0.19 |
| RxR | -1.68±0.22 | -1.20±0.22 | -1.44±0.22 | -1.05±0.22 |
| Year of birth | | | | |
| 1991 | ref | ref | ref | ref |
| 1992 | -0.34±0.23 | -0.59±0.24 | -0.30±0.24 | -0.46±0.25 |
| 1993 | 1.78±0.19 | 1.59±0.18 | 1.75±0.20 | 1.60±0.20 |
| 1994 | 0.95±0.23 | -0.23±0.24 | 0.74±0.24 | -0.17±0.26 |
| 1995 | 0.74±0.23 | -0.46±0.25 | 0.34±0.26 | -0.47±0.27 |
| 1996 | 0.75±0.28 | 0.61±0.28 | 0.63±0.29 | 0.55±0.29 |
| Gender | | | | |
| Females | ref | ref | ref | ref |
| Males | 0.43±0.10 | 0.34±0.10 | 0.37±0.10 | 0.29±0.11 |
| Age of dam | | | | |
| <=2yrs | ref | ref | ref | ref |
| =3 yrs | -0.07±0.17 | -0.20±0.17 | -0.29±0.18 | -0.35±0.19 |
| =4 yrs | 0.01±0.18 | -0.18±0.19 | -0.21±0.19 | -0.26±0.20 |
| =5 yrs | 0.19±0.18 | 0.09±0.18 | 0.04±0.19 | 0.02±0.19 |
| >=6yrs | -0.15±0.19 | -0.27±0.20 | -0.31±0.20 | -0.33±0.21 |

see Table 5.4 first two columns

Table 5.9: *Parameter estimates for the baseline covariates in the post-weaning period from a shared frailty model with time-varying body weight, PCV and LFEC.*

| | Time varying body weight | | | |
|---|---|---|---|---|
| | only | with PCV | with LFEC | with PCV & LFEC |
| | est±s.e. | est±s.e. | est±s.e. | est±s.e. |
| Genotype | | | | |
| DxD | ref | ref | ref | Ref |
| Dx(DxR) | -0.54±0.14 | -0.55±0.14 | -0.45±0.14 | -0.50±0.15 |
| DxR | -0.91±0.19 | -0.85±0.20 | -0.78±0.21 | -0.78±0.21 |
| RxD | -1.22±0.21 | -1.05±0.20 | -1.22±0.22 | -1.11±0.22 |
| Rx(RxD) | -1.67±0.19 | -1.39±0.18 | -1.57±0.19 | -1.37±0.19 |
| RxR | -1.73±0.22 | -1.29±0.22 | -1.51±0.22 | -1.16±0.22 |
| Year of birth | | | | |
| 1991 | ref | ref | ref | Ref |
| 1992 | -0.40±0.24 | -0.60±0.24 | -0.38±0.25 | -0.51±0.25 |
| 1993 | 1.58±0.19 | 1.46±0.19 | 1.58±0.21 | 1.47±0.20 |
| 1994 | 0.89±0.23 | -0.17±0.24 | 0.68±0.24 | -0.16±0.26 |
| 1995 | 0.71±0.23 | -0.40±0.25 | 0.30±0.26 | -0.46±0.27 |
| 1996 | 0.42±0.28 | 0.40±0.28 | 0.37±0.29 | 0.34±0.29 |
| Gender | | | | |
| Females | ref | ref | ref | ref |
| Males | 0.52±0.10 | 0.39±0.10 | 0.46±0.11 | 0.34±0.11 |
| Age of dam | | | | |
| <=2yrs | ref | ref | ref | ref |
| =3 yrs | -0.03±0.17 | -0.20±0.17 | -0.27±0.19 | -0.35±0.19 |
| =4 yrs | 0.15±0.18 | -0.13±0.19 | -0.07±0.20 | -0.21±0.20 |
| =5 yrs | 0.33±0.18 | 0.11±0.18 | 0.20±0.19 | 0.07±0.19 |
| >=6yrs | 0.00±0.19 | -0.20±0.20 | -0.14±0.20 | -0.25±0.21 |

see Table 5.4 last two columns

# Chapter 6

# Joint modelling of repeated measurements and event time

## 6.1 Introduction

In the Diani data set introduced in Chapter 4 we saw that the lambs were followed up until they were approximately one year of age and that the traits BWT, PCV and FEC were periodically recorded over this period of time. In this study, as in many longitudinal studies where individuals are followed over time, the data for each individual can be grouped into three categories: (1) the elapsed time to an event ($T_i$) such as death; (2) repeated measurements ($\boldsymbol{Y}_i$) of a time-dependent variable e.g. (PCV, FEC, BWT); (3) additional covariates ($\boldsymbol{X}_i$) that may affect both the repeated measurement and the time-to-event processes. This covariate information may be available at the baseline (e.g. genotype, gender, year of birth, animal sire) or can vary with time (e.g. rainfall amount). When modelling of the repeated measurements is of interest, one may focus, e.g. on how the measurements change with time; on how the parameter estimates are influenced by drop-out of individuals during the course of the study; or on how the measurements may be affected by the additional covariates. From a time-to-event process point of view, the interest may focus on how the time to the event is affected by both the repeated measurement process and the additional covariates as in Chapter 5. A vast amount of literature exists on the methods suitable for either approach. Repeated measurements are com-

monly analysed using linear mixed effects models (Laird and Ware, 1982). These models are attractive for several reasons, one of them being the ability to easily accommodate unbalanced designs, especially regarding the timing and frequency of observations. The models also allow for an explicit partitioning of variability and estimation of fixed effects. In particular, at least two sources of variability are readily identified: between- and within-individual variation. The between-individual variability is often modelled by a vector of correlated, individual random effects.

To analyse event-time data the Cox PH model is often the method of choice. In particular if time-dependent covariates are considered, the corresponding partial likelihood function for the Cox model assuming $r$ distinct ordered failure times $t_{(1)} < ..... < t_{(r)}$ is given by

$$PL(\boldsymbol{\beta}) = \prod_{i=1}^{r} \frac{\exp(\boldsymbol{x}_{(i)}^T \boldsymbol{\beta} + \varphi Y_{(i)}(t_{(i)}))}{\sum_{j \epsilon R_i} \exp(\boldsymbol{x}_j^T \boldsymbol{\beta} + \varphi Y_j(t_{(i)}))} \tag{6.1}$$

where $R_i$ is the risk set of all subjects alive prior to the $i^{\text{th}}$ failure and $Y_j(t_{(i)})$ is the observed value of the time-dependent covariate for the $j^{\text{th}}$ individual at the time of the $i^{\text{th}}$ failure. In addition $\boldsymbol{x}_{(i)}$ and $Y_{(i)}(t_{(i)})$ are respectively the covariate vector and time-dependent covariate value for the individual whose failure time is $t_{(i)}$. Further $\boldsymbol{\beta}$ is the vector of the parameters associated with the fixed effects while $\varphi$ assesses the effect of the current observed value of the repeated measurement on the risk of the event. From the above partial likelihood, it is observed that this model requires the knowledge of the repeated measurements for all subjects in the risk set at the time of each failure, which does not occur practically. In most longitudinal studies, the subjects fail on a continuous basis while the repeated measurements are recorded at only discrete time points. In the Diani data set, measurements were only recorded as described in Section 4.3. Thus, no measurement of the covariate exists for the members in the risk set when failure occurs in between these time points (say between two months and time of weaning). There are a number of approaches to handle this problem. Often, what is done is to use the nearest preceding value of the covariate and treat it as the observed value at the time of failure, resulting in a piece-wise constant covariate process ( as in Chapter 5). That is, the last value is carried forward (LVCF). One other additional problem associated with using the covariate process in the Cox PH model is the presence of measurement error. This measurement error can be thought of consisting both of laboratory error and short-term biological variability. Failure to account for this error and for any missing observation has been shown to cause the estimated regression parameters in the Cox PH model to be

biased to the null with a bias magnitude that is proportional to the size of measurement error (Prentice 1982, Dafni and Tsiatis, 1998).

In the last ten years many methods, which simultaneously use the information available in both the time-to-event and the repeated measurement processes, have been proposed in medical research. In particular, several models have been developed in the area of acquired immunodeficiency syndrome (AIDS) research (De Grutolla and Tu, 1994, Tsiatis et al., 1995, Faucett and Thomas, 1996, Wulfsohn and Tsiatis, 1997, Wang and Taylor, 2001) and in schizophrenia studies (Henderson et al., 2000, Xu and Zeger, 2001). A detailed review of research work in joint modelling of times to an event and repeated measurements is given in Wood (2002). Several advantages of joint modelling of the repeated measurement and the time-to-event processes have been highlighted in the literature: (1) the repeated measurements can be extrapolated from the observed measurement times to the specific event time in a way that utilises the entire measurement history; (2) the time-to-event is allowed to depend on the 'true' but unknown value of the repeated measurement, thus making an adjustment for the measurement error, which in turn leads to reduced bias of the parameter estimates of the Cox model; and (3) the repeated measurement process is adjusted for any loss of information arising from death or loss of individuals.

In this chapter we use the joint modelling approach to model the time to death of the lambs and the repeated measurements of PCV, BWT and FEC. To this aim, the methodology proposed by Henderson et al. (2000) is used. The need for joint models to model survival and performance traits in animals studies is discussed in Ducrocq (1999b). In Section 6.2 we give a brief background on the methodology of linear mixed effects model as well as on the joint model of Henderson et al. (2000). In Section 6.3 we adopt the joint model to the analysis of the Diani data set and the results are presented in Section 6.4. The concluding remarks are given in Section 6.5.

## 6.2 Model formulation

Most joint modelling approaches in the literature are formulated using standard methods that are used in modelling time-to-event data and repeated measurements separately. In Chapter 2 we looked at some of the methods that are used for time-to-event data. In this section we first discuss briefly one of the standard methods used in modelling

repeated measurements. Then we summarise briefly how the time-to-event and repeated measurements models are linked into a joint model.

### 6.2.1 Linear mixed effects models for repeated measures

Data sets resulting from follow-up studies are often highly unbalanced, with subjects having unequal number of measurements. Moreover, the data have complex correlation structure due to repeated measurements for each individual. As a result such data are not ideally suited to analysis by classical least squares techniques and linear mixed effects models (Laird and Ware, 1982) are now standard tools for analysing such complex hierarchical data. Let $Y_{ij}$ denote the observed $j^{\text{th}}$ measurement for the $i^{\text{th}}$ individual recorded at time $t_{ij}$ $(i = 1, \ldots, N,\ j = 1, \ldots, n_i)$ and let $\boldsymbol{Y_i}^T = (Y_{i1}, Y_{i2}, \ldots, Y_{in_i})$. Then a linear mixed effects model is written as

$$\boldsymbol{Y_i} = \boldsymbol{X}_{1i}\boldsymbol{\beta}_1 + \boldsymbol{Z}_i\boldsymbol{b_i} + \boldsymbol{\varepsilon_i} \tag{6.2}$$

where $\boldsymbol{X}_{1i}$ and $\boldsymbol{Z}_i$ are $n_i \times p$ and $n_i \times q$ design matrices, $\boldsymbol{\beta}_1$ is a $p_1 \times 1$ vector containing the fixed effects, and $\boldsymbol{b_i}$ is a $q \times 1$ vector of the random effects. It is assumed that the vector of random effects $\boldsymbol{b_i}$ is $N(\boldsymbol{0}, \boldsymbol{D})$, i.e., it is normally distributed with mean zero and variance-covariance matrix $\boldsymbol{D} = (d_{kl})$, where $d_{kl} = \text{Cov}(b_{ik}, b_{il})$. Furthermore, it is assumed that $\boldsymbol{b_i}$ is independent from the vector of residual random errors $\boldsymbol{\varepsilon_i}$. The residual errors are assumed to be $N(\boldsymbol{0}, \boldsymbol{\Sigma_i})$, with variance-covariance matrix $\boldsymbol{\Sigma_i}$ depending on $i$ only via its size $(n_i \times n_i)$. It then follows that, marginally, $\boldsymbol{Y_i}$ is normally distributed with mean $\boldsymbol{X}_{1i}\boldsymbol{\beta}_1$ and variance-covariance matrix $\boldsymbol{V}_i = \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i^T + \boldsymbol{\varepsilon_i}$. In model (6.2), $\boldsymbol{\varepsilon_i}$ captures the within-individual variability, while the between-individual variability is modelled through the random effects $\boldsymbol{b_i}$. In particular if

$$\boldsymbol{Z}_i^T = \begin{pmatrix} 1 & 1 & \ldots & 1 \\ t_{i1} & t_{i2} & \ldots & t_{in_i} \end{pmatrix}$$

then model (6.2) is known as a random *intercepts* and *slopes* model (see Verbeke and Molenberghs, 2000, p. 25). The underlying assumption of this model is that the measurements increase linearly in time, but for each individual the linear trend has its own intercept $(b_{i1})$ and slope $(b_{i2})$. Further, if $\text{Var}(\varepsilon_{ij}) = \sigma_e^2$ and $\text{Cov}(\varepsilon_{ik}, \varepsilon_{il}) = 0$, then the

assumed covariance function of the response for this model is

$$Cov(Y_{ik}, Y_{il}) = \begin{cases} d_{11} + d_{22}t_{ik}t_{il} + d_{12}(t_{ik} + t_{il}) + \sigma_e^2 & \text{if } k = l \\ d_{11} + d_{22}t_{ik}t_{il} + d_{12}(t_{ik} + t_{il}) & \text{if } k \neq l, \end{cases} \tag{6.3}$$

which is quadratic over time. That is, the covariance function between any two measurements from the same individual is a quadratic function.

