THESIS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

Correlated random effects models for clustered survival data

Frank Eriksson

Department of Mathematical Sciences
Division of Mathematical Statistics
Chalmers University of Technology and University of Gothenburg
Göteborg, Sweden 2012

Correlated random effects models for clustered survival data
Frank Eriksson
ISBN 978-91-7385-670-6

Doktorsavhandlingar vid Chalmers tekniska högskola
Ny serie nr 3351
ISSN 0346-718X

Department of Mathematical Sciences
Division of Mathematical Statistics
Chalmers University of Technology and University of Gothenburg
SE-412 96 Göteborg
Sweden
Telephone +46 (0)31 772 1000

Typeset with LaTeX.
Printed in Göteborg, Sweden 2012

Correlated random effects models for clustered survival data

Frank Eriksson

Department of Mathematical Sciences
Division of Mathematical Statistics
Chalmers University of Technology and University of Gothenburg

Abstract

Frailty models are frequently used to analyse clustered survival data in medical contexts. The frailties, or random effects, are used to model the association between individual survival times within clusters.

Analysis of survival times of related individuals is typically complicated because follow up on an event type of interest is censored by events of secondary interest. Treating such competing events as independent may yield an incorrect analysis when the random effects associated with other event types are dependent of the event type of interest. We study two related inferential procedures for dependent data where the frailties of the type specific hazards may be correlated between competing event types.

Routine registers offer possibilities to study covariate effects on survival times for rare diseases, for which large cohorts are required. However, the vast amount of data and the clustering of related individuals pose statistical challenges. In the first paper we adapt maximum likelihood methods for semiparametric transformation regression models to a cohort register subsampling design. This approach drastically reduces the computing times with a minor loss of efficiency, and results in practically useful estimation procedures.

In the second paper we propose an estimator of covariate effects based on the observed intensities, where the nonparametric baseline hazards are profiled out. Thereby we reduce the problem to finite dimensions, where e.g. the covariance matrix is more directly estimated. A set of frailty structures for paired competing risks data based on sums of gamma variables is investigated through simulations.

We establish the asymptotic properties of the estimators and present consistent covariance estimators. Worked examples are provided for illustration.

Key words: Survival analysis; frailty model, competing risks; random effects; case-cohort; routine register; semiparametric

List of Papers

The thesis is based on the following papers.

I. Eriksson, F. *Semiparametric transformation models for clustered survival data from routine registers.*

II. Eriksson, F. *A frailty model for paired competing risks survival data.*

Acknowledgments

I would like to express my gratitude towards my supervisors Dragi Anevski and Marita Olsson, and to my examiner Olle Nerman. I also direct a special thanks to Thomas Scheike for inviting me to visit him in Copenhagen, and for guidance and help while writing the second paper of this thesis. Furthermore I am thankful for Daniel Ahlberg, Ottmar Cronie, Fredrik Lindgren and Kyle Raymond for helpful suggestions and discussions.

Frank Eriksson

Contents

Part I

Introduction

1

Survival analysis

In survival analysis the response variable is the time T^* from some well defined time origin to a specific event. T^* can for instance be the life length of an individual, or the age at onset of a disease, or the time from treatment of a disease to relapse. Typically some event times are incompletely observed due to censoring. The most commonly encountered censoring is *right censoring* when we only observe an individual up to a possibly random censoring time C, i.e. we observe $T = T^* \wedge C$ and an indicator $\Delta = I(T = T^*)$ of whether or not censoring has occurred before the event time of interest. This may be because the subject has still to experience the event when the study is closed or because the individual is lost for follow-up due to other reasons. We assume that there is a maximum observation time $\tau < \infty$ and that all individuals still alive at this age are censored. A concept related to right-censoring frequently encountered in practice is *left-truncation* where an individual is only included in the sample conditionally on having survived till some given entry time V. Many registers used in epidemiological studies have left-truncated life times because they only include individuals or families that were alive at a given date. The Danish twin registry for example only includes twins that were both alive when the cancer registration started in 1943.

It is not obvious at first glance how to incorporate censored and truncated observations into inference for the distribution of T. Estimation based only on the complete data may give biased results, so the censored observations need to be taken into account. Modelling of the hazard rate λ, the event rate at time t conditional on survival until time t, has proved to be highly successful for this purpose.

The hazard rate may be interpreted as the instantaneous individual failure

rate among those at risk and is given by

$$\begin{aligned} \lambda(t) &= \lim_{dt\to 0} \frac{1}{dt} P(t \le T^* < t+dt | T^* \ge t) \\ &= -\frac{S'(t)}{S(t)}, \end{aligned} \tag{1.1}$$

where $S(t) = P(T^* > t)$ is the survival function of T^*, the probability that the event of interest has not happened at time t. From (1.1), by integration and using $S(0) = 1$, we see that the survival function may be calculated from the hazard rate as

$$\begin{aligned} S(t) &= \exp\left(-\int_0^t \lambda(s)ds\right) \\ &= \exp\left(-\Lambda(t)\right), \end{aligned} \tag{1.2}$$

where $\Lambda(t) = \int_0^t \lambda(s)ds$ is called the cumulative hazard rate up to time t. Note that by (1.1) and (1.2), the survival function and hazard rate are alternative but equivalent representations and both completely specify the distribution of T^*.

A convenient representation of survival data is by the *counting process*

$$N(t) = I(T \le t, \Delta = 1)$$

which jumps to one at T^* if the event is not censored and otherwise stays $N(t) = 0$ throughout. The dynamics of $N(t)$ is described by its *intensity process*

$$R(t)\lambda(t)$$

which is the product of the hazard function and the *at risk* process

$$R(t) = I(t \le T),$$

or $R(t) = I(V < t \le T)$ if there is left-truncation, indicating whether the individual is observed to be at risk just before time t. The interpretation of the intensity process is that

$$R(t)\lambda(t)dt = E[dN(t)|\mathscr{F}_{t-}], \tag{1.3}$$

the conditional expectation of the increment of $N(t)$ over a very small time interval $[t, t+dt)$ given the "history" $\mathscr{F}_{t-}$ in $[0, t)$. $\mathscr{F}_{t-}$ represents the available data just before time t, and contains information e.g. of $N(s)$ for $s < t$, and possibly other counting processes if there is dependence among the individuals in the sample. Mathematically $(\mathscr{F}_t)_{t\ge 0}$ is a *filtration*, i.e. an increasing right-continuous family of σ-algebras, and both $N(\cdot)$ and $R(\cdot)$ are adapted to $\mathscr{F}_t$.

1.1 Regression models for failure time data

A typical goal of a survival study is to relate the effect of explanatory variables on survival. It is convenient to build a regression model using the hazard rate as target function. The model can then be used to examine various hypotheses about the impact of risk factors or estimate regression parameters that relate to the lifetimes, taking into account that some of the lifetimes are censored.

The most popular survival model is Cox's proportional hazards model under which the hazard rate for an individual with covariate vector $\mathbf{X}$ takes the form

$$\lambda(t|\mathbf{X}) = \lambda_0(t)e^{\beta_0^T\mathbf{X}},$$

where β_0 is a vector of regression parameters and $\lambda_0(\cdot)$ is a baseline hazard rate describing the shape of the hazard as a function of time. The model is thus *semiparametric* in that the baseline hazard rate is treated nonparametrically, while a parametric form is assumed for the covariate effect. Correspondingly, the parameter (β_0, λ_0) contains an infinite dimensional component λ_0 in addition to the finite dimensional covariate effect vector β_0 of particular interest.

