Statistical modelling of repeated and multivariate survival data

Wintrebert, C.M.A.

CHAPTER 1

Introduction: Survival Analysis and Frailty Models

This dissertation consists of a general introduction on survival analysis and frailty models, followed by three accepted and two submitted papers which can be read as self-contained papers. It will end with a general summary.

1.1 Introduction: survival analysis

This thesis is about survival analysis, which is the statistical analysis of survival data. Survival data is a term used for describing data that measure the time to a given event of interest. The name survival data arose because originally events were most often deaths. The term survival data is now used for all kind of events. In all cases, the event can be seen as a transition from one state to another. In medical studies, often the main emphasis is the timing of this event.

1.1.1 Probability tools

In this section, the probability tools usually encountered in survival analysis and their properties are described.

Let T be the time variable, considered as a positive real valued variable, having a continuous distribution with finite expectation. For applications, this variable represents the time being in a given state or the time between two events. Several functions characterize the distribution of T:

- $f(t), t \geq 0$ is the probability density of T;
- $S(t) = P(T > t) = \int_t^\infty f(x)dx = 1 - F(t)$, is the survival function, which is the probability of an individual surviving beyond time t ($F(t)$ is the cumulative distribution function);
- the hazard function defined for $t > 0$: $\lambda(t) = f(t)/S(t) = \lim_{\delta t \to 0} \frac{P(t \leq T < t+\delta t | T \geq t)}{\delta T} = \frac{-\partial S(t)/\partial t}{S(t)} = -\partial \frac{\ln S(t)}{\partial t}$, which represents the probability that an individual alive at t experiences the event in the next period δt.

- The cumulative hazard function $\Lambda(t) = \int_0^t \lambda(x)dx$ is a useful quantity in survival analysis because of its relation with the hazard and survival functions: $S(t) = \exp(-\Lambda(t))$.

1.1.2 Censored and truncated data

Survival data are also distinguished from other data because the survival time is not always observed. This peculiar feature, often present in survival data, is known as censoring. This means that sometimes it is only known that T is larger than some time (censoring time) C. In that case, we say that the data are right censored. Analogously, the data are said to be left censored if it is only known that the survival time T is smaller than C. The data are interval censored if it is only known that the survival time falls in some known interval. In this thesis, we only consider right and/or interval censored data and make the assumption that the censoring time C and the survival time T are independent.

Some survival studies may contain truncated data. Left truncated data occur when individuals enter a study at a particular time-point and are followed from this entry time until the individual is censored or the event occurs. Right truncated data occur when only the individuals having experienced the event of interest are observable.

1.1.3 Common estimators of the survival function

Many parametric models (Weibull, lognormal, normal etc..) can be used to estimate the survival function (Klein and Moeschberger, 1997b). The non-parametric approaches: Kaplan and Meier (1958) and Aalen (1978) or Nelson (1969) are more often used in medical applications. The Kaplan-Meier estimator is written as the following product-limit estimator:

$$\hat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{d_i}{R_i}\right)$$

where (t_i, d_i) are the data of individual i, t_i representing the time to event or the time to censoring and d_i is the corresponding censoring indicator ($d_i = 1$ in case of event and $d_i = 0$ in case of censoring); R_i is the number of individuals still at risk at time t_i (still alive and uncensored just before t_i). The variance of the Kaplan-Meier estimator can be estimated by the Greenwood's formula:

$$\hat{V}[\hat{S}(t)] = \hat{S}(t)^2 \sum_{t_i \leq t} \frac{d_i}{R_i(R_i - d_i)}.$$