Model (6.2) has been used extensively to analyse repeated measurements arising from animal breeding programs (Foulley and Quaas, 1995, Jamrozik and Schaeffer, 1997, Meyer, 1992, 1999). In these applications, more emphasis has been placed on the covariance structure of the random effects $(\boldsymbol{b_i}, \boldsymbol{\varepsilon_i})$, in order to capture different sources of variability, such as those due to maternal, paternal and environmental effects. To estimate the parameters of model (6.2), various approaches can be applied. The most commonly used is the classical method of maximum likelihood (ML), which results in generalised least square (GLS) estimates for $\boldsymbol{\beta}$. This method of estimation however leads to underestimation of the variance parameters involved in $\boldsymbol{D}$ and $\boldsymbol{\Sigma_i}$. As an alternative, the restricted maximum likelihood estimation (REML) can be used, which remedies this problem.

### 6.2.2 A joint model for repeated measurements and time-to-event

Let $\boldsymbol{Y_i}^T = (Y_{i1}, Y_{i2}, ..., Y_{in_i})$ be the vector of the repeated measurements for the $i^{\text{th}}$ individual measured at times $\boldsymbol{t}_i^T = (t_{i1}, \ldots, t_{in_i})$. Let $T_i^o = \min(T_i, C_i)$ and $\delta_i$ denote, respectively, the time-to-event and the censoring indicator for the $i^{\text{th}}$ individual. The observed data available for the $i^{\text{th}}$ individual are thus $(T_i^o, \delta_i, \boldsymbol{Y_i}, \boldsymbol{t}_i, \boldsymbol{X}_{1i}, \boldsymbol{x}_{2i})$, where $\boldsymbol{X}_{1i}$ denotes the matrix of the observed values of covariates believed to influence the repeated measurements $\boldsymbol{Y_i}$, while $\boldsymbol{x}_{2i}$ is the incidence vector of the covariates believed to affect the time-to-event process.

Further assume that for the time-to-event the model of choice is the Cox PH model

$$\lambda_i(t) = \lambda_0(t) \exp(\boldsymbol{x}_{2i}^T \boldsymbol{\beta}_2) \tag{6.4}$$

where $\lambda_0(t)$ is the baseline hazard function common to all individuals and $\boldsymbol{\beta}_2$ is the vector of unknown parameters associated with the incidence vector $\boldsymbol{x}_{2i}$. Henderson et al. (2000) have proposed a model for the joint analysis of both the time-to-event and repeated measurements. They postulate a latent (unobserved) bivariate Gaussian process

$W_i(t) = \{W_{1i}(t), W_{2i}(t)\}$ such that the repeated measurements model depends on $W_{1i}(t)$ and the event time model on $W_{2i}(t)$. In particular for the repeated measurements process, consider a model of the general form

$$\boldsymbol{Y_i} = \mu_i(\boldsymbol{t}_i) + W_{1i}(\boldsymbol{t}_i) + \boldsymbol{\varepsilon_i} \tag{6.5}$$

where $\boldsymbol{\varepsilon_i}$ is a $N(\mathbf{0}, \boldsymbol{\Sigma_i})$ error vector such that $\boldsymbol{\Sigma_i}$ is a diagonal matrix and $\text{Var}(\epsilon_{ij})=\sigma_e^2$. Further, $\mu_i(\boldsymbol{t}_i)$ is the systematic component, which can be described by a linear model. As a basic example for the latent process $W_{1i}(t)$, Henderson et al. (2000) consider $W_{1i}(t) = U_{1i} + U_{2i}t$, where $(U_{1i}, U_{2i})$ is a bivariate normal random vector with zero mean and variance-covariance

$$\boldsymbol{D}_1 = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}.$$

One can observe that $\mu_i(\boldsymbol{t}_i)$ in model (6.5) corresponds to $\boldsymbol{X}_{1i}\boldsymbol{\beta}$ in model (6.2) while $W_{1i}(\boldsymbol{t}_i)$ corresponds to $\boldsymbol{Z}_i\boldsymbol{b_i}$, with $\boldsymbol{b_i} \equiv (U_{1i}, U_{2i})^T$.
On the other hand, the time-to-event is modelled through a Cox proportional hazards model

$$\lambda_i(t) = \lambda_0(t) \exp\left\{\boldsymbol{x}_{2i}^T\boldsymbol{\beta}_2 + W_{2i}(t)\right\}. \tag{6.6}$$

It is assumed that the repeated measurement and time-to-event processes are conditionally independent given $W_i(t)$. However, in order to induce association between the two processes, $W_{2i}(t)$ is related to particular components of $W_{1i}(t)$. This is achieved via the general equation

$$W_{2i}(t) = \varphi_1 U_{1i} + \varphi_2 U_{2i} + \varphi_3 W_{1i}(t). \tag{6.7}$$

For example, a joint model with $W_{2i}(t) = \varphi_1 U_{1i} + \varphi_2 U_{2i}$, would allow both the random intercept and slope to affect the risk of the event. In general, exponentiating the estimates of $\varphi_1$, $\varphi_2$ and $\varphi_3$ gives respective hazard ratios of death associated with the random intercept, random slope and the current predicted value of the repeated measurement, respectively.
The parameters of the models for the repeated measurement process and the time-to-event process are then estimated jointly by maximising the observed joint likelihood of the data, as detailed in Wood (2002) and Wulfsohn and Tsiatis (1997).

## 6.3 Application

We now describe the application of the joint model described in the previous section to the Diani data set introduced in Chapter 4. Separate analyses of the repeated measurements of PCV, LFEC and BWT were performed. Survival times of lambs that survived beyond one year, or those of lambs that were stolen, were censored at one year and at the last recorded observation, respectively.

In Chapter 5 we reported an average lamb mortality of 19% in the pre-weaning period and 31% in the post-weaning period. The age at death during the post-weaning period ranged from 3 to 12 months (median 6.4 months). The number of repeated measurements recorded from weaning ranged from 1 to 8 (median 6) per lamb with 1994 having the most post-weaning measurements. Figures 4.1 to 4.3 show scatter plots of the measurements recorded from one to 12 months for PCV and LFEC and those from birth to 12 months for BWT, across the six years. In each plot individual profiles for a randomly selected sample of 15 lambs are highlighted. Although all animals were weighed and sampled on the same day, ages varied as a result of lambs being born within a period of about 20-40 days.

In the joint models with either PCV, BWT or LFEC as repeated measurements, fixed effects of genotype (6 levels), year of birth (6 levels) and sex (2 levels) were included in the repeated measurements component of the joint model. Each of the traits was assumed to be curvilinear over time (see Figures 4.1 to 4.3). In the joint model with PCV, age of dam (5 levels) was considered as a baseline covariate for the time-to-event component only, but not for the repeated measurements, where it was found not to be significant.

Consequently in such a case, we can define $\boldsymbol{\beta}_1^T = [\mu, \alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \xi_1, \xi_2, \xi_3, \xi_4, \xi_5, \xi_1]$, where $\alpha_m$ $(m = 1, \ldots, 5)$ are the binary indicators capturing the breed effects, $\xi_k$ $(k = 1, \ldots, 5)$ are the indicators for year of birth and $\varpi_1$ is the binary indicator for males. As a result, the repeated measurements model can be written as

$$\boldsymbol{Y_i} = \boldsymbol{X}_{1i}\boldsymbol{\beta}_1 + \eta_1 \boldsymbol{t}_i + \eta_2 \boldsymbol{t}_i^* + W_{1i}(\boldsymbol{t}_i) + \boldsymbol{\varepsilon_i} \tag{6.8}$$

where $\boldsymbol{X}_{1i}$ is the $n_i \times 12$ design matrix corresponding to $\boldsymbol{\beta}_1$, $(\eta_1, \eta_2)$ are the parameters associated with the time trend and $\boldsymbol{t}_i^{*T} = (t_{i1}^2, \ldots, t_{in_i}^2)$ is the vector of the quadratic times. Let $\boldsymbol{\beta}^T = \big(\boldsymbol{\beta}_1^T, \eta_1, \eta_2\big)$, and $\boldsymbol{X}_{1i}(\boldsymbol{t}_i) = (\boldsymbol{X}_{1i} \mid \boldsymbol{t}_i \mid \boldsymbol{t}_i^*)$ be the $n_i \times 14$ design matrix

corresponding to $\boldsymbol{\beta}$. Model (6.8) can then be re-written as

$$\boldsymbol{Y_i} = \boldsymbol{X}_{1i}(\boldsymbol{t}_i)\boldsymbol{\beta} + W_{1i}(\boldsymbol{t}_i) + \boldsymbol{\varepsilon_i}. \tag{6.9}$$

To specify the survival component of the joint model, let $\boldsymbol{\beta}_2^T = [\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \xi_1, \xi_2, \xi_3, \xi_4, \xi_5, \varpi_1, a_1, a_2, a_3, a_4]$, where $a_l$ $(l = 1, \ldots, 4)$ are the binary indicators coding the dam age groups (with levels $\leq 2$ years, 3, 4, 5 and $\geq 6$ years). The model for survival time is then given by

$$\lambda_i(t) = \lambda_0(t) \exp\left\{\boldsymbol{x}_{2i}^T\boldsymbol{\beta}_2 + W_{2i}(t)\right\}, \tag{6.10}$$

where $\boldsymbol{x}_{2i}$ is the incidence vector of size 15 for the $i^{\text{th}}$ individual associated with the parameter vector $\boldsymbol{\beta}_2$.

For all the three traits the following settings for $W_{1i}$ and $W_{2i}$ were considered:

(S1) $W_{1i}(t) = U_{1i}$, $W_{2i} = 0$;

(S2) $W_{1i}(t) = U_{1i} + U_{2i}t$, $W_{2i} = 0$;

(S3) $W_{1i}(t) = U_{1i}$, $W_{2i} = \varphi W_{1i}$;

(S4) $W_{1i}(t) = U_{1i} + U_{2i}t$, $W_{2i} = \varphi_1 U_{1i} + \varphi_2 U_{2i} + \varphi_3 W_{1i}(t)$.

Settings (S1) and (S2) assume independence between the repeated measurement and survival processes. Settings (S3) and (S4) correspond to (S1) and (S2), respectively, with respect to the structure of $W_{1i}(t)$, but allow for dependence between the processes (joint models).