When the covariates are time independent, the interpretation of the β_0 vector is particularly easy. Assume that we observe two individuals with covariate vectors $\mathbf{X}$ and $\tilde{\mathbf{X}}$, respectively. Then the ratio of their hazard rates is

$$\frac{\lambda(t|\mathbf{X})}{\lambda(t|\tilde{\mathbf{X}})} = \exp\left(\beta_0^T(\mathbf{X} - \tilde{\mathbf{X}})\right), \tag{1.4}$$

which is constant over time. Hence the name proportional hazards. The proportion (1.4) is called the relative risk or hazard ratio of the two individuals. For example, if the covariate vectors of two individuals differ only by a binary covariate, then the risk of experiencing the event for the individual with $x = 1$ relative to the individual with $x = 0$ is e^{β_0}.

The Cox model has had a monumental success in applied work. In some applications, however, the proportional hazards assumption may not be reasonable. A popular alternative is the proportional odds model that constrains the ratio of the odds of survival associated with two sets of covariate values to be constant over time. Consequently, the ratio of the hazards converge to one with time. This is different from the proportional hazards model that constrains the hazard ratio to be constant while the odds ratio tends to zero or infinity.

Both the proportional hazards and the proportional odds models are examples of semiparametric transformation models. In this broad class of models the failure time T^* is related to $\mathbf{X}$ by

$$H(T) = -\beta_0^T\mathbf{X} + \varepsilon, \tag{1.5}$$

where $H(\cdot)$ is a continuous unspecified increasing function and ε is a residual with a known parametric distribution. The choices of the extreme value and

standard logistic distributions yield the proportional hazards and proportional odds models respectively.

The more general model (1.5) can be extended to allow time-dependent covariates by specifying that the cumulative hazard function, conditional on the covariate history $\overline{\mathbf{X}}(t) = \{\mathbf{X}(s) : s \leq t\}$, takes the form

$$\Lambda\left(t|\overline{\mathbf{X}}(t)\right) = G\left(\int_0^t e^{\beta_0^T \mathbf{X}(s)} \lambda_0(s) ds\right), \tag{1.6}$$

where the transformation G is a continuously differentiable and strictly increasing function (Zeng and Lin, 2007a) and $\lambda_0(\cdot)$ is an arbitrary positive function. Specifying the function G while leaving the function λ_0 unspecified is equivalent to specifying the distribution of ε while leaving the function H unspecified.

One class of transformations is the Box-Cox transformations,

$$G(x) = \begin{cases} [(1+x)^\rho - 1]/\rho, & \rho > 0, \\ \log(1+x), & \rho = 0. \end{cases}$$

For $\rho > 1$ the covariate effect increase over time, for $\rho < 1$ the covariate effects decrease over time. Another useful set of transformations is the logarithmic transforms given by

$$G(x) = \begin{cases} \log(1+rx)/r, & r > 0, \\ x, & r = 0. \end{cases}$$

For $r > 0$, the covariate effects always decrease over time, with a higher rate of decrease for larger r. The choice $\rho = 1$ or $r = 0$ yields the proportional hazards model while the choice $\rho = 0$ and $r = 1$ yields the proportional odds model.

Expression (1.6) can generate very general models, but this generality often comes with a problem of a lack of transparency of the role of covariates. Except in special cases, it is typically difficult to look at the expression for the cumulative hazard and gain any intuitive insight into how covariates influence the hazard.

1.2 Competing risks

When studying a specific cause of death, the observation of the disease may be preceded by other events, the occurrence of which prevents us from observing the disease of interest. This *competing risks* situation is the rule rather than the exception in epidemiological follow-up studies. In the competing risks framework the observable information for an individual is the time to first event among the possible competing reasons.

A naive analysis could consider death without the disease of interest as independent censoring, assuming that the censoring mechanism is independent

of the event type of interest conditional on the covariates. However, violation of the independent censoring assumption may produce biased estimates of covariate effects.

One method for describing a model for competing risks data is to specify the cause specific hazards (Prentice et al., 1978). With T^* the survival time and κ a stochastic variable that registers the type of death $\kappa \in \{1, \ldots, K\}$ the cause specific hazard function is

$$\lambda_k(t) \quad = \quad \lim_{dt \to 0} \frac{1}{dt} P(t \leq T^* < t + dt, \kappa = k | T^* \geq t).$$

A competing risks model can be described by specifying all the cause specific hazards. Based on the cause specific hazards various consequences of the model can be derived. One such summary statistic is the cumulative incidence function for cause $k = 1, \ldots, K$, defined as the probability of dying from cause k before time t

$$F_k(t) \quad = \quad P(T^* \leq t, \kappa = k) = \int_0^t \lambda_k(s) S(s-) ds,$$

where $S(t) = P(T^* > t)$ is the survival function. The survival function is expressed in terms of the hazards as

$$S(t) \quad = \quad \exp\left(- \int_0^t \sum_{k=1}^{K} \lambda_k(s) ds \right).$$

Note that the cumulative incidence function for cause k depends on the other cause specific hazard functions. The cause-specific hazard function and cumulative incidence function provide different perspectives for cause-specific failure times. The effect of a covariate on the two measures can be very different. There is no longer a one-to-one correspondence between the cumulative incidence and cause-specific hazard.

We can estimate the cumulative incidence function for a specific cause by modelling and estimating the cause specific hazards, but this requires models of the hazards for all causes. The cumulative incidence can alternatively be modelled directly by the subdistribution approach (Fine and Gray, 1999). The subdistribution approach does not demand models for the other causes, but instead modelling of the censoring distribution is required.

1.3 Correlated event time data

Clustered failure time data arise when subjects are sampled in clusters so that the failure times within the same cluster tend to be correlated. Medical examples include the age at onset of a genetic disease among family members with

families serving as clusters. Sometimes one would assume a simple structure with a common distribution for all individuals in a cluster, while in other situations the structure may be rather complex. For instance, when considering the lifetimes of parents and children in a family, individuals within the same cluster are not exchangeable and we have to distinguish between levels.

There are two main approaches to modelling cluster effects, marginal and conditional. The choice depends mainly on the purpose of the study. In marginal models the covariate effects are specified unconditionally and we assume that the regression model holds marginally for each individual, but that individuals within groups are associated. For the conditional approach we assume instead that the model holds for each individual conditional on some unobserved effect, which is modelled as random.

The marginal approach is well suited for the situation where one aims at estimating regression effects on the population level, and only have to deal with correlation to get valid standard errors to ensure correct inference. Then the cluster structure is ignored when estimating the covariance effects and is only used to derive correct standard errors. This approach is closely linked to the generalized estimating equations methodology (Liang and Zeger, 1986). Marginal models do not make any assumptions regarding the dependence structure. It can be seen as an advantage that we do not have to rely on a specific structure, but on the other hand such models cannot be used for assessment of dependence. We will focus on conditional models.

1.3.1 Conditional models

Random effects have been suggested to model two different but related sources of variation in event time data. Vaupel et al. (1979) introduced a random effect into a survival model to address the issue of variation due to unobserved variables. They introduced the term *frailty* and applied the model in a demographic setting to adjust for population heterogeneity. In this setting the frailty accounts for unobserved individual covariates that are not included in the study either because of practical circumstances, or because they are not known to be risk factors. These covariates are not observed and must be considered random and integrated out.