As an alternative, Nelson (1969) and in an other context Aalen (1978) estimated the cumulative hazard by the formula:

$$\hat{\Lambda}(t) = \sum_{t_i \leq t} \frac{d_i}{R_i}.$$

The estimated variance of the Nelson-Aalen estimator is due to Aalen (1978) and is estimated by:

$$\sigma^2_{\Lambda(t)} = \sum_{t_i \leq t} \frac{d_i}{R_i^2}.$$

When treating survival data and thus censored or truncated data extra care is needed to construct likelihood functions. Suppose we have a random sample of pairs (T_i, d_i), $i = 1, ..., n$ the likelihood function is written as

$$L = \prod_{i=1}^{n} Pr[t_i, d_i] = \prod_{i=1}^{n} [S(t_i)]^{1-d_i} [f(t_i)]^{d_i}.$$

This equation can also be simplified as:

$$L = \prod_{i=1}^{n} \exp(-\Lambda(t_i))[\lambda(t_i)]^{d_i}$$

1.1.4 Cox-regression model

Important aim in many clinical studies is to investigate the relation between the survival time and some risk factors called covariates. These risk factors might be fixed variables, or they may change over time (then called time-dependent covariates). Their influence on the survival is of great interest for clinicians and bio-statisticians and can be estimated by statistical models. The usual model for this kind of data is the so-called Cox-model, or the proportional hazards model. In this model, the relative risk is described parametrically and the hazard function non-parametrically. In this model, the hazard function for individual i is written as:

$$\lambda_i(t) = \lambda_0(t) \exp(\beta^T X_i).$$

$\lambda_0(t)$ is a baseline hazard function, left unspecified; $\exp(\beta^T X_i)$ is the relative risk of individual i, where X_i is the covariate vector of individual i. Cox (1975) proposed the partial likelihood method to estimate the β parameter of this model. The partial likelihood is a product over the uncensored failure times written as:

$$L(\beta) = \prod_{i=1}^{n} \left(\frac{\exp(\beta^T X_i)}{\sum_{j \in R_i} \exp(\beta^T X_j)} \right)^{d_i},$$

where each factor can be interpreted as the conditional probability that individual i dies at time t_i, given the risk set R_i.

An important fact is that $\lambda_0(t)$ cancels out. The first and second derivatives of the log likelihood of the model can be derived. Parameter estimates can then be obtained by maximizing $L(\beta)$ using e. g. the Newton-Raphson procedure. Subsequently the

cumulative baseline hazard function $\Lambda_0(t)$ is estimated as in Breslow (1972). Several goodness of fit tests have been developed for the Cox model (Andersen, 1985; Commenges and Andersen, 1995; Schoenfeld, 1980). Martingale residuals provide the basis for a number of procedures that access model adequacy as well as model form see, e.g. (Barlow and Prentice, 1988; Fleming and Harrington, 1991; Grambsch et al., 1995; Verweij et al., 1998).

1.1.5 Martingale Residuals and counting process approach

Martingale residuals are useful for survival analysis. The martingale residual of individual i is defined as follows:

$$MR_i = d_i - \hat{\Lambda}(T_i).$$

They may be interpreted as the difference between "observed" and "expected" number of events for an individual.

The counting process approach replaces the pair of variables (T_i, d_i) with the following pair of functions $(N_i(t), Y_i(t))$ where $N_i(t)$ counts the number of events in $[0,t]$ for unit i and $Y_i(t)$ indicates if unit i is at risk of having an event at time t. Right-censored survival data are also included in this formulation as a special case; $N_i(t) = I(\{T_i \leq t, d_i = 1\})$ and $Y_i(t) = I(\{T_i \geq t\})$. In the proportional hazards model, the intensity process $\alpha_i(t; X_i)$ for $N_i(t)$ can be written as

$$\alpha_i(t; X_i) = \alpha_0(t) \exp(\beta^T X_i) Y_i(t).$$

Note that in order to avoid confusions, only in this section the intensity process is called α.