To obtain parameter estimates for the fixed effects, variance components and the association parameters of the joint models (S3) and (S4) specified above, a program in SAS (Renard et al., 2002) was written. Estimates from either setting (S1) and (S2) were computed using PROC MIXED (for repeated measurements) and PROC PHREG (for survival time) in SAS.

Estimates of the standard errors for all parameter estimates in the joint models were obtained by using the jackknife method. This was achieved by leaving out the observations for lambs from the same sire and then re-fitting the model to the remaining observations. Classically, jackknife estimation method provides reliable estimates of the standard errors if the observations omitted are independent from those that are left in. When observations for lambs from the same sire are left out, so is the genetic component of these lambs. This genetic component is assumed here to be stronger in the contribution to the lamb characteristics (e.g. survival, BWT) than the environmental components, which are shared by

lambs born in the same year. This is supported by the findings in Baker et al. (2003). These authors show that for the analysed data set, the differences observed in heritability estimates of PCV and FEC for Dorper compared to Red Masaai-sired lambs were due to the differences in genetic variance rather than the environmental variability.

## 6.4 Results

Below we report the results of the fitted models for PCV, BWT and FEC for the above settings. The results of the joint model (settings (S2) and (S4)) will be compared to those of the corresponding independent model ((settings (S1) and (S3)) for both the repeated measurements and survival estimates. The findings were as follows.

### 6.4.1 Packed cell volume from one month

Initially, we considered the repeated measurements for PCV collected from one month to one year of age. The parameter estimates for these measurements for settings (S1) and (S3) are given in Table 6.1. When fitting the model corresponding to setting (S2), a non-positive definite estimate of the variance-covariance matrix $\boldsymbol{D}_1$ (see Section 6.2.2) was obtained. On further investigation, it was discovered that the PCV repeated measurements were negatively correlated, with the correlation increasing in the absolute value over time. This negative serial correlation cannot be captured by a model with a random intercept and random slope, as specified under setting (S2). Therefore, for the PCV measurements collected from one month to one year of age, the models for settings (S2) and (S4) could not be fitted.

*Setting (S1)*: Under this setting, which assumes independence between PCV measurements and survival time, the Dorper (DxD) breed had the lowest mean PCV from one month to one year of age, which was between 0.1 to 1.9% units lower than for other genotypes (see the 'Repeated measurements model - S1 column in Table 6.1). This difference increased as the Red Maasai genotype in the lambs increased, with the Red Maasai having the highest mean PCV. The linear and quadratic time effects were both significant($P<0.001$) implying an average non-linear trend in PCV. The trend is as indicated in Figure 4.1, which shows a general sharp decline in PCV after one month in all

Table 6.1: *Estimates from independent (S1) and joint models (S3) for repeated measurements of PCV(%) from one month to 12 months and survival.*

| | Repeated measurements model | | | | Survival model | | | |
|---|---|---|---|---|---|---|---|---|
| | S1 | | S3 | | S1 | | S3 | |
| | est | s.e. | est | s.e. | est | s.e. | est | s.e. |
| Fixed effects | | | | | | | | |
| Intercept | 31.730 | 0.238 | 31.457 | 0.394 | - | - | - | - |
| time(months) | -2.856 | 0.052 | -2.866 | 0.119 | - | - | - | - |
| time*time | 0.217 | 0.005 | 0.217 | 0.011 | - | - | - | - |
| Genotype | | | | | | | | |
| DxD | ref | | ref | | ref | | ref | |
| Dx(DxR) | 0.132 | 0.231 | 0.237 | 0.281 | -0.476 | 0.121 | -0.514 | 0.136 |
| DxR | 0.355 | 0.332 | 0.449 | 0.392 | -0.569 | 0.172 | -0.583 | 0.209 |
| RxD | 1.406 | 0.260 | 1.603 | 0.381 | -0.872 | 0.169 | -0.945 | 0.197 |
| Rx(RxD) | 1.668 | 0.224 | 1.935 | 0.313 | -1.341 | 0.141 | -1.390 | 0.184 |
| RxR | 1.937 | 0.279 | 2.230 | 0.356 | -1.342 | 0.177 | -1.407 | 0.226 |
| Year of birth | | | | | | | | |
| 1991 | ref | | ref | | ref | | ref | |
| 1992 | -0.412 | 0.222 | -0.373 | 0.303 | -0.023 | 0.185 | -0.048 | 0.209 |
| 1993 | -0.340 | 0.219 | -0.470 | 0.248 | 0.921 | 0.149 | 0.948 | 0.157 |
| 1994 | -2.867 | 0.252 | -3.098 | 0.270 | 1.358 | 0.167 | 1.361 | 0.185 |
| 1995 | -3.178 | 0.251 | -3.384 | 0.291 | 1.328 | 0.165 | 1.361 | 0.187 |
| 1996 | 1.097 | 0.257 | 1.008 | 0.306 | 0.724 | 0.197 | 0.769 | 0.248 |
| Gender | | | | | | | | |
| Females | ref | | ref | | ref | | ref | |
| Males | -0.512 | 0.138 | -0.558 | 0.156 | 0.220 | 0.088 | 0.232 | 0.115 |
| Age of dam | | | | | | | | |
| <=2yrs | - | - | - | - | ref | | ref | |
| =3 yrs | - | - | - | - | -0.277 | 0.158 | -0.271 | 0.160 |
| =4 yrs | - | - | - | - | -0.544 | 0.167 | -0.505 | 0.181 |
| =5 yrs | - | - | - | - | -0.368 | 0.159 | -0.278 | 0.182 |
| >=6yrs | - | - | - | - | -0.509 | 0.168 | -0.461 | 0.150 |
| Variances | | | | | | | | |
| $\sigma_e^2$ | 25.817 | 0.398 | 25.905 | 0.839 | - | - | - | - |
| $\sigma_1^2$ | 3.037 | 0.283 | 2.804 | 0.327 | - | - | - | - |
| Association | | | | | | | | |
| $\varphi$ | - | - | - | - | 0.0 | | -0.236 | 0.020 |

years except 1996 followed by a slight rise. The lambs born in 1992-1995 had on average a lower PCV (0.3 to 3.2%) than those born in 1991. The mean PCV was the highest in 1996. On average, male lambs had lower PCV than female lambs.

By exponentiating the estimates given in the 'Survival model - S1 column in Table 6.1,

one can see that, as compared to the Dorper, the relative mortality hazard of the other genotypes ranged from $\exp(-0.476) = 0.62$ to $\exp(-1.34) = 0.26$. The Rx(RxD) and RxR breeds had the lowest, and similar, mortality. The hazard of the lambs born in the years 1993–1996 was statistically significantly higher than that of the lambs born in 1991 and ranged from 2.2 to 4.0. Male lambs had a higher mortality hazard than females while the hazard ratio decreased with increasing age of dam.

*Setting (S3)*: This setting corresponds to (S1), but assumes dependence between PCV measurements and survival time. As compared to (S1), the differences in the mean PCV, as compared to the Dorper breed, increased slightly for all other genotypes. For instance, the estimated mean PCV from weaning for the non-Dorper genotypes was 0.6-3.4% units higher than for the Dorper breed (see the 'Repeated measurements model - S3 column in Table 6.1). This might be the result of the adjustment of the analysis of the repeated measurements for the variation in death rates. The estimated time trend parameters for the repeated measurements model for (S1) setting were similar to those obtained for (S3). Relative to the mortality hazard for the Dorper breed, the hazard ratio for the non-Dorper genotypes now ranged between 0.60 to 0.24, as compared to (S1) setting (see the 'Survival model - S3 column in Table 6.1). Significant negative estimates ($P<0.001$) were obtained for the association parameters ($\varphi$ in Tables 6.1) for the survival model under (S3). This indicates that the mortality hazard decreased with increasing PCV.

### 6.4.2 Packed cell volume from weaning

As the critical period for assessing genetic resistance to endoparasites in lambs is between weaning and 12 months of age, the analysis of the PCV repeated measurements for the period from weaning onwards was also considered. In this analysis, the survival time was re-defined by using weaning as the time of origin. Consequently, in this analysis only the animals alive at the time of weaning were considered. The results for settings (S1)–(S3) and (S2)–(S4) are given in Tables 6.2 and 6.3, respectively.

*Settings (S1) and (S2)*: Similar trends in the repeated measurement model, as those reported for the analysis of data from one month of age, were observed when PCV measurements were considered from weaning (see Table 6.2). However unlike from one month of age, the time trend had a more moderate negative slope estimate. This corresponds to Figure 4.1, which shows a gradual decline in PCV after weaning.

Table 6.2: *Estimates from independent (S1) and joint models (S3) for repeated measurements of PCV (%) from weaning to 12 months and survival.*

| | Repeated measurements model | | | | Survival model | | | |
|---|---|---|---|---|---|---|---|---|
| | S1 | | S3 | | S1 | | S3 | |
| | est | s.e. | est | s.e. | est | s.e. | est | s.e. |
| Fixed effects | | | | | | | | |
| Intercept | 26.179 | 0.285 | 25.603 | 0.451 | - | - | - | - |
| time(months) | -1.406 | 0.054 | -1.419 | 0.084 | - | - | - | - |
| time*time | 0.146 | 0.006 | 0.147 | 0.009 | - | - | - | - |
| Genotype | | | | | | | | |
| DxD | ref | - | ref | - | ref | - | ref | - |
| Dx(DxR) | 0.361 | 0.287 | 0.563 | 0.399 | -0.492 | 0.142 | -0.584 | 0.190 |
| DxR | 0.978 | 0.412 | 1.140 | 0.580 | -0.678 | 0.200 | -0.707 | 0.293 |
| RxD | 1.756 | 0.318 | 2.206 | 0.465 | -0.975 | 0.195 | -1.169 | 0.242 |
| R(RxD) | 2.405 | 0.276 | 3.045 | 0.408 | -1.641 | 0.176 | -1.839 | 0.243 |
| RxR | 2.866 | 0.344 | 3.430 | 0.454 | -1.332 | 0.201 | -1.527 | 0.242 |
| Year of birth | | | | | | | | |
| 1991 | ref | - | ref | - | ref | - | ref | - |
| 1992 | 0.884 | 0.269 | 0.947 | 0.388 | 0.037 | 0.228 | -0.077 | 0.278 |
| 1993 | -1.160 | 0.263 | -1.678 | 0.345 | 1.468 | 0.174 | 1.637 | 0.209 |
| 1994 | -5.170 | 0.313 | -5.678 | 0.466 | 1.570 | 0.203 | 1.640 | 0.252 |
| 1995 | -5.666 | 0.310 | -6.082 | 0.419 | 1.533 | 0.203 | 1.677 | 0.250 |
| 1996 | 1.974 | 0.316 | 1.901 | 0.359 | 0.531 | 0.268 | 0.458 | 0.276 |
| Gender | | | | | | | | |
| Females | ref | - | ref | - | ref | - | ref | - |
| Males | -0.477 | 0.169 | -0.616 | 0.212 | 0.300 | 0.104 | 0.342 | 0.150 |
| Age of dam | | | | | | | | |
| <=2yrs | - | - | - | - | ref | - | ref | - |
| =3 yrs | - | - | - | - | -0.377 | 0.177 | -0.385 | 0.165 |
| =4 yrs | - | - | - | - | -0.728 | 0.187 | -0.669 | 0.183 |
| =5 yrs | - | - | - | - | -0.549 | 0.178 | -0.437 | 0.203 |
| >= 6yrs | - | - | - | - | -0.810 | 0.196 | -0.698 | 0.209 |
| Variances | | | | | | | | |
| $\sigma_e^2$ | 15.315 | 0.293 | 15.390 | 0.451 | | | | |
| $\sigma_1^2$ | 6.465 | 0.427 | 6.318 | 0.542 | | | | |
| Association | | | | | | | | |
| $\varphi$ | - | - | - | - | 0.0 | - | -0.303 | 0.023 |

The relative mortality hazard in the post-weaning period exhibited similar pattern as in the analysis of data from one month of age, but now the Rx(RxD) had the lowest mortality ($\exp(-1.641) = 0.19$) when compared to the Dorper breed. The hazard of the lambs born

in the years 1993–1995 was now five times higher while it was 70% higher for lambs born in 1996 when compared to that of the lambs born in 1991.