Clayton (1978) suggested a random effect to account for variation that stems from unobserved common risk factors. When the frailty is integrated out correlation is induced among event times within groups of related individuals. Groups sharing some risk factor might be a family, a pair of twins or patients from the same hospital. The methodology is suitable also for repeated measurements on the same individual.

Although conceptually different, dependencies between intensities of competing risks can also be modelled by introducing unobserved random effects. Here the basic independence unit, which in the formulation above is the cluster,

e.g. a family, is now each subject. Associated with each individual, is a number of processes, one for each cause of death. A model is assumed for each cause specific intensity conditional on an unobserved random effect. The difference from the clustered survival setting is that we will always observe at most one event for each group since individuals can die only once.

The frailty is often modelled as an unobserved mean one random variable acting multiplicatively on the baseline hazard. We illustrate conditional modelling by considering a frailty of this type in a simple model where all members of the same cluster share frailty variable.

Assume right-censored competing risks data. Let T_{ij}^* and C_{ij} be the failure and censoring times for the jth individual in the ith cluster $i = 1, \ldots, n$, $j = 1, \ldots, m$ and let $\mathbf{X}_{ij}$ be a vector of covariates associated with this individual. We collect all failure and censoring times and covariates for cluster i in the variables $\mathbf{T}_i^*$, $\mathbf{C}_i$ and $\mathbf{X}_i$, respectively. In addition we assume the presence of some unobserved random effect V_i. Censoring, conditional on V_i and covariates, is assumed to be independent and noninformative on V_i. We assume that

$$(\mathbf{T}_i^*, \mathbf{C}_i, \mathbf{X}_i, V_i)\,,$$

$i = 1, \ldots, n$, are independent and identically distributed variables.

Denote the right-censored failure time $T_{ij} = T_{ij}^* \wedge C_{ij}$ and let $R_{ij}(t) = I(t \leq T_{ij})$ and $N_{ij} = I(T_{ij} \leq t, T = T^*)$ denote the individual at-risk process and counting process, respectively. We collect the at-risk and basic counting processes of cluster i in the vectors $\mathbf{R}_i$ and $\mathbf{N}_i$ and define the observed history of cluster i by

$$\mathscr{F}_t^i \quad = \quad \sigma\left\{\mathbf{N}_i(s), \mathbf{R}_i(s), \mathbf{X}_i(s) : \ 0 \leq s \leq t\right\}.$$

We define also the conditional history of cluster i where we pretend that we also observe V_i,

$$\mathscr{H}_t^i \quad = \quad \sigma\left\{\mathbf{N}_i(s), \mathbf{R}_i(s), \mathbf{X}_i(s), V_i : \ 0 \leq s \leq t\right\}.$$

Note that the histories are nested, i.e. $\mathscr{F}_t^i \subseteq \mathscr{H}_t^i$ for all t.

A conditional model is specified by assuming that the intensity of $N_{ij}(t)$ with respect to the conditional filtration $\mathscr{H}_{t-} = \bigvee_{i=1}^n \mathscr{H}_{t-}^i$, the smallest filtration that contains $\mathscr{H}_{t-}^i$, $i = 1, \ldots, n$, takes the form

$$R_{ij}(t) V_i \lambda_{ij}(t)$$

for some λ_{ij} that may depend on covariates.

The conditional history involves the unobserved frailties and cannot be used directly for inference. Instead we can rely on the observed history $\mathscr{F}_t =$

$\bigvee_{i=1}^{n} \mathscr{F}_t^i$. By the innovation theorem (Andersen et al., 1993, Section II.4.2), the intensity of $N_{ij}(t)$ with respect to the observed history is

$$\begin{aligned} E\left[R_{ij}(t)V_i\lambda_{ij}(t)|\,\mathscr{F}_{t-}\right] &= E\left[R_{ij}(t)V_i\lambda_{ij}(t)|\,\mathscr{F}_{t-}^i\right] \\ &= R_{ij}(t)E\left[V_i|\,\mathscr{F}_{t-}^i\right]\lambda_{ij}(t), \end{aligned} \tag{1.7}$$

where the first equality follows from the independence across clusters.

1.3.2 Shared gamma frailty

The classical shared frailty model (Clayton, 1978) for clustered survival data assumes that a gamma distributed frailty variable with mean one and unknown variance ν is shared within clusters. The value $\nu = 0$ corresponds to independence, and a high value of ν correspond to a high correlation between the survival times.

The gamma distribution is a mathematically convenient choice as the conditional expectation in (1.7) can be computed in closed form as

$$E\left[V_i|\,\mathscr{F}_{t-}^i\right] = \frac{1+\nu\sum_{j=1}^{m} N_{ij}(t-)}{1+\nu\sum_{j=1}^{m}\int_0^{t-} R_{ij}(s)\lambda_{ij}(s)ds}.$$

1.3.3 Additive gamma frailty

In a shared frailty model, frailty is defined as a measure of the relative risk which the cluster share. Thus the frailty variable is associated with groups of individuals rather than individuals. Yashin et al. (1995) developed a correlated frailty model for bivariate survival data. In this model the frailties for the individuals within a cluster are not necessarily identical, as they are in the shared frailty model, but they are still correlated.

In the model of Yashin et al. (1995) the frailty for individual j, $j = 1, 2$, in a pair is split into two components,

$$Z^{(j)} = Z_0 + Z_j,$$

where Z_0 is a common shared component and Z_j is an individual component. The variables Z_0, Z_1 and Z_2 are assumed to be independent and gamma distributed with different shape parameters, but the same scale parameter. Let ν denote the variance of Z_0 and ν^* the variance of Z_1 and Z_2. Yashin et al. (1995) argue that the correlation

$$\mathrm{Corr}(Z^{(1)}, Z^{(2)}) = \frac{\nu}{\nu+\nu*}$$

is a proper index of the correlation between the survival times.

Yashin et al. (1995) discussed the model in the context of classical twin studies involving monozygotic and dizygotic twins. In this context Z_0 represents genetic and shared environmental effects while Z_1 and Z_2 describes non-shared environmental effects.

A similar model was used by Zahl (1997) in a competing risks setting to assess the excess hazard for patients with malignant melanoma and colon cancer. The set up is the same as in the previous example with three independent gamma variables, but now the hazards act on the same individual. Note that as both intensities act on the same individual at most one of the counting processes N_{i1} and N_{i2} associated with individual i can have a jump.

Korsgaard and Andersen (1998) and Petersen et al. (1996) extended the correlated frailty model to more general structures of genetic and environmental effects.

1.3.4 Normal random effects

In view of the transformation model (1.5) it is appealing to add random effects that act on the linear scale. A natural choice is the model that, conditionally on a mean zero multivariate normal effect $\mathbf{b}_i$, relates the failure time of individual j in cluster i to covariate vectors $\mathbf{X}_{ij}$ and $\mathbf{Z}_{ij}$ by

$$H(T_{ij}^*) = -\beta_0^T \mathbf{X}_{ij} + \mathbf{Z}_{ij}^T \mathbf{b}_i + \varepsilon,$$

with H and ε defined as in (1.5).

Unlike gamma frailties, normal random effects have unrestricted covariance matrices. This flexibility is a big advantage of this model. A restriction is that we have to rely on approximations or numerical methods when integrating out the normal random effects as these integrals cannot be written in closed form. When the dimension of the random effect is high and the sample size is large this can be rather computationally demanding.