The estimated martingale residual for unit i at time t for the former model is thus defined as

$$\hat{M}_i(t) = N_i(t) - \int_0^t Y_i(s) \exp(\beta^T X_i) d\hat{\Lambda}_0(s),$$

where $\hat{\Lambda}_0(s)$ is the Breslow (1972) estimator given by

$$\hat{\Lambda}_0(t) = \int_0^t \frac{dN_\cdot(s)}{\sum_{i=1}^n Y_i(s) \exp(\hat{\beta}^T X_i)}$$

where $N_\cdot(t) = \sum_{i=1}^n N_i(t)$. Finally, denoting the estimated martingale residuals at $t = \infty$ as $\hat{M}_i(\infty) = \hat{M}_i$ we come back to the first expression given in this section:

$$\hat{M}_i = N_i(\infty) - \int_0^\infty Y_i(s) \exp(\hat{\beta}^T X_i) d\hat{\Lambda}_0(s) = "observed"_i - "expected"_i.$$

1.2 Frailty models

The concept of frailty provides a suitable way to introduce random effects in the model to account for association and unobserved heterogeneity. In its simplest form, a frailty is an unobserved random factor that modifies multiplicatively the hazard function of an individual or a group or cluster of individuals.

1.2.1 Introduction

Vaupel et al. (1979) introduced the term frailty and used it in univariate survival models. Clayton (1978) promoted the model by its application to multivariate situation on chronic disease incidence in families.

A random effect model takes into account the effects of unobserved or unobservable heterogeneity, caused by different sources. The random effect, called frailty and denoted here by Z is the term that describes the common risk or the individual heterogeneity, acting as a factor on the hazard function. Two categories of frailty models can be pointed out. The first one is the class of univariate frailty models that consider univariate survival times. The second one is the class of multivariate frailty models that take into account multivariate survival times.

1.2.2 Univariate frailty models

Univariate frailty models take into account that the population is not homogeneous. Heterogeneity may be explained by covariates, but when important covariates have not been observed, this leads to unobserved heterogeneity. Vaupel et al. (1979) introduced univariate frailty models (with a gamma distribution) into survival analysis to account for unobserved heterogeneity or missing covariates in the study population. The idea is to suppose that different patients possess different frailties and patients more "frail" or "prone" tend to have the event earlier that those who are less frail. The model is represented by the following hazard given the frailty:

$$\lambda(t|Z,X) = Z\lambda(t|X).$$

$\lambda(t|X)$ can be equal to the baseline hazard function $\lambda_0(t)$, or when we consider covariates $\lambda(t|X)$ may be equal to $\lambda_0(t)\exp(\beta^T X)$ (in a Cox regression model). The baseline hazard function $\lambda_0(t)$ can be chosen non-parametrically, or parametrically (Weibull, exponential, Gompertz, piecewise constant,...).

An important point is that the frailty Z is an unobservable random variable varying over the sample which increases the individual risk if $Z > 1$ or decreases if $Z < 1$.

The model can also be represented by its conditional survivor function:

$$S(t|Z,X) = \exp(-Z\int_0^t \lambda(u|X)du) = \exp(-Z\Lambda(t|X)),$$

where $\Lambda(t|X) = \int_0^t \lambda(u|X)du$. $S(t|Z,X)$ represents the fraction of individuals surviving until time t given Z and given the vector of observable covariates X.

Note that until now the model is described at individual level, but this individual model is not observable. That is the reason why it is essential to consider the model at a population level. The survival of the total population is the mean of the individual survival functions.

Many calculations can be done based on the Laplace transform. Hougaard (1984) demonstrated the importance of the Laplace transform for these calculations. The Laplace transform of a random variable Z is defined as

$$L(s) = \int \exp(-sz)g(z)dz = E[\exp(-sZ)]$$

where $g(z)$ is the density of Z. The integral is over the range of the distribution. The marginal survivor function can be calculated by

$$S(t|X) = \int S(t|Z,X)g(z)dz = E[S(t|Z,X)] = L(\Lambda(t|X)).$$

An important point is the identifiability of univariate frailty models. Univariate frailty models are not identifiable from the survival information alone. However, Elbers and Ridder (1982) proved that a frailty model with finite mean is identifiable with univariate data, when covariates are included in the model.