In the repeated measurements model with both random intercepts and slopes, i.e., under (S2) setting (see the 'Repeated measurements model - S2 column in Table 6.3), similar trends for the fixed effects parameter estimates were observed as in the simpler, random-intercept-only model (see the 'Repeated measurements model - S1 column in Table 6.2). However, the ranges of the estimates were reduced. Including a random slope, accounts for any variability that may be due to the rate of change, thus leading to smaller differences in the estimates. In this model the random intercept and slope were negatively correlated ($\sigma_{12} = -1.82$). This implies that lambs with a high PCV at weaning had a more rapid decline in PCV than those with a low PCV. The estimated variance component for the random intercept ($\sigma_1^2$) was 2 times larger than that in the simpler model (Table 6.2), but due to negative correlation the total variability in the two models is similar. Thus for any two time points under setting (S2), the estimated covariance is obtained by substituting the estimates of the variance components and residual error (from Table 6.3) into equation (6.3). This result should be almost equivalent to the sum of the estimates of $\sigma_1^2$ and $\sigma_e^2$ from the simpler (S1) model.

*Settings (S3) and (S4)*: As compared to (S1) and (S2) settings, the joint models constructed under both (S3) (see the 'Repeated measurements model - S3 column in Table 6.2) and (S4) (see the 'Repeated measurements model - S4 column in Table 6.3) settings, the differences in the mean PCV, increased slightly for all other genotypes relative to the Dorper. For instance, for (S3) setting (Table 6.2), the estimated mean PCV from weaning for the non-Dorper genotypes was 0.6-3.4% units higher than for the Dorper breed. For (S4) setting (Table 6.3), the difference was between 0.3% and 3.0%. This increase might be a result of the adjustment of the analysis of the repeated measurements for the variation in death rates.

Relative to the mortality hazard for the Dorper breed, the hazard ratio for the non-Dorper genotypes now ranged from 0.56 ($= \exp(-0.584)$) to 0.21 ($= \exp(-1.569)$) for both (S3) and (S4) settings (see the 'Survival model - S3 column in Table 6.2 and the 'Survival model - S4 column in Table 6.3). These estimates are lower than the estimates obtained for the corresponding (S1) and (S2) survival models. For (S3) setting, the mortality hazard for lambs born in 1993–1995 was about five times higher as compared to 1991 (Table 6.2). For (S4), the ratio was similar to setting (S3) for 1993, while for 1994 and 1995 it was

Table 6.3: *Estimates from independent (S2) and joint models (S4) for repeated measurements of PCV(%) from weaning to 12 months and survival.*

| | Repeated measurements model | | | | Survival model | | | |
|---|---|---|---|---|---|---|---|---|
| | S2 | | S4 | | S2 | | S4 | |
| | est | s.e. | est | s.e. | est | s.e. | est | s.e. |
| Fixed effects | | | | | | | | |
| Intercept | 25.817 | 0.273 | 25.844 | 0.444 | - | - | - | - |
| time(months) | -1.465 | 0.052 | -1.562 | 0.081 | - | - | - | - |
| time*time | 0.160 | 0.006 | 0.164 | 0.009 | - | - | - | - |
| Genotype | | | | | | | | |
| DxD | ref | - | ref | - | ref | - | ref | - |
| Dx(DxR) | 0.169 | 0.267 | 0.333 | 0.354 | -0.492 | 0.142 | -0.594 | 0.200 |
| DxR | 0.726 | 0.387 | 0.842 | 0.489 | -0.678 | 0.200 | -0.634 | 0.284 |
| RxD | 1.585 | 0.290 | 1.892 | 0.451 | -0.975 | 0.195 | -1.221 | 0.277 |
| Rx( RxD) | 2.070 | 0.254 | 2.521 | 0.404 | -1.641 | 0.176 | -1.843 | 0.257 |
| RxR | 2.586 | 0.319 | 2.991 | 0.440 | -1.332 | 0.201 | -1.569 | 0.304 |
| Year of birth | | | | | | | | |
| 1991 | ref | - | ref | - | ref | - | ref | - |
| 1992 | 1.460 | 0.237 | 1.392 | 0.404 | 0.037 | 0.228 | -0.260 | 0.340 |
| 1993 | -0.779 | 0.239 | -1.208 | 0.358 | 1.468 | 0.174 | 1.678 | 0.282 |
| 1994 | -3.685 | 0.286 | -4.311 | 0.631 | 1.570 | 0.203 | 1.343 | 0.403 |
| 1995 | -3.631 | 0.290 | -4.285 | 0.661 | 1.533 | 0.203 | 1.396 | 0.471 |
| 1996 | 2.101 | 0.286 | 1.998 | 0.365 | 0.531 | 0.268 | 0.582 | 0.308 |
| Gender | | | | | | | | |
| Females | ref | - | ref | - | ref | - | ref | - |
| Males | -0.539 | 0.154 | -0.597 | 0.188 | 0.300 | 0.104 | 0.357 | 0.170 |
| Age of dam | | | | | | | | |
| <=2yrs | - | - | - | - | ref | - | ref | - |
| =3 yrs | - | - | - | - | -0.377 | 0.177 | -0.425 | 0.184 |
| =4 yrs | - | - | - | - | -0.728 | 0.187 | -0.685 | 0.205 |
| =5 yrs | - | - | - | - | -0.549 | 0.178 | -0.453 | 0.227 |
| >=6yrs | - | - | - | - | -0.810 | 0.196 | -0.676 | 0.213 |
| Variances | | | | | | | | |
| $\sigma_e^2$ | 12.901 | 0.272 | 12.641 | 0.401 | | | | |
| $\sigma_1^2$ | 15.894 | 0.990 | 15.380 | 1.517 | | | | |
| $\sigma_{12}$ | -1.817 | 0.145 | -1.590 | 0.236 | | | | |
| $\sigma_2^2$ | 0.276 | 0.024 | 0.261 | 0.028 | | | | |
| Association | | | | | | | | |
| $\varphi_1$ | - | - | - | - | 0.0 | - | -0.273 | 0.097 |
| $\varphi_2$ | - | - | - | - | 0.0 | - | -1.986 | 0.505 |
| $\varphi_3$ | - | - | - | - | 0.0 | - | -0.251 | 0.087 |

about slightly reduced. This is in agreement with Figure 4.1; the lambs born in 1994 and 1995 had much lower PCV measurements at weaning than those born in 1991. The former however increased over time while the latter decreased. Thus, adjusting for the PCV evolution over time results in a slight decrease in the mortality hazard for 1994–1995. On the other hand, the lambs born in 1993 and 1991 had almost similar PCV measurements at weaning. However the decrease in 1993 over time was much sharper (larger negative slope) than in 1991. Adjusting for this sharp decrease translates into a higher mortality hazard for 1993. Finally, as compared to (S4), much higher hazard ratios are observed for 1994 and 1995 in (S3) model, as the latter model adjusts the risk only for level of PCV over time.

Significant negative estimates (P<0.001) were obtained for all the association parameters ($\varphi$ in Table 6.2, and $\varphi_1$–$\varphi_3$ in Table 6.3) for the survival model under both (S3) and (S4) setting. This indicates that the mortality hazard decreased with increasing PCV. Thus after weaning, in (S3) setting, lambs with PCV measurements higher than the average had a lower mortality hazard than those with lower PCV measurements. The standard deviation of the distribution of the random intercepts in the repeated measurements part of the joint model for (S3) setting was estimated as 2.54 ($=\sqrt{6.465}$). Thus, the model predicts that for every (random) increase by one standard deviation in the PCV, the risk of death decreases by 0.54 ($=1-\exp(-0.303*2.54)$) (see Table 6.2). For the (S4) model a large negative estimate was obtained for $\varphi_2$, which corresponds to the random individual slope. The standard deviation of the distribution of the random slopes was estimated to equal 0.53 ($=\sqrt{0.276}$). Thus, the model indicates that, for every increase of one standard deviation in the rate of change of PCV, the mortality hazard decreases by 0.65 ($=1-\exp(-1.986*0.53)$).

### 6.4.3 Body weight

In this analysis, measurements of body weight from birth to one year of age were used. Models were fitted using all four settings. The parameter estimates of the fixed effects were almost similar for (S1)–(S3) and (S2)–(S4) settings. We thus report only the estimates of the models constructed under (S2) and (S4) setting, which are shown in Table 6.4.

*Setting (S2)*: The Dorper (DxD) breed had the highest mean BWT, which was between 0.02 to 0.65 kgs higher than for other genotypes (see the 'Repeated measurements model

- S2 column in Table 6.4). There was a non-linear trend in change of the body weight over time. On average, the lambs born in 1994–1996 were lighter than those born in 1991 and 1992. Male lambs were on average 0.1 kg heavier than the females. Notably, lambs born to older dams were much heavier in body weight (0.7 to 1.1 kg) than those to young dams.

As compared to the Dorper, the relative mortality hazard of the other genotypes ranged from 0.61 (=exp(−0.494)) to 0.27 (=exp(−1.302)), with the Red Maasai (RxR) and Rx(RxD) having the lowest hazard. An increased mortality hazard was noted for the years 1993–1996. Lambs born to ewes $\geq 3$ years of age had lower hazard than those born to younger ewes.

*Setting (S4)*: Adjusting the repeated measurement process for the variation in death rates had a only slight effect on the parameter estimates of the (S4) models when compared to (S2) setting.

The relative mortality of the genotypes now ranged from 0.61 to 0.24 with the RxR genotype having the lowest hazard mortality. This result indicates that despite being lighter in body weight when compared to the other genotypes the Red Maasai demonstrates better performance in terms survival. The age of dam effect was non-significant. This could be due to the fact that in this analysis, we account for the low body weight of lambs born to young dams, which biologically is due to low milk production of the dam in her first parity.

Negative estimates of the parameters relating the random components of the repeated measurements model to the survival model were observed (see the estimates for $\varphi_1$–$\varphi_3$ in Table 6.4). In particular, there was a significant negative association only between random growth rate ($\varphi_2$, P<0.001) and risk of death. This shows that animals who had weight profiles with increasing slope had reduced risk of death. In a reduced model with $\varphi_1$ and $\varphi_3$ constrained to zero (results not shown), the estimates obtained for the association parameter $\varphi_2$ was $\hat{\varphi}_2 = -3.262$ (s.e.=0.389). The standard deviation of the random slope in this reduced model was equal to 0.28. Thus for every increase by one standard deviation, the change in mortality hazard associated to the rate of change in BWT was reduced by 0.60 (=1-exp($-3.262 \times 0.28$)).