Consider the simple case with a one-dimensional shared normal random effect b_i for clustered individuals and a conditional proportional hazards model. In this conditional hazard

$$\exp\left(\beta_0^T \mathbf{X}_{ij} + b_i\right) \lambda_0(t)$$

the random effect enters multiplicatively on the hazard as a log-normal frailty. This looks very much like the shared frailty model discussed above, but unlike these the log-normal frailty does not have mean 1 in the current formulation. Thus, it is not straightforward to compare the models.

2

Semiparametric inference

Counting processes and martingales have traditionally been the main tools when studying asymptotics in survival analysis. Let $N(t)$ denote a generic process counting the number of events that have occurred for some unit of interest up to time t. We can construct a counting process per individual (as we did in the previous chapter), or cluster, and another process counting the number of events for all individuals under study. The counting processes can be decomposed into a deterministic model part, the *compensator* A, and a random noise part M such that

$$M(t) = N(t) - A(t)$$

is a martingale. A martingale with respect to some filtration $\mathscr{F}_t$ is characterized by the relation

$$E\left[dM(t)|\,\mathscr{F}_{t-}\right] = 0$$

for all t. Many interesting quantities in survival analysis, such as score functions, can be written as stochastic integrals of the form

$$\int_0^t D(s)dM(s), \tag{2.1}$$

where D is a *predictable* stochastic process. Informally, the process D is predictable if the value $D(t)$ is known given the history $\mathscr{F}_{t-}$ just prior to time t. Integrals of the form (2.1) are, under some conditions, themselves martingales and asymptotic theory can often be established by Robelledo's martingale central limit theorem (Andersen et al., 1993, p. 83).

Martingale methods have an appealing conceptual foundation, but are not always applicable. In particular, if the integrand in (2.1) is not predictable, then the integral is not a martingale. This is the case for example if D contains weights that depend on events that might not have occurred at time t.

Often (2.1) can alternatively be viewed as an empirical process and large sample properties then follow by modern empirical process techniques (van der Vaart and Wellner, 1996). There is a price to be paid for this however, as empirical processes in this setting pose the strong restriction of independence of sampling units (e.g. individuals or clusters), whereas martingales allow more complex dependencies on the past. Martingale techniques are applicable in instances where the censoring mechanism depends on what happened previously to any individuals or clusters, even though this set up is clearly non-i.i.d. We conclude that none of the methods can fully replace the other. We refer to Andersen et al. (1993) and Aalen et al. (2008) for further reading on martingale methods in event history analysis. In Section 2.3.1 we briefly introduce some key concepts from empirical process theory.

In this chapter we investigate two methods for handling the infinite dimensional parameter $\Lambda_0(t) = \int_0^t \lambda_0(s)ds$ in a semiparametric frailty model. Both methods estimate Λ_0 by a step function with positive jumps at uncensored event times only. With the first method we reduce the problem to finite dimensions by profiling out Λ_0, while the latter involves joint maximization of the finite dimensional parameter and the jump sizes of the cumulative baseline hazard. We outline the estimation procedures for a conditional proportional hazards model in a clustered survival setting, but the methods apply more generally. We then briefly introduce a few key theorems that are useful for establishing the asymptotic properties of the two estimators.

2.1 Recursive estimating equations

Consider the right-censored and clustered survival setting of Section 1.3.1. Assume that the intensity of the basic counting process $N_{ij}(t)$ associated with individual j of cluster i has a proportional hazards form

$$R_{ij}(t)V_i e^{\beta_0^T \mathbf{X}_{ij}} \lambda_0(t)$$

with respect to the unobserved conditional hazard $\mathscr{H}_t$. Let $\theta = (\beta, \gamma) \in \mathbb{R}^d$ denote the finite dimensional parameter, where γ pertains to the frailty distribution. We write subscript 0 for the true value of any parameter. The goal is to make inference for θ_0 and Λ_0.

From (1.3) and (1.7) the increment of N_{ij} at time t has expectation

$$\begin{aligned} E\left[dN_{ij}(t)\middle|\mathscr{F}_{t-}\right] &= E\left[E\left[dN_{ij}(t)\middle|\mathscr{H}_{t-}\right]\middle|\mathscr{F}^i_{t-}\right] \\ &= E\left[R_{ij}(t)V_i e^{\beta_0^T \mathbf{X}_{ij}} \lambda_0(t)dt\middle|\mathscr{F}^i_{t-}\right] \\ &= R_{ij}(t)E\left[V_i\middle|\mathscr{F}^i_{t-}\right] e^{\beta_0^T \mathbf{X}_{ij}} \lambda_0(t)dt, \end{aligned}$$

conditional on the observed filtration $\mathscr{F}_{t-}$. Thus

$$dM_{ij}(t) \quad = \quad dN_{ij}(t) - R_{ij}(t)E\left[V_i|\,\mathscr{F}^i_{t-}\right] e^{\beta_0^T \mathbf{X}_{ij}} \lambda_0(t) \tag{2.2}$$

are zero mean stochastic processes for all $i = 1, \ldots, n$, $j = 1, \ldots, m$.

Equation (2.2) suggests estimating Λ_0 for fixed θ by a Breslow-type step function. More specifically, let $\tau_1 < \tau_2 < \ldots < \tau_N$ denote the uncensored event times arranged in increasing order (assuming no ties). We estimate the jump size at time τ_l by

$$\Delta\hat{\Lambda}_n(\tau_l;\theta) \quad = \quad \frac{1}{\sum_{i=1}^n \sum_{j=1}^2 R_{ij}(t)E\left[V_i|\,\mathcal{F}^i_{\tau_{l-1}};\theta,\hat{\Lambda}_n\right] e^{\beta^T \mathbf{X}_{ij}}},$$

where the conditional expectation is taken assuming parameter values $(\theta, \hat{\Lambda}_n)$. The estimator is recursive as it depends on $\hat{\Lambda}_n$, but only at times up to and including τ_{l-1}.

Assuming for the moment that Λ_0 is known, equation (1.3) again suggests estimating θ by solving for θ in the estimating equation

$$\mathbb{U}_n(\theta, \Lambda_0, \tau) \quad = \quad \mathbf{0}, \tag{2.3}$$

where

$$\mathbb{U}_n(\theta, \Lambda, t) = n^{-1} \sum_{i=1}^n \sum_{j=1}^m \int_0^t D(s;\theta,\Lambda) \left\{ dN_{ij}(s) - R_{ij}(s)E\left[V_i|\,\mathcal{F}^i_{s-};\theta,\Lambda\right] e^{\beta^T \mathbf{X}_{ij}} d\Lambda(s) \right\}$$

and $D(t;\theta,\Lambda)$ is some bounded d-dimensional vector valued predictable process. The true value of Λ_0 is unknown so we replace it by the estimator $\hat{\Lambda}_n$ in (2.3) in order to get the estimating equation

$$\mathbb{U}_n(\theta) \quad = \quad \mathbb{U}_n(\theta, \hat{\Lambda}_n, \tau) = \mathbf{0}. \tag{2.4}$$

We denote the solution to (2.4) by $\hat{\theta}_n$. By using $\mathbb{U}_n(\hat{\theta}_n) = 0$ and the mean value theorem we can make the usual linearization

$$\begin{aligned} \mathbb{U}_n(\theta^o) \quad &= \quad -\left(\mathbb{U}_n(\hat{\theta}_n) - \mathbb{U}_n(\theta_0)\right) \\ &\approx \quad -\nabla_\theta \mathbb{U}_n(\theta_0) \left(\hat{\theta}_n - \theta_0\right), \end{aligned} \tag{2.5}$$

where $\nabla_\theta \mathbb{U}_n(\theta_0)$ is the gradient of $\mathbb{U}_n(\theta)$ with respect to θ evaluated at θ_0.