Many distributions can be chosen for the frailty, but the most common frailty distribution is the gamma distribution. The gamma distribution has been widely applied as a mixture distribution (Clayton, 1978; Hougaard, 2000; Oakes, 1982a; Vaupel et al., 1979; Yashin et al., 1995). From a computational and analytical point of view the gamma distribution is convenient, because it is easy to derive the closed form expressions of survival, density and the hazard function. This is due to the simplicity of the Laplace transform, which is the reason why this distribution has been used in most applications published so far. The density function of the gamma distribution gamma(z;θ,β) is given by $g(z) = \theta^\beta z^{\beta-1}\exp(-\theta z)/\Gamma(\beta)$ where $\theta > 0$, $\beta > 0$ and $z > 0$. θ is a scale parameter and β is called a shape parameter. For identifiability, we suppose $\theta = \beta$ which implies $EZ = 1$ and $varZ = 1/\theta$.

An other distribution which can be chosen for the frailty is the positive stable distribution (Hougaard, 1986a). A distribution is strictly stable if the sum of independent random variables from the distribution normalized follows the same distribution. Suppose $Z_1, ..., Z_n$ i.i.d, the distribution of the sum of $Z_1, ..., Z_n$ is stable if for each n, there exists a constant c_n, with $D(Z_1 + ... + Z_n) = D(c_n Z_1)$ where $D(Z)$ means the distribution of Z. The constants satisfy $c_n = n^{1/\alpha}$, for some $\alpha \in (0,2]$. For $\alpha = 2$, the stable distribution has finite variance and is the normal distribution. For $\alpha = 1$, the degenerate distribution is obtained. The stable distribution on the positive numbers has

$\alpha \in (0,1]$ and apart from scale factors have Laplace transform:

$$L(s) = E[\exp(-sZ)] = \exp(-s^{\alpha})$$

$(s \geqslant 0)$. This distribution is denoted $P(\alpha,\alpha,0)$. Note that the frailty model using this distribution is not identifiable in the univariate case, because the mean does not exist. Unidentifiability is also easily seen from the marginal survival function: $S(t|X) = \exp((-\Lambda_0(t)\exp(X\beta))^{\alpha}) = \exp(-\alpha\Lambda_0(t)\exp(X\beta))$, where the frailty parameter (α) acts as a multiplicative factor which is confounded by $\Lambda_0(t)$.

Other distributions which are sometimes applied for the frailty distribution are the well-known normal, the lognormal (McGilchrist and Aisbett, 1991), the three-parameter distribution (PVF) (Hougaard, 1986b), the compound poisson distribution (Aalen, 1988, 1992) and inverse gaussian distribution. The effect of different frailty distributions is investigated by Congdon (1995).

The role of shared frailty is more useful when we consider multivariate survival times.

1.2.3 Multivariate frailty models

A very common situation in survival analysis is clustered or repeated data. Clustered data are for instance data where individuals are divided in groups likes family or study centres. Repeated data are seen in case of longitudinal data, concerning multiple recurrences of an event for the same individual. The difficulty of working with this kind of data is due to the dependence of individuals within groups, or repeated measures within individuals. The dependence usually arises because individuals in the same group are related to each other or because of the recurrence of an event for the same individual. Multivariate frailty models have been used frequently for modelling dependence in multivariate time-to-event data (Clayton, 1978; Hougaard, 2000; Oakes, 1982a; Yashin et al., 1995). The aim of the frailty is to take into account the presence of the correlation between the multivariate survival times.

Constant shared frailty models

In this situation, individuals j in a group i are supposed to share the same frailty Z_i. The conditional hazard for individual j in group i is:

$$\lambda(t_{ij}|Z_i) = Z_i\lambda(t_{ij}),$$

where $\lambda(t_{ij}) = \lambda_0(t_{ij})\exp(\beta X_{ij})$ in the cox-regression model. The Z_i are independent identically distributed following a chosen distribution, like in the univariate frailty models. This model is therefore an extension of the preceding described model.