Table 6.4: *Estimates from independent (S2) and joint models (S4) for repeated measurements of BWT (kg) from birth to 12 months and survival.*

| | Repeated measurements model | | | | Survival model | | | |
|---|---|---|---|---|---|---|---|---|
| | S2 | | S4 | | S2 | | S4 | |
| | est | s.e. | est | s.e. | est | s.e. | est | s.e. |
| Fixed effects | | | | | | | | |
| Intercept | 3.980 | 0.128 | 4.004 | 0.146 | - | - | - | - |
| time (months) | 2.330 | 0.016 | 2.270 | 0.045 | - | - | - | - |
| time*time | -0.101 | 0.001 | -0.101 | 0.003 | - | - | - | - |
| Gentotype | | | | | | | | |
| DxD | ref | - | ref | - | ref | - | ref | - |
| Dx(DxR) | -0.019 | 0.101 | -0.023 | 0.109 | -0.494 | 0.108 | -0.499 | 0.130 |
| DxR | -0.255 | 0.150 | -0.248 | 0.159 | -0.630 | 0.157 | -0.634 | 0.165 |
| RxD | -0.158 | 0.114 | -0.168 | 0.131 | -0.809 | 0.149 | -0.910 | 0.194 |
| Rx(RxD) | -0.326 | 0.098 | -0.351 | 0.125 | -1.263 | 0.124 | -1.240 | 0.160 |
| RxR | -0.649 | 0.127 | -0.658 | 0.142 | -1.302 | 0.157 | -1.431 | 0.203 |
| Year of birth | | | | | | | | |
| 1991 | ref | - | ref | - | ref | - | ref | - |
| 1992 | -0.808 | 0.102 | -0.814 | 0.160 | 0.073 | 0.161 | 0.239 | 0.163 |
| 1993 | 0.293 | 0.100 | 0.439 | 0.138 | 0.763 | 0.136 | 0.096 | 0.219 |
| 1994 | -2.170 | 0.120 | -2.147 | 0.140 | 1.384 | 0.148 | 1.448 | 0.162 |
| 1995 | -1.499 | 0.120 | -1.423 | 0.145 | 1.246 | 0.149 | 1.066 | 0.188 |
| 1996 | -2.277 | 0.129 | -2.320 | 0.060 | 0.763 | 0.173 | 1.164 | 0.325 |
| Gender | | | | | | | | |
| Females | ref | - | ref | - | ref | - | ref | - |
| Males | 0.094 | 0.062 | 0.075 | 0.118 | 0.202 | 0.079 | 0.393 | 0.097 |
| Age of dam | | | | | | | | |
| <=2yrs | ref | - | ref | - | ref | - | ref | - |
| =3 yrs | 0.670 | 0.115 | 0.643 | 0.115 | -0.187 | 0.144 | -0.139 | 0.147 |
| =4 yrs | 1.154 | 0.118 | 1.115 | 0.116 | -0.490 | 0.152 | -0.320 | 0.162 |
| =5 yrs | 1.096 | 0.119 | 1.068 | 0.109 | -0.333 | 0.145 | -0.101 | 0.163 |
| >=6yrs | 0.956 | 0.127 | 0.916 | 0.129 | -0.433 | 0.152 | -0.246 | 0.148 |
| Variances | | | | | | | | |
| $\sigma_e^2$ | 2.536 | 0.038 | 2.521 | 0.091 | - | - | - | - |
| $\sigma_1^2$ | 0.710 | 0.058 | 0.750 | 0.052 | - | - | - | - |
| $\sigma_{12}$ | 0.167 | 0.013 | 0.208 | 0.017 | - | - | - | - |
| $\sigma_2^2$ | 0.069 | 0.004 | 0.082 | 0.007 | - | - | - | - |
| Association | | | | | | | | |
| $\varphi_1$ | - | - | - | - | 0 | - | -0.092 | 0.389 |
| $\varphi_2$ | - | - | - | - | 0 | - | -3.361 | 1.609 |
| $\varphi_3$ | - | - | - | - | 0 | - | 0.027 | 0.079 |

### 6.4.4 Faecal egg count

The repeated measurements for the log-transformed FEC (LFEC) from one month to one year of age were initially considered. As was the case with PCV, a negative correlation that increased over time was observed when (S2) setting was used. This problem was not resolved by using the simpler (S1) setting. Considering LFEC measurements from weaning onwards did not resolve the problem either as the negative correlation observed in (S2) was large among measurements collected towards the end of the one year period. The random structure ($W_{1i}$) of the repeated measurements models that are assumed under (S1) and (S2) settings in general assume a smooth trend for the repeated measurements. Any serial correlation present in these measurements may be captured using stochastic processess (e.g., the non-stationary Gaussian process, Diggle, 1988) which are used for this purpose in linear mixed models.

## 6.5 Discussion

The repeated measurements of PCV, BWT and FEC were previously analysed by Baker et al. (2003). In that analysis the survival pattern of the animals was not taken into account, and these authors chose to analyse the data for each time point separately. They concluded that the Red Maasai had higher resistance (lower FEC) and higher resilience (higher PCV) than Dorpers. In Chapter 5, on the other hand, we studied the survival of each genotype and introduced the effects of BWT, PCV and FEC as time-varying covariates in shared frailty models, with the frailty defined as a random effect of sire. Introduction of PCV and FEC as time-varying covariates in that analysis in models with BWT (time-invariant or time-varying) reduced the magnitude of the sire variance, confirming the moderate levels of heritability reported by Baker et al. (2003).

In the analysis presented in the current chapter, it has been decided to analyse the individual repeated measurements jointly with the survival process. By doing so, parameter estimates in both components of the joint model generally increased in absolute order of magnitude, as compared to the models assuming independence between the two processes. For instance, in the joint model with PCV as the repeated measurement, the relative risks of death in the different genotypes, compared with the Dorper were altered from 0.61 to

0.19 for (S1) and (S2) models to 0.56 to 0.21 for both the (S3) and (S4) models. Thus, adjustment for the evolution of PCV widened the comparative relative risk between the Dorpers and the Red Maasai. On the other hand, adjustment for the death rates, showed wider differences in the PCV repeated measurements among the genotypes. For instance, the Dorpers had on average much lower PCV than the other genotypes in the (S3) and (S4) models when compared to the corresponding (S1) and (S2) models.

In general, repeated measurements such as PCV, BWT or FEC are only recorded at specific time points. When such variables are used in a proportional hazards model as time-varying covariates, the standard method is to impute the missing observations by using the last observed value, what results in a piece-wise constant profile. This was the approach undertaken in Chapter 5. Prentice (1982) however shows that this approach leads to biased model parameters, and the presence of any measurement error in the covariate attenuates the estimates towards zero. On the other hand, in the joint analysis, the repeated measurements are rather imputed by values resulting from modelling the repeated measurement process over time. Hence, estimated 'true' values of the repeated measurements are used at each time point. This could explain the larger absolute order of magnitude observed in the parameter estimates obtained from the joint models. In addition, the effect of other characteristics of the patterns of the repeated measurements on the risk of death can be considered. The effects of these characteristics is well demonstrated in this study when repeated measurements of PCV recorded from weaning were considered. Including a random slope had a dramatic effect on the post-weaning risks of death for the years 1993 to 1995 when compared to 1991.

The type of joint models that we propose here are not exhaustive. As observed with LFEC, the random intercepts and random slopes are not able to capture the intrisinc patterns of this trait when considered from either one month or time of weaning. However the 'smooth' BWT profiles from the time of birth are adequately captured with random intercept and slope models. In general random intercepts and slopes provide a representation of the dominant part of the evolution of the profiles but do not capture the more subtle behaviour. This behaviour can be captured using, e.g., autocorrelated stochastic processes. In fact, Henderson et al. (2000) did propose the use of a non-stationary Gaussian process in their approach. Unfortunately, due to the lack of appropriate software, this solution is not yet available in practice.

In the time-to-event component of the joint model clustering nature of any measurements

can also be incorporated in the model. To this end an additional latent process (say $W_3(t)$) independent of $W_2(t)$ could be used to induce the clustering. Clusters could for instance consist of the time to death measurements of lambs from the same sire. Henderson et al. (2000) do indeed propose inclusion of a frailty term in the time-to-event component of the joint model. Implementation of such an analysis in the current study was also hampered by software limitations, which still continues to limit the use of joint models.

# Chapter 7

# Concluding remarks and further research

## 7.1 Methodology

The main theme of this thesis is the study of random effects survival models. We mainly focus on shared frailty models. In the first part of this thesis, consisting of Chapters 2 and 3, we discuss methodological issues for such models.

In Chapter 2 we present a review on likelihood estimation methods of the semi-parametric frailty model. We consider the gamma and log-normal frailty distributions. This review gives us a better understanding of the common ground for these estimation methods.

Although frailty models have become the standard approach for analysing multivariate time-to-event data in the last two decades, the asymptotic theory for these models is still not well developed. The consistency and asymptotic distribution theory of the estimators from a semi-parametric gamma-frailty model (with no covariates) are discussed in Murphy (1994, 1995) while Parner (1998) studies the model with covariates. Murphy and van der Vaart (1997) obtain the asymptotic distribution of the likelihood ratio statistic for the two-sided testing problem of no heterogeneity in the semi-parametric frailty model considered by Murphy (1994, 1995). In Chapter 3 we derive the asymptotic distribution of the likelihood ratio and score statistic for testing the one-sided problem of no heterogeneity. Testing for heterogeneity is a non-standard testing problem as the variance (heterogene-

ity) parameter is on the boundary of the parameter space under the null hypothesis. Such problems have been studied in the recent past in the area of linear mixed models (Self and Liang, 1987, Stram and Lee, 1994, 1995, Verbeke and Molenberghs, 2003) but not in the context of frailty models. In this chapter, we prove that the likelihood ratio test statistic for no heterogeneity has an asymptotic distribution which is a 50:50 mixture of a point mass at zero and a chi-square distribution with one degree of freedom. To this end, a shared gamma frailty model with a Weibull baseline hazard was used. Further, the score statistic for a complete data model (Weibull-gamma frailty) and with no covariates is shown to have the same asymptotic distribution as the likelihood ratio statistic.

There are diverse ways to extend these issues that we discuss in this part of the thesis. An interesting follow-up on the review we carry out in Chapter 2 could be a comparison through simulation of the various approaches for a specific frailty distribution. Such a simulation study could give, for each specific approach, insight on the bias and spread of the estimates. The results that we derive in Chapter 3 can however be extended in a number of ways.

Using the likelihood expressions given in Murphy (1995) and Murphy and van der Vaart (1997), Theorem 4 on censored data can be extended to more complex parametric and semi-parametric frailty models. The related study of score tests for censored data is a further interesting topic as discussed in Section 3.8.

For completeness, Wald-type test statistics for testing hypotheses (3.1) may be employed as well. Robertson, Wright and Dykstra (1988) construct a Wald statistic for the situation where the alternative hypothesis is described by inequalities. A useful further research topic is to study the distributional behaviour of the test for heterogeneity under local alternatives converging to the null hypothesis at rate $n^{-\frac{1}{2}}$. As in the two-sided testing problems, it is expected that the test statistics will have the same power characteristics under the local circumstances.

A further relevant issue is to provide information on good finite sample approximation of the mixing properties which can be used to improve the asymptotic 50:50 mixture a point mass at zero and a chi-square distribution with one degree of freedom. In semi-parametric frailty models, the situation is made more complex by the presence of nuisance parameters and/or functions under the null hypothesis and censoring. Bootstrapping the distribution of the test statistic can provide an alternative to the asymptotic distribution.

Extending the results to other frailty distributions can also be a topic for further inves-

tigation. Ferreira and Garcia (2002) show in a simulation study that the heterogeneity (variance parameter) estimator from a semi-parametric log-normal frailty model with no covariates has an asymptotic normal distribution. In view of this result and that of Murphy (1995) (gamma frailty model) we can conjecture that the likelihood ratio test statistic for no heterogeneity for a log-normal frailty model will also have an asymptotic 50:50 mixture of a point mass at zero and a chi-square distribution with one degree of freedom. The constructive proof of this result will be more complicated as the observable log-likelihood for the log-normal frailty model has no closed form as noted in Chapter 2.

## 7.2 Advanced models for analysing animal breeding data

In this second part of the thesis, we have focused on the application of advanced methods for analysing the animal breeding data. We describe the data in Chapter 4 and we also discuss the methods that have been used so far in analysing these data. In this chapter we also highlight the need and the motivation for the advanced techniques that we use in Chapters 5 and 6.

In Chapter 5 we apply the shared frailty models to assess the effects that are associated with the risk of mortality among the lambs. Semi-parametric gamma and log-normal frailty models were used and parameter estimates obtained using the penalized partial likelihood estimation method. The lambs from the same sire were assumed to constitute a cluster. Our findings concur with other previous analyses (Baker et al., 1999, 2003). We have shown that the Red Maasai perform better than the Dorper in terms of survival and had about a three quarter lower risk of mortality than the Dorper in both the pre- and post-weaning periods. Based on our findings in Chapter 3 we have assessed whether there is between cluster variation among the times-to-event of the lambs from different sires. The null hypothesis is not rejected in both the pre-weaning and post-weaning periods.