Under appropriate conditions on the underlying distribution, the estimating function $\mathbb{U}_n(\theta_0, \Lambda_0, t)$ evaluated at the true values θ_0 and Λ_0 is a martingale

with respect to the observed filtration, or a sum of independent and identically distributed random variables. However, the asymptotic analysis is complicated by the fact that Λ_0 is estimated by the recursively constructed estimator. Based on the powerful and elegant product integration theory it is possible to show that $\mathbb{U}_n(\theta_0, \hat{\Lambda}_n, t)$ is asymptotically equivalent to a martingale, or a sum of independent and identically distributed variables. Normality then follows from the martingale central limit theorem or the classical central limit theorem. In Section 2.3.5 we review the theorem from product integration theory used here.

Arguments similar to those above were used to describe the asymptotics in a shared frailty proportional hazards model by Gorfine et al. (2009, 2006). Scheike et al. (2010) used the method for estimating haplotype effects in a proportional hazards model. Recently Martinussen et al. (2011) successfully used a similar technique within an additive Aalen gamma frailty framework.

2.2 Semiparametric maximum likelihood

Most efficient estimation approaches for semiparametric models are based on modifications of maximum likelihood estimators. Consider again a model with proportional hazards conditional on a frailty with density μ parametrized by γ. The likelihood is found by integrating out the random effect in the likelihood based on the conditional hazards as

$$\prod_{i=1}^{n} \int \prod_{j=1}^{m} \left(\mathbf{v}_i e^{\beta^T \mathbf{X}_{ij}} \lambda(T_{ij}) \right)^{\Delta_{ij}} \exp\left(-\mathbf{v}_i e^{\beta^T \mathbf{X}_{ij}} \Lambda(T_{ij}) \right) \mu(\mathbf{v}_i) d\mathbf{v}_i \tag{2.6}$$

If we restrict Λ to be absolutely continuous then a very high peak of λ at an uncensored event time would yield an arbitrarily large likelihood and there is no maximizer of the likelihood. Instead we maximize over all increasing right continuous functions and replace $\lambda(t)$ with the jump size at t, $\Delta\Lambda(t)$. The best choice among the discrete distributions are Λ that jump at the points T_{ij} with $\Delta_{ij} = 1$ only. This reduces the infinite dimensional problem to identifying jump sizes $\Delta\Lambda(T_{ij})$ that maximize the modified likelihood.

We obtain the nonparametric likelihood L_n by replacing $\lambda(T_{ij})$ by $\Delta\Lambda(T_{ij})$, $i = 1, \ldots, n$, $j = 1, \ldots, m$, in (2.6). The maximizer

$$(\hat{\theta}_n, \hat{\Lambda}_n) = \operatorname{argmax} L_n(\theta, \Lambda) \tag{2.7}$$

is referred to as the nonparametric maximum likelihood estimate (NPLME). Due to the complications resulting from the presence of the infinite dimensional parameter, what we treat as a likelihood here is not really a likelihood in the sense of products of densities. Thus, we need to verify that the NPLME indeed behaves like a maximum likelihood estimate, i.e. we wish to establish consistency, asymptotical normality and efficiency.

Murphy (1994, 1995) used empirical process theory to prove consistency, asymptotic normality and efficiency for the NPMLE in the shared gamma frailty model without covariates. Her work was generalized to the correlated gamma-frailty model allowing for covariates by Parner (1998). Many authors have since used similar techniques for various models. We will briefly outline the method of proof for the asymptotic normality. Please bear in mind that despite the common general scheme, the technical details can be very different from model to model. See Zeng and Lin (2007a, 2010) for a thorough exposition of NPLME for semiparametric transformation models.

To prove asymptotic normality of parametric maximum likelihood estimators we usually consider a system of estimating equations of the same dimension as the parameter. The solution is asymptotically normal if the system is appropriately differentiable. A semiparametric model would require infinitely many estimating equations. As shown by van der Vaart (1998, section 25.12) and van der Vaart (1999, Lecture 10), it turns out that we can proceed much in the same way as with a finite dimensional system, provided that we substitute functional analysis for multivariate calculus. The system is linearized in the estimators by a Taylor expansion around the true parameter, and the limit distribution involves the inverse of the derivative.

To set up the system of estimating equations, consider the set

$$\mathcal{H} = \{\mathbf{h} = (\mathbf{h}_\theta, h_\Lambda) : \mathbf{h}_\theta \in \mathbb{R}^d, h_\Lambda \in BV[0,\tau], \|\mathbf{h}\|_{\mathcal{H}} \leq 1\},$$

where $BV[0,\tau]$ is the class of real valued functions of bounded variation in $[0,\tau]$ and $\|\mathbf{h}\|_{\mathcal{H}} = \|\mathbf{h}_\theta\| + \|h_\Lambda\|_V$, where $\|h_\Lambda\|_V$ denotes the total variation of h_Λ in $[0,\tau]$. Define

$$\psi(\theta, \mathcal{A})[\mathbf{h}] = \mathbf{h}_\theta^T \ell_\theta(\theta, \Lambda) + \ell_\Lambda(\theta, \Lambda)[h_\Lambda], \tag{2.8}$$

where ℓ_θ is the score function for θ and ℓ_Λ is a score operator for Λ. The finite dimensional parameter can be perturbed in the usual way and $\mathbf{h}_\theta^T \ell_\theta$ is the ordinary score function for $\mathbf{h}_\theta^T \theta$ treating Λ as fixed. The operator ℓ_Λ is a little more involved. For each fixed (θ, Λ) and $h_\Lambda \in BV[0,\tau]$, $\ell_\Lambda(\theta, \Lambda)[h_\Lambda]$ corresponds to the score function for the one-dimensional submodel given by $\epsilon \mapsto (\theta, \int (1 + \epsilon h_\Lambda) d\Lambda)$ and can be found as the directional derivative of the log likelihood in the direction h_Λ. Each choice of $(\mathbf{h}_\theta, h_\Lambda)$ in (2.8) corresponds to an estimating equation for (θ, Λ).

We identify $(\hat{\theta}_n - \theta_0, \hat{\Lambda}_n - \Lambda_0)$, as a random element in $\ell^\infty(\mathcal{H})$, the space of bounded real valued functions on $\mathcal{H}$, by defining its value at $(\mathbf{h}_\theta, h_\Lambda)$ as $\mathbf{h}_\theta^T(\hat{\theta}_n - \theta_0) + \int h_\Lambda d(\hat{\Lambda}_n - \Lambda_0)$. Weak convergence will follow if we can verify the conditions of van der Vaart and Wellner (1996, Theorem 3.3.1) that is reviewed in Section 2.3.2.

When all parameters can be estimated at $n^{1/2}$ rate we may treat the NPMLE as a parametric log-likelihood with θ and the jump sizes of Λ at the

observed failure times as the parameters. The asymptotic covariance matrix of the NPMLEs for these parameters can be estimated by inverting the observed information matrix. Alternatively the covariance of the finite dimensional parameter θ may conveniently be estimated by semiparametric profile likelihood theory, see Section 2.3.3.