The model assumes that all time observations are independent given the values of the frailties. In other words, it is a conditional independence model. The value of Z

is constant over time and common to the individuals in the group and thus responsible for creating dependence. The interpretation of this model is that the between-groups variability (the random variation of Z) leads to different risks for the groups, which then show up as dependence within the group. In the case of gamma distribution for Z, I remember that $EZ = 1$ and $varZ = 1/\theta$. So, small value of θ reflect a greater degree of heterogeneity among groups and a stronger association within groups. The association between group members as measured by kendall's τ is $\tau = \frac{1}{1+2\theta}$, and large value of θ corresponds to the case of independence.

Note that the frailty models with multivariate survival data are identifiable in almost all cases.

It is assumed that there is independence between groups and between the times for the same value of i, owing to the common value Z_i of Z. Thus if the Z's do not vary then there is independence between the time observations.

Example of constant shared frailty model: the gamma frailty model

A first and common approach is to define the hazard function as:

$$\lambda(t_{ij} \mid Z_i) = Z_i \lambda_0(t_{ij}) \exp(\beta^t X_{ij}), i = 1, ..., n; j = 1, .., k_i$$

which is the hazard function of the j^{th} individual of group i given the frailty of group i (Z_i), where $\lambda_0(t_{ij})$ is an arbitrary baseline hazard rate and X_{ij} is the corresponding covariate vector. The frailty Z is supposed to follow a gamma distribution $g(z; \theta, \theta)$. The joint survival function for the k_i individuals within the i^{th} group is easily written by:

$$\begin{aligned} S\{t_{i1}, .., t_{ik_i}\} &= Pr(T_{i1} > t_{i1}, ..., T_{ik_i} > t_{ik_i}) \\ &= \int_0^\infty \prod_{j=1}^{k_i} Pr(T_{ij} > t_{ij} \mid Z_i) g(z_i) dz_i \\ &= [1 + \frac{1}{\theta} \sum_{j=1}^{k_i} \Lambda_0(t_{ij}) \exp(\beta^t X_{ij})]^{-\theta}. \end{aligned} \tag{1.1}$$

In this model, the estimates of $\beta, \theta, \Lambda_0(t)$ are obtained by using the EM (Expectation-Maximization) algorithm (Dempster et al., 1977). The EM algorithm is the main tool for estimation in frailty models in a frequentist framework and provides a means of maximizing complex likelihoods. The likelihood considered is the full likelihood we would have if the frailties were observed. This likelihood is easily manipulable and written as follows: $l_{full} = l_1(\theta) + l_2(\Lambda_0)$ where

$$\begin{cases} l_1(\theta) = n[\theta log\theta - log\Gamma(\theta)] + \sum_{i=1}^n [(D_i + \theta - 1) log Z_i - \theta Z_i] \\ l_2(\Lambda_0, \beta) = \sum_{i=1}^n \sum_{j=1}^{k_i} d_{ij} [\beta^t X_{ij} + log\lambda_0(t_{ij})] - Z_i \Lambda_0(t_{ij}) \exp(\beta^t X_{ij}). \end{cases}$$

In the E step the expected value of the full likelihood is completed given the current estimates of the parameters and the observable data. In the M step the estimates of the parameters which maximize the expected value of the full likelihood from the E step are obtained. For more details see Klein and Moeschberger (1997b).

If one assumes a parametric form for $\lambda_0(t_{ij})$, then, ML estimates are available by maximizing the log likelihood directly. In this following parametric example, the weibull distribution is chosen. This model is called the gamma-weibull frailty model:

$$\begin{aligned} L_i &= Pr((t_{i1}, d_{i1}), ..., (t_{ik_i}, d_{ik_i})) \\ &= \int Pr((t_{i1}, d_{i1}), ..., (t_{ik_i}, d_{ik_i}) \mid Z_i) g(z_i) dz_i \\ &= \int \prod_{j=1}^{k_i} [z_i \lambda_0(t_{ij}) \exp(\beta^t X_{ij})]^{d_{ij}} \exp(-z_i \Lambda_0(t_{ij}) \exp(\beta^t X_{ij})) g(z_i) dz_i \qquad (1.2) \\ &= \prod_{j=1}^{k_i} [\lambda_0(t_{ij}) \exp(\beta^t X_{ij})]^{d_{ij}} \frac{\theta^\theta}{\Gamma(\theta)} \frac{\Gamma(D_i + \theta)}{(\theta + \sum_{j=1}^{k_i} \Lambda_0(t_{ij}) \exp(\beta^t X_{ij}))^{D_i + \theta}} \end{aligned}$$