As discussed in Section 5.6 we do not calculate heritability estimates in this study. It would have been a useful addition to the application of the shared frailty model to the animal breeding data had it been feasible. As noted therein, the heritability expressions that exist in the literature have been derived from parametric frailty models and may not be appropriate when a Cox PH frailty model is used. In view of this, there is need for further research for an appropriate expression for calculating heritability when the latter

model is used.

The area of diagnostics for assessing the fit of a frailty model is another in need of further research. As seen earlier, we have fitted several models to the data but carried out no formal assessment of their adequacy. Very little has been done in this area of diagnostics. Glidden (1999) discusses diagnostics for the gamma frailty model by proposing techniques that are based on the means (2.12) defined in Chapter 2. A plot of $\hat{W}_G(t) = G^{-1/2} \sum \{\exp \hat{w}_i - 1\}$ against time is then used to assess the adequacy of the model. For the log-normal frailty distribution little or nothing has been done and certainly this is an area that is still in need of further research.

In Chapter 6 we have modelled jointly time-to-death of the lambs and repeated measurements of the traits PCV and BWT. These traits together with FEC were used as time-varying covariates in the shared frailty model analysis of Chapter 5. In the analyses carried out in these two chapters, the RxR is shown to have the lowest relative hazard when compared to the Dorper, when the weight of the lambs is taken into account (Tables 5.8, 5.9 and 6.4). This is despite the fact that the RxR is much lighter in body weight when compared to the other genotypes. This result demonstrates the better performance in terms survival for this genotype.

In Table 7.1 below we present the results from the survival component of the joint model under setting (S4) for the repeated measurements of PCV (from weaning) and BWT (from birth) together with those from a Cox PH model with a time-varying covariate for these measurements. Under the (S4) setting the individual profiles of the repeated measurements are assumed to have a linear trend. For, the time-dependent model, a linear term for the covariate is used. A direct comparison of these numerical results would be erroneous as the methodological aspect of the two models are different. The survival component of the joint model can be viewed as a conditional model (conditioned on the random effects) while a Cox PH model with a time-dependent covariate can be viewed as a population average model. Previous studies (Prentice, 1982) have shown that presence of measurement error in time-dependent covariates in a Cox PH model leads to biased parameter estimates. This drawback is however corrected within the joint model (Faucett and Thomas, 1996, Wang and Taylor, 2001). In Chapter 6 we have shown that with the joint model, the particular characteristics of the repeated measurements that are specifically associated with the risk of mortality are captured. This was well demonstrated when PCV measurements were considered from weaning.

Table 7.1: *Parameter estimates from a joint model (S4) and Cox PH model with time dependent covariate, for repeated measurements of PCV(%) from weaning and BWT(kg) from birth.*

| | Packed cell volume | | | | Body weight | | | |
|---|---|---|---|---|---|---|---|---|
| | S4 | | Time dependent | | S4 | | Time dependent | |
| | est | s.e. | est | s.e. | est | s.e. | est | s.e. |
| Genotype | | | | | | | | |
| DxD | ref | - | ref | - | ref | - | ref | - |
| Dx(DxR) | -0.594 | 0.200 | -0.406 | 0.136 | -0.499 | 0.130 | -0.512 | 0.102 |
| DxR | -0.634 | 0.284 | -0.621 | 0.194 | -0.634 | 0.165 | -0.782 | 0.153 |
| RxD | -1.221 | 0.277 | -0.690 | 0.180 | -0.910 | 0.194 | -0.948 | 0.138 |
| Rx(RxD) | -1.843 | 0.257 | -1.094 | 0.163 | -1.240 | 0.160 | -1.313 | 0.116 |
| RxR | -1.569 | 0.304 | -0.751 | 0.193 | -1.431 | 0.203 | -1.519 | 0.150 |
| Year of birth | | | | | | | | |
| 1991 | ref | - | ref | - | ref | - | ref | - |
| 1992 | -0.260 | 0.340 | -0.184 | 0.221 | 0.239 | 0.163 | 0.337 | 0.075 |
| 1993 | 1.678 | 0.282 | 1.455 | 0.164 | 0.096 | 0.219 | -0.178 | 0.148 |
| 1994 | 1.343 | 0.403 | 0.039 | 0.211 | 1.448 | 0.162 | 0.618 | 0.126 |
| 1995 | 1.396 | 0.471 | -0.050 | 0.216 | 1.066 | 0.188 | 0.707 | 0.145 |
| 1996 | 0.582 | 0.308 | 0.561 | 0.255 | 1.164 | 0.325 | 0.648 | 0.144 |
| Gender | | | | | | | | |
| Females | ref | - | ref | - | ref | - | ref | - |
| Males | 0.357 | 0.170 | 0.214 | 0.097 | 0.393 | 0.097 | 0.190 | 0.165 |
| Age of dam | | | | | | | | |
| <=2yrs | ref | - | ref | - | ref | - | ref | - |
| =3 yrs | -0.425 | 0.184 | -0.392 | 0.169 | -0.139 | 0.147 | 0.017 | 0.136 |
| =4 yrs | -0.685 | 0.205 | -0.578 | 0.177 | -0.320 | 0.162 | 0.072 | 0.144 |
| =5 yrs | -0.453 | 0.227 | -0.363 | 0.170 | -0.101 | 0.163 | 0.156 | 0.138 |
| >=6yrs | -0.676 | 0.213 | -0.703 | 0.187 | -0.246 | 0.148 | 0.012 | 0.146 |
| Association | | | | | | | | |
| $\varphi_1$ | -0.273 | 0.097 | - | - | -0.092 | 0.389 | - | - |
| $\varphi_2$ | -1.986 | 0.505 | - | - | -3.361 | 1.609 | - | - |
| $\varphi_3$ | -0.251 | 0.087 | - | - | 0.027 | 0.079 | - | - |
| Time dependent covariate | - | - | -0.190 | 0.010 | - | - | -0.307 | 0.015 |

As discussed in Section 6.5 the joint models that we propose are not exhaustive. We were unsuccessful in our attempt to model FEC jointly with survival. The random intercept and slope models that we use were unable to capture the more elaborate structure that was observed with the FEC repeated measurements. It is possible that this behaviour could be captured using, for example, autocorrelated stochastic processes. Extension of the current software to incorporate such processes that have been proposed for capturing

serial correlation in the framework of joint models (Henderson et al., 2000, Wang and Taylor, 2001) is an open area in need of further research. Within this line the more elaborate variance structures that are currently used to analyse repeated measurements for animal breeding data can also be incorporated.

In this thesis our main focus was on random effects models for survival data. Ideally we would have added a frailty term to the survival component of the joint model. This was not possible owing to software limitations.

# References


Aalen, O.O. (1988). Heterogeneity in survival analysis. *Statistics in Medicine*, **7**, 1121–1137.

Aalen, O.O. (1994). Effects of frailty in survival analysis. *Statistical Methods in Medical Research*, **3**, 227–243.

Agresti, A. (1996). *An Introduction to Categorical Data Analysis*. John Wiley.

Andersen, P.K. and Borgan, O. (1985). Counting process approach for life history data (with discussions). *Scandinavian Journal of Statistics*, **12**, 98–158.

Andersen, P.K., Klein, J.P., Knudsen, K.M. and Palacios, R.T. (1997). Estimation of variance in Cox's regression model with shared gamma frailties. *Biometrics*, **53**, 1475–1484.

Artin, M. (1991). *Algebra*. Prentice Hall.

Baker, R.L. (1998). A review of genetic resistance to gastrointestinal nematode parasites in sheep and goats in the tropics and evidence for resistance in some sheep and goat breeds in sub-humid coastal Kenya. *Animal Genetics Resources Information Bulletin*, **24**, 13–30.

Baker, R.L., Mwamachi, D.M., Audho, J.O. and Thorpe, W. (1994). Genetic resistance to gastrointestinal nematode parasites in Red Maasai sheep in Kenya. *Proceedings of the Fifth World Congress on Genetics Applied to Livestock Production, Guelph*, **20**, 277–280.

Baker R.L., Mwamachi, D.M., Audho, J.O., Aduda, E.O. and Thorpe, W. (1999). Ge-

netic resistance to gastro-intestinal nematode parasites in Red Maasai, Dorper and Red Maasai x Dorper ewes in the sub-humid tropics. *Animal Science*, **69**, 335–344.

Baker, R.L., Nagda, S., Rodriguez-Zas, S.L., Southey, B.R., Audho, J.O., Aduda, E.O. and Thorpe, W. (2003). Resistance and resilience to gastrointestinal nematode parasites and productivity of Red Maasai, Dorper and Red Maasai x Dorper crossbred lambs in the sub-humid tropics. *Animal Science*, **76**, 119–136

Barger, I.A. (1999). The role of epidemiological knowledge and grazing management for helminth control in small ruminants. *International Journal for Parasitology*, **29**, 41–47.

Bjarnason, H. and Hougaard, P. (2000). Fisher information for two gamma frailty bivariate Weibull models. *Lifetime Data Analysis*, **6**, 59–71.

Brent, R.P. (1973). *Algorithms for minimization without derivatives*. Prentice Hall.

Cheng, P.E. (1987). A nearest neighbour hazard rate estimator for randomly censored data. *Communications in Statistics-A, Theory and Methods*, **16**, 613–625.

Chernoff, H. (1954). On the distribution of the likelihood ratio. *Annals of Mathematical Statistics*, **25**, 573–578.

Claeskens, G. (2002). Restricted likelihood ratio lack of fit tests using mixed spline models. Technical Report (submitted).

Clarke, F.H. (1983). *Optimization and non smooth analysis*. John Wiley.

Cloete, S.W.P., Snyman, M.A., and Herselman, M.J. (2000). Productive performance of Dorper sheep. *Small Ruminant Research*, **36**, 119–135.

Collet, D. (1994). *Modelling survival data in medical research*. Chapman and Hall.

Colosimo, E.A. and Oliveira, M.D. (2002). On identifiability assumptions for the random effects models in survival analysis. Technical Report.

Cortinas Abrahantes, J. and Burzykowski, T. (2003). A version of EM algorithm for proportional hazards model with random effects. Technical Report.

Cox, D.R. (1972). Regression models and lifetables (with discussions). *Journal of the Royal Statistical Society B*, **39**, 1–38.

Cox, D.R and Oakes D. (1984). *Analysis of survival data*. Chapman and Hall.

Dafni, U.G. and Tsiatis, A.A. (1998). Evaluating surrogate markers of clinical outcome when measured with error. *Biometrics*, **54**, 1445–1462.

DeGrutolla, V., and Tu, X.M. (1994). Modelling progression of CD4-lymphocytes count and its relation to survival time. *Biometrics*, **50**, 1003–1014.

Diggle, P. J. (1988). An approach to the analysis of repeated measures. *Biometrics*, **44**, 659–671.

Duchateau, L., Janssen, P., Lindsey, P., Legrand, C., Nguti, R. and Sylvester, R. (2002). The shared frailty model and the power for heterogeneity tests in multicenter trials. *Computational Statistics and Data Analysis*, **40**, 603–620.

Ducrocq, V. (1999a). Two years of experience with the French genetic evaluation of dairy bulls on production-adjusted longevity of their daughters. Proceedings of the International Workshop on EU Concerted Action: Genetic Improvement of Functional Traits in Cattle (GIFT), *Interbull Bulletin*, **21**, 60–67.

Ducrocq, V. (1999b). Topics that may deserve further attention in survival analysis applied to dairy cattle breeding : some suggestions. Proceedings of the International Workshop on EU Concerted Action: Genetic Improvement of Functional Traits in Cattle (GIFT), *Interbull Bulletin*, 180–189.

Ducrocq, V. (2000). Genetic improvement of laying hens viability using survival analysis. *Genetic Selection Evolution*, **32**, 23–40.