2.3 Inferential tools

2.3.1 Empirical processes

Consider a random sample $X_1, \ldots, X_n$ from a probability measure P on an arbitrary sample space $\mathcal{X}$. For a measurable function $f : \mathcal{X} \mapsto \mathbb{R}$, we write $\mathbb{P}_n f$ for the expectation of f under the empirical measure and Pf for the expectation under P,

$$\mathbb{P}_n f = \frac{1}{n} \sum_{i=1}^{n} f(X_i) \text{ and } Pf = \int f dP.$$

A class $\mathcal{F}$ of measurable functions $f : \mathcal{X} \mapsto \mathbb{R}$ is P-Glivenko-Cantelli if

$$\sup_{f \in \mathcal{F}} |(\mathbb{P}_n - P) f| \stackrel{a.s.}{\to} 0.$$

The *empirical process* evaluated at f is defined as $\mathbb{G}_n f = n^{1/2}(\mathbb{P}_n f - Pf)$. A class $\mathcal{F}$ of measurable real valued functions $f : \mathcal{X} \mapsto \mathbb{R}$ is P-Donsker if the sequence of processes $\{\mathbb{G}_n f : f \in \mathcal{F}\}$ converges weakly to a tight limit process in $\ell^\infty(\mathcal{F})$, the space of bounded functions on $\mathcal{F}$. The limit process $\{\mathbb{G} f : f \in \mathcal{F}\}$ is a zero mean Gaussian process with covariance $E[f(X)g(X)] - E[f(X)]E[g(X)]$ for $f, g \in \mathcal{F}$. $\mathbb{G}$ is known as the P-Brownian bridge.

Verifying that a class of functions is P-Glivenko-Cantelli or P-Donsker can be achieved by entropy calculations (van der Vaart and Wellner, 1996). Fortunately, we do not need to calculate entropy for each new problem as there are a number of methods to determine if a class is P-Donsker based on whether the class is built up of classes that are known to be P-Donsker. For example, if $\mathcal{F}$ and $\mathcal{G}$ are P-Donsker, then $\{f \wedge g : f \in \mathcal{F}, g \in \mathcal{G}\}$, $\{f \vee g : f \in \mathcal{F}, g \in \mathcal{G}\}$ and $\{f + g : f \in \mathcal{F}, g \in \mathcal{G}\}$ are also P-Donsker. Moreover, if $\mathcal{F}$ and $\mathcal{G}$ are bounded P-Donsker, then $\{fg : f \in \mathcal{F}, g \in \mathcal{G}\}$ is P-Donsker (van der Vaart and Wellner, 1996, Section 2.10). Furthermore, if $\mathcal{F}$ is P-Donsker, then it is also P-Glivenko-Cantelli.

2.3.2 A Z-theorem from van der Vaart and Wellner

A Z-estimator $\hat{\eta}_n$ is the approximate zero of a data-dependent function. Let the parameter space be H and let $\Psi_n : H \mapsto \mathbb{L}$ be a data dependent function between two normed spaces with norms $\|\cdot\|$ and $\|\cdot\|_{\mathbb{L}}$ respectively. If

$\|\Psi_n(\hat{\eta}_n)\|_{\mathbb{L}} \xrightarrow{P} 0$, then $\hat{\eta}_n$ is a Z-estimator. Usually Ψ_n is an estimator for some fixed function $\Psi : H \mapsto \mathbb{L}$ such that $\Psi(\eta_0) = 0$ for some parameter of interest $\eta_0 \in H$.

Arguments for proving asymptotic normality of semiparametric maximum likelihood estimator are often based on the following master theorem for Z-estimators.

Theorem 1 (van der Vaart and Wellner (1996), Theorem 3.3.1)**.** *Let Ψ_n and Ψ be random maps and a fixed map, respectively, from H into a Banach space $\mathbb{L}$ such that*

$$\frac{\|\sqrt{n}\,(\Psi_n - \Psi)\,(\hat{\eta}_n) - \sqrt{n}\,(\Psi_n - \Psi)\,(\eta_0)\|_{\mathbb{L}}}{1 + \sqrt{n}\,\|\hat{\eta}_n - \eta_0\|} \xrightarrow{P} 0,$$

and such that the sequence $\sqrt{n}\,(\Psi_n - \Psi)\,(\eta_0)$ converges in distribution to a tight random element Z. Let $\eta \mapsto \Psi(\eta)$ be Fréchet differentiable (van der Vaart and Wellner, 1996, Example 3.9.2) at η_0 with a continuously invertible derivative $P_0\dot{\Psi}_0$.

If $\Psi(\eta_0)$ and the random sequence $\hat{\eta}_n \in H$ satisfies $\sqrt{n}\Psi_n(\hat{\theta}_n) \xrightarrow{P} 0$ and $\|\hat{\eta}_n - \eta_0\| \xrightarrow{P} 0$, then

$$\left\|\sqrt{n}\dot{\Psi}_{\eta_0}(\hat{\eta}_n - \eta_0) + \sqrt{n}\,(\Psi_n - \Psi)\,(\eta_0)\right\|_{\mathbb{L}} \xrightarrow{P} 0.$$

2.3.3 Profile likelihood

Consider inference for the finite dimensional parameter θ in a semiparametric model with parameter (θ, Λ). Estimation of θ in a semiparametric model is more taxing, meaning that the information is worse, than under any parametric submodel. If the information for a *regular* estimator is equal to the minimum of the information over all efficient estimators for all parametric submodels, then the estimator is called semiparametric efficient. A parametric model which achieves this minimum, if such a model exists, is called a *least favorable submodel*. For a definition of a regular estimator we refer to van der Vaart (1999, Lecture 2) and settle for claiming that most commonly encountered estimators are regular. Nonparametric maximum likelihood generally yields semiparametric efficient estimators.

The semiparametric log profile likelihood is defined as the semiparametric log likelihood, but where the infinite dimensional component is profiled out,

$$pl_n(\theta) = \sup_{\Lambda} \log L_n(\theta, \Lambda). \tag{2.9}$$

By taking the supremum in two steps, we note that the maximizer of (2.9) is the first component of the NPMLE of θ.

Murphy and van der Vaart (2000) showed that under regularity conditions, the profile likelihood admits an expansion around the maximum likelihood estimator $\hat{\theta}_n$ of the form

$$\begin{aligned} \log pl_n(\tilde{\theta}_n) &= \log pl_n(\hat{\theta}_n) - \frac{1}{2} n(\tilde{\theta}_n - \hat{\theta}_n)^T \tilde{I}(\tilde{\theta}_n - \hat{\theta}_n) \\ &\quad + o_P(n^{1/2}\|\tilde{\theta}_n - \tilde{\theta}_n\| + 1)^2, \end{aligned}$$

where $\tilde{I}$ is the *efficient information* for estimating θ, for any $\tilde{\theta}_n \xrightarrow{P} \theta_0$.

The asymptotic expansion suggests that the semiparametric profile likelihood asymptotically can be treated much like an ordinary likelihood. In particular, under some conditions the maximum profile likelihood estimator is consistent, asymptotically normal and efficient. Differentiation of the profile likelihood yields consistent estimators of the efficient information matrix. A profile likelihood ratio statistic can be compared to percentiles of the χ^2 distribution to produce asymptotic hypothesis tests.

2.3.4 Weighted nonparametric maximum likelihood

The following development of Breslow and Wellner (2007) extends the ideas of the previous sections to data sets sampled in two phases. Typically the first phase sample contains incomplete information for a very large cohort. When using all subjects from the phase one sample is infeasible we can choose a subsample, the phase two sample, for further analysis. Based on the information from the first phase we might want to overrepresent subjects believed to hold more statistical information or otherwise ascertain enough subjects of specific characteristics. In the setting with routine registers, the first phase typically corresponds to the full register and the second phase to carefully selecting a subset from the register for further analysis.