Because in the weibull situation, $\lambda_0(t_{ij}) = \alpha\beta t_{ij}^{\alpha-1}$ and the corresponding cumulative baseline hazard $\Lambda_0(t_{ij}) = \beta t_{ij}^{\alpha}$ the final expression of the likelihood is then easily derived, and also the log likelihood.

Usually the log likelihood is directly maximized using Newton-Raphson procedures and estimates of the variability of the parameter estimates are obtained by inverting the information matrix.

Limitations of the constant shared frailty models

The study and use of the constant shared frailty model confront us with its three principal limitations.

First, in most of the cases, a one-dimensional frailty can only imply a positive correlation within group. However, there are some situations in which the association is negative like time to response to treatment and survival.

Secondly, the model constrains the unobserved factors to be the same within a group of clustered observations implying constant correlation between all individuals in a cluster, and also to be the same during follow-up. This is unsatisfactory in many situations, because not always reflecting the reality.

Finally, the dependence parameter and the population heterogeneity are determined at the same time, and can be confounded. This can lead to difficulty in the interpretation.

These limitations suggest further developments of the frailty approach.

Correlated frailty models

There exists a need for more flexibility in modelling correlation. Most of the correlated frailty models developed until now are bivariate frailty models and applied for example

on twin data. Indeed, these models extend the idea of individual frailty to bivariate case and include shared frailty models as special cases. The novelty and difficulty in these models is that related individuals have different but dependent frailties. Such frailties are often constructed using independent additive components with one common component for both frailties. The identifiability conditions in the case of correlated gamma frailty models are discussed by Yashin and Iachine (1999).

Yashin et al. (1995) assumed gamma distributed frailties, Vaida and Xu (2000) suggested a bivariate frailty model in a slightly different setting, dos Santos et al. (1995) used a combination of a shared lognormal and a gamma frailty model on breast cancer data. Zahl (1997) used several correlated gamma frailty models to model the excess hazard. Li (2002) proposed a multivariate gamma frailty model in a genetic situation.

1.3 Introduction of the next chapters: Outline of the thesis

The emphasis of this thesis lies on complex survival data and on the modelling of this kind of data. Statistical models are developed or adapted and applied to five different real data sets, which all contain repeated censored measurements. To take into account the correlation between these repeated measurements, a frailty is considered in all statistical analysis used. Extensions of and alternatives for frailty models are considered. Special attention is paid to the role of the frailty and the effect of its use.

The centre-effect on survival after bone marrow transplantation is studied in chapter 2. Models which are able to take into account a time-dependent frailty are proposed and compared.

In chapter 3 survival analysis approaches are used for modelling an ecological capture-recapture data set. A joint model of breeding and survival on the Kittiwake bird is developed using frailty models.

In chapter 4, the emphasis lies on the frailty model used in a genetic context. Our model is applied on age at onset of Huntington disease. Correlation structure between different kinds of family members such as siblings are tested and estimated with martingale residuals on the Cox regression model including known risk factors as the number of CAG-repeats.

Chapter 5 concerns the estimation of the correlation between processes with frailties. The approach is applied on the Dutch part of the data set from the Caprie trial, involving cardiac, cerebral and peripheral atherosclerosis.

In chapter 6, the point of interest is the marginal survivor curve in different simulated balanced and unbalanced longitudinal situations. The frailty approach is compared to a weighted approach.

Finally, in chapter 7 a general summary can be found.