Ducrocq, V., Quaas, R.L., Pollak, E.J., and Casella, G. (1988). Length of productive life of dairy cows II. Variance component estimation and sire evaluation. *Journal of Dairy Science*, **71**, 3071–3079.

Ducrocq, V. and Casella, G. (1996). A Bayesian analysis of mixed survival models. *Genetic Selection Evolution*, **28**, 359–362.

Elbers, C and Ridder, G. (1982). True and spurious duration dependence: the identifiability of the proportional hazard model. *Review of Economic Studies*, **49**, 402–411.

Faucett, C. and Thomas, D. (1996). Simultaneously modelling censored data and repeatedly measured covariates. *Statistics in Medicine*, **15**, 1663–1685.

Ferguson, T. (1996). *A course in large sample theory*. Chapman and Hall.

Ferreira, A. and Garcia, N.L. (2002). Simulation study for misspecification on the frailty model. *Brazilian Journal of Probability and Statistics*, **15**, 121–134.

Fleming, T.R. and Harrington, D. (1991). *Counting processes and survival analysis*. John Wiley.

Foulley, J.L. and Quaas, R.L. (1995). Heterogeneous variances in Gaussian linear mixed models. *Genetic Selection Evolution*, **27**, 211–228.

Gatenby, R.M. (1986). *Sheep production in the tropics and sub-tropics*. Tropical Agricultural Series, Longman Group Ltd.

Geyer, C.J. (1994). On the asymptotics of constrained $M$-estimation. *Annals of Statistics*, **22**, 1993–2010.

Gill, R.D. (1984). Understanding Cox's regression model: a martingale approach. *Journal of the American Statistical Association*, **79**, 441–447.

Gilmour, A.R., Cullis, B.R., Welham, S.J. and Thompson, R. (1999). *ASREML reference manual*. New South Wales Agriculture Biometrics Bulletin, **3**.

Gray, G.D., Woolaston, R.R. and Eaton, B.T. (1995). *Breeding for resistance to infectious diseases of small ruminants*. Australian Centre for International Agricultural Research Monograph No. 34.

Gray, R.J. (1992). Flexible methods for analyzing survival data using splines, with application to breast cancer prognosis. *Journal of the American Statistical Association*, **87**, 942–951.

Heckman, J.J. and Singer, B. (1984). A method for minimizing the impact of distributional

assumptions in econometric models for duration data. *Econometrica*, **52**, 271–320.

Henderson, R., Diggle, P. and Dobson, A. (2000). Joint modelling of longitudinal measurements and event time data. *Biostatistics*, **1**, 465–480.

Henryon, M., Berg, P., Jensen, J. and Andersen, S. (2001). Genetic variation for resistance to clinical and subclinical diseases exists in growing pigs. *Animal Science*, **73**, 375–387.

Honoré, B.E. (1993). Identification results for duration models with multiple spells. *Review of Economic Studies*, **60**, 241–246.

Hougaard, P. (1995). Frailty models for survival data. *Lifetime Data Analysis*, **1**, 255–273.

Hougaard, P. (2000). *Analysis of multivariate survival data.* Springer.

Huster, W.J., Brookmeyer, R. and Self, S.G. (1989). Modelling paired survival data with covariates. *Biometrics*, **45** , 145–156.

Iverson, H.K. and Randles, R.H. (1989). The effect on convergence of substituting parameter eatimates into U-statistics and other families of statistics. *Probability Theory and Related Fields*, **81**, 453–471.

Ibrahim, J.G., Chen M. and Sinha, D. (2001). *Bayesian survival analysis.* Springer.

Jamrozik, J. and Schaeffer, L.R. (1997). Estimates of genetic parameters for a test day model with random regressions for yield traits of first lactation Holsteins. *Journal of Dairy Science*, **80**, 762–770.

Kalbfleisch, J.D. and Prentice, R.L. (1980). *The statistical analysis of failure time data.* John Wiley.

Klein, J.P. (1992). Semi-parametric estimation of random effects using the Cox model based on the EM algorithm. *Biometrics*, **48**, 795–806.

Klein, J.P. and Moeschberger, M.L. (1997). *Survival Analysis: Techniques for censored and truncated data.* Springer.

Korsgaard, I.R., Andersen, A.H. and Jensen, J. (1999). Discussion of heritability of sur-

vival traits. Proceedings of the International Workshop on EU Concerted Action: Genetic Improvement of Functional Traits in Cattle (GIFT), *Interbull Bulletin*, **21**, 31–35.

Kortram, R.A., Lenstra, A.J., Ridder, G. and van Rooij, A.C.M. (1995). Constructive identification of the mixed proportional hazards model. *Statistica Neerlandica*, **49**, 269–281.

Laird, N.M. and Ware, J.H. (1982). Random-effects Models for Longitudinal Data. *Biometrics*, **38**, 963–974.

Lancaster, T (1990). *The econometric analysis of transition data.* Cambridge University Press.

Lawless, J.F. (1982). *Statistical models and methods for lifetime data.* John Wiley.

Louis, T.A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society B*, **44**, 190–200.

Müller, H.G. and Wang, J.L. (1990). Nonparametric analysis of changes in hazard rates for censored survival data: an alternative to change-point models. *Biometrika*, **7**, 305–314.

Müller, H.G. and Wang, J.L. (1994). Hazard rate estimation under random censoring with varying kernels and bandwidths. *Biometrics*, **50**, 61–76.

Mathsoft 1999. *S-Plus 2000 User's Guide : Statistics.* Mathsoft, Seattle.

McGilchrist, G.A. (1993). REML estimation for survival models with frailty. *Biometrics*, **49**, 221–225.

Meyer, K. (1992). Variance components due to direct and maternal effects for growth traits of Australian beef cattle. *Livestock Production Science*, **31**, 179–204.

Meyer, K. (1999). Estimates of genetic and phenotypic covariance functions for post-weaning growth and mature weight of beef cows. *Journal of Animal Breeding and Genetics*, **116**, 181–205.

Milne, C. (2000). The history of the Dorper sheep. *Small Ruminant Research*, **36**, 99–102.

Morgan, B.J.T. (1992). *Analysis of quantal response data.* Chapman and Hall.

Mukasa-Mugerwa E., Lahlou-Kassi, A., Anindo, D, Rege, J.E.O., Tembely, S., Markos T. and Baker R.L. (2000). Between and within breed variation in lamb survival and the risk factors associated with major causes of mortality in indigenous Horro and Menz sheep in Ethiopia. *Small Ruminant Research*, **37**, 1–12.

Murphy, S.A. (1994). Consistency in a proportional hazards model incorporating a random effect. *Annals of Statistics*, **22**, 712–731.

Murphy, S.A. (1995). Asymptotic theory for the frailty model. *Annals of Statistics*, **23**, 182–198.

Murphy, S.A. and van der Vaart, A.W. (1997). Semi-parametric likelihood ratio inference. *Annals of Statistics* **25**, 1471–1509.

Nguti, R., Janssen, P., Rowlands, J., Audho, J. and Baker, L. (2003). Survival of the Red Maasai, Dorper and crossbred lambs in the sub-humid tropics. *Animal Science*, **76**, 3–17.

Nguti, R., Claeskens, G. and Janssen, P. (2003). One-sided tests in shared frailty models. Technical Report (under revision).

Nguti, R., Burzykowski, T., Rowlands, J., Renard, D. and Janssen, P. (2003). Joint modelling of repeated measurements and event time: application to survival and performance traits of lambs bred in sub-humid tropics. Technical Report.

Over, H.J., Jansen, J., and Von Olm, P.W. (1992). Distribution and impact of helminth diseases of livestock in developing countries. *FAO Animal Production and Health Paper No. 96.*

Parner, E. (1998). Asymptotic theory for the correlated gamma-frailty model. *Annals of Statistics*, **26**, 183-214.

Petersen, J.H. (1998). An additive frailty model for correlated life times. *Biometrics*, **54**, 646–661.

Prentice, R.L. (1982). Covariate measurement errors and parameter estimation in a failure time regression model. *Biometrika*, **69**, 331–342.

Ramlau-Hansen, H. (1983). Smoothing counting process intensities by means of kernel functions. *Annals of Statistics*, **11**, 453–466.

Renard, D., Geys, H., Molenberghs, G., Burzykowski, T., Buyse, M., Bijnens, L. and Vangeneugden, T. (2002). Validation of a longitudinally measured surrogate marker for a time-to-event endpoint. *Journal of Applied Statistics*, **29**, 235–247.

Ripatti, S. and Palmgren, J. (2000). Estimation of multivariate frailty models using penalized partial likelihood. *Biometrics*, **56**, 1016–1022.

Ripatti, S., Larsen, K. and Palmgren, J. (2002). Maximum likelihood inference for multivariate frailty models using a Monte Carlo EM Algorithm. *Lifetime Data Analysis*, **8**, 349–360.

Robertson, T., Wright, F.T. and Dykstra, R.L. (1988). *Order-restricted statistical inference*. John Wiley.

Self, S.G. and Liang, K.-Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under non-standard conditions. *Journal of the American Statistical Association*, **82**, 605–610.

Sen, P.K. and Silvapulle, M.J. (2002). An appraisal of some aspects of statistical inference under inequality constraints. *Journal of Statistical Planning Inference*, **107**, 3–43.

Silvapulle, M.J. and Silvapulle, P. (1995). A score test against one-sided alternatives. *Journal of the American Statistical Association*, **90**, 342–349.

Sleeper, L.A. and Harrington, D.P. (1990). Regression splines in the Cox model with application to covariate effects in liver disease. *Journal of the American Statistical Association*, **85**, 941–949.

Smith, W. D. (1999). Prospects for vaccines of helminth parasites of grazing ruminants. *International Journal for Parasitology*, **29**, 17–34.

Southey, B.R., Rodriguez-Zas, S.L. and Leymaster, K.A. (2001). Survival analysis of lamb mortality in a terminal sire composite population. *Journal of Animal Science*, **79**, 2298–2306.

Stram, D.O. and Lee, J.W. (1994). Variance components testing in the longitudinal mixed effects models. *Biometrics*, **50**, 1171–1177.

Stram, D.O. and Lee, J.W. (1995). Correction to "Variance components testing in the longitudinal mixed effects models". *Biometrics*, **51**, 1196.

Tanner, M and Wong, W.H. (1983). The estimation of the hazard function from randomly censored data by kernel method. *Annals of Statistics*, **11**, 989–993.

Therneau, T.M., and Grambsch, P.M. (2000). *Modelling survival data: extending the Cox model*. Springer.

Therneau, T.M., Grambsch, P.M. and Pankratz, V.S. (2003). Penalized survival models and frailty. *Journal of Computational and Graphical Statistics*, **12**, 156–175.

Therneau, T.M. (2003). On mixed-effect Cox models, sparse matrices, and modelling data from large pedigrees. Technical Report.

Traore, A., Peacock, C.P., Mack, S. and Agyemang, K. (1985). Mortality of lambs in African traditional livestock production systems. *Veterinary Research Communications*, **9**, 295–301.

Tsiatis, A.A., DeGruttola, V. and Wulfsohn, M.S. (1995). Modelling the relationship of survival to longitudinal data measured with error. An application to survival and CD4 counts in patients with AIDS. *Journal of the American Statistical Association*, **90**, 27–37.

Vaida, F. and Xu, R. (2000). Proportional hazards model with random effects. *Statistics in Medicine*, **19**, 3309–3324.

Vaupel, J., Manton, K.G. and Stallard, E. (1979). The impact of heterogeneity in individual frailty and the dynamics of mortality. *Demography*, **16**, 439–454.

Verbeke, G. and Molenberghs, G. (2000). *Linear mixed models for longitudinal data.* Springer.

Verbeke, G. and Molenberghs, G. (2003). The use of score tests for inference of variance components. *Biometrics*, **59**, 254–262.

Vu, H.T.V. and Zhou, S. (1997). Generalization of likelihood ratio tests under nonstandard conditions. *Annals of Statistics*, **25**, 897–916.

Waller, P. J. (1997). Anthelmintic resistance. *Veterinary Parasitology*, **72**, 391–412.