Assume that the first phase sample consists of independent draws $X_1, ..., X_n$ from the probability distribution P on the sample space $\mathcal{X}$, and that the cohort is partitioned into S strata depending on information available in the phase one sample. Let $\xi_i = 1$, $i = 1, \ldots, n$, indicate whether observation i was included in the subsample of the second phase and let $\pi_i = P(\xi_i = 1)$. The probabilities π_i depend on stratum membership of observation i. Define

$$\mathbb{P}_n^\pi f = \frac{1}{n} \sum_{i=1}^{n} \frac{\xi_i}{\pi_i} f(X_i)$$

as the expectation of the measurable function $f : \mathcal{X} \mapsto \mathbb{R}$ under the inverse probability weighted (IPW) empirical measure. Define the IPW empirical process

$$\begin{aligned} \mathbb{G}_n^\pi &= \sqrt{n}\,(\mathbb{P}_n^\pi - P) \\ &= \sqrt{n}(\mathbb{P}_n - P) + \sqrt{n}(\mathbb{P}_n^\pi - \mathbb{P}_n). \end{aligned}$$

If the population proportion of stratum s members, v_s, is positive for $s = 1, \ldots, S$, then Breslow and Wellner (2007, section 4) implies that

$$\mathbb{G}_n^{\pi} \xrightarrow{\mathcal{L}} \mathbb{G} + \sum_{s=1}^{S} \sqrt{v_s \frac{1-p_s}{p_s}} \mathbb{G}_s$$

in $\ell^{\infty}(\mathcal{F})$, where $(\mathbb{G}, \mathbb{G}_1, \ldots, \mathbb{G}_S)$ is a vector of independent Brownian bridge processes, all indexed by a P-Donsker class $\mathcal{F}$. Specifically, $\mathbb{G}_s$ is a P_s-Brownian bridge process indexed by $\mathcal{F}$, where P_s denotes P conditional on membership of stratum s. Breslow and Wellner (2007, Proposition B.1) further states that if $\mathcal{F}$ is P-Donsker then $\mathcal{F}$ is P_s-Donsker on stratum s, $s = 1, \ldots, S$.

2.3.5 Product integrals

Ordinary integration is a generalization of summation. Similarly, product integration generalizes the taking of products. A product integral is a product of many terms most of them being one or very close to one. Suppose $\mathbf{K}(t)$ is a $p \times p$ matrix valued function of time t. Suppose further that $\mathbf{K}$ is right continuous with left hand limits (cadlag) and of bounded variation. Let I denote the identity matrix. The product integral of $\mathbf{K}$ over the interval $[0, t]$ is defined as

$$\prod_{(s,t]} (I + d\mathbf{K}(s)) = \lim_{max|t_i - t_{i-1}| \to 0} \prod (I + (\mathbf{K}(t_i) - \mathbf{K}(t_{i-1})))$$

where as always the limit is taken over a sequence of ever finer partitions $0 = t_0 < t_1 < \ldots < t_k = t$ of the time interval $[0, t]$. An extensive exposition on product integrals and their use in survival analysis can be found in Gill and Johansen (1990).

The product integral arises as the solution to Volterra integral equations.

Theorem 2 (Andersen et al. (1993), Theorem II.6.3)**.** *Let* $\mathbf{V}$, $\mathbf{W}$ *be* $k \times p$ *matrix cadlag functions. For given* $\mathbf{W}$*, the unique solution* $\mathbf{V}$ *of the inhomogenous integral equation*

$$\mathbf{V}(t) = \mathbf{W}(t) + \int_0^t \mathbf{V}(s-)\mathbf{K}(ds)$$

is

$$\mathbf{V}(t) = W(0) \prod_{u \in [0,t]} (I + d\mathbf{K}(u)) + \int_0^t \mathbf{W}(ds) \prod_{u \in (s,t]} (I + d\mathbf{K}(u)) .$$

3

Register data

Many epidemiological cohort studies investigate the effects of certain covariates for a relatively rare disease. When the disease of interest is rare, a large cohort is required in order to accumulate sufficiently many cases to provide informative conclusion about the covariate effects. This will usually require a long period of time and tend to be very expensive.

The use of existing routine administrative registers in epidemiological studies can cut total research costs considerably. In the Nordic countries there are several registers of high quality that can be linked by the unique personal identification number assigned to each permanent resident used across all registers.

The central population registers collect and update information received from several different sources, e.g. marriages and divorces, migration. Each individual can be linked to parents and children via the national multi-generation registers.

Causes of disease can be identified in the registers of disease. Examples include the national cancer registers and causes of death registers. The hospital discharge registers are collected from all public and private hospitals, and are based on inpatient care periods. The registers include information on the length of stay in the hospital, diagnoses and procedures during hospitalization. The medical birth registers include information on mother's background, maternal health during pregnancy and delivery, medical interventions and newborn's outcome up to the age of seven days. Since the medical birth registers are routinely combined with the central population registers and the cause-of-death registers, they are complete in terms of births and deaths. In studies of disease inheritance, adoption and twin registers are favorite sources of data.

As routine registers are becoming increasingly common worldwide, the possibilities to use administrative data in epidemiological research is expanding. So is the need for statistical methods analysing such data.

3.1 Cohort sampling designs

Standard use of regression models requires inclusion of covariate information on all individuals in a cohort even when only a small fraction of these actually experience the event of interest. As noted in the previous section, when studying rare diseases the cohorts must necessarily be large and an analysis based on the full cohort may demand unreasonable computer power resources or time. Thus, when working with routine register data, a study design allowing for estimation of covariate effects without having to collect all data on all members of the cohort is desirable.

When the disease of interest is rare, the contribution of non failures (controls), in terms of statistical power may be close to negligible compared to that of failures. Sampling designs that include only a portion of the controls and overrepresent cases may drastically reduce sample sizes but still be sufficient to give reliable answers of the questions of interest. There are two important classes of case-control sampling designs: nested case-control sampling and case-cohort sampling.

3.1.1 Nested case-control sampling

In a nested case-control design, one selects, whenever an event occurs, a typically small number of controls among those at risk. The set consisting of these controls together with the case is called the sampled risk set. Covariate information is collected on the individuals in the sampled risk sets, but are not needed for the other individuals in the cohort.

The selection of controls is done independently at the different event times, so that subjects may serve as controls for multiple cases, and cases may serve as controls for other cases that experienced an event when the case was at risk. A crucial assumption is that at any time we do not make use of any information on events in the future.

If a surrogate measure of the exposure of main interest is available for everyone, then this information can be incorporated into the sampling process so that we obtain a more informative sample of controls. This stratified nested-case control design is called counter-matching and is described in Langholz and Borgan (1995).

3.1.2 Case cohort sampling

Prentice (1986) proposed the case-cohort design under which one observes covariates for each individual from a random sample of the cohort, selected at the beginning of the study, and all individuals experiencing an event. In contrast to the nested-case control design the same individuals are used as controls at

all event times when they are at risk. Subjects are included in the subcohort with probability depending on whether or not they experience the event.

It is well known that one can improve the efficiency of the parameter estimates by stratifying according to the covariates of the members in the cohort. Borgan et al. (2000) present large sample results for stratified case-cohort estimators in the proportional hazards model. The asymptotic covariance matrix can be split into two components; the cohort covariance matrix and a covariance matrix due to sampling the subcohort from the full cohort.

4

Summary of Papers

Paper I: Semiparametric transformation models for clustered survival data from routine registers

In this work we propose inferential procedures that can considerably reduce the resources needed to analyse clustered survival data from routine registers. We sample from registers with unequal inclusion probabilities in order to achieve an informative subsample of a modest size, so that it can be analysed with reasonable resources. The sampling is performed in two stages and is similar to the stratified case-cohort design. When considering large registers, even if the cases are small in proportion they may be big in numbers and we might want to sample cases as well. This is readily achieved by our design.

The weights we use depend on stratum membership and are typically not determined until an individual experiences an event or is censored. Such weights are certainly not predictable and martingales are of no help. It turns out that the inverse probability weighted empirical process techniques of Breslow and Wellner (2007) are exactly what we need.

We consider the general class of semiparametric transformation models, allowing for competing risks, with clustering between individuals and causes induced by random effects. Consistency and asymptotic normality of the nonparametric maximum likelihood estimator in this model were derived by nonparametric maximum likelihood by Zeng and Lin (2010). We combine the work of Zeng and Lin (2010) and Breslow and Wellner (2007) and derive similar results for estimation based on two-phase sampled data. An asymptotic likelihood ratio test for testing hypothesized values of one or more regression parameters is also given.

We suggest consistent estimators of the asymptotic variance of the IPW maximum likelihood estimator. The variance is the sum of two components.

The first component is the usual variability of an estimator based on random sampling from an infinite population whereas the second component represents the additional variability from selecting only a subsample in the second phase.

We present an extensive simulation study to illustrate the performance of the methods. We also apply the procedure on real world data in two worked examples.

Paper II: A frailty model for paired competing risks survival data

Epidemiological studies of survival times of related individuals are typically complicated because multiple types of events occur and follow-up of some of the events is censored by the onset of the other events. Failure from other causes can only be treated as non-informative censoring if the causes are independent.

In this work we present a semiparametric estimator for paired individuals under the risk of competing causes, where dependencies of failure times within pairs and across causes are modelled by unobserved frailties. We estimate the regression parameters by a score-type function based on the observed cause specific hazards where the baseline hazard functions are profiled out, thus reducing the dimensionality of the score vector greatly. This approach has previously been used in a clustered but non-competing setting by Gorfine et al. (2006). The method possesses desirable properties, such as a non iterative procedure for estimating the cumulative hazard function and a direct consistent covariance estimator.

Large sample properties are derived using product integration theory. The estimator is shown to be consistent and asymptotically normal. We discuss subsampling from routine registers when an analysis based on the full cohort is intractable.

In a simulation study we illustrate the performance of the proposed method for simple models with an additive gamma frailty structure. The same models are also used for illustration on a real data set on prostate cancer in twins. The estimator was implemented as an R program written in C.

In contrast to the nonparametric maximum likelihood estimator of Zeng and Lin (2007b) as discussed in a clustered competing risks setting by Gorfine and Hsu (2011), estimation by the Newton-Rhapson algorithm is directly applicable. We avoid inverting a potentially large matrix when estimating the standard errors, even in situations when profile likelihood methods are not applicable (e.g. when subsampling from registers as discussed in the accompanying paper of this thesis). We conjecture that our model is readily adapted to handle left truncated data, as it is based on the observed intensities.

Bibliography

Aalen, O., Borgan, Ø., and Gjessing, H. (2008). *Survival and Event History Analysis: A Process Point of View*. Springer-Verlag.

Andersen, P., Borgan, Ø., Gill, R., and Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer-Verlag, New York.

Borgan, Ø., Langholz, B., Samuelsen, S., Goldstein, L., and Pogoda, J. (2000). Exposure stratified case-cohort designs. *Lifetime Data Analysis*, 6(1):39–58.

Breslow, N. and Wellner, J. (2007). Weighted likelihood for semiparametric models and two-phase stratified samples, with application to Cox regression. *Scandinavian Journal of Statistics*, 34:86–102.

Clayton, D. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*, 65(1):141–151.

Fine, J. and Gray, R. (1999). A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association*, 94(446):496–509.

Gill, R. and Johansen, S. (1990). A survey of product-integration with a view toward applications in survival analysis. *Annals of statistics*, 18:1501–1555.

Gorfine, M. and Hsu, L. (2011). Frailty-based competing risks model for multivariate survival data. *Biometrics*, 67(2):415–426.

Gorfine, M., Zucker, D., and Hsu, L. (2006). Prospective survival analysis with a general semiparametric shared frailty model - a pseudo full likelihood approach. *Biometrika*, 93(3):735–741.

Gorfine, M., Zucker, D., and Hsu, L. (2009). Case-control survival analysis with a general semiparametric shared frailty model: A pseudo full likelihood approach. *Annals of statistics*, 37(3):1489–1517.

Korsgaard, I. and Andersen, A. (1998). The additive genetic gamma frailty model. *Scandinavian Journal of Statistics*, 25(2):225–269.

Langholz, B. and Borgan, Ø. (1995). Counter-matching: A stratified nested case-control sampling method. *Biometrika*, 82(1):69–79.

Liang, K. and Zeger, S. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22.

Martinussen, T., Scheike, T., and Zucker, D. (2011). The Aalen additive gamma frailty hazards model. *Biometrika*, 98(4):831–843.

Murphy, S. (1994). Consistency in a proportional hazards model incorporating a random effect. *Annals of Statistics*, 22:712–731.

Murphy, S. (1995). Asymptotic theory for the frailty model. *Annals of Statistics*, 23:182–198.

Murphy, S. and van der Vaart, A. (2000). On profile likelihood. *Journal of the American Statistical Association*, 95:449–465.

Parner, E. (1998). Asymptotic theory for the correlated gamma-frailty model. *Annals of Statistics*, 26(1):183–214.

Petersen, J., Andersen, P., and Gill, R. (1996). Variance components models for survival data. *Statistica Neerlandica*, 50:193–211.

Prentice, R. (1986). A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika*, 73(1):1–11.

Prentice, R., Kalbfleisch, J., Peterson, A., Flournoy, N., Farewell, V., and Breslow, N. (1978). The analysis of failure times in the presence of competing risks. *Biometrics*, 34(4):541–554.

Scheike, T., Martinussen, T., and Silver, J. (2010). Estimating haplotype effects for survival data. *Biometrics*, 66(3):705–715.

van der Vaart, A. (1999). Semiparametric statistics. In *Ecole d'Ete de Probailites de St. Flour XXIX*, volume 1781 of *Lectures on probability theory and statistics*, pages 331–457. Springer.

van der Vaart, A. and Wellner, J. (1996). *Weak Convergence and Empirical Processes*. Springer-Verlag.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press.

Vaupel, J., Manton, K., and Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, 16(3):439–454.

Yashin, A., Vaupel, J., and Iachine, I. (1995). Correlated individual frailty. *Mathematical Population Studies*, 5:145–159.

Zahl, P. (1997). Frailty modelling for the excess hazard. *Statistics in medicine*, 16:1573–1585.

Zeng, D. and Lin, D. (2007a). Maximum likelihood estimation in semiparametric models with censored data. *Journal of the Royal Statistical Society B*, 69:507–564.

Zeng, D. and Lin, D. (2007b). Semiparametric transformation models with random effects for recurrent events. *Journal of the American Statistical Association*, 102:167–180.

Zeng, D. and Lin, D. (2010). A generalized asymptotic theory for maximum likelihood estimation in semiparametric regression models with censored data. *Statistica Sinica*, 20:871–910.