Wallin, L., Strandberg, E., Philipsson,J. and Dalin, G. (2000). Estimates of longevity and causes of culling and death in Swedish warm blood and cold blood horses. *Livestock Production Science*, **63**, 275–289.

Wang, Y. and Taylor, J.M.G. (2001). Jointly modelling longitudinal and event time data with application to acquired immunodeficiency syndrome. *Journal of the American Statistical Association*, **96**, 895–905.

Wassell, J.T., Wojciechowski, W.C. and Landen, D.D. (1999). Recurrent injury event-time analysis. *Statistics in Medicine*, **18**, 3355–3363.

Watson, G.S. and Leadbetter, M.R. (1964). Hazard analysis I. *Biometrika*, **51**, 175–184.

Wei, L.J., Lin, D.Y. and Weissfeld, L. (1989). Regression analysis of multivariate incomplete failure data by modelling marginal distributions. *Journal of the American Statistical Association*, **84**, 1065–1073.

Wilson, R. T. (1991). Small ruminant production and the small ruminant genetic resource in tropical Africa. *FAO Animal Production and Health Paper No. 88.*

Wilson, R.T., Traore, A. and Mukasa-Mugerwa, E. (1993). Mortality and morbidity of African small ruminants under various management systems. *In: Penn, G. (Ed), Pathologie Caprine et Productions, 2e Colloque Internationale de Niort. Etudes et Syntheses de l'IEMVT*, **42**, 208–236.

Wood, A. (2002). Joint modelling of longitudinal and time-to-event data. *PhD Thesis,*

*Lancaster University, UK.*

Woolaston, R. R. and Baker, R. L. (1996). Prospects of breeding small ruminants for resistance to internal parasites. *International Journal for Parasitology*, **26**, 845–855.

Wulfsohn, M.S. and Tsiatis, A.A. (1997). A joint model for survival and longitudinal data measured with error. *Biometrics*, **53**, 330–339.

Xu, J. and Zeger, S.L. (2001). Joint analysis of longitudinal data comprising of repeated measurements and time to events. *Journal of Applied Statistics*, **50**, 375–387.

Yazdi, M.H., Rydhmer, L., Cederberg, E.R., Lundeheim, N. and Johansson, K. (2000). Genetic study of longevity in Swedish Landrace sows. *Livestock Production Science*, **63**, 255–264.

Yazdi, M.H., Visscher, P.M., Ducrocq, V. and Thompson, R. (2002). Heritability, reliability of genetic evaluations and response to selection in proportional hazards models. *Journal of Dairy Science*, **85**, 1563–1577.

# Samenvatting

In Subsahara Afrika worden op kleine boerderijen schapen en geiten gehouden voor de vleesproductie. Deze dieren grazen in open natuurlijke weilanden of in weilanden die toebehoren aan een lokale gemeenschap. Ongeveer dertig procent van de opgroeiende dieren bereiken de volwassen leeftijd niet omdat ze geïnfecteerd worden door endoparasieten. Alhoewel er technieken bestaan om de wormeitjes op besmette weilanden te vernietigen, is het gebruik ervan beperkt omwille van de kost. Daarom is het nodig om dieren te kweken die zich goed aanpassen aan een omgeving waarin deze parasieten veelvuldig voorkomen. Dit is op lange termijn de meest rendabele manier om de productiviteit te verhogen.

De gegevens, die in dit proefschrift bestudeerd worden, komen van een kweekprogramma dat in de periode 1991-1996 werd uitgevoerd in het International Livestock Research Institute (ILRI) in Nairobi (Kenya). Het objectief van het programma was om de genetische resistentie van Red Masaai schapen, Dorper schapen en gekruiste rassen tegen endoparasieten te onderzoeken. De lammeren werden maximaal over een periode van één jaar opgevolgd. Gebaseerd op deze gegevens is de genetische resistentie en resilintie van lammeren bestudeerd door gebruik te maken van gemengde lineaire modellen voor de variabelen packed cell volume (PCV), fecale wormeitelling (FEC : faecal egg count) en lichaamsgewicht (BWT : bodyweight). Deze variabelen zijn voor elk lam op een aantal tijdsmomenten gemeten. Uit analyse van de gegevens (Baker et al., 1994, 1999, 2003) is gebleken dat Red Masaai schapen een hogere resilintie (hogere PCV waarde) en een hogere resistentie (lagere FEC) hebben dan Dorper schapen.

Binnen het experiment is ook informatie aanwezig over de schapen die stierven binnen één jaar. We beschikken met andere woorden over de overlevingstijden van de schapen

(de tijd van hun geboorte tot hun dood). In deze context is 'dood' de gebeurtenis die ons interesseert. De dood van het lam is het moment waarop het lam 'faalt', we gebruiken daarom ook de term 'faaltijd' (time-to-event) naast de term 'overlevingstijd'.
De studie van deze overlevingstijden is het hoofddoel van dit proefschrift. In het experiment zijn er groepen (clusters) van lammeren met dezelfde ram als vader. Overlevingstijden van lammeren in dezelfde cluster, dus met dezelfde vader, zijn derhalve gecorreleerd met elkaar. We noemen dit soort gegevens multivariate faaltijden of multivariate overlevingstijden. Dit soort clustering kan ook voorkomen binnen één subject, zoals blijkt uit het volgende voorbeeld. Noteren we voor een schaap de tijden tussen opeenvolgende aanvallen van trypanosomose of slaapziekte, dan is het realistisch om te onderstellen dat de faaltijden gecorreleerd zijn (we spreken in dit geval over recurrente faaltijden).
In de laatste jaren zijn er gepaste statistische modellen beschreven om dit soort gegevens te modelleren en te analyseren (Klein en Moeschberger, 1997, Therneau en Grambsch, 2000, Hougaard, 2000). Om de correlatie tussen de gegevens van eenzelfde groep (cluster) of equivalent de heterogeniteit tussen de clusters te beschrijven, wordt een frailty factor (een stochastische veranderlijke met één als gemiddelde) als extra factor toegevoegd aan het model met proportionele risicofuncties (het proportional hazards model of het Cox regressiemodel). Alle lammeren die dezelfde ram als vader hebben, delen hetzelfde frailty effect en we spreken daarom over shared frailty modellen. In het Cox model wordt het effect van de regressoren parametrisch en de referentie risicofunctie (de baseline hazard function) niet-parametrisch gemodelleerd en we spreken daarom van een semi-parametrisch model. Modelleert men de referentie risicofunctie wel parametrisch, bijvoorbeeld met een Weibull risicofunctie, dan spreekt men over een parametrisch model.

In de literatuur zijn een aantal methoden beschreven om shared frailty modellen statistisch te analyseren. Om de onderliggende methodologie goed te begrijpen wordt in Duchateau et al. (2000) een grondig overzicht gegeven. Meer concreet bekijken we daar een multicentra klinische studie. In een dergelijke studie zijn de centra of de hospitalen de clusters waarbinnen de overlevingstijden van patinten gecorreleerd zijn. De bedoeling van deze publicatie was om het effect te onderzoeken van bepaalde design aspecten van een multicentra studie op de kwaliteit van de schatters van de parameters in het model. Via simulaties werd nagegaan hoe de vertekening en de spreiding van de schattingen van de heterogeniteitsparameter (de variantie van de frailty dichtheid) benvloed wordt door

de relatie tussen het aantal centra en het aantal patinten per centrum in de studie. In dit proefschrift is het methodologisch overzicht uit deze publicatie opgenomen, dit is de inhoud van Hoofdstuk 2. De bedoeling van dit hoofdstuk is het basismateriaal aan te reiken dat nodig is om de verdere hoofdstukken van dit proefschrift goed te kunnen onderbouwen.

Een belangrijke vraag die opduikt bij het gebruik van frailty modellen is na te gaan of er al dan niet heterogeniteit bestaat tussen de clusters. Zoals hierboven reeds gezegd, wordt heterogeniteit gedefinieerd als de variantie van de frailty dichtheid. We willen daarom toetsen of die variantie nul is (geen heterogeniteit) dan wel strikt positief (wel heterogeniteit). Dit eenzijdig toetsingsprobleem is geen standaard probleem omdat, onder de nulhypothese, de heterogeniteitsparameter op de rand van de parameterruimte ligt. Het construeren van gepaste toetsen voor dit probleem heeft ruime aandacht gekregen voor gemengde lineaire modellen (Self en Liang, 1987, Stram en Lee, 1994, 1995, Verbeke en Molenberghs, 2003). Binnen frailty modellen is de toetsingtheorie niet ontwikkeld. Gebaseerd op simulaties is in Duchateau et al. (2002) het vermoeden geformuleerd dat de asymptotische verdeling van de toets die gebaseerd is op de verhouding van de aannemelijkheidsfuncties onder resp. de nulhypothese en de alternatieve hypothese een 50:50 mengeling is van een puntmassa in nul en een chi-kwadraat verdeling met één vrijheidsgraad. In Hoofdstuk 3 geven we een bewijs van dit vermoeden voor parametrische shared frailty modellen. Binnen dat hoofdstuk bekijken we ook het limietgedrag van de verdeling van een gepaste score toets voor heterogeniteit.

In het tweede gedeelte van dit proefschrift geven we, in Hoofdstuk 4, een gedetailleerde beschrijving van de gegevens die binnen het hierboven beschreven kweekprogramma zijn verzameld. In dit hoofdstuk geven we ook een samenvatting van de statistische analyses die voor deze data zijn uitgevoerd. Verder tonen we aan dat de data set nog heel wat ongebruikte informatie bevat. Concreet zullen we bestuderen hoe de overlevingsgegevens gemodelleerd kunnen worden met behulp van frailty modellen. Door gebruik te maken van kernschatters (niet-parametrische schatters) van de referentie risicofunctie tonen we in Hoofdstuk 5 aan dat voor de gegevens uit dit kweekprogramma het niet-parametrisch specificeren van de risicofunctie inderdaad te verkiezen is boven het parametrisch modelleren. De kernschatters die hier gebruikt worden zijn gedefinieerd in Müller en Wang (1994) (zie ook Tanner en Wong, 1983).

Zoals reeds voorheen beschreven, beschikken we voor PCV, BWT en FEC voor ieder lam over longitudinale gegevens. Deze kenmerken zijn, over het algemeen genomen, ongeveer maandelijks beschikbaar over de periode waarin het lam geobserveerd werd (in het beste geval tot de leeftijd van één jaar). Daarom hebben we voor de frailty modellen, die we in Hoofdstuk 5 gebruiken, deze kenmerken opgenomen als tijdsafhankelijke covariaten. Het is immers duidelijk dat deze kenmerken informatie bevatten omtrent het risico om te sterven. Zo zullen bijvoorbeeld zieke dieren typisch een lage PCV en een hoge FEC waarde hebben; en zwakke dieren zullen vaak een laag lichaamsgewicht hebben.

Recent zijn, binnen het toepassingsdomein van de medische statistiek, gezamenlijke modellen ontwikkeld die toelaten om de associatie te beschrijven tussen de overlevingstijden en het longitudinale meetproces (De Grutolla en Tu, 1994, Tsiatis et al., 1995, Faucett en Thomas, 1996, Wulfsohn en Tsiatis, 1997, Henderson et al., 2000, Wang en Taylor, 2001, Renard et al., 2002). In Hoofdstuk 6 tonen we in een korte methodologische beschrijving aan hoe de associatie tussen de overlevingstijden en de longitudinale kenmerken PCV, BWT en FEC kan beschreven worden. Hiertoe worden de longitudinale kenmerken gemodelleerd als gemengde lineaire modellen en de overlevingstijden als modellen met proportionele risicofuncties. Het verbinden van de twee modellen gebeurt door de associatie tussen de modellen te beschrijven door een latent Gaussisch proces dat afhangt van de random effecten die optreden in de modellering van de longitudinale kenmerken. Zodoende wordt de overlevingstijd in het gezamenlijk model bekeken als een overlevingsmodel met subject specifieke random effecten.